本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

数据科学导论

Introduction to Data Science

Task2: 实验报告

陈恩红，黄振亚
Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html

助教：罗胤玻，李宁
ds_intro2026@163.com

实验（2026年12月08日）

- 两个实验方式
  - **实验方式：参加指定问题的数据比赛（组队）**
    - 参加课程推荐的比赛（推荐）
    - 自己寻找比赛

- 重点：大家在实践中熟悉和应用数据科学知识，锻炼团队合作能力，**只要在报告中叙述清楚、内容合理即可**
  - 学习：分析问题、解决问题、代码实践、团队协作、报告撰写
  - 项目组成员、任务分工和组织、个人总结收获

实验

- **大数据竞赛**：组队(1~2人)参加给定的比赛，最后将做题思路、结果以及比赛排名以报告形式提交
- **报告内容**
  - 比赛名称
  - 队伍名
  - 问题定义
  - 做题思路，模型设计
  - 比赛排名
  - 团队成员分工
  - 个人总结和感悟

组队要求

- 可以单挑，可以组队（1-2人），建议组队
- 组队成员
  - 课上同学
- 注明个人分工
- **组队节点：2026年10月20日**

实验报告评分要求

- 问题介绍与理解
- 团队协作：个人分工明确合理
- 实验过程：认真度、工作量、思路合理性
- 报告条理：是否条理清晰，内容充足
- 是否迟交，是否有抄袭
  - **需提交代码及运行说明**，提交的代码必须可运行，也必须**附有自己运行时的log文件**。如代码或log文件雷同判定为抄袭。
  - log文件要求记录代码运行的中间结果（如数据处理过程和训练时的loss值），要求每一条log都记录相应的时间。示例：

```
2024-09-22 19:10:49 - INFO - Epoch: 1
2024-09-22 19:10:49 - INFO - Learning Rate: 4e-05
2024-09-22 19:10:51 - INFO - Epoch: 1, Step: 0, Train Loss: 0.0102, Precsion: 0.0059, Recall: 0.7903, F1: 0.0117, A
2024-09-22 19:10:59 - INFO - Epoch: 1, Step: 100, Train Loss: 0.4072, Precsion: 0.0101, Recall: 0.7451, F1: 0.0199,
2024-09-22 19:11:07 - INFO - Epoch: 1, Step: 200, Train Loss: 0.1537, Precsion: 0.0215, Recall: 0.6254, F1: 0.0415,
2024-09-22 19:11:14 - INFO - Epoch: 1, Step: 300, Train Loss: 0.0834, Precsion: 0.0357, Recall: 0.4798, F1: 0.0664,
2024-09-22 19:11:22 - INFO - Epoch: 1, Step: 400, Train Loss: 0.0502, Precsion: 0.0539, Recall: 0.3615, F1: 0.0938,
```

实验报告内容要求细则

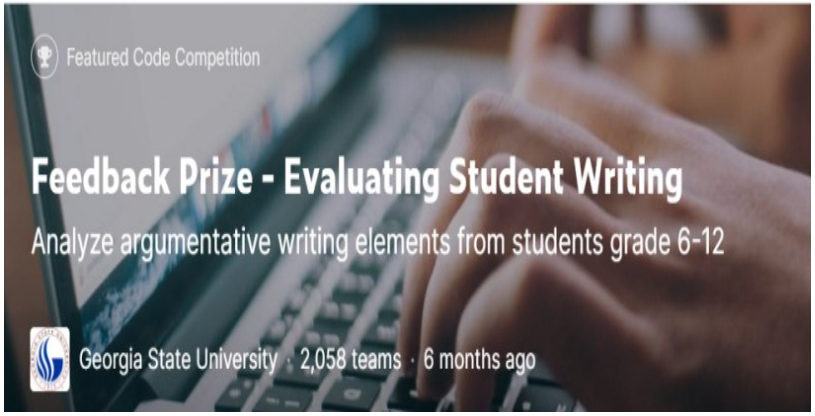
- 数据分析：对问题与数据的分析、特征的处理，需说明将原数据每一行转化为模型输入的完整处理过程。
- 模型：模型的选择依据以及模型原理说明、模型的输出形式。
- 训练：训练中是否以及如何调参、是否尝试并比较多种模型。
- loss：报告中需要讲述loss函数的含义（即为什么最小化这个函数就可以达到分类、回归预测的目的）。
- 最终模型的预测结果展示与输入输出样例展示。
- 模版不限

