



基于核范数正则化的抗癌药物组合 协同作用预测*

史磊晶¹⁾ 王波¹⁾ 张杉^{1,2)} 任福全¹⁾ 李玉双^{1)**}⁽¹⁾ 燕山大学理学院, 秦皇岛 066004; ⁽²⁾ 中车唐山机车车辆有限公司, 唐山 063000

摘要 目的 抗癌药物联合疗法是一种很有前途的治疗策略。针对特定癌症类型, 选择高度协同的药物组合, 对提高癌症疗效至关重要。然而, 确定具有协同作用的药物组合是一项复杂而困难的工作。本研究旨在完全以数据驱动、计算建模的方式优化抗癌药物组合高通量虚拟筛选, 为“旧药重新定位新组合”提供理论参考。**方法** 借鉴矩阵填充思想, 构建了基于核范数正则化的计算模型 NNRM, 用于预测抗癌药物组合的协同得分和协同状态。针对固定细胞系构造对称的协同得分观测矩阵; 采用分折技巧将观测矩阵稀疏化; 借助“交替方向乘子法”和“软阈值估计”求解模型。**结果** 将 NNRM 应用于 O'Neil 团队发布的数据集, 预测的协同得分与观测值之间的均方根误差为 14.78, 预测的协同状态准确率为 0.94, 优于随机森林 (RF) 和支持向量机 (SVM), 完全可以与深度学习模型相媲美。此外, NNRM 预测的部分缺失值结果与已有研究或临床实践相吻合。**结论** NNRM 可实现大规模、批量预测抗癌药物组合的协同作用, 极大地降低了已有模型对数据的要求和计算成本, 缩短了高通量虚拟筛选的测试时间, 可以作为抗癌药物组合高通量虚拟筛选的可选择工具。

关键词 抗癌药物组合, 协同作用, 高通量虚拟筛选, 矩阵填充, 核范数正则化, 计算模型

中图分类号 Q7, Q81

DOI: 10.16476/j.pibb.2022.0252

耐药性一直是癌症治疗的主要障碍^[1], 需要超越“一类疾病、一种药物、一个靶点”的治疗范式, 才能实现抗癌精准治疗。研究表明, 药物组合治疗可以增强肿瘤细胞杀伤力, 降低毒性并避免耐药性的快速发作, 使患者产生更持久的反应^[2]。例如, 对于携带 BRAF V600E 突变的黑色素瘤患者, 组合使用达拉非尼和曲美替尼能够延迟肿瘤复发^[3]。药物组合治疗可引起 3 种不同反应, 组合效应大于独立效应的总和 (协同作用)、组合效应等于独立效应的总和 (累加作用)、组合效应小于独立效应的总和 (拮抗作用)。组合治疗的主要目标是获得药物协同作用^[2]。对于特定患者, 哪些药物组合治疗可能对其有效, 在备选方案中应该优先选用哪种组合, 是临床上必须要面对和解决的问题。然而, 确定具有协同作用的药物组合是极具挑战性的。临床筛选耗时、昂贵且易出错; 高通量筛选虽然能够在合理的时间内以低成本确定药物组合, 但很难用它测试完整的组合空间^[4]。

为解决上述问题, 2012 年 DREAM 挑战计划与

美国国家癌症研究中心 (NCI) 合作, 提出了 NCI-DREAM 药物协同排序预测挑战。Bansal 等^[5]基于实验剂量-反应曲线的评分指标对提交的 32 种解决方法进行了系统分析。为了加快对药物组合协同机制的了解, 2015 年, DREAM 挑战计划与 AstraZeneca 和 Sanger 研究所合作, 发起了又一项更大规模的 AZ-S DREAM 挑战, 发布了 11 576 个协同筛查实验数据。Menden 等^[6]使用严格的框架评估了 160 个计算策略, 为科学界提供了完整数据和方法学基准。此外, O'Neil 团队和美国国家癌症研究所 ALMANAC 分别于 2016 年和 2017 年独立组织了另外两项高通量协同筛查。O'Neil 团队发布的数据涵盖了 38 种药物和 39 种癌细胞系, 执行的高通量筛选涵盖了可能药物对中的 83%^[7]; ALMANAC 系统评估了 5 000 多种 FDA 批准的抗癌

* 河北省自然科学基金 (A2020203021) 和河北省引进留学人员基金 (C20200365) 资助项目。

** 通讯联系人。

Tel: 15232356135, E-mail: yushuangli@ysu.edu.cn

收稿日期: 2022-05-29, 接受日期: 2022-07-15

药物对于 60 种特征明确的肿瘤细胞系 (NCI-60) 的组合反应^[8]。

以上连续开展的高通量协同筛查均使用了不同的实验设计和方案, 为开发和评估协同预测计算模型提供了强大的数据支撑。在诸多的计算策略中, 深度学习技术因其能够从大规模数据中提取抽象信息而备受研究者青睐, 如基于 O'Neil 团队发布的数据, Preuer 等^[9] 将癌细胞系基因表达与药物特征作为输入信息, 首次构建深度神经网络模型 (DeepSynergy) 预测药物组合协同作用, 结果优于多种经典的机器学习模型; Ma 等^[10] 在此工作基础上引入主成分分析 (Deep learning+PCA) 进行特征降维, 在保证预测效果的同时降低了模型计算成本; 陈希等^[11] 提出了 MulinputSynergy 模型, 整合癌细胞系的基因表达、基因突变、拷贝数变异和药物化学特征, 利用卷积神经网络进行特征降维, 借助残差前馈神经网络准确、快速地定位高度协同的药物组合; Zhang 等^[12] 使用自动编码器编码基因表达、基因突变和拷贝数变异, 将编码信息与药物特征输入深度神经网络 (AuDNNsynergy), 不仅取得了比 DeepSynergy 更好的预测结果, 而且识别出与药物协同相关联的基因特征。这些工作证明了使用深度学习技术和大量辅助数据预测药物协同的可行性, 衍生的代价则是高昂的计算成本。

本文利用 O'Neil 团队发布的高通量协同筛查数据, 借鉴矩阵填充思想, 构建了基于核范数正则化的抗癌药物组合协同作用预测模型 NNRM。该模型不需要任何其他辅助信息, 仅从抗癌药物组合协同筛查数据出发, 构造对称的协同得分观测矩阵, 利用分析技巧将观测矩阵稀疏化, 采用“交替方向乘子法 (ADMM)”和“软阈值估计”求解模型, 不仅取得了与深度学习技术相媲美的预测结果, 而且大大降低了计算成本。

1 材料与方法

1.1 数据来源和数据处理

O'Neil 团队发布的高通量协同筛查数据涵盖了来自 7 种不同组织的 39 个癌细胞系和 38 种药物 (24 种批准药物和 14 种试验药物)。其中, 22 种药物 (14 种试验药物和 8 种批准药物, 统称为穷举组药物) 两两组合 (不考虑药物自身的组合, 不考虑药物组合的顺序, 以下同), 得到 231 种药物对, 剩余的 16 种批准药物 (统称为补充组药物) 与 22 种穷举组药物组合, 构成 352 种药物对, 总计 583

