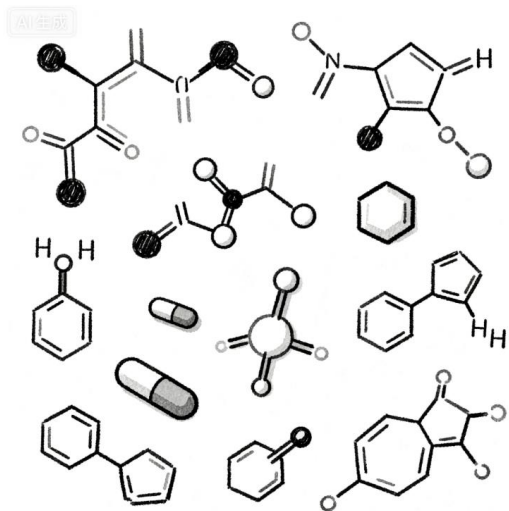


第七章 药物大模型

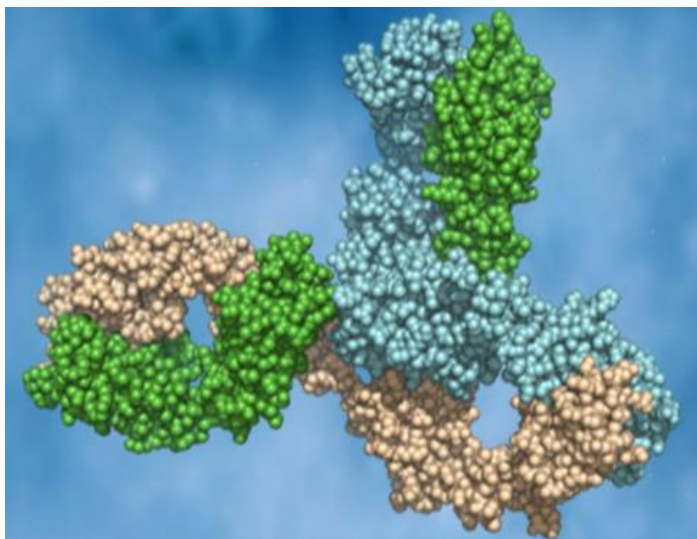
药物的主要分类

小分子药物



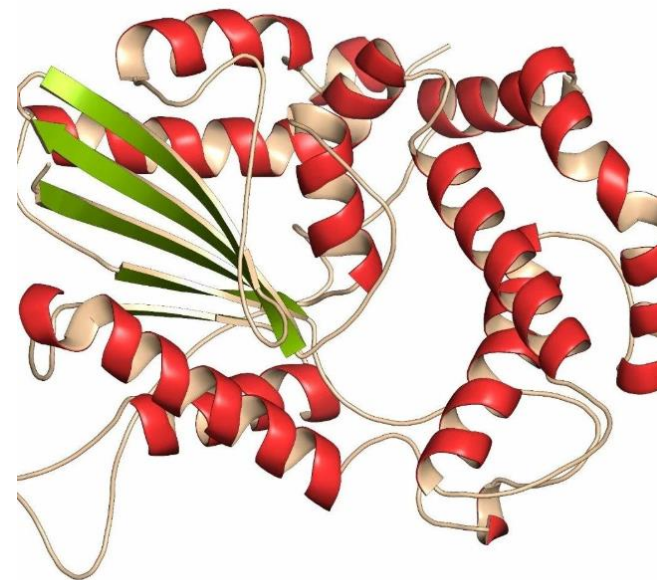
低分子量，结构明确，成本低

生物制剂/大分子



生物体系产生，分子量大、结构复杂

肽类药物



介于小分子和大蛋白之间，近年来逐渐受重视

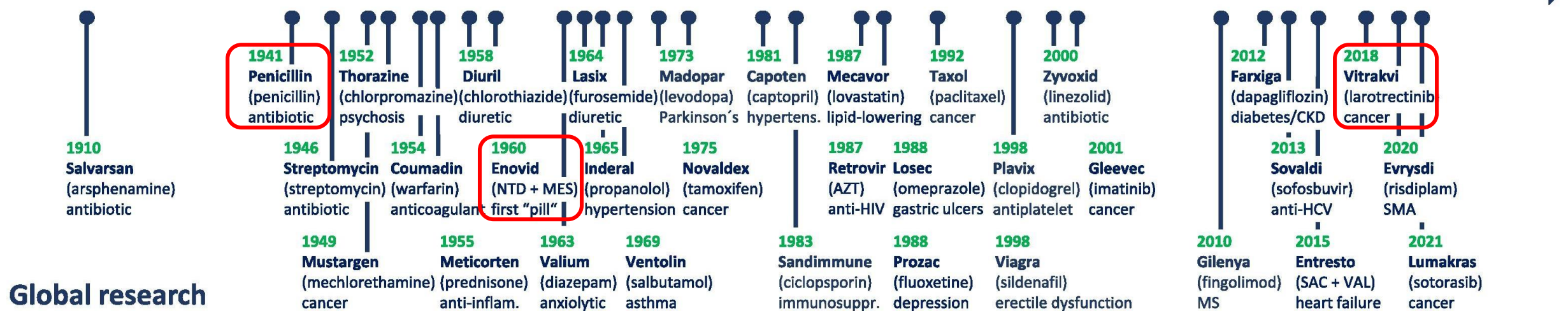
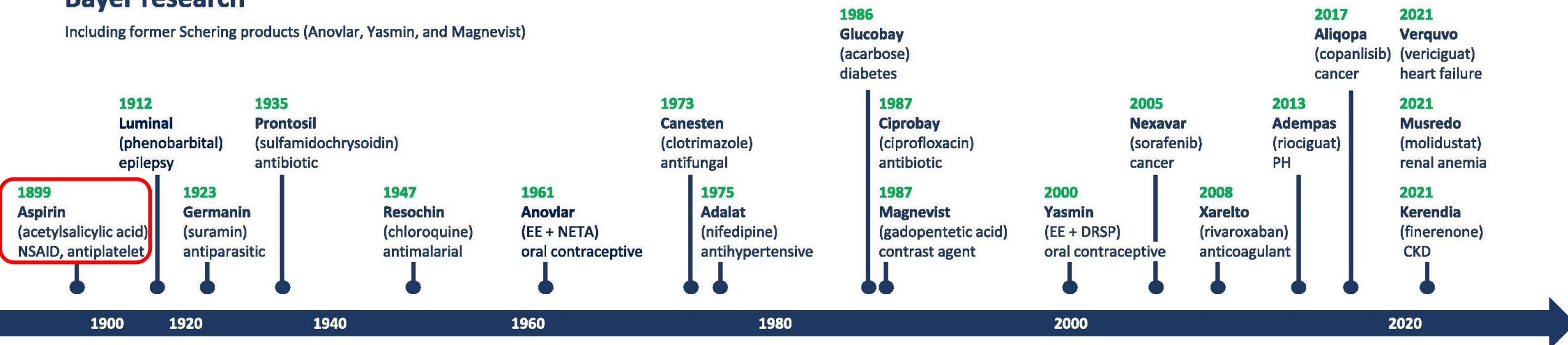
[肽类药物：现状、进展与未来展望 - 知乎](#)

小分子药物

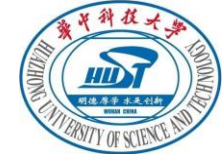
小分子作为生物医学研究中的化学探针至关重要。在过去一个世纪里，传统小分子药物一直是药物研究的主导。目前小分子药物仍**占临床与上市新药的大多数**，适合口服、合成成熟、成本与规模化优势明显，且**尤其适合基于算法与计算的药物设计**。

Bayer research

Including former Schering products (Anovlar, Yasmin, and Magnevist)



Global research



基于计算的小分子药物发现

计算机辅助药物设计 (CADD) : [计算机辅助药物设计教程 文档](#)

