生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.210676

Jun. 25, 2022, 38(6): 2201-2212

·医药生物技术·

基于多组学数据的肿瘤药物敏感性预测

杨晨雨[1,2]，刘振浩[2,3]，代培斌[4]，张钰[1,2]，黄鹏杰[1,2]，林勇[1]，谢鹭[2]

1 上海理工大学 健康科学与工程学院，上海 200093
2 上海市生物医药技术研究院 基因组与生物信息研究所，上海 201203
3 中南大学湘雅医院，湖南 长沙 410008
4 同济大学 医学院，上海 200092

杨晨雨，刘振浩，代培斌，张钰，黄鹏杰，林勇，谢鹭. 基于多组学数据的肿瘤药物敏感性预测. 生物工程学报，2022, 38(6): 2201-2212.
YANG CY, LIU ZH, DAI PB, ZHANG Y, HUANG PJ, LIN Y, XIE L. Predicting tumor drug sensitivity with multi-omics data. Chin J Biotech, 2022, 38(6): 2201-2212.

摘　要：肿瘤药物敏感性预测在指导患者临床用药方面具有重要意义。本文基于癌症药物敏感性基因组学数据库 (genomics of drug sensitivity in cancer, GDSC) 198 种药物的细胞系敏感性 IC_{50} 数据，通过 Stacking 集成学习构建了包含基因表达、基因突变、拷贝数变异数据的多组学癌症药物敏感性预测模型。采用多种特征选择方法对基因特征进行降维，使用 Stacking 方法集成 6 种初级学习器和 1 种次级学习器进行建模，采用 5 折交叉进行模型验证。预测结果中 AUC 大于 0.9 的占比为 36.4%，在 0.8–0.9 之间的占比为 49.0%，最低 AUC 为 0.682。基于 Stacking 构建的多组学预测模型较已有单组学和多组学模型的准确性和稳定性具有优势。多组学整合预测药物敏感性优于单一组学。特征基因功能注释和富集分析解析了肿瘤对 sorafenib 潜在的耐药机制，从生物学角度提供了模型可解释性及其应用于临床用药指导的价值。

关键词：集成学习；Stacking；特征选择；多组学；肿瘤耐药机制；sorafenib

Received: September 4, 2021; **Accepted:** February 9, 2022; **Published online:** February 15, 2022
Supported by: National Natural Science Foundation of China (31301092, 31800700); Shanghai Municipal Health Commission, and Collaborative Innovation Cluster Project, China (2019CXJQ02)
Corresponding authors: LIN Yong. Tel: +86-21-55271119; E-mail: yong_lynn@usst.edu.cn
XIE Lu. Tel: +86-21-20283705; E-mail: xielu@sibpt.com
基金项目：国家自然科学基金 (31301092, 31800700)；上海市卫健委协同创新集群项目 (2019CXJQ02)

Predicting tumor drug sensitivity with multi-omics data

YANG Chenyu[1,2], LIU Zhenhao[2,3], DAI Peibin[4], ZHANG Yu[1,2], HUANG Pengjie[1,2], LIN Yong[1], XIE Lu[2]

1 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
2 Institute for Genome and Bioinformatics, Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai 201203, China
3 Xiangya Hospital, Central South University, Changsha 410008, Hunan, China
4 School of Medicine, Tongji University, Shanghai 200092, China

Abstract: The prediction of tumor drug sensitivity plays an important role in clinically guiding patients' medication. In this paper, a multi-omics data-based cancer drug sensitivity prediction model was constructed by Stacking ensemble learning method. The data including gene expression, mutation, copy number variation and drug sensitivity value (IC_{50}) of 198 drugs were downloaded from the GDSC database. Multiple feature selection methods were applied for dimensionality reduction. Six primary learners and one secondary learner were integrated into modeling by Stacking method. The model was validated with 5-fold cross-validation. In the prediction results, 36.4% of drug models' AUCs were greater than 0.9, 49.0% of drug models' AUCs were between 0.8–0.9, and the lowest drug model's AUC was 0.682. The multi-omics model for drug sensitivity prediction based on Stacking method is better than the known single-omics or multi-omics model in terms of accuracy and stability. The model based on multi-omics data is better than the single-omics data in predicting drug sensitivity. Function annotation and enrichment analysis of feature genes revealed the potential resistance mechanism of tumors to sorafenib, providing the model interpretability from a biological perspective, and demonstrated the model's potential applicability in clinical medication guidance.

Keywords: ensemble learning; Stacking; feature selection; multi-omics; tumor resistance mechanism; sorafenib

癌症是一种复杂疾病，肿瘤组织和微环境具有高度异质性。不同的患者亚群采用相同的癌症治疗方案反应存在差异：2006 年至 2018 年经美国食品药品监督管理局 (Food and Drug Administration, FDA) 批准的抗癌药物适应症的平均反应率仅为 40%[1]，癌症个性化治疗和精准用药为改善患者治疗情况及预后等提供可能。然而，该领域的两个主要挑战包括预测个体对不同治疗的反应和确定药物敏感性的分子生物标记物。随着高通量测序技术的发展与数据准确性的提高[2]，复杂的机器学习方法的不断提出[3]，增加了个性化精准治疗的可能性：通过机器学习对海量的肿瘤相关数据进行训练学习，结合患者的临床数据进行预测。

现有的预测药物敏感性的方法依赖于经不同药物处理的肿瘤细胞系转录组等组学数据，从中抽提特征来对药物反应进行回归或分类建模。线性回归是机器学习的基本模型之一，它运行快速并且易于解释。在实际的应用中，它经常与正则化惩罚项一起使用，减少过拟合问题。一些经