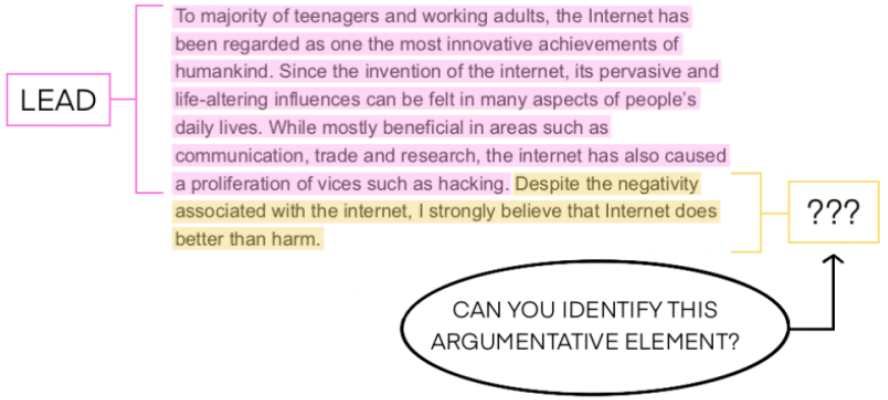
比赛平台

- 比赛平台-供了解
  - CCF BDCI
    - https://www.xir.cn/competition/races/BDCI2026/competition
  - 天池
    - https://tianchi.aliyun.com/competition
  - Kaggle
    - https://www.kaggle.com/competitions
  - 会议竞赛
    - KDD CUP （“大数据世界杯”、数据挖掘领域“奥运会”）
      - https://kdd2026.kdd.org/call-for-kdd-cup-proposals/
    - NeurIPS 2026 Competition Track
      - Scientific AI, Physics, and Engineering
      - Foundation Models, Reasoning, and AI Evaluation

比赛实例

- Kaggle 2022
  - Feedback Prize - Evaluating Student Writing
    - 背景：该比赛由佐治亚州立大学(GSU)和The Learning Agency Lab提出，专注于开发基于学习的工具和社会公益项目的科学。目前有很多自动写作反馈工具，但它们都有局限性，往往不能识别写作结构，比如论文中的导语、立场、论点、论据、反论据等文章元素。
    - 要求：将在6 -12年级学生写的文章中找出学生写作中的元素，更具体地说，你需要自动分割文本，并对议论文的结构元素进行分类。
    - 数据量：**15, 600 篇文章**

