Multimodal Single-Cell Integration （中文版）

https://www.kaggle.com/competitions/open-problems-multimodal/overview

Results

Prize: Silver Medal

Rank: Top 4%

Competition Goal

- 预测随着骨髓干细胞发育成更成熟的血细胞，DNA、RNA和蛋白质测量值如何在单个细胞中共同变化
- 数据集来自四个人类供体的5个时间点的CD34+造血干细胞和祖细胞（HSPC）

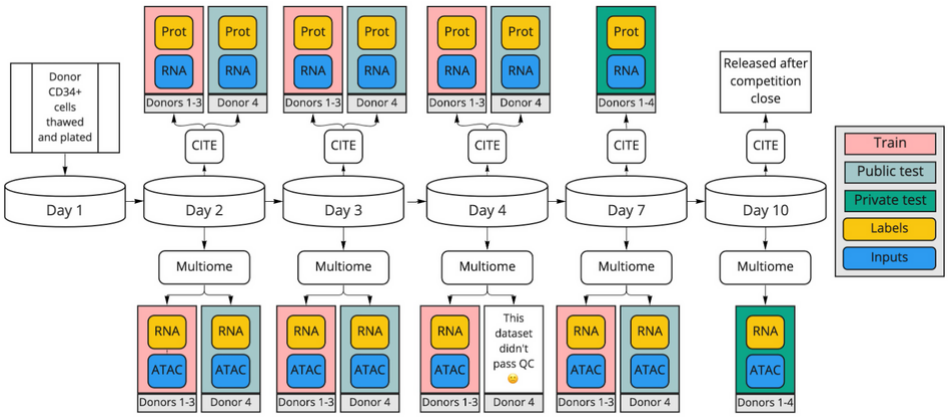
Context

- 数据集由两**种测定技术**分别各自测量**两种模式**得出：
  - Multiome试剂盒测量染色质可及性（DNA）和基因表达（RNA）
  - CITEseq试剂盒测量基因表达（RNA）和表面蛋白质水平
- 遵循分子生物学的中心法则：**DNA->RNA->Protein**，参赛者有**两个任务**：
  - 对于Multimo样本：给定DNA可及性，预测RNA
  - 对于CITEseq样本：给定RNA，预测蛋白质水平
- 数据的粒度是细胞，主键是cell_id：
  - 一共有30000个细胞
  - 8类细胞（7个已知类型，1个未知类型）

Data

- 任务一：CITEseq
  - 输入：
    - 70988×22050，连续值，高维度，稀疏矩阵
    - 行是不同的细胞，列是RNA的gene expression
    - 进行library-size normalization和log1p transformation
  - 输出：
    - 70988×140，连续值，稠密矩阵
    - RNA是2万多维
    - 蛋白质水平是140维
- 任务二：Multiome
  - 输入：
    - 105942×228942，更高维度，稀疏矩阵
    - 行是不同细胞，列是chromatin accessibility
    - 通过TF-IDF将数据变成连续值，log(TF)×log(IDF)
  - 输出：
    - 105942×23418，连续值，稠密矩阵
    - 染色质是22万多维
    - RNA是2万多维
- 训练集：
  - 细胞来自捐赠者13176、31800和32606
  - CITEseq样本包含day2-4
  - Multiome样本包含day2-4和7
- 测试集（公榜）：
  - 细胞来自捐赠者27678
  - CITEseq样本包含day2-4
  - Multiome样本包含day2-3和7
  - 相当于来一个新人，考验模型迁移到新捐赠者身上的表现

- 私榜：
  - 细胞来自四个捐赠者
  - CITEseq样本只包含day7
  - Multiome样本只包含day10
  - 考验模型迁移到新的时间节点上的表现



★ 亮点：

- 数据维度特别大
- 数据集的划分很考验模型的泛化能力

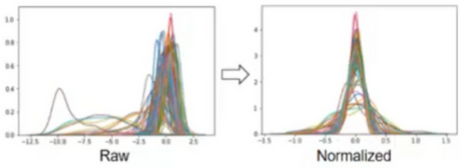
Prediction & Evaluation Metric

- 因为一个预测值是多维的，所以采用皮尔森相关系数进行评价，越相关值越靠近1
- 如果一个样本所有维度的预测值都一样，相关系数取-1
- 最后取每个样本的相关系数的平均值，越靠近1越好

```
def correlation_score(y_true, y_pred):
    corrsum = 0
    for i in range(len(y_true)):
        corrsum += np.corrcoef(y_true[i], y_pred[i])[1,0]
    return corrsum / len(y_true)
```