种药物组合。Preuer 等^[9] 借助 Combenefit 批处理程序^[13] 计算了 Loewe Additivity 模型^[14], 给出 583 种药物组合的协同得分。

针对每个细胞系, 构造 38×38 的对称观测矩阵 M : 将 38 种药物以相同顺序进行横、纵排列, 设为 $D_1, \dots, D_{38}$ 。若 D_i 和 D_j 有观测值 (协同得分) S , 则 $M_{ij} = M_{ji} = S$; 若没有观测值, 则 $M_{ij} = M_{ji} = 0$ 。由此得到包含 1 166 个非零元和 278 个零元 (38 个对角零元和 240 个非对角零元) 的对称矩阵。其中, 非对角零元恰好对应缺失的 16 种补充组药物的两两组合。观测矩阵的第 i 列恰好对应以 D_i 为成员的所有药物组合。

1.2 基于核范数正则化算法的预测模型 NNRM

1.2.1 NNRM 的构建与求解

本文的研究目的是依据细胞系观测矩阵中的部分协同得分来预测/填充其余位置上的协同得分, 该问题属于矩阵完备问题^[15], 其思想是假设少量潜在的变量可以解释大部分数据, 具体可描述为:

$$\begin{aligned} \min \quad & \text{rank}(X) \\ \text{s.t.} \quad & X_{ij} = M_{ij}, (i, j) \in \Omega \end{aligned} \quad (1)$$

其中, $X \in R^{38 \times 38}$ 是预测矩阵, $\text{rank}(X)$ 为矩阵 X 的秩, $M \in R^{38 \times 38}$ 是观测矩阵, $\Omega = \{(i, j) | M_{ij} \neq 0, i \in \{1, \dots, 38\}, j \in \{1, \dots, 38\}\}$ 表示非零空间。

已知公式 (1) 描述的问题是一个 NP 难题, 可转化为如下线性约束凸规划问题^[16]:

$$\begin{aligned} \min \quad & \|X\|_* \\ \text{s.t.} \quad & X - Y = 0 \\ & Y \in \{VIV^T = M_\Omega\} \end{aligned} \quad (2)$$

其中, $\|X\|_*$ 是核范数, 即矩阵 X 的奇异值之和, X 、 Y 、 V 和 M 的维数相同。

进一步将公式 (2) 描述的问题转化为无约束最小化问题, 即最小化如下增广拉格朗日目标函数:

$$L(X, Y, Z, \beta) = \|X\|_* + \langle Z, X - Y \rangle + \frac{\beta}{2} \|X - Y\|_F^2 \quad (3)$$

其中, $Z \in R^{38 \times 38}$ 是线性约束的拉格朗日乘子; $\beta > 0$ 是惩罚参数; $\langle \cdot, \cdot \rangle$ 表示内积。

最后, 用 ADMM^[17-18] 求解优化问题 $\min L(X, Y, Z, \beta)$, 即先求解 Y 极小化问题; 再利用交替迭代的 Soft-impute 算法^[19] 求解 X 极小化问题; 最后迭代更新变量 Z 。更新规则如公式 (4)

所示:

$$\begin{cases} Y^{t+1} = \operatorname{argmin} L(X^t, Y^t, Z^t, \beta) \\ X^{t+1} = \operatorname{argmin} L(X^t, Y^{t+1}, Z^t, \beta) \\ Z^{t+1} = Z^t + \beta(X^{t+1} - Y^{t+1}) \end{cases} \quad (4)$$

具体执行过程如下:

第一步, 初始矩阵 X^t , Z^t 为全零矩阵, 迭代步

数 $t = 0$.

第二步, 交替更新变量 Y 、 X 和 Z .

a. 固定 X 和 Z , 利用公式 (5) 更新 Y :

$$Y^{t+1} = \begin{cases} M_{ij}, (i, j) \in \Omega \\ T_{ij}^{t+1}, \text{otherwise} \end{cases}, T^{t+1} = X^t + \frac{1}{\beta} Z^t \quad (5)$$

b. 固定 Y 和 Z , 利用公式 (6) 更新 X :

$$\begin{aligned} X^{t+1} &= U^{t+1} \hat{\Sigma}^{t+1} V^{t+1}, \hat{\Sigma}^{t+1} = D_{1/\beta}(\Sigma^{t+1}) = \\ &\operatorname{diag}\left\{\left(\sigma_j^{t+1} - \frac{1}{\beta}\right)_+, \Sigma^{t+1} = \right. \\ &\left. \operatorname{diag}(\sigma_1^{t+1}, \sigma_2^{t+1}, \dots, \sigma_{\rho^{t+1}}^{t+1}) \right\} \end{aligned} \quad (6)$$

其中, U 和 V 是 $Y^{t+1} - \frac{1}{\beta} Z^t$ 奇异值分解得到的正交矩阵, $D_{1/\beta}(\Sigma^{t+1})$ 为奇异值的软阈值化, 软阈值运

$$\operatorname{算}(\sigma_j - \frac{1}{\beta})_+ = \begin{cases} \sigma_j - \frac{1}{\beta}, & \text{若 } \sigma_j > \frac{1}{\beta} \\ 0, & \text{其他} \end{cases}, \text{这里 } \sigma_j \text{ 是 } \Sigma^{t+1}$$

对角线上的元素。

c. 固定 Y 和 X , 利用公式 (7) 更新 Z :

$$Z^{t+1} = Z^t + \gamma \beta (X^{t+1} - Y^{t+1}) \quad (7)$$

第三步, 满足迭代公式 (8), 算法终止。

$$\frac{\|X^{t+1} - X^t\|_F}{\|X^t\|_F} < \varepsilon \quad (8)$$

参数说明: 对于参数 β , 只要其大于 0 都可以保证 ADMM 的收敛性。借鉴文献 [20-21] 的做法, 限制对 β 的讨论, 以便清楚地揭示 ADMM 的主要思想, 这里取 $\beta = 2.5/\sqrt{mn}$ ($m=n=38$, 即观测矩阵的行数和列数)。参数 $0 < \gamma < (\sqrt{5} + 1)/2$ 。由于调节 γ 的大小对预测结果影响不大, 这里取 $\gamma=1.6$ 。迭代终止参数 $\varepsilon = 10^{-6}$ 。

为使观测矩阵 M 满足低秩、稀疏条件, 将 M 随机分成 10 折 $M1, \dots, M10$ (图 1), 每折维数均与 M 相同, 且只有非零空间的 10% 有观测值。分别把其中 9 折的并集 (如 $M - M10$) 输入到 NNRM 中, 得到部分预测值 (如 $X10$)。一次 10 折之后得到一个完整的预测矩阵 $X = X1 + X2 + \dots + X10$, 其中缺失值部分取 10 折结果的平均值。

