Quantitative Structure-Mutation-Activity Relationship Tests (QSMART) Model for Protein Kinase Inhibitor Response Prediction

Liang-Chin Huang[1], Wayland Yeung[1], Ye Wang[2], Huimin Cheng[2], Aarya Venkat[3], Sheng Li[4], Ping Ma[2], Khaled Rasheed[4], Natarajan Kannan[1,3*]

1 Institute of Bioinformatics, University of Georgia, Athens, GA, USA
2 Department of Statistics, University of Georgia, Athens, GA, USA
3 Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA
4 Department of Computer Science, University of Georgia, Athens, GA, USA

* nkannan@uga.edu

Abstract

Predicting drug sensitivity profiles from genotypes is a major challenge in personalized medicine. Machine learning and deep neural network methods have shown promise in addressing this challenge, but the "black-box" nature of these methods precludes a mechanistic understanding of how and which genomic and proteomic features contribute to the observed drug sensitivity profiles. Here we provide a combination of statistical and neural network framework that not only estimates drug IC_{50} in cancer cell lines with high accuracy ($R^2 = 0.861$ and RMSE = 0.818) but also identifies features contributing to the accuracy, thereby enhancing explainability. Our framework, termed QSMART, uses a multi-component approach that includes (1) collecting drug fingerprints, cancer cell line's multi-omics features, and drug responses, (2) testing the statistical significance of interaction terms, (3) selecting features by Lasso with Bayesian information criterion, and (4) using neural networks to predict drug response. We evaluate the contribution of each of these components and use a case study to explain the biological relevance of several selected features to protein kinase inhibitor response in non-small cell lung cancer cells. Specifically, we illustrate how interaction terms that capture associations between drugs and mutant kinases quantitatively contribute to the response of two EGFR inhibitors (afatinib and lapatinib) in non-small cell lung cancer cells. Although we have tested QSMART on protein kinase inhibitors, it can be extended across the proteome to investigate the complex relationships connecting genotypes and drug sensitivity profiles.

Introduction

Protein kinases are a class of signaling proteins, greatly valued as therapeutic targets for their key roles in human diseases, such as cancer [1]. For decades, chemotherapy has served as part of a standard set of cancer treatments; however, the resistance of cancer cells to chemotherapy is still a major clinical challenge [2]. Mutations in protein kinase are known to play important roles not only in drug resistance [3] but also in drug sensitivity [4]. Depending on the structural location, mutations can have varying impacts on drug sensitivity. For example, non-small cell lung cancer (NSCLC) cells

harboring either the EGFR T790M or L858R mutation respectively leads to resistance or hypersensitivity to the cancer drug gefitinib [5, 6], while those with EGFR T790M/L858R double mutant are only resistant to gefitinib [7]. As mutations impact the efficacy of different cancer drugs, there is a need to incorporate structural knowledge in drug response prediction methods.

To facilitate the understanding of the molecular mechanisms that cause drug sensitivity and drug resistance in cancer cells, the Genomics of Drug Sensitivity in Cancer (GDSC) Project [8] recently screened the drug responses of 266 anti-cancer drugs against ∼1,000 human cancer cell lines and provided the largest publicly available drug response dataset. Moreover, to broaden the pharmacologic annotation for human cancers, the Cancer Cell Line Encyclopedia [9] (CCLE) provided pharmacologic profiles for 24 drugs across 504 cancer cell lines. By utilizing these datasets, several prediction models were built to pursue a more accurate drug response estimation by different types of approaches, from traditional statistical models, network-based models, to the recent machine learning methods and state-of-the-art neural networks (Table 1). These approaches include (1) statistical models: MANOVA [10] and generalized linear models (regularization: ridge [11–14], elastic net [11–13, 15], Lasso [11–13], and mixture [16]), (2) network-based models [17–24], (3) random forests [25, 26], (4) support vector machine (SVM) [21, 27, 28] and other kernelized methods [29–31], and (5) neural networks: artificial neural network (ANN) [32], convolutional neural network (CNN) [33–35], recurrent neural network (RNN) [35], and other deep neural networks (DNN) [36, 37].

Over the years, new techniques continue to emerge and the samples of drug response have increased constantly; nevertheless, existing prediction models still cannot achieve high performance to realize "precision" medicine goals. The prediction performances measured by the coefficient of determination (R^2) are in the range from 0.25 to 0.78. More recently, deep neural networks with multiple hidden layers such as CDRscan [33], tCNNS [34], and MCA [35] have been proposed that achieve R^2 higher than 0.8 (R^2 = 0.84, 0.83, and 0.86, respectively). However, most of the cancer cell line features used in previous studies were based on gene expression profiles and did not explicitly consider associations between drugs and the structural location of mutations (Table 1). Consequently, the molecular mechanisms of drug-protein interactions cannot be inferred from these models, and thus these models have limited explainability.

The trade-off between prediction performance and explainability is an issue not only for CDRscan, tCNNS, and MCA but also for other existing machine learning approaches, thus the Defense Advanced Research Projects Agency (DARPA) recently launched the Explainable Artificial Intelligence (XAI) program [38] to facilitate building explainable models while maintaining prediction performance. In recognition of the interest in building explainable AI models, we built the Quantitative Structure-Mutation-Activity Relationship Tests (QSMART) model by (1) introducing interaction terms that capture drug-mutation relationships into quantitative structure-activity relationship (QSAR) model and using traditional statistical tests to identify significant interaction terms, and (2) utilizing a feature selection method to obtain highly informative interactions (Fig 1). Note that these interactions are statistical terms that cannot be interpreted as physical interactions directly. These two steps are equivalent to moving two hidden layers outside the neural networks "black box" for increasing the explainability of the prediction model. When applied on a subset of protein kinase inhibitors (PKIs), the QSMART achieves prediction accuracy comparable or better than state-of-the-art deep neural network methods (overall R^2 = 0.861, AUC = 0.981, and RMSE = 0.818). Although building fully explainable models is beyond the scope of this study, we show that some level of statistically significant relationships between drugs and mutated residues contributes to drug sensitivity in cancer cell lines.

Table 1. Current drug response prediction approaches.

Date	Author	Best model (Compared models)	Cancer cell line feature	Drug response		Validation	Performance
				GDSC	CCLE		
2013.04.30	Menden et al. [32]	ANN (RF)	MUT, CNV	✓		8-fold CV	$R^2 = 0.72$
2014.03.03	Geeleher et al. [11]	GLM	EXP	✓		LOOCV	AUC = 0.81
2015.01.01	Jang et al. [12]	GLM (PLS, SVM, PCA, RF)	MUT, EXP, CNV, CLS	✓	✓	5-fold CV	r = ∼0.5
2015.06.30	Dong et al. [27]	SVM	EXP		✓	10-fold CV	Accuracy = ∼0.8
2015.09.29	Zhang et al. [17]	Network (EN)	EXP	✓	✓	LOOCV	r = 0.6
2016.03.31	Gupta et al. [28]	SVM	MUT, EXP, CNV		✓	LOOCV	r = 0.78
2016.09.01	Ammad-ud-din et al. [29]	Kernel (GLM)	PWY	✓		5-fold CV	ρ = ∼0.22
2016.12.28	Nguyen et al. [10]	MANOVA (RF)	EXP	✓		10-fold CV	MCC = 0.18
2017.01.09	Stanfield et al. [18]	Network (Kernel)	MUT, PPI	✓	✓	LOOCV	AUC = 0.881
2017.07.15	Ammad-ud-din et al. [13]	GLM (PLS, SGL, RF, SVM)	EXP, PWY	✓		LOOCV	$\rho = 0.375$
2017.08.28	Geeleher et al. [14]	Ridge	EXP	✓		10-fold CV	$\rho = 0.48$
2017.09.12	Rahman et al. [25]	RF	EXP	✓	✓	3-fold CV	AUC = ∼0.3
2017.11.13	Ding et al. [36]	DNN (EN, SVM)	MUT, EXP, CNV	✓	✓	25-fold CV	AUC = 0.87
2018.03.08	He et al. [30]	Kernel (EN, Ridge, RF)	EXP	✓		3-fold CV	Precision = ∼0.35
2018.06.11	Chang et al. [33]	CNN (RF, SVM)	SNP	✓		5% leave-out	$R^2 = 0.843$
2018.07.01	Cichonska et al. [31]	Kernel	SNP, MET, EXP, CNV	✓		10-fold CV	r = 0.858
2018.09.14	Le et al. [19]	Network (Kernel)	MUT, EXP	✓	✓	5-fold CV	r = 0.804
2018.09.14	Juan-Blanco et al. [20]	Network	MUT, EXP		✓	LOOCV	AUC = ∼0.72
2018.10.10	Yang et al. [21]	Network + SVM (Kernel)	MUT, MET, CNV, PPI	✓		5-fold CV	AUC = 0.788
2018.12.07	Liu et al. [22]	Network	EXP	✓	✓	10-fold CV	r = 0.73
2019.01.22	Wei et al. [23]	Network	EXP	✓	✓	LOOCV	r = 0.63
2019.01.31	Wang et al. [15]	EN	EXP, PWY		✓	10-fold CV	MSE = ∼2.8
2019.01.31	Chiu et al. [37]	DNN (LR, SVM, PCA)	MUT, EXP	✓		10% leave-out	r = ∼0.86
2019.02.27	Li et al. [16]	Mixture (GLM, RF)	EXP		✓	20% leave-out	r = 0.882
2019.07.11	Lind et al. [26]	RF (SVM, ANN)	MUT	✓		5-fold CV	r = 0.86
2019.07.29	Liu et al. [34]	CNN (ANN)	MUT, CNV	✓		10% leave-out	$R^2 = 0.826$
2019.10.16	Manica et al. [35]	CNN + RNN (RF, SVM)	EXP, CNV, PPI	✓		5-fold CV	$R^2 = 0.86$
2019.11.04	Oskooei et al. [24]	Network (RF, LR)	EXP, PPI	✓		30-fold CV	r = ∼0.9

ANN: artificial neural network; AUC: area under the ROC curve; CCLE: Cancer Cell Line Encyclopedia; CLS: cancer classification; CNN: convolutional neural network; CNV: copy number variation; CV: cross-validation; EN: elastic net; EXP: gene expression; GDSC: Genomics of Drug Sensitivity in Cancer; GLM: generalized linear model, including ridge, elastic net, and lasso regression; DNN: deep neural networks; LOOCV: leave-one-out cross-validation; LR: linear regression; MCC: Matthews correlation coefficient; MET: methylation; MSE: mean squared error; MUT: gene-level mutation (i.e. whether the gene is mutated or not); PCA: principal component analysis; PLS: partial least squares; PPI: protein-protein interaction; PWY: pathway; r: Pearson correlation coefficient; R^2: coefficient of determination; RF: random forests; ρ: Spearman's rank correlation coefficient; RNN: recurrent neural network; SGL: sparse group lasso; SNP: single nucleotide polymorphism; SVM: support vector machine.

