HINGRL: predicting drug–disease associations with graph representation learning on heterogeneous information networks

期刊：Briefings in Bioinformatics　　日期：2022　　分区：二区

中国科学院大学

HINGRL: 预测药物-疾病关联与异构信息网络上的图表示学习

问题与方法

背景：

蛋白质表达的变化与疾病表现和药物作用直接相关。

问题：

1.以往方法专注于构建各种类型的异构网络，忽略了分子的内在特征。

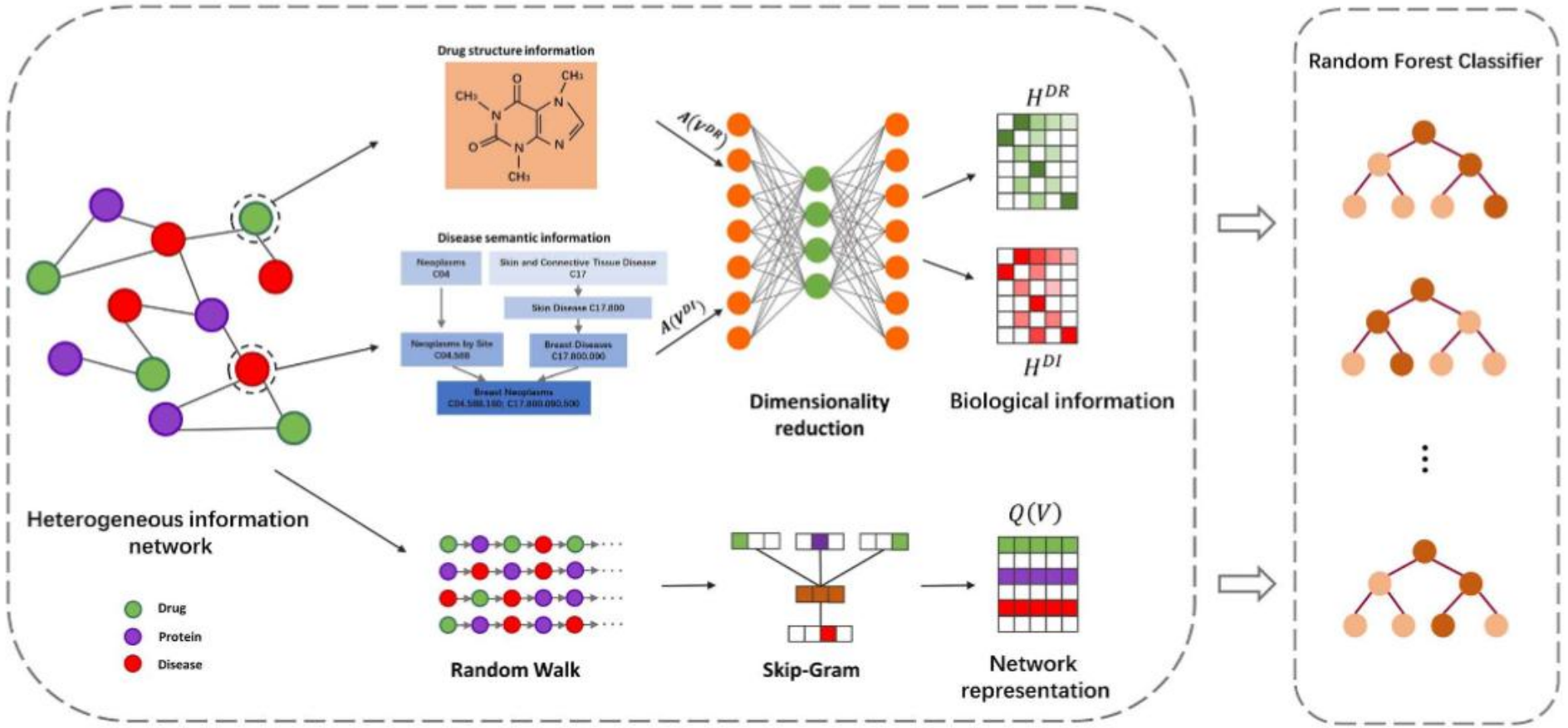
2.缺乏同时考虑同一异构网络中，药物和疾病的网络拓扑和生物知识的模型，也忽略蛋白质的作用。