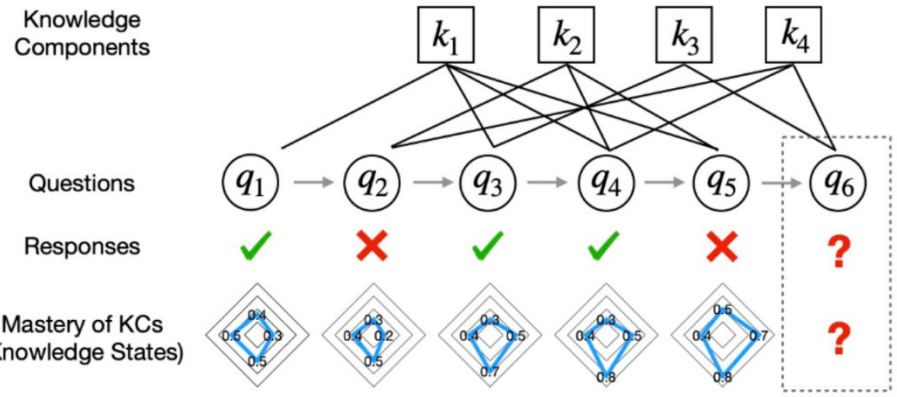




比赛实例

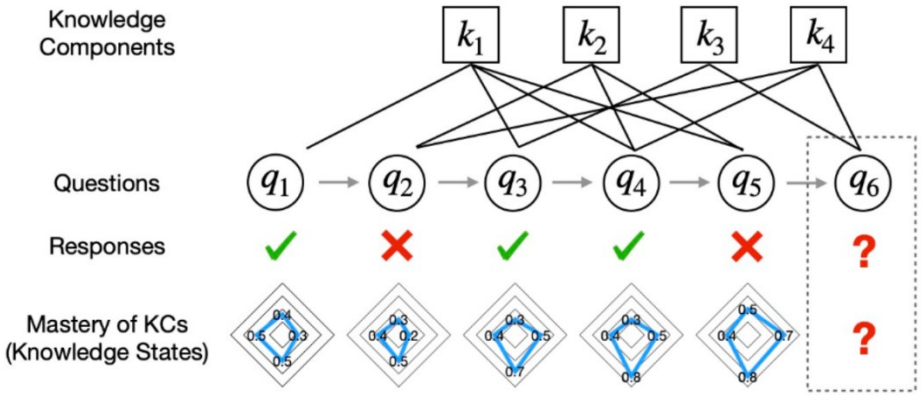
- AAAI2023 KT比赛
  - Global Knowledge Tracing Challenge
    - 背景：知识追踪（KT）是利用学生的历史学习交互数据来建模他们随时间变化的知识掌握情况，以便预测他们未来的答题表现的任务。这种预测能力可以潜在地为学生提供个性化的服务（比如推荐适合学生能力的试题），这对于构建下一代智能个性化教育至关重要。
    - 数据量：18,066名学生在7,652道题目上的5,549,635条答题记录
    - 中国科学技术大学、新加坡科技研究局A*STAR、网易等 50多支队伍参赛

排名	团队	学生答题表现预测(AUC)
1	中科大	0.8178
2	A*STAR	0.8167
3	网易	0.8166



比赛实例

- 解决方案 --- 数据方面
  - 除了最基本的问题-答案对，还使用了以下信息来增强交互表征
    - 问题类别
    - 问题本身内容长度
    - 答案的长度与相对长度
    - 问题相关的知识点，以及该知识点在整个层级知识点树中的位置
    - 历史 8 次互动和未来 8 次互动的间隔



比赛实例

- ICML 2024 数学自动推理比赛
  - Automated Math Reasoning
    - 背景：数学推理是人类智慧中最为高级的形式之一。人类开发了形式化的语言，用以严格描述数学问题并推导数学知识。机器学习领域的研究者们也在努力开发具有数学推理能力的神经模型，旨在达到与人类相似的推理水平。
    - 要求：
      - 数学证明自动形式化：给定一个问题陈述及其自然语言的证明，生成相应的形式化陈述和证明。
      - 数学证明自动去形式化：给定一个问题陈述的形式化陈述和证明，生成相应的自然语言问题陈述及其自然语言的证明。



比赛实例

- 数据示例

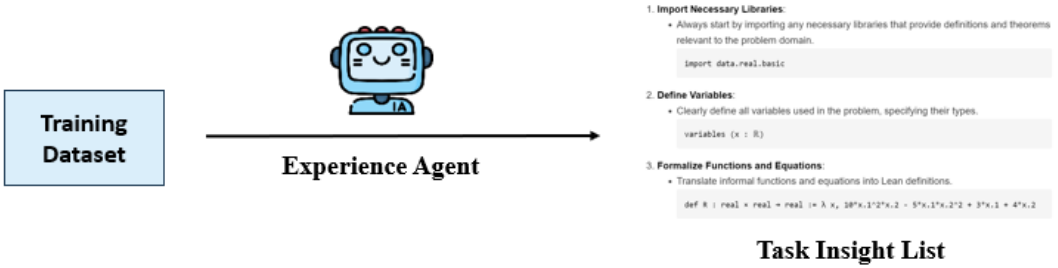
{
"**problem_name**": "correct_by_msg__ELEM_theorem_proving_1st_grade_15_round2",
"**informal_statement**": "If a two-digit number is formed with 4 in the tens place and 3 in the ones place, prove that the number is 43.",
"**informal_proof**": "We know that the number in the tens place represents tens and the number in the ones place represents ones. So, if 4 is in the tens place, it represents 4 * 10 = 40. If 3 is in the ones place, it represents 3. So the total number is 40 + 3 = 43.",
"**formal_proof**": "def ten (n : ℕ) : ℕ := n * 10\ndef one (n : ℕ) : ℕ := n\ntheorem two_digit_number : ten 4 + one 3 = 43 :=\nbegin\n have h1 : ten 4 = 40, by refl, \n have h2 : one 3 = 3, by refl, \n rw [h1, h2],\n exact add_comm 3 40 \nend"