Results

Framework for protein kinase inhibitor response prediction

The overall objective of this study is to emphasize the contribution of adding interaction terms that capture drug-mutation relationships and to show how these interaction terms could help explain the mechanism of drug resistance/sensitivity. The framework we proposed in this study includes four main components: (1) the substructure fingerprints of protein kinase inhibitor (PKI) and cancer cell line's multi-omics data, including from low-level features, such as residue mutations, to high-level features, such as perturbed biological processes, (2) F-test for identifying significant drug-mutation relationships and other interaction effects, (3) a feature selection method: Lasso with Bayesian information criterion (BIC) control, and (4) a machine learning method to predict PKI response: neural networks (Fig 1). This framework has flexibility in adapting different materials and methods in each component. To implement this framework, we collected a dataset containing ∼0.2 million drug responses (IC_{50} in a logarithmic scale; "IC_{50}" hereinafter) from GDSC, split them into 23 sub-datasets according to the primary site where the cancer cell line originated, and then built a cancer-centric model for each

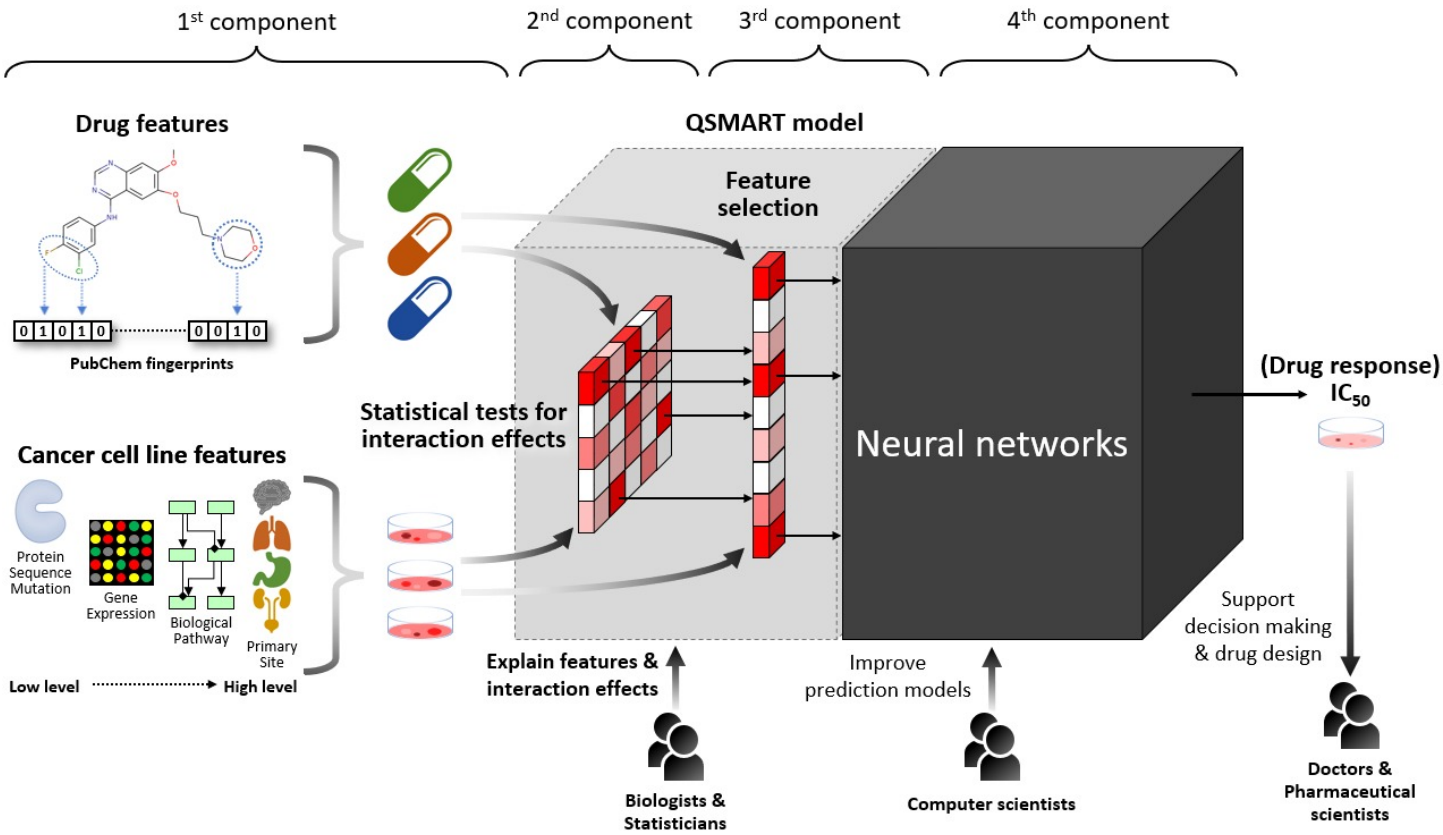

Fig 1. The framework of using the QSMART model with neural networks to predict protein kinase inhibitor response in cancer cell lines. Four main components of this framework: (1) drug features, cancer cell line features, and drug responses, (2) statistics tests for interaction terms, (3) a feature selection method for identifying highly informative features, and (4) a machine learning method for predicting drug response.

sub-dataset. The overall prediction performance of our proposed framework and the evaluation of each component are described below.

Overall performance of QSMART model with neural networks

For each cancer-centric model, Table 2 summarizes the number of PKI responses, the total number of features (including drug features, cancer cell line features, and interaction terms), the number of nodes in the first and second hidden layers of neural networks, and prediction performance (R^2). Additional measurements of prediction performance (RMSE and AUC), cancer cell line features at seven feature levels, interaction terms, and training iterations are shown in S1 Table. By using the QSMART model with neural networks, we can predict the PKI response in 23 cancer types with accuracies ranging from $R^2 = 0.805$ to 0.880. Fig 2a presents IC_{50} vs. predicted IC_{50} plot for all types of cancer cell lines (overall RMSE = 0.818 and $R^2 = 0.861$, which means these prediction models can explain 86% of the variation in PKI responses). Although we designed three types of neural network architectures in this study: single-layer, double-layer, and complex-double-layer (see Materials and methods), we found that the prediction models for all the 23 cancer types can achieve $R^2 > 0.8$ by using either single-layer or double-layer architecture. Based on Occam's razor principle [39], we chose the simplest single-layer architecture.

Residual analysis was then performed to assess the appropriateness of our prediction

Table 2. Prediction performances of using QSMART model with different machine learning methods.

Cancer type	#IC_{50}	#All Features	#Drug Features	#Cancer features		#Interactions		#Nodes		Performance (R^2)					
				Residue	Others	DxM	Others	1st	2nd	NN	RF	SVM	Lasso	ANOVA	MCA
AG	2,971	62	38	0	5	9	10	38	8	**0.815**	0.362	0.243	0.293	0.672	0.656
Bone	3,410	84	52	0	13	4	15	10	0	**0.856**	0.483	0.316	0.370	0.693	0.819
Breast	4,706	129	70	5	26	12	16	26	6	**0.880**	0.527	0.452	0.496	0.702	0.814
CNS	4,250	114	65	0	23	11	15	11	0	**0.858**	0.548	0.399	0.465	0.774	0.851
Cervix	1,044	37	29	0	3	1	4	7	0	**0.864**	0.552	0.389	0.455	0.809	0.824
Endometrium	1,073	33	21	0	4	4	4	11	4	**0.878**	0.358	0.279	0.328	0.769	0.832
Haematopoietic	4,204	119	58	3	24	28	6	11	0	**0.858**	0.518	0.378	0.429	0.679	0.807
Kidney	2,458	73	51	0	3	0	19	9	0	**0.836**	0.537	0.347	0.415	0.794	0.820
Large intestine	4,628	141	53	10	14	50	14	12	0	**0.814**	0.468	0.449	0.495	0.736	0.794
Liver	1,348	48	35	0	4	2	7	7	0	0.836	0.575	0.301	0.377	0.730	**0.859**
Lung (NSCLC)	9,205	207	72	7	35	47	46	15	0	**0.854**	0.466	0.470	0.513	0.728	0.819
Lung (others)	7,206	162	58	2	16	46	40	30	6	**0.859**	0.381	0.428	0.470	0.725	0.791
Lymphoid	13,302	291	72	54	30	86	49	18	0	**0.873**	0.449	0.448	0.495	0.758	0.834
Oesophagus	3,337	91	58	0	17	4	12	10	0	**0.841**	0.509	0.391	0.452	0.771	0.838
Ovary	3,502	113	64	2	18	9	20	11	0	**0.844**	0.532	0.471	0.522	0.741	0.810
Pancreas	2,421	84	60	0	7	0	17	10	0	**0.833**	0.591	0.419	0.492	0.784	0.816
Pleura	1,431	36	23	0	5	0	8	11	4	0.805	0.263	0.243	0.303	0.776	**0.837**
Skin	5,732	132	64	9	21	15	23	12	0	**0.875**	0.602	0.398	0.458	0.754	0.800
Soft tissue	1,938	63	45	0	10	2	6	8	0	**0.818**	0.540	0.333	0.404	0.758	0.786
Stomach	2,327	83	49	0	13	16	5	20	5	0.836	0.490	0.319	0.392	0.720	**0.842**
Thyroid	1,352	33	25	0	5	0	3	6	0	0.830	0.538	0.359	0.398	0.798	**0.853**
UAT	3,856	126	74	1	13	13	25	12	0	**0.869**	0.653	0.545	0.600	0.792	0.841
Urinary tract	1,454	68	47	0	5	9	7	9	0	**0.863**	0.558	0.344	0.433	0.754	0.847
Overall	87,155									**0.861**	0.496	0.429	0.460	0.755	0.823

