



组会汇报

 汇报人: Lilian

2021/12/23



文献来源

nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 08 June 2021](#)

MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification

[Tongxin Wang](#), [Wei Shao](#), [Zhi Huang](#), [Haixu Tang](#), [Jie Zhang](#), [Zhengming Ding](#)  & [Kun Huang](#) 

MOGONET 使用图卷积网络集成多组学数据，实现对患者分类和生物标志物的识别



Abstract

- ✓ 为了充分利用组学技术的进步，更全面地了解人类疾病，需要新的计算方法来对多种类型的组学数据进行综合分析。
- ✓ 作者提出了一种新的**多组学集成方法**，称为**多组学图卷积网络（MOGONET）**，用于生物医学分类。
- ✓ MOGONET **联合探索特定组学学习（omics-specific learning）和跨组学相关学习（cross-omics correlation learning）**，以实现有效的多组学数据分类。
- ✓ 本文章证明 MOGONET 在使用 mRNA 表达数据、DNA 甲基化数据和 microRNA 表达数据的不同生物医学分类应用中优于其他最先进的监督多组学综合分析方法。此外，MOGONET 可以从与所研究的生物医学问题相关的不同组学数据类型中识别重要的生物标志物。

Introduction

为什么要用多组学数据？

- ✓ 组学技术的快速发展使得个体化医学能够利用具有前所未有细节的分子水平数据。例如mRNA 表达量、DNA 甲基化和 microRNA 表达量，可基于同一组样本中获得多组学（multi-omics）数据。
- ✓ 但每种组学技术只能捕捉生物复杂性的一部分。只有整合多种类型的组学数据，才可以微生物过程提供更全面的视角。例如，在疾病研究中集成多组学数据，可提高患者临床结果预测的准确性。

Introduction

先前研究的不足

- ✓ 以前，由于收集和注释数据的费用高昂，以及缺乏关于疾病亚型的知识，**带标记的生物医学数据很少**。因此，大多数现有的多组学整合方法侧重于**无监督的方法**，在没有额外的表型信息时，试图从已确定的样本群中提取生物学见解。
 - ✓ 随着带有详细注释的组学数据增加，出现**有监督分类方法**，主要有两类：
 1. **基于特征连接的方法**通过直接将多组学的输入数据特征合并，来训练分类模型，从而集成了不同的组学数据；
 2. **基于集成的方法**综合了不同分类器的预测结果，每个分类器都对同组学数据单独训练。
- 然而，这些方法没有考虑不同组学数据之间的相关性，可能导致预测结果偏向于某些组学数据的影响。
- ✓ 随着深度学习在各种任务中的不断进步，越来越多的多组学集成方法开始利用深度神经网络(NN)的高学习能力和灵活性。但现有方法基于全连接网络，**没有通过相似性网络有效地利用样本之间的相关性**。



Introduction

本文内容概述

- 最近的一些计算方法集中在通过同时整合不同类型的基因组数据来预测癌症基因或识别癌症基因模块，但缺乏有效地结合基因特征网络和矩阵的方法。
以往的方法要么是将单一维度的分数与PPI网络集成在一起，而不能处理多维节点特征向量，要么是仅使用多维向量编码特征，而不包括基因-基因网络；
- 很少有方法同时结合多维节点向量和基因-基因相互作用的图表示；然而，这些方法缺乏可解释性。**可解释性对于评估与癌症相关的基因的分子起源、检测潜在的人为因素以及提高建模方法的可信度非常重要**

Introduction

- ✓ 作者提出的MOGONET，是一种用于生物医学应用中分类任务的多组学数据分析框架。
- ✓ MOGONET 在标签空间将特定组学学习与多组学综合分类相结合。
- ✓ 具体来说，MOGONET 利用**图卷积网络 (GCN)** 进行特定于组学的学习。与全连接 NN 相比，GCN 利用组学特征和相似性网络描述的样本之间的相关性来获得更好的分类性能。
- ✓ 展示了集成多种组学数据类型的必要性，以及通过综合消融研究**将 GCN 和 VCDN(View Correlation Discovery Network) 结合**用于多组学数据分类的重要性。
- ✓ 我们证明了 MOGONET 可以识别与所研究的生物医学问题相关的重要组学特征和生物标志物。

Datasets

为了证明 MOGONET 的有效性，作者使用四种不同的数据集将所提出的方法应用于四种不同的生物医学分类任务：

1. 用于阿尔茨海默病 (**AD**) 患者与正常对照 (NC) 分类的 ROSMAP,
2. 用于低级别胶质瘤分级的 **LGG**。
3. 用于肾癌类型分类的 **KIPAN** ;
4. 用于乳腺癌浸润性癌 (**BRCA**) PAM50 亚型分类的 BRCA。

使用三种类型的组学数据（即mRNA表达数据（mRNA）、DNA甲基化数据（meth）和miRNA表达数据（miRNA））进行分类，以提供有关疾病的全面和补充信息。我们的研究仅包括具有匹配 mRNA 表达、DNA 甲基化和 miRNA 表达数据的样本。

Dataset	Categories	Number of features mRNA, meth, miRNA	Number of features for training mRNA, meth, miRNA
ROSMAP	NC: 169, AD: 182	55,889, 23,788, 309	200, 200, 200
LGG	Grade 2: 246, Grade 3: 264	20,531, 20,114, 548	2000, 2000, 548
KIPAN	KICH: 66, KIRC: 318, KIRP: 274	20,531, 20,111, 445	2000, 2000, 445
BRCA	Normal-like: 115, Basal-like: 131, HER2-enriched: 46, Luminal A: 436, Luminal B: 147	20,531, 20,106, 503	1000, 1000, 503

