组会汇报

汇报人：Lilian

2021/12/9

文献来源

nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00325-y

Check for updates

Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms

Roman Schulte-Sasse[1], Stefan Budach[1], Denes Hnisz[1] and Annalisa Marsico[1,2] ✉

整合多组学数据，利用图卷积网络识别新的癌症基因及其相关分子机制

内容概述

- ✓ 研究团队开发了一款**基于图卷积网络（GCN）**的机器学习算法——EMOGI（Explainable Multiomics Graph Integration）。
- ✓ 该算法集成了从患者样本中生成的数以万计的数据集，这些数据集包括**突变的DNA序列数据、DNA甲基化、单个基因活性以及细胞通路中蛋白质相互作用信息。**在这些数据中，深度学习算法可检测导致癌症发展的模式和分子原理。
- ✓ 作者识别出了165个新的癌症基因，它们**不一定具有复发性改变(harbour recurrent alterations)，但会与已知的癌症基因相互作用**；
- ✓ 所有这些新发现的癌基因都与已知的著名癌基因有紧密相互作用。而且细胞实验证实它们对肿瘤细胞的生存至关重要。
- ✓ 这种方法可以为精确肿瘤学开辟新的途径，并应用于预测其他复杂疾病的生物标志物。

Introduction

面临的问题

尽管数以千计的癌症基因组序列有助于癌症相关基因的鉴定，但仍然存在以下挑战：

1、在几种肿瘤类型中确定的癌症基因数量仍然很低，

2、许多在肿瘤发生中发挥重要作用的基因在其DNA序列水平上没有改变，而是通过诸如DNA甲基化、拷贝数改变等各种细胞机制失调

这些非突变的癌症依赖基因引起了人们的极大兴趣，因为它们中的许多是转录和表观遗传调控因子，可以通过小分子治疗进行靶向治疗。

- ✓ 蛋白质-蛋白质相互作用(PPI)网络中包含的信息在预测癌症基因时非常重要
- ✓ 生物网络可以被看作是图，其中节点代表基因，节点之间的连接代表基因-基因的相互作用，而组学数据水平可以被视为基因的特征向量

Introduction

已有方法

- 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

该研究团队开发了一种基于图卷积网络的机器学习方法：**EMOGI**，通过将突变、拷贝数变化、DNA甲基化和基因表达等多组学泛癌数据与蛋白质-蛋白质相互作用（PPI）网络相结合来预测癌症基因。

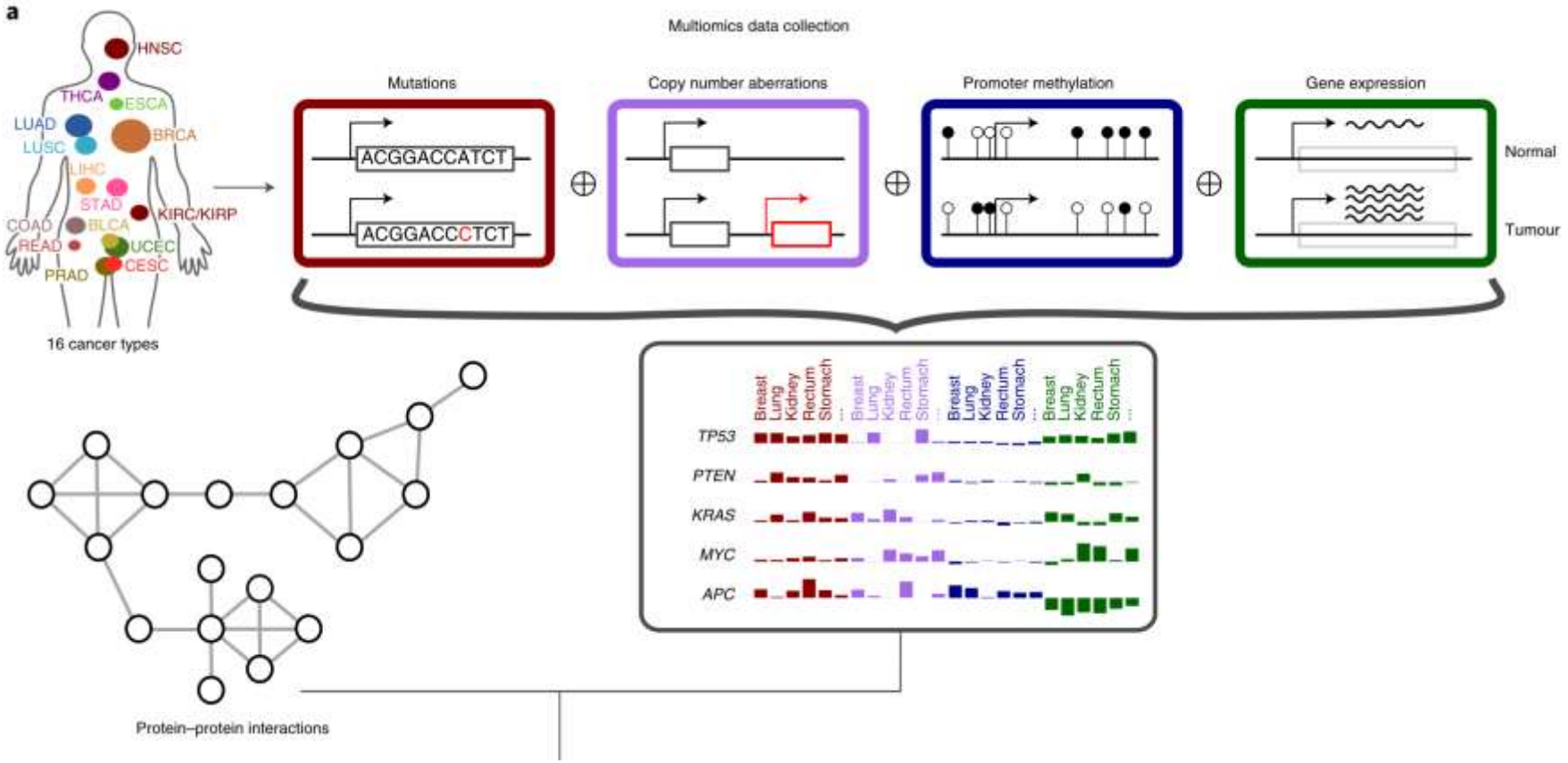
该研究团队预测了165个新的癌症基因，**这些基因与PPI网络中已知的癌症驱动因素相互作用，而不是自身高度突变**，新的预测还丰富了通过功能丧失筛查确定的必需基因。该方法能够找到由不同的分子改变而非高突变率定义的癌症基因类别，**阐释了基因如何促进或阻碍肿瘤的发展和进展。**