典的惩罚项线性回归模型，如岭回归[4]、最小绝对值收敛和选择算子 (least absolute shrinkage and selection operator, lasso) 回归还有弹性网络回归已应用于药物反应预测，例如，Huang 等构建了具有 lasso 惩罚项的正则化线性回归模型[5]。基于深度学习 Sharifi-Noghabi 等提出了 MOLI 方法来对体细胞突变、拷贝数变异、基因表达在内的多组学数据进行建模预测药物敏感性[6]。其他还有基于贝叶斯推理[7]、矩阵分解[8]、深度学习[9]等方法来构建的药物敏感性预测模型[10-11]。

方法的研究促进了人工智能在药物组学数据上的应用与发展，通过人工智能算法对药物数据进行大规模筛选有助于预测患者对不同治疗药物的反应进而实现个性化的癌症治疗方案[12-13]。由于癌症生长、扩散和对治疗产生反应的生态环境变化和进化机制的复杂性，肿瘤微环境成为肿瘤治疗的一个重要关注点，药物的作用机制变得更为复杂。整合多组学数据能更好地模拟出肿瘤在体内的状态，对预测肿瘤药物敏感性更具意义。本研究通过 IC_{50} 值排序将样本分为药敏组和耐药组，基于基因表达数据、突变数据和拷贝数变异数据使用多种特征选择方法实现特征降维，使用 Stacking 集成方法进行建模预测药物敏感性。模型构建中使用网格搜索方法进行调参，通过 5 折交叉验证方法确定最优参数值，以此创建每种药物独立的预测模型。基于多组学数据建立的 Stacking 模型，使用数据集经过二八划分后保留的测试集数据和 TCGA 外部数据集对模型的泛化能力进行验证，准确性及效率较其他模型均有一定程度的提升。

1 材料与方法

本研究中涉及所有分析均使用 R 软件 (版本 4.0.3) 和 Python 软件 (版本 3.8.5) 进行，单组学和多组学模型的构建方法以及对模型后续的分析流程如图 1 所示。

1.1 实验数据

本研究涉及药物敏感数据主要来源于癌症药物敏感性基因组学数据库 (genomics of drug sensitivity in cancer, GDSC, https://www.cancerrxgene.org)[14]，包括 198 个药物化合物、809 个细胞系组成的 135 242 条药物-细胞系敏感性 IC_{50} (half maximal inhibitory concentration) 数据。IC_{50} 是药物反应指标半抑制浓度，即对指定的生物过程 (或该过程中的某个组分例如酶、受体、细胞等) 抑制一半时所需的药物或抑制剂的浓度。IC_{50} 值越低，意味着细胞对药物越敏感。细胞的活力测定使用 DNA 染料 (Syto60) 或代谢测定法 (刃天青或 CellTiter-Glo) 测定细胞活力[15]。

GDSC 数据库中 E-MTAB-3610 数据集和 EGAS00001000978 数据集分别包括了上述药敏反应细胞系的基因表达、基因突变和拷贝数变异等数据。整合多组学数据与药敏数据匹配，保留共有细胞系数据，按药物进行数据分割用于后续分析。

1.2 数据处理和特征基因降维

1.2.1 样本筛选

从 E-MTAB-3610 数据集收集到的 1 018 个肿瘤细胞系 RNA-seq 数据样本，本研究首先基于基因表达数据对样本进行无监督的一致性聚类分析，分类模型采用的方法通过 R 包‘cola’实现[16]，使用概率近似正确 (probably approximately correct, PAC) 方法计算最佳聚类数目并进行分类。采用 PCA、t-SNE 两种方法对肿瘤细胞系样本进行聚类分析，根据聚类结果剔除与实体瘤存在显著差异的非实体瘤血液来源样本数据。

1.2.2 样本分组

出于疾病治疗及患者用药的角度，药物是否对其有效，即预测患者对药物是“敏感”还是

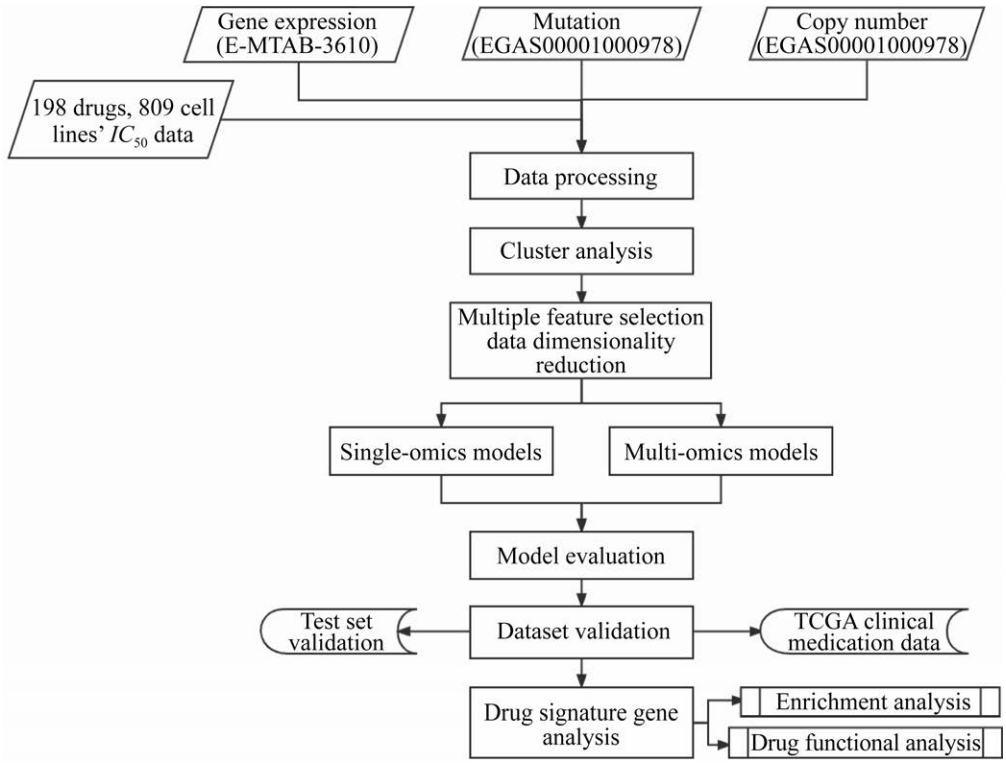

图 1　Stacking 模型构建方法与分析流程图

Figure 1　Flow chart of Stacking model construction and analysis.