- ✓ LGG、KIPAN 和 BRCA 的组学数据以及 LGG 患者的分级信息是通过 Broad GDAC Firehose 从癌症基因组图谱计划 (TCGA) 中获取的。
- ✓ ROSMAP 数据集中的不同组学数据来自 AMP-AD

Preprocessing

1、对于 DNA 甲基化数据，仅保留与 Illumina Infinium HumanMethylation27 BeadChip 中探针相对应的探针

2、进一步过滤掉没有信号（零均值）或低方差的特征。

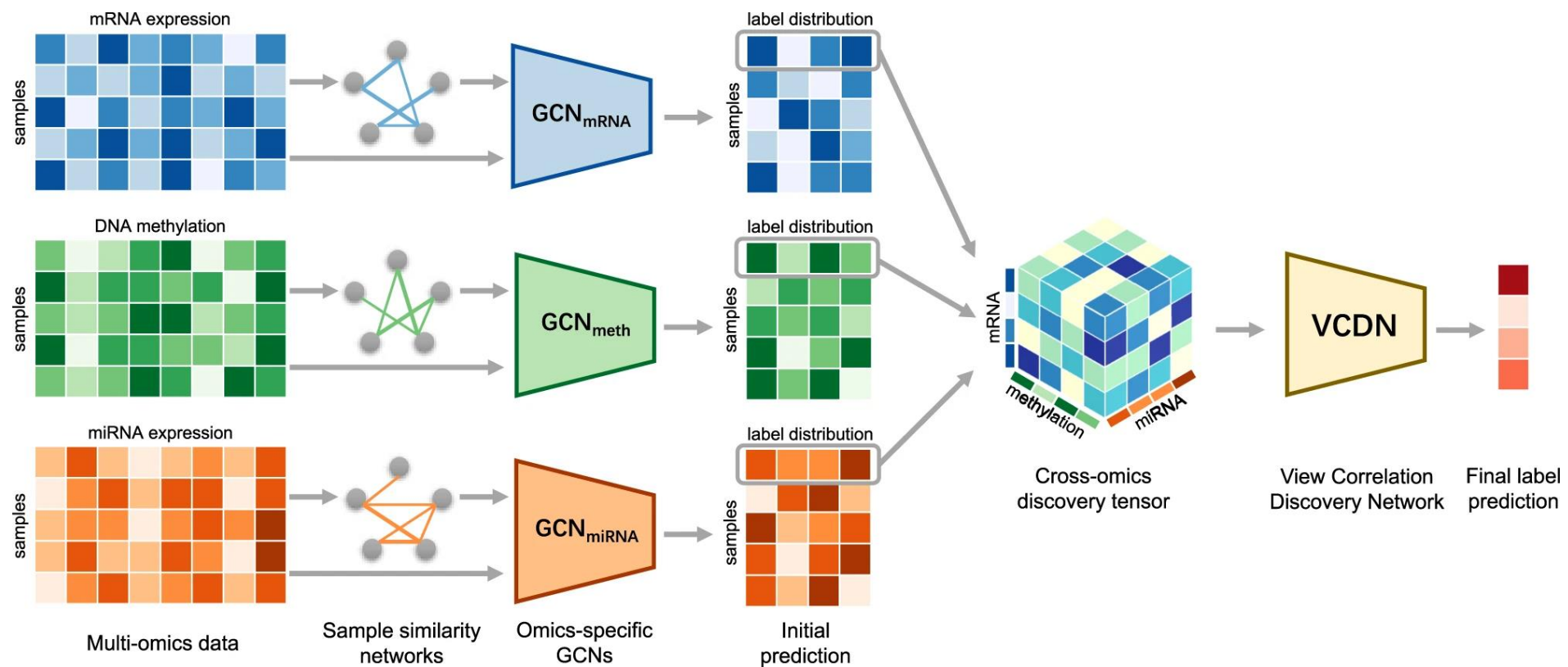
对不同类型的组学数据应用了不同的方差过滤阈值（mRNA 表达数据为 0.1，DNA 甲基化数据为 0.001）；

对于 miRNA 表达数据，只过滤掉没有变化的特征（方差为零），因为可用的特征由于 miRNA 的数量很少而受到限制。

3、最后，我们通过线性变换将每种类型的组学数据单独缩放到 $[0, 1]$ 以训练 MOGONET。

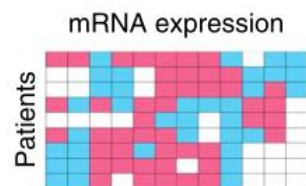


MOGONET网络框架

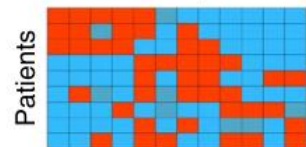


SNF(similarity network fusion)相似网络融合算法

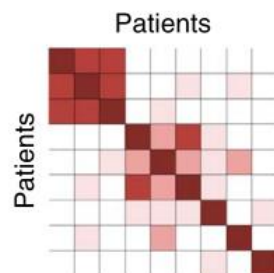
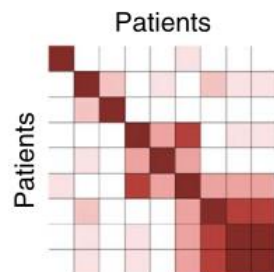
a Original data



