多组学-深度神经网络-药物预测

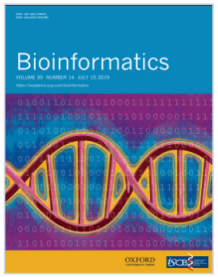
汇报人：Lilian

2022/2/18

文献来源

Bioinformatics

Issues Advance articles Submit ▼ Purchase Alerts About ▼ All Bioinforma



Volume 35, Issue 14
July 2019

MOLI: multi-omics late integration with deep neural networks for drug response prediction

Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins ✉, Martin Ester ✉

Bioinformatics, Volume 35, Issue 14, July 2019, Pages i501–i509,
https://doi.org/10.1093/bioinformatics/btz318
Published: 05 July 2019

MOLI：使用深度神经网络进行多组学数据后期集成以用于药物反应预测

研究背景

- ✓ 精准肿瘤学是利用基因组数据为肿瘤患者量身定做治疗方案。目前，只有11%接受精准肿瘤学治疗的患者能够进入临床试验，只有5%的患者从精准肿瘤学中获益。

精准肿瘤学的概念及现状

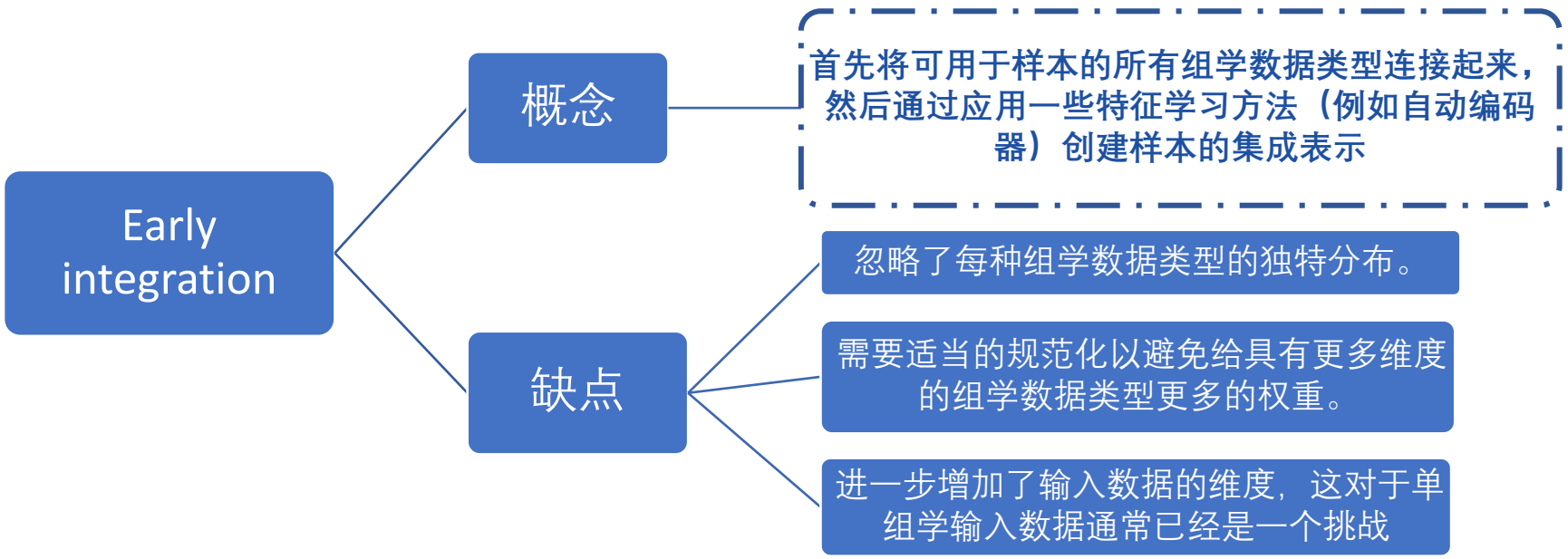
- ✓ 药物反应预测的最终目的是检测药物的临床效用，而药物反应面临的一大挑战就是体内数据集没有足够的患者记录和药物反应信息；**药物预测中信息量最大的数据类型是基因表达。**有研究表明，**整合额外组学数据可以提高预测精度，由此引出了如何整合额外组学数据的问题。**