方法：

1.将蛋白质相关的关联和药物和疾病的生物学知识整合到原始的药物-疾病关联网络中；

2.结合药物和疾病的网络拓扑和生物知识学习节点的特征表示。

HINGRL 整体框架



B-data、F-data：药物-疾病、药物-蛋白质、蛋白质-疾病以及各自二者关联的信息。
构建正负样本相等的数据集，避免不平衡问题。

HIN 模型

$$HIN = \{\mathbf{V}, \mathbf{A}, \mathbf{E}\},$$

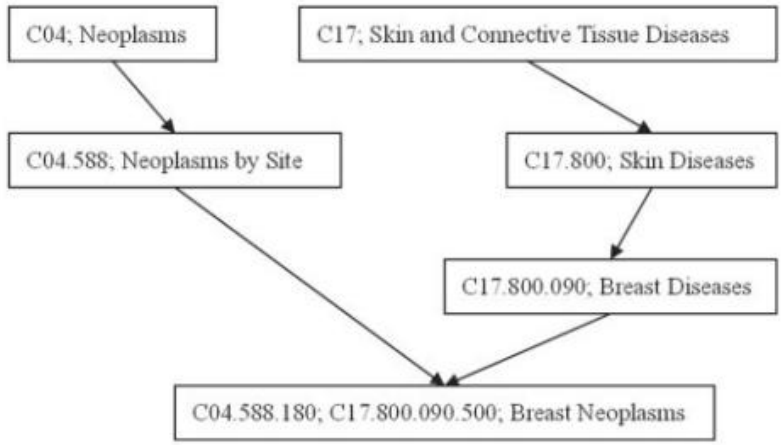
$$\mathbf{V} = \{V^{DR}, V^{DI}, V^{PR}\}$$

$$\mathbf{A} = \{A^{DR}, A^{DI}\}$$

$$\mathbf{E} = \{E^{DD}, E^{DP}, E^{PD}\}$$

$$\mathbf{A}^{DR} \in \mathbb{R}^{N \times K}$$

$$\mathbf{A}^{DI} \in \mathbb{R}^{M \times M}$$



Adr：药物特征矩阵 {0,1}
由SMILES得到化学描述符，利用RDKit判断特定化学结构的存在。

Adi：疾病相似度矩阵 [0,1]
概率矩阵，
由MeSH树结构通过大量计算得到。

As an example, the DV value of 'breast neoplasms' is 1.0 (breast neoplasms) + 0.5 (breast diseases) + 0.5 (neoplasms by site) + 0.5 × 0.5 (neoplasms) + 0.5 × 0.5 (skin diseases) + 0.5 × 0.5 × 0.5 (skin and connective tissue diseases) = 2.6250.

生物知识提取

MeSH：医学主题词同义词库
利用医学主题描述符提取疾病的生物学信息，
使用DAG描述每种疾病。

$$DAG_{V_a^{DI}} = (V_a^{DI}, F(V_a^{DI}), E(V_a^{DI}))$$

$$\begin{cases} D_{V_a^{DI}}(V_t^{DI}) = 1 \text{ if } V_a^{DI} = V_t^{DI} \\ D_{V_a^{DI}}(V_t^{DI}) = \max\left\{\gamma \times D_{V_a^{DI}}(V_t^{DI'}) \mid V_t^{DI'} \in \text{children of } V_t^{DI}\right\} \text{ if } V_a^{DI} \neq V_t^{DI} \end{cases}$$

$$DV(V_a^{DI}) = \sum_{V_t^{DI} \in F(V_a^{DI})} D_{V_a^{DI}}(V_t^{DI})$$

即a疾病的语义值

$$\text{Sim}(V_a^{DI}, V_b^{DI}) = \frac{\sum_{V_t^{DI} \in F(V_a^{DI}) \cap F(V_b^{DI})} \left(D_{V_a^{DI}}(V_t^{DI}) + D_{V_b^{DI}}(V_t^{DI})\right)}{DV(V_a^{DI}) + DV(V_b^{DI})}$$