The best performance for each cancer type is highlighted in bold. The performance of each machine learning method, except for ANOVA and MCA [35], is based on 10-fold cross-validation. The performance of MCA is based on its prediction for PKI response (S1 Data). AG: autonomic ganglia; ANOVA: analysis of variance; CNS: central nervous system; DxM: drug-mutation interaction term; MCA: multiscale convolutional attentive; NN: neural networks; NSCLC: non-small cell lung cancer; R^2: coefficient of determination; RF: random forests; SVM: support vector machine; UAT: upper aerodigestive tract; #IC_{50}: the number of drug responses; #Nodes: the number of nodes in the first and second hidden layers of neural networks.

models. The overall residual plot (Fig 2b) shows that there is no specific U shape, inverted U shape, or funnel shape, which means our prediction models need no more higher-order features to capture the variation of drug responses (S1 Fig shows residual plots for all the 23 cancer types). To further confirm the prediction model's ability to classify drug responses into two categories (sensitive vs. non-sensitive), we chose thresholds to define actual IC_{50} as sensitive or non-sensitive. Comparing to the single threshold used in a previous study [33] (IC_{50} = -2), we set multiple thresholds (-4, -3, -2, -1, and 0) and averaged the results to avoid overestimating the prediction performance. The resulting ROC curves for 23 cancer types and the overall curve are shown in Fig 2c. The overall AUC is 0.981 and comparable to a recent deep neural network-based study [33] (AUC $>$ 0.98). AUC for each cancer-centric model is available in S1 Table.

Contribution of different feature categories

To roughly estimate the contribution of different feature types to the prediction accuracy, we split the features into three different categories: drug features, cancer cell line features, and interaction terms. We used the same neural network architecture (the number of nodes in the first and second hidden layers) in each cancer-centric model, and then built prediction models by using the split feature sets. Across the 23 cancer types, this experiment showed that using drug features alone to predict PKI response outperformed using cancer cell line features or interaction terms alone (overall R^2 = 0.661, 0.126, and 0.152, respectively; Fig 2d). The contribution of interaction terms to prediction performance was significant (p-value = 0.0041, Wilcoxon signed-rank test) in comparison to cancer cell line features. Although it was partially due to the number of

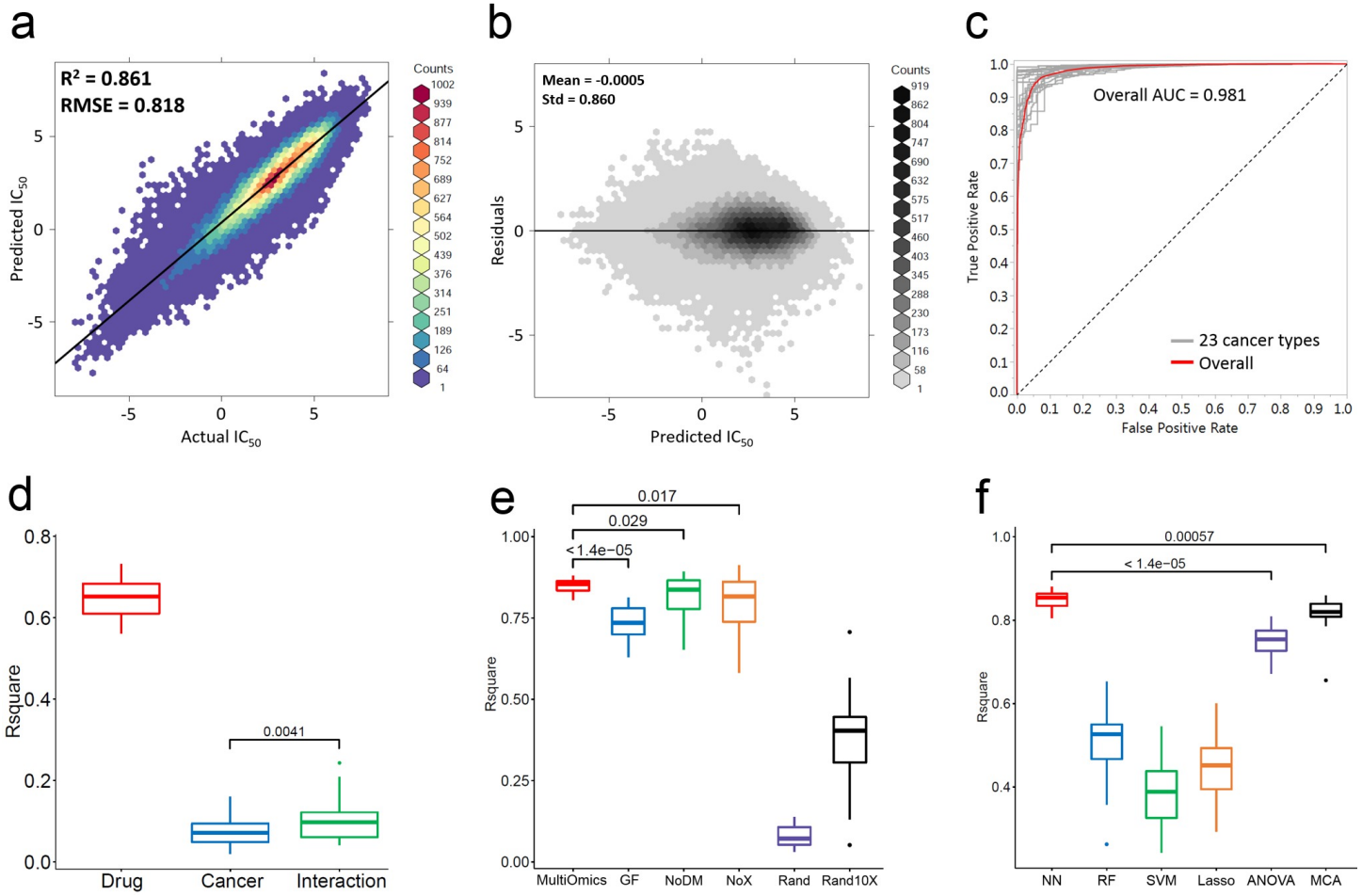

Fig 2. Prediction performances of different datasets and different prediction models. Wilcoxon signed-rank test is performed to compare prediction performances and the p-value is shown in each box plot. (a) Comparison between actual IC_{50} (x-axis) and the IC_{50} predicted by using QSMART with neural networks across all cancer types (y-axis). The hexbin plot is drawn by an R package "ggpubr" [40]. A fitted regression line is shown. (b) Residual analysis for the models using QSMART with neural networks across all cancer types. X-axis: predicted IC_{50}; y-axis: residuals, defined as actual IC_{50} minus predicted IC_{50}. (c) AUC curves of 23 cancer-centric models and an overall AUC. (d) The prediction performances of split QSMART models. (e) The prediction performances of using different datasets and different feature selection methods. GF: genomics fingerprints; NoDM: no drug-mutation interaction terms; NoX: no interaction terms; Rand: random selection; Rand10X: randomly select 10 times the number of features in the QSMART model. (f) The prediction performances of using different statistical or machine learning methods. NN: neural networks; RF: random forests; SVM: support vector machine; ANOVA: analysis of variance; MCA: multiscale convolutional attentive.

selected drug features being more than those of the other two feature categories, the main reason was that the drug features were more informative in cancer-centric models. Since the entire training dataset was split into 23 cancer-centric datasets, the similarity among cancer cell lines in one dataset was higher than the similarity among PKIs and thus the drug features had higher variation and higher entropy.

Assuming that the features from different categories in a full model are independent and can explain the variation of drug response independently, the summation of the prediction performances of split models (the R^2_{SSP} in Table 3) would ideally be the upper limit of a full model. However, Table 3 shows that the prediction performances R^2_{Full} are even higher than R^2_{SSP} for 14 cancer types, which implies that the

Table 3. Prediction performances of using split QSMART models with neural networks.

Cancer type	#Nodes		Split QSMART models						Performance comparison		
			Drug		Cancer cell line		Interaction		Full model	Split models	Difference
	1st	2nd	#Features	R^2_{Drug}	#Features	R^2_{Cancer}	#Features	$R^2_{Interaction}$	R^2_{Full}	R^2_{SSP}	R^2_{Full}-R^2_{SSP}
AG	38	8	38	0.611	5	0.044	19	0.041	0.815	0.696	0.119
Stomach	20	5	49	0.611	13	0.053	21	0.062	0.836	0.726	0.110
Breast	26	6	70	0.629	31	0.070	28	0.073	0.880	0.771	0.109
Pleura	11	4	23	0.614	5	0.043	8	0.061	0.805	0.718	0.088
Liver	7	0	35	0.652	4	0.020	9	0.078	0.836	0.751	0.086
Haematopoietic	11	0	58	0.599	27	0.092	34	0.098	0.858	0.789	0.070
Oesophagus	10	0	58	0.699	17	0.027	16	0.050	0.841	0.776	0.066
Soft tissue	8	0	45	0.561	10	0.100	8	0.104	0.818	0.765	0.053
CNS	11	0	29	0.683	3	0.072	5	0.055	0.858	0.810	0.048
Urinary tract	9	0	47	0.673	5	0.105	16	0.048	0.863	0.826	0.037
Lung (NSCLC)	15	0	72	0.610	42	0.084	93	0.128	0.854	0.822	0.031
Skin	12	0	64	0.685	30	0.041	38	0.122	0.875	0.848	0.027
Bone	10	0	52	0.607	13	0.111	19	0.112	0.856	0.830	0.026
Lung (others)	30	6	58	0.610	18	0.121	86	0.104	0.859	0.834	0.024
Pancreas	10	0	60	0.717	7	0.058	17	0.061	0.833	0.835	-0.002
Thyroid	6	0	25	0.713	5	0.067	3	0.053	0.830	0.833	-0.003
UAT	12	0	74	0.732	14	0.061	38	0.080	0.869	0.873	-0.004
Endometrium	11	4	21	0.709	4	0.076	8	0.099	0.878	0.884	-0.006
Ovary	11	0	64	0.648	20	0.092	29	0.122	0.844	0.861	-0.017
Kidney	9	0	51	0.666	3	0.074	19	0.126	0.836	0.866	-0.030
Lymphoid	18	0	72	0.661	84	0.097	135	0.149	0.873	0.907	-0.034
Cervix	7	0	65	0.669	23	0.033	26	0.244	0.864	0.946	-0.081
Large intestine	12	0	53	0.574	24	0.160	64	0.209	0.814	0.943	-0.129
Overall				0.661		0.126		0.152	0.861	0.940	-0.079

R^2_{Full}: the performance of using a full QSMART model with neural networks shown in Table 2; R^2_{SSP}: the sum of split model performances ($R^2_{SSP} = R^2_{Drug} + R^2_{Cancer} + R^2_{Interaction}$). AG: autonomic ganglia; CNS: central nervous system; NSCLC: non-small cell lung cancer; UAT: upper aerodigestive tract; #Nodes: the number of nodes in the first and second hidden layers of neural networks.

synergistic prediction performances (R^2_{Full} - R^2_{SSP}) are potentially derived from the higher-order interactions performed by neural networks. Interestingly, we found that the neural network architectures of the models with the top four synergistic effects are all double-layer neural networks, instead of single-layer neural networks, which also supports our hypothesis that the synergistic prediction performance is derived from higher-order interactions.