药物反应预测的目的及现状

研究背景

✓ 多组学数据研究面临的一大问题就是如何将数据进行有效的整合，目前主要有两种方法:**早期集成（early integration）和晚期集成（late integration）**。

多组学数据整合的方法



研究背景

Late integration

概念

针对每种组学数据类型分别学习特征，然后将这些特征集成到一个统一的表示中，用作分类器或回归器的输入

特点

优点是它适用于每种组学数据类型的唯一分布，它可以对每种数据类型采用单组学规范化，并且不会增加输入空间的维数。

本文工作

本文提出了一种基于**深度神经网络**的多组学**晚期整合**方法MOLI。

MOLI将**体细胞突变、拷贝数畸变和基因表达数据**作为输入，预测给定药物的反应。

输入

MOLI包含多个前馈编码子网络，每个编码子网络输入相应的组学数据。将编码子网络中学习到的特征拼接成一种表示。将拼接的表示作为分类子网络的输入，用于预测药物反应。整个网络以端到端的方式进行训练。

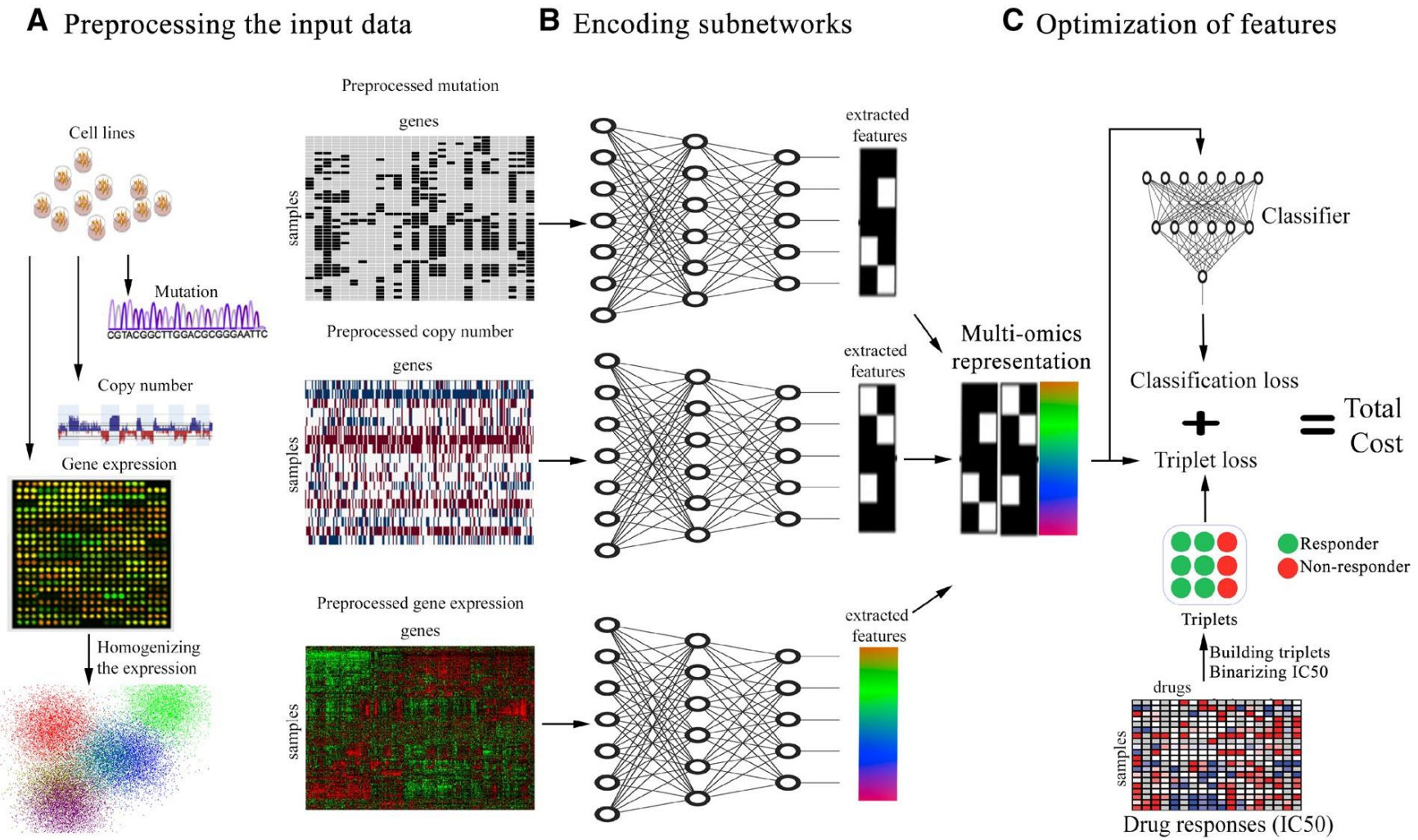
分类子网络的成本函数结合了三元损失和二元交叉熵损失。前者使得响应样本间的表达更相似，响应样本与非响应样本的表示更不同，后者使得这种表示对IC50值更有预测性。

