



Sequence motifs

 汇报人: Lilian



文献来源

nature biotechnology 影响因子: 34.714

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

[nature](#) > [nature biotechnology](#) > [analyses](#) > article

[Published: 27 July 2015](#)

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

[Babak Alipanahi](#), [Andrew DeLong](#), [Matthew T Weirauch](#) & [Brendan J Frey](#)

[Nature Biotechnology](#) **33**, 831–838 (2015) | [Cite this article](#)

166k Accesses | **1009** Citations | **268** Altmetric | [Metrics](#)



内容概述

介绍了一项在基因组领域中使用卷积神经网络的开创性工作。文中表明可以使用深度学习从实验数据中确定序列特异性，无论是在体外数据训练还是体内测试中，深度学习都胜过其他最新方法。

作者根据此设计了一个软件——DeepBind，是全自动的，每个实验可处理数百万个序列。



Introduction

- ✓ DNA 和 RNA 结合蛋白在基因调控中发挥核心作用，包括转录和选择性剪接。
- ✓ 了解 DNA 和 RNA 结合蛋白的序列特异性对于开发生物系统中的调节过程模型和识别致病变异至关重要。

面临的问题

➤ 数据形式多种多样

“Protein binding microarrays (PBMs) and RNAcompete assays provide a **specificity coefficient** for each probe sequence, whereas chromatin immunoprecipitation (ChIP)-seq provides a **ranked list of putatively bound sequences of varying length**, and HT-SELEX11 generates a **set of very high affinity sequences**. ”

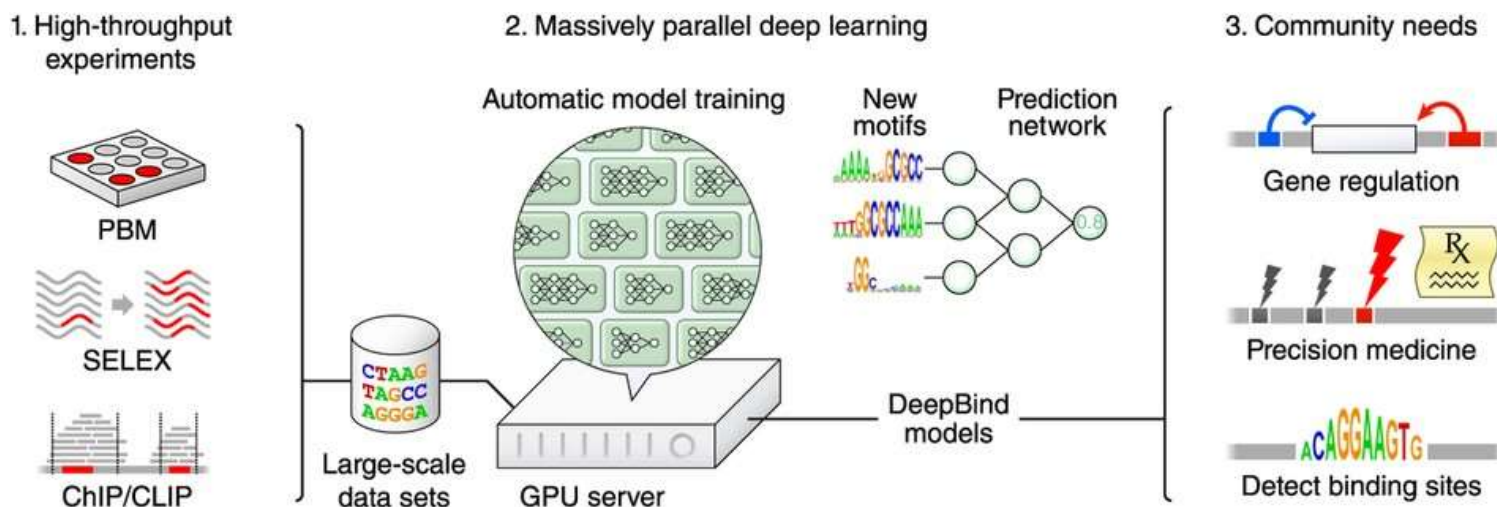
➤ 数据量大

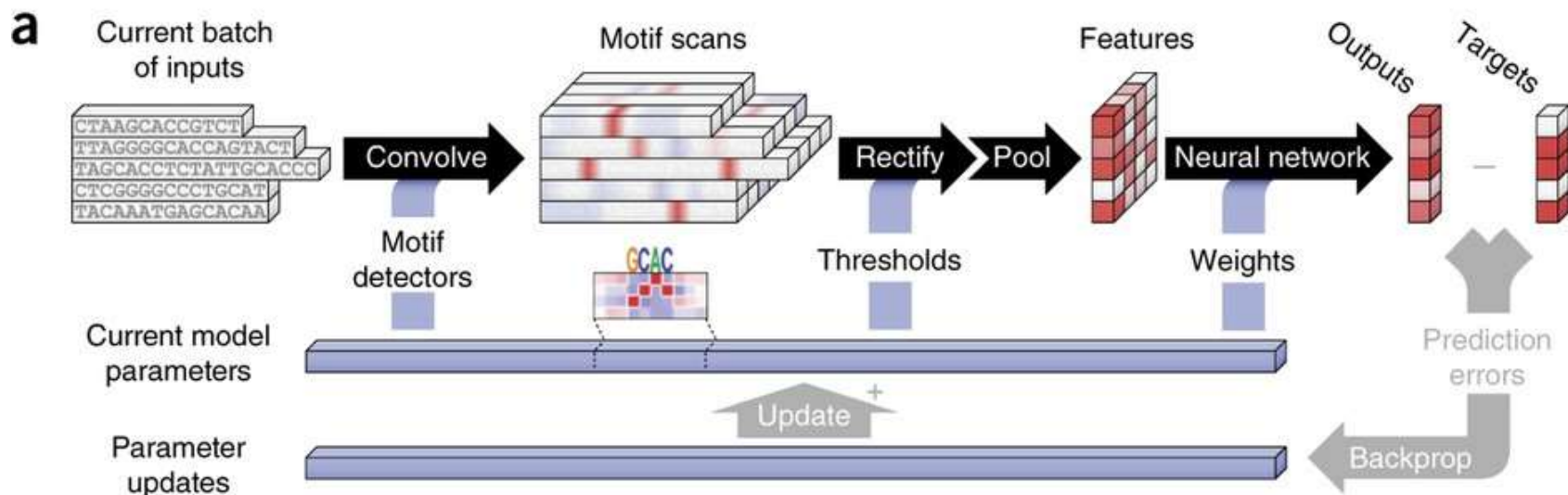
➤ 每种数据采集技术都有其自身的缺陷、偏差和局限性，需要发现相关的特异性。

“For example, ChIP-seq reads often localize to “hyper-ChIPable” regions of the genome near highly expressed genes.”