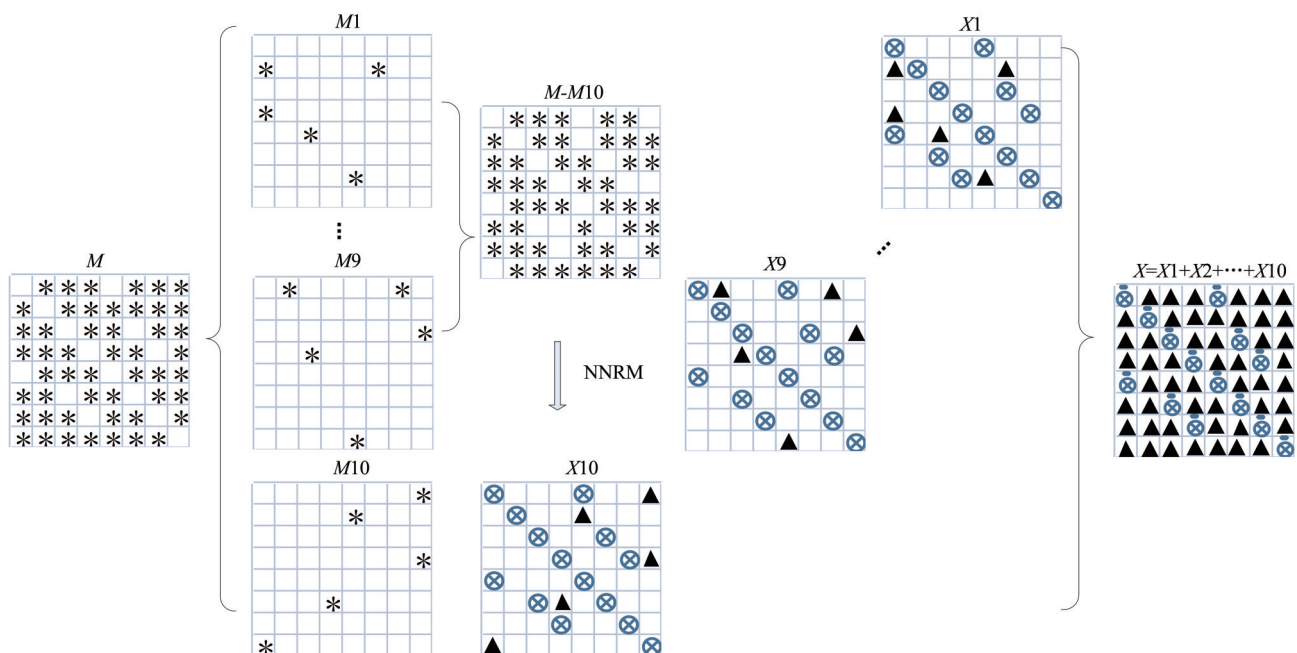


Fig. 1 Ten-fold-based synergy prediction

* Represents the observed synergy score, $\blacktriangle$ represents the predicted synergy score corresponding to the non-zero space, $\otimes$ represents the predicted synergy score corresponding to the missing value, $\oplus$ represents the average predicted synergy score corresponding to the missing value.

1.2.2 模型预测性能的评估指标

利用均方误差 (MSE)、均方根误差 (RMSE)、皮尔森相关系数 (PCC) 和斯皮尔曼秩相关系数 (SCC) 衡量模型的回归性能。

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2} \quad (10)$$

$$PCC = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (11)$$

$$SCC = \frac{\sum_{i=1}^N (s_i - \bar{s})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^N (s_i - \bar{s})^2 \sum_{i=1}^N (t_i - \bar{t})^2}} \quad (12)$$

其中, N 为药物组合数, x_i 为第 i 个药物组合的预测协同得分, y_i 为第 i 个药物组合的观测协同得分, $\bar{x}$ 为 N 个药物组合的预测协同得分的均值, $\bar{y}$ 为 N 个药物组合的观测协同得分的均值, s_i 表示第 i 个药物组合在预测协同得分排序中的位序, t_i 表示第 i 个药物组合在观测协同得分排序中的位序, $\bar{s}$ 和 $\bar{t}$ 表示所有位序的均值。SCC 是 PCC 的一种非参数形式, 它衡量了预测向量 (以 x_i 为分量) 和观测向量 (以 y_i 为分量) 之间的依赖性。

高度协同的药物组合因其应用价值高而备受关注。采用与文献 [9] 相同的定义方式, 将协同得分不低于 30 的药物组合定义为高度协同, 低于 30 的为非高度协同。这里利用准确率 (ACC)、受试者工作特征曲线下的面积 (ROC_AUC)、PR 曲线下的面积 (PR_AUC) 进一步衡量模型的分类性能:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

其中, (TN) TP 为 (非) 高度协同的药物组合被预测为 (非) 高度协同的组合数量, (FP) FN 为 (非) 高度协同的药物组合被预测为 (高度) 非高度协同的组合数量。

2 结 果

针对每个细胞系构造的对称观测矩阵, 一次 10 折后得到一个预测矩阵, 随机分折操作重复 10 次, 得到 10 个预测矩阵 X , 取这 10 个矩阵的平均值得到“平均预测矩阵”。结果显示, 平均预测矩阵对称位置上的预测得分差别不大。为进一步使其满足“对称”特点, 将平均预测矩阵上、下三角对称位置上的元素赋值为二者预测得分的平均值, 得到“最终预测矩阵”。为方便评估指标的计算, 分别将观测矩阵和最终预测矩阵中的上三角及对角线元素赋值为 0, 下三角元素不变, 依次“连接”矩阵的各列, 得到两个 1 444 维的列向量; 再将观测矩阵对应的列向量中的零元去除, 得到压缩的 583 维“观测向量”, 依据观测向量中每个分量的具体位置, 得到压缩的 583 维“预测向量”; 最后, 依据两个向量计算各项评估指标。为保证预测的稳定性, 对每个细胞系, 通过不同的药物排列顺序构造了 3 个观测矩阵, 重复执行实验 3 次。结果显示, 3 次实验的计算结果差别不大, 故将 3 次结果的平均值作为最终预测结果。

2.1 整体预测结果分析

583 种药物组合对于 39 个细胞系的协同得分预测值与观测值的分布如图 2a 所示。尽管存在部分异常点, 但整体上二者之间存在较高的相关性 ($PCC=0.73$), 而预测值的绝对值普遍低于观测值的绝对值。

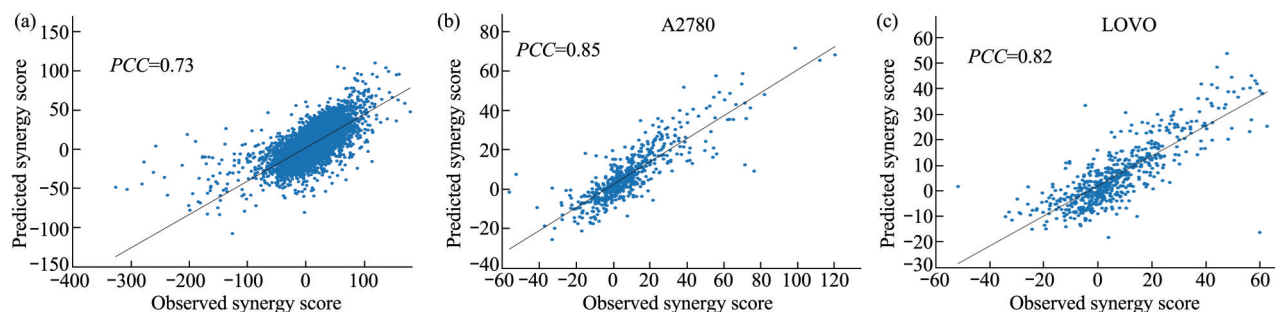


Fig. 2 The correlation analysis between predicted and observed synergy scores based on 583 drug combinations

(a) The distribution of predicted and observed synergy scores to 39 cell lines. (b) The distribution of predicted and observed synergy scores to the cell line A2780. (c) The distribution of predicted and observed synergy scores to the cell line LOVO.

