天津大學
Tianjin University

Report

Priceless lab

汇报人：Lilian

2022/6/16

01 Reference

AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks

Wan Xiang Shen [1,2], Yu Liu[3,4], Yan Chen[1], Xian Zeng[5], Ying Tan[1,6], Yu Yang Jiang[1,7,*] and Yu Zong Chen [1,7,*]

[1]The State Key Laboratory of Chemical Oncogenomics, Key Laboratory of Chemical Biology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, P.R. China, [2]Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore 117543, Singapore, [3]Institute for Health Innovation & Technology, National University of Singapore 117543, Singapore, [4]Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore 117543, Singapore, [5]Department of Biological Medicines & Shanghai Engineering Research Center of Immunotherapeutics, School of Pharmacy, Fudan University, Shanghai 201203, P.R. China, [6]Shenzhen Kivita Innovative Drug Discovery Institute, Shenzhen 518110, P.R. China and [7]Institute of Biomedical Health Technology and Engineering, Shenzhen Bay Laboratory, Shenzhen 518132, P.R. China

使用特征聚合的多通道网络增强和可解释的低样本组学深度学习

文献来源：Nucleic Acids Research/2022.5.6/新加坡国立大学、清华大学

Abstract & introduction

基于组学的生物医学学习经常依赖于高维（多达数千）和低样本量（数十到数百）的数据，这对高效的深度学习 (DL) 算法提出了挑战，尤其是对于低样本的组学研究。

在这里，**开发了一种无监督的新型特征聚合工具 AggMap 来聚合和映射基于它们的内在相关性**，将组学特征转化为多通道 2D 空间相关的类图像特征图 (Fmaps)。AggMap 在随机基准数据集上表现出强大的特征重建能力，优于现有方法。

以 AggMap 多通道 Fmap 作为输入，新开发的多通道 DL AggMapNet 模型在 18 个低样本组学基准任务上优于最先进的机器学习模型。AggMapNet 在学习噪声数据和疾病分类方面表现出更好的鲁棒性。

BioHULM 数据

生物医学研究经常依赖于来自组学（转录组学、蛋白质组学、代谢组学）分析的**高维、无序特征、低样本量和多平台** (highdimensional, unordered feature, low-sample size, and multiplatform， BioHULM) 数据。

尽管深度学习 (DL) 在学习复杂数据方面具有优势，但传统的机器学习 (ML) 方法是最近基于 BioHULM 的生物医学研究的主要工具 。

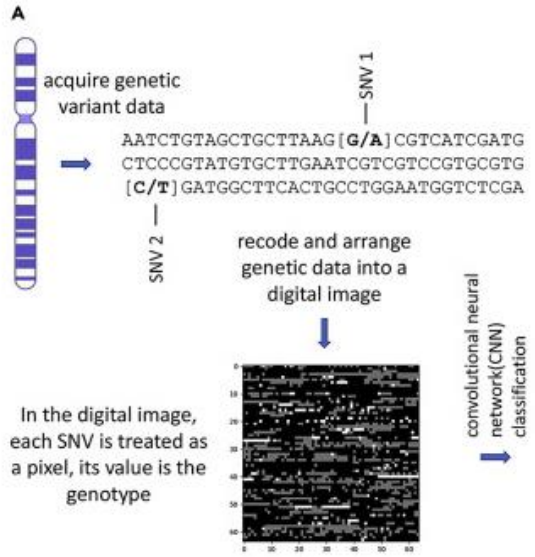
BioHULM 数据应用于 DL的两个问题：

- **过度拟合大量的超参数**
- **可解释性**

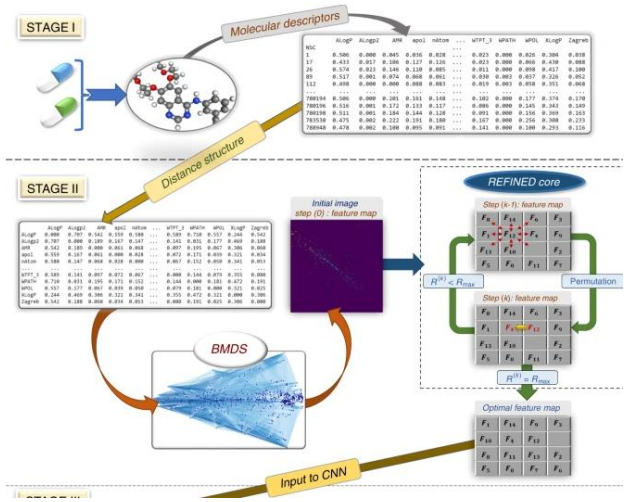
BioHULM 任务需要高性能和可解释的 DL 算法。

研究进展

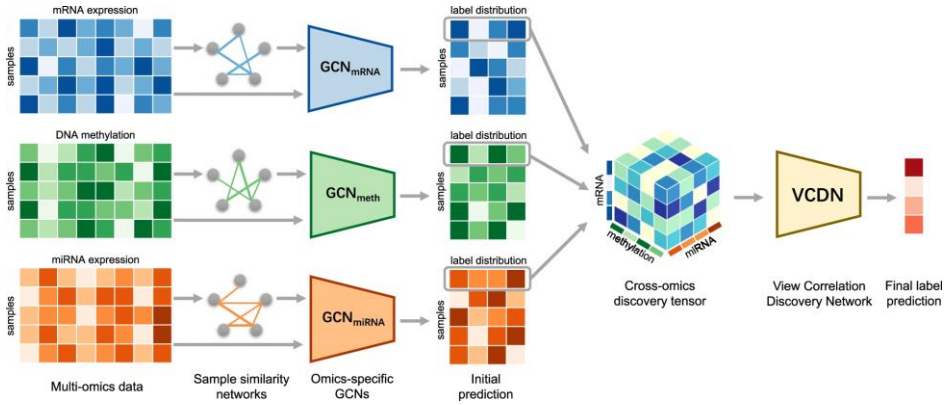
为了充分缓解 BioHULM 学习中的“维度灾难”问题，最近的研究集中在将 1D 无序数据转换为**基于遗传位置、数据邻域或功能关系**。



Artificial image objects for classification of schizophrenia with GWAS-selected SNVs and convolutional neural network
2021.8.13



