讨论班汇报

图在生物化学领域中的应用

报告人：周玉晨

报告时间：2023.04.14

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

大纲

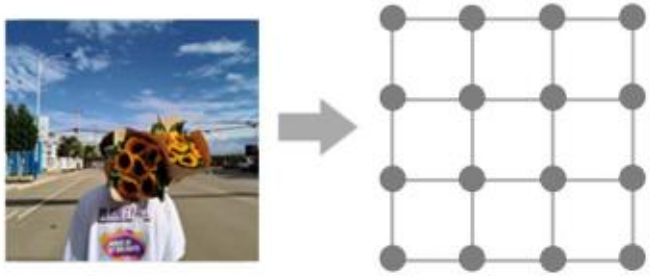
1. 背景概述

2. 前沿应用

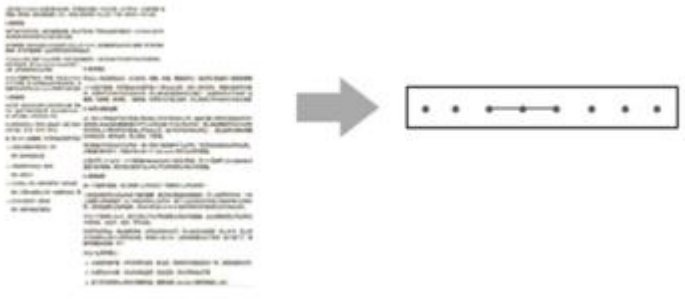
背景概述

图的优势

Image

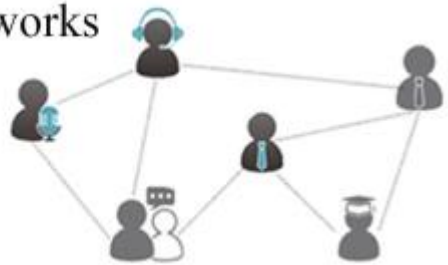


Text

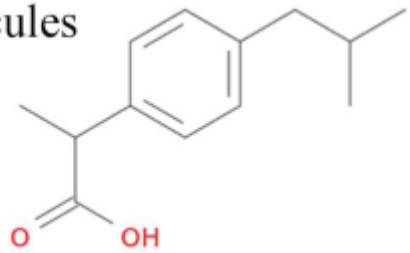


图像和文本都是欧式空间中的数据，样本之间服从独立同分布（iid）假设（规则化数据）

Social networks



Drug molecules

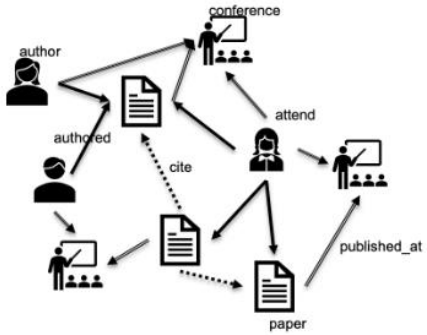


图是非欧空间中的数据，节点之间存在依赖关系，不服从独立同分布（iid）假设，节点的邻居数量不固定（非规则化数据），用于建模事物之间的多样化联系，考虑拓扑信息

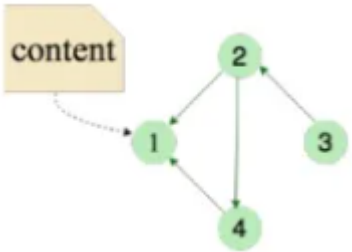
背景概述

图的类型

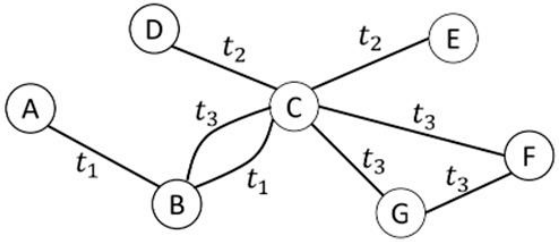
同质图



异质图



属性图

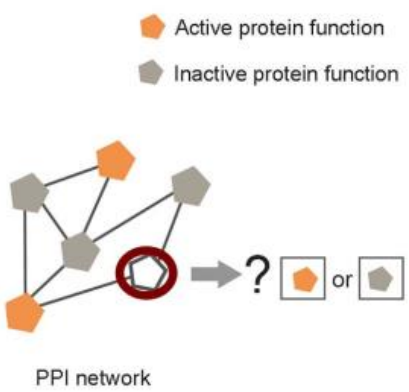


时序图

可以根据实际需求构建不同类型的图来满足不同场景下的建模需求

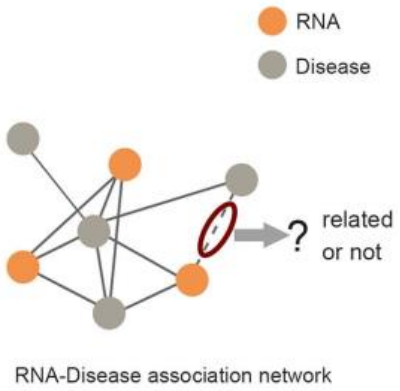
背景概述

常见图分析任务



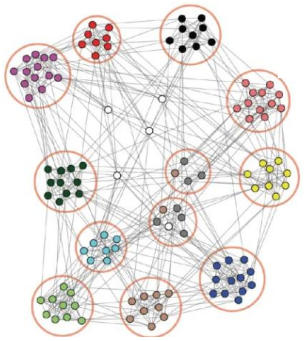
PPI network

节点分类



RNA-Disease association network

链接预测



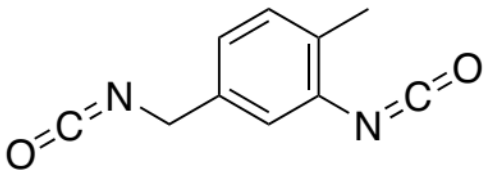
社区检测

图生成

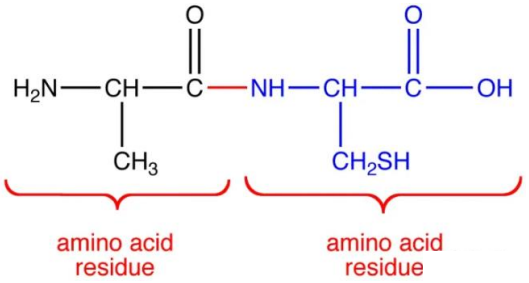
现实世界不同领域下所涉及到图数据的任务都可以转化成上述常见的图分析任务

背景概述

生物化学领域中的图数据

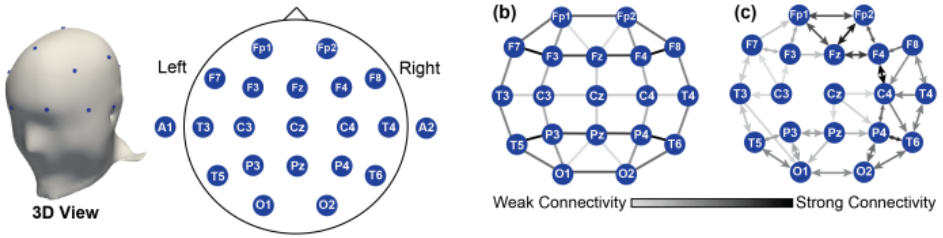


化学分子图

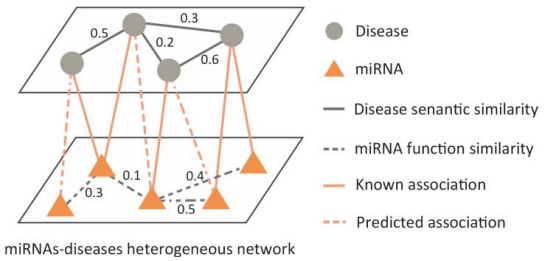


蛋白质结构图

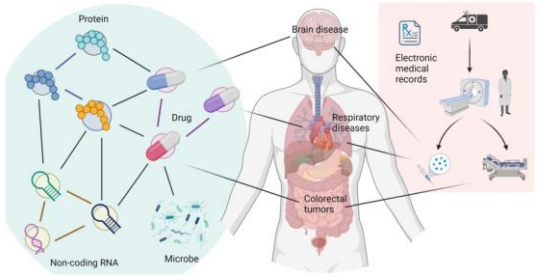
蛋白质相互作用图



脑结构图



疾病-RNA二分图



生物医学知识图谱

背景概述

生物化学领域中常见的图分析任务

任务级别	图分析任务	相关文献
节点级别任务	蛋白质功能预测 疾病蛋白质发现 RNA分类 医学图像分割	Graph Neural Networks for Predicting Protein Functions. CAMSAP 2019. Disease protein prediction with graph convolutional networks. Genetics. 2004. ncRNA Classification with Graph Convolutional Networks. Arxiv 2019. Structure-based function prediction using graph convolutional networks. Nat. Commun. 2019.
连边级别任务	蛋白质相互作用预测 药物相互作用预测 药物-疾病关系预测 疾病-miRNA关系预测 疾病-基因关系预测 药物反馈预测	Graph convolutional networks to explore drug and disease relationships in biological networks. Comput. Sci. 2017. Inferring disease-associated Piwi-interacting RNAs via graph attention networks. ICIC 2020. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics. 2020. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses . Nat. Commun. 2021.
图级别任务	分子图生成 蛋白质结构预测 蛋白质结构设计 药物设计 分子性质预测	Variational graph autoencoders for miRNA-disease association prediction. Methods. 2020. Graphvae: towards generation of small graphs using variational autoencoders. ICANN 2018. Junction tree variational autoencoder for molecular graph generation. ICML 2018. MolGAN: An implicit generative model for small molecular graphs. Arxiv 2018.

背景概述

国内外知名研究机构

国内

1. 清华大学
2. 北京大学
3. 浙江大学
4. 中国科学技术大学
5. 中山大学
6. 中国科学院计算所
7. 复旦大学
8. 湖南大学

……

国外

1. 宾夕法尼亚大学（美国）
2. 麦吉尔大学（澳大利亚）
3. 加利福尼亚大学（美国）
4. 澳洲国立大学（澳大利亚）
5. 新加坡国立大学（新加坡）
6. 麻省理工大学（美国）
7. 斯坦福大学（美国）
8. 浦项工科大学（韩国）
9. 剑桥大学（英国）
10. 南加州大学（美国）
11. 德克萨斯农工大学（美国）
12. 哈佛大学（美国）
13. 耶鲁大学（美国）
14. 约翰霍普金斯大学（美国）
15. 利兹大学（英国）
16. 奥本大学（美国）
17. 伊利诺伊大学（美国）
18. 谷歌DeepMind（美国）

……

前沿应用

单体型组装和病毒准种重构

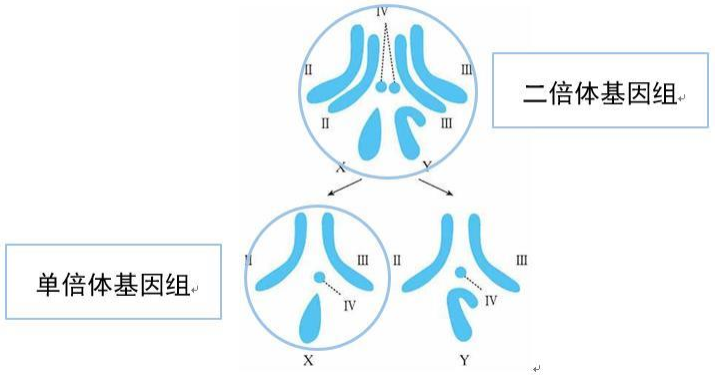
Graph Coloring via Neural Networks for Haplotype Assembly and Viral Quasispecies Reconstruction

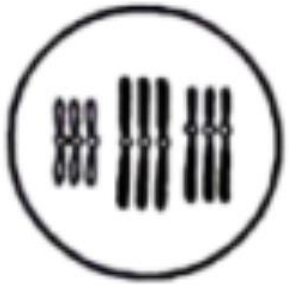
Hansheng Xue,[1] Vaibhav Rajan,[2] and Yu Lin[1]*
[1]School of Computing, Australian National University, Canberra, Australia
[2]School of Computing, National University of Singapore, Singapore
{hansheng.xue,yu.lin}@anu.edu.au, vaibhav.rajan@nus.edu.sg

(2022 NIPS)

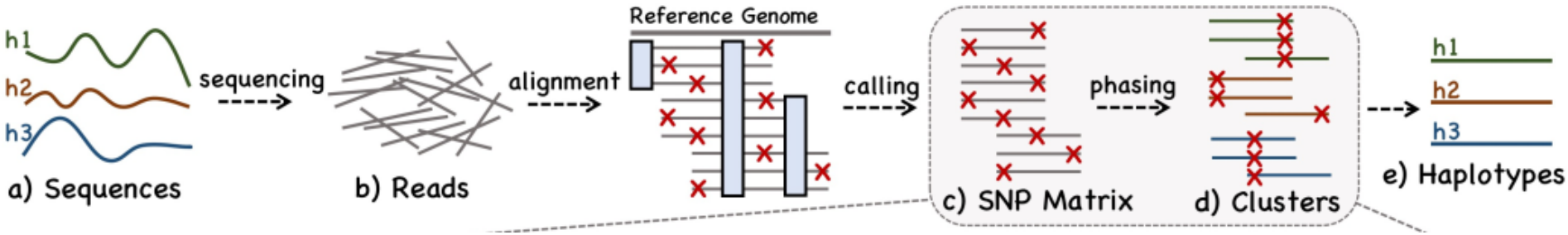
前沿应用

单体型组装和病毒准种重构





三倍体基因组

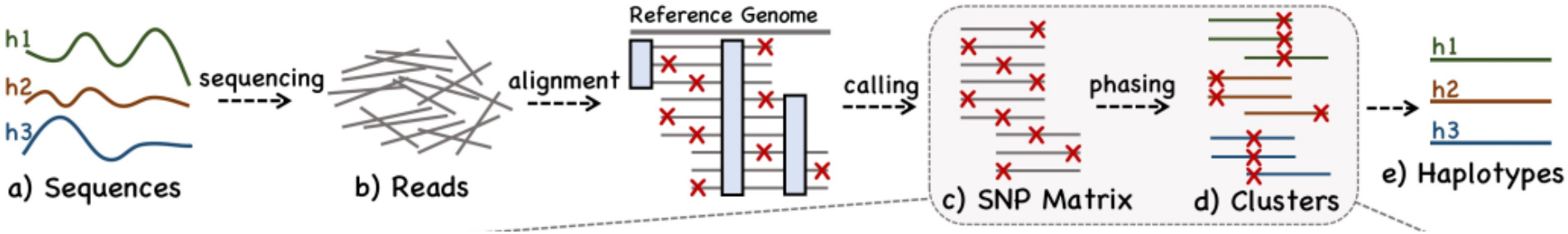


1. 理想状态下，属于同一单倍型下的reads应该保持一致，但是现实中由于sequencing错误的出现，会出现不一致
2. 利用最小错误纠正分数（MEC）来度量同一单倍型不同reads之间的差异
3. 优化MEC分数是一个NP-hard的问题

$$\mathrm{MEC}(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{R_j \in C_i} HD(\mathcal{H}_i, R_j)$$