成本函数

MOLI是第一个使用深度神经网络的端到端晚期集成方法。

- 整体工作流程如下：

每个编码子网络学习其组学数据类型的特征，并将学习到的特征连接。



多组学数据预处理

预测药物反应的分类器子网络

Data

数据集

1.肿瘤药物敏感性基因组学（GDSC）数据集

包含1000多个癌细胞系的多组学数据和265种靶向和化疗药物的反应数据。

2.异种移植（PDX）百科全书数据集

包含300多个不同癌症类型的PDX模型及34种靶向和化疗药物的反应数据。

3.TCGA数据集

包含一万多名不同癌症类型患者的肿瘤样本的谱数据及部分患者药物反应。

数据预处理

Gene expression profiles

基因表达数据

原始强度从 ArrayExpress (E-MTAB-3610) 获得，用于 GDSC 数据集进行 RMA 标准化 、对数转换并聚合到基因水平。

PDX 和所有 TCGA 数据集的基因表达值被转换为 TPM 并进行对数转换。PDX 样本的 FPKM 值转换为 TPM 并进行对数转换。

在每个数据集中，我们排除了方差最小的 5% 基因，假设它们没有提供信息。

基因表达：将基因表达的值进行了标准化

数据预处理

Somatic copy number profiles

体细胞拷贝数

从 TCGA 数据集的基因组分割文件中删除不可靠的片段，并为每个基因分配一个与其重叠片段的强度对数比相对应的值。如果基因重叠多个片段，则保留最极端的对数比值。

与 TCGA 不同，GDSC 和 PDX 数据集提供了总拷贝数的基因水平估计。为了使这些数据与 TCGA 具有可比性，本文为每个基因计算其拷贝数的对数除以样本中拷贝中性状态的倍性。

最后，对于所有四个数据集，将基因级拷贝数估计值二值化，将零分配给拷贝中性基因，将零分配给所有重叠缺失或扩增的基因。

体细胞拷贝数：缺失或扩增基因赋值为1 ，其余赋值为0 。

数据预处理

Somatic point mutations

体细胞点突变

体细胞点突变：体细胞点突变的基因赋值为1，其余赋值为0。

训练数据： GDSC细胞系MOLI性能实验中使用的药物种类：多西他赛、顺铂、吉西他滨、紫杉醇、厄洛替尼和西妥昔单抗

迁移学习性能实验中使用的药物种类（靶向EGFR通路的药物）： 西妥昔单抗、厄洛替尼、阿法替尼、吉非替尼和拉帕替尼。使用这些药物的多组学数据创建一个大的训练集（ > 3000个样本）。