表1展示了NNRM与5种模型的预测结果(各指标均为平均值, MulinputSynergy仅计算了31个细胞系, 故未纳入比较)。NNRM在6个指标上均优于随机森林(RF)和支持向量机(SVM); 相比DeepSynergy, NNRM在3个指标上均有提升(MSE、RMSE、ACC), 其余3个指标持平; 相比Deep learning+PCA, NNRM仅在1个指标(RMSE)上有优势; 相比AuDNNsynergy, NNRM

仅在2个指标(RMSE、ACC)上有优势。需要强调的是, 这5种模型均使用了大量的辅助特征。其中, AuDNNsynergy使用了93 019个辅助特征^[12]; Deep learning+PCA使用了超过14 000个辅助特征^[10]; DeepSynergy、RF和SVM使用了8 371个辅助特征^[9]。可见, NNRM在没有使用任何辅助特征的情况下, 以低计算成本获得了与深度学习相媲美的预测结果。

Table 1 The prediction result comparison between NNRM and other four models

Model	MSE	RMSE	PCC	ACC	ROC_AUC	PR_AUC
NNRM	252.64	14.78	0.73	0.94	0.90	0.59
AuDNNsynergy	241.00	15.46	0.74	0.93	0.91	0.63
Deep learning+PCA	241.20	15.46	0.76	—	—	—
DeepSynergy	255.49	15.91	0.73	0.92	0.90	0.59
RF	307.56	17.49	0.65	0.92	0.87	0.55
SVM	398.39	19.92	0.50	0.76	0.81	0.42

2.2 基于细胞系的预测结果分析

39个细胞系的MSE的变化范围为68.53~945.08(图3a), 5个(13%)细胞系大于450, 67%的细胞系小于200; 39个细胞系的PCC范围为0.48~0.87, 只有2个(5%)细胞系小于0.60, 67%的细胞系大于0.70(图2b, c); 39个细胞系的SCC范围为0.59~0.83, 平均值为0.73, 其中9个(23%)细胞系大于0.80, 64%的细胞系大于0.70, 97%的细胞系大于0.60。AuDNNsynergy和Deep learning+PCA的SCC范围为0.56~0.81, 平均值为0.73, 55%的细胞系大于0.70, 92%的细胞系大于0.60。综上, NNRM对于大部分细胞系的预测结果较好, 且SCC指标整体优于AuDNNsynergy和Deep learning+PCA。

进一步将39个细胞系按组织分为7类: LUNG(肺, 包含7个细胞系)、BREAST(乳腺, 包含6个细胞系)、LARGE_INTESTINE(大肠, 包含8个细胞系)、SKIN(皮肤, 包含6个细胞系)、PLEURA(胸膜, 包含1个细胞系)、OVARY(卵巢, 包含9个细胞系)和PROSTATE(前列腺, 包含2个细胞系)。7个组织的PCC(该组织下所有细胞系的PCC的平均值)分别为0.72、0.75、0.70、0.67、0.67、0.79、0.74; SCC分别为0.74、0.73、

0.70、0.70、0.67、0.77、0.72。由图3b可以看出, 在组织层面, NNRM在这两个指标下的预测趋势基本一致; 7个组织的整体预测结果较好, 其中OVARY的预测结果最理想, PLEURA的预测结果最差。因PLEURA仅包含1个细胞系, 故不能说明NNRM不适于预测该组织。结合图3a发现, 属于同一组织的细胞系, NNRM的预测结果不尽相同, 甚至会出现较大的差异, 如LUNG组织中细胞系NCIH2122和NCIH23的MSE分别为945.08和899.66, 而NCIH520为98.99。这种现象同样出现在AuDNNsynergy中, 但是NCIH2122在Deep learning+PCA和DeepSynergy中的预测相对来说比较理想。

2.3 基于药物的预测结果分析

由于3次实验的计算结果差别不大, 故针对每种药物, 依据实验一的计算结果进行分析。分别以38种药物为固定成员的所有药物组合与39个细胞系的预测值与观测值的MSE分布如图4a所示。38种药物的MSE变化范围为67.65~624.05, 7种(18%)药物大于400, 47%的药物小于200; 38种药物的PCC范围为0.55~0.79, 2种(5%)药物低于0.60, 42%的药物高于0.70。特别地, 包含药物SUNITINIB的组合对于4个细胞系(占比10%)的

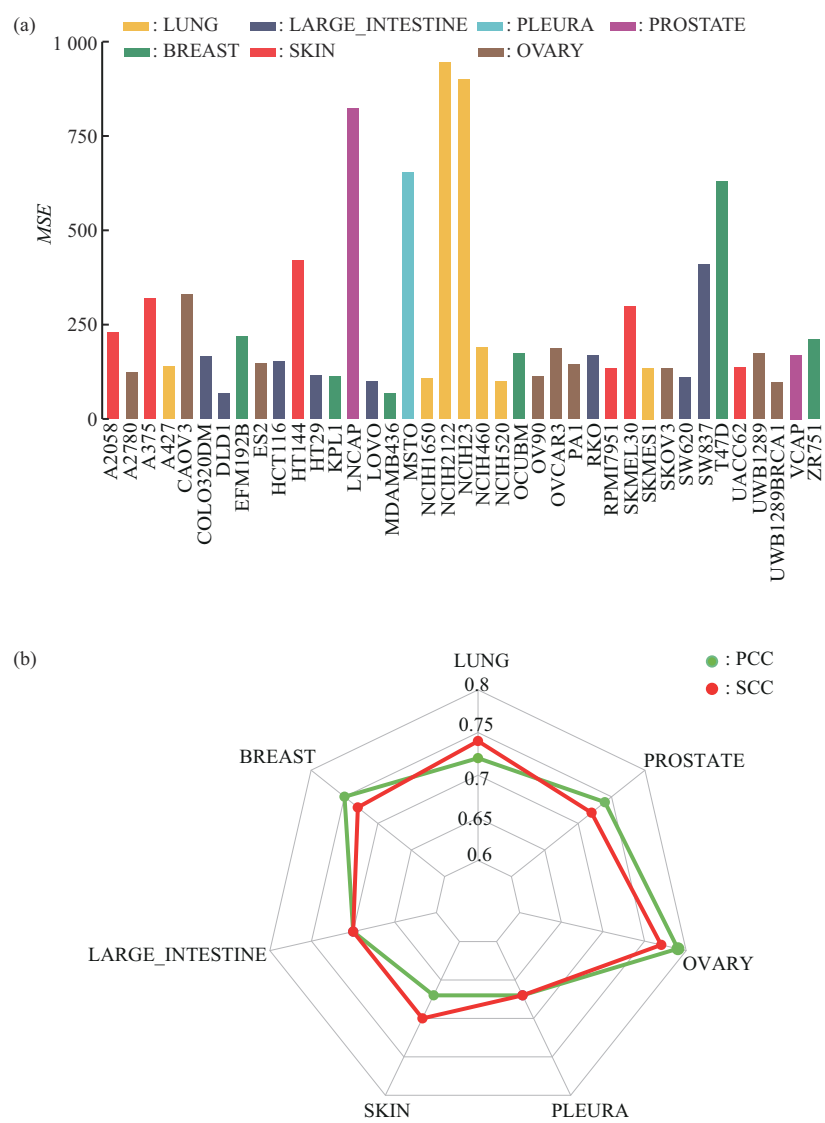


Fig. 3 The prediction results of 39 cell lines
(a) MSE distribution of 39 cell lines. (b) The distributions of PCC and SCC of seven types of tissues.