REFIEND-CNN
2021 .9.1/nature

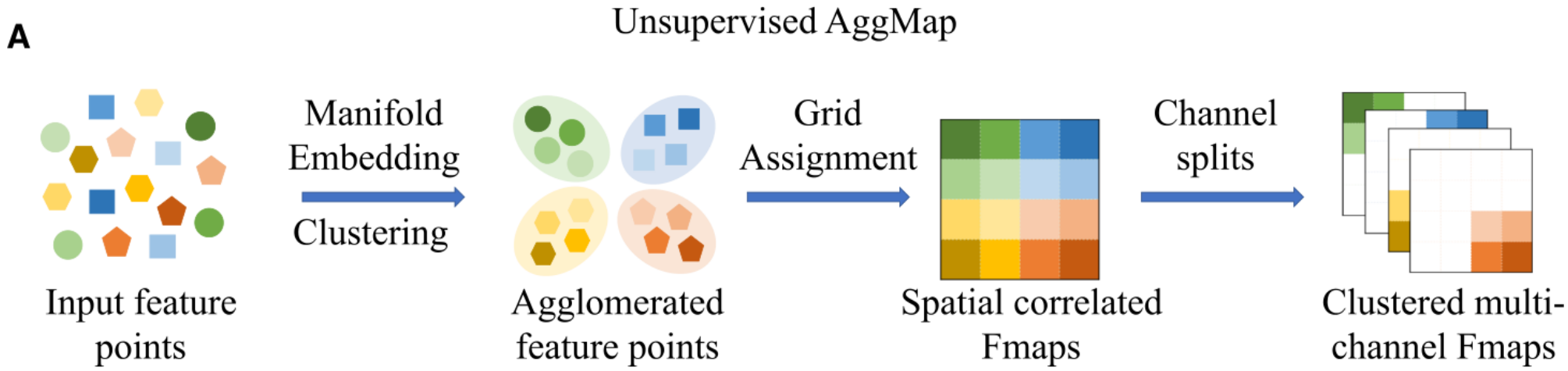


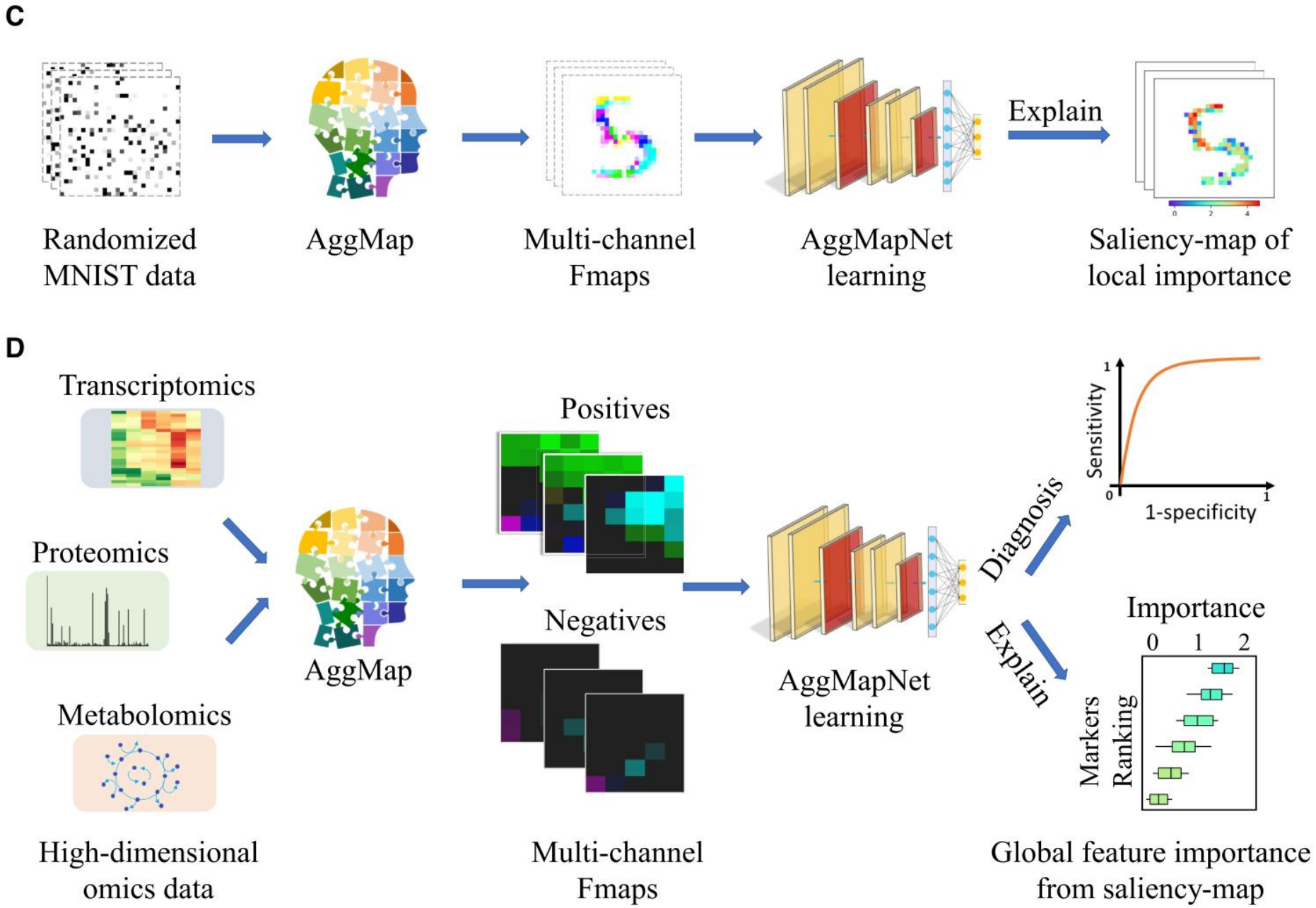
MOGONET 2021 .7.8/nature communications

这种转换可以使用卷积神经网络 (CNN) 进行有效的深度学习，但它们缺乏关于输入特征点的聚类组的丰富通道信息。

本文工作

在这项工作中，为了增强对低样本组学数据的基于 CNN 的有效学习，**开发了一种新的无监督特征聚合工具 AggMap**，用于将单个无序 BioHULM 特征点 (FP) 聚合和映射到空间相关的多通道 2D Fmap（图[1A])。AggMap 被定义为一个拼图求解器，因为它基于它们的内在相似性和拓扑结构来解决无序 FP 的拼图。我们还构建了一个新的多通道 CNN 架构 AggMapNet,用于从 AggMap Fmaps 中增强和可解释地学习 BioHULM。




C
Randomized MNIST data
AggMap
Multi-channel Fmaps
AggMapNet learning
Explain
Saliency-map of local importance
D
Transcriptomics
Proteomics
Metabolomics
High-dimensional omics data
AggMap
Positives
Negatives
Multi-channel Fmaps
AggMapNet learning
Diagnosis
Sensitivity
1-specificity
Explain
Importance
Markers Ranking
Global feature importance from saliency-map