Introduction

- ✓ DeepBind可以解决以上问题，实现：
- (i) 它可以应用于微阵列和测序数据；
 - (ii) 它可以通过在图形处理单元 (GPU) 上并行实现从数百万个序列中学习；
 - (iii) 能很好地概括各种技术；
 - (iv) 它可以容忍中等程度的噪声和错误标记的训练数据；
 - (v) 它可以完全自动地训练预测模型，从而减少手动调整的需要。

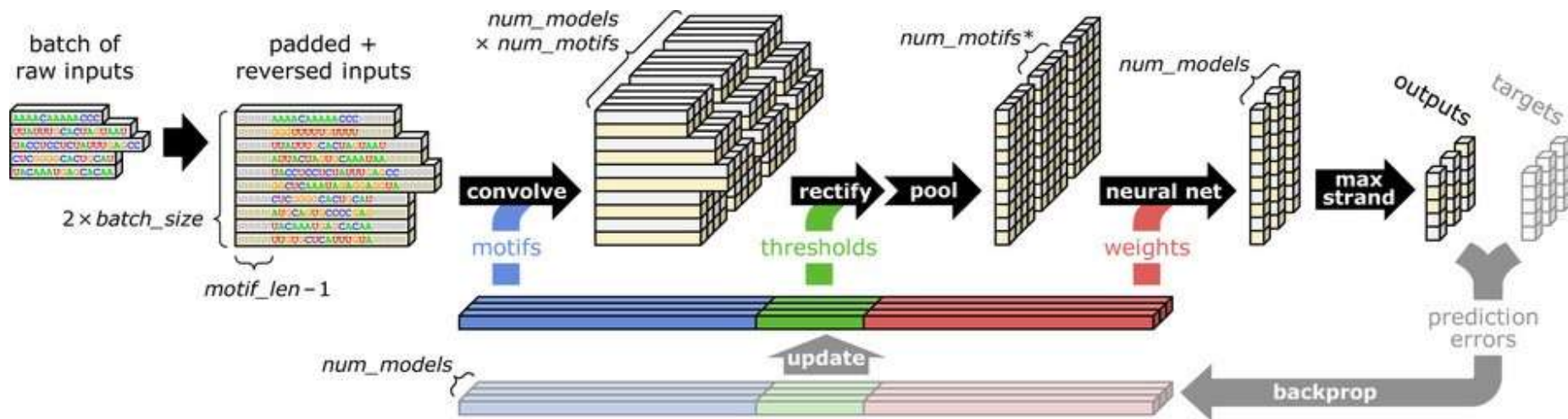




- ✓ 把基因组序列窗口当作一个图，与由具有三个颜色通道(R,G,B)像素组成的图像不同，DeepBind把基因组序列看作是由(A, C, G, T)或(A, C, G, U)四个通道组成的定长序列窗口。因此DNA蛋白结合位点预测问题就类似于图片二分类问题。
- ✓ 输入一条包含ATCG的序列，会返回一个标量值，值越高，说明序列是该种转录因子结合序列的可能性越高。
- ✓ 单个DeepBind模型并行处理五个独立序列，使用当前的模型参数来预测每个序列的单独分数。



Convolution.



Specifically, given a genomic sequence $s = (s_1, \dots, s_n)$ and a maximum motif detector length m , our convolution stage begins by converting s to a padded "one-of-four" representation, stored as an $(n + 2m - 2) \times 4$ array S in the obvious way:

$$S_{i,j} = \begin{cases} .25 & \text{if } s_{i-m+1} = N \text{ or } i < m \text{ or } i > n - m \\ 1 & \text{if } s_{i-m+1} = j^{\text{th}} \text{ base in (A, C, G, T)} \\ 0 & \text{otherwise} \end{cases}$$

Convolution.

For example, if $s = \text{ATGG}$ and motif detector length is $m = 3$ then the representation is

$$S = \begin{bmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{bmatrix}$$

The output of our convolution stage is an $(n + m - 1) \times d$ array X where d is the number of tunable motif detectors within the DeepBind model. Element $X_{i,k}$ is essentially the score of motif detector k aligned to position i of padded sequence S . The tunable motif detectors (all length m) are stored in an $d \times m \times 4$ array M where element $M_{k,j,l}$ is the coefficient of motif detector k at motif position j and base l . Specifically, the expression $X = \text{conv}_M(s)$ computes a discrete cross-correlation between padded sequence S and each detector M_k , where

$$X_{i,k} = \sum_{j=1}^m \sum_{l=1}^4 S_{i+j,l} M_{k,j,l}$$

Convolution.

For example, if $s = \text{ATGG}$ and motif detector length is $m = 3$ then the representation is