前沿应用

单体型组装和病毒准种重构



现有方法的不足：

1）没有利用不同reads之间的隐式关系

2）现有方法分两阶段（编码和聚类），这造成结果的不稳定

3） SNP矩阵的稀疏性给现有方法带来了巨大的挑战

本文贡献：

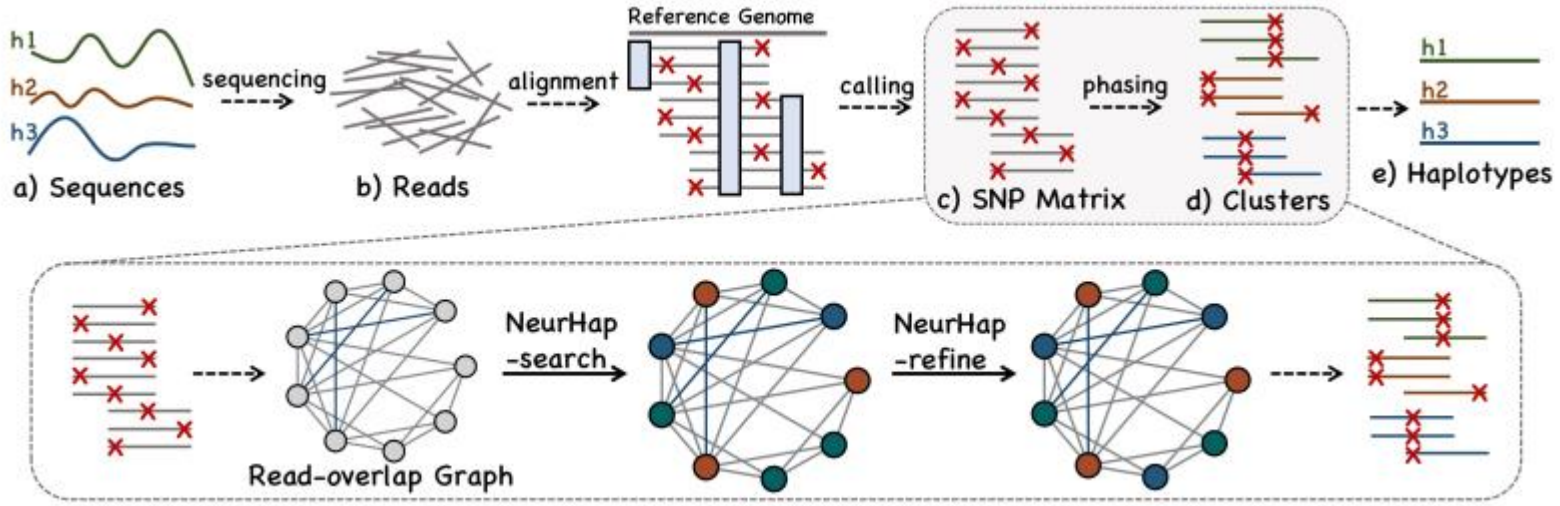
1）将单倍型解析问题转化为图染色问题，不同颜色代表不同的单倍型

2）作者提出了一个基于图表示学习的算法和联合优化策略来解决上述问题

3） 通过大量实验来验证所提NeuralHap方法的有效性

前沿应用

单体型组装和病毒准种重构



构建Read-overlap图：

1）一致：两个read至少在p个位置重合，且重合位置具有相同的等位基因

2）矛盾：两个read至少在q个重合位置具有不同的等位基因

3）模糊：没有充足的证据证明两个read一致或矛盾

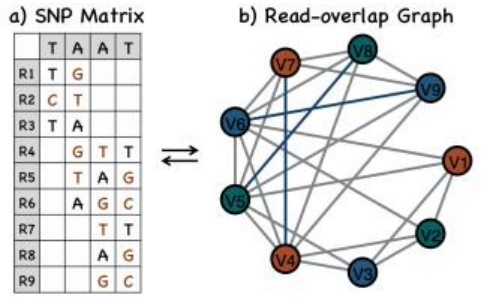
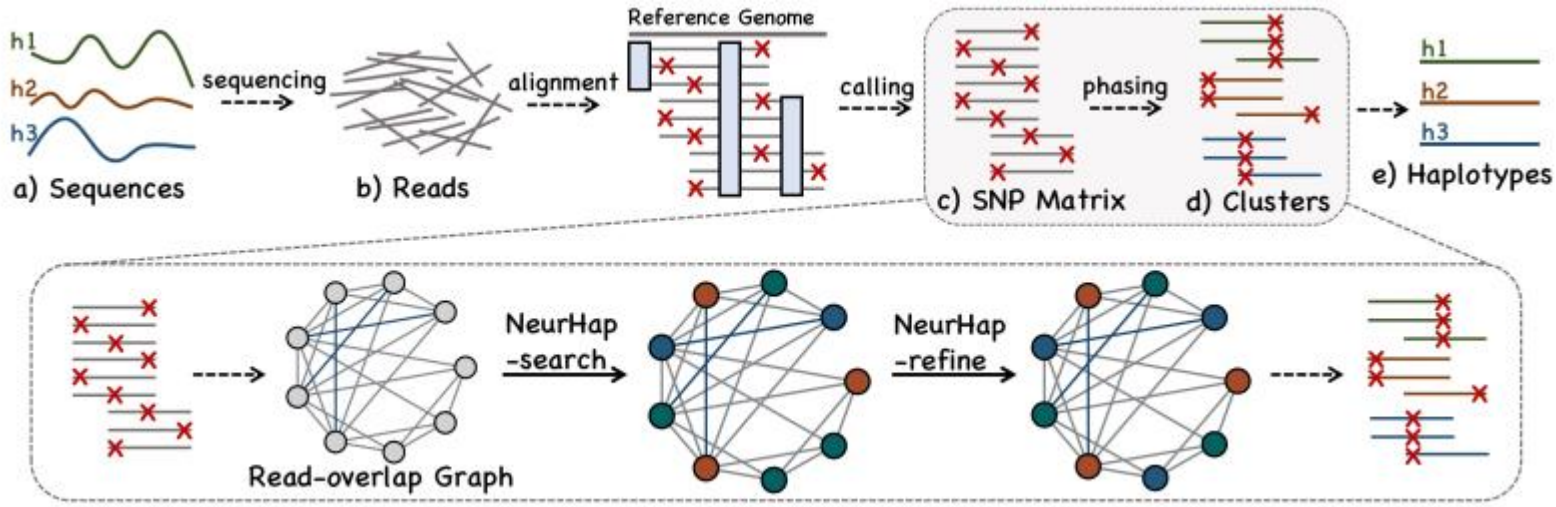


Figure 2: A toy example of constructing *read-overlap graph* with conflict edges (in grey) and consistent edge (in blue).

前沿应用

单体型组装和病毒准种重构



Read-overlap图上的图染色问题：

$$\min\ \mathrm{MEC}(c(v_1), c(v_2), \ldots, c(v_n)) = \min \sum_{i=1}^{k} \sum_{c(v_j)=i} \mathrm{HD}(\mathcal{H}_i, R_j)$$

$$\text{s.t.,} \begin{cases} \forall (v_i, v_j) \in \mathcal{E}_{\neq}, c(v_i) \neq c(v_j) \\ \forall (v_i, v_j) \in \mathcal{E}_{=}, c(v_i) = c(v_j) \end{cases}$$

$i, j \in \{1, \ldots, k\}$ K个类簇

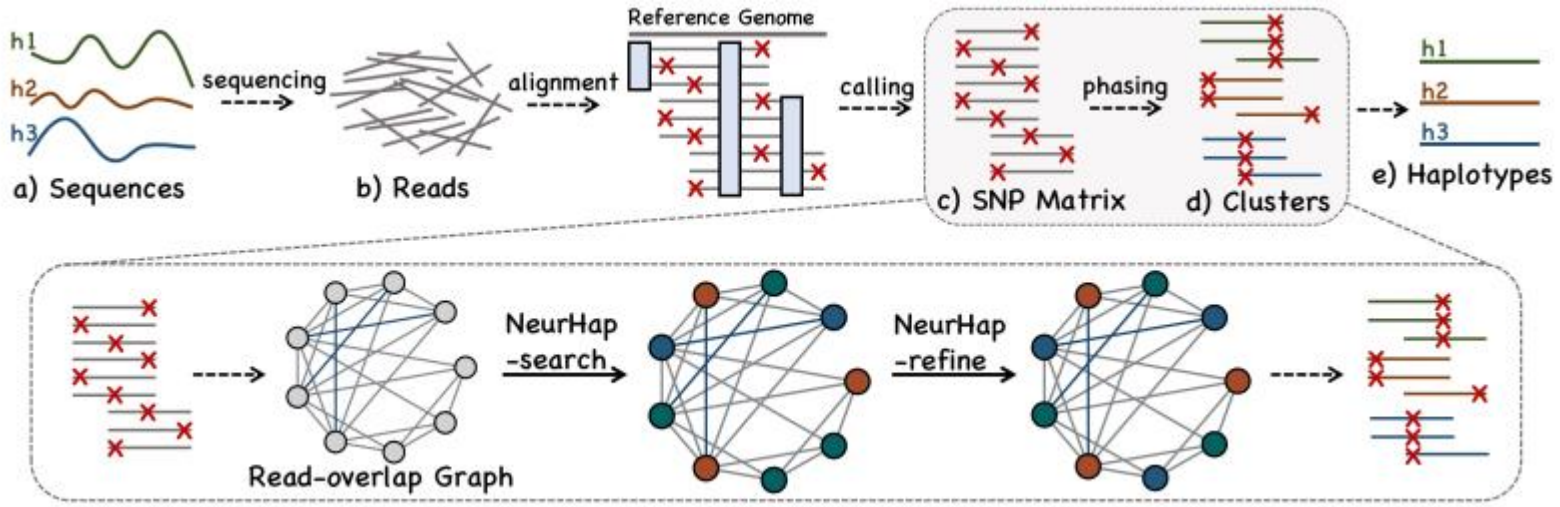
$\mathcal{M}_{\neq}(i,j) = 1 \text{ if } i \neq j$

$\mathcal{M}_{=}(i,j) = 1 \text{ if } i = j$

$$\mathcal{L} = -\frac{1}{|\mathcal{E}_{\neq}|} \sum_{(v_i,v_j)\in\mathcal{E}_{\neq}} \log(P(v_i)\mathcal{M}_{\neq}P(v_j)^{\mathsf{T}}) - \lambda \cdot \frac{1}{|\mathcal{E}_{=}|} \sum_{(v_i,v_j)\in\mathcal{E}_{=}} \log(P(v_i)\mathcal{M}_{=}P(v_j)^{\mathsf{T}})$$

前沿应用

单体型组装和病毒准种重构



全局搜索：仅使用了conflict edge

$$\mathrm{h}(v_i) = \mathrm{COMBINE}(\mathrm{m}(v_i), \mathrm{h}(v_i)),$$
$$\mathrm{m}(v_i) = \mathrm{AGGREGATE}(\{\mathrm{MESSAGE}(\mathrm{h}(v_i), \mathrm{h}(v_j)) : v_j \in N(v_i)\}).$$

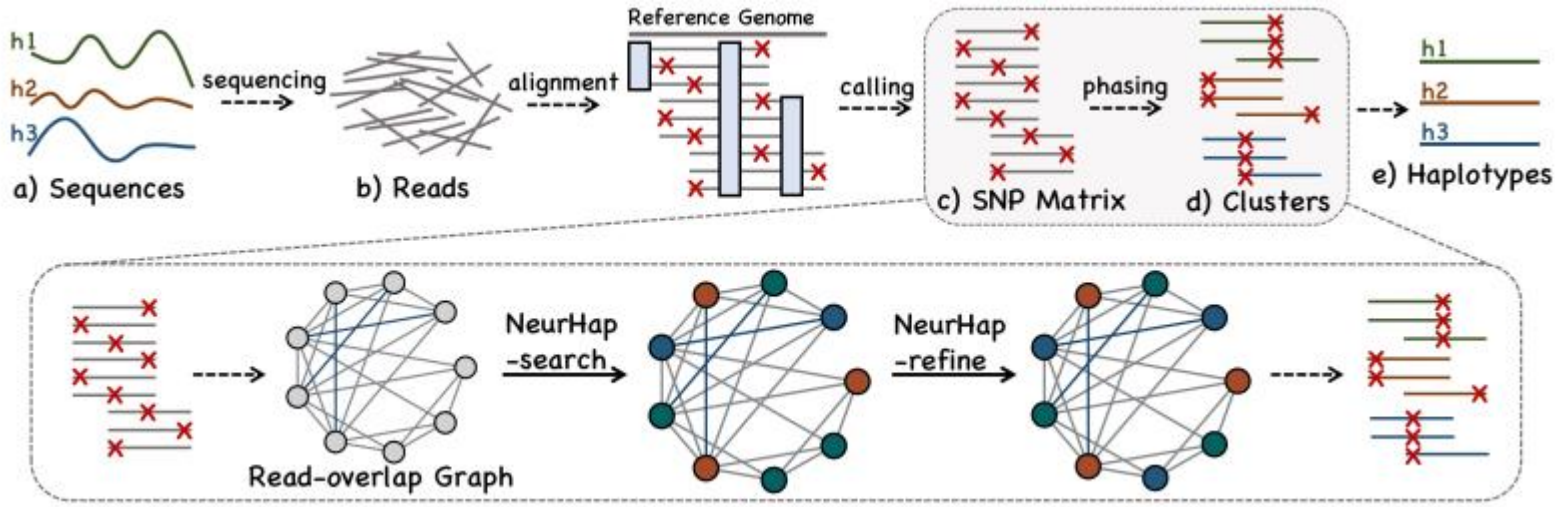
AGGREGATE：均值聚合器 $\mathrm{m}(v_i) = \frac{1}{|N(v_i)|}\sum_{v_j \in N(v_i)} \mathrm{h}(v_j)$

$$\mathrm{MESSAGE}(\mathrm{h}(v_i), \mathrm{h}(v_j)) = \mathrm{MLP}(\mathrm{h}(v_i)||\mathrm{h}(v_j))$$

$$\hat{P}(v_i) = \mathrm{DEC}(\mathrm{h}(v_i))$$

前沿应用

单体型组装和病毒准种重构



局部细化：由于没有优化MEC分数，上述过程可能出现多个满足条件的染色方案，因此定义了局部细化步骤来找到最可能的染色方案

具体来说，如果可以在不违反与相邻顶点的任何相关冲突约束的情况下为单个顶点分配与其当前颜色不同的颜色，并且通过改变获得了更低的MEC分数，则颜色被改变。

前沿应用

单体型组装和病毒准种重构

Table 1: Performance comparison on Sim-Potato data.

