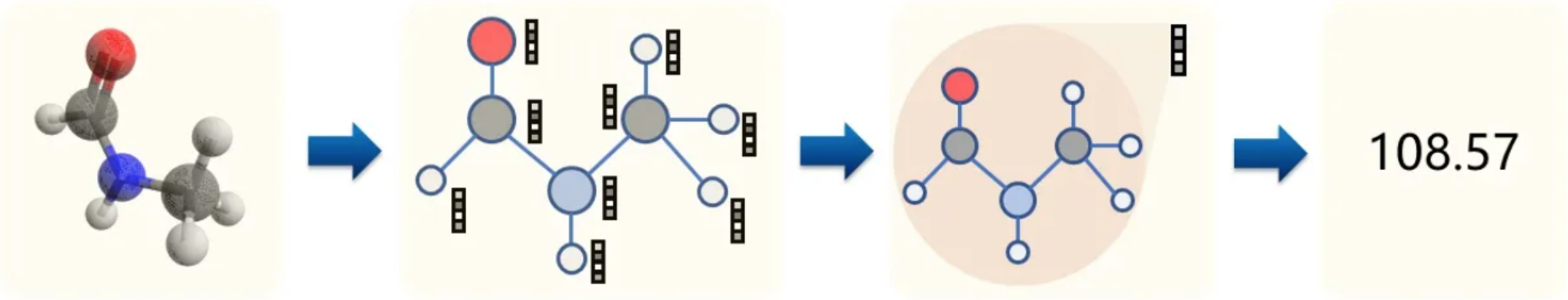
分子属性预测实战任务

GNN 分子属性预测流程

Pipeline



分子图（由SMILE获得）　　原子级别表征　　图级别表征　　预测

思考：Molecular property prediction challenges

1.Self-supervised learning methods 常见自监督工作中自监督任务对分子3D信息学习依旧不充分。

2. Graph convolution methods with experience 通过一些向模型注入专业知识的方法，引入经验数据和专家知识能加强模型的表征，但目前的工作没有很好应用。

3. 1D, 2D, and 3D data fusion and selection methods 如何确定合适的数据类型（或多种类型的最佳组合）仍然是一个悬而未决的问题。

4. 真实实验数据不足。 迁移学习、多任务学习和元学习都被用于解决某些属性的实验数据不足的问题。

5.The interpretability of DL models 与图像处理中的传统任务不同，大多数分子相关任务都是高度专业化的，需要化学专家来分析潜在机制，例如分子亚结构的作用。分子相关任务的这些特征与DL模型的“黑箱”性质有些矛盾。因此，改进DL模型的解释始终是重要的。

实战任务：尝试在分子数据集（TOX21）上进行分子属性预测任务

目标是预测给定分子的目标毒性（graph level 分类任务）

数据集：可以通过深度学习框架 DeepChem加载TOX21数据集，也可直接使用发在群里的CSV数据集文件。可以使用RDkit 转换SMILES 为分子图。

指标：ROC-AUC

数据集分割：random splitting。 training, validation , test 为8比1比1

方法：使用基于图的方法。允许使用别人的预训练模型，但是推荐自己从头训练。可以参考文章MoleculeNet: a benchmark for molecular machine learning

(此工作预计两到三周完成，后续在此基础上尝试解决一个前页提出的挑战)

Tox21

The “Toxicology in the 21st Century” (Tox21) initiative created a public database measuring toxicity of compounds, which has been used in the 2014 Tox21 Data Challenge. This dataset contains qualitative toxicity measurements for 8k compounds on 12 different targets, including nuclear receptors and stress response pathways.

Random splitting is recommended for this dataset.

The raw data csv file contains columns below:“smiles” - SMILES representation of the molecular structure“NR-XXX” - Nuclear receptor signaling bioassays results“SR-XXX” - Stress response bioassays results

please refer to https://tripod.nih.gov/tox21/challenge/data.jsp for details.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	NR-AR	NR-AR-LE	NR-AhR	NR-Arom	NR-ER	NR-ER-LB	NR-PPAR	SR-ARE	SR-ATAD5	SR-HSE	SR-MMP	SR-p53	mol_id	smiles					
2	0	0	1			0	0	1	0	0	0	0	TOX3021	CCOc1ccc2nc(S(N)(=O)=O)sc2c1					
3	0	0	0	0	0	0	0		0		0	0	TOX3020	CCN1C(=O)NC(c2ccccc2)C1=O					
4								0		0			TOX3024	CC[C@]1(O)CC[C@H]2[C@@H]3CCC4=CCCC[C@@H]4[C@					
5	0	0	0	0	0	0	0		0		0	0	TOX3027	CCCN(CC)C(CC)C(=O)Nc1c(C)cccc1C					
6	0	0		0	0	0	0	0	0	0		0	TOX3028	CC(O)(P(=O)(O)O)P(=O)(O)O					
7	0	0	0	0	0	0	0		0	0	0	0	TOX5110	CC(C)(C)OOC(C)(C)CCC(C)(C)OOC(C)(C)C					
8	0	0	0	0	0	0	0	0	0	0	0	0	TOX6619	O=S(=O)(Cl)c1ccccc1					
9	0		0		1			1	0	1	0	1	TOX25232	O=C(O)Cc1cc(I)c(Oc2ccc(O)c(I)c2)c(I)c1					
10	0	0	0	0	0	0	0	0	0	0		0	TOX22514	OC[C@H](O)[C@@H](O)[C@H](O)CO					
11								0		0			TOX22517	CCCCCCCC(=O)[O-].CCCCCCCC(=O)[O-].[Zn+2]					
12	0	0	0		0	0	0		0			0	TOX25236	NC(=O)c1ccc[n+]([C@@H]2O[C@H](COP(=O)([O-])OP(=C					
13	0	0			0	0	0	0	0	0	0	0	TOX25238	O=c1[nH]c(=O)n([C@H]2C[C@H](O)[C@@H](CO)O2)cc1I					
14	0	0	0	0	0	0	0	0	0	0	0	0	TOX6612	CC(C)COC(=O)C(C)C					
15	0	0	0	0	0	0	0	0	0	0	0	0	TOX6615	C=C(C)C(=O)OCCOC(=O)C(=C)C					
16	0	0	0		0	0		1	0	0	0		TOX15748	Cl/C=C\C[N+]12CN3CN(CN(C3)C1)C2					
17	0	0	0	0	0	0	1	0	0	0	0	0	TOX14833	O=C([O-])Cc1cccc2ccccc12					
18	0	0	0		0					0			TOX26529	CCCCCCCCCCCOCC(O)CN					
19	0	0	0	0		0	0		0		0	0	TOX26524	CCN(CC)C(=O)c1cccnc1					