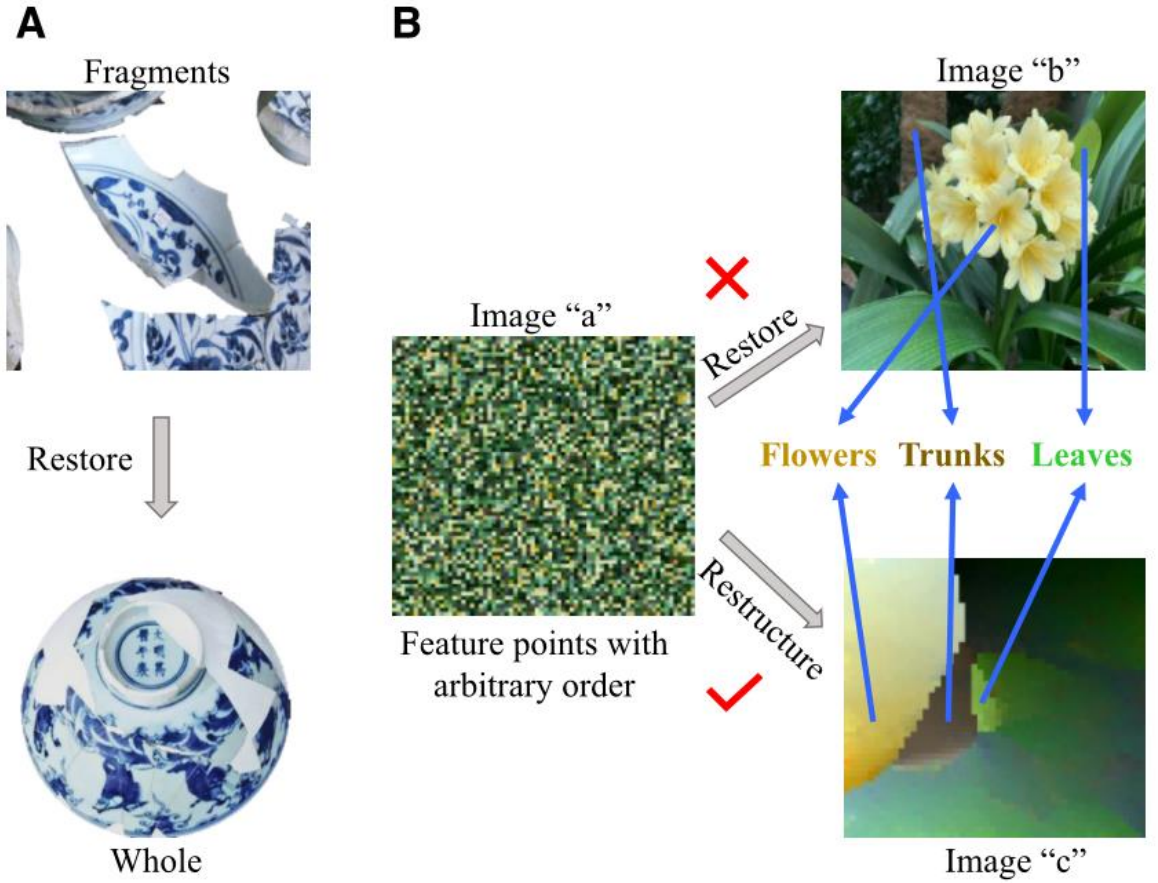
AggMap 对学习 BioHULM 数据的有用性：

- AggMap 多通道 Fmap 在 AggMapNet 对多个数据集的学习上显示出比单通道 Fmap 显着的改进和更好的鲁棒性。
- AggMap 优于现有的 2D 特征工程方法，例如 Lyu-reshape 和 Bazgir-REFINED (基于 RNA-seq 的泛癌分类的多任务。在细胞周期数据集中，AggMap 可以通过聚合和分组 FP 轻松获取特定阶段的基因。
- 多通道 AggMapNet 优于六个 ML 模型，它们是 k-最近邻 (kNN)、L2 正则化多项逻辑回归 (LGR)、随机森林 (RF)、旋转森林 (RotF)、Xgboost (XGB) 和 LightGBM (GBM)) 在 18 个低样本转录组基准数据集中的大多数中。
- 基于 AggMapNet 中开发的 Simply-explainer，我们进一步探索了用于 COVID-19 检测和严重性预测的重要生物标志物。那些确定的 COVID-19 相关生物标志物与文献报道的发现或生物学机制高度一致

Materials and methods

AggMap 特征重构的动机：

人类能够对破碎的碎片对象进行逻辑还原，例如解决拼图游戏或恢复文化财产，如图[2A]所示。这种能力源于预先学习的先验知识，可以根据片段的相关性和边缘连接来连接和组合片段。这些知识是通过各种碎片恢复过程学习的。