小分子特别适合机器学习 (ML) 方法 :

1. 化学结构模块化 : 可转换成计算机可读的格式
2. 数据集大 : 合成方法、理化特性和蛋白质-靶点相互作用的大量历史数据集早已广泛使用

使用计算方法进行小分子药物发现的早期尝试 :

1. QSAR (定量构效关系) : 基于分子描述符与统计/机器学习模型建立 “结构 → 活性” 的关系
适用于系列化合物的活性预测和构效关系分析。解释性较好，外推能力差
[基于构效关系模型的药物设计 \(QSAR\) 建立3d qsar模型-CSDN博客](#)
2. 药效团模型 : 抽象出小分子里对结合靶点关键的化学结构或特征，用于虚拟筛选等
[药效团模型构建 \(上篇\) - 知乎](#)
[Drug Design by Pharmacophore and Virtual Screening Approach](#). Pharmaceuticals (Basel), 2022, 15, 646.
3. **分子对接 (Molecular Docking)** : 将小分子虚拟地放入靶点结构中，评分结合位点适配度
用于虚拟筛选和结合构象预测，是CADD的主力工具。结构基础，评分函数与柔性处理仍有挑战
[分子对接 \(molecular docking\) 技术介绍 - 知乎](#)
4. **虚拟筛选 (Virtual Screening)** : 将上述方法结合用于百万甚至上亿的化合物库中进行优先级排序候选
[虚拟筛选基础知识 - 知乎](#)
[虚拟筛选 | DrugFlow Docs](#)