}

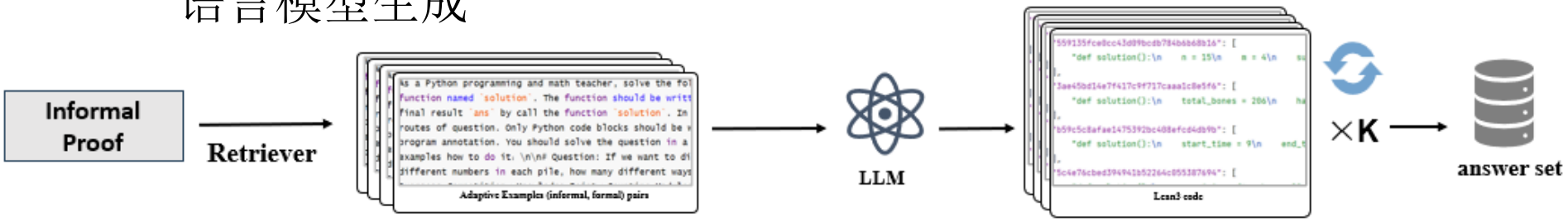
- 结果评估
  - Rouge-L，BLUE，Lean 3 代码通过率
  - Rouge-L，BLUE

比赛实例

- 解决方案
  - 1. 使用大语言模型（GPT-4）总结训练数据集中的经验和方法



  - 2. 在训练集中寻找与当前问题类似的例子，以此为示例辅助大语言模型生成



  - 3. 大语言模型生成多个推理路径后，根据验证结果（如检测代码是否通过编译、模型自验证等）和多数投票决定最终答案

天池介绍

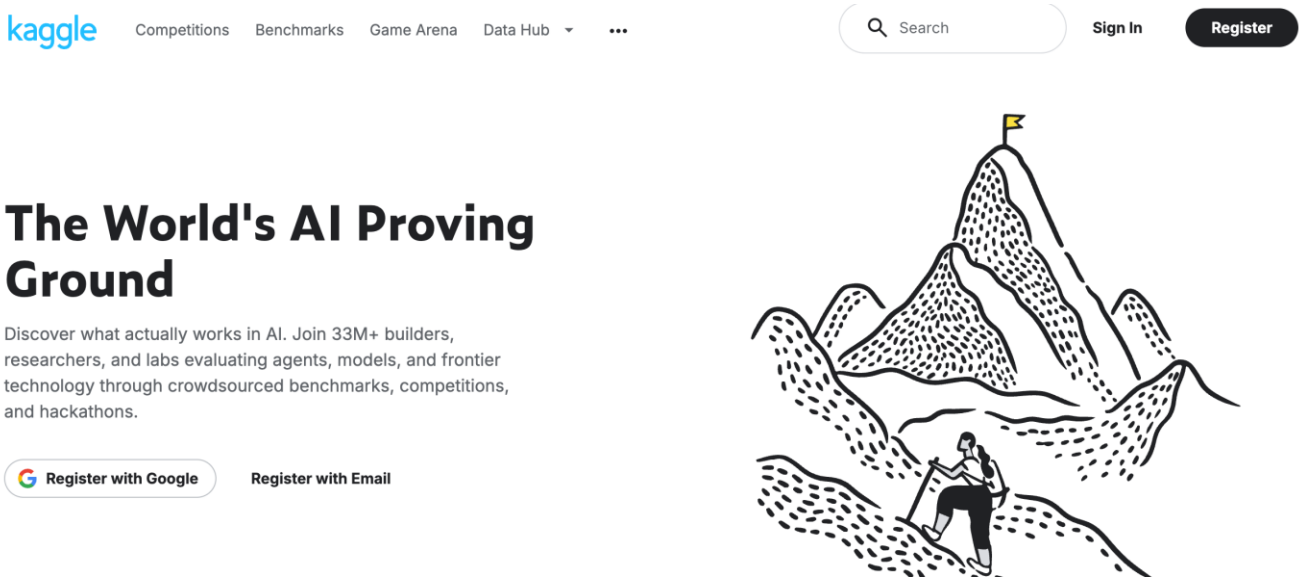
- 天池
  - 阿里云天池竞赛平台是一个汇聚 AI 大模型赛、数据算法赛等多类型赛事，同时提供学习课程、活动参与及相关云计算资源与服务支持的大数据竞赛与学习平台。
  - 里面包括很多大模型，推荐系统，数据算法等经典赛事