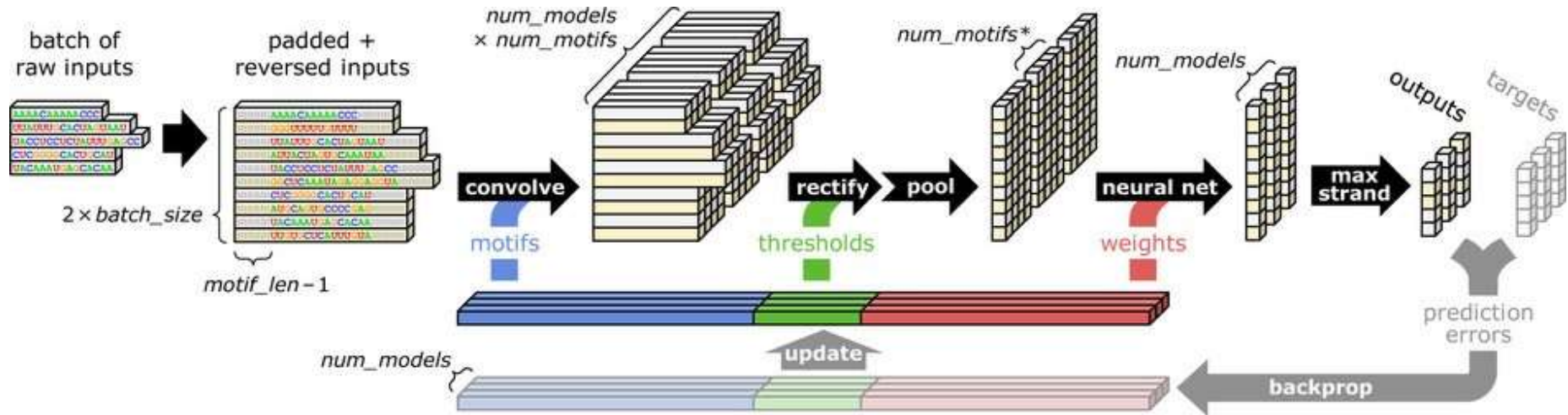
$$S = \begin{bmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{bmatrix}$$

The output of our convolution stage is an $(n + m - 1) \times d$ array X where d is the number of tunable motif detectors within the DeepBind model. Element $X_{i,k}$ is essentially the score of motif detector k aligned to position i of padded sequence S . The tunable motif detectors (all length m) are stored in an $d \times m \times 4$ array M where element $M_{k,j,l}$ is the coefficient of motif detector k at motif position j and base l . Specifically, the expression $X = \text{conv}_M(s)$ computes a discrete cross-correlation between padded sequence S and each detector M_k , where

$$X_{i,k} = \sum_{j=1}^m \sum_{l=1}^4 S_{i+j,l} M_{k,j,l}$$



Rectification.



rectified linear unit (ReLU) layer

$$Y_{i,k} = \max(0, X_{i,k} - b_k).$$

Pooling.

two choices for the pooling stage:

- max pooling——DNA-binding proteins

$$z_k = \max(Y_{1,k}, \dots, Y_{n,k}).$$

- max pooling and average pooling——RNA-binding protein (RBP模型)

$$z_{2k+0} = \max(Y_{1,k}, \dots, Y_{n,k})$$

$$z_{2k+1} = \text{avg}(Y_{1,k}, \dots, Y_{n,k})$$

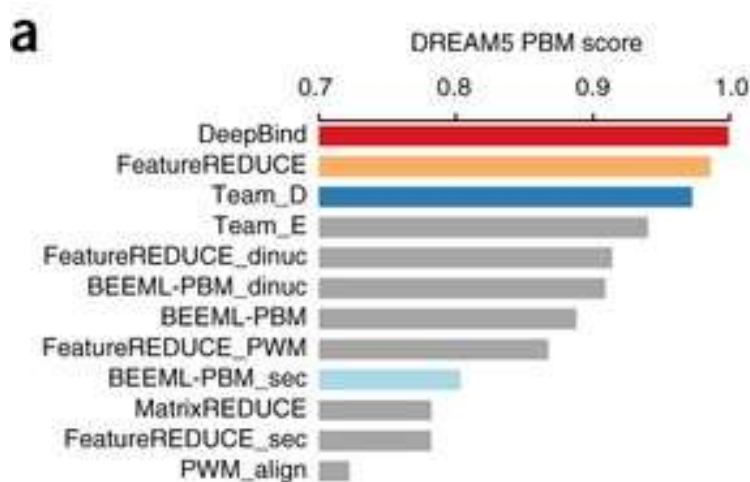
Neural network、Back-propagation stages

■ ■ ■ Ascertaining DNA sequence specificities

Data collection

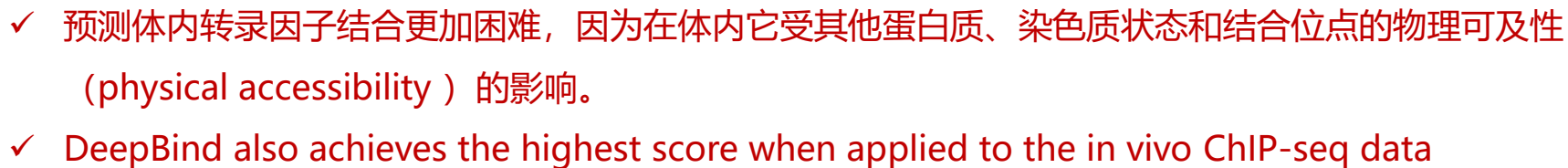
- 使用了来自 Weirauch 等人修订的 **DREAM5 TF-DNA Motif** Recognition Challenge 的 PBM 数据，包含86 种不同的小鼠转录因子；

Step 1: DREAM5 PBM training and evaluation



- ✓ Each algorithm's final score is the average of Pearson correlation-based score and AUC-based score.
- ✓ 值越靠近1，表现效果越好；