基于AI的小分子药物发现——分子表征

基于人工智能的药物设计 (AIDD) : 基于人工智能的药物设计(Artificial Intelligence-based Drug Design)

- SMILES编码** : 字符串数据, 简单易用; 一个分子可以对应多个SMILES; SMILES单个字符改变会产生无效分子
- SELFIES编码** : SMILES的改进版本, 一个分子唯一对应且鲁棒性100%
- 分子指纹** : 二进制或数值向量描述分子中亚结构的存在和缺失, 输入是SMILES编码
- InChi编码** : IUPAC开发的标准编码, 可唯一标识分子; 编码复杂, 不易直接用于学习
- 分子图** : RDKit将分子SMILES转换为2D分子图, 或用分子力场生成合理3D结构
- Encoders** : 深度学习模型如GNN、RNN、Transformer将分子结构编码成向量

SMILES和图编码对比: SMILES本身就是分子图深度优先遍历后得到的一维表示, 理论效果不会弱于图网络。因此实际一般会用SMILES或SELFIES

Imatinib Mesylate ($C_{30}H_{35}N_7O_4S$)

1D Representation

SMILES:

CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5.CS(=O)(=O)O

ECFP:

216545222: 2, 337875000: 1, 353395765: 2,
368350036: 1, 847957139: 1, 847961216: 2,
...
3275683399: 1, 3918336191: 2, 3975275337: 1]

MACCS:

000000000.....10100110101111111
1110

Mathematical Representation:

Differential geometry
Algebraic graph
Algebraic topology

2D Representation

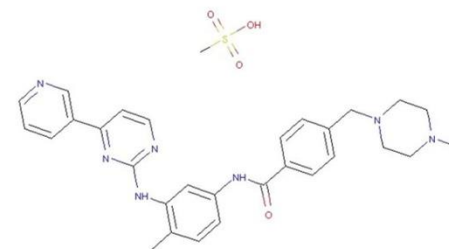
$$\begin{bmatrix} C & 0 & 1 & 0 & & 1 & 1 & 0 \\ C & 1 & 0 & 0 & \dots & 1 & 1 & 0 \\ C & 0 & 0 & 0 & & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & \vdots \\ O & 1 & 1 & 0 & & 0 & 0 & 1 \\ O & 1 & 1 & 0 & \dots & 0 & 0 & 0 \\ O & 0 & 0 & 0 & & 1 & 0 & 0 \end{bmatrix}$$

Adjacent Matrix

$$\begin{bmatrix} C & 1 & 0 & \dots \\ C & 1 & 0 & \dots \\ C & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ O & 0 & 1 & \dots \\ O & 0 & 1 & \dots \\ O & 0 & 1 & \dots \end{bmatrix}$$

Feature Matrix

2D Molecular Graph



2D Molecular Image

3D Representation

$$\begin{bmatrix} C & 0 & 1 & 0 & & 1 & 1 & 0 \\ C & 1 & 0 & 0 & \dots & 1 & 1 & 0 \\ C & 0 & 0 & 0 & & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots & \vdots \\ O & 1 & 1 & 0 & & 0 & 0 & 1 \\ O & 1 & 1 & 0 & \dots & 0 & 0 & 0 \\ O & 0 & 0 & 0 & & 1 & 0 & 0 \end{bmatrix}$$

Adjacent Matrix

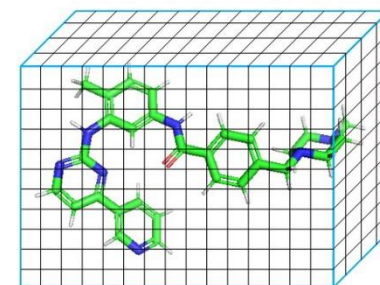
$$\begin{bmatrix} C & 6.16 & 0.32 & 0.78 \\ C & 7.85 & 0.67 & -0.90 \\ C & 7.12 & -0.56 & 1.58 \\ \vdots & \vdots & \vdots & \vdots \\ O & -0.83 & 0.25 & -1.04 \\ O & -6.61 & -0.26 & -0.56 \\ O & -1.83 & -0.23 & -0.11 \end{bmatrix}$$