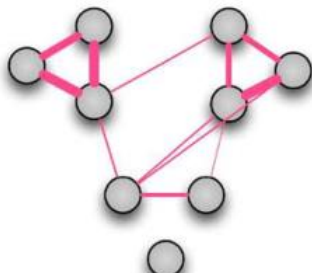
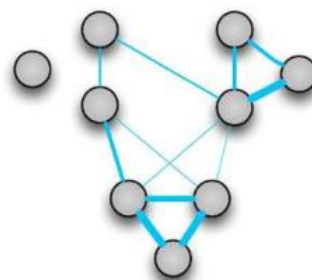
DNA methylation



b Patient similarity matrices



c Patient similarity networks



$$W(i, j) = \exp \left(-\frac{\rho^2(x_i, x_j)}{\mu \epsilon_{i, j}} \right)$$

$\rho(i, j)$: 患者 x_i 与患者 x_j 的欧几里得距离

$W(i, j)$: 患者 x_i 与患者 x_j 的相似度矩阵 ($n \times n$)

$$S(i, j) = \begin{cases} \frac{W(i, j)}{\sum_{k \in N_i} W(i, k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

$S(i, j)$ 是第 i 个患者与其 k 个最邻近患者的相似度

GCNs for omic-specific learning

By viewing each sample as a node in the sample similarity network, the goal of each GCN in MOGONET is to learn a function of features on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ $\mathbf{G}\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{A} \in \mathbb{R}^{n \times n}$

$$\begin{aligned}\mathbf{H}^{(l+1)} &= f(\mathbf{H}^{(l)}, \mathbf{A}) \\ &= \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}),\end{aligned}\tag{1}$$

where $\mathbf{H}^{(l)}$ is the input of the l th layer and $\mathbf{W}^{(l)}$ is the weight matrix of the l th layer. $\sigma(\cdot)$ denotes a non-linear activation function. For effective training of GCNs, Kipf and Welling^{[55](#)} further modified the adjacency matrix \mathbf{A} as:

$$\widetilde{\mathbf{A}} = \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} = \hat{\mathbf{D}}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \hat{\mathbf{D}}^{-\frac{1}{2}},\tag{2}$$



GCNs for omic-specific learning

In MOGONET, the original adjacency matrix \mathbf{A} is constructed by calculating the cosine similarity between pairs of nodes and edges with cosine similarity larger than a threshold ϵ are retained. Specifically, A_{ij} , which is the adjacency between node i and node j in the graph, is calculated as:

$$A_{ij} = \begin{cases} s(\mathbf{x}_i, \mathbf{x}_j), & \text{if } i \neq j \text{ and } s(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where \mathbf{x}_i and \mathbf{x}_j are the feature vectors of node i and node j , respectively.

$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$ is the cosine similarity between node i and j . The threshold ϵ is determined given a parameter k , which represents the average number of edges per node that are retained including self-connections:

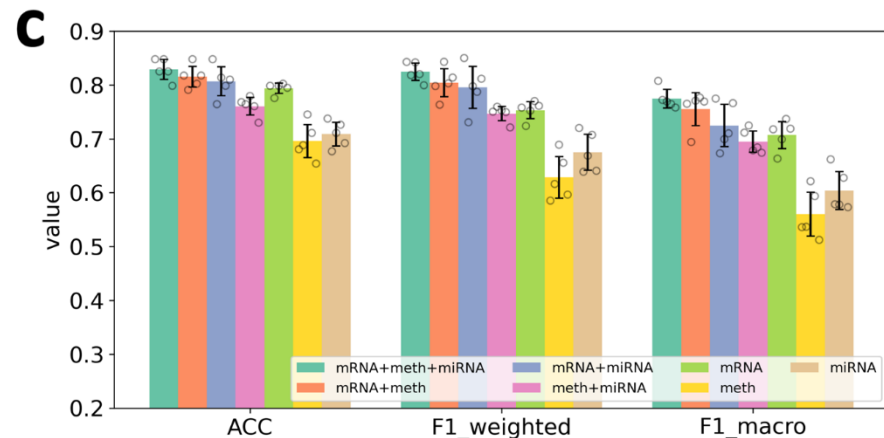
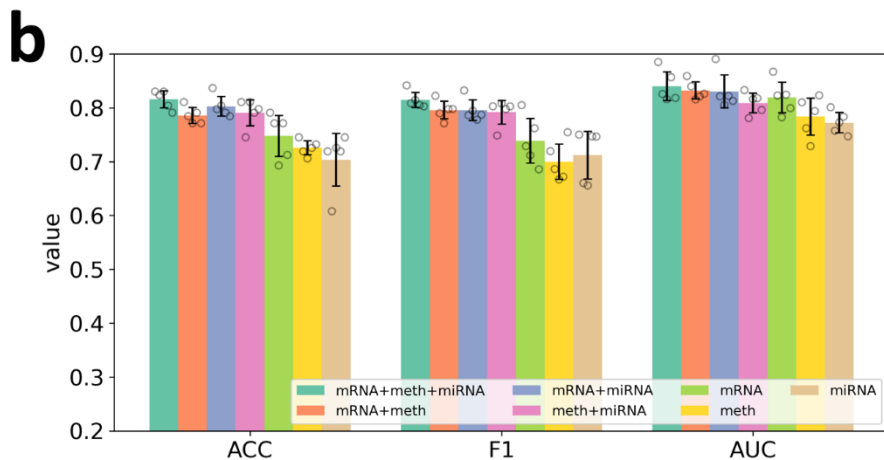
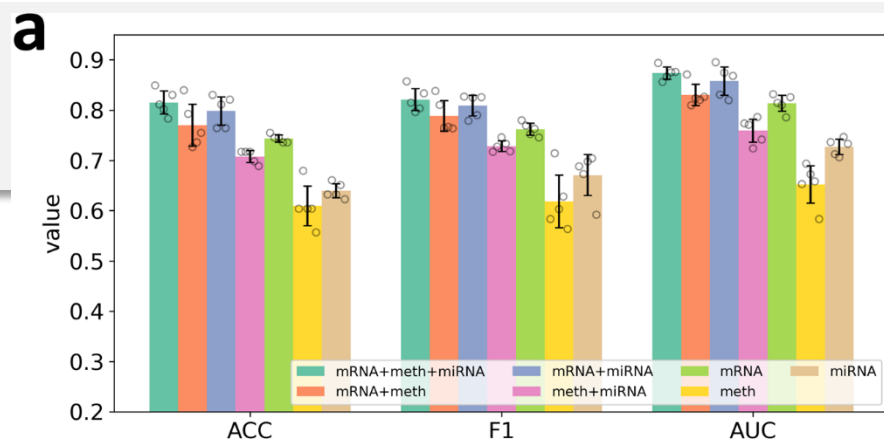
$$k = \sum_{i,j} I(s(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon) / n, \quad (4)$$