验证数据：体外(PDX)和体内(TCGA患者)，作者使用5种化疗药物和2种靶向治疗药物对MOLI进行了验证。

MOLI

MOLI 假设为每种组学数据类型提供相同基因的值。MOLI 的网络由以下子网络组成。

1、它有多个前馈编码子网络，每个输入组学数据类型一个。

2、每个编码子网络接收其相应的组学数据并将其编码到学习的特征空间中。从编码子网络中学习到的特征通过连接集成到一个表示中。

3、具有 Sigmoid 激活函数的分类层用作预测药物反应的分类。

整个网络使用结合了分类损失和三元组损失的成本函数以端到端的方式进行训练。

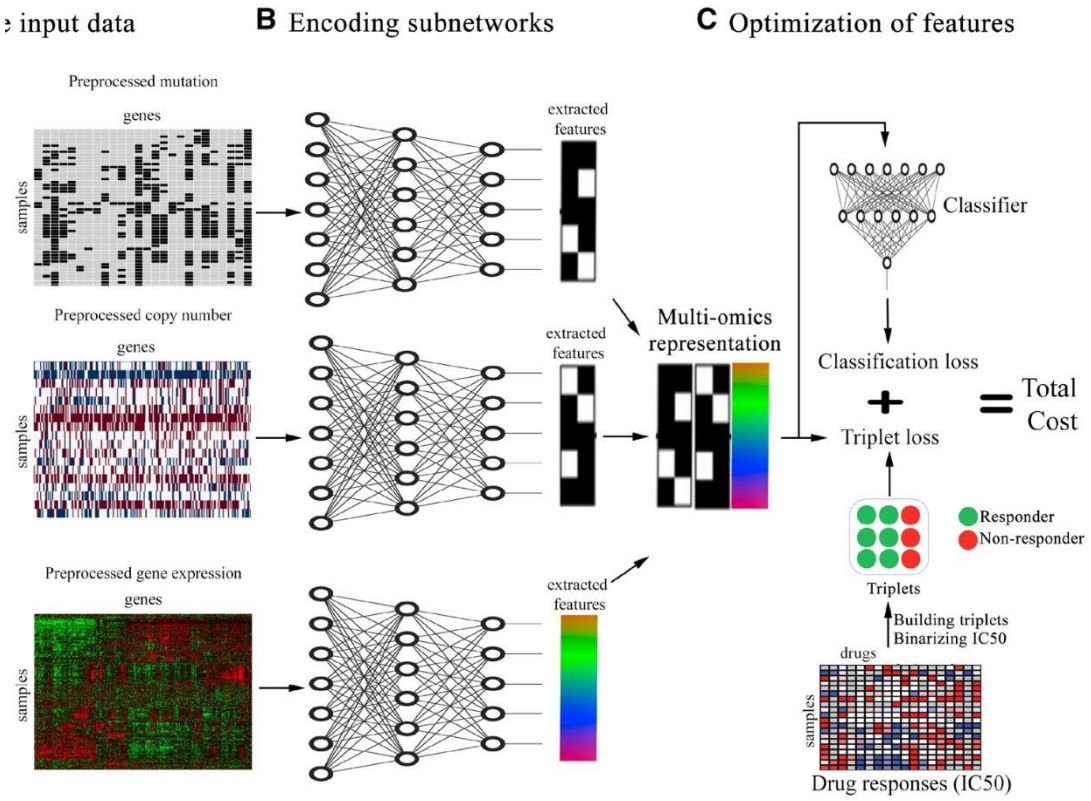
Learning features by encoding sub-networks

通过编码子网络学习特征

为了学习输入中每种组学数据类型的特征，设计了单独的编码前馈子网络来将输入空间映射到特征空间。

1、使用X_M，X_E和X_C表示突变，拷贝数畸变和基因表达数据，维度为N×D，其中N是样本的数量，D是基因的数量。

2、每个编码子网络的基本结构为**全连接层**，激活函数为**Relu**，均包含dropout和批标准化。将这些子网络分别表示为f_M (X_M)、f_C (X_C)和f_E (X_E)。