Data collection and preprocessing

我们收集了29446份TCGA样本的突变、拷贝数、DNA甲基化和基因表达数据，涵盖16种不同的癌症类型

Study Name	Cancer Type	Tumor	Same Tissue Normal	TCGA	Synapse Filtered	Tumor	Same Tissue Normal	GTEx Tissue	
BRCA	Breast invasive carcinoma	796	96	1044	Yes	1109	113	218	TRUE
GBM	Glioblastoma multiforme	153	2	396	Yes	169	5	1403	FALSE
OV	Ovarian serous cystadenocarcinoma	10	-	433	Yes	379	-	108	FALSE
LUAD	Lung adenocarcinoma	475	32	569	Yes	535	59	374	TRUE
UCEC	Uterine Corpus Endometrial Carcinoma	439	46	542	Yes	552	35	90	TRUE
KIRC	Kidney renal clear cell carcinoma	324	160	339	Yes	538	72	36	TRUE
HNSC	Head and Neck squamous cell carcinoma	530	50	510	Yes	502	44	70	TRUE
LGG	Brain Lower Grade Glioma	534	-	513	Yes	529	-	-	FALSE
THCA	Thyroid carcinoma	515	56	496	Yes	510	58	355	TRUE
LUSC	Lung squamous cell carcinoma	370	42	497	Yes	502	49	374	TRUE
PRAD	Prostate adenocarcinoma	503	50	498	Yes	499	52	119	TRUE
SKCM	Skin Cutaneous Melanoma	472	2	470	Yes	470	1	974	FALSE
COAD	Colon adenocarcinoma	315	38	433	Yes	480	41	203	TRUE
STAD	Stomach adenocarcinoma	395	2	441	Yes	375	32	204	TRUE
BLCA	Bladder Urothelial Carcinoma	419	21	412	Yes	414	19	11	TRUE
LIHC	Liver hepatocellular carcinoma	380	50	375	No	374	50	136	TRUE
CESC	Cervical squamous cell carcinoma and en	309	3	305	Yes	306	3	11	TRUE
KIRP	Kidney renal papillary cell carcinoma	275	45	288	Yes	288	32	36	TRUE
SARC	Sarcoma	265	4	255	No	263	2	621	FALSE
LAML	Acute Myeloid Leukemia	140	-	149	Yes	151	-	456	FALSE
ESCA	Esophageal carcinoma	186	16	184	No	162	11	790	TRUE
PAAD	Pancreatic adenocarcinoma	185	10	183	Yes	178	4	197	FALSE
PCPG	Pheochromocytoma and Paraganglioma	181	3	179	No	180	3	-	FALSE
READ	Rectum adenocarcinoma	99	7	158	Yes	167	10	173	TRUE
TGCT	Testicular Germ Cell Tumors	150	-	150	No	150	-	203	FALSE
THYM	Thymoma	124	2	123	No	119	58	-	FALSE
KICH	Kidney Chromophobe	66	-	66	No	65	24	36	FALSE
ACC	Adrenocortical carcinoma	80	-	92	No	79	-	159	FALSE
MESO	Mesothelioma	87	-	83	No	86	-	-	FALSE
UVM	Uveal Melanoma	80	-	80	No	80	-	-	FALSE
DLBC	Lymphoid Neoplasm Diffuse Large B-cell L	48	-	37	No	48	-	-	FALSE
UCS	Uterine Carcinosarcoma	57	-	57	No	56	-	90	FALSE
CHOL	Cholangiocarcinoma	36	9	51	No	36	9	-	FALSE
SUM		8998	746	10408	20	10351	786	7447	0

Data collection and preprocessing



Data collection and preprocessing

- ✓ 基因突变频率(Gene mutation frequencies)——SNVs
- ✓ 拷贝数畸变（CNAs，copy number aberrations）
- ✓ DNA甲基化改变启动子区域（DNA methylation changes in promoter regions.）
- ✓ 基因表达的变化（Gene expression changes.）
- ✓ PPI networks.

Concatenation of the omics and network features.

每个基因分配一个16 × 4维的载体，其中16为癌症类型的数量，4为四组类型的值，即计算每个癌症类型的SNVs、CNAs、差异甲基化和差异表达。

所有四组数据集分别进行预处理，然后连接形成N行64列的泛癌症矩阵。所有或部分组学类型中缺失的PPI网络基因值被设为零。

GCNs

图卷积网络 将卷积神经网络框架扩展到位于非规则网格上的数据。与图像相反，图中的节点可以具有可变数量的邻居和局部拓扑，这对分类结果产生很重要的影响；

与卷积神经网络类似，GCN 在信号 $x \in \mathbb{R}^N$ （N 是图中的节点数）上扫描，以识别节点局部邻域中的模式。对于整张图来说，由于需要计算拉普拉斯矩阵，所以采用近似的方法

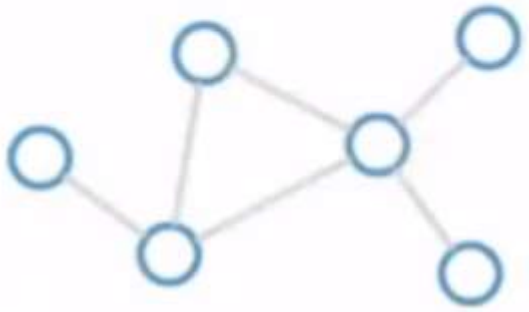
通过近似谱图卷积，GCN 能够通过将图拉普拉斯算子与特征矩阵 $X \in \mathbb{R}^{N \times p}$ 相乘来平均节点周围的相邻信息，其中 N 表示图中节点的数量，p 表示 特征维度。GCN 中每一层的简单传播规则可以定义为：

$$H^{(l+1)} = \sigma\left(LH^{(l)}W^{(l)}\right)$$

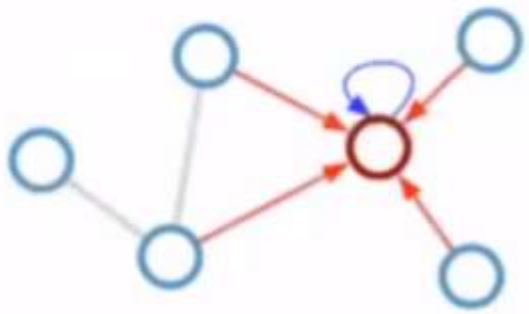
$$H^{(l+1)} = \sigma\left(LH^{(l)}W^{(l)}\right)$$