Step 2: DREAM5 ChIP-seq evaluation



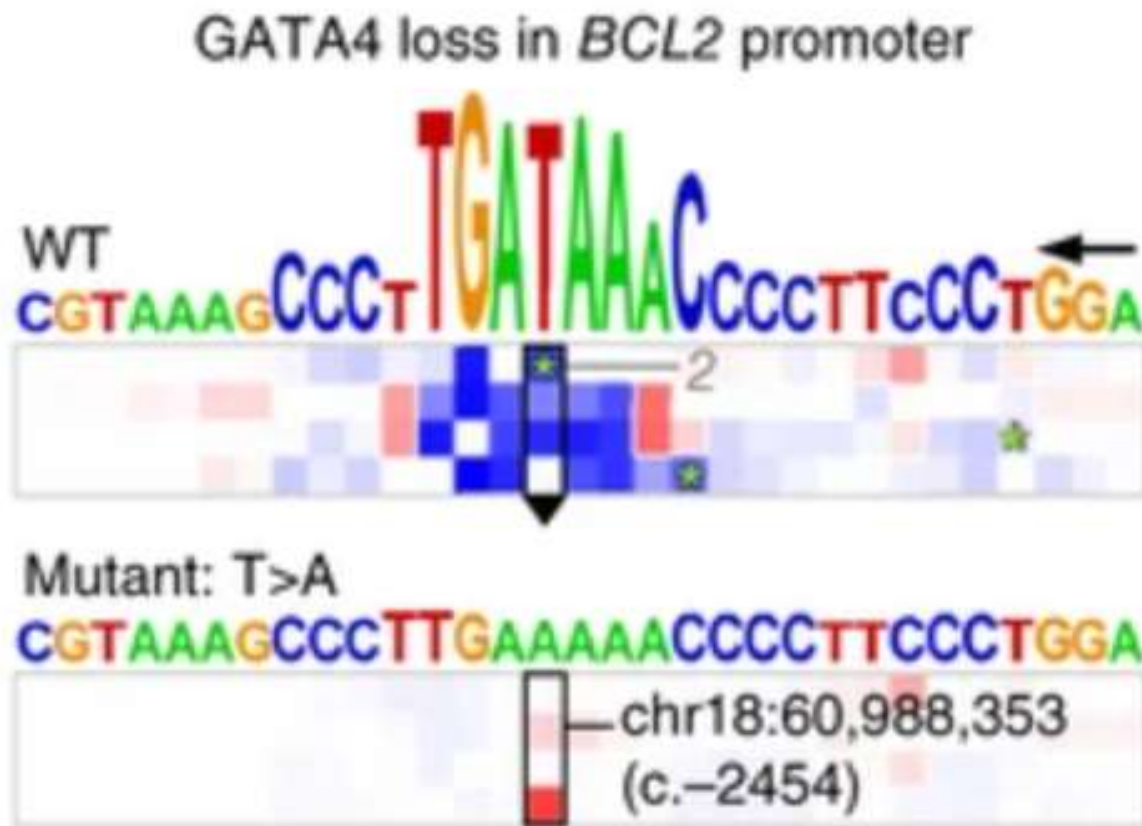
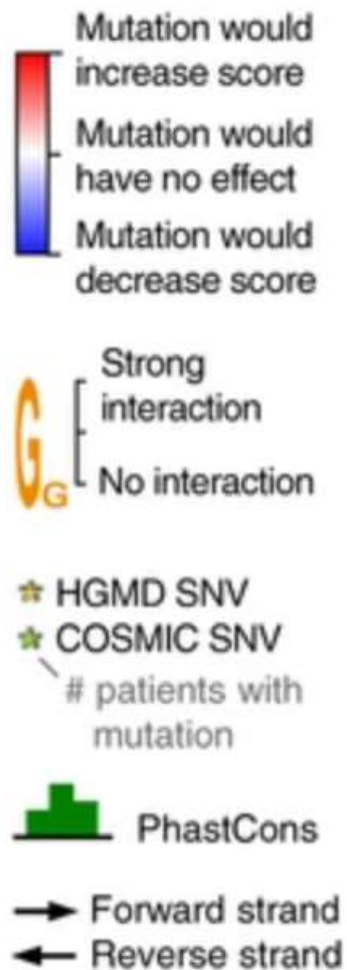


Identifying and visualizing damaging genetic variants (识别和可视化破坏性遗传变异)

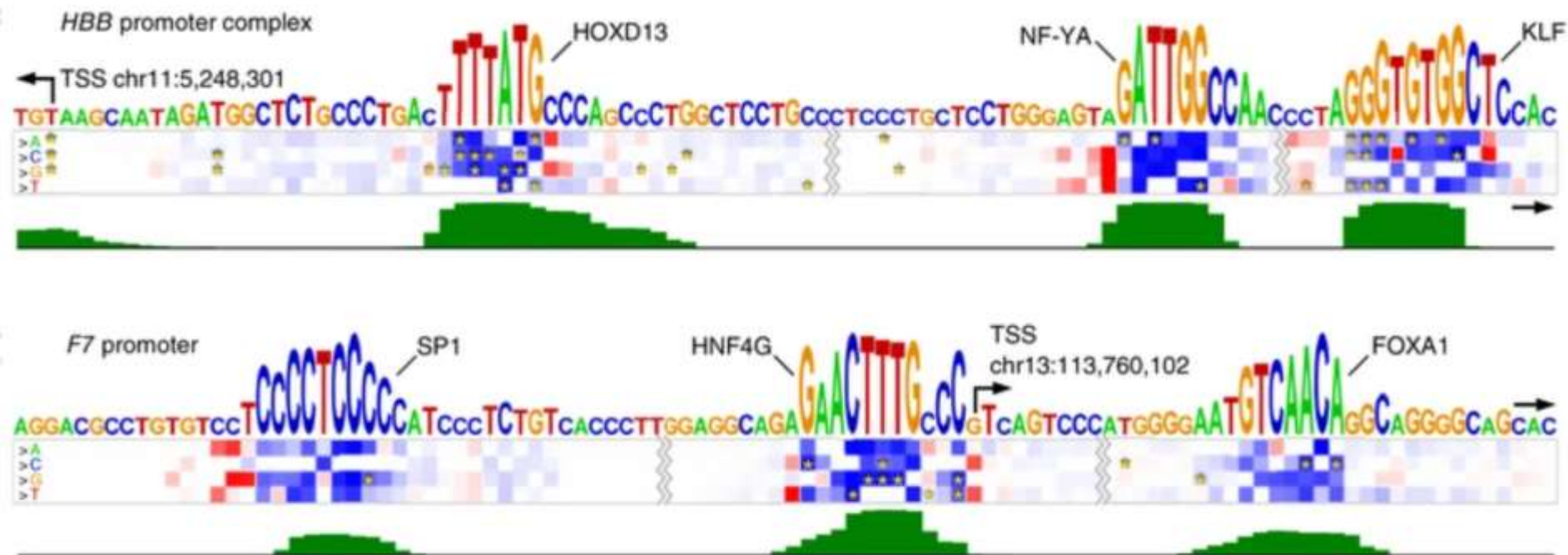
- ✓ 创建或废除结合位点的遗传变异可以改变基因表达模式并可能导致疾病;
- ✓ 精准医学的一个有前途的方向是使用结合模型来识别、分组和可视化可能改变蛋白质结合的变异;
- ✓ 为了利用DeepBind 探索遗传变异的影响, 作者开发了一种称为 “突变图 (mutation map) ” 的可视化方法;
- ✓ 使用近 600 个使用 ChIP-seq 和 HT-SELEX 数据训练的 DeepBind 模型检查了启动子内的变体。