3D Coordinates

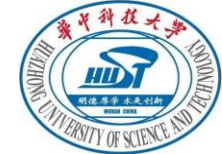
$$\begin{bmatrix} C & 1 & 0 & \dots \\ C & 1 & 0 & \dots \\ C & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ O & 0 & 1 & \dots \\ O & 0 & 1 & \dots \\ O & 0 & 1 & \dots \end{bmatrix}$$

Feature Matrix

3D Molecular Graph



3D Molecular Grid



基于AI的小分子药物发现——生成式模型微调

一般不会自己从头训练大模型，而是微调已有的预训练模型：

化学小分子预训练模型（CPM）

1. **基于序列的模型**：RNN/Transformer在SMILES上训练，生成新SMILES

将SMILES/SELFIES当作化学语言，token序列的建模与生成和文本生成高度类似，transformer的长距依赖与生成能力对复杂分子结构建模有优势

MolGPT：[JCIM2021 | MolGPT：基于Transformer-Decoder的分子生成-CSDN博客](#)

MolGen：[ICLR 2024 | MolGen: 化学反馈引导的预训练分子生成 - 知乎](#)

2. 基于图的模型：直接在分子的图结构上训练并生成分子，提高合法性和化学合理性

3. VAE/GAN/强化学习RL/**扩散模型**：分别在潜在空间建模、多目标优化（如活性+合成可行性）、或高质量采样上各有所长

通用大语言模型

直接拿大参数模型大力出奇迹

MolGen参数量为355,578,041（**0.3B**），DeepSeek参数量为**671B**，即便是用于微调的DeepSeek蒸馏模型也有**8B**

示例：Transformer构建小分子药物生成模型

药物研发里，最核心的是新分子生成，可以是有条件生成（如根据目标靶点生成新的小分子），也可以是无条件生成（如先用某个应用场景的数据微调模型后再生成大批新分子）

1. 搭建简易的基于Transformer的小分子药物生成模型（train_selfies_transformer.py）

```
29 # -----
30 # Tokenizer for SELFIES
31 # -----
32 SELFIES_TOKEN_RE = re.compile(r'\[.*?\]')
33
34 > class SelfiesTokenizer: ...
109
110 # -----
111 # Dataset
112 # -----
113 > class SelfiesDataset(Dataset): ...
126
127 > def collate_fn_pad(batch, pad_id: int): ...
135
136 # -----
137 # Transformer LM model (autoregressive)
138 # -----
139 > class TransformerLM(nn.Module): ...
177
178 # -----
179 # Lightning Module
180 # -----
181 > class SelfiesLitModule(pl.LightningModule): ...
230
231 # -----
232 # Utilities: load data
233 # -----
234 > def load_selfies_file(path): ...
243
244 # -----
245 # Main training orchestration
246 # -----
247 > def main(args): ...
316
```

Tokenizer：分词器，如中文分词器把句子分为单个字或词，SELFIES字符串也可以按照一定规则拆分，拆分规则定义在SELFIES_TOKEN_RE变量里。如
[C][C][=C][Ring1][=Branch2] → [C] [C] [=C] [Ring1] [=Branch2]