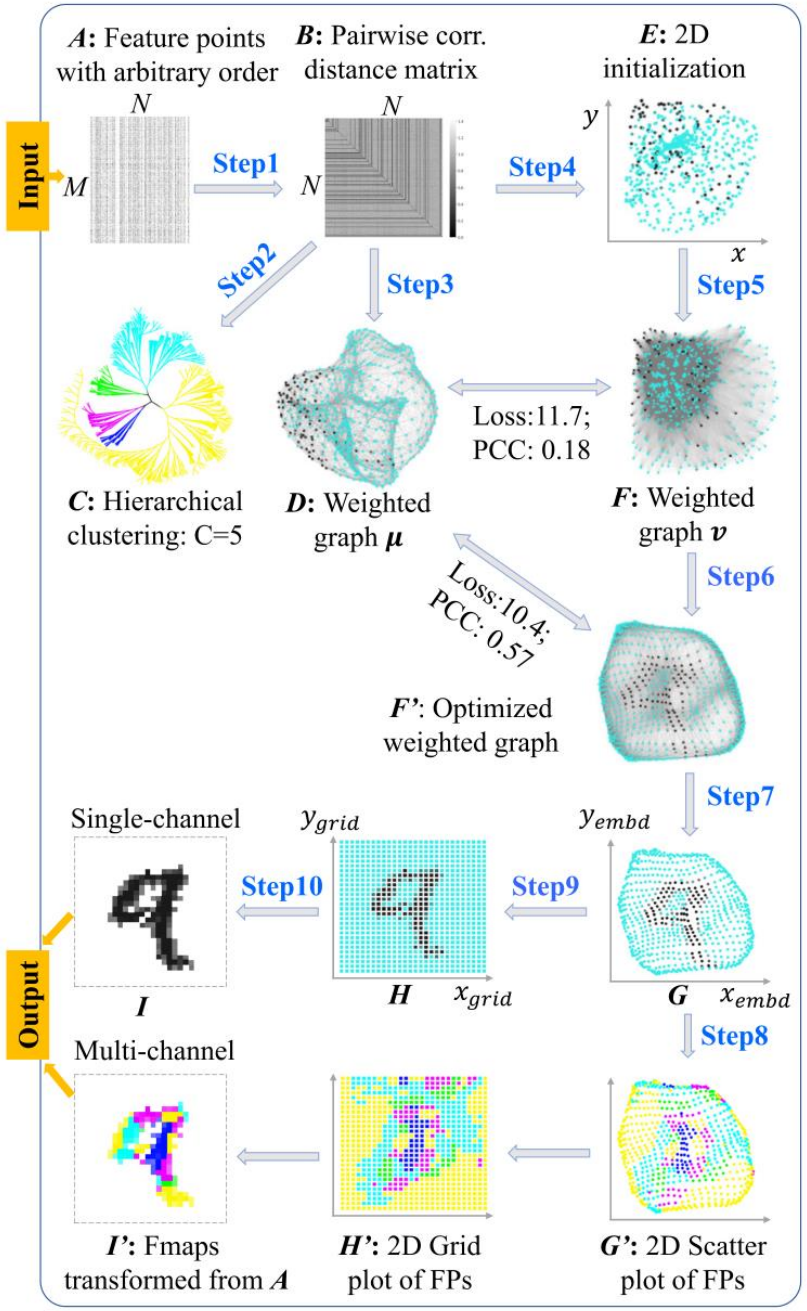


Theoretical basis of unsupervised AggMap

为了将无序的 FP 重构为结构化的 Fmap，自监督 AggMap 需要一个度量来衡量 FP 之间的相似性，一种嵌入 FP 的方法，以及一种将嵌入的 FP 分配到规则网格的算法。

在 AggMap 中，这些任务分别由相关度量、基于流形的 UMAP 方法和线性分配 Jonker-Volgenant (J-V) 算法执行。

UMAP 最初是通过将样本嵌入低维空间来开发降维的。它可以有效地聚合相似的 FP，同时保留**它们对本地和全局数据结构的相对接近度**，导致在现实世界数据中降维的 SOTA 性能 。**默认情况下，UMAP 在 AggMap 中用于将 FP 而不是样本嵌入到 2D 空间中。**



A: $Input$: $M \times N, N: features, M: samples$

B: $d_{corr}(x_i, x_j) = 1 - r(x_i, x_j), i, j \in N$

C: $n\_cluster = 5,\ linkage = 'complete'$

D: $\mu_{i|j} = exp(-(d_{corr}(x_i, x_j) - \rho_i) / \sigma_i)$

$\mu_{ij} = \mu_{i|j} + \mu_{j|i} - \mu_{i|j}\mu_{j|i},\ i, j \in N$

E: $d'_{(i,j)} = exp(-d_{corr(i,j)}{}^2)$

$E_{(x,y)} = spectral\ init\ embedding\ (d'_{(i,j)})$

$d_{(i,j)} = \|x - y\|_2,\ i, j \in N$

F: $v_{ij} = (1 + a * d_{(i,j)}^{2b})^{-1}$

F': $v_i = v_i - lr * \frac{\partial CE}{\partial v_i},\ i \in n_{epochs}$

Step6: Minimize the CE to optimize layout:

Loss: $CE(\mu, v) =$

$\sum_{a \in A}(\mu(a) \log\left(\frac{\mu(a)}{v(a)}\right) + (1 - \mu(a))\log(\frac{1-\mu(a)}{1-v(a)}))$

Optimizer: SGD

Step9: Minimize the CM to assign FPs into grid:

Cost Matrix: $CM(embd, grid)_{(i,j)} =$

$(x_i^{embd} - x_j^{grid})^2 + (y_i^{embd} - y_j^{grid})^2$

Optimizer: LAPJV

AggMap 的拟合有九个步骤，使用随机 MNIST FP 的重构（像素是随机排列的 MNIST）作为示例。

基于规则网格H(H‘)，可以将输入的MXN FFP变换为形状为(M，w，h，c)的标准四维张量，其中M、w、h和c分别为输入样本数、Fmap的宽度、高度和通道。

步骤1

步骤1，给定一个输入表格数据A，其形状为M*N,M为样本，N为特征数量。AggMap通过相关距离测量 FP 的成对距离，生成距离矩阵B。

成对相关距离定义为dcorr(xi,xj)：

$$r\left(x_i,\ x_j\right)=\frac{\sum_{a=1}^{M}\left(x_i^a-\overline{x_i}\right)\left(x_j^a-\overline{x_j}\right)}{\sqrt{\sum_{a=1}^{M}\left(x_i^a-\overline{x_i}\right)^2\sum_{a=1}^{M}\left(x_j^a-\overline{x_j}\right)^2}} \quad (1)$$

$$d_{corr}\left(x_i,\ x_j\right)=1-r\left(x_i,\ x_j\right),i,\ j\ \in N;\ a\in M \quad (2)$$

步骤2

对 FP 进行层次聚类，根据计算的B生成聚类C，其中使用了**完整的链接**，**默认的簇数为5。**

这种聚类操作将 FP 分成不同的组（簇），每个集群被单独嵌入到一个单独的 Fmap 通道中，用于通过 CNN 分类器进行特征组特定或特征选择性学习。由于多通道彩色图像比灰度图像包含更多信息，因此多通道 AggMap Fmap 具有更丰富和可区分的模式。

为了在 AggMap 工具中可视化多通道 Fmap，每个通道的 FP 被用不同的颜色着色，颜色的亮度对应于 FP 值。FP 集群的最佳数量是一个超参数（在 AggMap 超参数部分中描述）。

- **单连接(single linkage)：**计算每一对簇中最相似两个样本的距离，并合并距离最近的两个样本所属簇;
- **全连接(complete linkage):**通过比较找到分布于两个簇中最不相似的样本(距离最远)，从而来完成簇的合并;
- **平均连接(average linkage):**合并两个簇所有成员间平均距离最小的两个簇;ward连接:合并的是使得SSE增量最小的两个簇;

步骤3

步骤3是UMAP图构建的第一阶段，但与默认的UMAP使用**欧氏距离构建加权拓扑k-邻域图不同，AggMap使用相关距离B以指数概率分布构建加权图D**：

$$\mu_{i|j} = exp\left(-\left(d_{corr}\left(x_i,\ \ x_j\right) - \ \rho_i\right)/\sigma_i\right),$$

$$\mu_{ij} = \ \mu_{i|j} + \mu_{j|i} - \mu_{i|j}\mu_{j|i}$$

高维概率对称化

确保流形的本地连通

为每个数据点提供了局部自适应的指数内核，因此距离度量因点而异。

图D是无向加权图，其邻接矩阵由μi j给出，这种结构提供了数据的适当的模糊拓扑表示。

步骤4、步骤5

为了在步骤4中初始化FP的2D坐标，AggMap使用频谱布局来初始化嵌入E，AggMap利用**相关距离B**来初始化嵌入E。为了确保更均匀的初始化嵌入，AggMap首先将B的这个距离矩阵dcorr(i，j)转换成指数亲和度矩阵d'(i，j)。