k 表示保留在相似性网络中的每个样本的平均边数。

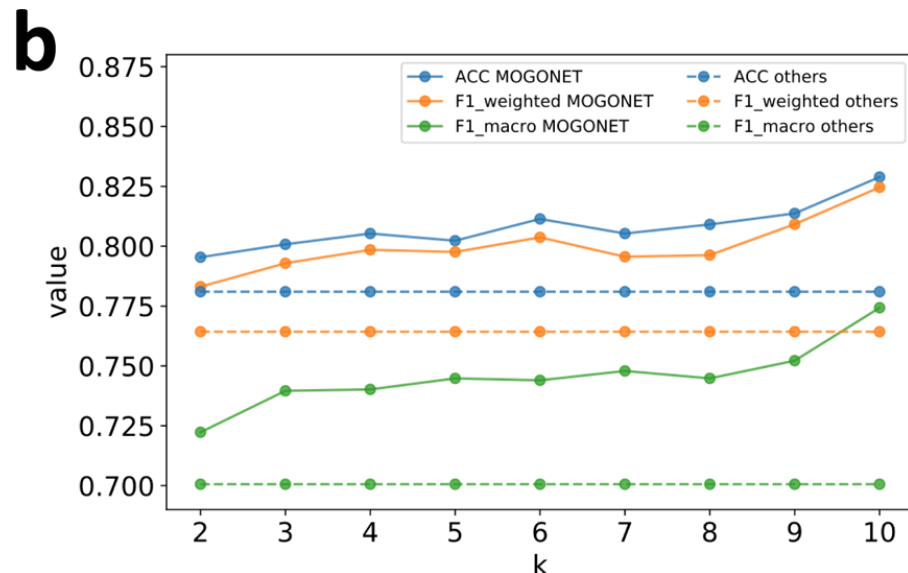
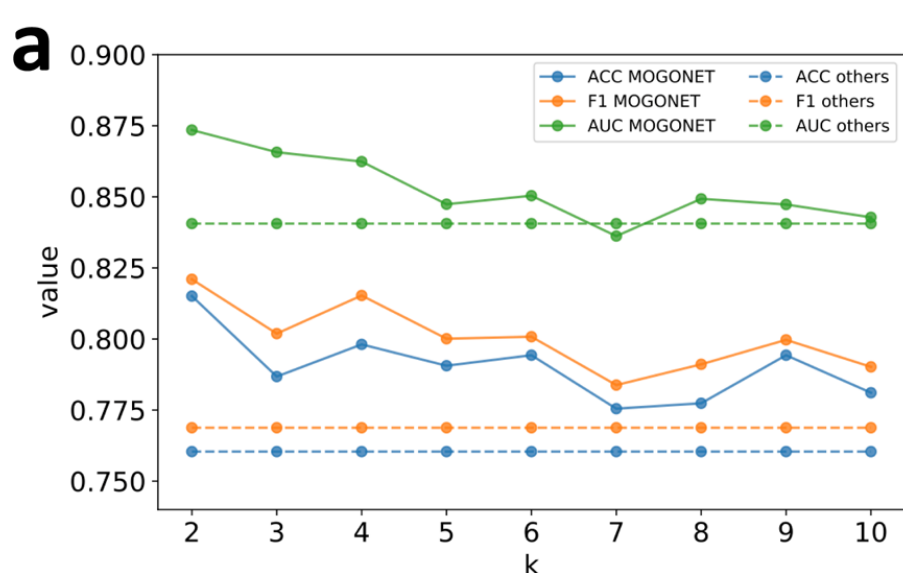


Performance of MOGONET under different omics data types

- ✓ 右图：使用MOGONET，基于单个或多个组学，在三种不同数据集ROSMAP（阿兹海默）LGG（神经胶质瘤）和 BRAC（乳腺癌）上的分类表现。
- ✓ MOGONET 可以扩展以适应不同数量的组学数据类型
- ✓ 不论使用那种评价指标，代表三种数据集的绿柱子都是最高的。
- ✓ 证明了 GCN 在组学数据分类问题中的有效性以及使用 VCDN 进行组学数据跨组学学习的有效性



Performance of MOGONET under different hyper-parameter k



1. k表示保留在相似性网络中的每个样本的平均边数
2. 观察到超参数k确实影响 MOGONET 的分类性能，因为性能随着k的变化而波动。
3. 然而，MOGONET 仍然对k的变化具有鲁棒性，因为它在不同的k值下始终优于现有方法。

Important biomarkers identified by MOGONET

1. 识别生物标志物对于解释结果和理解生物学应用中的潜在生物学至关重要。
2. 由于 MOGONET 的输入在预处理期间被缩放到 $[0, 1]$ ，可以通过将其设置为零来从特征中移除信号。因此，**特征对分类任务的重要性可以通过特征去除后的性能下降来衡量。**
3. 通过将特征分配为零并计算与使用所有特征相比在测试集上的分类性能下降情况来分析每个特征在不同类型组学数据中的贡献，**性能下降最大的功能被认为是最重要的功能。**