Model	#Cov 5X	#Cov 10X	#Cov 20X	#Cov 30X
H-PoP	429.0±64.1	933.9±103.6	1782.2±161.8	2826.9±180.7
AltHap	610.9±259.3	722.3±179.1	649.3±369.4	1148.2±509.9
GAEseq	153.7±20.3	261.6±58.7	372.8±74.5	496.9±128.7
CAECseq	96.2±26.9	141.4±40.7	254.2±99.7	372.9±148.9
NeurHap	**29.9**±5.7	**51.9**±8.2	**92.6**±10.6	**142.0**±23.6

Table 2: Performance comparison on Real-Potato data.

Sample	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Avg.
Reads	240	389	274	115	141	398	295	284	489	449	-
SNPs	294	238	83	23	176	198	456	424	236	410	-
H-PoP	705	525	132	4	240	982	981	766	793	1413	654.1±435.6
AltHap	746	572	192	9	299	1295	1021	982	811	1311	723.8±451.1
GAEseq	231	406	97	2	180	873	558	441	592	712	409.2±266.6
CAECseq	229	393	103	**1**	172	859	522	430	593	698	400.0±260.9
NeurHap	**178**	**343**	**93**	**1**	**163**	**857**	**499**	**384**	**561**	**632**	**371.6**±268.9

前沿应用

蛋白质结构预测

Research Track Paper

KDD '21, August 14–18, 2021, Virtual Event, Singapore

Geometric Graph Representation Learning on Protein Structure Prediction

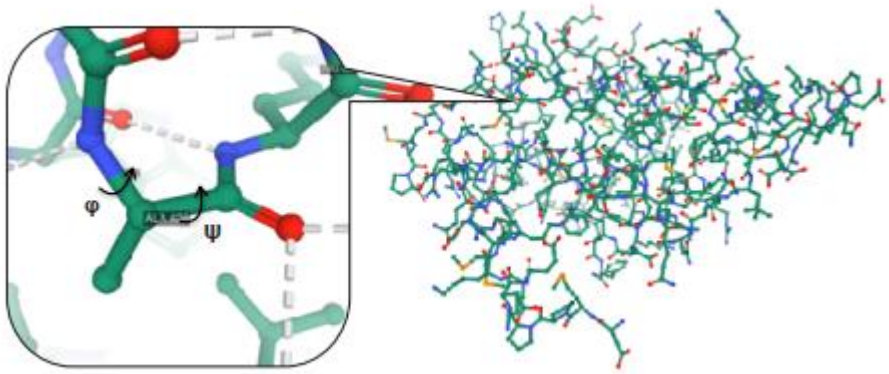
Tian Xia
tianxia@auburn.edu
Auburn University
Auburn, AL, USA

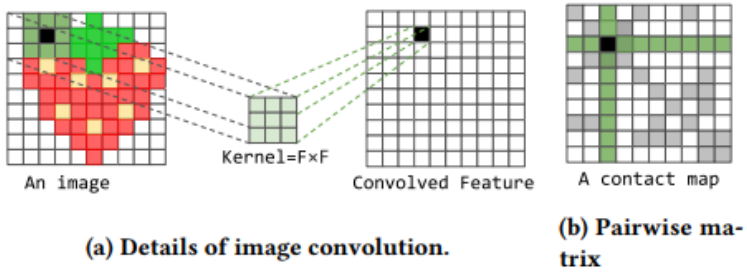
Wei-Shinn Ku
weishinn@auburn.edu
Auburn University
Auburn, AL, USA

(2021 KDD)

前沿应用

蛋白质结构预测





(a) Details of image convolution.

(b) Pairwise matrix

任务目标：根据氨基酸序列预测蛋白质3D结构

任务挑战：

1）蛋白质距离矩阵无法提供充足的结构信息来满足蛋白质结构建模的需求，需要二面角信息（距离信息+旋转信息）

2）捕获蛋白质结构中的非局部关系是很具挑战性的，CNN仅能捕获局部关系

3）氨基酸序列长度的变化以及过大的蛋白质结构规模很难进行处理，传统基于CNN的方法采用padding和cropping

本文贡献：

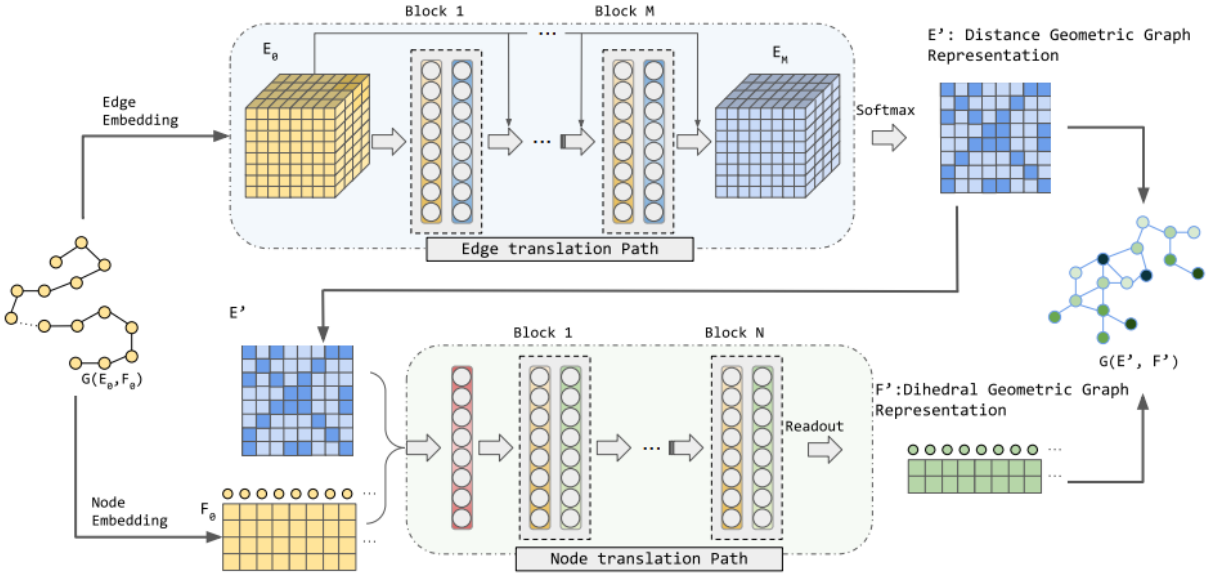
1）将蛋白质结构建模问题转换成3D图表示问题，其中节点为残基，边为残基间的点对信息

2）提出一个新颖的框架来生成蛋白质3D结构图，不仅包含距离信息还包含旋转角度信息

3）提出一个新颖的几何图卷积来处理远距离关系，并且能够处理变长的氨基酸序列和规模庞大的蛋白质结构图

前沿应用

蛋白质结构预测



模型输入：链式图 $G(\mathcal{V}, \mathcal{E}, E, F)$，E是边属性矩阵（co-evolution information, distance potential, and inter-residue coupling score），F是节点属性矩阵（position-specific scoring matrix, predicted secondary structure, solvent accessibility）

模型输出：几何图 $G(\mathcal{V}', \mathcal{E}', E', F')$ ，节点属性 F' 的前两维包含了旋转角度信息

编码链式图，生成几何图的节点表示

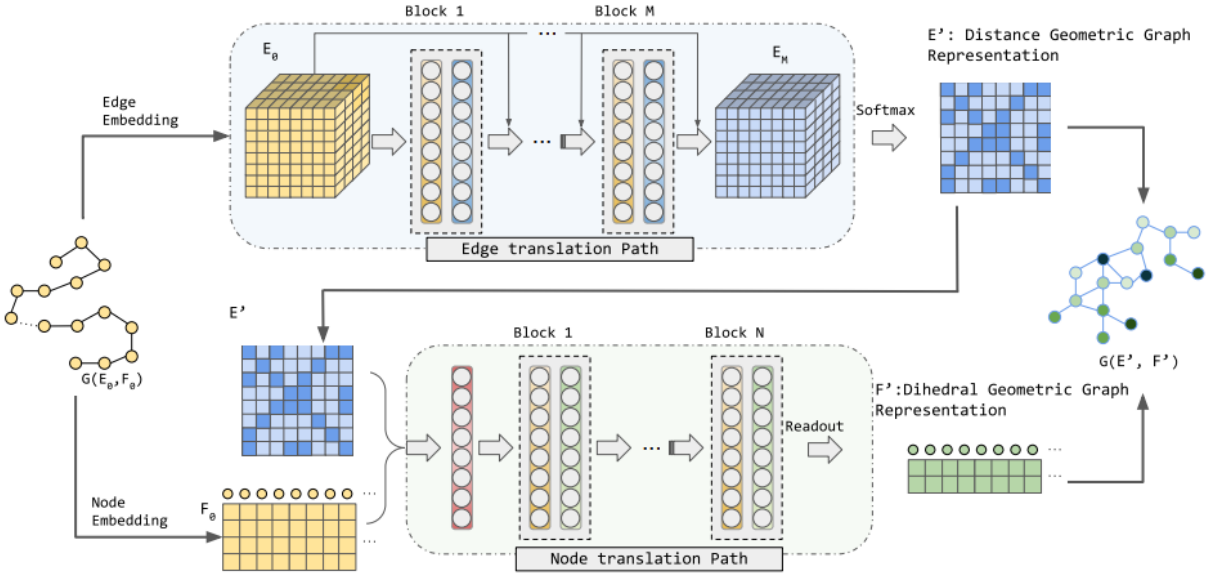
$$G(\mathcal{V}, \mathcal{E}, E, F) \rightarrow G(E').$$

模型包括节点转换路径和边转换路径分别预测几何图中节点和边的属性信息

$$G(\mathcal{V}, \mathcal{E}, E, F) \rightarrow G(F').$$

前沿应用

蛋白质结构预测



目标函数：

1）边属性预测 $L(E,E') = -\frac{1}{L\times L}\sum_{i=1}^{L}\sum_{j=1}^{L} y_{ij}\times\log\hat{y}_{ij}$ $\hat{y}_{ij}$ 残基i与残基j之间预测的距离label

2）节点属性预测 $L(F,F') = L(F_1,F_1') + L(F_2,F_2')$

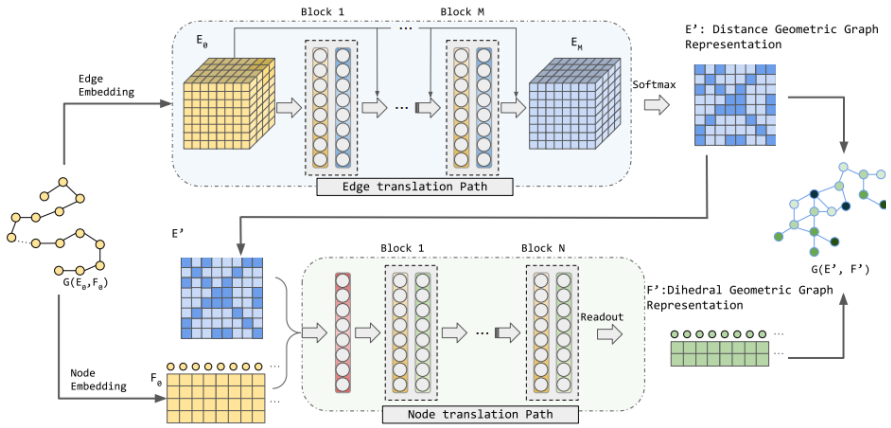
$$= \frac{1}{L}\sum_{i=1}^{L}\left(\sin(\phi_i)-\sin(\hat{\phi}_i)\right)^2 + \frac{1}{L}\sum_{i=1}^{L}\left(\cos(\phi_i)-\cos(\hat{\phi}_i)\right)^2$$

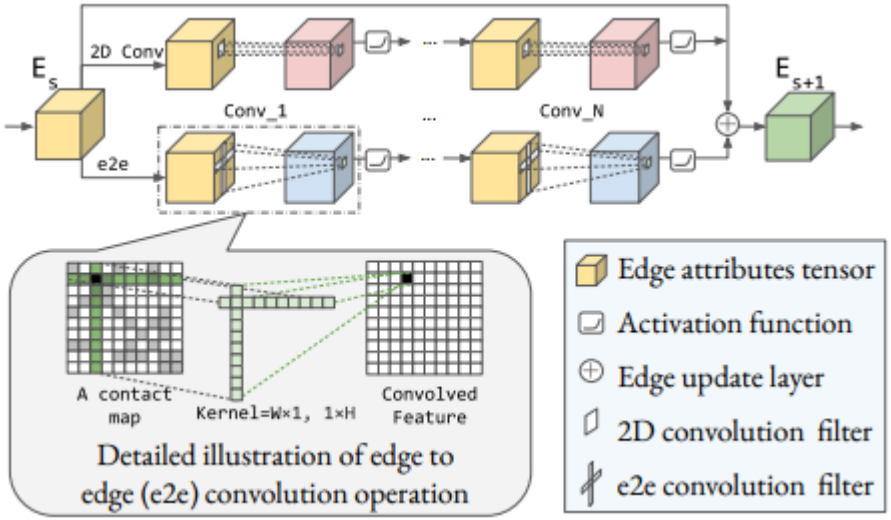
$$+ \frac{1}{L}\sum_{i=1}^{L}\left(\sin(\psi_i)-\sin(\hat{\psi}_i)\right)^2 + \frac{1}{L}\sum_{i=1}^{L}\left(\cos(\psi_i)-\cos(\hat{\psi}_i)\right)^2$$

$$L = L(E,E') + \lambda\times L(F,F')$$

前沿应用

蛋白质结构预测





边转换路径：捕获局部和远距离点对关系

1）边-边卷积：远距离关系

2）2D卷积：局部关系

3）边更新：带有残差连接和softmax

节点转换路径：利用节点属性和边转换路径输出的边属性信息来获取几何图中节点属性信息（旋转角度）

1）消息传递：$M_t(h_v^t, h_w^t, e_{vw}) = A_{e_{vw}} h_w^t$

2）节点更新：$h_i^{t+1} = GRU(h_i^t, m_i^{t+1})$

3）Readout：$\hat{v} = R(\{h_v^T | v \in G\})$

$v_i = (v_{a,i}, v_{b,i}, v_{c,i}, v_{d,i})$ $\sin\phi_i, \cos\phi_i, \sin\psi_i, \cos\psi_i$