$$d'_{(i,j)} = exp\left(-d_{corr(i,j)}{}^2\right)$$

最后，利用拉普拉斯特征映射(LE)算法将矩阵d'(i，j)用于谱嵌入。LE使用图拉普拉斯的频谱分解来找到数据的低维表示：

$$E_{(x,\ y)} = LE\_Spectral\,Embedding\left(d'_{(i,j)}\right) \qquad (6)$$

$$d_{(i,j)} = \|x - y\|_2, \quad i, j \in N, \qquad (7)$$

$$v_{ij} = \left(1 + a * d_{(i,j)}^{2b}\right)^{-1},$$

步骤6-9

图F的布局优化。由于存在两个加权图D和F，AggMap通过最小化两个拓扑表示D和F之间的误差来优化图F到F'的布局。交叉熵作为代价函数

$$CE(\mu,\ v) = \sum_{a \in A}\left(\mu(a)\log\left(\frac{\mu(a)}{v(a)}\right)\right.$$
$$\left.+(1-\mu(a))\log\left(\frac{1-\mu(a)}{1-v(a)}\right)\right),$$

$$v_i = v_i - lr * \frac{\partial CE}{\partial v_i},\quad i \in n_{epochs},$$

步骤7：利用优化布局的图F'生成2D嵌入结果G。同时，步骤8通过步骤2中定义的组将G分组到G‘中。G’中的每个颜色是一个聚类组，如C所示。

在步骤9中通过线性分配算法将2D嵌入的FP分配到2D规则网格H(H’)中。J-V算法(22)在分配时保留2D嵌入的邻域关系，而FP被分配到网格点。

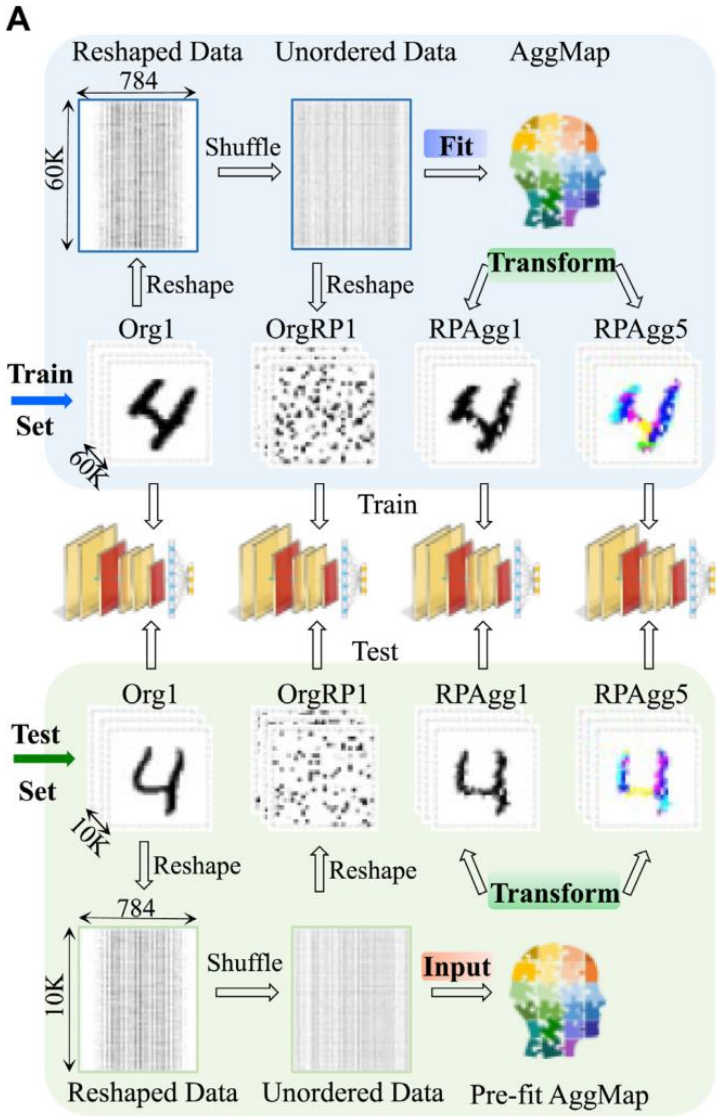
$$CM(embd,\ grid)_{(i,j)} = \left(x_i^{embd} - x_j^{grid}\right)^2$$
$$+\left(y_i^{embd} - y_j^{grid}\right)^2,\ i,\ j\ \in N,$$