Omics data type	Biomarkers
mRNA expression (8)	<i>NPNT, CDK18, KIF5A, SPACA6, TCEA3, SYTL1, ARRDC2, APLN</i>
DNA methylation (5)	<i>TMC4, AGA, HYAL2, CCL3, TTC15</i>
miRNA expression (17)	<i>hsa-miR-423-3p, hsa-miR-33a, hsa-miR-640, hsa-miR-362-3p, hsa-miR-491-5p, hsa-miR-206, hsa-miR-548b-3p, hsa-miR-127-3p, hsa-miR-106a, hsa-miR-17, hsa-miR-424, hsa-miR-577, hsa-miR-873, hsa-miR-651, hsa-miR-199b-5p, hsa-miR-192, hsa-miR-199a-5p, hsa-miR-H1</i>

- ✓ 于 AD 患者分类，MOGONET 将 8 个 mRNA 特征、5 个 DNA 甲基化特征和 17 个 miRNA 特征确定为前 30 个重要生物标志物
- ✓ 早期和晚期 AD 患者的海马和内侧额回中 hsa-miR-423 的表达水平显著改变，其中海马和内侧额回都是主要受 AD 病理影响的区域。

Important biomarkers identified by MOGONET

Omics data type	Biomarkers
mRNA expression (15)	<i>SOX11, AMY1A, SLC6A15, FABP7, SLC6A14, SLC6A2, FGFBP1, DSG1, UGT8, ANKRD45, PI3, SERPINB5, COL11A2, ARHGEF4, SOX10</i>
DNA methylation (9)	<i>GPR37L1, MIR563, OR1J4, ATP10B, KRTAP3-3, FLJ41941, TMEM207, CDH26, MT1DP</i>
miRNA expression (6)	<i>hsa-mir-205, hsa-mir-187, hsa-mir-452, hsa-mir-20b, hsa-mir-224, hsa-mir-204</i>

- ✓ 对于 BRCA PAM50 亚型分类, MOGONET 将 15 个 mRNA 特征、9 个 DNA 甲基化特征和 6 个 miRNA 特征确定为前 30 个重要生物标志物
- ✓ MOGONET 鉴定的高排名基因和 miRNA 也已被证明与乳腺癌有关, 例如, SOX11对调节定义基底样亚型的许多基因的表达至关重要, 而且SOX11与基底样乳腺肿瘤的侵袭和迁移有关。



文献来源

nature methods

[Explore content](#) ▼

[About the journal](#) ▼

[Publish with us](#) ▼

[Subscribe](#)

[nature](#) > [nature methods](#) > [articles](#) > article

[Published: 26 January 2014](#)

Similarity network fusion for aggregating data types on a genomic scale

[Bo Wang](#), [Aziz M Mezlini](#), [Feyyaz Demir](#), [Marc Fiume](#), [Zhuowen Tu](#), [Michael Brudno](#), [Benjamin Haibe-Kains](#)

& [Anna Goldenberg](#) 

用于在基因组规模上聚合数据类型的相似网络融合

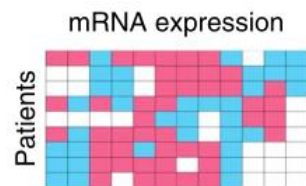


内容概述

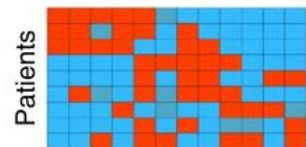
- ✓ 最近的技术使收集各种类型的全基因组数据具有成本效益。需要计算方法来组合这些数据，以创建给定疾病或生物过程的综合视图。
- ✓ 相似性网络融合 (SNF) 通过为每种可用数据类型构建样本（例如患者）网络，然后将这些网络有效地融合到一个代表所有基础数据的网络中，从而解决了这一问题。
- ✓ 例如，为了在给定一组患者的情况下创建疾病的综合视图，SNF 计算并融合从每个数据类型分别获得的患者相似性网络，利用数据的互补性。
- ✓ 作者使用 SNF 结合了五个癌症数据集的 mRNA 表达、DNA 甲基化和 microRNA (miRNA) 表达数据。

SNF(similarity network fusion)相似网络融合算法

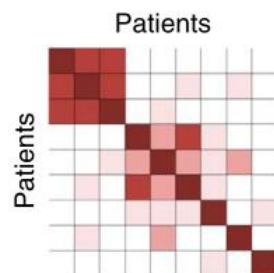
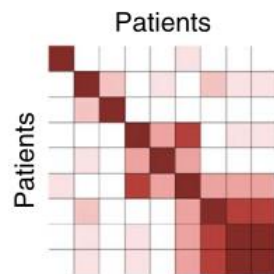
a Original data



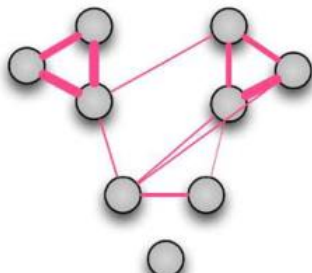
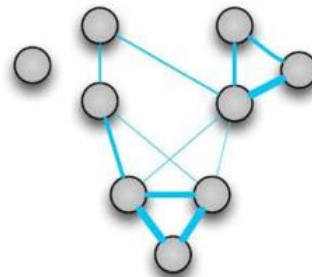
DNA methylation



b Patient similarity matrices



c Patient similarity networks



$$W(i, j) = \exp \left(-\frac{\rho^2(x_i, x_j)}{\mu \epsilon_{i, j}} \right)$$

$\rho(i, j)$:患者 x_i 与患者 x_j 的欧几里得距离

$W(i, j)$:患者 x_i 与患者 x_j 的相似度矩阵 ($n \times n$)

$$S(i, j) = \begin{cases} \frac{W(i, j)}{\sum_{k \in N_i} W(i, k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases}$$