$$A_a^{DI} = [\text{Sim}(V_a^{DI}, V_b^{DI})] \ (1 \leq b \leq M).$$

即a疾病与b疾病的语义值相似度

自编码器

优点：解决原始数据的冗余和稀疏问题，避免过拟合，保留主要特征，提高泛化能力。

编码器：把高维度的输入x 编码成低维度的隐变量z

$$H^{DR} = \sigma\left(WA\left(V^{DR}\right) + b\right)$$

解码器：把编码过后的输入z 解码为高维度的x ‘

$$\left(A^{DR}\right)' = \sigma\left(W'H^{DR} + b'\right)$$

优化目标：解码器的输出能够完美地或者近似恢复出原来的输入，称为重建误差函数。

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left\|A\left(V^{DR}\right)_i - A\left(V^{DR}\right)'_i\right\|^2$$

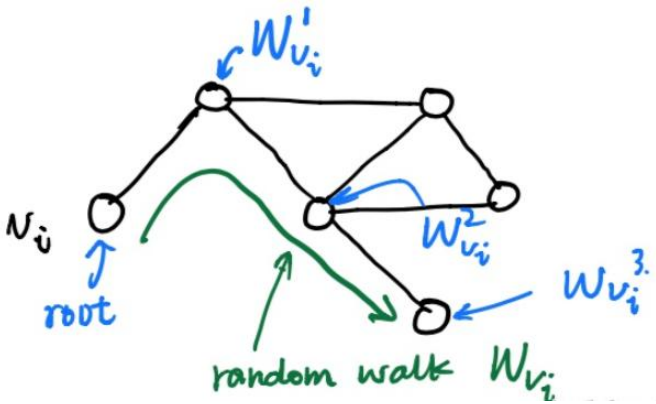
生物信息提取向量： $\mathbf{H} = [H^{DR}, H^{DI}]$

异构网络表示

$$\mathbf{V} = \{V^{DR}, V^{DI}, V^{PR}\} \qquad \mathbf{E} = \{E^{DD}, E^{DP}, E^{PD}\}$$

deepWalk算法主要分为随机游走和生成表示向量两个部分。

随机游走(random walk)，即在网络上不断重复地随机选择游走路径，最终形成一条贯穿网络的路径。
优点：局部游走，多个随机游走同时进行，减少采样时间。



$$Pr(w_n | w_0, w_1, \ldots, w_n - 1)$$

优化目标就是在012已知，求下一个是3的概率。

DeepWalk

$$\mathbf{V} = \{V^{DR}, V^{DI}, V^{PR}\} \qquad \mathbf{E} = \{E^{DD}, E^{DP}, E^{PD}\}$$

输入：多对节点

随机游走序列定义： from v0 to vi−1 (1 ≤ i ≤ |V|) is denoted as {v0 , v1 , · · · , vi−1}

已知随机游走序列，预测下一个节点是vi的概率：

$$\Pr\left(v_i | \left(v_0, \cdots, v_{i-1}\right)\right)$$

但节点本身很难计算，于是引入映射函数；

目标：得到每个节点的向量表示。

$$\Pr\left(v_i | \left(\Phi\left(v_0\right), \cdots, \Phi\left(v_{i-1}\right)\right)\right)$$