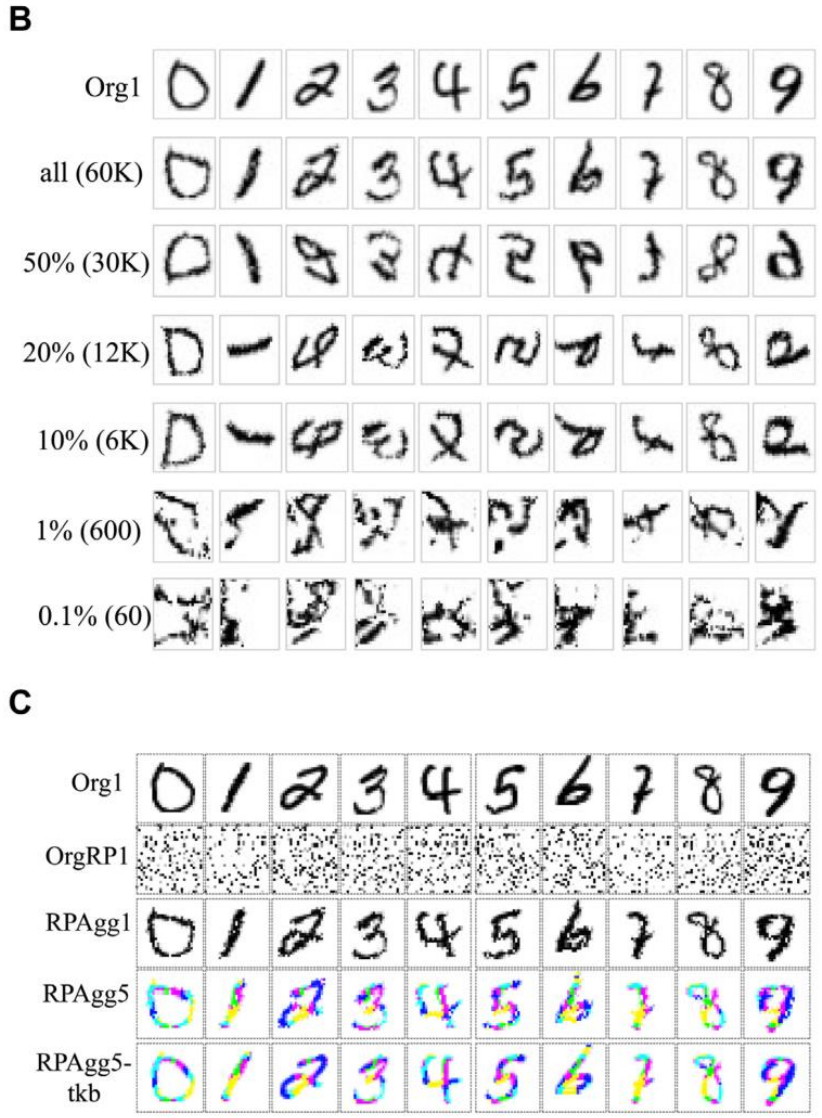
数据集和评估指标

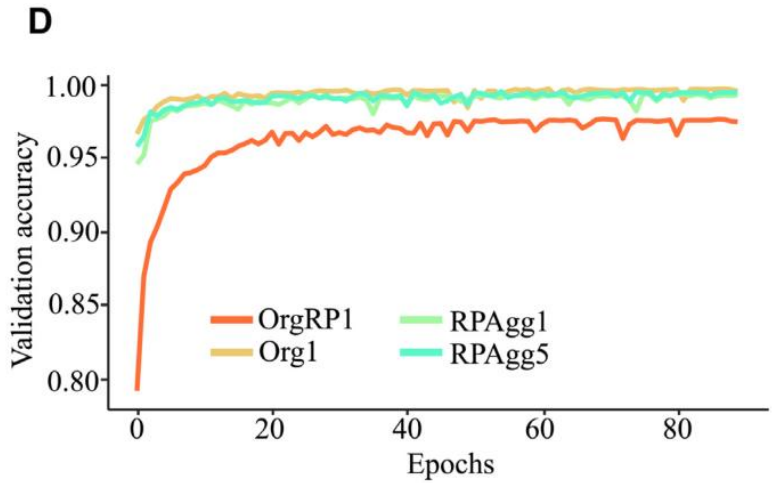
Table 1. Summary of the datasets in this study

Project	Datatype	Dataset	Num. of samples	Num. of features
Proof-of-Concept	Image data	MNIST (24): handwritten digits.	70K images including 10 classes: 60K training set, 10K test set.	28×28 grayscale images, 684 pixels.
	Image data	Fashion-MNIST (38): Zalando's article images.	70K images including 10 classes: 60K training set, 10K test set.	28×28 grayscale images, 684 pixels.
Cell-cycle	Transcriptomics	CCTD-U (39): cell-cycle transcriptome data of U2OS cells	5 different phases of cell cycle (G1, G1/S, S, G2, M) in biological replicates	5162 RNA-seq genes expression of U2OS cells during cell-cycle progression
Pan-Cancer	Transcriptomics	TCGA-T (12): The Cancer Genome Atlas (TCGA) of 33 cancer types.	10446 samples, including 33 cancer types from Pan-Cancer Atlas, the number of samples for each class is ranged from 45 to 1212, with an average of 317. The sample sizes for 15 tumor types are less than 200.	10 381 normalized-level3 RNA-Seq gene expression data.
		TCGA-S (5): TCGA cancer in different stages.	TCGA cancer stage (10 datasets), 249-554 patients in each of 9 datasets, 1,134 in 1 dataset.	17 970 "O" genes with Z-score transformed RNA-Seq gene expression data.
		TCGA-G (5): TCGA cancer in different grades.	TCGA cancer grade (8 datasets), 179-554 patients in each of 8 datasets.	
COVID-19	Proteomics	COV-D (1): Proteomic MALDI-MS data of COVID-19 nasal swabs	363 samples, 211 SARS-CoV-2 positives, and 151 negatives that are from 3 different labs.	88 nasal swabs MALDI-MS signal peaks.
	Proteomics & Metabolomics	COV-S (2): Multi-omics data of COVID-19 sera.	41 patients, including 31 in the training set (18 non-severe and 13 severe) and an independent cohort of 10 patients (6 non-severe and 4 severe).	1486 markers from the sera samples, including 649 proteins and 847 metabolites.