突变图传达两种类型的信息:

- 对于给定的序列, 突变图通过碱基字母的高度显示每个碱基对 DeepBind 分析的重要性。
- 突变图包括一个大小为 $4 \times n$ 的热图, 其中 n 是序列长度, 表示每个可能的突变将增加或减少结合分数的程度。



- ✓ BCL-2启动子中丢失的GATA4结合位点,
- ✓ BCL-2与肿瘤的发生、发展、治疗等有着密切的关系,
- ✓ 可能在卵巢颗粒细胞肿瘤中起作用。



- ✓ HGMD SNV 破坏了 *HBB* 和 *F7* 启动子中的几个转录因子结合位点;
- ✓ *HBB* 基因上出现突变而导致的遗传性贫血症、同样的 *F7* 启动子与血友病的发生也有着密切的联系;
- ✓ 通过 Deepbind 检测出来的几个突变的转录因子位点分别可能导致 β -地中海贫血和血友病。

Conclusions

- ✓ 虽然没有统一的衡量序列特异性预测质量的指标，但我们发现 DeepBind 在各种数据集和评估指标方面超越了现有技术。
- ✓ 文章结果表明，在体外训练的 DeepBind 模型在对体内数据进行评分方面效果很好，这表明**能够捕捉核酸结合相互作用的真实特性；**
- ✓ DeepBind 可以很好地扩展到大型数据集，对于 ChIP-seq 和 HT-SELEX，作者发现可以**从其他技术因计算原因而丢弃的序列中学习有价值的信息。**



文献来源

nature

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

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

Article | [Open Access](#) | [Published: 22 September 2021](#)

Biologically informed deep neural network for prostate cancer discovery

[Haitham A. Elmarakeby](#), [Justin Hwang](#), [Rand Arafeh](#), [Jett Crowdis](#), [Sydney Gang](#), [David Liu](#), [Saud H. AlDubayan](#), [Keyan Salari](#), [Steven Kregel](#), [Camden Richter](#), [Taylor E. Arnoff](#), [Jihye Park](#), [William C. Hahn](#) & [Eliezer M. Van Allen](#)

[Nature](#) **598**, 348–352 (2021) | [Cite this article](#)

35k Accesses | **1** Citations | **198** Altmetric | [Metrics](#)

NATURE

期刊影响因子™

2020

49.962

五年

54.637

JCR 学科类别	类别排序	类别分区
MULTIDISCIPLINARY SCIENCES 其中 SCIE 版本	1/72	Q1



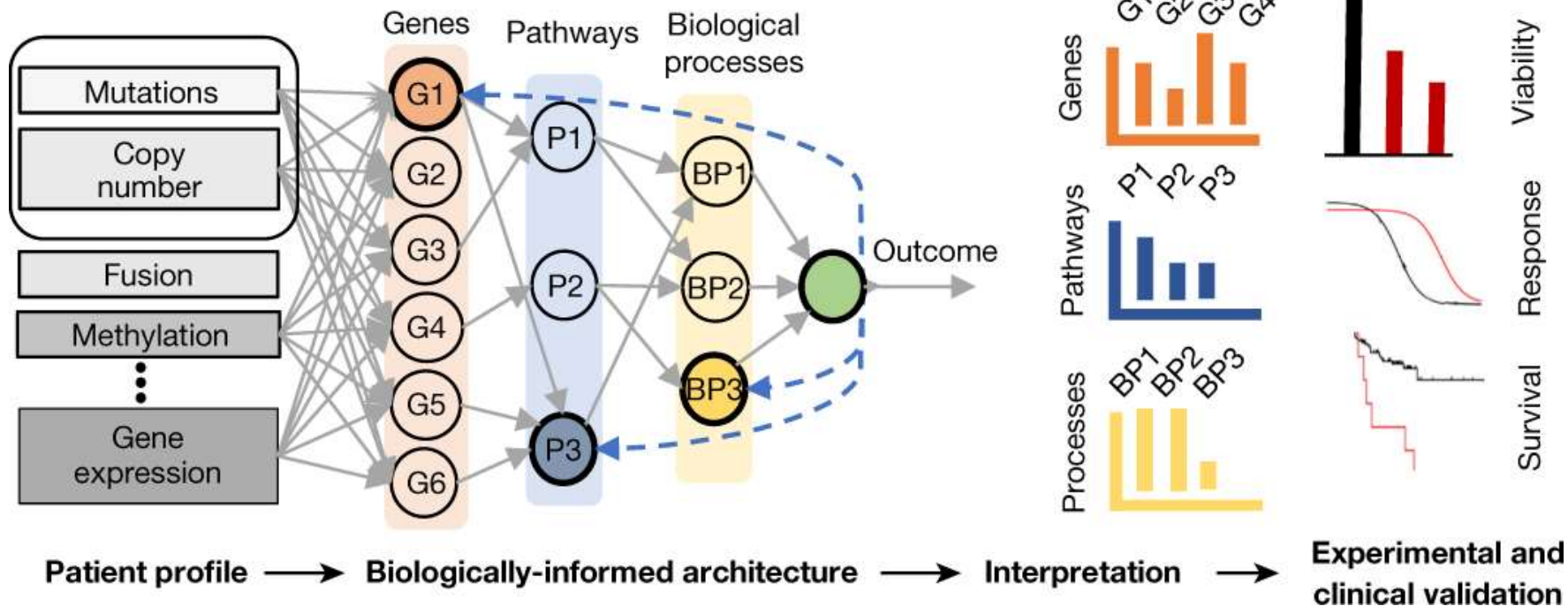
内容概述

- ✓ 在治疗癌症患者时，肿瘤学家想要预测患者的病程，以做出关键的治疗决定。
- ✓ 了解肿瘤独特的分子特征可以帮助指导这些决定，为癌症是生长缓慢还是具有侵袭性和致命性或者是会抵制治疗提供线索。
- ✓ 新的分子谱技术产生了大量关于肿瘤的信息，但医生们一直在努力将所有这些数据转化为有意义的预后。
- ✓ 美国布罗德研究所和丹娜-法伯癌症研究所的研究人员开发出一种新的模型，可以**区分致命的前列腺癌和那些不太可能导致症状或死亡的前列腺癌的基因组特征。**
- ✓ 它还可以帮助临床医生**预测前列腺癌患者的肿瘤是否会扩散到身体的其他部位，或者随着时间的推移变得对治疗变得更具抵抗性。**
- ✓ 这种称为P-NET的模型还能**识别可能与疾病进展有关的分子特征、基因和生物通路。**