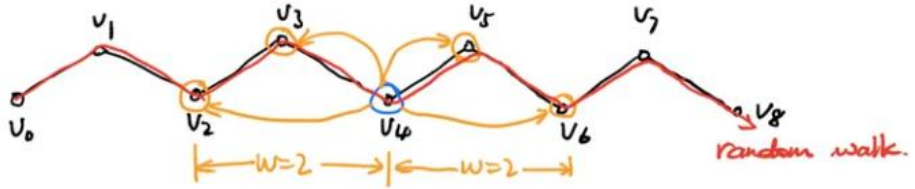
学习网络的隐藏信息，能够将图中的节点表示为一个包含潜在信息的向量。

异构网络提取向量：$Q = \Phi(\{V^{DR}, V^{DI}\})$

skip-gram模型：

1.不使用上下文预测缺失词，而使用缺失词预测上下文。

2.同时考虑左边窗口和右边窗口。



$\Pr(\{v_2, v_3, v_5, v_6\} | \Phi(v_4))$

rw 中的观 v_2, v_3, v_5, v_6 给定 v_4

$$\underset{\Phi}{\text{minimize}} \; -\log \Pr(\{v_{i-w}, \cdots, v_{i-1}, v_{i+1}, \cdots, v_{i+w}\})$$

$$|\Phi\left(v_i\right)) = \prod_{\substack{j=i-w, \\ j \neq i}}^{i+w} \Pr\left(v_j | \Phi(i)\right) \qquad (13)$$

$$P(w_{t+j} | w_t) = \frac{\exp(v_{w_{t+j}} \cdot v_{w_t})}{\sum_{i=1}^{V} \exp(v_i \cdot v_{w_t})}$$

DeepWalk: online learning of social representations

期刊：SIGKDD（ACM Special Interest Group on Knowledge Discovery in Data） 日期：2014

石溪大学

DeepWalk：社会表征的在线学习

算法流程

$\mathbf{H} = [H^{DR}, H^{DI}]$ $\mathbf{Q} = \Phi(\{V^{DR}, V^{DI}\})$

分类器：随机森林分类器
输入：药物疾病对表示向量
输出：P矩阵表示药物疾病关联的预测结果。{0,1}

Algorithm 1: The complete procedure of HINGRL.

Input: graph $HG(V, A, E)$.
representation sizes: d_1, d_2
the number of random walks: n
random walk length k
context size: w
the number of trees: t
Output: the relationships matrix $\mathbf{P} \in \mathbb{R}^{E^{DD}}$ of node v_i and node v_j, $v_i, v_j \in V$
1: Initialization: $\mathbf{P}$
2: Calculate the attribute similarity information of drugs $A(V^{DR})$
3: Calculate the attribute similarity information of diseases $A(V^{DI})$
4: Dimensionality reduction for $A(V^{DR})$ and $A(V^{DI})$
5: $H^{DR} = \textbf{\textit{AutoEncoder}}(A(V^{DR}), d_1)$
6: $H^{DI} = \textbf{\textit{AutoEncoder}}(A(V^{DI}), d_1)$
7: $\mathbf{H} = \begin{bmatrix} H^{DR} \\ H^{DI} \end{bmatrix}$
8: Learned the network representation of nodes
9: $\mathbf{Q} = \textbf{\textit{DeepWalk}}(E, d_2, n, k, w)$
10: Trained the prediction model by RF classifier
11: **for each** $e_{ij} =< v_i, v_j > \in E^{DD}$ do
12: the features matrix of nodes $\mathbf{X} = [\mathbf{H}(V) \ \mathbf{Q}(V)]$
13: $\mathbf{P} = \textbf{\textit{Random Forest Classifier}}([\ \mathbf{X}(v_i) \quad \mathbf{X}(v_j)\ ], t)$
14: **end for**
15: **Predicted unknown drug-disease associations in P**

实验

Dataset	Methods	AUC	AUPR	MCC	F1-score		
					Precision	Recall	F1-score
B-dataset	deepDR	0.8205	0.8043	0.2987	0.8814	0.2345	0.3704
	DTINet	0.8324	0.8472	0.2994	**0.9710**	0.1783	0.3012
	LAGCN	0.8790	0.1448	0.1917	0.0689	0.6931	0.1253
	HINGRL	**0.8835**	**0.8768**	**0.6012**	0.7971	**0.8063**	**0.8017**
F-dataset	deepDR	0.8553	0.8871	0.5609	0.9564	0.5241	0.6762
	DTINet	0.8220	0.8721	0.2081	**1.0000**	0.0841	0.1545
	LAGCN	0.8462	0.0068	0.0542	0.0058	0.6653	0.0115
	HINGRL	**0.9363**	**0.9446**	**0.7340**	0.8868	**0.8402**	**0.8625**