https://tianchi.aliyun.com/competition/

Kaggle 介绍

- Kaggle
  - Kaggle 是面向全球的数据科学与人工智能竞赛平台，提供竞赛、公开数据集、在线 Notebook 和学习课程，支持数据分析、模型开发与交流学习。
  - 里面包括很多大模型，推荐系统，数据算法等经典赛事



https://www.kaggle.com/competitions

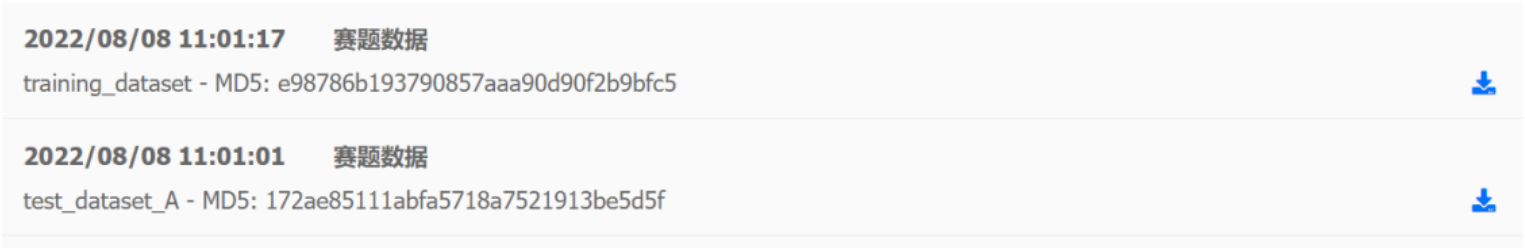
实验题目（推荐赛题）

实验题目（推荐赛题）

- 现提供以下实战题目和若干训练数据集
  - Kaggle比赛--Spaceship Titanic：利用乘客信息开展二分类预测。(长期开放)
  - Kaggle比赛--Geekhaven AI/ML Wing Selection Contest 2026：结合商品类别、品牌、描述及规格开展价格回归预测。(已结束，支持补交)
  - Kaggle比赛--Enveda CASMI 2026：根据质谱数据识别小分子结构，并对候选结构进行排序。(2026-12-14)

实验基本信息

- 数据集：训练集+测试集



- 常用评价指标
  - ✓ 回归任务：RMSE, MAE, NRMSE…
  - ✓ 分类任务：ACC, AUC, Recall@K, MRR@K...
  - ✓ 主办方自行定义指标：F1、NDCG等
- A/B榜：评分排名时测试数据分割为A/B两份，分别评分并生成对应排行榜，目的是为了**防止对测试数据过拟合**
  - A榜在“提交开放阶段”对提交结果自动评分并排名,生成A榜
  - B榜在“提交截止阶段”对提交结果自动评分并排名,生成B榜，**确定决赛资格**