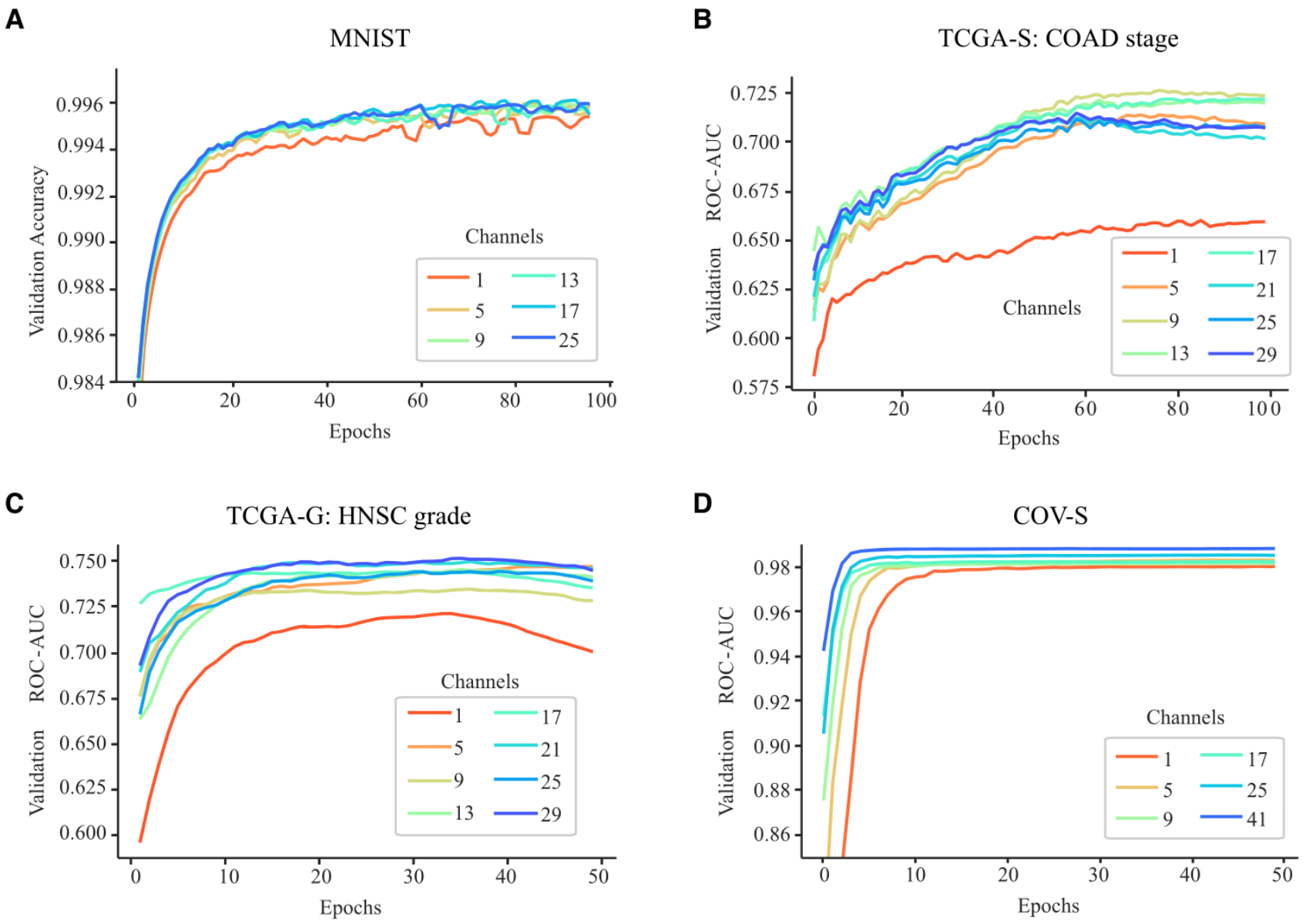
Results AggMap 良好的特征重构能力



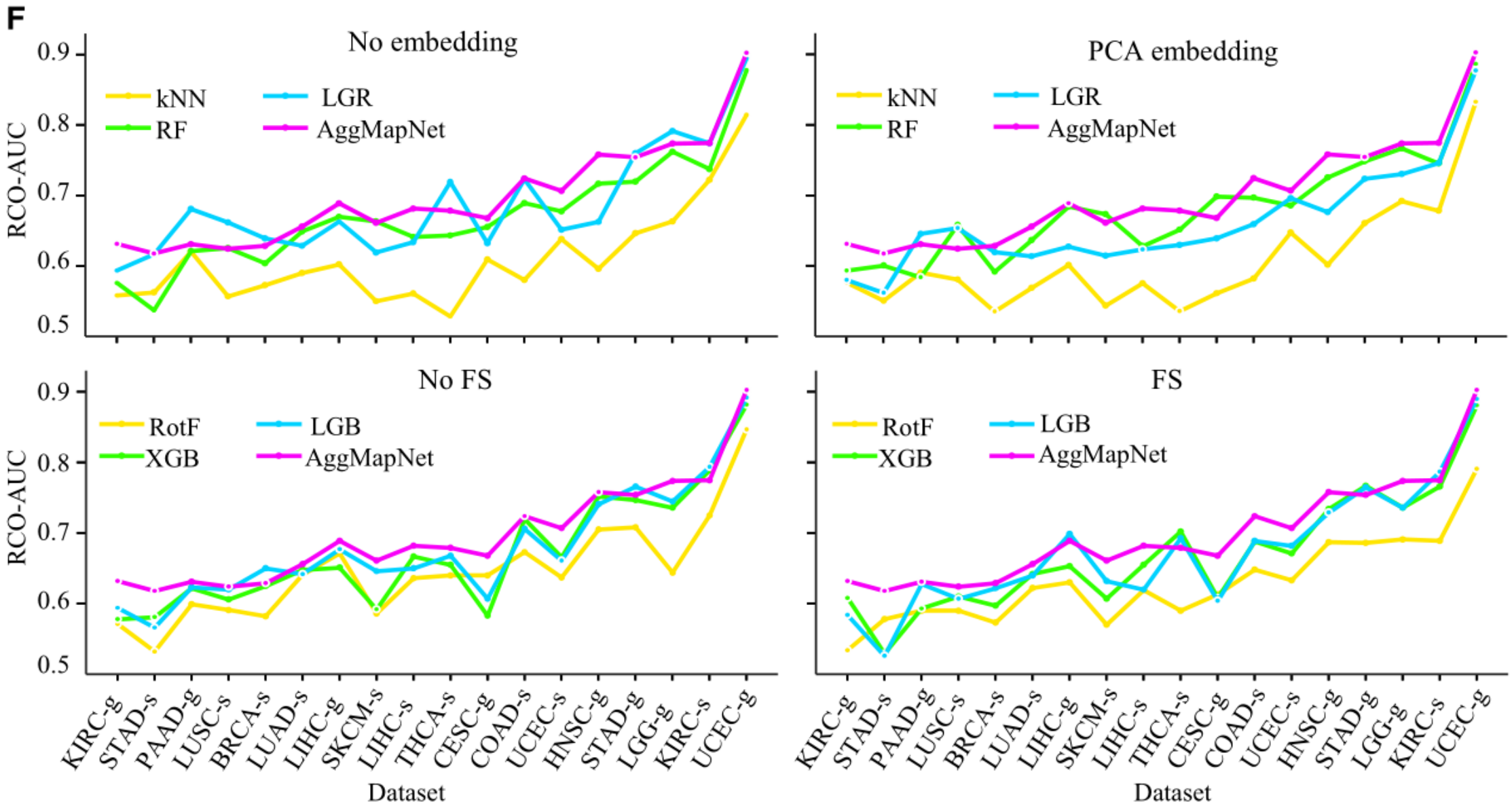




测试了通道数对 MNIST、TCGA-S COAD、TCGA-G HNSC 和 COV-S 四个代表性数据集的影响



AggMap 和 AggMapNet 学习模型在转录组数据 TCGA-T、 TCGA-S 和 TCGA-G 上的表现



Reference

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open Access | Published: 01 September 2020

Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks

Omid Bazgir, Ruibo Zhang, Saugato Rahman Dhruba, Raziur Rahman, Souparno Ghosh & Ranadip Pal