预测 *PCC* 均高于 0.78, 包含药物 METFORMIN、VINORELBINE 和 MK-8669 的组合对于 3 个细胞系 (占比 8%) 的预测 *PCC* 均高于 0.83。38 种药物的 *SCC* 变化范围为 0.54~0.76, 5 种 (13%) 药物低于 0.60, 47% 的药物高于 0.70。综上, NNRN 对于大部分药物的预测结果较好, 但整体低于 AuDNNsynergy 和 Deep learning+PCA。

将 38 种药物按作用的靶点类型分为 8 类, EGFR (包含 2 种药物)、PARP (包含 2 种药物)、MULT-KINASE (包含 3 种药物)、TOPOISOMERASE II (包含 2 种药物)、DNA (包含 6 种药物)、MICROTUBULES (包含 3 种药物)、TOPOISOMERASE I (包含 2 种药物) 和 OTHER

(包含 18 种药物)。8 类靶点的 *PCC* (该靶点下所有药物的 *PCC* 的平均值) 分别为 0.67、0.59、0.73、0.77、0.64、0.67、0.67、0.69; *SCC* 分别为 0.68、0.57、0.74、0.73、0.62、0.73、0.65、0.68。在药物靶点层面, NNRN 在这两个指标下的预测趋势基本一致; 除靶点 PARP 外, 其他靶点的预测结果较为理想 (图 4b)。属于同一靶点的药物, NNRN 的预测结果不尽相同, 甚至会出现较大的差异, 如属于靶点 MULT-KINASE 的药物 SUNITINIB 和 DASATINIB 的 *MSE* 分别为 67.65 和 457.36 (图 4a)。这种现象同样出现在 AuDNNsynergy 和 Deep learning+PCA 中。

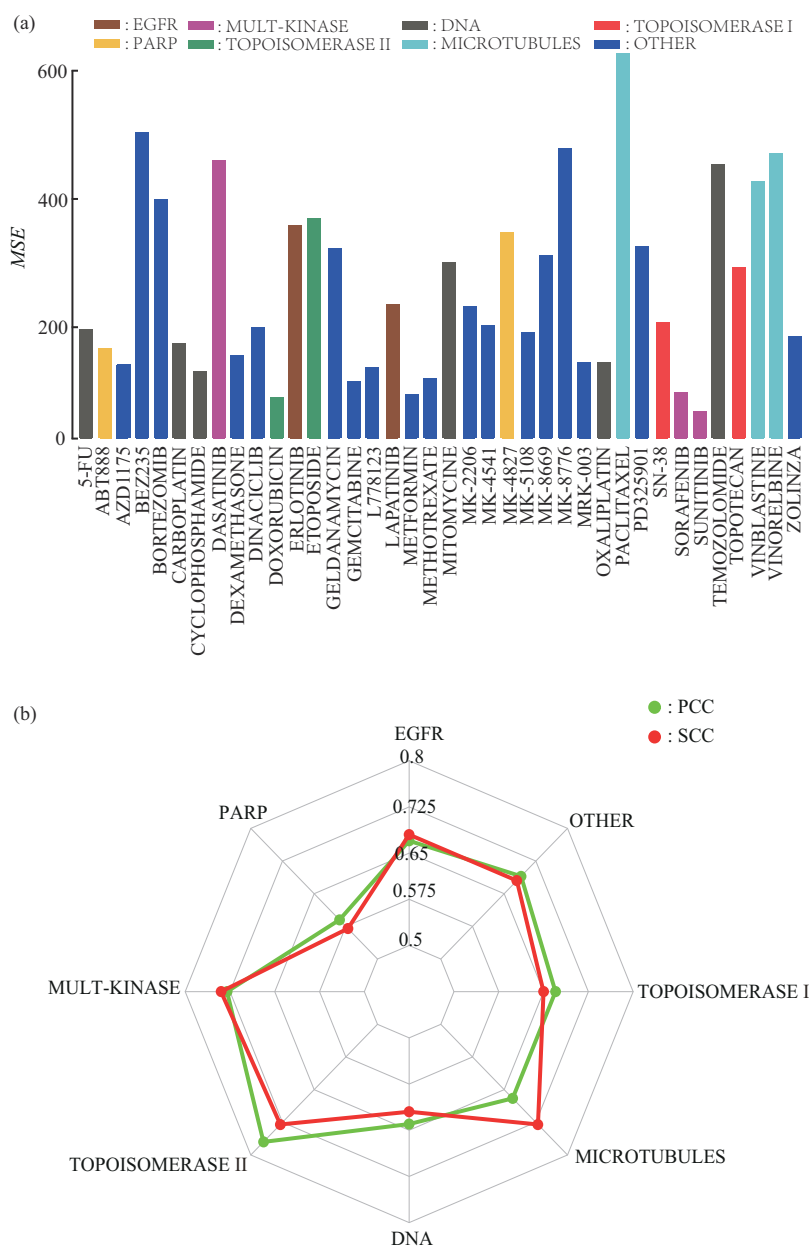


Fig. 4 The prediction results of 38 drugs

(a) MSE distribution of 38 drugs. (b) The distributions of PCC and SCC of eight types of targets.