Feature Engineering

- **CITEseq**：
  - 维度太高，必须进行降维或者特征筛选
  - 去掉常数列
  - 根据业务背景对基因名进行name_matching，去重生成important_cols
  - 将训练集和测试集拼接在一起，对除去important_cols之外的特征采用截断SVD进行降维，从22050降到512维
  - 对数据进行归一化（normalization）：

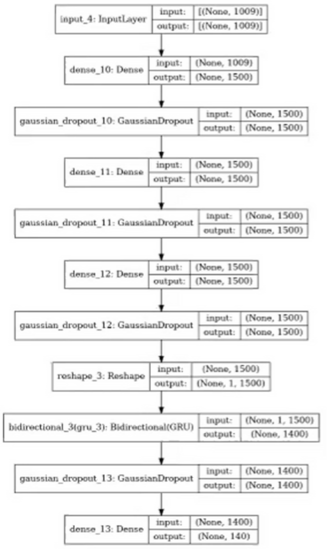


- **Multiome**：
  - 方案一：
    - 拼接训练集和测试集
    - 截断SVD降维
    - 数据归一化
    - 筛选降维归一化后信息量最高的64个特征
  - 方案二：
    - 拼接训练集和测试集
    - 数据归一化
    - 数据p2正则化（所有样本除以自己的p2范数）+ 对数变换（log1p）
    - 截断SVD降维
    - 数据归一化
    - 筛选降维归一化后信息量最高的100个特征
  - 两个方案提取的信息是不一样的，可以分别用来训练两个模型，然后进行融合

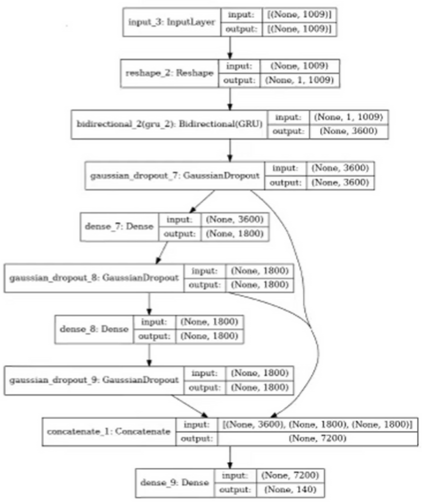
Modelling and Evaluation

- **CITEseq**：
  - 树模型：LightGBM+MultiOutputRegressor，直接输入处理后的特征，用sckopt简单进行调参
  - 深度学习：两个GRU （Gated Recurrent Unit），分别用不同的dense层扩充特征维度，gaussian dropout+GRU+concatenate（拼接多种处理后的数据，类似ResNet）
    - 采用相关系数loss作为损失函数，使用adam优化器训练模型，并使用Optuna进行调参

LightGBM+GRU



GRU:



- **Multiome**：
  - 深度学习：Multi-Layer Perceptron
  - GELU可以看作是RELU和dropout思想的结合，主要目的是为激活函数引入了随机性，使得模型训练过程更加鲁棒

    MLP

```python
class MultiNet(nn.Module):
    def __init__(self, dim_mod1, dim_mod2):
        super(MultiNet, self).__init__()
        self.input_ = nn.Linear(dim_mod1, 2048)
        self.fc = nn.Linear(2048, 2048)
        self.fc1 = nn.Linear(2048, 512)
        self.dropout1 = nn.Dropout(p=0.25)
        self.dropout2 = nn.Dropout(p=0.2)
        self.dropout3 = nn.Dropout(p=0.25)
        self.output = nn.Linear(512, dim_mod2)
    def forward(self, x):
        x = F.gelu(self.input_(x))
        x = self.dropout1(x)
        x = F.gelu(self.fc(x))
        x = self.dropout2(x)
        x = F.gelu(self.fc1(x))
        x = self.dropout3(x)
        x = F.gelu(self.output(x))
        return x
```

- **模型验证**：
  - 设计交叉验证策略，贴合真实的测试情况：新受试者+新时间点