“耐药”对治疗更有实际意义。因此，我们将药物敏感预测问题转化为二分类问题进行研究。考虑样本的均衡性，按照如下方法选取样本：对每种药物，将细胞系样本按照 IC_{50} 值从高到低进行排序，取前 30%的样本作为药物应答反应耐药样本，后 30%作为药物应答反应敏感样本，剔除中间 40%样本，将研究由回归问题转化为二分类问题处理。将处理好的样本进行随机划分，80%的样本作为训练集训练模型，剩余的 20%样本作为测试集用于模型泛化性能验证。

1.2.3　特征降维

输入特征数量的增加会提高模型的计算复杂度并降低预测的准确性，因此需要对特征进行降维分析。首先对样本中无显著差异的干扰特征进行去噪，我们使用过滤法，根据不同组的数据特点设置不同的阈值剔除低方差的特征：对 RNA-seq 数据，我们将阈值设定为 1；对突变数据，我们将阈值设置为 0.05；对拷贝数数据，我们将阈值设置为 0.5。

不同的特征选择方法优缺点不同，我们采用多种特征选择方法组合来挑选特征。使用多种特征选择方法能有效避免单一方法的局限导致的肿瘤相关基因特征遗漏。本研究使用 Python 语言中 sklearn 机器学习库中的特征选择模块实现特征基因降维。使用了 4 种特征选择方法：(1) 根据样本的方差分析 F 值 (F-value)；(2) 单变量线性回归测试；(3) 基于 lasso 的回归；(4) 卡方检验。重点说明如下：单变量线性回归测试，用于测试多个回归变量的每一个个体效应的线性模型。首先，用公式 (1) 计算每个特征量和目标因变量之间的样本相关系数，

计算出 F-score 后，通过公式 (2) 进行转换，此时 f 服从 F(1, n–2)。统计中常用公式 (3) 来检验正态假定下两个变量之间的相关性，值越高表明相关性越大。

$$r_i = \frac{(X_i - \overline{X_i})^T * (y - \bar{y})}{std(X_i) * std(y)} \quad (1)$$

$$f = \frac{r_i^2}{1 - r_i^2} * (n - 2) \quad (2)$$

$$t = \sqrt{n-2}\frac{r_i}{\sqrt{1 - r_i^2}} \quad (3)$$

Lasso 算法，是一种同时进行特征选择和正则化的回归分析方法，旨在增强统计模型的预测准确性和可解释性。Lasso 算法加入 L1 正则化，计算损失函数参见式 (4)。

$$J = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda\omega_1 \quad (4)$$

Lasso (L1 正则化)等价于参数 ω 的先验概率分布满足拉普拉斯分布，更容易使得权重变为 0。针对不同的数据，我们选择适用的方法，每个组学数据保留 200 个相关性排名靠前的特征基因作建模特征：对 RNA-seq 数据，本研究使用基于 lasso 的回归、样本方差分析 F 值和单变量线性回归测试；对突变数据，本研究使用样本方差分析 F 值和卡方检验；对拷贝数变异数据，本研究使用基于 lasso 的回归和样本方差分析 F 值的方法。

1.3 模型构建

采用 Stacking 集成方法构建机器学习模型预测药物敏感性。Stacking 通过各种机器学习算法的差异来保证初级学习器的复杂性，次级学习器用于总结各初级学习器的预测结果。Stacking 集成学习方法预测准确性更高，过拟合风险较低。本研究中 Stacking 集成方法集成了 6 种初级学习器：决策树、随机森林、梯度提升树、极限树、梯度下降分类、岭分类器。使用逻辑回归模型作为次级学习器进行单组学建模和多组学建模。对于单组学模型，以处理后的基因表达数据为特征输入；对于多组学数据，以处理后的基因表达数据、突变数据、拷贝数变异数据为特征输入。

模型的调参过程我们采用网格搜索的方法。本研究针对所选模型，选择子树的数量 (n_estimators)、最大树深 (max_depth) 2 个参数进行网格搜索调优，这两个参数都是对模型在未知数据上的评估性能影响程度最高的 2 个参数。n_estimators 参数本研究选取 40、70、100、130、160 五个值范围，max_depth 参数选取 none、3、6、9 四个值范围。

1.4 模型性能评估

训练集的划分采用五折交叉验证的方法：将数据集分成 5 份，不重复地每次取其中的 1 份用作测试集验证，剩余的 4 份作训练集训练模型，5 次结果的均值作为算法预测的结果。交叉验证对数据进行多次划分，能消除单次划分时数据划分不平衡造成的过拟合影响。除了使用常用的 ROC 曲线下面积 (area under curve, AUC) 和准确度 (accuracy) 作为衡量二分类模型优劣的评价指标，为能更好地描述模型预测性能，参考了精确率 (precision)、召回率 (recall) 和 F1 分数 (F1-Score) 分类精确度指标。后续使用划分出的测试集对训练后模型的泛化性能进行验证。

1.5 外部数据集验证

通过 TCGAbiolinks 工具[17]获取 TCGA 中 2 572 条病人肿瘤用药治疗数据，以及对应病人的基因表达数据、基因突变数据和拷贝数变异数据。其中，保留通过单一药物进行治疗的肿瘤病人数据作为外部数据集验证模型。