2.4 高通量虚拟筛选结果分析

针对每个细胞系，将药物组合协同值按照从大到小的顺序排序，依据排序定义两个指标 TopPN 和 TopON，分别为预测的前 N 个有效组合中包含真实最佳药物组合（协同得分最高）的细胞系数与总细胞系数之比，以及观测的前 N 个有效药物组合中包含预测最佳药物组合的细胞系数与总细胞系数之比。分别取 N 为 1、3、5、20、40、60、90

（表2）。对于每个细胞系，只要测试其预测排名前 90 的药物组合，就能以 90% 的概率筛选出真实的最佳药物组合，对比原本需要测试 583 个药物组合，减少 85% 的测试工作量；90% 的细胞系预测的最佳药物组合均出现在其观测药物组合的前 40，对比原本需要测试 583 个药物组合，减少 93% 的测试工作量。结果表明，NNRM 有助于高通量虚拟筛选，大大降低了测试成本。

Table 2 Virtual screening performance of the NCRM model

N	1	3	5	20	40	60	90
TopPN	3%	18%	26%	54%	74%	82%	90%
TopON	2%	21%	38%	74%	90%	95%	97%

2.5 缺失值预测结果分析

对于缺失观测值的药物组合，NCRM 预测出部分组合对于来自卵巢组织的特定细胞系产生（高度）协同作用。例如，3 个组合依托泊苷（ETOPOSIDE）+地塞米松（DEXAMETHASONE）、依托泊苷+卡铂（CARBOPLATIN）、依托泊苷+环磷酰胺（CYCLOPHOSPHAMIDE）均对细胞系 CAO V3 高度协同，预测得分分别为 75.0、49.5、37.6。NCRM 预测出地塞米松还与其他多种药物同样产生了（高度）协同作用，这与 2020 版美国临床肿瘤学会指南“支持地塞米松可作为免疫检查点抑制剂联合化疗的预防性止吐药物之一^[22]”相呼应。研究发现，以依托泊苷+地塞米松为基础的方案可用于治疗结外 NK/T 细胞淋巴瘤^[23]，是否可用于治疗卵巢癌还未见相关报道，值得进一步研究。卵巢恶性肿瘤诊断与治疗指南（2021 年版）明确指出，依托泊苷+卡铂适用于 IB~III 期无性细胞肿瘤术后患者，以及亟需降低化疗毒性的部分患者，依托泊苷+异环磷酰胺（环磷酰胺异构体）可作为复发卵巢恶性生殖细胞肿瘤的二线化疗方案^[24]，证实了本文的预测结果。此外，长春瑞滨（VINORELBINE）+紫杉醇（PACLITAXEL）对细胞系 OV90 的预测得分为 33.0，长春碱（VINBLASTINE，与长春瑞滨的抗癌机制相同）+紫杉醇对细胞系 SKOV3 和 ES2 的预测得分分别为 20.6、20.5，这些预测结果与“长春瑞滨和紫杉醇可用作铂敏感复发上皮性卵巢癌的二线化疗备选方案^[24]”相吻合。依托泊苷+吉西他滨（GEMCITABINE）对细胞系 OVCAR3 的预测得分为 14.8，二者联合应用可作为复发性卵巢癌患者化疗的新选择^[25]。

其他组织也有类似的预测结果。如紫杉醇+阿霉素（DOXORUBICIN）对于来自乳腺组织的细胞系 NCIH460 的预测得分为 10.65，已有文献报道该组合对于治疗晚期乳腺癌效果明显^[26]。另外两种组合依托泊苷+卡铂、紫杉醇+羟基喜树碱（SN-38）分别对来自乳腺组织的细胞系 KPL1 和 SKMES1 产生较高的预测得分（23.85、20.25）。已

知依托泊苷+卡铂是小细胞肺癌的标准治疗方案^[27]，紫杉醇+羟基喜树碱治疗晚期食道癌疗效确切^[28]，这两种组合是否适用于乳腺癌的治疗还有待进一步验证。二甲双胍（METFORMIN）+羟基喜树碱对来自皮肤组织的细胞系 RPMI7951 的预测得分为 11.15，由这两种药物自组装的纳米材料已被证实具有免疫和化疗协同治疗功能^[29]。拓扑替康（TOPOTECAN）+紫杉醇对于来自肺组织的细胞系 A427 的预测得分为 14.45，两者联合治疗能有效改善老年晚期非小细胞肺癌患者的生存质量与功能状态^[30]，是一种潜在的治疗方案^[31]。依托泊苷+环磷酰胺对于来自前列腺组织的细胞系 VCAP 的预测得分为 12.85，研究表明两者联合治疗雄激素抵抗性前列腺癌有一定疗效^[32]。

3 讨 论

为探讨观测矩阵的构造形式对 NCRM 预测结果的影响，将 38 种药物以不同顺序进行横、纵排列，得到非对称观测矩阵。同样地，每个细胞系均构造 3 个非对称观测矩阵，重复执行实验 3 次，以 3 次结果的均值作为最终预测结果，发现其与对称观测矩阵的预测结果几乎一致。例如，细胞系 A2058 在两种情况下的预测 PCC 均为 0.71，MSE 分别为 231.51（非对称矩阵）和 231.55（对称矩阵），说明 NCRM 的预测不受观测矩阵构造形式的影响。由于对称矩阵的处理和计算更简便，故文中选择对称的观测矩阵进行实验。

为探讨观测矩阵的稀疏性对 NCRM 预测结果的影响，分别进行了将观测矩阵随机分为 3 折、5 折和 8 折的实验。以 10 折法中 PCC 最高的前 7 个细胞系和最低的后 7 个细胞系为例，不同折数对 14 个细胞系的 PCC 影响不大，但 MSE 有些许差别（表 3）。可见，观测数据在不同缺失率下对预测结果有一定影响。总体来说，折数越少，输入数据对应的观测矩阵越稀疏，模型的部分预测指标也会有所降低。

Table 3 Prediction result comparison of 14 representative cell lines with different folds

Fold number	Top 7 cell lines		Last 7 cell lines	
	MSE	PCC	MSE	PCC
3	150.34	0.82	484.95	0.56
5	136.29	0.83	465.80	0.58
8	130.41	0.84	456.85	0.59
10	128.78	0.84	456.42	0.58

VINORELBINE、DASATINIB、BEZ-235、BORTEZOMIB、GELDANAMYCIN), 其对应列的标准差大于40。图5b和5c分别展示了NCIH23的观测矩阵中16种补充组药物和22种穷举组药物对应列的观测值分布。众所周知, 在进行细胞系-药物反应实验时可能会出现异常值^[33], 进而影响药物协同计算, 导致部分协同观测值异常。为使细胞系的预测结果更稳定, 尝试将标准差大于40的列中的异常值删除。由于补充组药物的观测数据个



(a) The heat map of PCC among 39 cell lines based on observed synergy scores. (b) The distribution of observed synergy scores of 16 drugs in the supplemental group for the cell line NCIH23. (c) The distribution of observed synergy scores of 22 drugs in the exhaustive group for the cell line NCIH23.