Informative features for predicting PKI response: multi-omics data

To evaluate the first component of the framework in this study – drug features and cancer cell line's multi-omics data – referring to the features used in a previous study [33] which achieves a high prediction accuracy, we replaced our drug features and cancer cell line features with 3,072 molecular fingerprints and 44,364 genomic fingerprints (S2 Fig). According to the previous study's method, drug fingerprints were generated by PaDEL-descriptor [41] (a software to calculate molecular descriptors and fingerprints) and genomic fingerprints (genomic mutation positions) were obtained from COSMIC Cell Lines Project [42]. To make them comparable with our models, we used the same feature selection method to prioritize all the drug fingerprints and genomic fingerprints, selected the same total number of features in our model for each cancer type (shown in Table 2), and then used the same neural network architectures. The number of selected features and prediction performances are shown in S2 Table. The box plot in Fig 2e shows that the performance distribution of 23 cancer-centric models using multi-omics data is significantly higher than that of the models using genomic

fingerprints (overall $R^2 = 0.861$ vs. 0.752, p-value $< 1.4e\text{-}05$, Wilcoxon signed-rank test). Under the same sample size, the number of features, and neural network architectures, our experiment shows that compared to genomic fingerprints, multi-omics data are more informative features for predicting PKI response.

Contribution of interaction terms in the QSMART model

To evaluate the second component of the framework in this study – F-test for identifying significant interaction terms – we performed two experiments: removing drug-mutation interaction terms and removing all interaction terms. We utilized the feature selection method to prioritize all input features, selected the same total number of features in the original models shown in Table 2, and then used the same neural network architectures to train the new models. The results of these two experiments are shown in S3 Table and S4 Table, respectively. The box plot in Fig 2e shows that the full QSMART models significantly outperform the models without drug-mutation interaction terms (overall $R^2 = 0.861$ vs. 0.839, p-value = 0.029) and the models without any interaction terms (overall $R^2 = 0.861$ vs. 0.823, p-value = 0.017). Interestingly, compared to the full QSMART models, we found that the prediction models of some cancer types, such as upper aerodigestive tract and breast, achieved higher performance without using any interaction terms. We conjectured that it was because some informative high-order interactions were captured inside the neural network black box and thus compensated the lack of interaction terms in the input layer. However, using neural networks cannot guarantee that these informative but unexplainable high-order interactions will be captured under the limited number of samples and the training iteration we used. This fact is reflected in Fig 2e, which shows that the prediction performance is variable when the drug-mutation interaction terms are not used ($R^2 = 0.653$ to 0.892) or all the interaction terms are not used ($R^2 = 0.581$ to 0.912). In this paragraph, we validate that adding interaction terms significantly contributes to PKI response prediction; in a non-small cell lung cancer case study described below, we will illustrate how the statistical interaction terms are interpreted as potential physical drug-target interactions and thus increase the explainability of prediction models.

Feature selection method: Lasso with BIC control

We next evaluated the feature selection methods. Since building neural networks using all the features from the first two components of the framework cannot be finished within a reasonable amount of time, instead of removing the third component of the framework, we randomly selected the same number of features in the original model for each cancer type shown in Table 2. Then, we trained a new neural network model with the same architecture for each cancer type. For each cancer type, the number of randomly selected features along with prediction performances are shown in S5 Table. It is not surprising that the prediction performances dropped to $R^2 = 0.031$ to 0.138 (overall $R^2 = 0.125$). To further evaluate the feature selection method we used, we increased the number of randomly selected features to 10 times the original number. The performances increased to $R^2 = 0.052$ to 0.707 (S6 Table; overall $R^2 = 0.378$). Nevertheless, these performances are significantly lower than those of the original models that utilize Lasso with BIC control (Fig 2e). Although the second component of the framework in this study had provided a lot of statistically significant interaction terms, compared to random selections, Lasso with BIC control identified highly informative features for PKI response prediction.

Comparison of machine learning methods used in our framework

We next compared neural networks with three commonly used machine learning models for drug response prediction: random forests, SVM, and Lasso regression. Based on the same feature set as inputs, neural networks significantly outperformed other machine learning approaches (Table 2 and Fig 2f; overall $R^2 = 0.861$, 0.496, 0.429, and 0.460 for neural networks, random forests, SVM, and Lasso regression, respectively). Furthermore, based on the feature sets used to validate the contribution of previous components in the framework, neural networks also outperformed random forests, SVM, and Lasso regression (S2 Table-S6 Table). A previous study [33] also showed that neural networks outperformed random forests and SVM ($R^2 = 0.843$, 0.698, and 0.562 for DNN, random forests, and SVM, respectively) in drug response prediction. Interestingly, neural network was only slightly better than Lasso in overall performance when randomly selected features were used as inputs ($R^2 = 0.125$ vs. 0.116, p-value = 0.015, Wilcoxon signed-rank test).

In addition to the three commonly used machine learning approaches, we compared our models with two-way ANOVA analysis and MCA [35]. Two-way ANOVA analysis was applied to assess the ability of two factors – drug and cancer cell line – to explain the variation of drug response. Drug IDs and cancer cell line IDs represented different levels of drug and cancer cell line, respectively. The result of two-way ANOVA showed that these two factors could explain 67.2% to 80.9% of the drug response variation in different cancer types (Table 2; overall $R^2 = 0.755$). Compared to ANOVA and MCA, using the QSMART model with neural networks had significantly higher ability to explain the PKI response variation for 23 cancer types (p-value $<$ 1.4e-05 and p-value = 0.00057, respectively; Fig 2f).

Case study: non-small cell lung cancer

We next evaluated the contribution of different features in drug response prediction using non-small cell lung cancer as a case study. All 207 features in the NSCLC-specific QSMART model and their descriptions are listed in S2 Data. We choose several pertinent features and explain their biological relevance in this case study to demonstrate how scientists may use our prediction model to explain their findings.

Drug feature

"From_Sanger". This feature was introduced into the model to distinguish the assays done by Massachusetts General Hospital (0) or Wellcome Sanger Institute (1). This feature represents the batch effects among the laboratory experiments performed by these two institutes. On average, the PKI responses obtained from Massachusetts General Hospital showed lower drug sensitivity (higher IC_{50} value) than those from the Wellcome Sanger Institute in the NSCLC dataset (average actual $IC_{50} = 2.88$ vs. 2.41). To investigate these experimental batch effects, we increased one unit to this feature and held other features constant. Although holding other features constant is not possible in reality, from the mathematical point of view, the result showed that if we replace 0 with 1 for From_Sanger, the average IC_{50} predicted by our pre-trained model will reduce 0.65 (average predicted $IC_{50} = 2.87$ vs. 2.22, S2 Data). Interestingly, this feature was selected not only in the NSCLC model but also in other 22 cancer-centric models, meaning the batch effects were significant across the assays done by these two institutes.

Biological processes interaction

"GO_0030324_X_GO_0048675". This feature represents the multiplication of the number of mutations perturbing the biological process "lung development" (Gene Ontology ID:

GO:0030324) and the number of mutations perturbing "axon extension" (Gene Ontology ID: GO:0048675). Axon initiation, extension, and guidance are known to play important roles in cancer invasion and metastasis [43]. In the NSCLC dataset, there are eight cell lines with mutations in protein kinases associated with axon extension; among them, NCI-H1944 and NCI-H2030 are from patients with metastatic NSCLC. On average, the NSCLC cell lines with this interaction showed higher PKI responses than those without this interaction (average actual IC_{50} = 4.32 vs. 2.69) and those involved in "lung development" or "axon extension" alone (average actual IC_{50} = 3.20 or 2.07, respectively). Based on our prediction model, every unit increase in this interaction term is associated with a 0.45 unit increase in IC_{50} on average (average predicted IC_{50} = 2.73 vs. 3.18).

Protein-protein interaction

The NSCLC model contains 27 protein-protein interaction (PPI) terms. Each PPI was quantified by the product of the participants' gene expression level. Every unit of gene expression level increase in these 27 PPIs is contributed to -0.089 to 0.061 unit increase in IC_{50} on average. Gene enrichment analysis of the 27 genes in the TP53-centric subnetwork shown in Fig 3 revealed the overrepresentation of pathways associated with angiogenesis, inflammation, apoptosis, and axon guidance (performed by PANTHER [44], S7 Table). MAP4K4, one of the genes involved in the apoptosis signaling pathway, is an emerging therapeutic target in cancer [45], and its over-expression is a prognostic factor for lung adenocarcinoma [46]. MAP4K4 expression is up-regulated upon binding by p53, and it will then activate the JNK signaling pathway to drive apoptosis [47]. In the NSCLC dataset, when the expression of MAP4K4-TP53 interaction increases, average IC_{50} is slightly decreased (Pearson correlation = -0.10); in our pre-trained PKI response prediction model, every unit of gene expression level increase in MAP4K4-TP53 PPI is associated with 0.012 unit decrease in IC_{50} on average (average predicted IC_{50} = 2.727 vs. 2.715).