文章设计了一个五层的神经网络，层与层间的链接并不是全连接，而是比全连接要少的稀疏链接，比全连接网络有更少的参数，从而能够加快训练速度。该稀疏模型有超过71000多个结点，1400百万个权重。

内容概述



作者利用这种方法将生物信息，如基因和代谢或信号通路之间的已知关系，直接编码到他们的模型中。然后，他们利用1000多名前列腺癌患者的基因组序列和体细胞（即非遗传性）突变等数据训练P-NET，以便预测肿瘤是否具有侵袭性。

当他们使用来自其他前列腺癌患者的数据测试他们的模型时，他们发现它能正确区分80%的转移性肿瘤和原发的进展较慢的肿瘤。这表明这种经过训练的模型能够对新数据执行相同的功能。



文献来源


nature medicine

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

[nature](#) > [nature medicine](#) > [letters](#) > [article](#)

Letter | [Published: 07 January 2019](#)

Identifying facial phenotypes of genetic disorders using deep learning

[Yaron Gurovich](#) , [Yair Hanani](#), [Omri Bar](#), [Guy Nadav](#), [Nicole Fleischer](#), [Dekel Gelbman](#), [Lina Basel-Salmon](#), [Peter M. Krawitz](#), [Susanne B. Kamphausen](#), [Martin Zenker](#), [Lynne M. Bird](#) & [Karen W. Gripp](#)

使用深度学习识别遗传疾病的面部表型

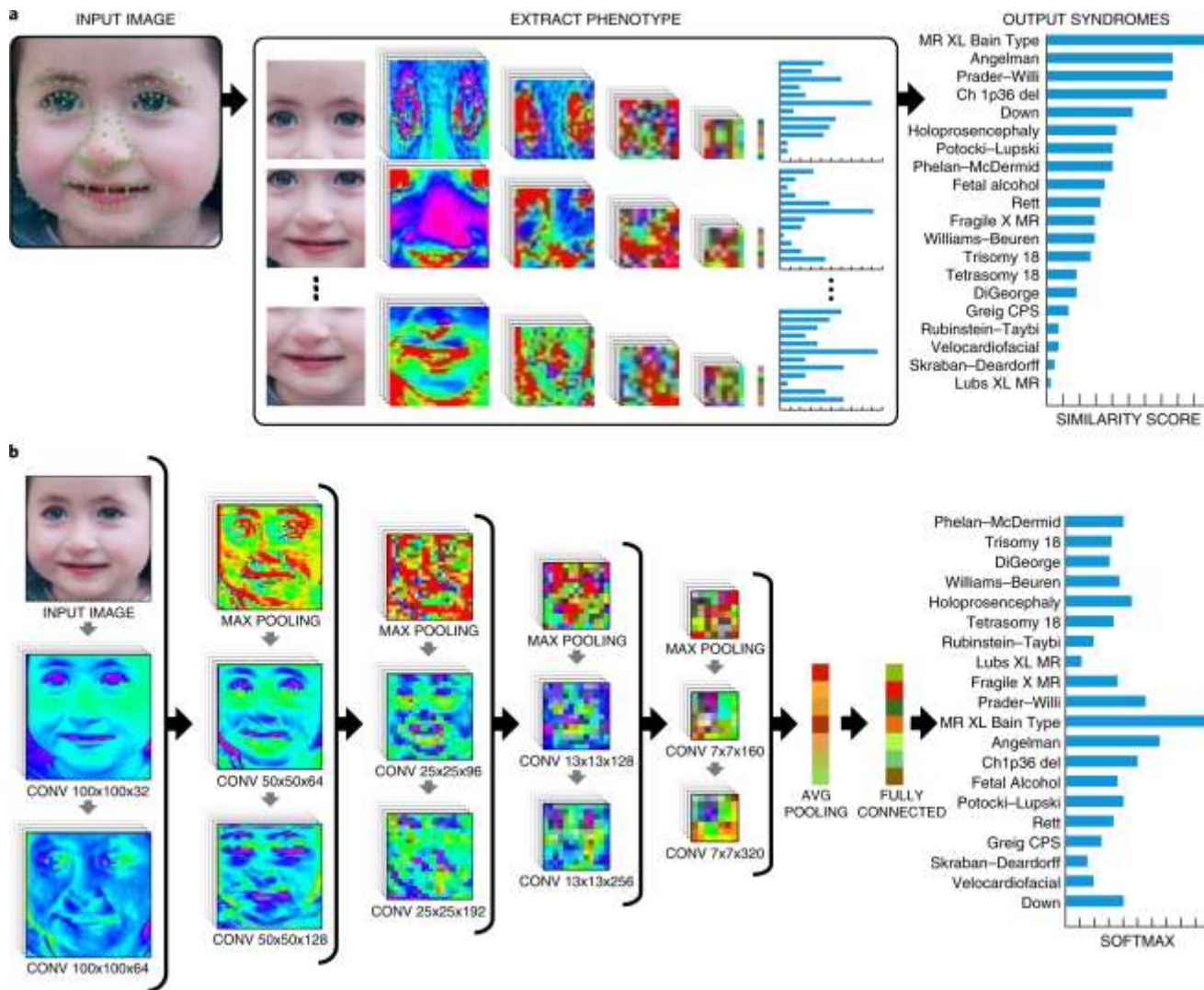


内容概述

- ✓ 综合遗传病影响 8% 的人口，许多综合征具有可识别的面部特征，这些特征对临床遗传学家具有重要意义。最近的研究表明，面部分析技术可达到专家及临床医生在综合征识别方面的能力；
- ✓ 美国数字医疗公司 FDNA 的研究人员提出了一种深度学习算法 DeepGestalt，可以帮助医生和研究人员通过分析人们的面部照片来发现罕见遗传病；
- ✓ **方法：DCNN**
- ✓ 在这篇论文中，研究者详细介绍了这项辅助诊断方法背后的技术——一个名为 Face2Gene 的智能手机 APP。该应用依靠深度学习算法和类脑神经网络来**区分人类照片中与先天性和神经发育障碍有关的独特面部特征**。利用从照片中推断出的模式，该模型可以定位到可能的诊断结果，并提供可能的选项列表。