$S(i, j)$ 是第 i 个患者与其 k 个最邻近患者的相似度



文献来源

nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 18 November 2021](#)

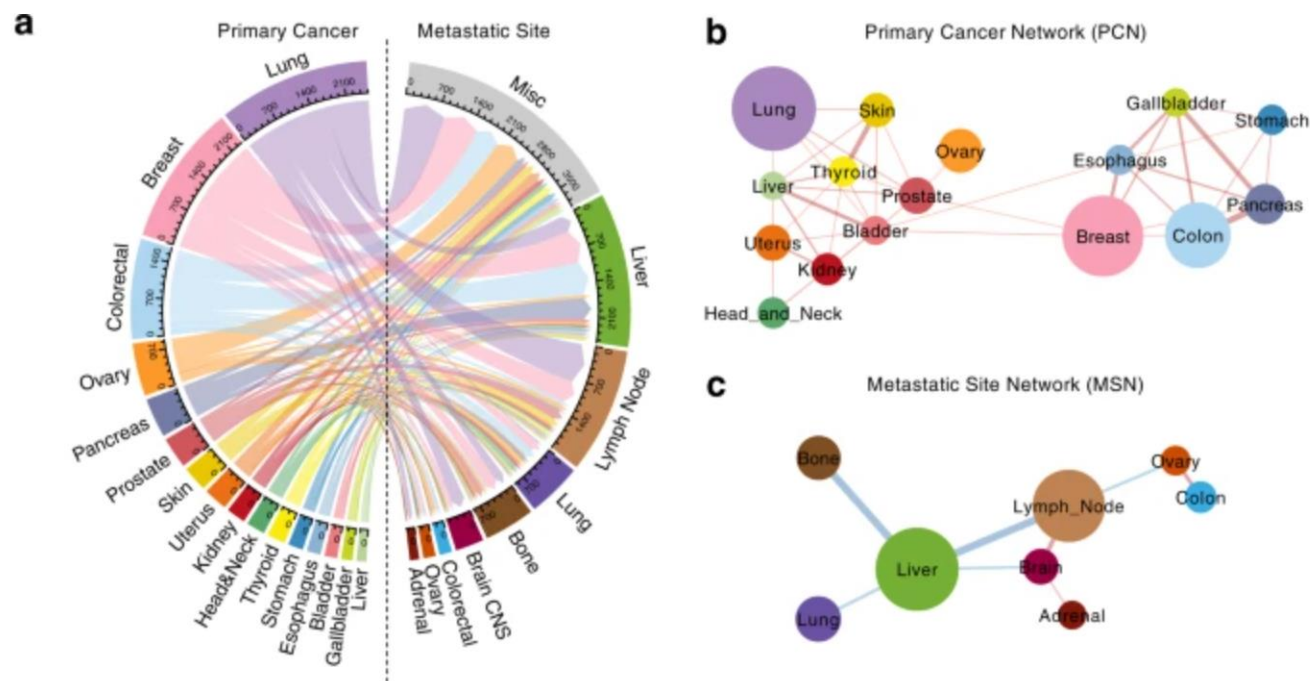
Machine learning of genomic features in organotrophic metastases stratifies progression risk of primary tumors

[Biaobin Jiang](#), [Quanhua Mu](#), [Fufang Qiu](#), [Xuefeng Li](#), [Weiqi Xu](#), [Jun Yu](#), [Weilun Fu](#), [Yong Cao](#) & [Jiguang Wang](#) 

向器官转移的基因组特征的机器学习对原发性肿瘤的进展风险进行分层



内容概述



- ✓ 开发一个转移网络模型（MetaNet），通过收集和分析总共 32,176 个泛癌 DNA 测序样本来评估转移风险和潜在的目标器官。
- ✓ 使用这个大数据队列，确定了与常见和器官转移相关的基因组生物标志物，并使用机器学习模型在早期阶段验证它们在转移风险评估中的效用，以筛选出具有更短无病生存率高于传统原发性患者。
- ✓ MetaNet 在区分乳腺癌和前列腺癌的转移灶和原发灶方面具有很高的准确性。