CDK13, an understudied protein kinase defined by NIH Illuminating the Druggable Genome (IDG) program [48] (S3 Data, last updated on June 11, 2019), participates in a 4-clique PPI module in the TP53-centric subnetwork (Fig 3). Its three PPIs in this module are all the features of the NSCLC-specific model. One of CDK13's PPI partners, AKAP4, is a biomarker for NSCLC, and its expression increase was associated with tumor stage [49]. In addition to NSCLC, AKAP4 is also a potential therapeutic target of colorectal cancer [50] and ovarian cancer [51], and it regulates the expression of the CDK family. In the NSCLC dataset, the expression of CDK13-AKAP4 interaction had a weak positive correlation with IC_{50} (Pearson correlation = 0.07); in the prediction model, every unit of gene expression level increase in CDK13-AKAP4 PPI is associated with 0.017 unit increase in IC_{50} on average (average predicted IC_{50} = 2.727 vs. 2.744).

Drug-mutation interaction

In total, there are 47 drug-mutation interaction terms in the NSCLC model, and they are located at 22 PKA positions represented by spheres in Fig 4a (PDB ID: 1ATP). Their impacts on IC_{50} are listed in S8 Table, sorted by absolute IC_{50} impact. The drug-mutation relationships located in the canonical ATP-binding pocket (highlighted by a dashed rectangle in Fig 4a) could be formed by type I or type II protein kinase inhibitors according to the protein structure's active or inactive conformation, respectively [52].

Fig 4b and Fig 4c respectively illustrate different binding modes of two EGFR inhibitors in our dataset (afatinib and lapatinib) that contribute to variable response in L858R mutant EGFR (L858 corresponds to position 187 in protein kinase A; PKA_187).

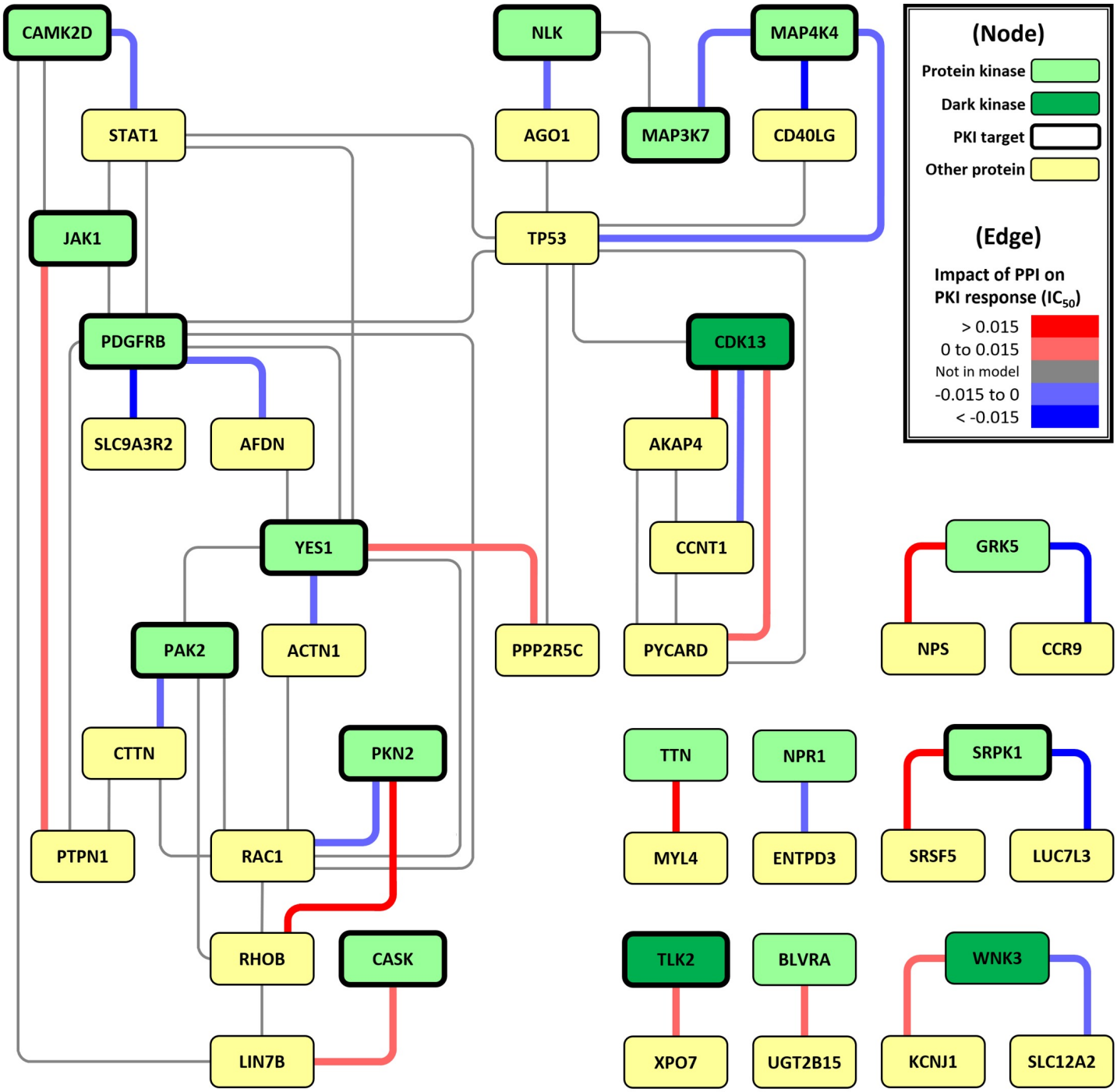

Fig 3. Protein-protein interaction network constructed by the interaction terms for predicting PKI response in NSCLC cell lines. Green node: protein kinase; dark green node: dark/understudied protein kinase; yellow node: other protein; the node with a thick border: known PKI target; red edge: the PPI with a positive impact on IC_{50}; light red edge: the PPI with a weak positive impact on IC_{50}; blue edge: the PPI with a negative impact on IC_{50}; light blue edge: the PPI with a weak negative impact on IC_{50}; gray edge: the PPI not in the prediction model.

H3255, an NSCLC cell line with EGFR L858R mutation, is hypersensitive to afatinib

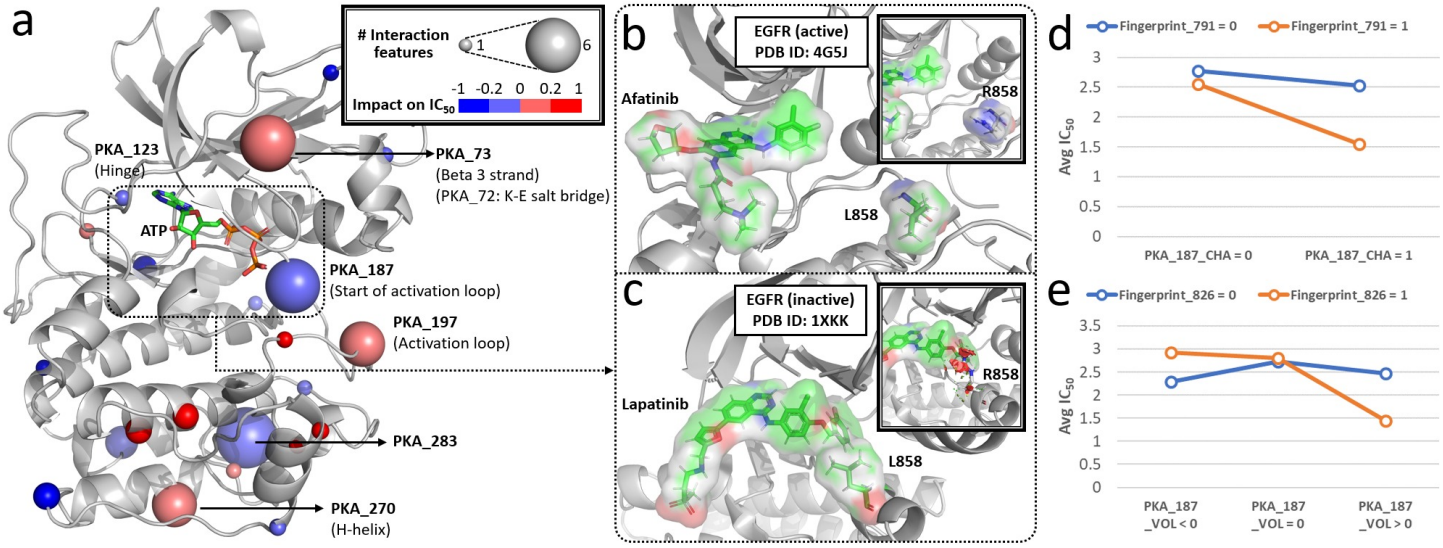

Fig 4. Distribution of drug-mutation relationships on the reference protein kinase A structure and interaction analyses. (a) Interaction hot spots are labeled and represented by larger spheres on the reference PKA structure (PDB ID: 1ATP, visualized by PyMol [53]). If a residue involved in multiple drug-mutation relationships, the median of their impacts on IC_{50} is chosen to represent the color of the sphere. Red sphere: the drug-mutation relationship with a positive impact on IC_{50}; light red sphere: the relationship with a weak positive impact on IC_{50}; blue sphere: the relationship with a negative impact on IC_{50}; light blue sphere: the relationship with a weak negative impact on IC_{50}. (b) and (c): Examples of two PKIs (afatinib and lapatinib) with different binding modes in the active (PDB ID: 4G5J) and inactive (PDB ID: 1XKK) conformations of EGFR crystal structures, respectively. The residue corresponding to PKA_187 – EGFR L858 – is labeled in each example; its arginine mutant form simulated by PyMol is illustrated. (d) and (e): Statistical interaction analyses for Fingerprint_791 vs. PKA_187_CHA and Fingerprint_826 vs. PKA_187_VOL in the NSCLC dataset, respectively.

treatment (IC_{50} = -4.35; average IC_{50} = 2.03 for all the NSCLC cell lines treated with afatinib). Notably, the L858R mutation can be accomodated in the active conformation of EGFR, but not in the inactive state due to steric hindrance [54].