$F_i = (F_{1,i}, F_{2,i})$ $\hat{F}_i = (arctan(\hat{v}_{a,i}/\hat{v}_{b,i}), arctan(\hat{v}_{c,i}/\hat{v}_{d,i}))$

前沿应用

蛋白质结构预测

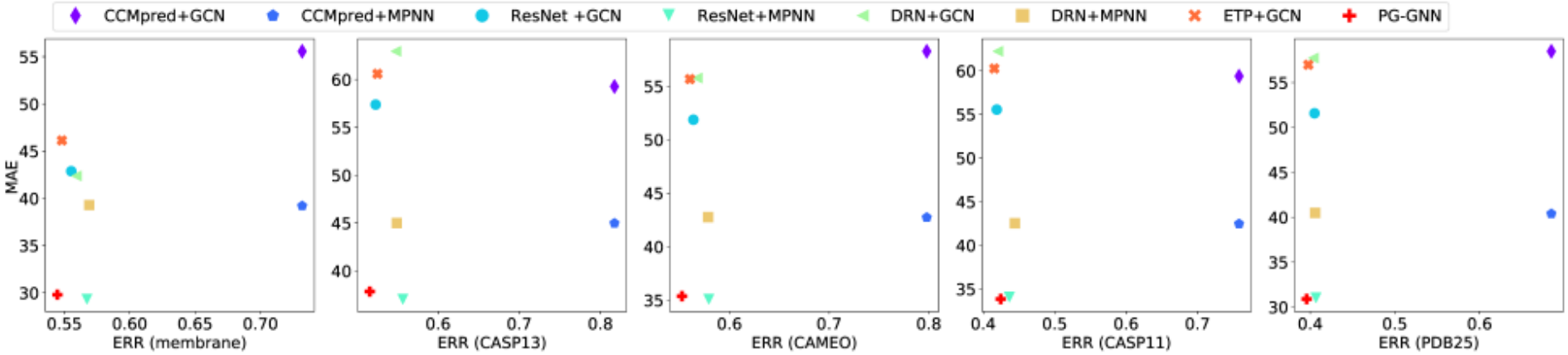
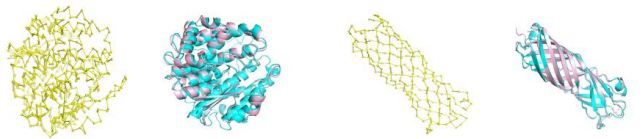
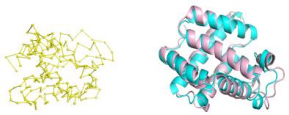


Figure 5: Overall performance comparison

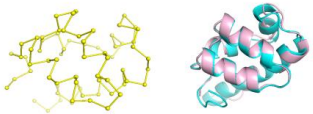


(a) 2c4j chain A

(b) 2x27 chain X



(c) 3pb8 chain B



(d) 4pzo chain A

前沿应用

蛋白质相互作用预测

Learning Unknown from Correlations: Graph Neural Network for Inter-novel-protein Interaction Prediction

Guofeng Lv, Zhiqiang Hu, Yanguang Bi, Shaoting Zhang
SenseTime Research
{lvguofeng, huzhiqiang, biyanguang, zhangshaoting}@sensetime.com

(2021 IJCAI)

预测蛋白质间的相互作用，如交换反应产物、参与信号传递机制、共同促进特定的生物体功能等，其对治疗靶点的识别、新药的研发有着非常重要的意义

$$\mathcal{X} = \{x_{ij} = \{p_i, p_j\} | i \neq j, p_i, p_j \in \mathcal{P}, I(x_{ij}) \in \{0, 1\}\}$$

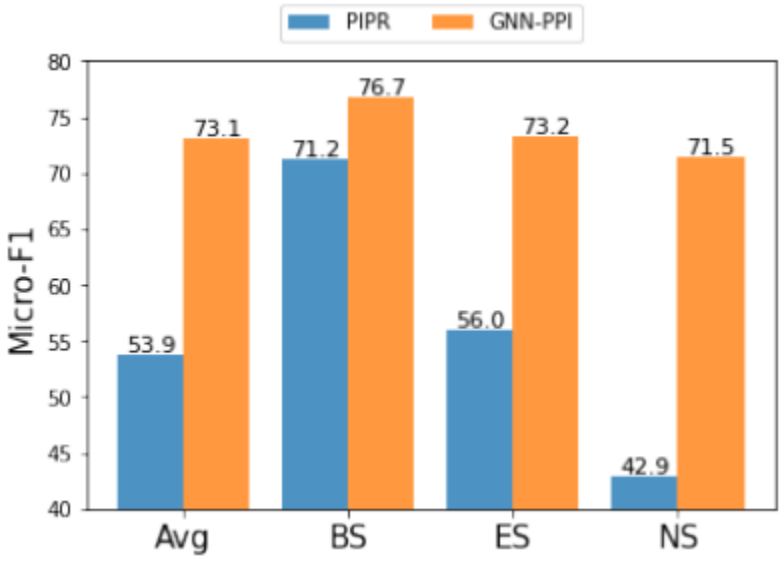
$$\mathcal{L} = \{l_0, l_1, ..., l_n\} \quad y_{ij} \subseteq \mathcal{L}$$

$$\mathcal{D} = \{(x_{ij}, y_{ij}) | x_{ij} \in \mathcal{X}\}$$

多类别标注问题

前沿应用

蛋白质相互作用预测



BS：两个蛋白质均在训练时可见

ES：只有一个蛋白质在训练时可见

NS：两个蛋白质都在训练时不可见

$$\mathcal{X}_{\mathrm{BS}} = \{x_{ij} | x_{ij} \in \mathcal{X}_{\mathrm{test}}, p_i, p_j \in \mathcal{P}_v\}$$
$$\mathcal{X}_{\mathrm{ES}} = \{x_{ij} | x_{ij} \in \mathcal{X}_{\mathrm{test}}, p_i \in \mathcal{P}_u, p_j \in \mathcal{P}_v \text{ or } p_j \in \mathcal{P}_u, p_i \in \mathcal{P}_v\}$$
$$\mathcal{X}_{\mathrm{NS}} = \{x_{ij} | x_{ij} \in \mathcal{X}_{\mathrm{test}}, p_i, p_j \in \mathcal{P}_u\}$$

1）ES、NS子集性能较差是模型性能下降的主要原因

2）当前评价体系采用蛋白质无关的随机策略划分训练集和测试集，这导致BS集占总测试集的92%，BS子集上的性能统治了整体性能，忽略了新蛋白质间的相互作用

3）现有方法忽视了蛋白质之间的相关关系

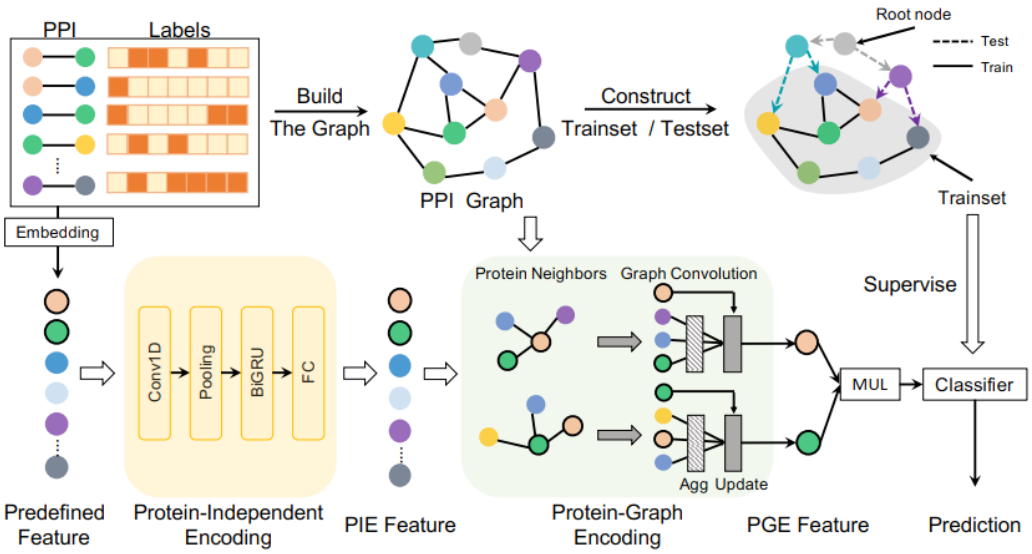
本文贡献：

1）设计了一个新的评估框架，充分考虑了新蛋白质间的相互作用 $|\mathcal{X}_{\mathrm{BS}}| \ll |\mathcal{X}_{\mathrm{ES}}| + |\mathcal{X}_{\mathrm{NS}}|$

2）提出了一个基于图神经网络的模型来建模蛋白质之间的相关关系

前沿应用

蛋白质相互作用预测



本文贡献：

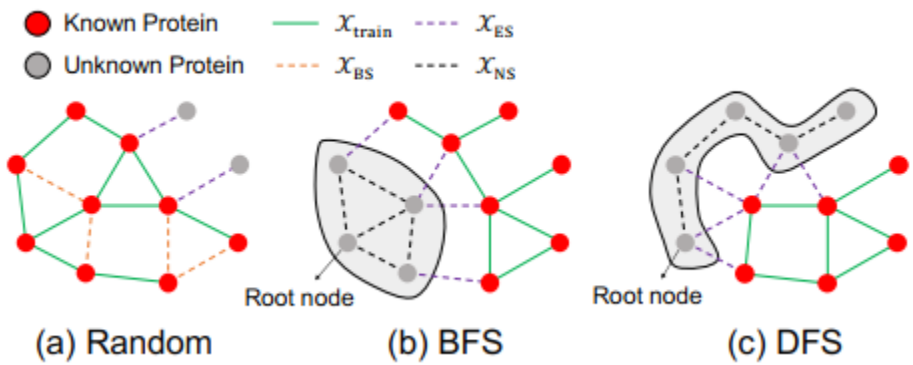
1）设计了一个新的评估框架，充分考虑了新蛋白质间的相互作用

2）提出了一个基于图神经网络的模型来建模蛋白质之间的相关关系

3）大量实验验证所提GNN-PPI模型对于新蛋白质间相互作用关系预测的有效性

前沿应用

蛋白质相互作用预测



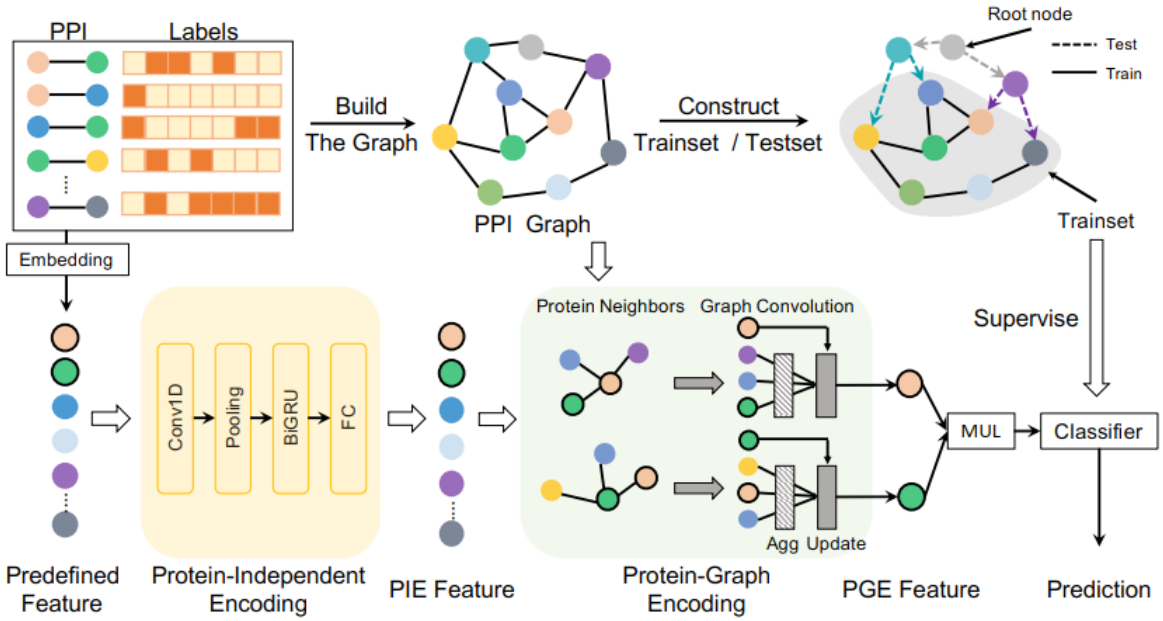
评估框架：利用DFS算法和BFS算法来构造测试集

1）BFS算法：蛋白质之间紧密联系，在PPI网络中以聚类形式存在

2）DFS算法：蛋白质在PPI网络中稀疏分布，彼此之间相互作用关系较少

前沿应用

蛋白质相互作用预测



蛋白质特征编码：编码蛋白质的属性特征和拓扑结构特征

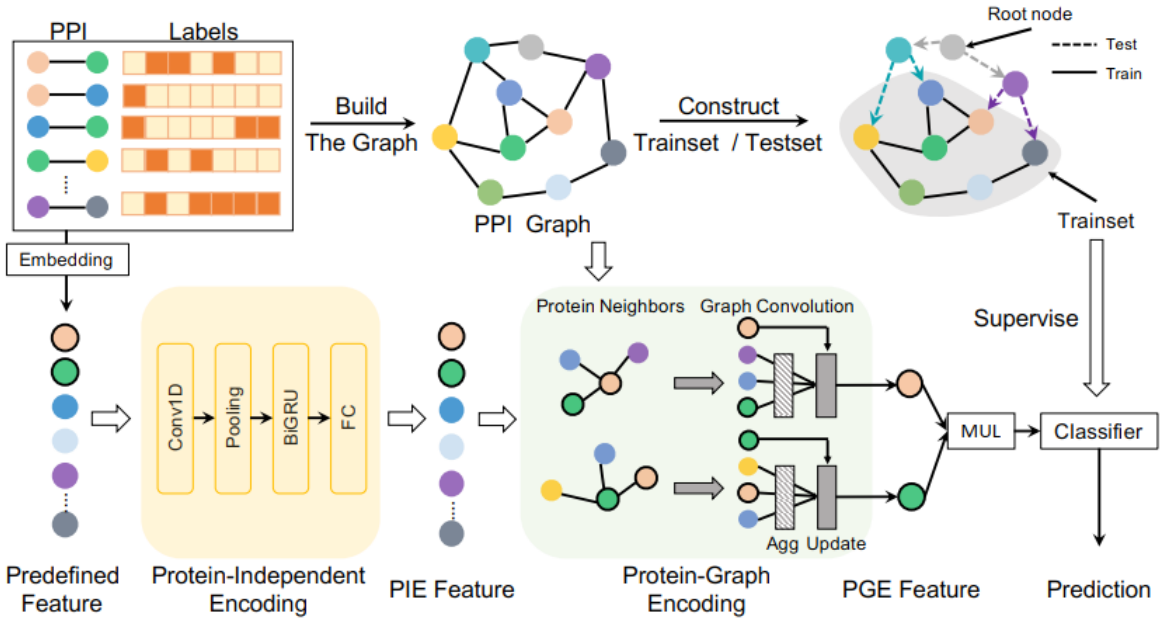
1）蛋白质独立编码：捕获基于氨基酸序列的蛋白质特征，其作为PPI网络中蛋白质的初始特征

