**赛题**

* **比赛背景描述**

CRISPR/Cas系统是普遍存在于细菌和古菌中的一类古老的免疫系统，用于抵御防御外源遗传元件入侵，包括使用多效应蛋白的1类系统和使用单蛋白效应域进行切割的2类系统。2类系统因其单蛋白的优势是成为应用最广泛的CRISPR工具，其可进一步细分为II、V和VI三种类型。Cas12a属于2类系统中的V型，属于单一gRNA引导的核酸内切酶，具有对目标DNA进行双链切割的功能，是一种应用前景广泛的基因编辑技术。要求同学们基于现有实验获得的Cas12a在不同内源位点的编辑效率，利用AI模型预测Cas12a在不同内源位点编辑的效率，提高在内源靶向位点编辑的精确性。

* **场景描述：****包括不限于数据制作场景（街道、校园），环境要求：（明，暗，室外，白天，夜晚灯光充足）**

无场景要求。

**数据集**

* **提供数据集和对应文件目录介绍。**

数据集名为gRNA-data.zip。

赛题关注Cas12a在不同内源位点编辑的效率的预测，Cas12a是当前广泛使用的基因编辑工具，通过使用AI模型进行不同内源性编辑活性的预测，从而加速基因编辑的效率与精确度。本次数赛题提供一万多条高通量实验测定的特定靶向序列及其对应的编辑效率，要求同学们使用实验数据建立模型预测所提供序列对应的编辑效率。其中输入数据为Cas12a靶向位置的碱基序列，输出数据为该位置序列对应的编辑效率。

* **解释数据的标注方案。**

'34 bp synthetic target and target context sequence(4 bp + PAM + 23 bp protospacer + 3 bp)'表示的是靶向序列, 'Indel freqeuncy(Background substracted, %)'表示该靶向位置对应的编辑效率。AI模型输入为靶向位置的碱基序列，输出为Cas12a在该序列的编辑效率。

* **划分训练集，验证集和测试集。**

train.xlsx、calibration.xlsx和test.xlsx分别为本次比赛提供的训练集、验证集和测试集，选手可以基于训练数据集开发模型，然后通过验证集对模型进行校正，最后利用测试集提交结果。

具体的数据量说明：train.xlsx包含10000条数据，calibration.xlsx包含500条数据，test.xlsx包含50条数据，其中test.xlsx仅包含靶向位置的碱基序列数据用于评价模型的准确度。

* **定义测试集中公榜和私榜占用的比例。**

公榜和私榜占用的比例为1:1。具体而言，同学们提交整个测试集的预测结果，我们使用测试集的一部分计算得分和排名，实时显示在公榜上，用于给同学们提供及时的反馈和动态展示比赛的进行情况；测试集的剩余部分用于计算同学们最终提交的模型的最终得分和排名，此即为私榜，在比赛结束后会揭晓。

* **如果使用公共API测试数据，提供链接，保证测试集对参赛者不可见。**

无该情况。

* **如果数据预处理和加载非常见类型（例如医学图像，或过于复杂，提供预处理脚本）。**

无该情况。

* **数据的评价指标，包含公式（例如语义分割的IoU，2D框的f1-score）。**

本次赛题采用R2作为评价指标，

其中为测试集对应的实验数据，为模型预测的值。

* **数据测试方法。需要详细的测试流程。包括不同指标的加权方案，所使用的数据集个数，是否忽略类别，或忽略某项指标。**

最终排行榜展示结果只保留R2的四位小数。

* **每天限制提交评测的次数。**

每支队伍每天最多提交**1**次，队伍提交次数上限为**30**次。

**算法：**

* **如果任务非常规，根据任务难度考虑是否提供Baseline模型。**

不提供Baseline模型。

* **是否限制模型的复杂度。**

不限制。

**保障：**

* **我们将建立参赛QQ群用于集中答疑，请安排专门人员在群内不定期答疑**

QQ群号为：441242969