An interaction analysis (Fig 4d) shows that the mutated residues involving charge difference at PKA_187 have significant interaction (p-value = 0.043, F-test) with Fingerprint_791, a drug substructure "NC1CCC(N)CC1" of afatinib). Based on our prediction model, every unit increase in the interaction term "PKA_187_CHA_X_Fingerprint_791", an interaction term with one of the highest impact on IC_{50} among all the drug-mutation interaction terms in the model (S8 Table), is associated with a 0.46 unit decrease in IC_{50} on average (average predicted IC_{50} = 2.73 vs. 2.27). Another interaction analysis (Fig 4e) shows that the mutated residues involving volume difference at PKA_187 have significant interaction (p-value = 0.035, F-test) with Fingerprint_826 (a drug substructure "OC1C(N)CCCC1" of afatinib). Every unit increase in "PKA_187_VOL_X_Fingerprint_826" is associated with a 0.01 unit decrease in IC_{50} on average (average predicted IC_{50} = 2.73 vs. 2.72). Since lapatinib lacks both substructures Fingerprint_791 and Fingerprint_826, we speculate that mutant EGFR in NSCLC cells with a larger, positively charged mutation at PKA_187 are resistant to lapatinib (the blue lines in Fig 4d and Fig 4e).

Discussion

In this study, we propose a PKI response prediction framework to accurately estimate IC_{50} values with a more explainable AI model. This framework includes four

components: (1) drug features, cancer cell line's multi-omics data, and PKI responses, (2) statistical tests for capturing interaction effects, (3) feature selection, and (4) neural networks. We validate the contribution of each component, showed high prediction performances, and used NSCLC dataset as a case study to explain several features. We systematically investigate the contributions of various interaction effects (such as protein-protein interactions, pathway-pathway interactions, and drug-mutation relationships) to drug response prediction.

The intrinsic limitation of drug response prediction is the unexplainable variation of drug response caused by different assays and experimental conditions. Currently, GDSC and CCLE are the two main sources for studying cancer drug response. Several previous studies about predicting drug response used data not only from GDSC but also from CCLE (Table 1). However, a previous study [20] pointed out that although GDSC and CCLE datasets shared 343 cancer cell lines and 15 drugs, the drug responses from these two datasets were poorly correlated. Thus, we chose to only use a single source in this study to minimize the unexplainable effect from different experimental environments. Nevertheless, this situation impeded us from finding appropriate independent testing set outside the GDSC data. Even though the drug response data we used were only from GDSC, the feature selection process showed that the drug feature "From_Sanger" was selected for all the 23 cancer-centric prediction models, meaning the batch effects were significant across the assays. While our studies were underway, we noticed that GDSC 8.0 was released. Compared with release 7.0, it contains 160 thousand more drug responses. However, this dramatic increase did not provide us a syncretic testing set since the old drug response dataset (called GDSC1 in release 8.0) and the new drug response dataset (called GDSC2) were generated based on different types of assays. Although the drug responses measured by different assays seem to have high correlation ($R = 0.838$ in Pearson correlation coefficient), unfortunately, it implies that even if we train a perfect model for GDSC1, the performance of predicting the drug responses in GDSC2 as an independent testing set would only be $R^2 = 0.702$ (S3 Fig panel a). Moreover, if we only focus on PKI responses between the two datasets, the correlation is reduced to 0.774 and $R^2 = 0.599$ (S3 Fig panel b). Furthermore, if we use our pre-trained models to predict the PKI response in GDSC2, the overall performance drops to $R^2 = 0.556$ (S3 Fig panel c).

In the case study, we illustrated the possibility of interpreting statistical interaction terms into potential physical interactions. When we investigated the contribution of protein-protein interactions to drug response prediction, the original purpose of utilizing biological knowledge (such as known PPIs from STRING [55]) was to narrow down the huge search space (a matrix of 30,000 proteins by 30,000 proteins). Consequently, this additional information also helped us explain the biological role of these statistical interaction terms with preliminary evidence. On the contrary, when we investigated the contribution of drug-mutation relationships to drug response prediction, we explored all the interactions between drug features and the mutations at reference PKA positions. Although restricting the mutations to be in the region around ATP-binding pocket (from PKA_47 to PKA_188, defined by the Kinase-Ligand Interaction Fingerprints and Structures (KLIFS) database [56]) could increase the probability of finding physical interactions among those statistical interaction terms, we will lose the opportunity to explore the potential interactions between PKIs and the allosteric binding sites outside the ATP-binding pocket.

In conclusion, by integrating multi-omics data, utilizing the innovative QSMART model, and employing neural networks, we not only can accurately predict PKI responses in cancer cell lines but also increase the explainability behind our prediction models. Compared to traditional QSAR models, the QSMART model proposed in this study further introduces different types of interaction terms. While we demonstrate our

model in protein kinase binding, the QSMART model can be applied to other protein families, such as G protein-coupled receptors (GPCRs) and ion channels. Moreover, the concept of QSMART model can also be broadly applied to other types of interactions, such as the protein-protein interaction that we had demonstrated, drug-drug interaction, glycosyltransferase-donor analog interaction, gene-environment interaction, and agent-host interaction.

Materials and methods

Protein kinase inhibitor

We define small-molecule (molecular weight < 900 daltons) protein kinase inhibitors (PKIs) in GDSC (release 7.0) [8] from a variety of publicly available, manually curated drug-target databases, and experimental data. The list of human protein kinases in this study is defined by ProKinO (version 2.0) [57]. Drug-kinase associations were extracted from DrugBank (version 5.1.0) [58], Therapeutic Target Database (TTD, last accessed on September 15th, 2017) [59], Pharos (last accessed on May 15th, 2018) [60], and LINCS Data Portal (last accessed on May 15th, 2018) [61]. We define a drug as a PKI if it is annotated as an "inhibitor", "antagonist", or "suppressor" in the drug-kinase associations. We also include the PKIs in LINCS Data Portal if their controls are less than 5% in KINOMEscan® assays. Based on these criteria, we define 143 small-molecule PKIs out of the 252 unique screened compounds in GDSC (S4 Data).

Drug response

GDSC provides the half-maximal inhibitory concentration values (IC_{50}, on a logarithmic scale) for 224,202 drug-cancer cell line pairs of drug sensitivity assays. These assays were performed by either the Wellcome Trust Sanger Institute or Massachusetts General Hospital Cancer Center. In this drug response dataset, there are 12,509 duplicated drug-cancer cell line pairs derived from 16 duplicated drugs. We measured the Pearson correlation coefficient between the IC_{50} values of each duplicated drug. Only afatinib and refametinib showed a strong positive correlation ($r > 0.7$); their IC_{50} values were then merged by their weighted means [62]. Drug responses of all other duplicated drugs were excluded from our study as they may have been assayed under different experimental conditions. The resulting dataset of 197,459 non-redundant drug responses consists of 236 drugs and 1,065 cancer cell lines. After filtering out non-PKIs, 109,856 non-redundant drug responses consisting of 135 PKIs and 1,064 cancer cell lines remained.

Drug features

Drugs' 2D structures were obtained from PubChem in SDF format. The CDK Descriptor Calculator GUI (version 1.4.6) [63] generated 881 PubChem fingerprints and 286 chemical descriptors including constitutional, topological, electronic, geometric, and bridge descriptors. Observing high multicollinearity within features, we removed redundant features and implemented the variance inflation factor (VIF) criterion [64] to reduce multicollinearity (for more details, see the Feature screening section below). After filtering, 92 PubChem fingerprints and 0 chemical descriptors remained.

To compare our prediction performances with those in a previous study [33], we used the same methods to generate (1) fingerprints, (2) extended fingerprints, and (3) graph-only fingerprints by PaDEL-descriptor (version 2.21) [41] for each drug. In total, there are 3,072 binary descriptors as drug features in the compared training sets. The

relatively large, unfiltered set of drug features are only used for comparison purposes in our study.

Cancer cell line features

Using mutation profiles for each cancer cell line sample provided by COSMIC Cell Lines Project (v87) [42], we incorporate 7 categories of multi-omics data to quantify the differences between wild type and mutant protein kinases: (1) residue-level: reference protein kinase A (PKA) position (from ProKinO), mutant type, charge, polarity, hydrophobicity, accessible surface area, side-chain volume, energy per residue [65], and substitution score (BLOSUM62 [66]); (2) motif-level: sequence and structural motifs of protein kinase (from ProKinO); (3) domain-level: subdomain in protein kinase (from ProKinO) and functional domain (from Pfam v31.0 [67]); (4) gene-level: the number of mutations in the genes encoding protein kinases, gene expression (from GDSC), and copy number variation (from COSMIC); (5) family-level: protein kinase family and group (from ProKinO); (6) pathway-level: reaction, pathway (from Reactome [68], last accessed on May 15th, 2018), and biological process (from AmiGO [69], last accessed on May 15th, 2018); and (7) sample-level: microsatellite instability, average ploidy, age, cancer originated tissue type, and histological classification (from COSMIC and Cellosaurus [70]).

The formula for generating all cancer cell line features is shown in S9 Table. Residue-level features of a cancer cell line were extracted from COSMIC mutants labeled as "Substitution - Missense". These features were then calculated if the mutation position could be aligned to the reference PKA position. This choice is based on an assumption that, for all protein kinases, mutations at equivalent positions will have similar effects on drug response. We further used two different types of weights, conservation score (KinView [71] with Jensen-Shannon divergence calculation [72]) and gene expression, to estimate the different effects of the same mutant type aligned to the same PKA position from different protein kinases.

Based on mutation position, the values of motif-level or domain-level features were calculated if it occurs in a specific motif or domain and its mutation description is "Substitution - Missense" or in-frame INDELs (insertions and deletions) in COSMIC. All mutation types, except for "Substitution - coding silent" and "Unknown", were taken into account for calculating the values of gene-level and higher-level features. For missing data, we assigned "Neutral" for copy number variation and "Unknown" for microsatellite instability and gender. No imputation was implemented for missing age.

QSMART model

The Quantitative Structure-Mutation-Activity Relationship Tests (QSMART) model was developed based on the QSAR model a variety of interaction terms. First, we built a basic model with all drug features and cancer cell line features as independent variables for estimating IC_{50}:

$$IC_{50} = \beta_0 + \sum_{i=1}^{I} \beta_{1i} D_i + \sum_{j=1}^{J} \beta_{2j} C_j + \epsilon, \tag{1}$$

where β_0 is the intercept, β_{1i} and β_{2j} represent the coefficients of the ith drug feature D_i and the jth cancer cell line feature C_j, and ϵ is the error term.