根据用药后疾病进展，TCGA 中药物反应分为完全缓解 (complete response, CR)、部分缓解

(partial response, PR)、疾病稳定 (stable disease, SD)、疾病进展 (progressive disease, PD) 四部分。我们将 CR 和 PR 归纳为肿瘤样本药物敏感，SD 和 PD 归纳为肿瘤样本耐药进行分析。

2 结果与分析

2.1 聚类结果分析

实体瘤与非实体瘤在转录水平上存在显著差异，而肿瘤的药物敏感性与耐药基因的表达水平高度相关。无监督一致性聚类 (图 2A) 以及 PCA 和 t-SNE 方法对肿瘤细胞系样本进行聚类分析 (图 2B–C) 发现非实体瘤 (血液组织来源) 与实体瘤存在显著异质性。为了提高预测模型的相关性和准确性，本研究剔除 383 个非实体瘤样本数据，保留 635 个实体瘤样本数据进行建模和后续分析。

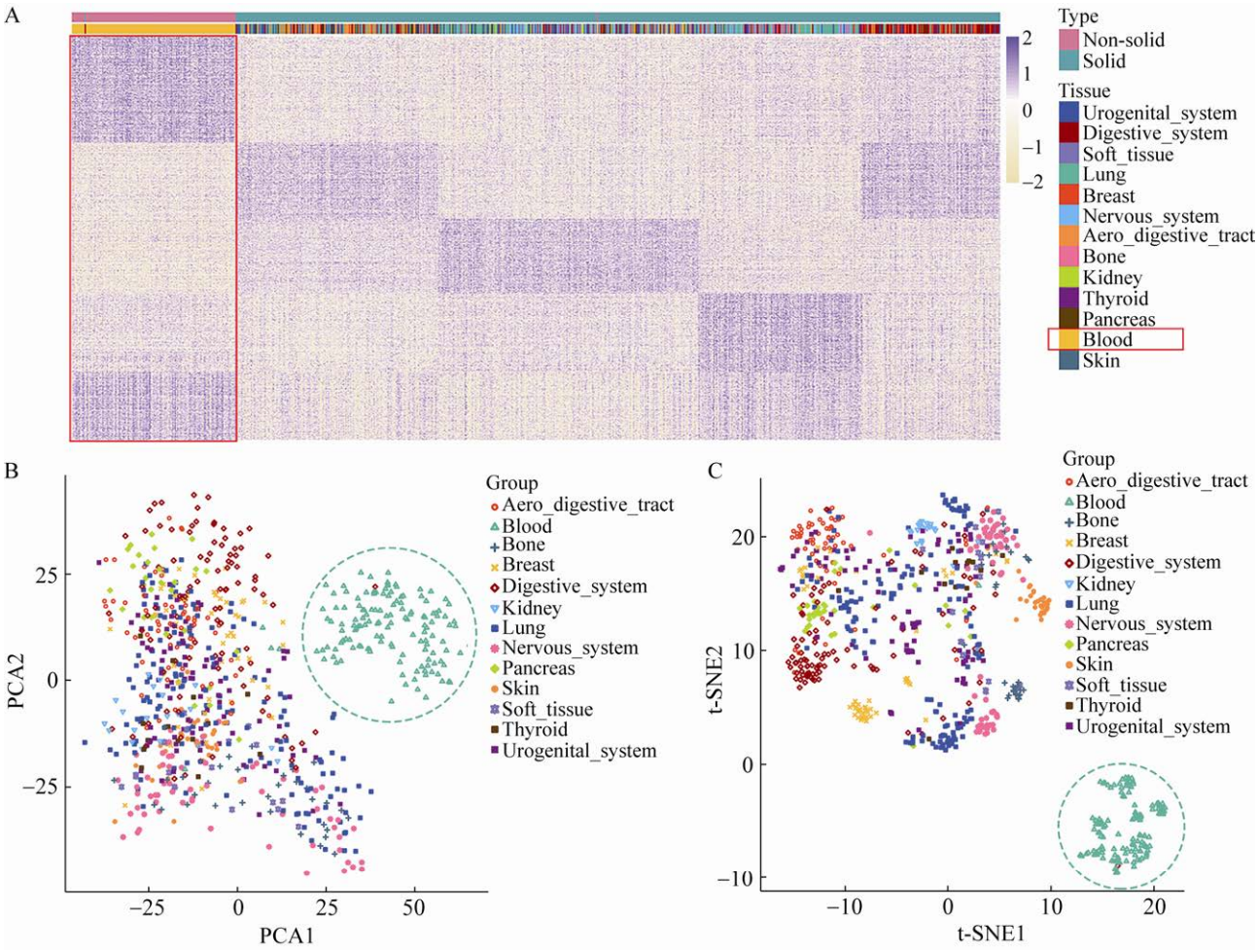

图 2 肿瘤细胞系样本转录水平聚类分析

Figure 2 Cluster analysis of tumor cell line samples on transcriptional level. (A) Unsupervised consistent cluster analysis. The number of clusters is 5, and the selected part is blood tissue as the clustering result. (B) Principal component analysis method of unsupervised clustering. The green triangle arrow is divided separately as blood tissue. (C) t-SNE method of unsupervised clustering. The green triangle arrow is divided separately as blood tissue. There are significant differences in gene expression levels between blood tissue and other tissues.