内容概述



- ✓ 该网络由 10 个卷积层组成，除最后一层外，所有层后面都跟有批量归一化和整流线性单元 (ReLU)。
- ✓ 在每对卷积 (CONV) 层之后，应用一个池化层（前四对之后的最大池化和第五对之后的平均池化）。
- ✓ 研究者们给算法输入了涵盖 216 种不同综合征的 17000 多张确诊病例的图像。在用新面孔进行测试时，该 APP 的最佳诊断猜测准确率达到了 65%。如果考虑多个预测结果，则 Face2Gene 的 top-10 准确率可以达到约 90%。



文献来源

Science

[Current Issue](#)

[First release papers](#)

[Archive](#)

[About](#) ▼

[Submit manuscript](#)

HOME > SCIENCE > VOL. 298, NO. 5594 > NETWORK MOTIFS: SIMPLE BUILDING BLOCKS OF COMPLEX NETWORKS



Network Motifs: Simple Building Blocks of Complex Networks

R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII, AND , U. ALON [Authors Info & Affiliations](#)

SCIENCE • 25 Oct 2002 • Vol 298, Issue 5594 • pp. 824-827 • DOI: 10.1126/science.298.5594.824

↓ 761 3,776





内容概述

- ✓ 给出network motif的定义与解释

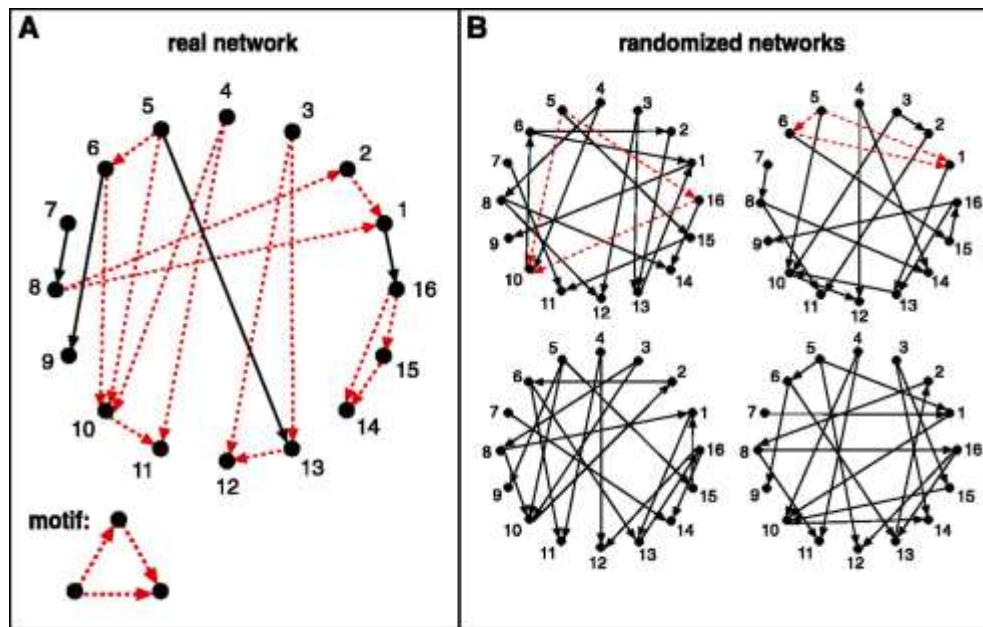
模体可以通俗地理解为网络中频繁出现的局部连接模式。模体不限于 3 个节点。4 个或 5 个节点组成的子网络结构，也可以看成是模体。

- ✓ 模体的普遍与跨学科性

- ✓ 如何检测模体的存在

- ✓ 了解常见的模体，找到复杂网络中的规律

- ✓ 5. 模体在复杂网络中的应用





文献来源

nature communications

Explore content ▾

About the journal ▾

Publish with us ▾

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

Open Access | [Published: 01 February 2016](#)

Network-based *in silico* drug efficacy screening

[Emre Guney](#), [Jörg Menche](#), [Marc Vidal](#) & [Albert-László Barábas](#) 

[Nature Communications](#) **7**, Article number: 10331 (2016) | [Cite this article](#)

26k Accesses | **186** Citations | **40** Altmetric | [Metrics](#)