Integrating learned features by late integration

通过后期集成来集成学习到的特征

利用后期集成方法并将不同单组学数据类型的学习特征连接起来，以获得一个多组学表示。例如，如果三个编码子网络的输出是三个M×N的特征矩阵，那么在连接之后，输出将是一个M×3N表示矩阵;

经过L2标准化层进一步平滑了集成表示;

多组学数据作为输入并返回集成表示，如下所示:

$$F(X_{\mathrm{M}}, X_{\mathrm{C}}, X_{\mathrm{E}}) = f_{\mathrm{M}}(X_{\mathrm{M}}) \oplus f_{\mathrm{C}}(X_{\mathrm{C}}) \oplus f_{\mathrm{E}}(X_{\mathrm{E}}),$$

⊕表示连接操作

Optimizing the learned features by the combined cost function

通过组合成本函数优化特征

MOLI的最后一个子网络的激活函数为Sigmoid，使用dropout和L2正则化。将这个分类器表示为g(.)

组合成本函数的第一部分是传统意义上的二元交叉熵损失函数.

$$\mathbb{L}_{\text{Classifier}} = -[Y \log g(F(X_{\text{E}}, X_{\text{M}}, X_{\text{C}})) + (1-Y)\log(1 - g(F(X_{\text{E}}, X_{\text{M}}, X_{\text{C}}))],$$

在成本函数中添加了三元组损失，以施加进一步的约束，进一步保证分类的准确性。这种约束迫使响应者彼此之间比与非响应者更相似。

使应答细胞系的表征与无应答细胞系的表征更相似但与非应答细胞系的表征不同

选择三元组（triplets）的方法:

离线选择和在线选择。在训练模型之前，离线选择会根据标签的值（在本例中为药物反应）构建三元组。在线选择在训练期间从每个小批量（*mini-batch*）的样本中选择三元组。

作者这里采用了在线方式。在线选择也包含两种：**软选择**是在输入样本/mini-batch所有可能的组合构建三元组。**硬选择**是只使用三元损失值高的三元组。

软选择为模型提供了更多的训练样本，但网络可能过于依赖简单的样本，在困难样例上表现不佳。硬选择训练样本少，在小的不平衡数据集上表现不佳。**作者采用了软选择的方法。**

✓ 三元损失函数（Triplet Loss）

T的取值有三种：锚、阳性、阴性。其中前两个是多组学数据药物响应者，最后一个是药物非响应者，需要满足以下条件：

$$d(F(Anchor_i), F(Positive_i)) \leq d(F(Anchor_i), F(Negative_i)),$$

其中d(.)为欧几里得距离。移项得：

$$d(F(Anchor_i), F(Positive_i) - d(F(Anchor_i), F(Negative_i)) \leq 0$$

为了避免零解，引入$\xi > 0$:

$$d(F(Anchor_i), F(Positive_i)) - d(F(Anchor_i), F(Negative_i)) + \xi \leq 0$$

我们希望锚和阴性的距离大于锚和阳性的距离。因此，第i个三元损失函数的值为：

$$\mathscr{L}^i_{Triplet} = max[d(F(Anchor_i), F(Positive_i) - d(F(Anchor_i), F(Negative_i)) + \xi, 0]$$