数还不到穷举组药物的 3/5, 所以将大于或 $Q3+1.5IQR$ 小于 $Q1-1.5IQR$ 的值认定为补充组药物的异常值, 将大于 $Q3+3IQR$ 或小于 $Q1-3IQR$ 的值认定为穷举组药物的异常值 (图 5b, c), 这里 $Q3$ 为上四分位数, $Q1$ 为下四分位数, $IQR=Q3-Q1$ 为四分位数差。分别计算删除异常值前后这 10 个细

胞系的观测值 (不包含异常值) 与对应预测值的 MSE 和 PCC , 发现 8 个细胞系的 MSE 下降, 1 个细胞系升高, 1 个细胞系保持不变; 6 个细胞系的 PCC 升高, 1 个细胞系降低, 3 个细胞系保持不变 (表 4)。可见, 删除异常值能够有效改善 NNRM 的预测效果。

Table 4 The changes of evaluation indexes of 10 representative cell lines before and after removing outliers					
Cell lines	Outlier removal	PCC	PCC status	MSE	MSE status
A375	No	0.64	Up	276.38	Down
	Yes	0.68		256.13	
CAOV3	No	0.80	—	314.51	Down
	Yes	0.80		308.59	
HT144	No	0.61	Up	222.14	Down
	Yes	0.69		189.68	
LNCAP	No	0.67	Down	654.21	Up
	Yes	0.63		721.01	
MSTO	No	0.72	Up	511.18	Down
	Yes	0.73		510.62	
NCIH23	No	0.59	Up	441.50	Down
	Yes	0.63		421.31	
NCIH2122	No	0.64	Up	322.43	Down
	Yes	0.67		310.04	
SKMEL30	No	0.62	—	295.40	—
	Yes	0.62		295.40	
SW837	No	0.71	Up	218.71	Down
	Yes	0.73		207.79	
T47D	No	0.77	—	605.74	Down
	Yes	0.77		597.83	

4 结 论

NNRM 降低了已有模型对数据的要求, 单纯利用部分药物组合的观测协同得分, 即可实现大规模、批量预测抗癌药物组合的协同作用和协同状态, 不仅取得了与深度学习技术相媲美的预测结果, 而且极大地降低了模型的计算成本, 缩短了高通量虚拟筛选的测试时间, 预测的部分缺失值结果与已有研究或临床实践相吻合。以上说明, NNRM 可以作为抗癌药物组合高通量虚拟筛选的可选择工具。后续可融合药物基因组学数据, 进一步提升 NNRM 的预测性能和适用范围。

参 考 文 献

[1] Assaraf Y G, Brozovic A, Goncalves A C, *et al.* The multi-factorial nature of clinical multidrug resistance in cancer. *Drug Resist Update*, 2019, **46** (Suppl 1): 100645

[2] Pemovska T, Bigenzahn J W, Superti-Furga G. Recent advances in combinatorial drug screening and synergy scoring. *Curr Opin Pharmacol*, 2018, **42**: 102-110

[3] Spain L, Julve M, Larkin J. Combination dabrafenib and trametinib in the management of advanced melanoma with BRAFV600 mutations. *Expert Opin Pharmaco*, 2016, **17**(7): 1031-1038

[4] Paltun B G, Kaski S, Mamitsuka H. Machine learning approaches for drug combination therapies. *Brief Bioinform*, 2021, **22**(1): bbab293

[5] Bansal M, Yang J, Karan C, *et al.* A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol*, 2014, **32**(12): 1213-1222

[6] Menden M P, Busulu K, Guan Y F, *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*, 2019, **10**(1): 2674

[7] O'Neil J, Bentia Y, Feldman L, *et al.* An unbiased oncology compound screen to identify novel combination strategies. *Mol Cancer Ther*, 2016, **15**(6): 1155-1162

[8] Holbeck S L, Camalier R, Crowell J A, *et al.* The National Cancer

- Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res*, 2017, **77**(13): 3564-3576
- [9] Preuer K, Lewis R, Hochreiter S, *et al.* DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics*, 2018, **34**(9): 1538-1546
- [10] Ma J, Motsinger-Reif A. Prediction of synergistic drug combinations using PCA-initialized deep learning. *BioData Min*, 2021, **14**(1): 46
- [11] 陈希, 秦玉芳, 陈明, 等. 基于多输入神经网络的药物组合协同作用预测. *生物医学工程学杂志*, 2020, **37**(4): 676-682
Chen X, Qin Y F, Chen M, *et al.* *J Biomed Eng*, 2020, **37**(4): 676-682
- [12] Zhang T, Zhang L, Payne P, *et al.* Synergistic drug combination prediction by integrating multiomics data in Deep Learning models. *Methods Mol Biol*, 2021, **2194**: 223-238
- [13] Veroli D G Y, Fornari C, Wang D, *et al.* Combeneft: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics*, 2016, **32**(18): 2866-2868
- [14] Loewe S. The problem of synergism and antagonism of combined drugs. *Arzneimittelforsch*, 1953, **3**(6): 285-290
- [15] Candes Emmanuel J, Tao T. The power of convex relaxation: near-optimal matrix completion. *IEEE T Inform Theory*, 2010, **56**(5): 2053-2080
- [16] Chen C, He B, Yuan X. Matrix completion via an alternating direction method. *IMA J Numer Anal*, 2012, **32**(1): 227-245
- [17] Buccini A. Fast alternating direction multipliers method by generalized Krylov subspaces. *J Sci Comput*, 2022, **90**(1): 1-23
- [18] Liu T, Sun M, Liu Y, *et al.* ADMM based low-rank and sparse matrix recovery method for sparse photoacoustic microscopy. *Biomed Signal Proces*, 2019, **52**(JUL.): 14-22
- [19] Cai J F, Candes Emmanuel J, Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM J Optimiz*, 2010, **20**(4): 1956-1982
- [20] Kontogiorgis S, Meyer R R. A variable-penalty alternating directions method for convex optimization. *Math Program*, 1998, **83**(1-3): 29-53
- [21] He B S, Yang H, Wang S L. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *J Optimiz Theory Appl*, 2000, **106**(2): 337-356
- [22] 潘半舟, 沈波. 《2020年美国临床肿瘤学会呕吐临床诊疗指南》解读. *中国医学前沿杂志(电子版)*, 2020, **12**(10): 28-35
Pan B Z, Shen B. *China J Front Med Sci (Electronic Version)*, 2020, **12**(10): 28-35
- [23] Wei L, Yang L, Cong J, *et al.* Using etoposide + dexamethasone-based regimens to treat nasal type extranodal natural killer/T-cell lymphoma-associated hemophagocytic lymphohistiocytosis. *J Cancer Res Clin*, 2021, **147**(1): 863-869
- [24] 中国抗癌协会妇科肿瘤专业委员会. 卵巢恶性肿瘤诊断与治疗指南(2021年版). *中国癌症杂志*, 2021, **31**(6): 490-500
Chinese Anti-Cancer Association Gynecological Oncology Committee. *China Oncol*, 2021, **31**(6): 490-500
- [25] 甘立菁, 王貽军, 曾哲超, 等. 吉西他滨联合依托泊苷治疗复发性卵巢癌效果观察. *肿瘤研究与临床*, 2014, **26**(8): 563-565
Gan L J, Wang Y J, Zeng Z C, *et al.* *Cancer Res Clin*, 2014, **26**(8): 563-565
- [26] Denlinger C S, Goldstein L J. Randomized parallel study of doxorubicin plus paclitaxel and doxorubicin plus cyclophosphamide as neoadjuvant treatment of patients with breast cancer. *J Clin Oncol*, 2005, **16**(3): 277-277
- [27] Nasr F, Khoury R E, Yehia I, *et al.* Small cell lung cancer: real-world data from two institutions in Lebanon for patients receiving carboplatin etoposide or atezolizumab in first line of treatment. *J Clin Oncol*, 2021, **39**(15_suppl): e20578
- [28] 林尤恩, 李立文, 林冬群. 紫杉醇联合羟基喜树碱治疗晚期食管癌的临床疗效观察. *中外医疗*, 2009, **28**(21): 18-20
Lin Y E, Li L W, Lin D Q. *China Foreign Med Treat*, 2009, **28**(21): 18-20
- [29] 卢春华, 蔡淑贤, 钟武坤, 等. 一种盐酸二甲双胍和7-乙基-10-羟基喜树碱自组装的纳米材料: 中国, CN201910872968.6. 2019-11-15
Lu C H, Cai S X, Zhong W K, *et al.* A self-assembled nanomaterial of metformin hydrochloride and 7-ethyl-10-hydroxycamptothecin: China, CN201910872968.6. 2019-11-15
- [30] 吕红梅, 王翠英. 白蛋白紫杉醇单用与联合拓扑替康治疗老年晚期非小细胞肺癌的疗效比较. *世界最新医学信息文摘*, 2017, **17**(22): 66
Lü H M, Wang C Y. *World Latest Med Inf*, 2017, **17**(22): 66
- [31] Eckardt, John R. Feasibility of oral topotecan plus intravenous paclitaxel in advanced non-small-cell lung cancer. *Oncology*, 2001, **61**(1): 30-34
- [32] 钱伟庆, 孙忠全, 宋建达. 足叶乙贰联合环磷酰胺治疗雄激素抵抗性前列腺癌. *中华泌尿外科杂志*, 2005, **26**(4): 267-269
Qian W Q, Sun Z Q, Song J D. *Chin J Urol*, 2005, **26**(4): 267-269
- [33] Ianevski A, Giri A K, Gautam P, *et al.* Prediction of drug combination effects with a minimal set of experiments. *Nat Mach Intell*, 2019, **1**(12): 568-577