Dataset：规定读取数据的方式和数据预处理

Model：这里用的是Transformer自回归语言模型，即一次预测一个词，并以前面预测过的词作为条件（写作文边写边参考前面的内容）

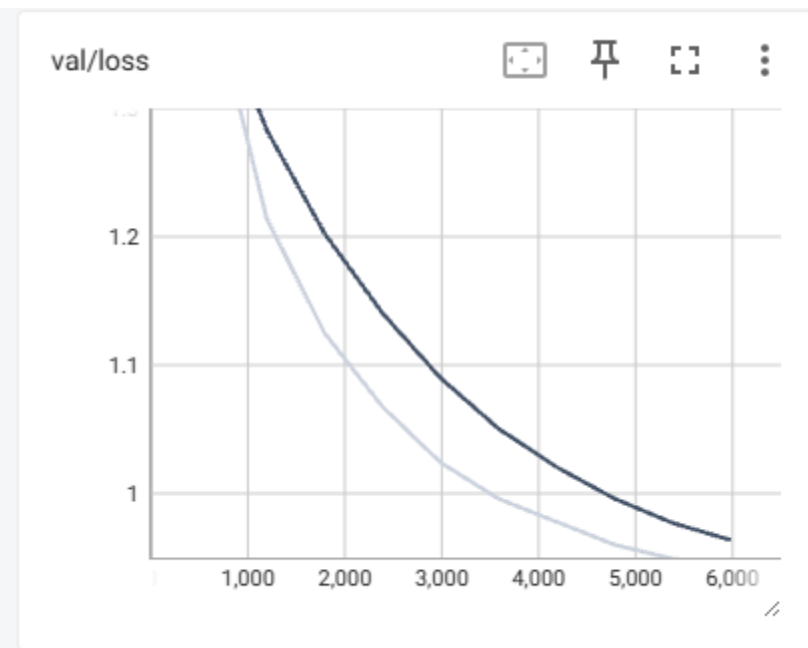
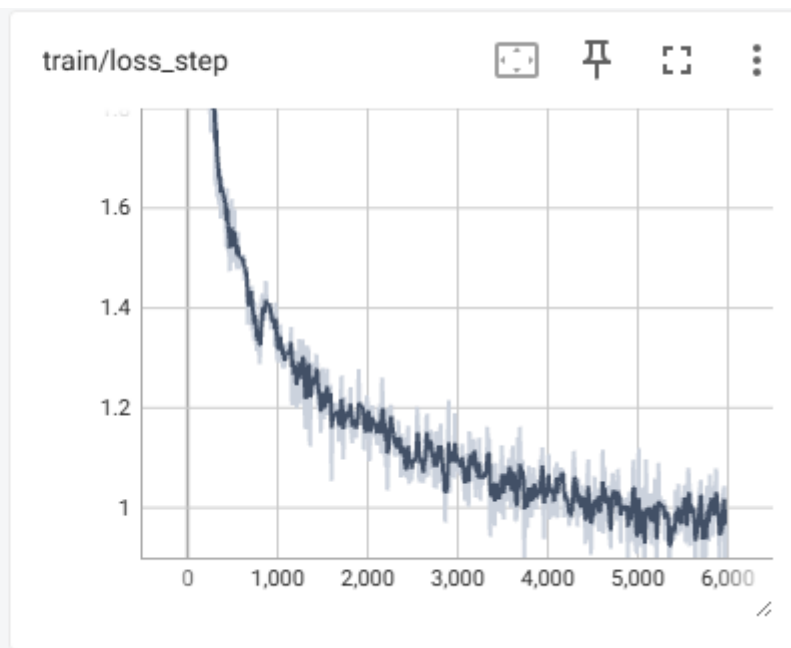
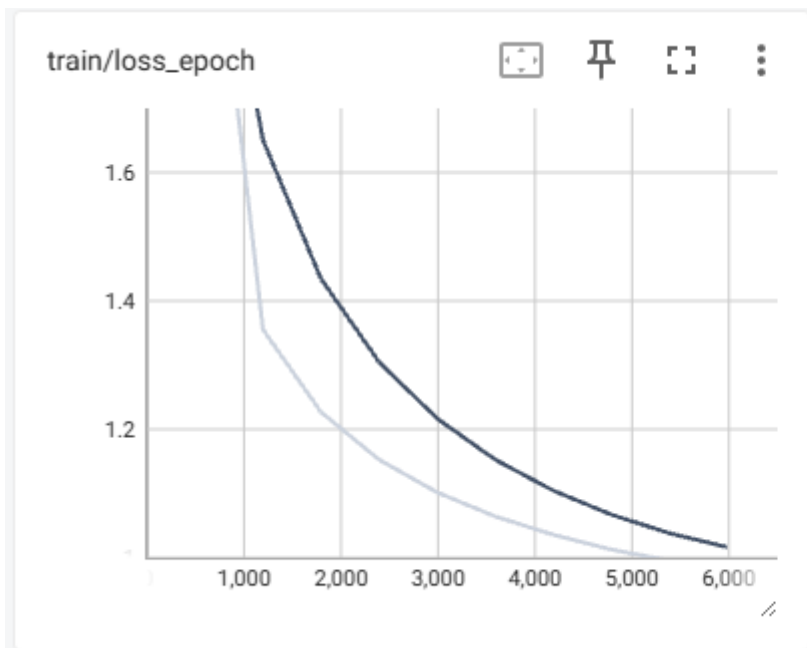
Lightning：Lightning Module管理模型

示例：Transformer构建小分子药物生成模型

2. 预训练模型训练结果

数据集：All_small_molecule_drugs_with_selfies.txt，整合了PubChem、ChEMBL、DrugBank的类药小分子的Selfies