Nature Communications **11**, Article number: 4391 (2020) | Cite this article

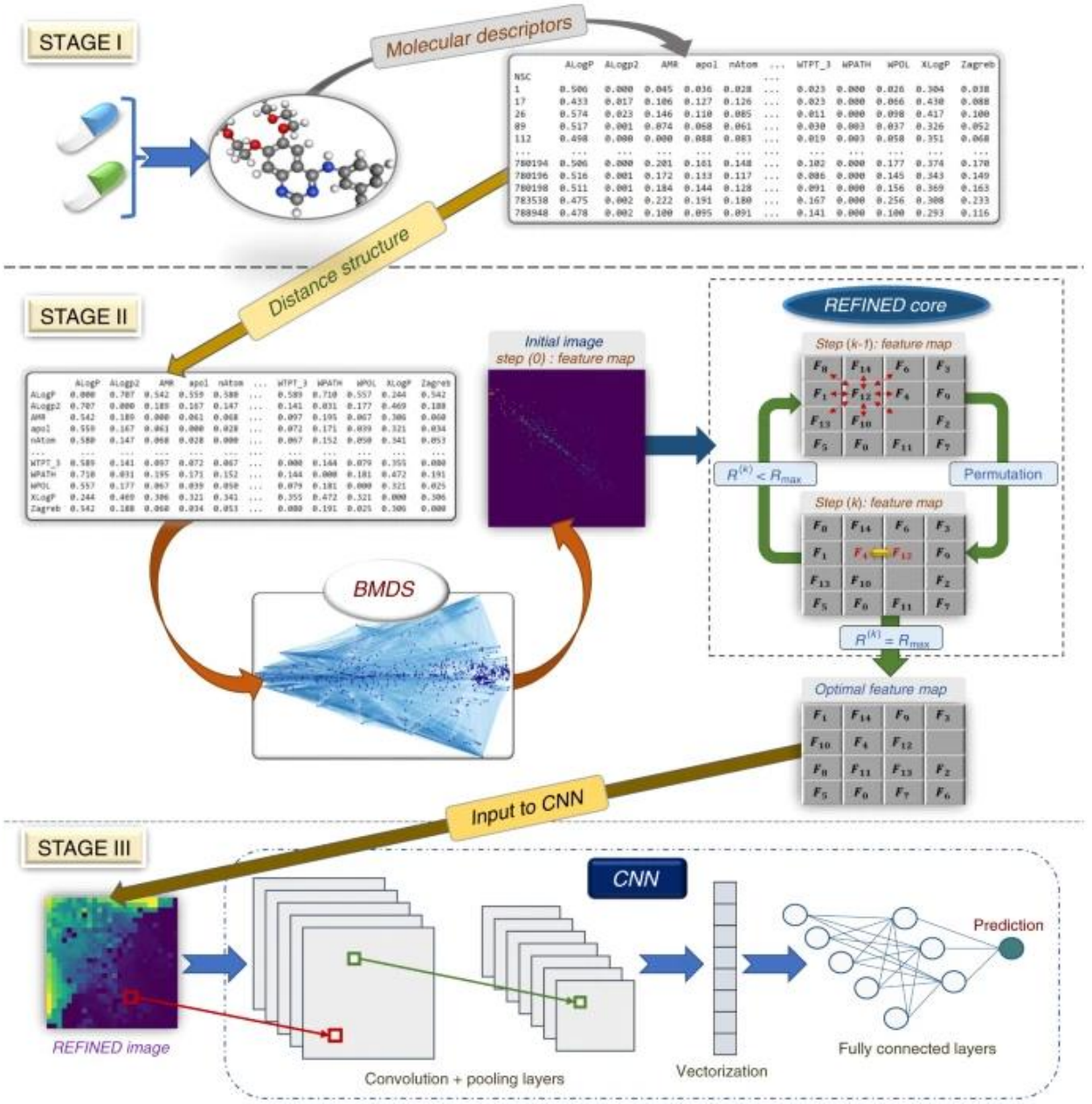
9807 Accesses | **26** Citations | **12** Altmetric | Metrics

将特征表示为具有邻域依赖性的图像，以与卷积神经网络兼容

2020.9,1/德克萨斯理工大学

使用卷积神经网络进行深度学习在基于图像的分类和增强方面显示出巨大的前景，但通常不适合使用没有空间相关性的特征进行预测建模。

本文提出了一种称为 REFINED 的特征表示方法（将特征表示为具有邻域依赖性的图像），以将高维向量排列成紧凑的图像形式，可用于基于 CNN 的深度学习。考虑特征之间的相似性，通过按照贝叶斯度量多维缩放方法最小化成对距离值来生成二维图像形式的简洁特征图。作者假设这种方法可以实现嵌入式特征提取，并且与基于 CNN 的深度学习相结合，可以提高预测准确性。


STAGE I
Molecular descriptors
STAGE II
Distance structure
Initial image
step (0) : feature map
BMDS
REFINED core
Step (k-1): feature map
Permutation
Step (k): feature map
Optimal feature map
Input to CNN
STAGE III
CNN
REFINED image
Convolution + pooling layers
Vectorization
Fully connected layers
Prediction

Reference

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 08 June 2022

Synonymous mutations in representative yeast genes are mostly strongly non-neutral

Xukang Shen, Siliang Song, Chuan Li & Jianzhi Zhang ✉

Nature (2022) | Cite this article

21k Accesses | **596** Altmetric | Metrics

代表性酵母基因的同义突变大多是强非中性的

2022.6.8/美国密歇根大学安娜堡分校生态与进化生物学系

- 尽管基因中的一些突变改变了该基因编码的蛋白质的氨基酸序列，但其他的——称为同义突变——对蛋白质序列没有影响。
- 同义突变并不总是具有选择性的中性或“沉默”效应。例如，基因的序列会影响其表达水平，这意味着同义突变会影响蛋白质丰度，基因转录的信使 RNA 的序列会影响其形状和稳定性。一些物种在其整个基因组中已经进化为比其他物种更多地使用某些密码子。这些现象都表明同义变化可以对细胞功能和有机体适应性产生微小但根本的影响。

- 作者着手系统地探索出芽酵母*酿酒酵母同义变化的影响*. 使用基因组编辑技术 CRISPR-Cas9 在 21 个已知具有不同功能和表达水平的酵母基因中创建了数千个同义和非同义突变。他们将携带这些突变中的每一个的菌株与非突变细胞一起培养，跟踪每个突变序列是否随着时间的推移在群体中变得或多或少频繁——这是它们对细胞适应性影响的指标。

表达水平　　适应度　　mRNA稳定性　　……

Reference

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 01 June 2022

A tissue-like neurotransmitter sensor for the brain and gut

Jinxing Li, Yuxin Liu, Lei Yuan, Baibing Zhang, Estelle Spear Bishop, Kecheng Wang, Jing Tang, Yu-Qing Zheng, Wenhui Xu, Simiao Niu, Levent Beker, Thomas L. Li, Gan Chen, Modupeola Diyaolu, Anne-Laure Thomas, Vittorio Mottini, Jeffrey B.-H. Tok, James C. Y. Dunn, Bianxiao Cui, Sergiu P. Pașca, Yi Cui, Aida Habtezion, Xiaoke Chen ✉ & Zhenan Bao ✉

Nature **606**, 94–101 (2022) | Cite this article

13k Accesses | **159** Altmetric | Metrics

用于大脑和肠道的类组织神经递质传感器

2022.6.1/斯坦福大学化学工程系

研究人员通过使用**基因工程荧光探针**在神经递质传感方面取得了很大进展。**生物电子神经接口**也被用于研究野生动物甚至人类，但这些设备主要集中在神经系统的电生理学。然而，研究神经化学的生物电子工具是有限的。它们往往是刚性和易碎的，并且可能导致对目标组织的不良刺激或炎症反应，使其不适合监测软组织。

需要能够监测中枢和外周神经系统中神经递质的自然时空动态的软生物电子接口，而不会干扰大脑和肠道等柔软和活动器官的生理机能。

快速扫描循环伏安法

快速升高和降低施加在探针上的电压，以反复氧化和还原目标神经递质，从而产生特定于神经递质的电流。

石墨烯作为我们的电极材料

作为多巴胺和血清素等单胺类神经递质氧化的催化剂。它还具有优异的电性能，生物相容性，受弯曲、拉伸和扭曲