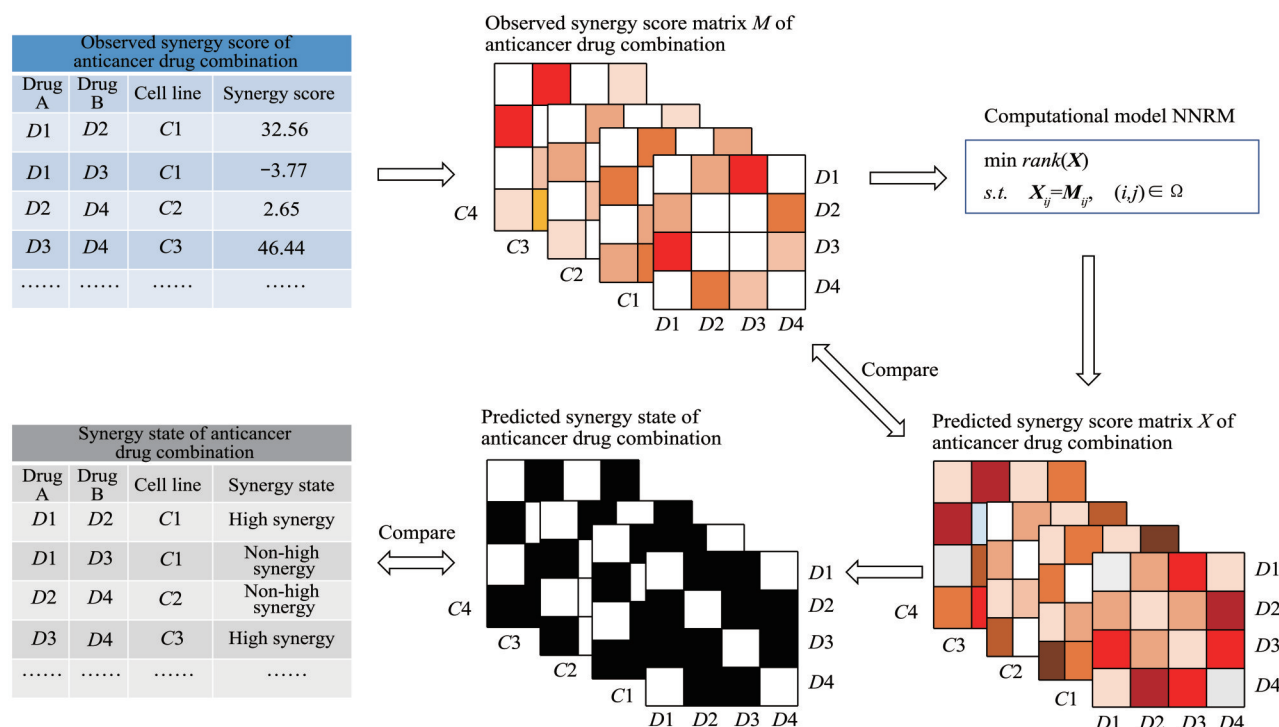
Predicting Synergistic Effect of Anticancer Drug Combinations Based on Nuclear Norm Regularization*

SHI Lei-Jing¹⁾, WANG Bo¹⁾, ZHANG Shan^{1,2)}, REN Fu-Quan¹⁾, Li Yu-Shuang^{1)**}

¹⁾School of Science, Yanshan University, Qinhuangdao 066004, China;

²⁾CRRRC TANGSHAN CO., LTD, Tangshan 063000, China)

Graphical abstract



Abstract Objective Anticancer drug combination therapies are a promising therapeutic strategy. Drug combinations exhibiting a highly synergistic effect are crucial to improve the treatment for specific cancers. However, identifying such combinations is very complicated and difficult due to the tremendous screening cost. The availability of large-scale high-throughput combination screening data provides opportunities for computational approaches. The purpose of this study is to optimize the high-throughput virtual screening of anticancer drug combinations in a completely data-driven and computational modeling way, and provide theoretical reference for “old drugs repositioning as new combinations”. **Methods** Inspired by the matrix completion, we present a nuclear norm regularization-based model, termed NNRM, to predict synergy scores and

* This work was supported by grants from Natural Science Foundation of Hebei, China (A2020203021) and Science Foundation for Returned Scholars of Hebei Province (C20200365).

** Corresponding author.

Tel: 86-15232356135, E-mail: yushuangli@ysu.edu.cn

Received: May 29, 2022 Accepted: July 15, 2022

synergy status of anticancer drug combinations. Symmetric observation matrixes of synergy scores were constructed for given cell lines; a folding technique was employed to sparse the observation matrix; alternating direction multiplier method and soft threshold estimation were applied to solve the model. **Results** NNRM achieved expected predictive result on the dataset released by O'Neil's team, the root mean square error of the synergy score prediction was 14.78, and the accuracy of the synergy status prediction was 0.94. It is not only significantly superior to the Random Forest and the Support Vector Machine, but also completely comparable to the state-of-the-art deep learning models including DeepSynergy, Deep learning+PCA and AuDNNsynergy. Moreover, NNRM effectively filled the missing synergy scores most of which are consistent with existing research or clinical practice. **Conclusion** NNRM could predict the synergistic effect of large-scale drug combinations in batches, which greatly lowers the data requirements by existing models, reduces the computational cost, and shortens the screening time. It indicates that NNRM is an alternative tool for high-throughput virtual screening of anticancer drug combinations.

Key words anticancer drug combination, synergistic effect, high-throughput virtual screening, matrix completion, nuclear norm regularization, computational model

DOI: 10.16476/j.pibb.2022.0252