2）引入PPI网络，将传统独立学习任务转换成图相关学习任务 $\mathcal{F}(x_{ij}|p_i,p_j,\theta) \rightarrow \hat{y}_{ij}$ $\mathcal{F}(x_{ij}|\mathcal{G},\theta) \rightarrow \hat{y}_{ij}$

3）利用GIN模型来学习节点的特征表示 $g_p^k = \mathrm{MLP}^k\left((1+\epsilon^k)\cdot g_p^{k-1} + \sum_{p'\in\mathcal{N}(p)} g_{p'}^{k-1}\right)$

前沿应用

蛋白质相互作用预测



多标签蛋白质相互作用预测：

1）利用点乘操作联合所预测两个蛋白质的特征

2）利用全连接层预测模型输出 $\hat{y}_{ij} = \mathrm{FC}(g_{p_i} \cdot g_{p_j})$

3）利用多任务二元交叉熵作为loss函数 $L = \sum_{k=0}^{n} \left(\sum_{x_{ij} \in \mathcal{X}_{\text{train}}} -y_{ij}^{k} \log \hat{y}_{ij}^{k} - (1 - y_{ij}^{k}) \log(1 - \hat{y}_{ij}^{k}) \right)$

前沿应用

蛋白质相互作用预测

Dataset	Partition Scheme	Methods						
		SVM	RF	LR	DPPI	DNN-PPI	PIPR	GNN-PPI
SHS27k	Random	75.35±1.05	78.45±0.88	71.55±0.93	73.99±5.04	77.89±4.97	83.31±0.75	**87.91**±0.39
	BFS	42.98±6.15	37.67±1.57	43.06±5.05	41.43±0.56	48.90±7.24	44.48±4.44	**63.81**±1.79
	DFS	53.07±5.16	35.55±2.22	48.51±1.87	46.12±3.02	54.34±1.30	57.80±3.24	**74.72**±5.26
SHS148k	Random	80.55±0.23	82.10±0.20	67.00±0.07	77.48±1.39	88.49±0.48	90.05±2.59	**92.26**±0.10
	BFS	49.14±5.30	38.96±1.94	47.45±1.42	52.12±8.70	57.40±9.10	61.83±10.23	**71.37**±5.33
	DFS	58.59±0.07	43.26±3.43	51.09±2.09	52.03±1.18	58.42±2.05	63.98±0.76	**82.67**±0.85
STRING	Random	-	88.91±0.08	67.74±0.16	94.85±0.13	83.08±0.11	94.43±0.10	**95.43**±0.10
	BFS	-	55.31±1.02	50.54±2.00	56.68±1.04	53.05±0.82	55.65±1.60	**78.37**±5.40
	DFS	-	70.80±0.45	61.28±0.53	66.82±0.29	64.94±0.93	67.45±0.34	**91.07**±0.58

Table 1: Performance of GNN-PPI against comparative methods over different datasets and data partition schemes. The reported results are mean±std micro-averaged F1 score over three repeated experiments. Results of SVM on STRING is omitted for unafforable running time.

Dataset	Partition Scheme	$\mathcal{X}_{\text{BS}}$		$\mathcal{X}_{\text{ES}}$		$\mathcal{X}_{\text{NS}}$		$\mathcal{X}_{\text{Avg}}$				
		PIPR	GNN-PPI	PIPR	GNN-PPI	PIPR	GNN-PPI	Proportion(BS/ES/NS)			PIPR	GNN-PPI
SHS27k	Random	83.12	88.31	64.48	74.28	35.29	33.33	92.2	7.5	0.3	81.58	87.11
	BFS	-	-	44.92	68.08	30.34	46.25	0.0	72.6	27.4	40.92	62.10
	DFS	-	-	58.25	72.22	48.77	63.22	0.0	88.6	11.4	57.17	71.19
SHS148k	Random	92.82	92.24	78.80	73.09	40.72	36.36	97.2	2.7	0.1	92.42	91.68
	BFS	-	-	62.80	72.51	73.82	77.02	0.0	69.7	30.3	66.13	73.88
	DFS	-	-	64.17	83.37	55.51	73.08	0.0	91.9	8.1	63.47	82.54
STRING	Random	94.32	95.42	61.65	77.68	33.33	57.14	99.7	0.3	0	94.23	95.37
	BFS	-	-	56.71	83.99	39.87	72.83	0.0	85.8	14.2	54.31	82.41
	DFS	-	-	68.61	90.38	55.22	87.07	0.0	94.3	5.7	67.84	90.19

Table 2: In-depth analysis between PIPR and GNN-PPI over BS, ES and NS subsets.

前沿应用

蛋白质功能预测

ARTICLE

https://doi.org/10.1038/s41467-021-23303-9 OPEN

Check for updates

Structure-based protein function prediction using graph convolutional networks

Vladimir Gligorijević[1✉], P. Douglas Renfrew[1], Tomasz Kosciolek[2,3], Julia Koehler Leman[1], Daniel Berenberg[1,4], Tommi Vatanen[5,6], Chris Chandler[1], Bryn C. Taylor[7], Ian M. Fisk[8], Hera Vlamakis[5], Ramnik J. Xavier[5,9,10,11], Rob Knight[2,12,13], Kyunghyun Cho[14,15] & Richard Bonneau[1,4,14,16✉]

(2021 Nature Communications)

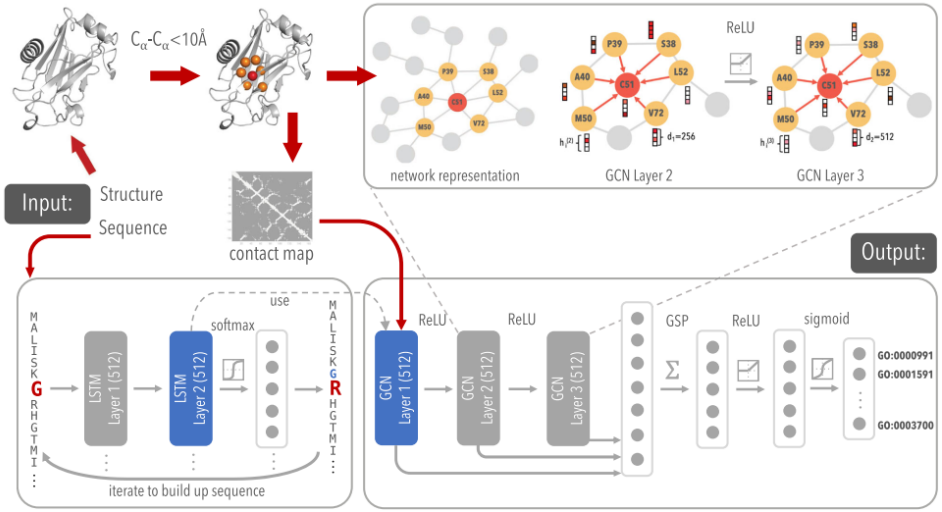
蛋白质功能：结合特异性、稳定性、生物化学反应的催化、运输、信号传导……

由于检测蛋白质功能实验的规模、设计和成本的考虑，大多数具有未知功能的蛋白质（即假设的蛋白质）不太可能通过实验进行验证

了解新发现的蛋白质的功能作用和研究其机制是后基因组时代最重要的生物学问题之一。

前沿应用

蛋白质功能预测



现有方法的不足：

1）大部分现有方法仅仅利用序列或基于序列的特征作为模型的输入来预测蛋白质功能，但是蛋白质的结构信息对于蛋白质功能的预测非常重要

2）虽然一些方法使用contact map和3D卷积神经网络来进行蛋白质功能预测，证明了蛋白质结构信息的重要性，但是以高分辨率存储和处理蛋白质结构并不是存储有效的，因为大部分3D空间未被蛋白质结构所占据