三元损失：

$$\mathscr{L}_{Triplet} = \sum_{i=1}^{T} \mathscr{L}^i_{Triplet}$$

总成本

$$J = \mathbb{L}_{\text{Classifier}} + \gamma \mathbb{L}_{\text{Triplet}},$$

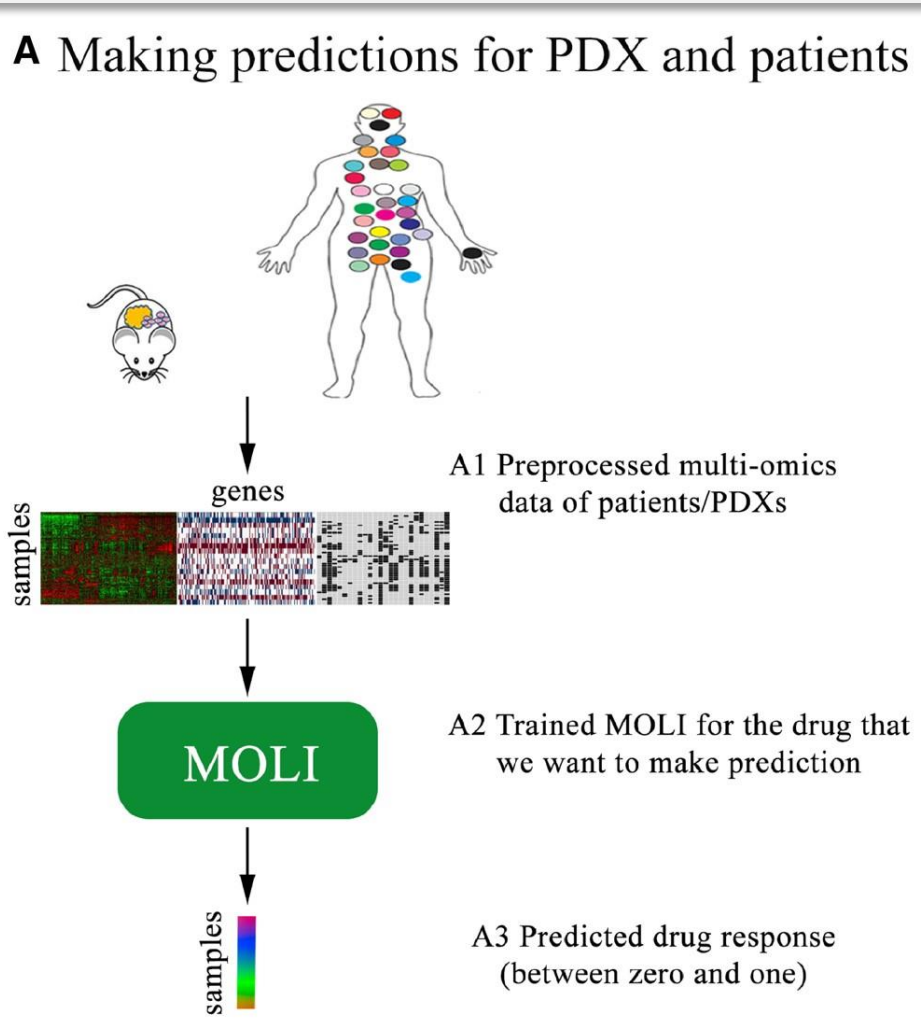
Transfer learning for targeted drugs

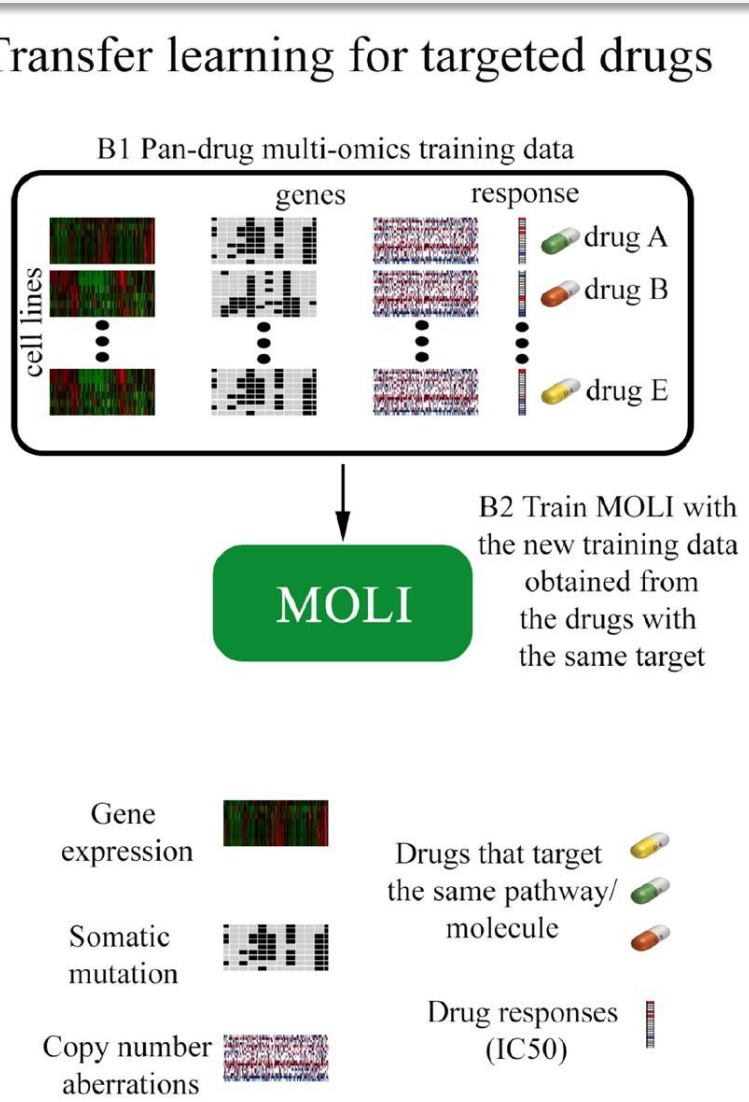
靶向药物的迁移学习

对于靶向药物，作者使用迁移学习，用泛药物训练MOLI。这种泛药物输入包含针对同一通路或分子的靶向药物族的多组学特征和药物反应。

一个MOLI模型是针对一个药物族进行训练的，预计此类药物会在细胞系中产生高度相关的反应，这种方法增加了训练数据集的大小。

作者评估了EGFR通路抑制剂的迁移学习，该方法适用于任何靶向药物族。





结合靶向同一通路或分子的靶向药物，制作MOLI的泛药物训练数据集

实验主要针对三个问题进行设计：

1.在PDX和患者数据的预测中，MOLI是否优于单组学和早期整合模型?

2.迁移学习是否对靶向药物有效，即接受泛药物数据训练的MOLI是否优于接受单药数据训练的MOLI？

3.对于靶向药物，MOLI预测的反应是否与该药物的靶向有关?

实验结果

MOLI和对照模型的AUC值

Method	PDX	PDX	PDX	PDX	TCGA	TCGA	TCGA	Input omics
Drug	Paclitaxel	Gemcitabine	Cetuximab	Erlotinib	Docetaxel	Cisplatin	Gemcitabine	
Geeleher *et al.* (2014)	0.52	0.59	*0.58*	*0.67*	0.59	0.62	0.53	Expression
Early integration via NMF	0.24	0.56	0.53	0.28	0.39	0.40	0.58	Multi