2.2 模型结果比较评估

本研究对 198 种药物进行了有效建模，并与经典分类模型预测方法及相关文献已报道方法进行了比较。将样本数据二八划分测试集与训练集，使用 5 折交叉验证的方法在训练集中训练模型，采用随机森林[18] (radom forest, RF)、极限树 (extra tree, ET)、支持向量机[19] (support vector machine, SVM)、bagging 集成方法、boosting 集成方法构建模型与本研究的 Stacking 集成方法模型进行比较。基于单一组学数据(转录组) 进行建模预测，模型效果没有显著优势。鉴于基因突变等基因组水平变异与皮肤癌(如黑色素瘤等) 发生和发展等密切相关，单一基因表达数据不能有效预测皮肤癌等肿瘤药物敏感性。去除皮肤癌样本重新建模发现，Stacking 集成方法预测药物敏感性 AUC 均值优于其他算法，且显著高于包含皮肤癌样本集数据时的预测结果 (0.874>0.800, 表 1)。部分实体瘤和皮肤癌样本包含大量基因突变并参与了肿瘤发生，基因突变引起的基因功能和结构改变可能导致了基因表达调控网络的异质性，从而影响了药物敏感性预测的准确性。整合多组学数据可以较好地模拟肿瘤组织和微环境的状态，有利于建模预测肿瘤药物敏感性并探究肿瘤耐药相关的生物学机制。整合基因表达、基因突变、拷贝数变异数据建立多组学整合预测模型。Stacking 集成学习方法构建模型 AUC 平均值为 0.874。较单一组学模型，多组学模型 Stacking 集成学习方法 AUC 值提升约 9.3%。模型预测结果显示，198 种药物中有 180 种药物 (90.9%) 的多组学模型预测 AUC 值高于单一组学模型。整合多组学数据构建模型预测效果优于单一组学模型。较其他机器学习方法，Stacking 集成学习方法 AUC 值提升约 3.9%–7.5%，预测效果显著高于其他模型 (图 3B, 表 2)。

2.3 模型效果验证

使用测试集数据对多组学模型的泛化性能进行验证 (图 3B)，结果显示 Stacking 模型在测试集上的验证效果最好，AUC 均值为 0.821，较其他算法有明显提升。测试集验证结果显示基于多组学数据的 Stacking 模型预测药物敏感性具有良好的泛化性能。

表 1 各模型的 AUC 均值和第一三四分位数取值范围

Table 1 The average value of AUC and the value range between first quartile to three quartiles of each model

Model methods	Single-omics including skin mean (Q1–Q3)	Single-omics excluding skin mean (Q1–Q3)	Multi-omics mean (Q1–Q3)
Stacking	0.800 (0.739–0.852)	0.874 (0.850–0.913)	0.874 (0.842–0.914)
ET	0.787 (0.747–0.816)	0.828 (0.791–0.851)	0.813 (0.785–0.839)
RF	0.778 (0.738–0.807)	0.819 (0.783–0.842)	0.822 (0.791–0.851)
SVM	0.817 (0.776–0.854)	0.846 (0.809–0.875)	0.827 (0.797–0.849)
Bagging	0.773 (0.738–0.805)	0.814 (0.779–0.844)	0.816 (0.789–0.844)
Boosting	0.816 (0.760–0.867)	0.859 (0.818–0.900)	0.841 (0.803–0.879)
Bayes	0.723 (0.677–0.783)	0.823 (0.782–0.848)	0.742 (0.712–0.785)
ElasticNet	0.717 (0.666–0.800)	0.834 (0.793–0.863)	0.772 (0.725–0.814)
Lasso	0.723 (0.668–0.799)	0.841 (0.794–0.864)	0.775 (0.739–0.817)
Perceptron	0.707 (0.648–0.789)	0.835 (0.796–0.862)	0.803 (0.765–0.838)

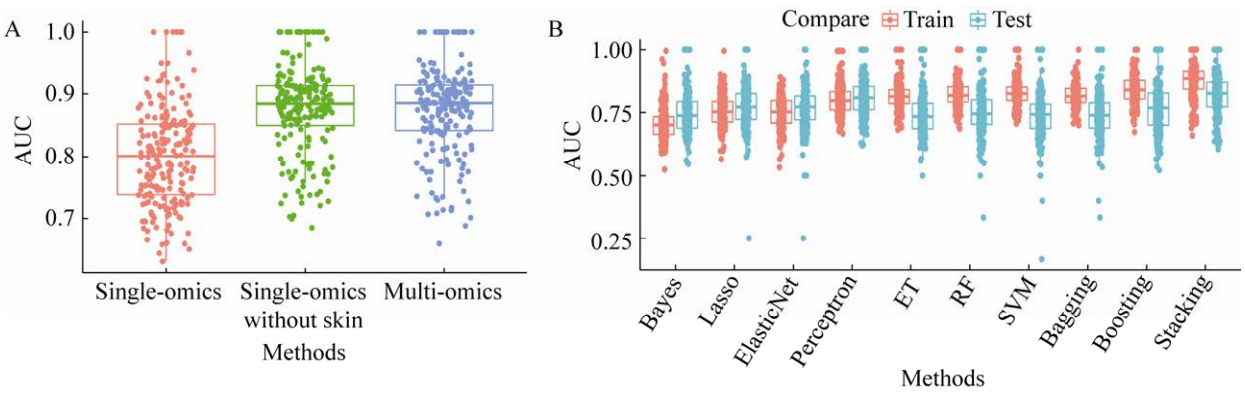

图 3　模型预测结果准确性比较

Figure 3　Comparison of model prediction accuracy. (A) Comparison of AUC values between including, excluding skin tissues in the single-omics model and multi-omics model. (B) The result of AUC values from different multi-omics prediction models comparing training sets with testing sets.