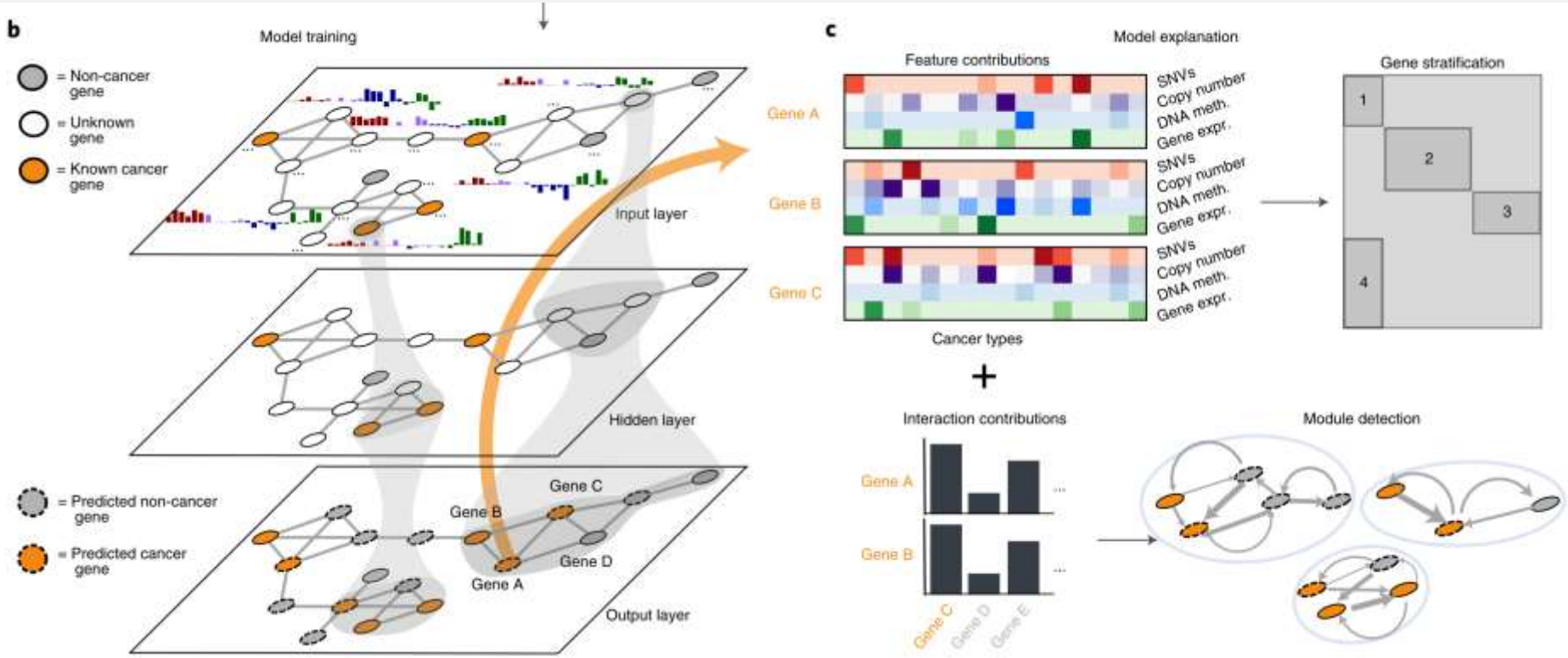
$$L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^{T}$$

无向图



计算更新红点



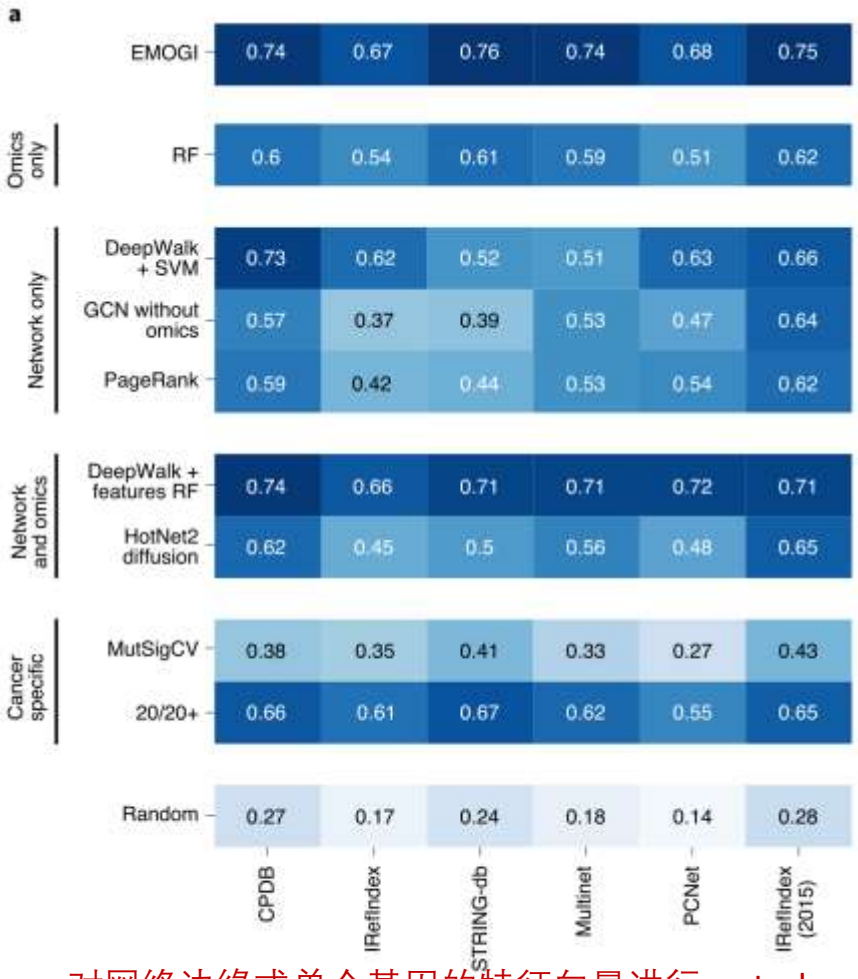


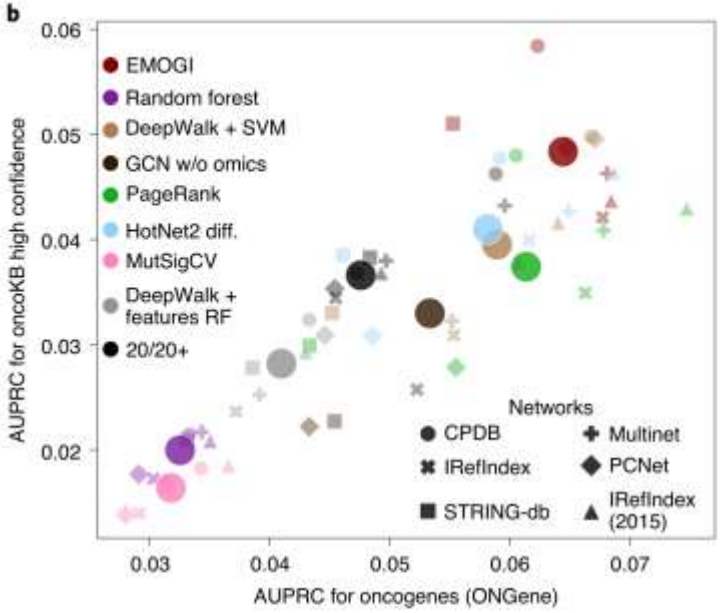
在EMOGI模型训练中，特征通过连续的图卷积层进行转换，输出层根据输出概率将基因分为预测癌症基因和非癌症基因。

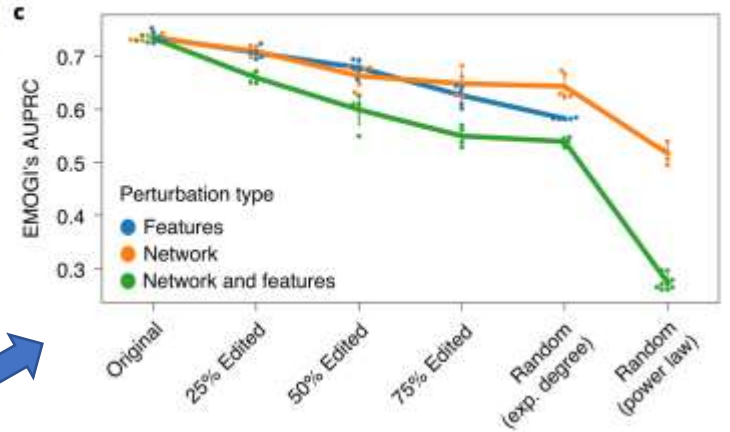
每个基因分类的最重要特征(包括癌症类型和相互作用及组学水平)是使用LRP（Layer-wise Relevance Propagation，相关性逐层传播）提取的。随后根据其特征贡献对基因进行聚类，每个基因的相互作用贡献用于检测癌症中具有重要基因-基因连接的模块。