Feature	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision	Recall	F1-score
HINGRL-A	83.06 ± 0.55	82.20 ± 0.53	50.33 ± 1.21	74.58 ± 0.59	76.33 ± 1.10	75.44 ± 0.67
HINGRL-B	87.65 ± 0.45	86.68 ± 0.55	58.94 ± 1.06	79.38 ± 0.34	79.60 ± 1.21	79.49 ± 0.65
HINGRL	**88.35 ± 0.41**	**87.68 ± 0.51**	**60.12 ± 1.02**	**79.71 ± 0.53**	**80.63 ± 1.33**	**80.17 ± 0.62**

Classifier	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision	Recall	F1-score
Gaussian NB	74.94 ± 0.71	71.65 ± 1.35	38.33 ± 1.32	69.07 ± 0.79	69.41 ± 1.00	69.24 ± 0.66
SVM	78.04 ± 0.72	76.80 ± 0.79	42.19 ± 1.48	70.83 ± 0.68	71.73 ± 1.39	71.27 ± 0.86
LR	78.69 ± 0.69	77.73 ± 0.56	42.75 ± 1.54	70.94 ± 0.82	72.41 ± 1.14	71.66 ± 0.79
KNN	80.17 ± 0.75	76.05 ± 1.01	45.19 ± 1.10	66.65 ± 0.57	**86.38 ± 0.74**	75.25 ± 0.44
RF	**88.35 ± 0.41**	**87.68 ± 0.51**	**60.12 ± 1.02**	**79.71 ± 0.53**	80.63 ± 1.33	**80.17 ± 0.62**

B\F dataset的 AUC 都明显优于现有模型。
对于HINGRL的强壮型：
1.丰富的生物信息表示；
2.使用了RF分类，可处理大量高维输入数据。

对于LAGCN，HIN的稀疏性导致性能不理想，而HINGRL通过图嵌入减少了影响。

H-A仅拥有生物学知识
H-B在A的基础上整合药物疾病网络。
仅依靠药物和疾病的生物学知识可能不足以实现药物重新定位的良好性能，蛋白质相关的关联从拓扑的角度丰富了异质信息。

总结

HINGRL提出基于HIN的模型，用于图表示学习预测潜在的药物-疾病关联。
首先将蛋白质整合到原始的药物-疾病关联网络中，利用网络拓扑和生物知识的角度捕捉药物和疾病的目标特征。用 RF 分类器完成其预测任务。
实验结果表明，HINGRL在准确性和鲁棒性方面比最先进的药物重新定位算法具有更好的性能。

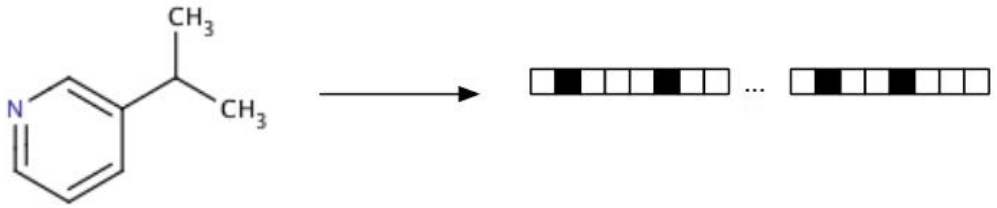
FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction

期刊：Briefings in Bioinformatics　　日期：2022　　分区：二区

华南理工大学

FP-GNN:用于增强分子性质预测的多功能深度学习架构

问题与方法



分子指纹是用于表示分子结构的一种方法，通过将分子转换为一种特殊的编码或特征向量，用于描述其结构、性质或相似性。

背景：

分子表征可以分为三大类：

分子描述符、分子指纹（ML）和分子图（GNN）。

问题：

1.对于传统ML很难从大量指纹特征中找到最重要的，而GNN面临建模数据集不足的限制。

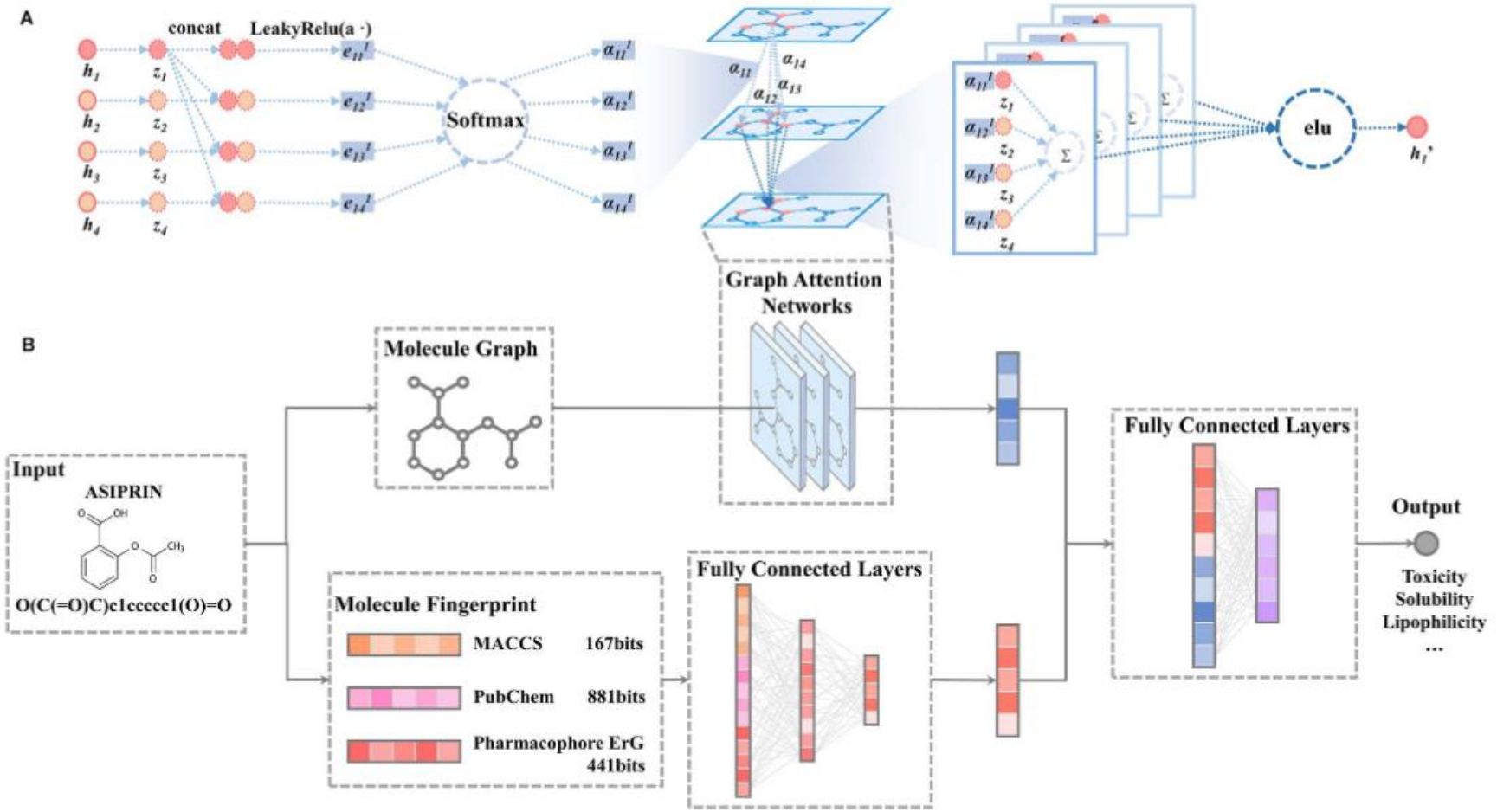
2.大多数研究称GNN优于ML，少数声称反之。

方法：

1.假设图和指纹捕获的信息不同甚至互补。

2.FP-GNN结合二者的混合分子表示。

整体框架


A
concat
LeakyRelu(a ·)
Softmax
elu
Graph Attention Networks
B
Molecule Graph
Input
ASIPRIN
O(C(=O)C)c1ccccc1(O)=O
Molecule Fingerprint
MACCS 167bits
PubChem 881bits
Pharmacophore ErG 441bits
Fully Connected Layers
Fully Connected Layers
Output
Toxicity
Solubility
Lipophilicity
...