文献来源



Cell Reports

Volume 26, Issue 12, 19 March 2019, Pages 3461-3474.e5



Resource

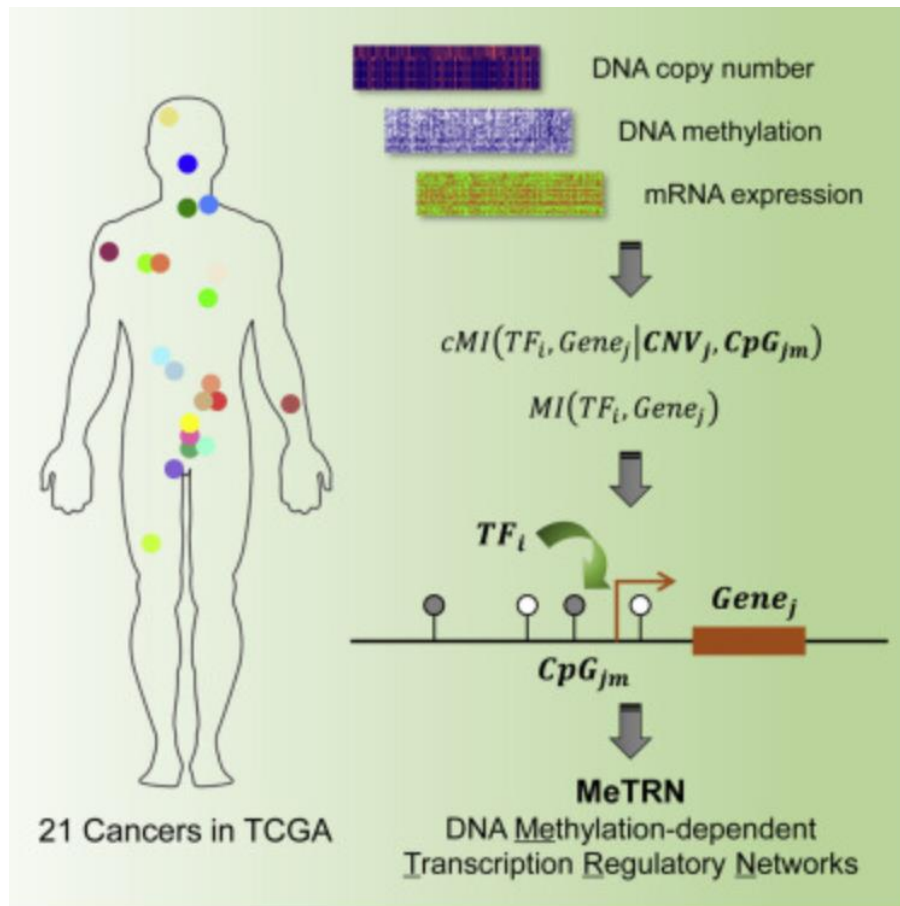
Dependency of the Cancer-Specific Transcriptional Regulation Circuitry on the Promoter DNA Methylome

Yu Liu^{1, 2, 3, 4, 6}, Yang Liu^{1, 3, 4, 5, 6}, Rongyao Huang^{1, 3, 4, 6}, Wanlu Song^{1, 2, 3, 4}, Jiawei Wang^{1, 3, 4}, Zhengtao Xiao^{1, 2, 3, 4}, Shengcheng Dong^{1, 3, 4}, Yang Yang^{1, 3, 4}, Xuerui Yang^{1, 3, 4, 7}  

癌症特异性转录调控网络对启动子DNA甲基化组的依赖



内容概述



- ✓ 文章使用公开数据库癌症基因组图谱（The Cancer Genome Atlas, TCGA）的肿瘤多组学数据，采用**基于信息论的数据挖掘策略**，首次以CpG位点解析度系统阐述了21种主要癌症中启动子区DNA甲基化组在肿瘤基因转录调控网络中的深度参与。
- ✓ 作者聚焦各类肿瘤中高度特异的启动子DNA甲基化组，使用TCGA项目**21种癌症**中总计7000余例肿瘤组织的**基因组、转录组、甲基化组**等数据，全面评估了各**DNA甲基化位点**对转录因子和靶基因之间的**调控关系**的影响。
- ✓ **文章中的分析显示，DNA甲基化参与的转录调控是基因表达过程的核心调控层级之一，许多重要的癌症相关基因强烈依赖于这种调控机制。**

文献来源





Cell Metabolism

Volume 33, Issue 12, 7 December 2021, Pages 2367-2379.e4



Article

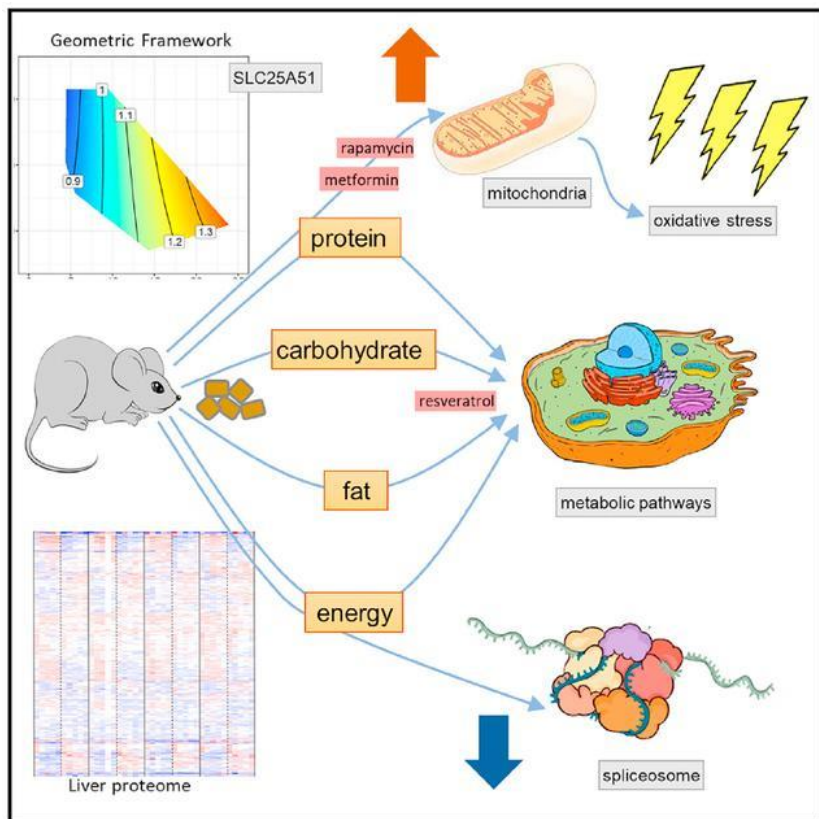
Nutritional reprogramming of mouse liver proteome is dampened by metformin, resveratrol, and rapamycin

David G. Le Couteur^{1, 2, 3}  , Samantha M. Solon-Biet¹, Benjamin L. Parker⁴, Tamara Pulpitel¹, Amanda E. Brandon^{1, 5}, Nicholas J. Hunt^{2, 3}, Jibril A. Wali¹, Rahul Gokarn¹, Alistair M. Senior¹, Gregory J. Cooney¹, David Raubenheimer^{1, 6}, Victoria C. Cogger^{2, 3}, David E. James^{1, 6}, Stephen J. Simpson^{1, 6, 7}  