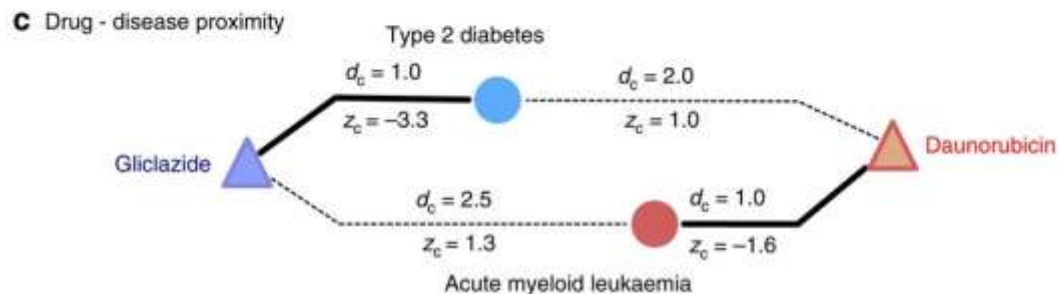
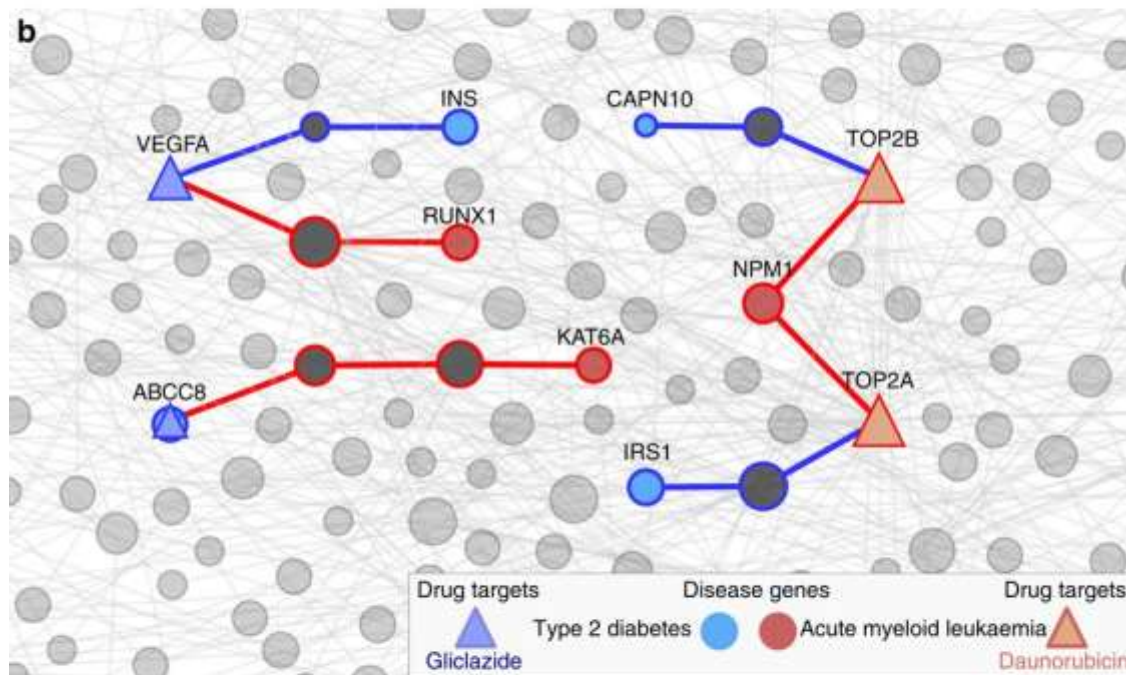
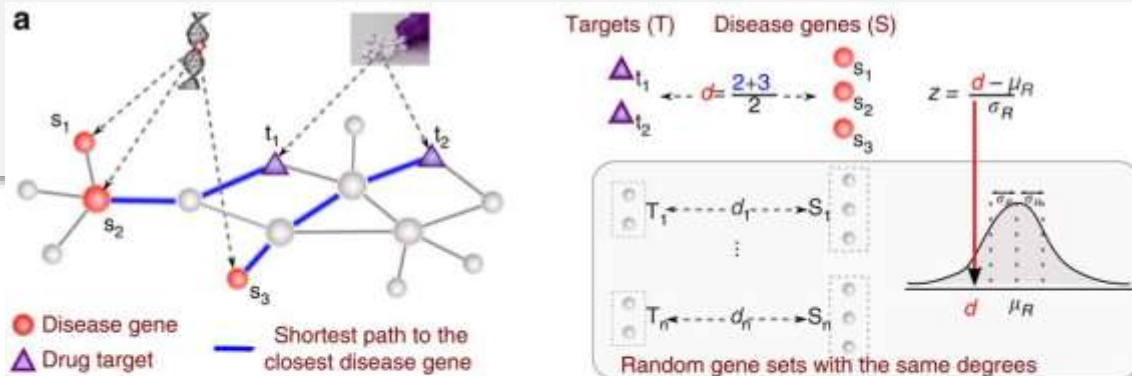
内容概述

- ✓ 大多数疾病的出现不能用单基因缺陷来解释，而是涉及不同基因组协调功能。因此，药物开发必须将重点从携带疾病相关突变的单个基因转移到基于网络的疾病机制视角。
- ✓ 这项研究中，引入了一个**无监督的基于网络的框架来分析药物与疾病之间的关系**。
- ✓ 为了研究药物靶标和疾病蛋白质之间的关系，作者开发了一种**相对接近度的度量方式**，用于基于网络的关系实现药物和疾病蛋白质（由与疾病相关的基因编码的蛋白质）之间关联的量化。



内容概述

作者提出的“接近度”的概念可作为治疗效果的一个很好的代表。





文献来源

nature medicine

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

[nature](#) > [nature medicine](#) > [perspectives](#) > [article](#)

Perspective | [Published: 07 January 2019](#)

A guide to deep learning in healthcare

[Andre Esteva](#) , [Alexandre Robicquet](#), [Bharath Ramsundar](#), [Volodymyr Kuleshov](#), [Mark DePristo](#), [Katherine Chou](#), [Claire Cui](#), [Greg Corrado](#), [Sebastian Thrun](#) & [Jeff Dean](#)

[Nature Medicine](#) **25**, 24–29 (2019) | [Cite this article](#)

52k Accesses | **696** Citations | **438** Altmetric | [Metrics](#)



内容概述

集中讨论了计算机视觉、自然语言处理、强化学习和通用方法中的深度学习；
本文描述了这些计算技术如何影响医学的几个关键领域，并探索如何构建端到端系统。

- ✓ **CNN** 可以接受各种医学图像的训练，包括放射学、病理学、皮肤病学和眼科。信息从左向右流动。CNN 获取输入图像并使用简单的操作（例如卷积、池化和完全连接层）将它们依次转换为扁平向量。输出向量（softmax 层）的元素代表疾病存在的概率。
- ✓ **自然语言处理 (NLP)** 侧重于分析文本和语音以从单词推断含义，在医疗保健领域，深度学习和语言技术为电子健康记录 (EHR) 等领域的应用提供动力。大型医疗机构的 EHR 可以捕获超过 1000 万患者在整个十年过程中的医疗交易。仅一次住院通常会产生约 150,000 条数据。从这些数据中获得的潜在好处是巨大的。
- ✓ **强化学习 (RL)** 是指一类旨在训练计算代理成功与其环境交互的技术，通常是为了实现特定目标。可以从深度强化学习中受益的一个医疗领域是机器人辅助手术 (RAS)。目前，RAS 在很大程度上依赖于外科医生以遥控方式引导机器人的器械。例如，计算机视觉技术（例如，用于物体检测/分割和立体视觉的 CNN）可以从图像数据重建开放伤口的景观，并且可以通过解决路径优化问题来生成缝合或打结的轨迹。



内容概述

✓ Generalized deep learning (广义深度学习)

现代基因组技术收集了各种各样的测量数据，从个人的 DNA 序列到血液中各种蛋白质的数量。深度学习有很多机会来改进用于分析这些测量的方法，这最终将帮助临床医生提供更准确的治疗和诊断。