Because the residue-level features of a cancer cell line represent the mutation status in the reference PKA structure and we are interested in investigating drug-mutation

relationships, we introduced drug-mutation interaction terms into the model:

$$IC_{50} = \beta_0 + \sum_{i=1}^{I} \beta_{1i} D_i + \sum_{j=1}^{J} \beta_{2j} C_j + \sum_{i=1}^{I} \sum_{k=1}^{K} \beta_{3ik} D_i M_k + \epsilon, \tag{2}$$

where β_{3ik} is the coefficient of the interaction term formed by the ith drug feature D_i and the kth residue-level feature M_k. Since all cancer cell line features contain residue-level features and other six feature categories, $\{C_1, ..., C_J\}$ is a superset of $\{M_1, ..., M_K\}$. Considering that the interaction terms formed by the substructures of drug and high-level cancer cell line features have no biological relevance, we did not incorporate all cancer cell line features as part of interaction terms. For example, we did not consider the interaction between a substructure "Fingerprint_1" and a biological process "lung development" because it is unexplainable.

In addition to using all cancer cell line features, we further introduced more types of interaction terms into the full QSMART model to capture the environment of a cancer cell line:

$$\begin{aligned} IC_{50} = \beta_0 + \sum_{i=1}^{I} \beta_{1i} D_i + \sum_{j=1}^{J} \beta_{2j} C_j + \sum_{i=1}^{I} \sum_{k=1}^{K} \beta_{3ik} D_i M_k + \\ \sum_{p=1}^{P} \beta_{4p} PPI_p + \sum_{q=1}^{Q} \beta_{5q} RECx_q + \sum_{r=1}^{R} \beta_{6r} PWYx_r + \sum_{s=1}^{S} \beta_{7s} GOx_s + \epsilon, \end{aligned} \tag{3}$$

where β_{4p}, β_{5q}, β_{6r}, and β_{7s} are the coefficients of the pth protein-protein interaction PPI_p, the qth reaction-reaction interaction $RECx_q$, the rth pathway-pathway interaction $PWYx_r$, and the sth biological processes interaction GOx_s, respectively. These four types of interaction terms are formed by all pairs of protein, reaction, pathway, and biological process features, respectively. More details about interaction terms are described below.

Interaction terms

Five types of interaction terms were introduced into the QSMART model: drug-mutation interaction, protein-protein interaction, reaction-reaction interaction, pathway-pathway interaction, and biological processes interaction. These interactions were not necessarily physical interactions; instead, they were predictors that show statistically significant contribution to explaining the variation of IC_{50} values. For drug-mutation interaction term, only the residues mapping to the reference PKA structure were considered to form interactions with drugs. To reduce the search space, prior biological knowledge was used to filter interactions with less biological relevance. For protein-protein interaction (PPI), we retain PPIs with scores higher than 700 in the STRING database [55]. Gene expression level was used as a weight for PPIs to roughly represent the protein abundance in cancer cell lines. For reaction, pathway, and biological processes interactions, we removed the interactions formed by two entities from the same biological process/pathway hierarchy. For instance, the interaction between the biological process "lung cell differentiation" (GO:0060479) and its parent "lung development" (GO:0030324) was removed since it is unexplainable. Each interaction term was tested individually by F-test using R (version 3.4.4) [73]. Significant interaction terms (FDR < 0.05) with no less than 30 non-zero values were taken for further feature selection.

Datasets

To reduce more potential sources of noise and bias, we further filter cancer cell lines from the PKI response dataset if (1) their mutation profiles are not detected by whole-genome sequencing,(2) they have less than 30 drug response entries, (3) their gene expression profile is not available, or (4) their mutation site does not map to a residue in the reference PKA position. The dataset was then split into 29 groups, stratified by cancer primary site. Groups with less than 1,000 responses (adrenal gland, biliary tract, placenta, prostate, salivary gland, small intestine, testis, and vulva) were excluded due to low statistical power. "Haematopoietic and lymphoid tissue", the largest group, was further divided into two subsets by primary histology: "haematopoietic neoplasm" and "lymphoid neoplasm". For the case study, we collected cancer cell lines for the non-small cell lung cancer (NSCLC) dataset from the lung cancer dataset if their histology subtype is adenocarcinoma, non-small cell carcinoma, squamous cell carcinoma, large cell carcinoma, giant cell carcinoma, or mixed adenosquamous carcinoma. Remaining lung cancer cell samples were classified as "lung (others)". We created cancer type-centric training sets by expanding the drug response dataset with drug features, cancer cell lines features, and significant interaction terms. Categorical data in the training sets were coded into dummy variables. As a result, we prepared 23 cancer type-centric training sets. The number of PKI response for each cancer type is shown in Table 2.

Feature screening

Observing high multicollinearity within the features in the first component of our prediction framework (Fig 1), we implemented the variance inflation factor (VIF) criterion [64] to remove highly correlated features. For the multiple regression model with f features, X_i $(i = 1, ..., f)$, the VIF for the ith feature can be expressed by:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{4}$$

where R_i^2 is the coefficient of determination of the regression between X_i and the remaining $f - 1$ features. $VIF_i > 5$ (i.e. $R_i^2 > 0.8$) is considered to be high collinearity [74] and X_i should be excluded from the model. We first prioritized drug features based on these rules: (1) the later PubChem fingerprint bit positions (complex patterns) have higher priorities than the earlier ones (simple elements), and (2) PubChem fingerprints have higher priorities than calculated chemical descriptors because fingerprints directly represent molecular substructures of the drug. Then, starting from higher priority features moving towards lower priority features, we implemented stepwise selection under VIF control.

Co-expressed genes in the same prediction model also exhibited collinearity. To address this issue, we also used the VIF criterion to filter co-expressed genes in each training set. We prioritize genes based on the Pearson correlation coefficient between their expression and IC_{50} values, then implemented stepwise selection under VIF control.

Feature selection

To combat the problem of p (the number of drug features plus cancer cell line features and interaction terms) >> n (the number of drug responses) in the training sets, we implemented Lasso [75] with Bayesian information criterion (BIC) [76] by an R package "HDeconometrics" [77] (the third component of our prediction framework in Fig 1) because Lasso is appropriate for estimating coefficients in high-dimensional space [78] and BIC provides an efficient approach to obtain the optimal lasso fit [79]. After feature

selection, the remaining number of selected features for each cancer type is shown in Table 2.

Neural network architecture

We built neural network models by using JMP® [80]. We designed three types of neural network architectures in this study: single-layer, double-layer, and complex-double-layer. The numbers of hidden layer nodes follow the geometric pyramid rule [81]. Given N input nodes, there are $\lceil N^{1/2} \rceil$ hidden nodes in a single-layer architecture. In a double-layer architecture, there are $\lceil N^{2/3} \rceil$ and $\lceil N^{1/3} \rceil$ hidden nodes in the first and second hidden layers, respectively. In a complex-double-layer architecture, there are N and $\lceil N^{1/2} \rceil$ hidden nodes in the first and second hidden layers, respectively. The nodes among the two layers are fully connected. Biases are introduced into the input and hidden layers. The activation function of every node in the neural network is a hyperbolic tangent function (TanH). Newton's method [82] is chosen as an optimizer by JMP.

To avoid overfitting, we implement 10-fold cross-validation, early stopping, and Lasso-style penalty function (absolute value penalty, i.e. L1 regularization [83]). Based on Occam's razor principle [39], we started from a single-layer model for each cancer type. If the performance (average R^2 of the intact validation sets across the 10 folds) is less than a threshold 0.8 in 200 iterations, we increased the iteration to 300; if the performance is still less than the threshold, we implemented a double-layer model for 200 iterations and so on until using a complex-double-layer model for 300 iterations. To increase the reproducibility of this study, fixed random seeds were assigned and all the codes for training and prediction models are available at https://github.com/leon1003/QSMART/.

Other machine learning models

We compared neural networks with three other prediction algorithms with 10-fold cross-validation: random forests [84], support vector machine (SVM) [85], and Lasso [75]. Random forests were implemented by WEKA (version 3.8.3) [86] with default settings ("maxDepth" = 0, "bagSizePercent" = 100). For each cancer type, the number of iterations was decided based on the iterations used for each of the pre-trained neural network models (200 or 300 iterations) shown in S1 Table. SVM was implemented by the SMOreg function (SVM for regression) of WEKA with default kernel ("PolyKernel") and optimizer ("RegSMOImproved") settings. Lasso was implemented by an R package "glmnet" [87] with the default parameter setting for Lasso regression (alpha = 1 and family = "gaussian"). Additionally, we also compared our prediction models with two-way ANOVA analysis and MCA [35]. Because the purpose of two-way ANOVA analysis implemented by R was to quantify how much two factors (drug and cancer cell line) can explain the variation of drug response (adjusted R^2 was used), the model used the drug and cancer cell line identifiers as inputs and did not undergo 10-fold cross-validation. The performance of MCA shown in Table 2 is based on its prediction for PKI response (details are available in S1 Data).

Supporting information

S1 Fig. Residual analyses for 23 cancer-centric models and the overall result of using QSMART with neural networks.

S2 Fig. Genome-wide mutational status (genomic fingerprints) across the 23 cancer types.

S3 Fig. Comparison between GDSC1 and GDSC2 in the GDSC release 8.0.

S4 Fig. Prediction performances of using QSMART model with neural networks for different PKI target groups.

S1 Table. The number of features at different feature levels and the prediction performance of neural networks.

S2 Table. Prediction performances of using genomic fingerprints.

S3 Table. Prediction performances of using no drug-mutation interaction terms.

S4 Table. Prediction performances of using no interaction terms.

S5 Table. Prediction performances of using random feature selection.

S6 Table. Prediction performances of using random 10X feature selection.

S7 Table. Pathway enrichment analysis.

S8 Table. Drug-mutation relationships and their impact on IC_{50} in NSCLC cells.

S9 Table. Cancer cell line features.

S1 Data. MCA's performance for PKI response prediction.

S2 Data. Features for Lung (NSCLC) dataset.

S3 Data. Understudied proteins.

S4 Data. PKI target groups and PKI structures.

Acknowledgments

Funding for N.K. (R01GM114409 and U01CA239106) from the National Institutes of Health is acknowledged. Funding for P.M. (R01GM122080 and DMS-1903226) from NIH and NSF is acknowledged.