二甲双胍、白藜芦醇和雷帕霉素抑制了小鼠肝脏蛋白质组的营养重编程
为了机体健康，是选择饮食还是药物？



内容概述



- ✓ 用小鼠模型系统考察了不同饮食方式分别与 Metformin, Rapamycin和Resveratrol三种小分子联用之后对于肝脏蛋白组的重塑作用, 并以此评估它们对于机体代谢健康的潜在功能。
- ✓ 研究小组设计了一项复杂的小鼠研究, 涉及40种不同的治疗方法, 每种治疗方法的蛋白质、脂肪和碳水化合物平衡、卡路里和药物含量各不相同。该研究旨在检查三种抗衰老药物和饮食对肝脏的影响, 因为肝脏是调节新陈代谢的关键器官。
- ✓ 作者发现, 饮食中的卡路里摄入量和常量营养素 (蛋白质、脂肪和碳水化合物) 的平衡对肝脏有很大的积极影响, 而有两种药物实际上抑制了均衡饮食对肝脏产生的积极影响。

我们吃的食物决定我们的健康状况, 不应该一味依靠药物

文献来源


nature communications

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature communications](#) > [articles](#) > article

Article | [Open Access](#) | [Published: 30 October 2020](#)

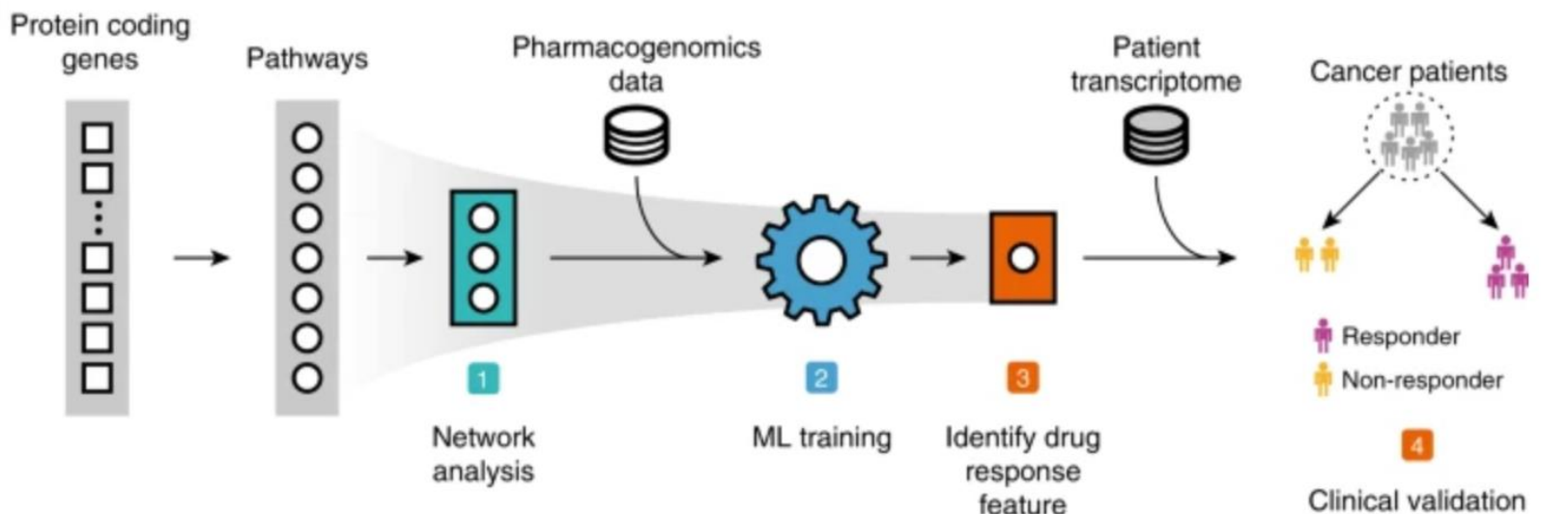
Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients

[JungHo Kong](#), [Heetak Lee](#), [Donghyo Kim](#), [Seong Kyu Han](#), [Doyeon Ha](#), [Kunyoo Shin](#)  & [Sanguk Kim](#) 

[Nature Communications](#) **11**, Article number: 5485 (2020) | [Cite this article](#)

11k Accesses | **23** Citations | **73** Altmetric | [Metrics](#)

结直肠和膀胱类器官模型中基于网络的机器学习可预测患者的抗癌药物疗效



- ✓ 通过基于网络的机器学习 (ML) 计算机识别药物反应生物标志物的总体框架。
- ✓ ML 的输入生物通路首先被过滤到蛋白质 - 蛋白质相互作用网络（绿色）中最接近药物靶标的那些。然后将近端通路用作训练 ML 模型（蓝色）的输入，揭示每个输入通路的预测性能。选择具有高预测性能（橙色）的通路作为生物标志物，用于将患者分为药物反应者和非反应者。
- ✓ **即使是患有相同癌症的患者，对抗癌药物的反应也不同。** 由于不必要的生物标记信息，机器学习存在基于错误信号进行学习的问题。
- ✓ 为了提高预测的准确性，研究小组引入了机器学习算法，该算法使用蛋白质相互作用网络，既可以与靶蛋白相互作用，也可以与药物靶点直接相关的单个蛋白质的转录组相互作用。它诱导学习功能上接近目标蛋白的蛋白质的转录组生成。因此，**只有选定的生物标记可以被学习，而不是传统机器学习必须学习的错误生物标记，从而提高了准确性。**