Early integration via DNNs	NSC	**0.66**	NSC	NSC	0.52	NSC	0.59	Multi
Feed forward net	0.68	0.48	0.43	0.37	**0.69**	0.44	*0.65*	Expression
MOLI complete	*0.69*	0.52	0.51	0.39	*0.63*	**0.75**	0.64	Expression
MOLI with classifier	NSC	0.55	0.46	NSC	0.58	0.6	**0.69**	Multi
MOLI complete	**0.74**	*0.64*	0.53	0.63	0.58	*0.66*	*0.65*	Multi
MOLI complete Pan-drug	NA	NA	**0.80**	**0.72**	NA	NA	NA	Multi

NA：非靶向药物
NSC：损失曲线或AUC曲线是波动的
Complete：MOLI包含分类损失和三元损失

三元组损失对提高预测性能的贡献

靶向药物的迁移学习显着提升性能

由表格可以得出的结论：

1. MOLI 在七个外部验证数据集中的四个中表现出更好的性能；

2.对于Erlotinib和Cetuximab，MOLI在接受泛药物输入时表现更好。

3.对于Paclitaxel和 Erlotinib，大多数对照模型要么表现不佳，要么出现NSC。可能的原因是样本数量少 。

4.在四种药物的早期整合模型中观察到了NSC，可能的原因是开始时的级联增加了维度，使自编码器和分类器在特征学习上更加困难。

Conclusion

提出了四个主要发现：

1. MOLI 在 AUC 和精确召回曲线下面积方面优于单组学（基因表达）预测性能。
2. MOLI 在 AUC 和精确召回曲线下面积方面优于使用早期集成的深度神经网络。
3. MOLI 及其组合成本函数优于单组学和多组学基线，只有分类损失。
4. MOLI 接受泛药物输入训练，采用迁移学习，优于针对 EGFR 的靶向治疗药物特定输入训练的 MOLI。