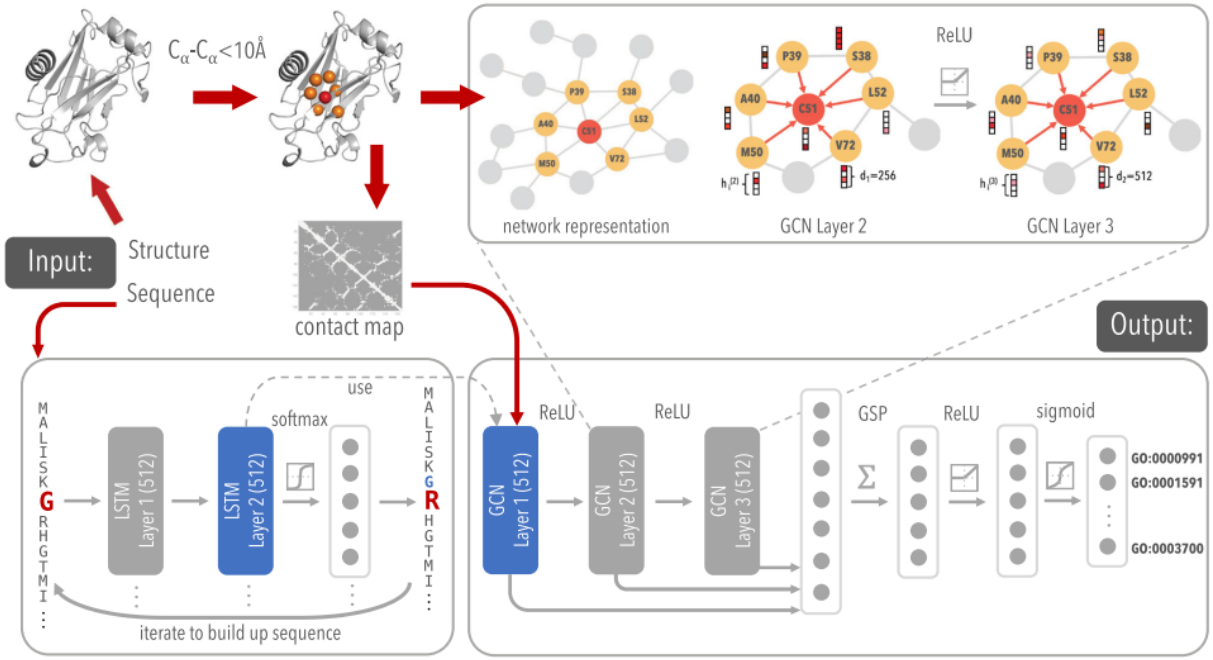
本文贡献：作者提出了一个基于GNN的蛋白质功能预测框架，其包含两个模块

1）LSTM模块提取氨基酸序列级别的特征

2）GNN提取蛋白质结构特征

前沿应用

蛋白质功能预测



构建contact map：

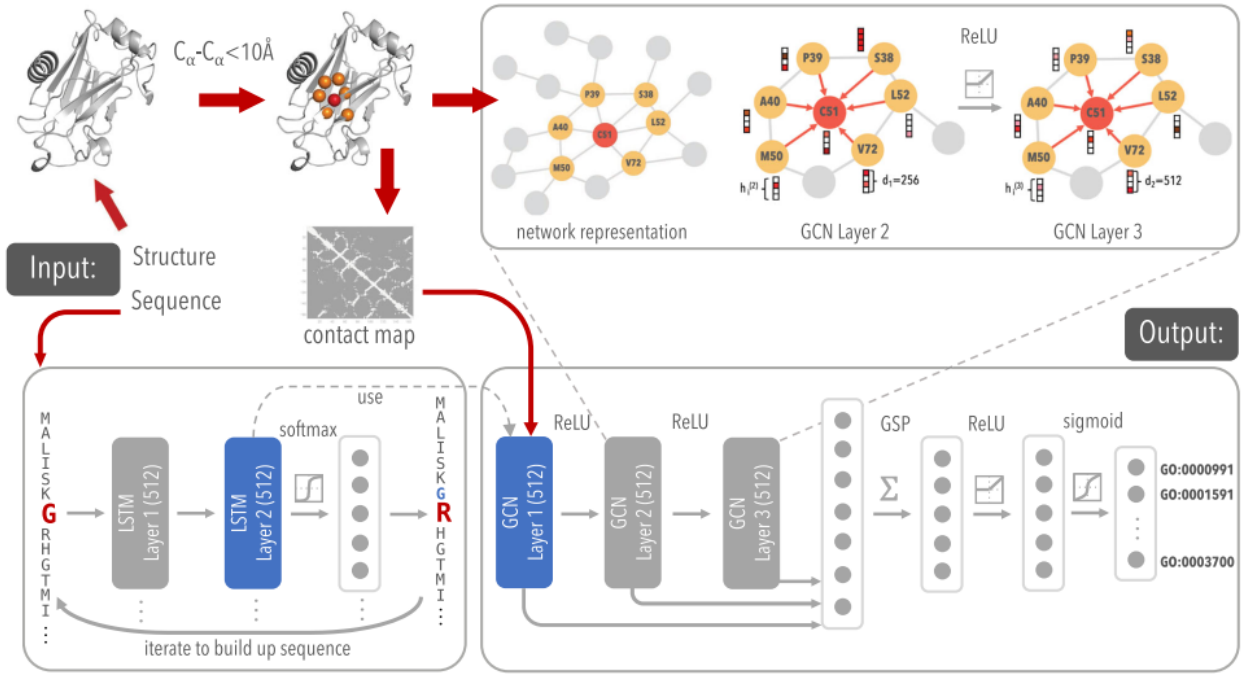
1）两个残基间 C_α 原子的距离小于10 Å

2）两个残基间任意两个原子的距离小于6.5 Å

3）两个残基间 C_β 原子的距离小于氨基酸对的相邻半径之和

前沿应用

蛋白质功能预测



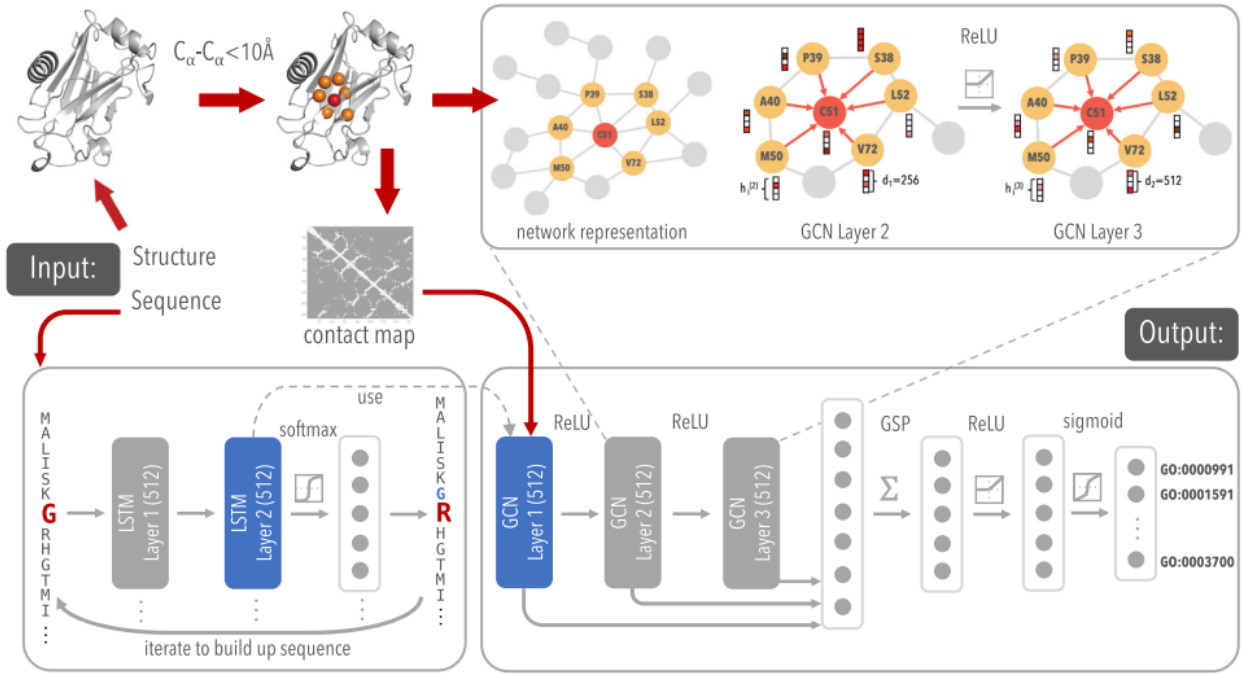
利用LSTM语言模型来学习残基级别特征：构建氨基酸序列预测任务来训练LSTM模型的参数（预训练）

$$\mathbf{H}^{\text{input}} = \text{ReLU}(\mathbf{H}^{LM}\mathbf{W}^{LM} + \mathbf{X}\mathbf{W}^{X} + \mathbf{b})$$

获取残基级别特征作为GNN模型的输入，在模型训练过程中LSTM模型的参数固定，不参与训练

前沿应用

蛋白质功能预测



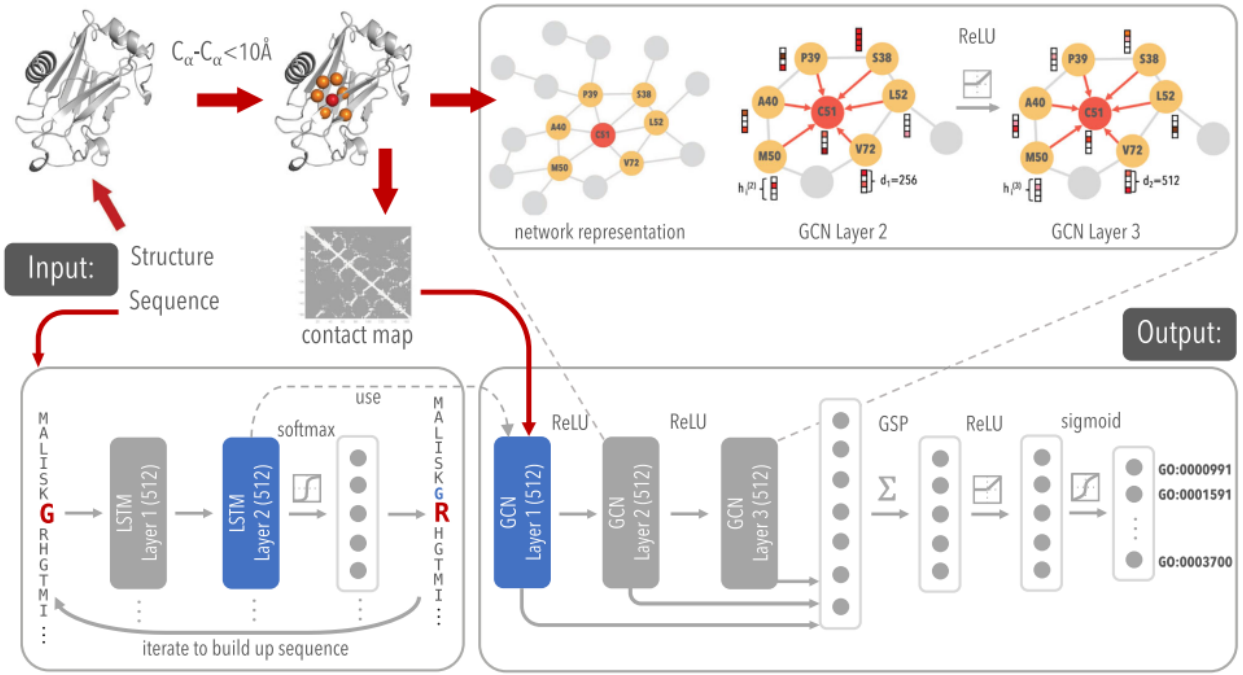
利用GCN学习图特征表示：利用GCN学习节点特征，拼接节点所有层特征后利用pooling操作得到蛋白质特征表示

$$\mathbf{H}^{l+1} = \mathrm{ReLU}(\tilde{\mathbf{D}}^{-0.5}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-0.5}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \qquad \mathbf{H} = [\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(N_l)}] \in \mathbb{R}^{L \times \sum_{i=1}^{L} c_i}$$

$$\mathbf{h}^{\mathrm{pool}} = \sum_{i=1}^{L} \mathbf{H}_{i:}$$

前沿应用

蛋白质功能预测



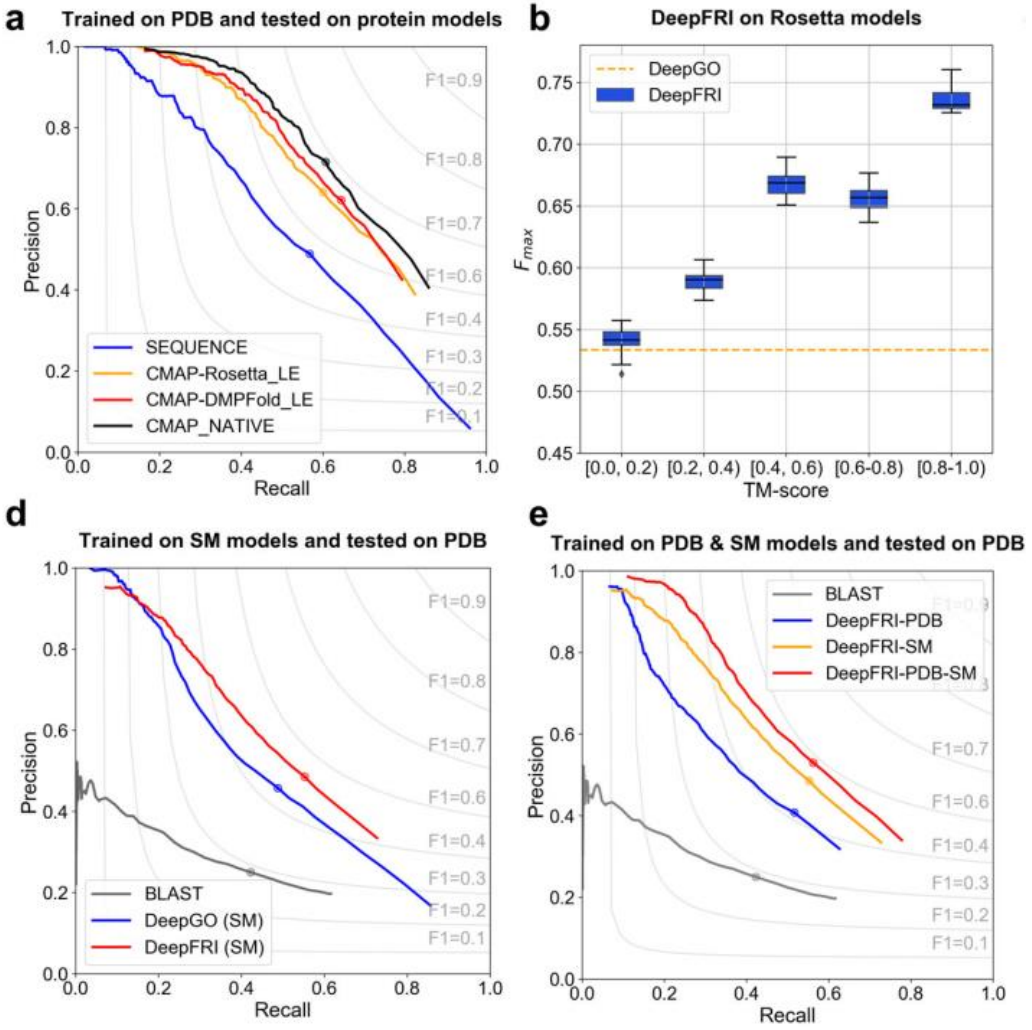
为了解决标签不平衡问题，利用带权二元交叉熵损失函数来训练模型参数

$$\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|GO|}\sum_{k=1}^{2} w_j y_{ijk}\log\left(\hat{y}_{ijk}\right) \qquad w_j = \frac{N}{N_j^+}$$

$y_{ij1} = 1$, if sample i is annotated with function j, and $y_{ij2} = 0$

前沿应用

蛋白质功能预测



前沿应用

疾病预测（抑郁症严重程度）

USING GRAPH REPRESENTATION LEARNING WITH SCHEMA ENCODERS TO MEASURE THE SEVERITY OF DEPRESSIVE SYMPTOMS

Simin Hong
School of Computing
University of Leeds, UK
scsho@leeds.ac.uk

Anthony G. Cohn
School of Computing
University of Leeds, UK
a.g.cohn@leeds.ac.uk

David C. Hogg
School of Computing
University of Leeds, UK
d.c.hogg@leeds.ac.uk

（2022 ICLR）

利用GNN来预测重度抑郁症（世界范围内超过3E人患有抑郁症，85%的抑郁症患者是没有经过测量诊断的，30%的抑郁症患者没有寻求治疗，只有10%的抑郁症患者得到治疗）

前沿应用

疾病预测（抑郁症严重程度）

i always feel irritated. i am lazy when i do not sleep well. my mood was just not right, i was always feeling down and depressed and lack of energy. i always want to sleep. i am lack of interest. i have gone to therapy, it has been useful for me in the past. i would love to talk to someone, i just feel like i do not have anyone so i do not depend on anyone. i have always felt depressed in my life, my symptoms were lack of energy, wanting to sleep a lot, lack of interest. my appetite was uncontrollable either lack of or i was just being gluttonous and eating the wrong things. i have notices those changes in my behavior……

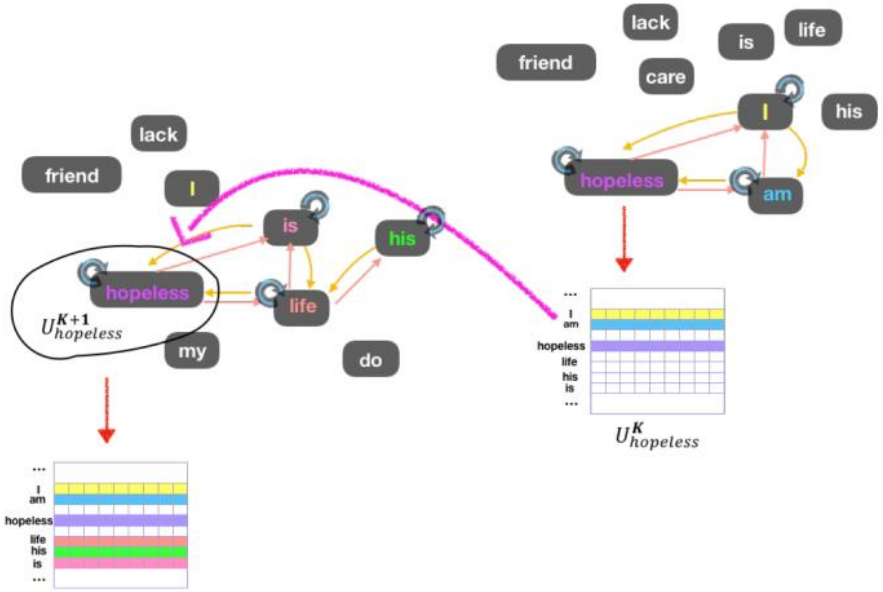
已有研究发现文本可以用来推断一个人的心理状态，例如抑郁症患者比正常人会使用更多的第一人称代词，他们对于自己更加关注，而与其他人的联系较少

在实际生活中，医生通常利用一份结构化的患者健康问卷（PHQ）来确定患者抑郁症的严重程度，作者根据患者回答中词的上下文关系来预测抑郁症的严重程度

将句子中的词建模成一个矩阵表示，编码当前词与其他词之间的上下文关系信息，同时利用GNN来捕获词与词之间的上下文关系，更新词嵌入矩阵

前沿应用

疾病预测（抑郁症严重程度）

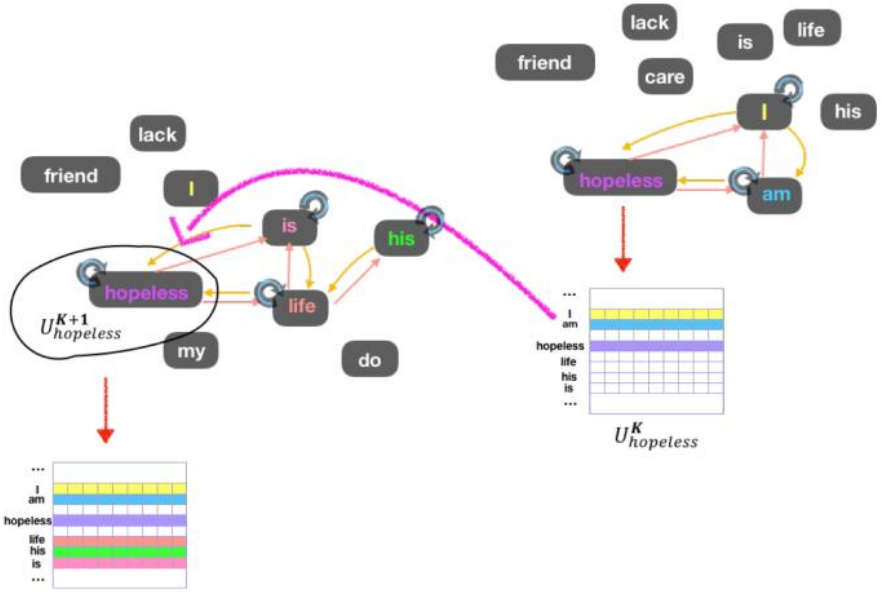


构建文本图：

文本中的词作为节点，句子中词与词的距离小于窗长时，构建词与词之间的连边（边也代表了共现关系）

前沿应用

疾病预测（抑郁症严重程度）



模型框架：

1）初始化：将文本中的每个词初始成一个词矩阵，当前词对应行随机初始化，其他行置为全0 $U_i \in \mathbb{R}^{n \times d}$ $v_i \in V$