表 2　多组学中各模型性能评估指标的平均值

Table 2　The average value of each model performance evaluation in multi-omics

Model methods	AUC	Accuracy	Precision	Recall	F1
Stacking	0.874	0.833	0.837	0.834	0.832
ET	0.813	0.740	0.756	0.717	0.729
RF	0.822	0.748	0.769	0.716	0.734
SVM	0.827	0.732	0.767	0.675	0.705
Bagging	0.816	0.741	0.759	0.714	0.728
Boosting	0.841	0.757	0.771	0.745	0.747
Bayes	0.742	0.750	0.757	0.722	0.745
ElasticNet	0.772	0.766	0.775	0.743	0.761
Lasso	0.775	0.775	0.781	0.769	0.771
Perceptron	0.803	0.792	0.800	0.778	0.791

此外，结合临床数据，本研究通过 TCGA 中肿瘤病人用药治疗情况进行外部数据模型验证。其中，2 572 条病人肿瘤用药治疗数据中保留 485 条肿瘤患者单一用药治疗数据用于模型预测分析。保留已有预测模型且样本量>10 的药物用药数据，筛选获得 cisplatin (98 例)、temozolomide (77 例)、gemcitabine (50 例)、fluorouracil (49 例)、sorafenib (12 例) 5 种药物进行验证。基于 Stacking 的多组学数据预测药物敏感性模型对 TCGA 中病人用药反应进行了有效预测 (表 3)，灵敏度指标是评价诊断准确度的基本指标之一，指当真实情况为肿瘤药物敏感时预测结果正确的能力，Stacking 模型在 5 种药物的预测过程中均显示出稳定的预测结果，在 4 种药物 (cisplatin, temozolomide, fluorouracil, sorafenib) 中预测结果均为最优，显示出模型具有可用于对患者临床药物治疗反应进行预测的潜力。

2.4　药物耐药性分析

经检索 OncoKB 数据库[20]中治疗水平为 level 1 和 level 2 (美国 FDA 认证) 的药物中有 16 种药物与本研究涉及药物一致，包含相应靶标及疾病类型等信息。根据预测效果排序，选取 sorafenib、alpelisib、dabrafenib、afatinib、lapatinib、crizotinib 等药物的预测结果相关信息进行列举 (表 4)。

Sorafenib 敏感性预测在多组学模型预测中 AUC 值为 0.876，与单一组学预测区别较小 (图 4A–B)。KIT 编码受体酪氨酸激酶是 sorafenib 的靶基因，在胃肠道间质瘤中反复发生突变[21]。KIT 通过 PI3K、MAPK 和 STAT 等通路引起细

表 3　在外部测试集中多组学各模型灵敏度评估指标

Table 3　Multi-omics model sensitivity evaluation index in the external test set

Machine methods	Cisplatin	Temozolomide	Gemcitabine	Fluorouracil	Sorafenib
RF	0.676	0.636	0.667	0.833	0
ET	0.622	0.545	0.889	0.722	0
SVM	0.784	0.545	0.611	0.500	0
Bagging	0.351	0.364	0.444	0.778	1
Boosting	0.581	0.545	0.389	0.556	1
Stacking	0.838	0.727	0.667	0.833	1

表 4　与 OncoKB 数据库匹配到的药物作用信息数据

Table 4　Drug effect information data matched with OncoKB database

Drug names	AUC of single-model	AUC of multi-model	Genes	Alterations	Cancer types
Sorafenib	0.833	0.876	KIT	A829P, A829P, C809G, D816, D820, N822, Y823D	Gastrointestinal stromal tumor
Alpelisib	0.784	0.884	PIK3CA	C420R, E542K, E545A, E545D, H1047L, H1047R, H1047Y, Q546E, Q546R, et al.	Breast cancer
Dabrafenib	0.702	0.776	BRAF	V600E,V600K	Anaplastic thyroid cancer, melanoma, non-small cell lung cancer
Afatinib	0.808	0.886	EGFR	Exon 19 deletion, L858R, G719, L861Q, S768I	Non-small cell lung cancer
Lapatinib	0.795	0.825	ERBB2	Amplification	Breast cancer, colorectal cancer

胞内信号传导增加，导致细胞的增殖和存活[22]。包括 sorafenib 在内的至少有 8 种靶向 KIT 的小分子酪氨酸激酶抑制剂 (tyrosine kinase inhibitors, TKI) 已获得 FDA 批准，每种 TKI 的疗效很大程度上取决于激活 KIT 突变的位置[23]。

基于 GO 数据库生物过程 (biological process, BP) 进行特征基因功能注释和富集分析 (图 4C)。Sorafenib 通过抑制细胞内多种丝/苏氨酸和酪氨酸激酶活性抑制肿瘤细胞的生长和血管的生成，以此发挥药物疗效。丝氨酸蛋白酶抑制蛋白基因家族中的基因如 *SERPINB2*、*SERPINB3*、*SERPINB4* 等 (共 10 个 *SERPINB* 家族基因在特征选择中作为特征被挑选出来) 作为 sorafenib 的特征基因，被显著富集在与酶活性相关的生物学过程中。干扰素反应相关的生物学过程显著富集，与 $CD8^+$ $Ki67^+$ T 细胞产生的 γ 干扰素是 sorafenib 药物反应的关键生物标记物[24-25]一致。对 sorafenib 中通过特征选择方法挑选出来的基因表达组特征基因，使用 R 语言‘cola’包中的‘adjust_matrix’函数删除方差较小的特征基因，这里我们设置的截止位阈值为 0.25，再使用‘limma’包的贝叶斯方法计算差异基因，最终得到 108 个显著差异基因，对这 108 个基因使用‘pheatmap’包进行热图的绘制 (图 4D)，针对 *SLC* 基因家族存在多个基因产生差异表达，结合已发表文献的相关研究，了解到 *SLC* 基因家族编码膜转运蛋白，参与药物的跨膜运输。*SLC* 基因家族 *SLC2A1*、*SLC14A1*、*SLC14A2*、*SLC16A3*、*SLC22A4*、*SLC35D2* 六个基因 (*SLC14A1*、*SLC14A2* 在基因突变和拷