训练10个epoch



示例：Transformer构建小分子药物生成模型

3. 预训练模型评估

MOSES：一个开源的基准测试平台，旨在标准化分子生成模型的训练、评估和比较。

[MOSES分子生成模型基准测试平台](#)

[Molecular Sets \(MOSES\): A Benchmarking Platform for Molecular Generation Models](#)

[GitHub - peteole/molecule_benchmarks](#)

```
=== Validity Metrics ===
Valid molecules: 1.000
Unique molecules: 1.000
Valid & unique: 1.000
Novel molecules: 1.000

=== Moses Metrics ===
Passing Moses filters: 0.330
SNN score: 0.223
Internal diversity (p=1): 0.898
Internal diversity (p=2): 0.859

=== Distribution Metrics ===
KL divergence score: 0.590
FCD score: 29.424
FCD (valid only): 29.424
```

	Name	Type	Params	Mode
0	model	TransformerLM	5.1 M	train
1	criterion	CrossEntropyLoss	0	train
5.1 M	Trainable params			
0	Non-trainable params			
5.1 M	Total params			
20.287	Total estimated model params size (MB)			
68	Modules in train mode			
0	Modules in eval mode			

可训练参数为5.1M

总结：生成的分子全部合法、不重复、新颖性强，但与训练集相似度低，可能过于偏离真实药物分布，且只有 33% 的分子通过MOSES的筛选规则

原因：模型小且简单，训练epoch少，采用的参数不合适（如temperature设为1导致结果过度发散）

示例：Transformer构建小分子药物生成模型

4. 微调预训练模型

用特定靶点的药物数据微调预训练模型，生成潜在的针对该靶点的新药

```

7 # 1. 加载小数据集
8 small_file = "target_selfies.txt"
9 > with open(small_file) as f: ...
11
12 # 用之前预训练保存的 tokenizer
13 tok = torch.load("selfies_tokenizer.pt", weights_only=False)
14 pad_id = tok.token_to_id("<pad>")
15 dataset = SelfiesDataset(small_data, tok, max_len=256)
16 > loader = DataLoader(dataset, batch_size=16, shuffle=True, ...
18
19 # 2. 加载预训练大模型
20 > model = SelfiesLitModule.load_from_checkpoint(...)
25
26 # 3. 用小数据继续训练（微调）
27 trainer = pl.Trainer(max_epochs=5, accelerator="gpu", devices=1)
28 trainer.fit(model, loader)
29
30 # 4. 简单生成函数（采样）
31 def generate(model, tok, max_len=100, temperature=1.0):
32     ... model.eval()
33     ... ids = [tok.token_to_id("<bos>")]
34 > ... for _ in range(max_len): ...
35     ... return tok.decode(ids)
36
37 # 5. 生成几个分子
38 for i in range(5):
39     ... print(generate(model, tok))

```

数据集target_selfies.txt为DrugBank里与CDK2 (P24941) 蛋白互作的所有药物小分子



```

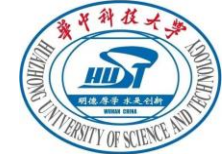
[C][O][C][=Branch1][C][=O][N][C][=N][C][Branch1][O][N][C][S][Branch1][C][C][=Branch1][C][=O][=O][=C][Branch1][=Bra
[H][C@@][Branch1][C][N][Branch2][Ring1][=N][C][C][=C][N][C][=C][C][=C][Ring1][#Branch2][Ring1][=Branch1][C]
[O][=C][Branch2][Ring1][#Branch1][N][S][C][=C][C][=N][C][=Branch1][C][=O][N][Ring1][#Branch1][C][Nd+3][=N][C][=C][
[F][C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1][C][=N][C][=N][C][=C][Branch1][Ring2][N][Ring1][Branch1
[N][C][=N][C][Branch1][O][N][C][C][=N][N][=N][C][C][C][Ring1][Branch1][=N][Ring1][=C]

```



下游验证：

1. Docking：计算机模拟结构对接，评估结合打分
2. 小分子合成 → 实验验证
3.



随堂小测 & 课外学习

- 作业：
 - 1.
 - 2.
- 参考资料：
 - 1. 学术论文：[Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the Bayer Chemical Research Laboratory](#). Drug Discov Today, 2022, 27, 1560-1574.
 - 2. 学术论文：[Molecular Sets \(MOSES\): A Benchmarking Platform for Molecular Generation Models](#). arXiv, 2018, arXiv:1811.12823.