带注意力机制的GNN

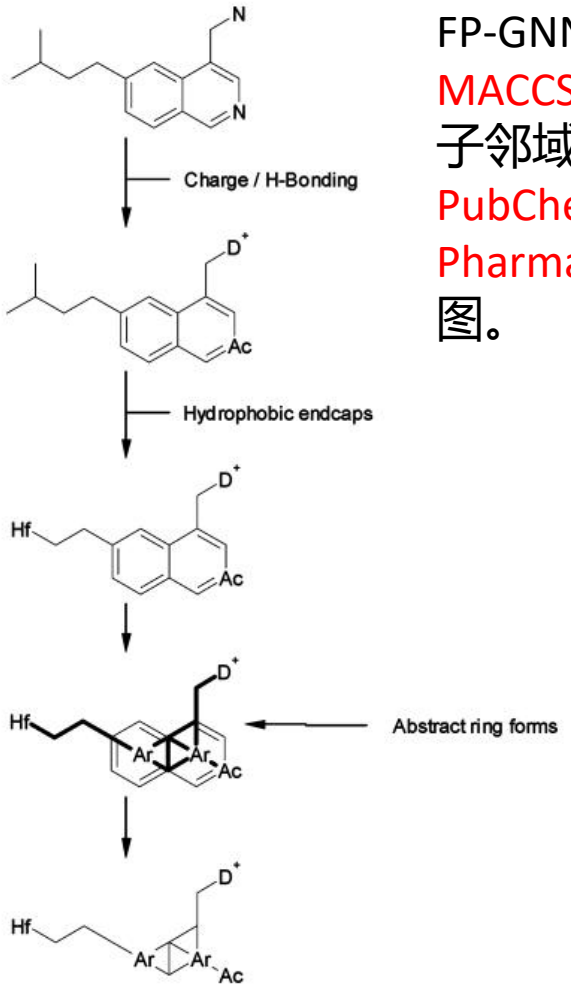
$$e_{ij} = LeakyRelu\left(a \cdot \left[W_1 h_i \| W_1 h_j\right]\right)$$

$$\alpha_{ij} = softmax\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in N(j)} \exp\left(e_{ik}\right)}$$

$$h_i' = elu\left(\sum_{j \in N(i)} \alpha_{ij} W_1 h_j\right)$$

$$h_i' = elu\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in N(i)} \alpha_{ij}^{k} W_1^{k} h_j\right)$$