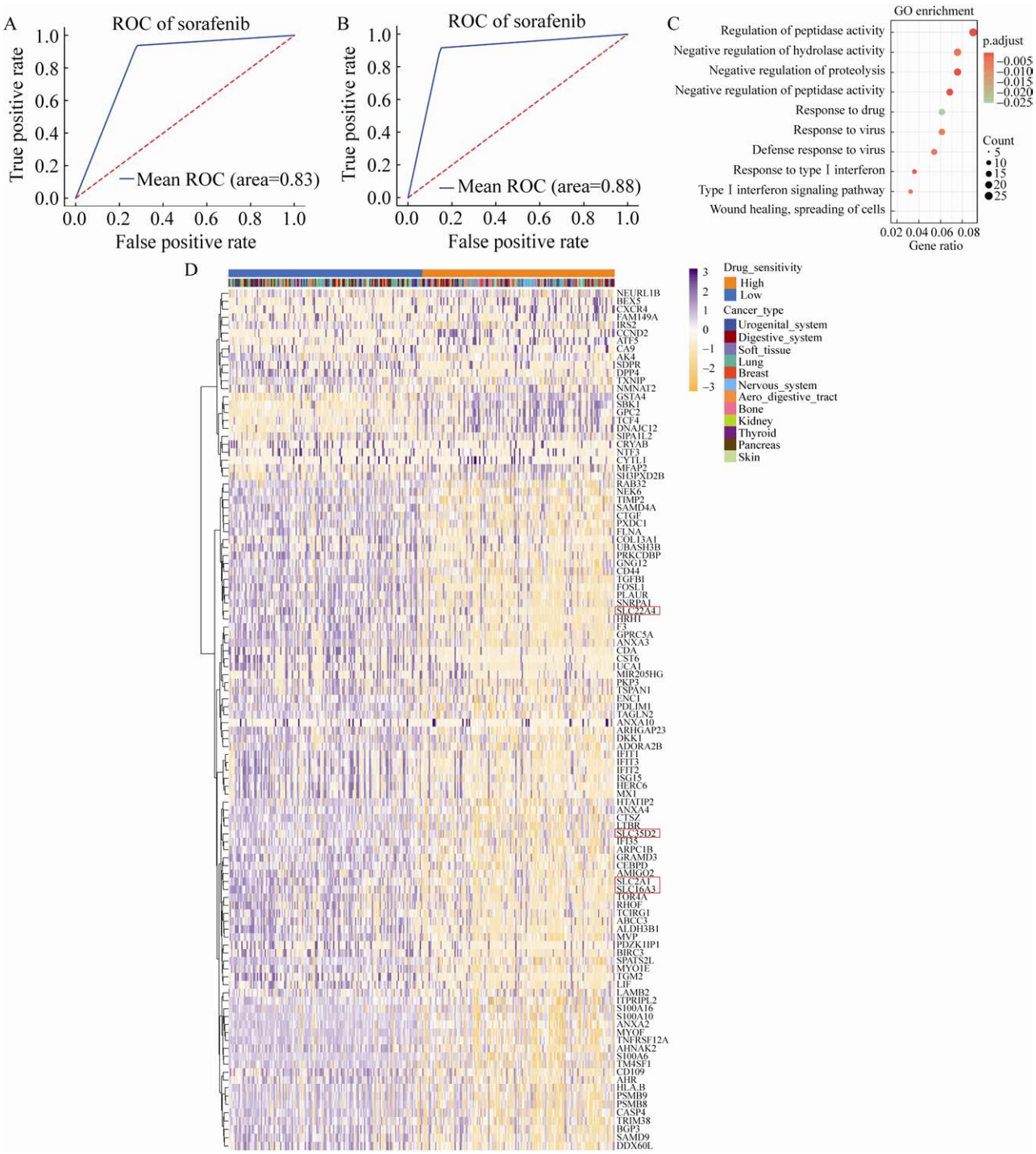

图 4　Sorafenib 预测结果 ROC 曲线与特征基因功能富集分析

Figure 4　Drug sensitivity prediction of sorafenib and functional enrichment analysis of feature genes. (A) The ROC curve for drug sensitivity prediction of sorafenib with the single-omics prediction model based on gene expression data. (B) The ROC curve for drug sensitivity prediction of sorafenib with the multi-omics prediction model. (C) Functional enrichment analysis of characteristic genes for drug sensitivity prediction of sorafenib. (D) Heat map showing 108 top up-regulated gene and down-regulated gene features selected from RNA-seq data (fold change≥2, P<0.05).

贝数变异的特征选择中被挑选出) 作为特征基因可能参与了细胞内药物浓度的调节，通过转运药物到细胞外导致肿瘤耐药。

3 讨论

本研究构建的模型使用 Stacking 集成方法采用双层预测结构，集成了决策树、随机森林、梯度提升树、极限树、梯度下降分类、岭回归分类器 6 种机器学习方法作为初级学习器，使用逻辑回归模型作为次级学习器。该模型与其他模型进行比较在多组学数据预测上准确性显著优于其他模型。