最后，本文分析了 MOLI 的生物学意义，并发现大量证据表明 MOLI 预测的反应与乳腺癌、肾癌、肺癌和前列腺癌的 TCGA 患者的 EGFR 通路中许多基因的表达水平具有统计学上的显着关联。

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open Access | Published: 19 January 2022

Global fine-scale changes in ambient NO_2 during COVID-19 lockdowns

Matthew J. Cooper ✉, Randall V. Martin, Melanie S. Hammer, Pieternel F. Levelt, Pepijn Veefkind, Lok N. Lamsal, Nickolay A. Krotkov, Jeffrey R. Brook & Chris A. McLinden

Nature **601**, 380–387 (2022) | Cite this article

7071 Accesses | **144** Altmetric | Metrics

新冠封城期间地面二氧化氮污染的变化

本文概述

NO2是空气污染的重要组成，人类暴露其中与多种不良健康结局有关，包括呼吸道感染、哮喘和肺癌。曾有报告称，为减少新冠肺炎（COVID-19）传播而采取的封城措施致大气和地面的NO2降低。

本文作者用高分辨率卫星图像计算了全球地面NO2浓度，以单个城市在2020年新冠疫情封城期间的估计与2019年作比较。然后他们使用卫星测量结果量化了超过200个城市的NO2水平变化，包括65个没有地面监测的城市，这些城市多数位于低收入地区。

作者发现，在封城条件严格的国家如北美和中国，平均国家级人口加权NO2浓度比其他地方多下降了29%。作者还发现，封城期间NO2的下降超出了最近排放控制带来的年平均同比下降，相当于全球15年的减排量。

因为新冠疫情封城致地面二氧化氮（NO2）浓度的下降，**依区域和排放部门差别很大，在封城水平更严格的国家，平均国家级人口加权NO2水平多降低了1/3。**这些发现改进了我们对NO2暴露评估的理解，提供了改进空气质量健康评估的机会。

nature cardiovascular research

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature cardiovascular research › letters › article

Letter | Published: 14 February 2022

Low depression frequency is associated with decreased risk of cardiometabolic disease

Michael C. Honigberg, Yixuan Ye, Lillian Dattilo, Amy A. Sarma, Nandita S. Scott, Jordan W. Smoller, Hongyu Zhao, Malissa J. Wood & Pradeep Natarajan ✉

Nature Cardiovascular Research **1**, 125–131 (2022) | Cite this article

857 Accesses | **1** Citations | **875** Altmetric | Metrics

低抑郁频率与心血管代谢疾病风险降低有关

本文概述

40多年来，人们一直知道心脏病患者中存在未被认识的抑郁症的普遍性。但尚不清楚抑郁是否促进心脏疾病发展，或其是否主要继发于临床疾病。

美国麻省总医院的Pradeep Natarajan和同事研究了英国生物信息库（UK Biobank）328152名欧洲祖先个体（年龄在40-69岁之间）的基因组。

作者利用这些数据生成了一个多基因风险评分——可用于改善心脏病风险预测的专门工具。

作者发现，**抑郁情绪负担较轻与冠心病、Ⅱ型糖尿病和心房颤动风险下降有关，下降程度分别为34%、33%和20%。**

这些关联独立于已知与精神健康状况不佳及心血管疾病风险均有关的生活方式因素，例如饮食、锻炼和吸烟。此外，**女性抑郁和冠心病之间的相关性高于男性。**

这项研究拓展了关于抑郁在促进心血管疾病中潜在作用的认识。但作者总结说，未来需要进一步研究以确定这种相关背后的机制，确定对预防性疗法的潜在影响。