Author Contributions

L.H., P.M., K.R., and N.K. designed the research. L.H. performed data integration. L.H., Y.W., H.C., and P.M. performed statistical analyses. L.H., Y.W., H.C., S.L., and K.R. performed machine learning methods. L.H., W.Y., Y.W., H.C., P.M., and N.K. analyzed the data and interpreted the results. L.H., W.Y., A.V., and N.K. wrote the manuscript. Y.W., H.C., S.L., P.M., and K.R. revised the manuscript. L.H. created the tables and figures. All authors reviewed the manuscript and approved the final version of the manuscript.

References

1. Arslan MA, Kutuk O, Basaga H. Protein kinases as drug targets in cancer. Curr Cancer Drug Targets. 2006;6(7):623–634.
2. Lehne G, Elonen E, Baekelandt M, Skovsgaard T, Peterson C. Challenging drug resistance in cancer therapy–review of the First Nordic Conference on Chemoresistance in Cancer Treatment, October 9th and 10th, 1997. Acta Oncol. 1998;37(5):431–439.
3. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nat Rev Cancer. 2013;13(10):714–726.
4. Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer. 2007;7(3):169–181.
5. Bell DW, Gore I, Okimoto RA, Godin-Heymann N, Sordella R, Mulloy R, et al. Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. Nat Genet. 2005;37(12):1315–1316.
6. Tracy S, Mukohara T, Hansen M, Meyerson M, Johnson BE, Janne PA. Gefitinib induces apoptosis in the EGFRL858R non-small-cell lung cancer cell line H3255. Cancer Res. 2004;64(20):7241–7244.
7. Pao W, Miller VA, Politi KA, Riely GJ, Somwar R, Zakowski MF, et al. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med. 2005;2(3):e73.
8. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):D955–961.
9. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–607.
10. Nguyen L, Dang CC, Ballester PJ. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. F1000Res. 2016;5.
11. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014;15(3):R47.

12. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput. 2014; p. 63–74.

13. Ammad-Ud-Din M, Khan SA, Wennerberg K, Aittokallio T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics. 2017;33(14):i359–i368.

14. Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. Genome Res. 2017;27(10):1743–1751.

15. Wang X, Sun Z, Zimmermann MT, Bugrim A, Kocher JP. Predict drug sensitivity of cancer cells with pathway activity inference. BMC Med Genomics. 2019;12(Suppl 1):15.

16. Li Q, Shi R, Liang F. Drug sensitivity prediction with high-dimensional mixture regression. PLoS ONE. 2019;14(2):e0212108.

17. Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. PLoS Comput Biol. 2015;11(9):e1004498.

18. Stanfield Z, Co?kun M, Koyuturk M. Drug Response Prediction as a Link Prediction Problem. Sci Rep. 2017;7:40321.

19. Le DH, Pham VH. Drug Response Prediction by Globally Capturing Drug and Cell Line Information in a Heterogeneous Network. J Mol Biol. 2018;430(18 Pt A):2993–3004.

20. Juan-Blanco T, Duran-Frigola M, Aloy P. Rationalizing Drug Response in Cancer Cell Lines. J Mol Biol. 2018;430(18 Pt A):3016–3027.

21. Yang J, Li A, Li Y, Guo X, Wang M. A novel approach for drug response prediction in cancer cell lines via network representation learning. Bioinformatics. 2019;35(9):1527–1535.

22. Liu H, Zhao Y, Zhang L, Chen X. Anti-cancer Drug Response Prediction Using Neighbor-Based Collaborative Filtering with Global Effect Removal. Mol Ther Nucleic Acids. 2018;13:303–311.

23. Wei D, Liu C, Zheng X, Li Y. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. BMC Bioinformatics. 2019;20(1):44.

24. Oskooei A, Manica M, Mathis R, Martinez MR. Network-based Biased Tree Ensembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer. Sci Rep. 2019;9(1):15918.

25. Rahman R, Matlock K, Ghosh S, Pal R. Heterogeneity Aware Random Forest for Drug Sensitivity Prediction. Sci Rep. 2017;7(1):11347.

26. Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS ONE. 2019;14(7):e0219774.

27. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15:489.

28. Gupta S, Chaudhary K, Kumar R, Gautam A, Nanda JS, Dhanda SK, et al. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. Sci Rep. 2016;6:23857.

29. Ammad-Ud-Din M, Khan SA, Malani D, Murumagi A, Kallioniemi O, Aittokallio T, et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32(17):i455–i463.

30. He X, Folkman L, Borgwardt K. Kernelized rank learning for personalized drug recommendation. Bioinformatics. 2018;34(16):2808–2816.

31. Cichonska A, Pahikkala T, Szedmak S, Julkunen H, Airola A, Heinonen M, et al. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics. 2018;34(13):i509–i518.

32. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS ONE. 2013;8(4):e61318.

33. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci Rep. 2018;8(1):8857.

34. Liu P, Li H, Li S, Leung KS. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinformatics. 2019;20(1):408.

35. Manica M, Oskooei A, Born J, Subramanian V, Saez-Rodriguez J, Rodriguez Martinez M. Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. Mol Pharm. 2019;.

36. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res. 2018;16(2):269–278.

37. Chiu YC, Chen HH, Zhang T, Zhang S, Gorthi A, Wang LJ, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genomics. 2019;12(Suppl 1):18.

38. Gunning D, Aha DW. DARPA's Explainable Artificial Intelligence Program. AI Magazine. 2019;40(2):44–58.

39. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Occam's Razor. Inf Process Lett. 1987;24(6):377–380. doi:10.1016/0020-0190(87)90114-1.

40. Kassambara A. ggpubr:"ggplot2" based publication ready plots. R package version 01. 2017;6.

41. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–1474.

42. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941–D947.

43. Chedotal A, Kerjan G, Moreau-Fauvarque C. The brain within the tumor: new roles for axon guidance molecules in cancers. Cell Death Differ. 2005;12(8):1044–1056.

44. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129–2141.

45. Gao X, Gao C, Liu G, Hu J. MAP4K4: an emerging therapeutic target in cancer. Cell Biosci. 2016;6:56.

46. Qiu MH, Qian YM, Zhao XL, Wang SM, Feng XJ, Chen XF, et al. Expression and prognostic significance of MAP4K4 in lung adenocarcinoma. Pathol Res Pract. 2012;208(9):541–548.

47. Miled C, Pontoglio M, Garbay S, Yaniv M, Weitzman JB. A genomic map of p53 binding sites identifies novel p53 targets involved in an apoptotic network. Cancer Res. 2005;65(12):5096–5104.

48. the Druggable Genome I. Understudied Proteins; 2019. `https://commonfund.nih.gov/idg/understudiedproteins`.

49. Gumireddy K, Li A, Chang DH, Liu Q, Kossenkov AV, Yan J, et al. AKAP4 is a circulating biomarker for non-small cell lung cancer. Oncotarget. 2015;6(19):17637–17647.

50. Jagadish N, Parashar D, Gupta N, Agarwal S, Purohit S, Kumar V, et al. A-kinase anchor protein 4 (AKAP4) a promising therapeutic target of colorectal cancer. J Exp Clin Cancer Res. 2015;34:142.

51. Kumar V, Jagadish N, Suri A. Role of A-Kinase anchor protein (AKAP4) in growth and survival of ovarian cancer cells. Oncotarget. 2017;8(32):53124–53136.

52. Gavrin LK, Saiah E. Approaches to discover non-ATP site kinase inhibitors. MedChemComm. 2013;4(1):41–51.

53. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8; 2015.

54. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, et al. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell. 2007;11(3):217–227.

55. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447–452.

56. Kooistra AJ, Kanev GK, van Linden OP, Leurs R, de Esch IJ, de Graaf C. KLIFS: a structural kinase-ligand interaction database. Nucleic Acids Res. 2016;44(D1):D365–371.

57. McSkimming DI, Dastgheib S, Talevich E, Narayanan A, Katiyar S, Taylor SS, et al. ProKinO: a unified resource for mining the cancer kinome. Hum Mutat. 2015;36(2):175–186.

58. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–D1082.

59. Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res. 2018;46(D1):D1121–D1127.

60. Nguyen DT, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, et al. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017;45(D1):D995–D1002.

61. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Res. 2018;46(D1):D558–D566.

62. Jones DC, Hallyburton I, Stojanovski L, Read KD, Frearson JA, Fairlamb AH. Identification of a Î^{o}-opioid agonist as a potent and selective lead for drug development against human African trypanosomiasis. Biochem Pharmacol. 2010;80(10):1478–1486.

63. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr Pharm Des. 2006;12(17):2111–2120.

64. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. vol. 112. Springer; 2013.

65. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res. 1999;27(1):368–369.

66. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89(22):10915–10919.

67. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–D432.

68. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–D655.

69. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288–289.

70. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. J Biomol Tech. 2018;29(2):25–38.

71. McSkimming DI, Dastgheib S, Baffi TR, Byrne DP, Ferries S, Scott ST, et al. KinView: a visual comparative sequence analysis tool for integrated kinome research. Mol Biosyst. 2016;12(12):3651–3665.

72. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23(15):1875–1882.

73. Team RC. type [; 2014].

74. Sheather S. A modern approach to regression with R. Springer Science & Business Media; 2009.

75. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B. 1994;58:267–288.

76. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978;6(2):461–464.

77. R GF. HDeconometrics: Implementation of several econometric models in high-dimension; 2016.

78. Bhlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. 1st ed. Springer Publishing Company, Incorporated; 2011.

79. Zou H, Hastie T, Tibshirani R, et al. On the "degrees of freedom" of the lasso. The Annals of Statistics. 2007;35(5):2173–2192.

80. Sall J, Stephens ML, Lehman A, Loring S. JMP start statistics: a guide to statistics and data analysis using JMP. Sas Institute; 2017.

81. Masters T. Practical Neural Network Recipes in C++. San Diego, CA, USA: Academic Press Professional, Inc.; 1993.

82. Deuflhard P. Newton methods for nonlinear problems: affine invariance and adaptive algorithms. vol. 35. Springer Science & Business Media; 2011.

83. Ng AY. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In: Proceedings of the Twenty-first International Conference on Machine Learning. ICML '04. New York, NY, USA: ACM; 2004. p. 78–. Available from: `http://doi.acm.org/10.1145/1015330.1015435`.

84. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.

85. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273–297. doi:10.1023/A:1022627411411.

86. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques. 4th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2016.

87. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.