本研究分别基于单组学数据 (基因表达) 和多组学数据 (基因表达、基因突变、拷贝数变异) 为特征，使用多种特征选择方法组合和网格搜索模型调参方法构建模型。针对样本的组织类别，在基因表达层面血液肿瘤与其他实体瘤之间存在显著差异。出于对模型准确性的考虑，本研究主要对实体肿瘤进行分析。Stacking 集成学习基于单一转录组学数据构建模型，有无皮肤肿瘤样本预测结果的差异显示，基因突变等参与了肿瘤发生发展调控和引起药物反应的异质性。整合多组学模型 AUC 平均值达到 0.874，较单组学的 0.800 提升约 9.3%。模型预测结果显示,198 种药物中有 180 种药物 (90.9%) 的多组学模型预测 AUC 值高于单一组学模型。较其他机器学习方法，Stacking 集成学习方法在多组学预测中 AUC 值提升约 3.9%–7.5%，整合多组学数据有利于肿瘤药物敏感性预测模型的构建。使用测试集数据对模型的泛化性能进行预测，结果显示 Stacking 模型的预测 AUC 在 0.821，预测效果显著高于其他模型。Sorafenib 药物敏感性预测相关的特征基因与药物作用机制显著相关，证明了特征选择方法的有效性。

模型评估与测试集的验证表明了该模型对患者肿瘤药物敏感性具有良好的预测能力。从 TCGA 上我们搜集到了相关的临床用药治疗情况，并下载了病人相关的 RNA-seq、基因突变和拷贝数变异数据用作外部测试集对模型进行临床应用的验证。验证结果表明，相对于其他模型预测结果的较大波动幅度，本研究的 Stacking 集成模型预测具有较强的准确性和临床应用价值。肿瘤的发生发展涉及复杂的生物学机制，生物体的药物反应需要多组学水平参与调控，通过多组学数据预测药物敏感性能为肿瘤对药物的反应进行预测，模型在临床实际的预测上具有较高的灵敏度，大概率保证对临床诊疗患者中存在的用药敏感者得到正确有效的及时治疗，为肿瘤病人个性化、精准化用药提供参考，有效提高病人的治疗效率。后续可以扩展为网页等工具形式为临床医生的用药决策提供支持。

本研究在肿瘤多组学数据预测药物敏感性方面取得进展，但仍存在不足：由于单一肿瘤类型细胞系样本有限，整合泛癌细胞系数据构建的模型在不同肿瘤类型中可能存在异质性；模型构建中，基因组学、转录组学、表观组学等组学数据的筛选组合，组学特征间存在的复杂调控机制等可能会影响模型的特异性和有效性。精准医学时代肿瘤多组学数据在不断累积，深度学习等人工智能的算法应用在不断拓展，因此高效的组合特征筛选方法和稳健的模型构建方法研究在肿瘤药物敏感性预测领域都仍然是值得持续关注的方向。

REFERENCES

[1] Chen EY, Raghunathan V, Prasad V. An overview of cancer drugs approved by the US food and drug administration based on the surrogate end point of response rate. JAMA Intern Med, 2019, 179(7): 915-921.

[2] Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell, 2015, 58(4): 586-597.

[3] Deo RC. Machine learning in medicine. Circulation, 2015, 132(20): 1920-1930.
[4] 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.
[5] Huang EW, Bhope A, Lim J, et al. Tissue-guided lasso for prediction of clinical drug response using preclinical samples. PLoS Comput Biol, 2020, 16(1): e1007607.
[6] Sharifi-Noghabi H, Zolotareva O, Collins CC, et al. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 2019, 35(14): i501-i509.
[7] Liu Q, Muglia LJ, Huang LF. Network as a biomarker: a novel network-based sparse Bayesian machine for pathway-driven drug response prediction. Genes, 2019, 10(8): 602.
[8] Emdadi A, Eslahchi C. DSPLMF: a method for cancer drug sensitivity prediction using a novel regularization approach in logistic matrix factorization. Front Genet, 2020, 11: 75.
[9] Chiu YC, Chen HIH, Gorthi A, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. Brief Bioinform, 2020, 21(6): 2066-2083.
[10] Kurilov R, Haibe-Kains B, Brors B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. Sci Rep, 2020, 10(1): 2849.
[11] Zhang NQ, Wang HY, Fang Y, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS Comput Biol, 2015, 11(9): e1004498.
[12] Feng F, Shen BH, Mou XQ, et al. Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. J Genet Genomics, 2021, 48(7): 540-551.
[13] Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol, 2014, 32(12): 1202-1212.
[14] Yang WJ, Soares J, Greninger P, 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-D961.
[15] Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. Cell, 2016, 166(3): 740-754.
[16] Gu Z, Schlesner M, Hübschmann D. Cola: an R/Bioconductor package for consensus partitioning through a general framwork. Nucleic Acids Res, 2021, 49(3): e15.
[17] Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res, 2016, 44(8): e71.
[18] Fang Y, Xu PR, Yang JL, et al. A quantile regression forest based method to predict drug response and assess prediction reliability. PLoS One, 2018, 13(10): e0205155.
[19] Huang C, Mezencev R, McDonald JF, et al. Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. PLoS One, 2017, 12(10): e0186906.
[20] Chakravarty D, Gao JJ, Phillips SM, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol, 2017, 1: 1-16.
[21] Hirota S, Isozaki K, Moriyama Y, et al. Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science, 1998, 279(5350): 577-580.
[22] Corless CL, Barnett CM, Heinrich MC. Gastrointestinal stromal tumours: origin and molecular oncology. Nat Rev Cancer, 2011, 11(12): 865-878.
[23] Yun CH, Mengwasser KE, Toms AV, et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. PNAS, 2008, 105(6): 2070-2075.
[24] Kalathil SG, Hutson A, Barbi J, et al. Augmentation of IFN-γ^+ $CD8^+$ T cell responses correlates with survival of HCC patients on sorafenib therapy. JCI Insight, 2019, 4(15): e130116.
[25] Tang WW, Chen ZY, Zhang WL, et al. The mechanisms of sorafenib resistance in hepatocellular carcinoma: theoretical basis and therapeutic aspects. Signal Transduct Target Ther, 2020, 5(1): 87.

(本文责编 郝丽芳)