课程推荐赛题
Kaggle比赛—太空泰坦尼克号

- **Kaggle：Spaceship Titanic**
- **任务介绍：** 在虚构的太空航行事故背景下，利用乘客的个人信息、舱位及消费记录，预测其是否被传送至另一维度。这是一项二分类任务。
- **评价指标：** 分类准确率（Accuracy），即预测正确的乘客数占总预测人数的比例，得分越高越好。
- **数据集：** 训练集包含 8,693 名乘客的记录及标签，测试集包含 4,277 名乘客的记录。数据包括出发星球、目的地、年龄、冷冻休眠状态、舱位及各项消费金额，目标是预测 Transported 字段。
- **比赛链接：** https://www.kaggle.com/competitions/spaceship-titanic

课程推荐赛题
Kaggle比赛—太空泰坦尼克号

- **数据说明：** train.csv 包含乘客特征及标签；test.csv 包含待预测乘客的特征；sample_submission.csv 提供提交格式。

字段	含义
`PassengerId`	乘客编号，包含同行群组及组内编号
`HomePlanet`、`Destination`	出发星球、目的地
`CryoSleep`、`Age`、`VIP`	冷冻休眠状态、年龄、是否购买 VIP 服务
`Cabin`	舱位信息，格式为“甲板 / 舱号 / 左右舷”
`RoomService`、`FoodCourt`、`ShoppingMall`、`Spa`、`VRDeck`	五类船上服务的消费金额
`Transported`	预测标签：True 表示被传送，False 表示未被传送

- 使用 train.csv 划分训练集与验证集，完成数据预处理、特征工程和模型比较；再对 test.csv 生成预测，按示例格式提交。
- **提交文件：** CSV 文件包含 PassengerId 和 Transported 两列

课程推荐赛题
Kaggle比赛—电商商品价格预测

- **Kaggle：Geekhaven AI/ML Wing Selection Contest 2026**
- **任务介绍：** 利用商品类别、品牌、描述、评分及规格等信息，预测商品零售价格。这是一项回归任务，重点在于从结构化字段、文本和半结构化规格中提取有效特征。
- **评价指标：** 均方根误差（RMSE），衡量预测价格与真实价格之间的偏差，得分越低越好。

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **数据集：** 数据包含不同零售类别的商品记录，存在缺失信息、不规范字段和长尾价格分布；训练集提供目标字段 price，测试集不包含该字段。
- **比赛链接：** https://www.kaggle.com/competitions/geekhaven-ai-ml-wing-selection-contest-2026

课程推荐赛题
Kaggle比赛—电商商品价格预测

- **数据说明：** train.csv 用于训练和验证，test.csv 用于生成预测，sample_submission.csv 提供提交格式。主要字段如下

字段	含义
`id`	商品唯一编号
`category`	多级商品类别，如“一级类别 > 二级类别 > 三级类别”
`brand`	商品品牌或制造商
`description`	商品描述文本
`rating`	商品评分，可能包含缺失值或“No rating”
`specs`	商品规格，采用类似 JSON 的字符串存储
`price`	**预测目标：商品零售价格，仅训练集提供**

- 建议依次完成缺失值处理、类别编码、描述文本特征提取和规格解析，比较不同特征组合与回归模型，并在独立验证集上评估效果。
- **提交文件：** CSV 文件仅包含 id、price 两列，行顺序与测试集一致。

课程推荐赛题
Kaggle--化学分子预测

- **Enveda CASMI 2026 - Molecule ID From Mass Spectra**
- **任务介绍：**质谱可以通过分子碎裂后产生的信号来帮助研究人员识别未知化合物。本比赛要求参赛者根据未知小分子的二级质谱数据，预测其对应的 二维化学结构，并使用 SMILES 字符串表示预测结果。
- **评价指标：**MRR@25 ↑，根据正确分子结构在 Top-25 候选中的排名计算倒数排名并取平均，正确答案排名越靠前得分越高。
- **数据集：**包含约 250 万条 MS/MS 质谱、27.5 万种分子结构，提供碎片 、峰强度、母离子质量、碰撞能量、加合物类型等信息，以及对应的真实 SMILES。
- **最终目标：**针对每个测试分子输出一个**按照置信度排序的 Top-25 SMILES 候选列表。**