$$H = \left(\frac{1}{N}\sum_{i=1}^{N} h_i'\right)$$



分子指纹

FP-GNN使用三种互补指纹，互补并全息表达分子特征。

MACCS：大多数原子特性、键特性和不同拓扑分离的原子邻域

PubChem：覆盖化学结构

Pharmacophore ErG：将样品结构的化学图转换为简化图。

$$FP = \left(FP_{PubChem} \| FP_{MACCS} \| FP_{Pharmacophore\ ErG}\right)$$

$$V' = W_2 \cdot FP + b$$

从两条路径收到的结果组合在一起并导入到完全连接的层中以产生最终输出。

D:氢键给体;Ac:氢键受体;Hf:疏水性基团;Ar:芳香环系;+/-:正负电荷。

Dataset	Split type	Metric	MoleculeNet (Graph) [27]	Chemprop (optimized) [20]	Attentive FP [54]	HRGCN+ [54]	XGBoost [54]	FP-GNN
BACE	Random	ROC–AUC		**0.898**	0.876	0.891	0.889	0.881
	Scaffold	ROC–AUC	0.806 (Weave)	0.857				**0.860**
HIV	Random	ROC–AUC		**0.827**	0.822	0.824	0.816	0.825
	Scaffold	ROC–AUC	0.763 (GC)	0.794				**0.824**
MUV	Random	PRC–AUC	**0.109 (Weave)**	0.053	0.038	0.082	0.068	0.090
Tox21	Random	ROC–AUC	0.829 (GC)	**0.854**	0.852	0.848	0.836	0.815
BBBP	Random	ROC–AUC		0.917	0.887	0.926	0.926	**0.935**
	Scaffold	ROC–AUC	0.690 (GC)	0.886				**0.916**
ClinTox	Random	ROC–AUC	0.832 (Weave)	0.897	0.904	0.899	**0.911**	0.840
SIDER	Random	ROC–AUC	0.638 (GC)	0.658	0.623	0.641	0.642	**0.661**
PDBbind-C	Random	RMSE		1.910				**1.876**
PDBbind-F	Random	RMSE		**1.286**				1.296
PDBbind-R	Random	RMSE		**1.338**				1.349
FreeSolv	Random	RMSE	1.150 (MPNN)	1.009	1.091	0.926	1.025	**0.905**
ESOL	Random	RMSE	0.580 (MPNN)	0.587	0.587	**0.563**	0.582	0.675
Lipophilicity	Random	RMSE	0.655 (GC)	0.563	**0.553**	0.603	0.574	0.625

Cell lines	Classification	Compounds	Task metric	Attentive FP [53]	GAT [53]	GCN [53]	MPNN [53]	XGBoost [53]	FP-GNN
MDA-MB-453	HER-2+[a]	440	ROC–AUC	0.872	0.812	0.866	0.715	0.810	**0.886**
SK-BR-3	HER-2+	2026	ROC–AUC	0.805	0.840	0.839	0.760	0.848	**0.852**
MDA-MB-435	HER-2+	3030	ROC–AUC	0.824	0.830	**0.858**	0.749	0.853	0.820
T-47D	Luminal A[b]	3135	ROC–AUC	0.812	0.763	0.819	0.751	0.821	**0.846**
MCF-7	Luminal A	29 378	ROC–AUC	0.845	0.800	0.833	0.843	0.826	**0.866**
MDA-MB-361	Luminal B[c]	367	ROC–AUC	0.938	0.896	0.955	0.972	**0.976**	0.905
BT-474	Luminal B	811	ROC–AUC	0.787	0.657	0.866	0.847	0.827	**0.868**
BT-20	TNBC[d]	292	ROC–AUC	0.735	0.721	0.740	0.784	0.740	**0.887**
BT-549	TNBC	1182	ROC–AUC	0.630	0.710	0.669	0.634	0.651	**0.807**
HS-578 T	TNBC	469	ROC–AUC	**0.830**	0.758	0.636	0.665	0.753	0.770
MDA-MB-231	TNBC	11 202	ROC–AUC	**0.870**	0.770	0.859	0.850	0.865	0.866
MDA-MB-468	TNBC	1986	ROC–AUC	0.875	0.875	0.887	0.858	**0.896**	0.888
Bcap37	TNBC	275	ROC–AUC	**0.858**	0.767	0.693	0.807	0.744	0.779
HBL-100	Normal cell line	316	ROC–AUC	0.645	0.641	0.658	0.701	0.776	**0.850**
Average				0.809	0.774	0.798	0.781	0.813	**0.849**

实验

FP-GNN
在8个学习任务中的3个上表现最好。结果都表明在预测分子方面是稳定的。

13 种乳腺癌细胞系和 1 种正常乳腺细胞系。
FP-GNN 在 14 个细胞系中的 8 个上表现最佳。

总结

FP-GNN 首先将基于分子图的图注意力网络和基于混合分子指纹的人工神经网络结合起来，以生成更全面的分子表示。在大量的经典数据集实验中，表现具有很强的竞争力。