2）消息传播层：

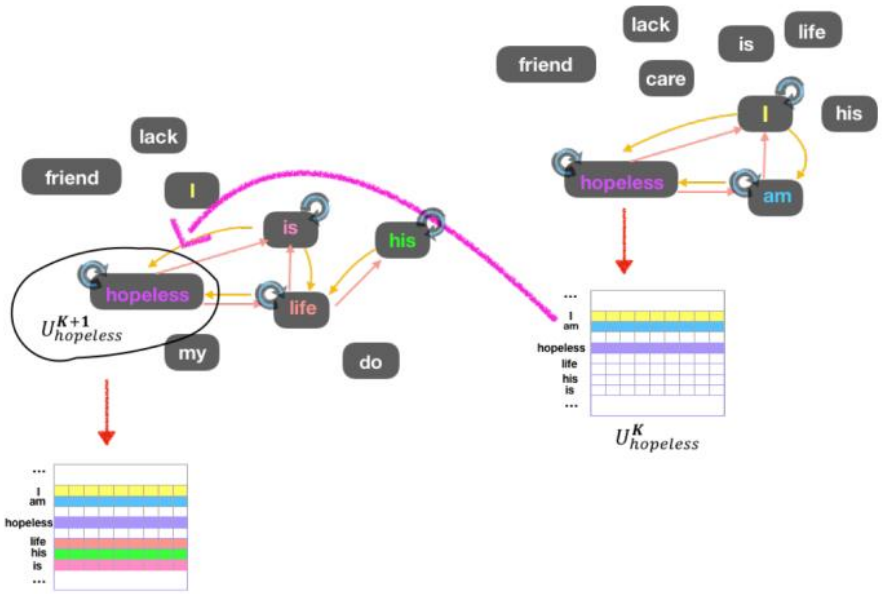
$$\hat{U}_i^{(k)} = U_i^{(k)} W_1^{(k)} + \frac{1}{n} \mathbf{1}\mathbf{1}^T U_i^{(k)} W_2^{(k)} + \frac{1}{n} \mathbf{1}_i \mathbf{1}^T U_i^{(k)} W_3^{(k)}$$

$$M^{(k)}\left(\hat{U}_i^{(k)}, \hat{U}_j^{(k)}\right) = \hat{U}_j^{(k)} + RELU\left(\left[\hat{U}_i^{(k)} || \hat{U}_j^{(k)}\right] W_4^{(k)}\right) W_5^{(k)}$$

$$U_i^{(k+1)} = \sum_{j \in N(i)} M^{(k)}\left(\hat{U}_i^{(k)}, \hat{U}_j^{(k)}\right) \in \mathbb{R}^{n \times d_{k+1}}$$

前沿应用

疾病预测（抑郁症严重程度）



模型框架：

1）初始化：将文本中的每个词初始成一个词矩阵，当前词对应行随机初始化，其他行置为全0 $U_i \in \mathbb{R}^{n \times d}$ $v_i \in V$

2）消息传播层：进行消息聚合以及词嵌入矩阵的更新

3）预测层：先利用最大池操作获取节点（词）的特征表示，然后利用readout函数获取文本图表示，文本图表示经过MLP预测最终PHQ分数

$$h_G = \sum_{v_i \in V} U_i^{(K)}$$

前沿应用

疾病预测（抑郁症严重程度）

Table 1: Comparison of machine learning approaches for measuring the severity of depressive symptoms on the DAIC-WOZ development set using mean absolute error (MAE). The task evaluated is: PHQ score regression. Modalities: A: audio, V: visual, L: linguistic(text), A+V+L: combination. The result marked with a * has been computed by us; the others are taken from the cited papers.

Regression: PHQ score		
Methods	Modalities	PHQ score MAE
Baseline Challenge (Valstar et al., 2016)	A+V	5.52
Gaussian Staircase Regression (Williamson et al., 2016)	A+V+L	4.18
LSTM (Haque et al., 2018)	A+V+L	5.18
LSTM (Alhanai et al., 2018)	A+L	5.1
DCGAN (Yang et al., 2020)	A	4.63
DepArt-Net (Du et al., 2019)	V	4.65
LSTM (Alhanai et al., 2018)	L	5.2
C-CNN (Haque et al., 2018)	L	6.14
BiLSTM(Lin et al., 2020)	L	3.88
Multi-level Attention network (Ray et al., 2019)	L	4.37
*GNN (Gilmer et al., 2017)	L	4.24
Our Proposed Approach	L	**3.76**

Table 2: Results of ablation studies. We apply 10-fold stratified cross-validation and give mean results. MAE: mean absolute error. RMSE: root mean squared error.

Metric	MAE		RMSE	
Setting	Train	Test	Train	Test
Original(schema-GNN)	3.85	3.52	4.48	4.32
i) Fast(schema-GNN)	4.28	3.86	5.14	4.51
ii) Without equivariant linear layers (Maron et al., 2018)	4.14	3.95	5.60	4.49
iii) Mean Reduction	4.32	5.16	4.45	5.12

前沿应用

分子图生成

SPANNING TREE-BASED GRAPH GENERATION FOR MOLECULES

Sungsoo Ahn[1], **Binghong Chen**[2], **Tianzhe Wang**[2], **Le Song**[3,4]
[1]POSTECH, [2]Georgia Institute of Technology, [3]Biomap, [4]MBZUAI
sungsoo.ahn@postech.ac.kr, {binghong, tianzhe}@gatech.edu,
dasongle@gmail.com

(2022 ICLR)

将分子图生成任务形式化为树的生成和残基间连边的构造

前沿应用

分子图生成

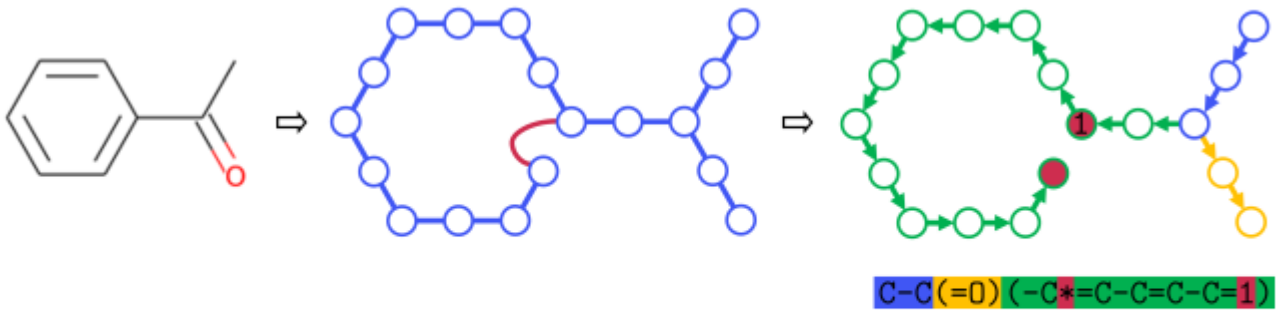


Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

两大挑战：

1）构造整个分子图的步骤需要尽可能少

2）确保生成的分子满足化学价态规则

本文贡献：

1）将分子图生成作为生成树和相应的残基边的组合，其中原子和键作为生成树中的节点。

2）预先定义违反分子图构造和化学价态规律的决策从而保证生成模型构造分子的合理性

3）提出一个基于Transformer的框架来构造生成模型

前沿应用

分子图生成

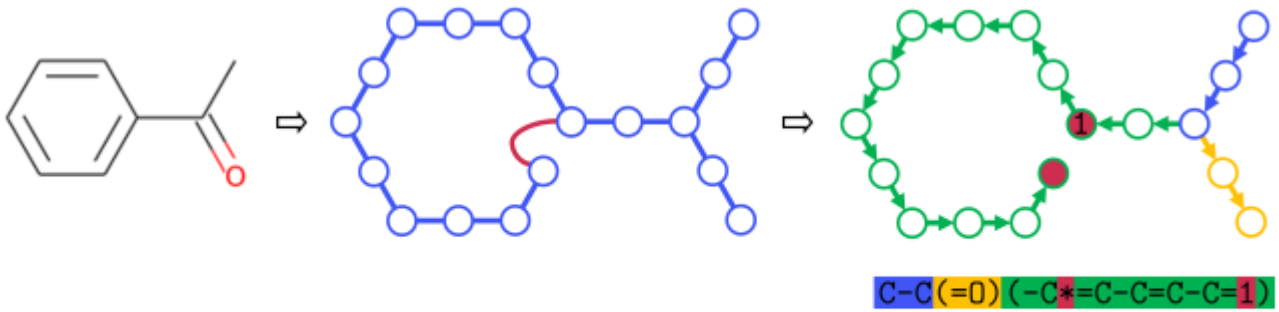


Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

两大挑战：

1）构造整个分子图的步骤需要尽可能少

2）确保生成的分子满足化学价态规则

本文贡献：

1）将分子图生成作为生成树和相应的残余边的组合，其中原子和键作为生成树中的节点。

2）预先定义违反分子图构造和化学价态规律的决策从而保证生成模型构造分子的合理性

3）提出一个基于Transformer的框架来构造生成模型

前沿应用

分子图生成

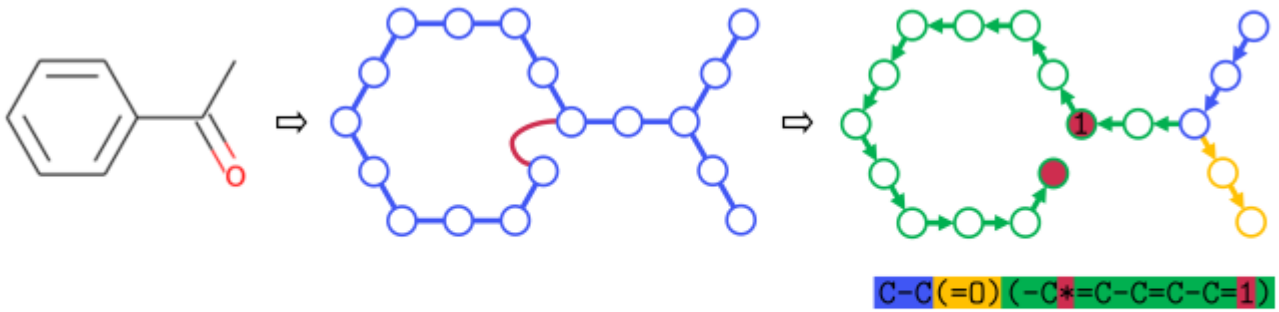


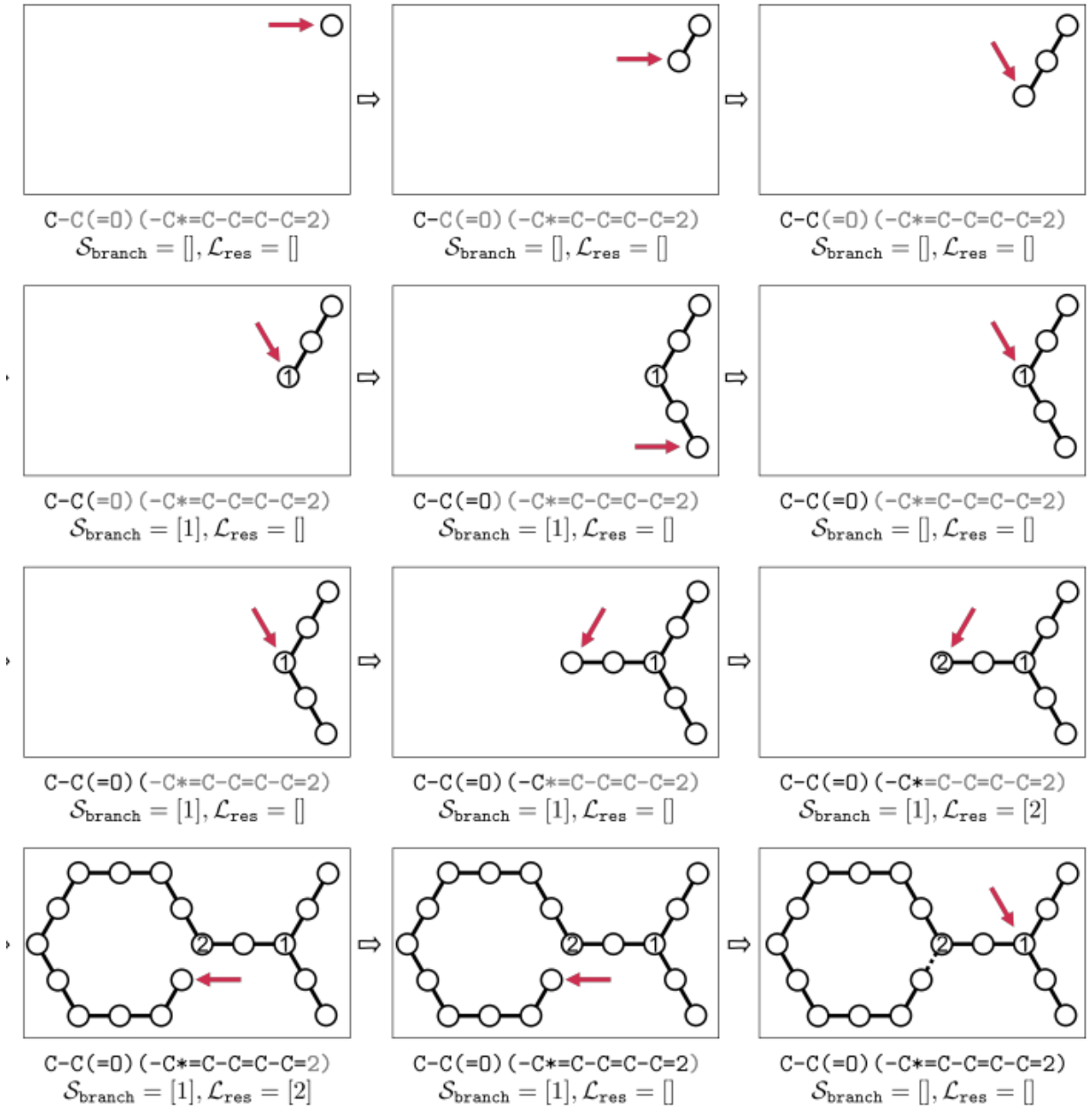
Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