课程推荐赛题
Kaggle--化学分子预测

- 训练数据示例：
  - SMILES 字符串
  - 各种化学实验量
  - ……

Row (0-based)	ingest_lib	normalized_smiles	inchikey	inchikey14
0	drug_plus	O=C1NC(=O)c2cc(Nc3ccccc3)c(Nc3ccccc3)cc21	AAALVYBICLMAMA-UHFFFAOYSA-N	AAALVYBICLMAMA
1	drug_plus	C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1	AAKJLRGGTJKAMG-UHFFFAOYSA-N	AAKJLRGGTJKAMG
2	drug_plus	CC1(C)CCC(C)(C)c2cc(C(O)C(O)=Nc3ccc(C(=O)O)cc3F)ccc21	AANFHDFOMFRLLR-UHFFFAOYSA-N	AANFHDFOMFRLLR
3	drug_plus	CN1C(=O)CN=C(c2ccccc2)c2cc(Cl)ccc21	AAOVKJBEBIDNHE-UHFFFAOYSA-N	AAOVKJBEBIDNHE
4	drug_plus	NC(=O)C1=CN(C2OC(COP(=O)(O)OP(=O)(O)OCC3OC(n4cnc5c(N)ncnc54)C(OP(=O)(O)O)C3O)C(O)C2O)C=CC1	ACFIXJIJDZMPPO-UHFFFAOYSA-N	ACFIXJIJDZMPPO



课程推荐赛题
Kaggle--化学分子预测

- 测试数据示例：
  - id
  - 各种化学实验量
  - ……

Row (0-based)	molecule_id	spectrum_id	ms2_mzs
0	m_005e53	s_4d19d522	▶ 183 items
1	m_005e53	s_5b6e28ff	▶ 119 items
2	m_005e53	s_6e514292	▶ 84 items
3	m_005e53	s_855ae19c	▶ 33 items
4	m_006153	s_3c74e9fe	▶ 219 items
5	m_006153	s_54245904	▶ 804 items
6	m_006153	s_59a80abf	▶ 343 items
7	m_006153	s_7151fa86	▶ 572 items
8	m_006153	s_86472581	▶ 192 items
9	m_006153	s_cb15f0ac	▶ 330 items
10	m_00b5aa	s_04dffc2a	▶ 244 items
11	m_00b5aa	s_3b617d66	▶ 176 items
12	m_01e922	s_4f6a16e5	▶ 31 items
13	m_01e922	s_cedea619	▶ 28 items
14	m_01e922	s_edf03be9	▶ 80 items
15	m_0259d4	s_234f7054	▶ 287 items

课程推荐赛题
Kaggle--化学分子预测

- **提示**：此次比赛的评测测试集分为三类数据，难度不同
  - Class 1：已有公开参考质谱；可以使用质谱相似度检索 / 最近邻 / 学习式检索
  - Class 2：没有参考质谱，但结构存在于 PubChem、COCONUT 等公开数据库；从数据库筛候选 + 模型排序
  - Class 3：PubChem 中都不存在；需要 de novo 分子结构预测
- 可以考虑针对不同难度进行不同的处理
- 比赛主页链接：
  - https://www.kaggle.com/competitions/enveda-CASMI26-molecule-id-mass-spectra/overview

实验报告（2026年12月08日）

- 当前任务
  - **10月20日前完成实验组队和选题，在线表格中填写组队信息和赛题信息**
  - **对于正在进行的比赛，注意比赛报名时间**

- 实验提交（12月08日）
  - 实验报告要求提交pdf格式文件
  - 将实验报告、代码和log文件压缩打包，命名为“姓名-学号-课程实验.zip”发送至课程邮箱，邮箱主题为“姓名-学号-课程实验”
  - 课程邮箱： ds_intro2026@163.com