- ✓ 例如，在CPDB)-PPI网络中，EMOGI识别出89%的KCG和约50%的候选癌症基因(CCGs)；
- ✓ 在六个不同的PPI网络中，它的表现优于所有其他方法

AUPRC values

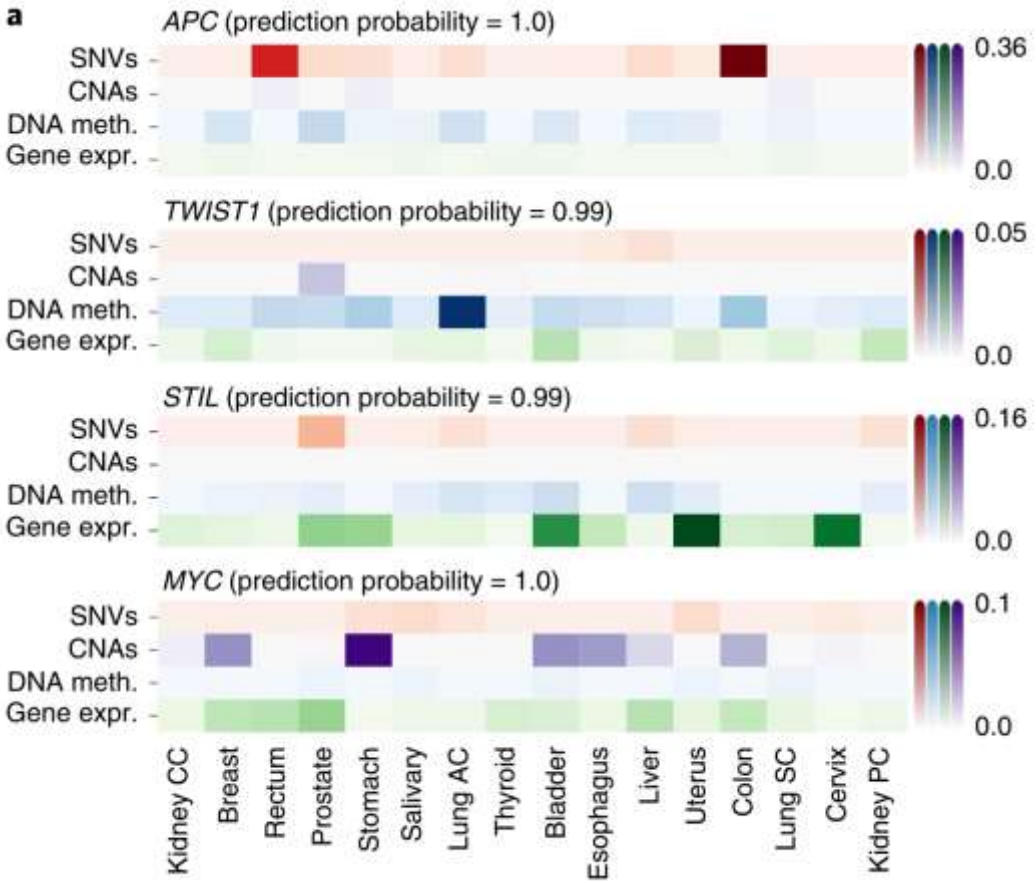






对网络边缘或单个基因的特征向量进行perturb
EMOGI受益于不同的数据表示和多组学集成。

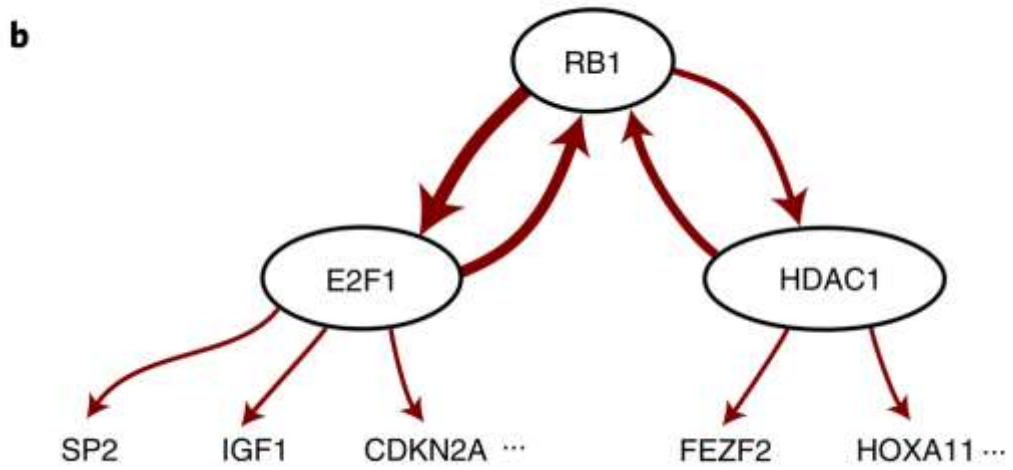
Layer-wise Relevance Propagation Applied to EMOGI to Extract Explanations for Genes



$$R_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j^{(l+1)}$$

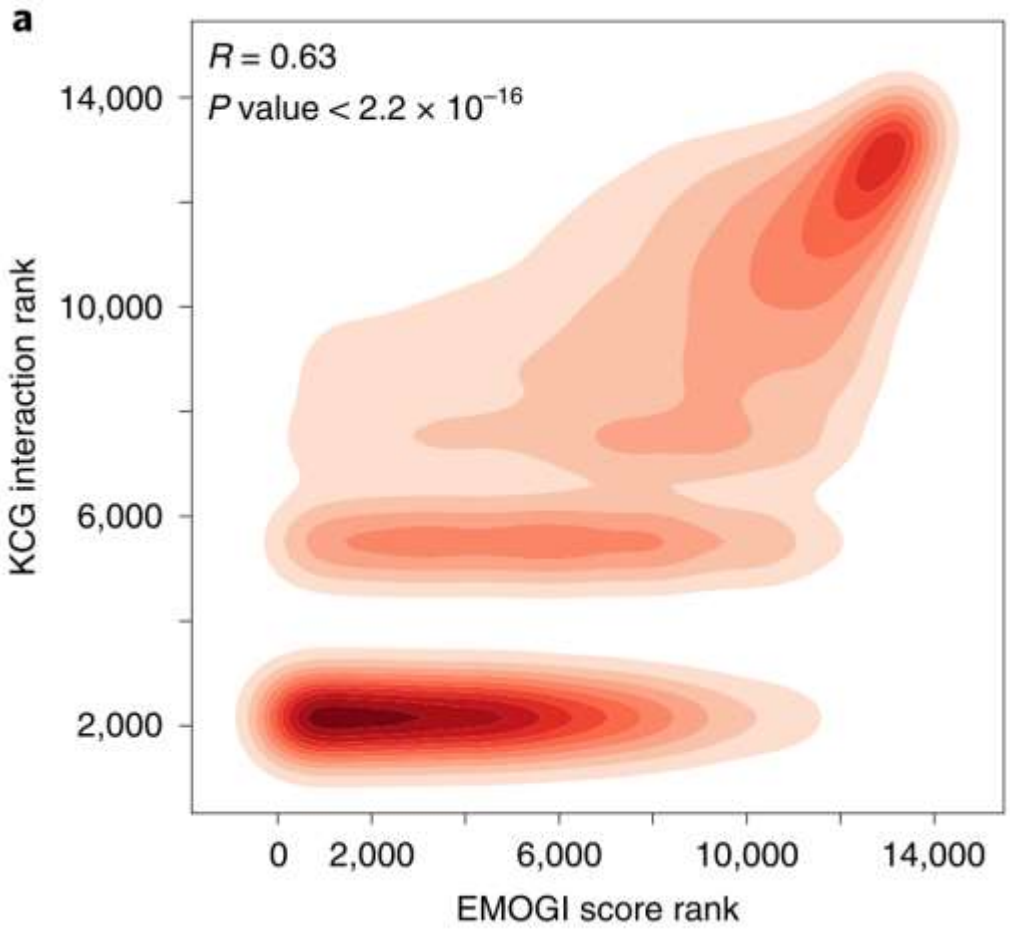
- ✓ 通过LRP规则提取了PPI网络中相互作用伙伴对个体基因分类的贡献，并用于为肿瘤发生提供更多的机制见解。
- ✓ 颜色越深，该类型对特定癌症的贡献就越大。
- ✓ 例如，作者分析了肿瘤抑制基因APC，表明该基因在结直肠癌中高频突变

Layer-wise Relevance Propagation Applied to EMOGI to Extract Explanations for Genes



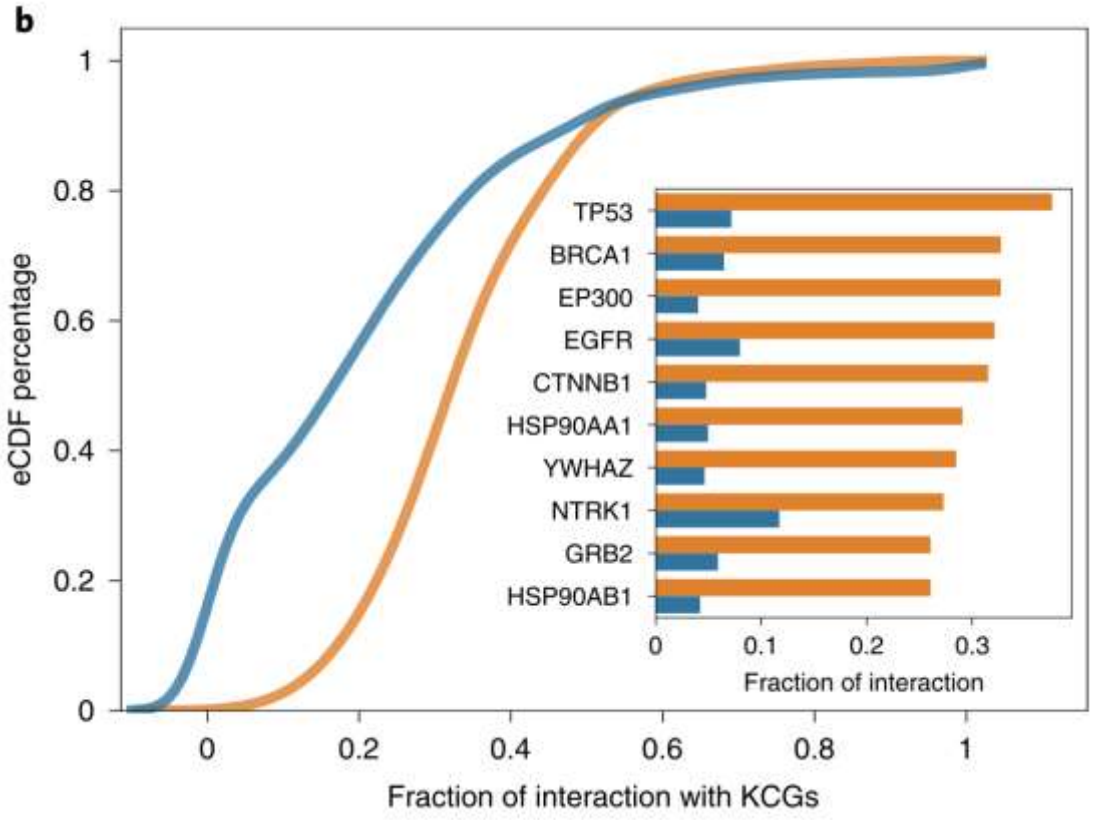
- ✓ 以RB1-E2F1-HDAC1复合物为例。基因A和B之间的直接优势表明基因提取作为重要的互动伙伴B的基因。例如,E2F1和HDAC1被部队最重要的网络邻居的分类癌症RB1基因,因此直接连接到它。同样，RB1是E2F1和HDAC1最重要的邻居。根据LRP，箭头越粗表示相互作用的重要性越高。CC,透明细胞;交流,腺癌;SC,鳞状细胞;电脑,乳头状细胞。

Newly predicted cancer genes.



- ✓ EMOGI评分(代表一个基因成为癌症基因的概率)与该基因与kcg相互作用的数量之间存在显著相关性
- ✓ 所有NPCGs都至少与KCG发生过一次相互作用

Newly predicted cancer genes.

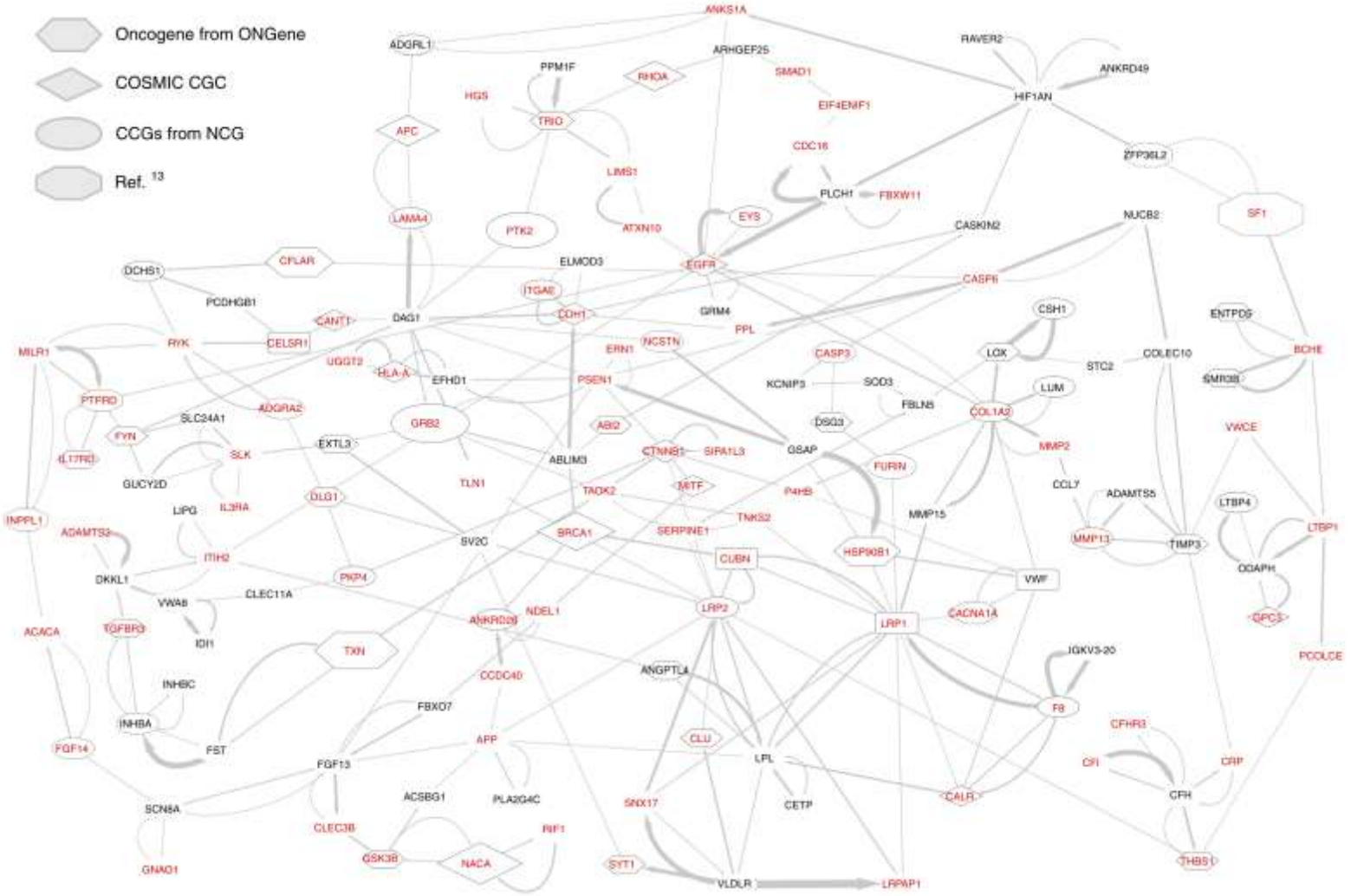


- ✓ EMOGI评分(代表一个基因成为癌症基因的概率)与该基因与kcg相互作用的数量之间存在显著相关性
- ✓ 所有NPCGs都至少与KCG发生过一次相互作用;
- ✓ 著名的癌症基因如TP53, EP300, BRCA1和EGFR在新预测的癌基因的十大相互作用基因中。

- ✓ 鉴定了一个包含149个基因的大SCC，该SCC对应于EMOGI模型用于执行癌症基因分类任务的核心相互作用组
- ✓ 有助于进一步加强我们在细胞通路水平上对癌症起始和发展的理解

癌基因互作网络

互作基因形成得的癌症网络模块有助于进一步加强我们在细胞通路水平上对癌症起始和进展的理解。



总结思路

- ✓ 作者结合多组学信息和基因互作网络信息，借助图卷积神经网络来构建预测癌基因模型，
- ✓ 对该模型进行了评估（与已有方法在多个数据集上表现进行对比）
- ✓ 最后分析了新预测的癌基因（LRP + 强连通相互作用网络）

文献来源

Cell Research

Explore content ∨ About the journal ∨ Publish v

nature › cell research › articles › article

Article | Open Access | Published: 23 February 2021

CELL RESEARCH

期刊影响因子™

2020	五年
25.617	25.924

JCR 学科类别	类别排序	类别分区
CELL BIOLOGY *其中SCIE 版本*	8/195	Q1

Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures

Lei Sun, Kui Xu, Wenze Huang, Yucheng T. Yang, Pan Li, Lei Tang, Tuanlin Xiong & Qiangfeng Cliff Zhang

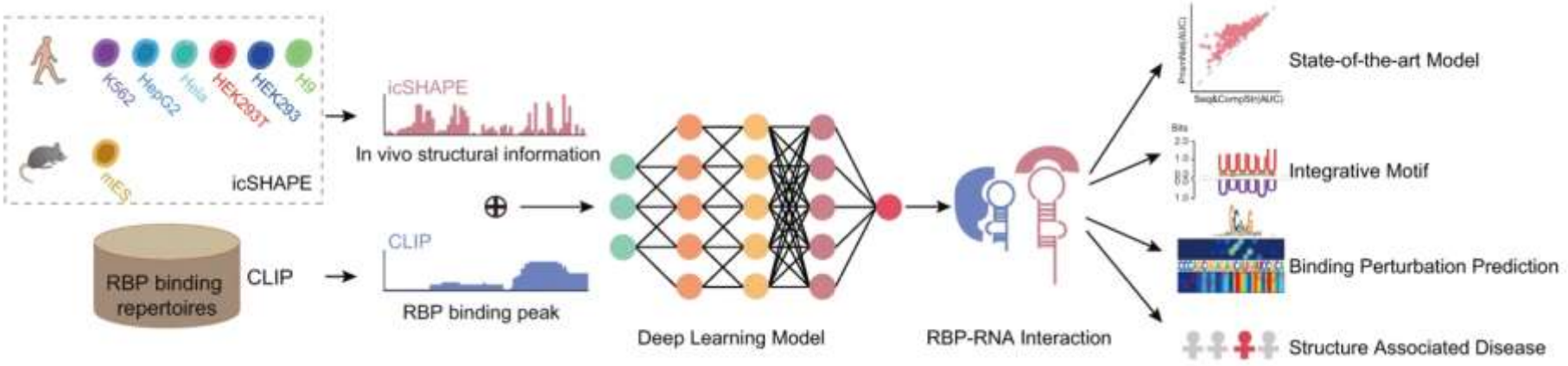
使用体内 RNA 结构、通过深度学习预测动态细胞蛋白质-RNA 相互作用

内容概述

- ✓ 与 RNA 结合蛋白 (RBP) 的相互作用是 RNA 功能和细胞调节不可或缺的一部分，并动态反映特定的细胞状况。
- ✓ 作者开发了一种深度学习工具PrismNet，整合了已有实验中体内 RNA 结构数据和匹配细胞的 RBP 结合数据，以准确预测各种细胞条件下的动态 RBP 结合.

- ✓ 张强锋研究组利用PrismNet模型，使用新冠病毒SARS-CoV-2在宿主细胞内的RNA基因组结构信息，预测了多个新冠病毒的宿主结合蛋白；从这些宿主蛋白出发，找到了一些对抑制新冠传播有效的重定位药物。这个研究再次证明了PrismNet的广阔应用前景

内容概述



- ✓ 作者通过整合细胞内RNA结构信息以及对应细胞系的RBP结合信息，利用**深度神经网络**，**构建了预测RBP结合位点的PrismNet模型。**
- ✓ PrismNet 架构使用**一个卷积层、一个二维残差块和一个通过最大池化连接的一维残差块**来捕获转录本中跨越大距离的序列和结构决定因素。
- ✓ 该模型在168个人类RBP结合的CLIP数据集上进行了训练学习和检验，发现其预测准确率**显著高于**之前仅仅利用RNA序列以及整合基于序列预测得到的RNA结构的方法，预测和CLIP实验结果的吻合度**甚至达到或超过同一条件下两个CLIP实验的吻合度。**
- ✓ 细胞内RNA结构信息对于预测准确率的提高起到了重要作用。有意思的是，作者**发现RNA结构信息对于提高双链结合蛋白预测准确率的帮助更大**

文献来源

nature machine intelligence

Explore content ∨ About the journal ∨ Publish with us ∨ Subscribe

nature > nature machine intelligence > articles > article

Article | Published: 18 January 2021

Deep neural networks identify sequence context features predictive of transcription factor binding

An Zheng, Michael Lamkin, Hanqing Zhao, Cynthia Wu, Hao Su & Melissa Gymrek ✉

Nature Machine Intelligence **3**, 172–180 (2021) | Cite this article

1880 Accesses | **3** Citations | **51** Altmetric | Metrics

深度神经网络识别预测转录因子结合的序列上下文特征

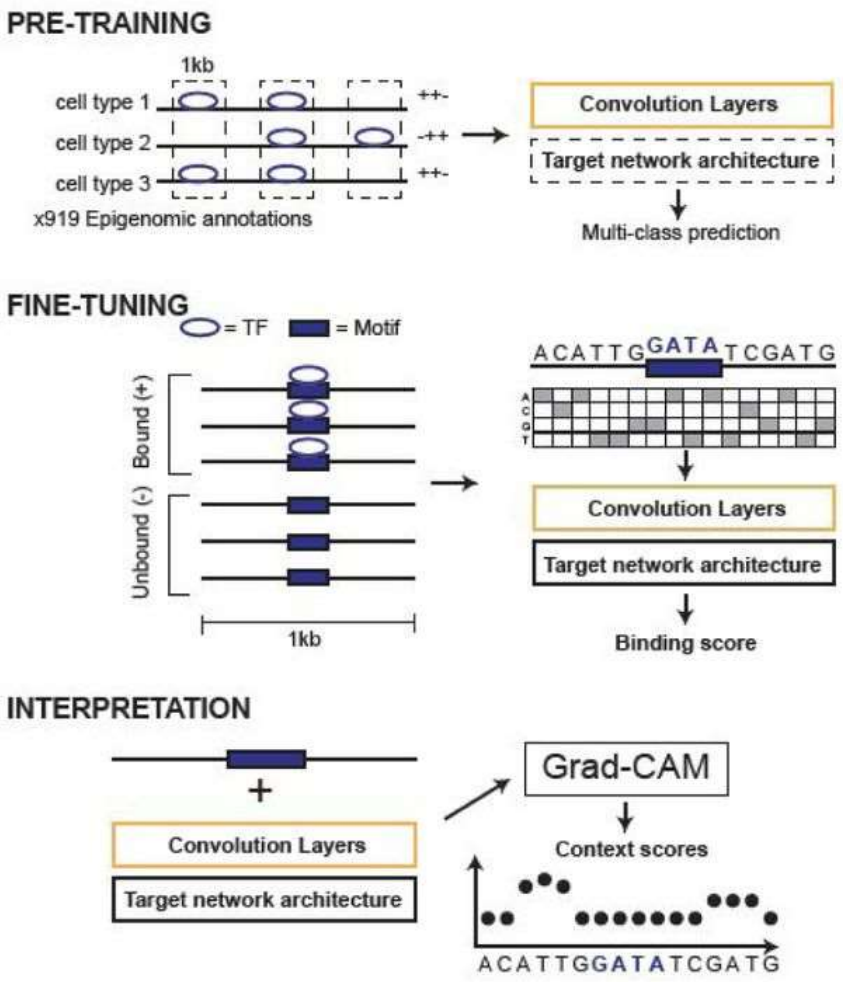
内容概述

- ✓ 研究表明，大多数TF具有独特的结合偏好，只能识别包含特定模式（即核心基序）的DNA序列。但是基序匹配序列和实验确定的结合位点之间通常只有部分重叠。**特定基序是否能够与TF结合取决于许多其他因素**，包括染色质可及性，核小体定位，与其他TF的协同和竞争结合等等。这些因素中有许多与TF基序周围的序列背景有关。
- ✓ 为了研究序列上下文在TF结合中的作用，我们开发了一个名为**AgentBind**的构架，**用于预测是否会结合基序实例，以及寻找对结合状态影响最大的特定核苷酸。**

- ✓ 研究提供了一种有价值的机器学习框架，可帮助解码TF结合其靶位点并可识别对结合作用最强的特定非编码核苷酸。这项技术未来的应用包括：新发现的TF和细胞类型，学习其他TF的细胞类型特定规则，研究TF结合区的选择信号

内容概述

a



- ✓ 研究人员使用这一构架预测**在淋巴母细胞细胞系中38个转录因子的基序结合**，评估了特定碱基对于序列的重要性，并表征最能预测结合的序列特征。
- ✓ 该模型架构包含三个步骤：**预演，微调，解释（pre-training, fine-tuning, interpretation）**,并应用DanQ作为模型体系结构。
- ✓ 首先利用pre-train DanQ模型从多个细胞获取外遗传性注释类型。随后将为每个TF建立一个二进制数据集：提取以基序实例为中心的1kb基因组序列，并根据与ChIP测序确定的结合位点重叠将每个序列标记为结合（阳性）与未结合（阴性）。
- ✓ 每个二进制数据集用于微调一个单独的预演模型，使其学习TF绑定的重要功能。
- ✓ 最后将使用一种名为Grad-CAM的模型解释方法对每个核苷酸对结合预测的贡献进行评分

文献来源

nature communications

Article | Open Access | Published: 30 November 2021

A meritocratic network formation model for the rise of social media influencers

Nicolò Pagan, Wenjun Mei ✉, Cheng Li & Florian Dörfler

Nature Communications **12**, Article number: 6865 (2021) | Cite this article

2446 Accesses | **51** Altmetric | Metrics

社交媒体影响者崛起的精英网络形成模型

内容概述

- ✓ 许多当今最常用的在线社交网络（例如 Instagram、YouTube、Twitter 或 Twitch）都基于**用户生产的内容 (UGC)**。
- ✓ 这些基于 UGC 的定向在线平台在很大程度上影响了我们的社会，例如，舆论两极分化或（错误）信息的传播。用户现在可以通过他们的 UGC迅速获得人气，并成为所谓的新影响者。
- ✓ 在探究“网红”崛起之路时，需要理解两个问题：**用户生产的内容与快速成长的“网红”之间有着怎样的关联？由此产生的社交网络有哪些特征？**

- ✓ 研究团队结合了现代社交平台的特点和既有模型，**提出了一个预测社交网络形成的简单数学模型。**在这个模型中，内容质量是主要影响因素，此外作者也考虑了用户的功利主义原则。用户会根据自己的兴趣，基于内容质量决定关注与否。
- ✓ 在一个由6000多名科学家组成的网络中，研究团队针对推特数据测试了他们的模型，从而获得了网络形成的动态过程。
- ✓ 结果表明，**这些用户想要提高自己接收的内容的质量，因此他们会持续搜索优质内容提供者，从而形成自己的社交平台网络。**在这个过程中，他们会重点评判内容与自身兴趣的一致性、内容同质性，以及质量有多高。

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 29 September 2021

Circadian autophagy drives iTRF-mediated longevity

Matt Ulgherait, Adil M. Midoun, Scarlet J. Park, Jared A. Gatto, Samantha J. Tener, Julia Siewert, Naomi Klickstein, Julie C. Canman, William W. Ja & Mimi Shirasu-Hiza ✉

Nature **598**, 353–358 (2021) | Cite this article

16k Accesses | **251** Altmetric | Metrics

昼夜节律自噬驱动 iTRF 介导的长寿

内容概述

- ✓ 限时喂养(每天固定时间禁食6～12小时)
- ✓ 限时进食（TRF）有助于改善代谢健康，可能有潜在的抗衰老作用，但是TRF潜在的分子作用机制尚不清楚。
- ✓ 在这里，作者为了开发果蝇的遗传工具和特征明确的衰老标记物，设计了一种间歇性TRF (iTRF) 饮食方案，**该方案显著的延长了果蝇的寿命，并推迟了肌肉和肠道中衰老标记物的出现。**

- ✓ iTRF的这种作用机制不同于其他的促长寿干预（热量限制、限制膳食蛋白质、抑制胰岛素样信号），也不依赖于肠道菌群，而是通过**增强生物钟调控的夜间自噬作用来发挥其健康效益的。**
- ✓ **iTRF能增强昼夜节律性的生物钟和自噬基因表达，且其作用也依赖于这二者；**
- ✓ **敲除核心生物钟基因或抑制自噬介导因子表达，以及改为白天禁食/夜间进食或在白天诱导自噬，均使iTRF失效；**
- ✓ 增强夜间的自噬，能在随意进食的果蝇中起到与iTRF一样的有益作用，是iTRF抗衰老、促长寿的主要机制。
- ✓ 这些发现为研发促进健康长寿的饮食和药物干预手段，提供了新的理论和证据支持。

文献来源

nature aging

Explore content ∨ About the journal ∨ Publish with us ∨ Subscribe

nature > nature aging > articles > article

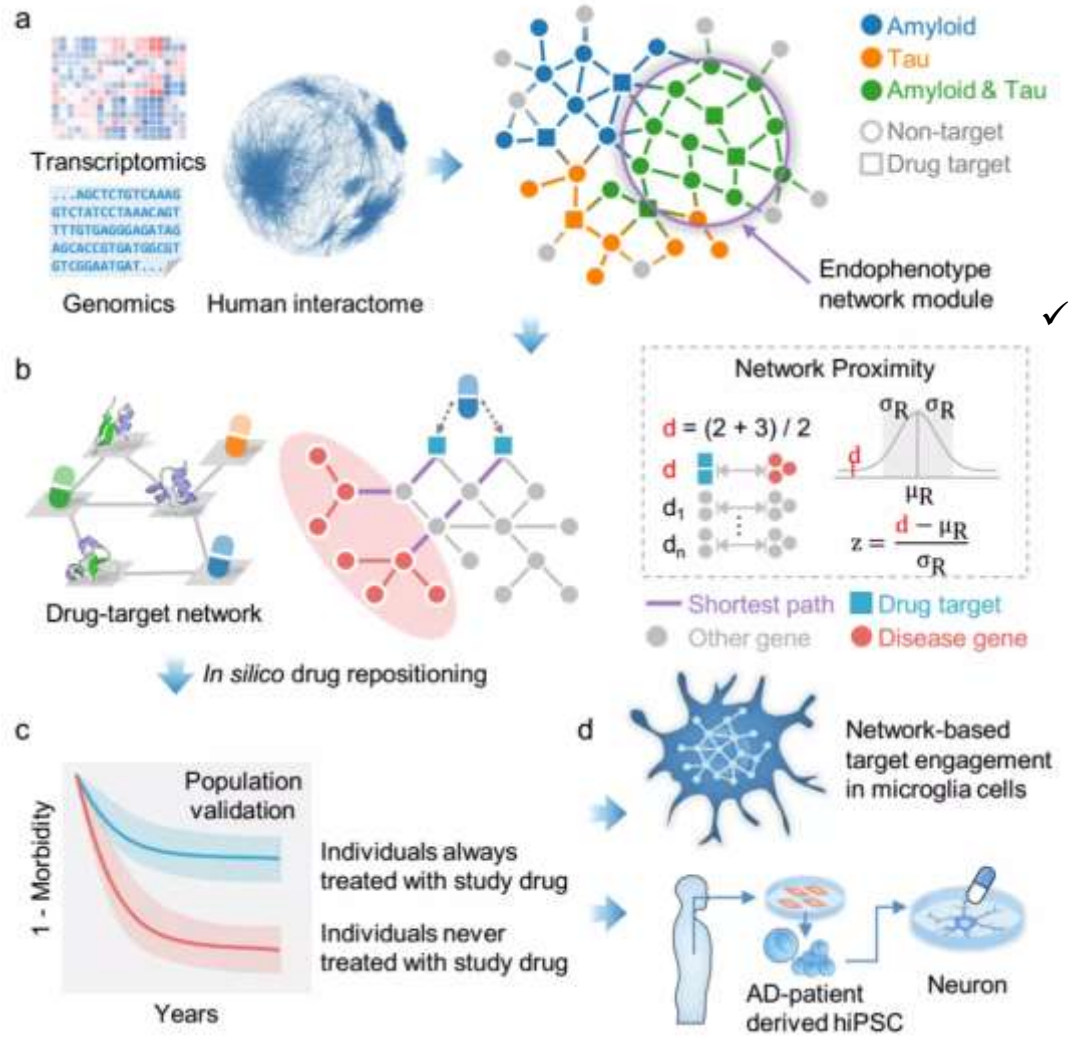
Article | Published: 06 December 2021

Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease

Jiansong Fang, Pengyue Zhang, Yadi Zhou, Chien-Wei Chiang, Juan Tan, Yuan Hou, Shaun Stauffer, Lang Li, Andrew A. Pieper, Jeffrey Cummings & Feixiong Cheng

基于内表型的计算机网络医学发现结合医疗病历数据挖掘确定西地那非为阿尔茨海默病候选药物

内容概述



- ✓ 作者们旨在运用先前提出的**内表型网络方法开展抗AD的药物重定位研究**，并结合大规模的临床电子病历数据挖掘评估候选药物与AD疾病风险之间的关联性，同时结合体外AD细胞模型研究其作用机理；

- ✓ 整合了AD遗传学数据和其他生物学数据，构建能够计算病理生物学特征的13个疾病“内表型模块”。团队将这些模块绘制成一个包含了35444个人类蛋白质互作的大型网络，随后计算了1600种FDA批准的药物种类的网络药物度，在阿尔茨默海中可以发现相关疾病模块与多种分子靶点进行物理相互作用。**西地那非是得分最高的药物之一，提示该药或能影响阿尔茨海默病。**