分子图表示：将分子图构造成二分图 $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{E})$ $\mathcal{A}$、$\mathcal{B}$ 代表原子节点集和化学键集，$\{a, b\} \in \mathcal{E}$ 代表原子和化学键间的连边，原子属性 $x_a \in \mathcal{X}_{\texttt{atom}}$，边属性 $x_b \in \mathcal{X}_{\texttt{bond}}$

分子图生成：为了生成分子图 $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{E})$，需要一系列操作 $d_1, \ldots, d_T$ 来生成树 $\mathcal{T} = (\mathcal{A}_\mathcal{T}, \mathcal{B}_\mathcal{T}, \mathcal{E}_\mathcal{T})$ 和残余边 $\mathcal{E}_R = \mathcal{E} \setminus \mathcal{E}_\mathcal{T}$，一共有七类操作可供选择

attach_atom	attach_bond	branch_start	branch_end	res_atom	res_bond	terminate
"C" $\in \mathcal{X}_{\texttt{atom}}$	"-" $\in \mathcal{X}_{\texttt{bond}}$	"("	")"	"*"	$d \in \mathcal{L}_{\texttt{res}}$	"[eos]"

attach_atom	attach_bond	branch_start	branch_end	res_atom	res_bond	terminate
"C" $\in \mathcal{X}_{\text{atom}}$	"-" $\in \mathcal{X}_{\text{bond}}$	"("	")"	"*"	$d \in \mathcal{L}_{\text{res}}$	"[eos]"



Algorithm 1 Tree-based generation of molecular graphs

Input: sequence of decisions $d_1, \ldots, d_T$.
Output: graph $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{E})$, atom attributes $\{x_a\}_{a \in \mathcal{A}}$, and bond attributes $\{x_b\}_{b \in \mathcal{B}}$
Set $\mathcal{A}_\mathcal{T} \leftarrow \emptyset$, $\mathcal{B}_\mathcal{T} \leftarrow \emptyset$, $\mathcal{E}_\mathcal{T} \leftarrow \emptyset$, $\mathcal{E}_R \leftarrow \emptyset$, and $\mathcal{T} \leftarrow (\mathcal{A}_\mathcal{T}, \mathcal{B}_\mathcal{T}, \mathcal{E}_\mathcal{T})$. ▷ *Initialize the empty graph.*
Set $\mathcal{L}_{\texttt{res}}$ as an empty list and $\mathcal{S}_{\texttt{branch}}$ as an empty stack.
for $t = 1, \ldots, T$ **do**
    if $d_t \in \mathcal{X}_{\texttt{atom}}$ **then** ▷ *Add a new atom vertex.*
        Create a new atom vertex a and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{b\}$ and $x_a \leftarrow d_t$.
        If $|\mathcal{B}_\mathcal{T}| > 0$, set $\mathcal{E}_\mathcal{T} \leftarrow \mathcal{E}_\mathcal{T} \cup \{\{a, i_{\texttt{point}}\}\}$. ▷ *Edge is added when tree is not empty.*
        Set $i_{\texttt{point}} \leftarrow a$.
    if $d_t \in \mathcal{X}_{\texttt{bond}}$ **then** ▷ *Add a new bond vertex.*
        Create a new bond vertex b and set $\mathcal{E}_\mathcal{T} \leftarrow \mathcal{E}_\mathcal{T} \cup \{\{b, i_{\texttt{point}}\}\}$, $\mathcal{B}_\mathcal{T} \leftarrow \mathcal{B}_\mathcal{T} \cup \{b\}$, and $x_b \leftarrow d_t$.
        Set $i_{\texttt{point}} \leftarrow b$.
    if $d_t =$ `"*"` **then** Insert $i_{\texttt{point}}$ into $\mathcal{L}_{\texttt{res}}$. ▷ *Add pointer vertex into the queue.*
    if $d_t \in \mathcal{L}_{\texttt{res}}$ **then** Pop d_t from $\mathcal{L}_{\texttt{res}}$ and update $\mathcal{E}_R \leftarrow \mathcal{E}_R \cup \{\{i_{\texttt{point}}, d_t\}\}$. ▷ *Add a new residual edge.*
    if $d_t =$ `"("` **then** Insert $i_{\texttt{point}}$ into $\mathcal{S}_{\texttt{branch}}$. ▷ *Add pointer vertex into the stack.*
    if $d_t =$ `")"` **then** Set $i_{\texttt{point}} \leftarrow \texttt{pop}(\mathcal{S}_{\texttt{branch}})$ ▷ *Update pointer vertex from the stack.*
Set $\mathcal{A} \leftarrow \mathcal{A}_\mathcal{T}$, $\mathcal{B} \leftarrow \mathcal{B}_\mathcal{T}$, and $\mathcal{E} \leftarrow \mathcal{E}_\mathcal{T} \cup \mathcal{E}_R$.

attach_atom	attach_bond	branch_start	branch_end	res_atom	res_bond	terminate
"C" $\in \mathcal{X}_{\text{atom}}$	"-" $\in \mathcal{X}_{\text{bond}}$	"("	")"	"*"	$d \in \mathcal{L}_{\text{res}}$	"[eos]"

前沿应用

分子图生成

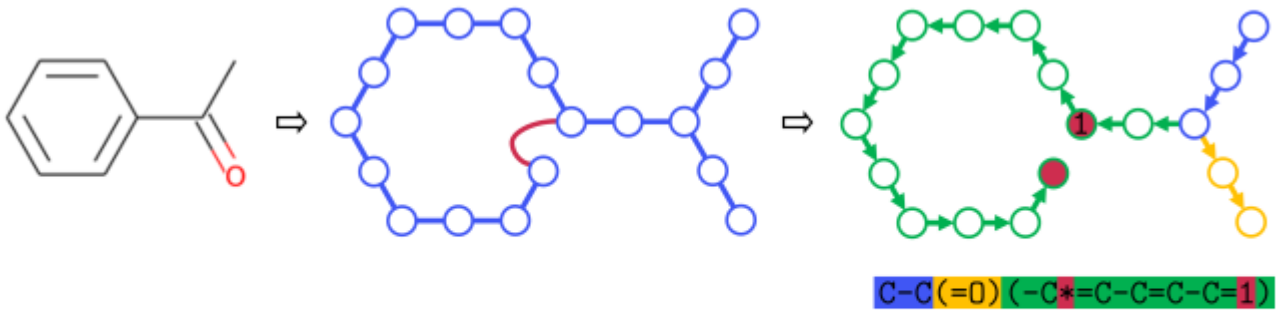


Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

过滤不符合规则的操作：

1）图生成的有效性：branch_end操作只能在栈S_{branch}为空时出现、res_atom和res_bond只能在指针指向原子节点和化学键节点的时候出现、 branch_end(branch_end)操作只能在指针指向原子节点的时候出现、栈内不包含重复的顶点指针

2）化学价态规则：$v(x_a) \geq \sum_{b \in \mathcal{N}(a)} o(x_b)$ ，$v(x_a)$ 是原子a的价态，$o(x_b)$ 是化学键的阶

前沿应用

分子图生成

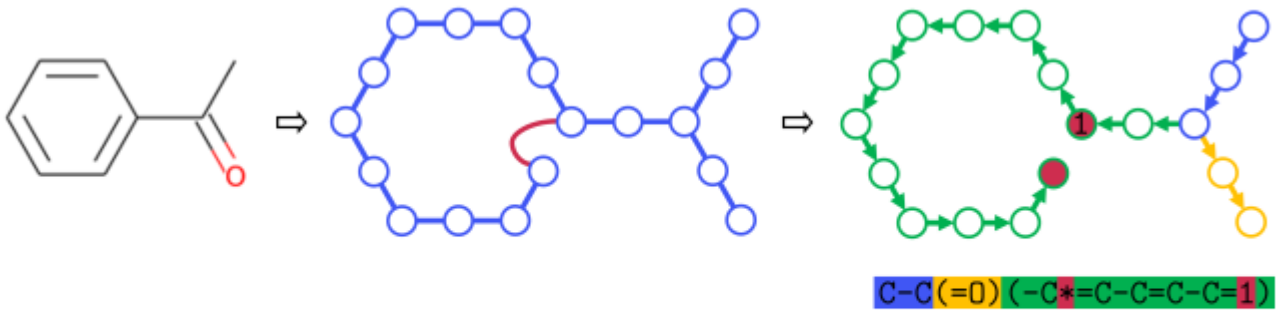


Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

基于transformer的模型架构：

$$Q = HW_Q, \quad K = HW_K, V = HW_V,$$

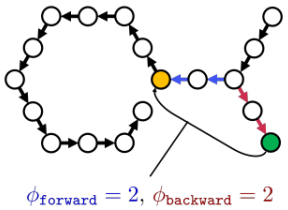
$$A = \frac{QK^\top}{\sqrt{\ell_K}} + P, \quad P_{t_1,t_2} = \mathbf{z}^{(1)}_{\phi_{\text{forward}}(t_1,t_2)} + \mathbf{z}^{(2)}_{\phi_{\text{backward}}(t_1,t_2)} + \mathbf{z}^{(3)}_{\phi_{\text{seq}}(t_1,t_2)},$$

$$\text{Attention}(H) = \text{SoftMax}(M \circ A)V,$$

M 是掩码矩阵防止模型在预测时利用未来的信息

P 是相对位置编码 $\phi_{\text{seq}}(t_1, t_2) = t_1 - t_2$

$\phi_{\text{forward}}(t_1, t_2), \phi_{\text{backward}}(t_1, t_2)$



前沿应用

分子图生成

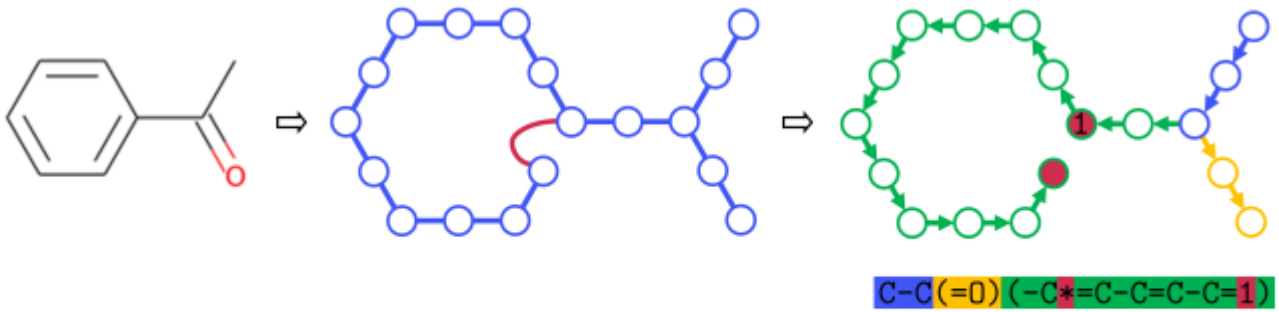


Figure 1: (left) Benzaldehyde, (middle) its spanning tree (blue) and residual edges (red), and the corresponding constructive decisions (right). Open circle represent atoms and bonds in the molecule.

预测层:

$$p(d) \propto \begin{cases} m_g(d) \cdot m_v(d) \cdot \exp(\mathbf{w}_d^\top h) & \forall d \in \mathcal{X}_{\text{atom}} \cup \mathcal{X}_{\text{bond}} \cup \{\text{"("}, \text{")"}, \text{"*"}\} \\ m_g(d) \cdot m_v(d) \cdot \exp\left(h_d^\top W_1 W_2^\top h\right) & \forall d \in \mathcal{L}_{\text{res}}, \end{cases}$$

$m_v(d), m_g(d)$ Mask掩码检测操作是否符合价态和图生成规则

前沿应用

分子图生成

Table 2: Experimental results on ZINC250K and QM9 datasets.

		ZINC250K			QM9		
METHOD	CORRECTABLE	VALID	UNIQUE	NOVEL	VALID	UNIQUE	NOVEL
GCPN (You et al., 2018)	✓	0.20	**1.0000**	**1.0000**	-	-	-
MRNN (Popova et al., 2019)	✓	0.65	0.9989	**1.0000**	-	-	-
GRAPHNVP (Madhawa et al., 2019)		0.426	0.948	**1.0000**	0.831	**0.992**	0.582
GRF (Honda et al., 2019)		0.734	0.537	**1.0000**	0.845	0.66	0.586
GRAPHAF (Shi et al., 2020)	✓	0.680	0.991	**1.0000**	0.67	0.9415	0.8883
MOFLOW (Zang & Wang, 2020)	✓	0.680	0.991	**1.0000**	0.8896	0.9853	0.9604
GRAPHCNF (Lippe & Gavves, 2021)		0.9635	0.9998	0.9998	-	-	-
GRAPHDF (Luo et al., 2021)	✓	0.8903	0.9916	**1.0000**	0.8267	0.9762	**0.9810**
SMILES-TRANSFORMER		0.9558	0.9998	0.9946	0.9908	0.9629	0.6939
STGG (ours)	✓	**0.9950**	0.9999	0.9989	**1.0000**	0.9676	0.7273

总结

一些探讨

1. 现在基于深度学习尤其是基于图深度学习的方法开始应用到生物化学领域，并成为热门研究方向，其对于节约成本，加快研究进程，辅助医学人员工作，发现与创造新物质有着重要的价值和意义

2.目前生物化学领域基于图的研究开始从2D空间向3D空间扩展，三维空间结构能为下游任务提供更多关键的信息

3.异质、时序、医学知识、多模态信息的引入是生物化学领域未来重点研究的方向

4.已有研究发现生物化学网络存在天然的稀疏性和无尺度分布特性，这给模型学习带来挑战

5.生物化学领域缺乏负样本并且已有数据存在大量的噪声，这增加了模型训练的难度

6.在生物化学领域，简单提供计算结果是完全不够的，模型需要具备良好的可解释性

7.计算机辅助的生物化学是未来非常重要的方向！！！

后续讨论班介绍

- 第一周：图神经网络在金融领域的应用-石逢钊
- 第二周：图神经网络赋能的知识图谱研究与应用-刘　瑜
- 第三周：图网络在生物化学领域的应用-周玉晨
- 第四周：图网络在社会网络中的应用-宋传承
- 第五周：图神经网络在推荐系统中的应用-吴咏萱

谢谢您的聆听！

敬请批评指正

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS