共有两个group project供同学们选择，每组可自行选择其中一个主题（每组成员不超过**4**人）。两个project都是**方法不限**的。

Group Project 1: Lux AI Challenge

1. 规则介绍

Lux AI Challenge是一个策略类的回合制生存游戏，两名玩家各操控一支队伍，在游戏开始时，每支队伍将获得一定的初始资源（工人单位、建筑单位），每回合工人单位可以在地图上进行移动/采集资源/建造建筑等操作，当夜晚回合到来时，需要消耗采集的资源来度过黑夜，当单位的资源不足以消耗时，单位将从地图中消失。在所有回合结束之后，拥有最多建筑单位的队伍获胜。

详细的规则介绍请参考：https://www.kaggle.com/nin7a1/lux-ai-rules-lux-ai

2. kaggle平台介绍

Lux AI Challenge主页：https://www.kaggle.com/c/lux-ai-2021，同学们需要注册kaggle账号进行登录并且加入Lux AI Challenge，同组的同学需要在kaggle平台上完成组队。

开始编写你的代码

- Tutorial: https://www.kaggle.com/stonet2000/lux-ai-season-1-jupyter-notebook-tutorial, 这份tutorial中，会提供大家一些相关的介绍链接，并指导大家如何实现一个简单的agent，怎么在kaggle平台上完成提交
- Baseline1: https://www.kaggle.com/aithammadiabdellatif/lux-ai-reinforcement-learning
- Baseline2: https://www.kaggle.com/huikang/lux-ai-working-title-bot
- Baseline3: https://www.kaggle.com/shoheiazuma/lux-ai-with-imitation-learning

以上3个baseline在排行榜上取得了不错的分数，大家可以在一些公开模型的基础上进行改进，但前提是必须了解这些模型的实现逻辑，切忌直接copy然后提交就不管了。大家可以在Code模块（https://www.kaggle.com/c/lux-ai-2021/code） 查看他人公开的模型。

合理利用讨论区

在Discussion模块（https://www.kaggle.com/c/lux-ai-2021/discussion）可以看到参加竞赛的其他人对竞赛相关内容的讨论，可以帮助大家加深理解、拓宽思路，如果存在代码/规则/思路等方面的疑问也可以直接在上面发布帖子寻求解答。

排行榜

在Leaderboard模块（https://www.kaggle.com/c/lux-ai-2021/leaderboard），可以看到自己以及其他人提交的agent在公开排行榜的分数排名情况，每当自己提交一个新的agent，系统会自动帮你匹配其他人的agent进行PK，按照胜负情况进行加减分，当几天之后，这个agent的分数便会趋于稳定，这时就可以查看你的agent的大致排名情况。

需要注意的是，最终的排名情况将由Private Leaderboard的分数排名情况决定，具体的规则是：在**2021.12.6**之后，所有用户将不能再进行提交，每只队伍可以选择2个agent进行最终的提交，在**2021.12.7~2021.12.20**之间，所有被选择的agent将进行为期两周的PK，以得到最终的分数排名。所以所有选择group project 1的同学必须在**2021.12.6**之前完成kaggle上的提交，请同学们考虑好时间安排再进行选择。

3. 提交要求

前提：完成kaggle上的组队，并且在kaggle平台上进行有效提交（在2021.12.6之前完成）；

每个小组应提交（在2021.12.31之前完成）：

1. 代码（注意代码规范，做一定的注释）
2. 报告
    - 提交pdf版本，中英文都可，限长度7页之内，模版不限
    - 在报告中详细描述模型的方法，如果使用了开源代码请进行标注，并且注上自己的改进/创新部分
    - 提供Public Leadboard以及Private Leadboard的分数排名截图
    - 标注小组同学分工情况
3. Presentation用到的PPT

4. 评分规则

- Presentation（20%）
- kaggle分数以及排名情况（20%）
- 报告（40%）
- 代码（20%）

5. 注意事项

- 不同小组的同学请勿在kaggle平台上互相分享自己的私人代码，否则可能会被kaggle官方移出排行榜
- kaggle提交的截止时间为**2021.12.6**，而group project提交的截止时间为**2021.12.31**

Group Project 2: RNA二级结构预测

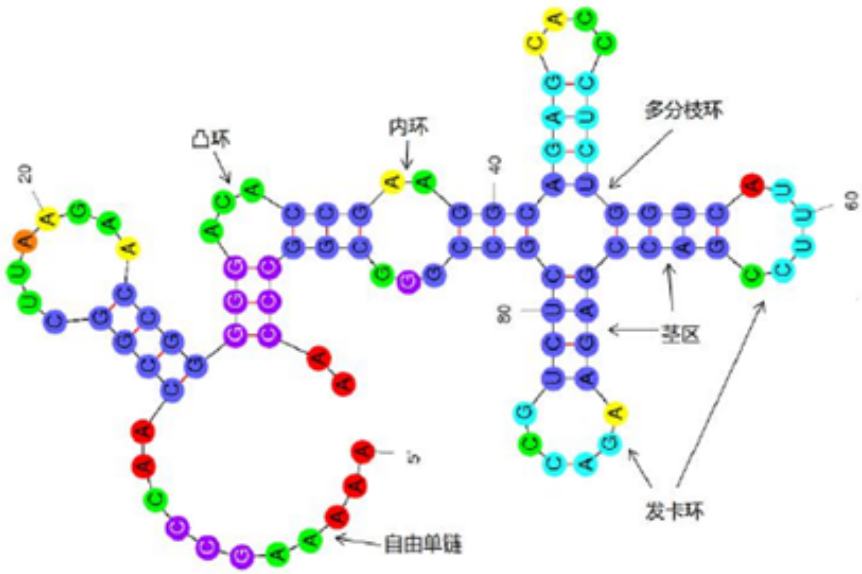
1. 背景

核糖核酸（RNA）是生物体内最为重要的聚合化合物之一。在基因表达的过程中，RNA是遗传信息从DNA转化为蛋白质的桥梁。与基因和蛋白质类似，RNA必须具有特定的空间结构才能发挥其功能。然而，使用实验方法测量RNA三级结构的方法昂贵且缓慢，因此，研究人员提出了从RNA一级结构（序列）去预测二级结构（平面结构），进而预测三级结构（空间结构）的流程。同时，研究表明，通过二级结构就可以获得部分RNA的功能信息。这可以帮助我们了解RNA在细胞中发挥调控功能的内在机制，并为实际生产生活中诸如医药、食品、环保等领域做出贡献。

2. RNA二级结构及其表示方法

2.1 RNA二级结构的基本构件

RNA由腺嘌呤（A）、尿嘧啶（U）、鸟嘌呤（G）、胞嘧啶（C）四种碱基组成，长度从几个碱基到上千个碱基不等。如图1所示，RNA二级结构包括单链，茎区，发卡环等基本构件（还有一种较为复杂的假结结构（Pseudoknot），该结构不可以被点括号表示法表示，且该结构的预测是目前RNA二级结构预测方法的最大难点之一，在本作业中可以不考虑）。



2.2 RNA二级结构的限制

与许多机器学习任务不同，RNA的二级结构预测需要考虑诸多的限制。这些限制可以总结如下：

- 仅有A-U，C-G，G-U（G-U配对极少，但是存在）
- 每个碱基仅能与至多一个其他碱基配对
- 不存在尖锐的环，即两个不同碱基A_i, A_j，若 $|i-j|<4$，则 A_i 与 A_j 不可能配对

2.3 RNA二级结构的表示方法

为了便于计算机的录入和使用，RNA二级结构一般有两种表示方式：

- 点括号表示（Dot-Brackets Notation）:这一方法将配对的两个碱基分别用左右括号表示，在序列中靠前的碱基对应左括号，靠后的碱基对应右括号。

- .ct文件：这一方法将RNA结构表示在一个文本文件中。其中，第1列与第6列是序列碱基的索引； 第2列则表示 RNA 序列中从5’端开始到3’端结束的各个碱基（A、U、G、C）的排列顺序； 第3列、 第4列分别表示序列中与之相邻的前一个碱基和后一个碱基的索引； 第 5 列表示 RNA 序列中是否存在与该位置碱基形成配对碱基对的碱基，其中数字‘0’ 表示该位置碱基是未配对碱基， 非‘0’ 表示该位置碱基存在配对碱基， 且用数字n表

示与之配对的碱基索引。如下表所示：

1	2	3	4	5	6
1	C	0	2	26	1
2	C	1	3	25	2
3	G	2	4	24	3
4	U	3	5	23	4
5	C	4	6	0	5
6	A	5	7	0	6
7	G	6	8	0	7
8	G	7	9	0	8
9	U	8	10	18	9
10	C	9	11	17	10
11	C	10	12	16	11
12	G	11	13	0	12
13	G	12	14	0	13
14	A	13	15	0	14
15	A	14	16	0	15
16	G	15	17	11	16
17	G	16	18	10	17
18	A	17	19	9	18
19	A	18	20	0	19
20	G	19	21	0	20
21	C	20	22	0	21
22	A	21	23	0	22
23	G	22	24	4	23
24	C	23	25	3	24
25	G	24	26	2	25
26	G	25	27	1	26

27	U	26	28	0	27
28	A	27	0	0	28

3. RNA二级结构数据库

由实验方法得到的真实RNA二级结构数据数量较少，推荐大家使用以下以真实结构为数据的数据库作为训练集：

- ArchiveII(Sloma & Mathews, 2016)：包括10种不同RNA的近4000个结构 http://rna.urmc.rochester.edu/pub/archiveII.tar.gz
- RNAStalign：包括8种RNA的约37000个结构，其中大约7000个结构包含假结。该数据库无公开链接，可以通过 https://drive.google.com/drive/folders/19KPRYJjjMJh1qdMhtmUoYA_ncw3ocAHc 下载
- RNASTRAND：包括5种RNA的约2493个结构。http://www.rnasoft.ca/strand （网站目前无法访问）

4. 任务

设计一个满足所有限制的RNA二级结构预测方法，输入RNA序列，输出预测出的结构（用点括号表示法表示）。

5. 提示

这一任务的最大难点在于，如何使得输出的二级结构能够满足2.2中的限制。该任务的传统方法使用动态规划算法计算最小自由能，其中的代表如LinearFold(https://pubmed.ncbi.nlm.nih.gov/31510672/)。随着深度学习的发展，使用不同backbone的深度学习方法被不断提出，例如CDPFold(https://pubmed.ncbi.nlm.nih.gov/31191603/)，E2EFold(https://arxiv.org/pdf/2002.05810.pdf)。大家可以在理解这些方法的原理后，参考其解决方法，做出一定的改进，但请务必不要直接copy。

6. 提交要求

每个小组应提交（在2021.12.31之前完成）：

1. 代码（注意代码规范，做一定的注释）
2. 报告
    - 提交pdf版本，中英文都可，限长度7页之内，模版不限
    - 在报告中详细描述模型的方法，如果用了公开模型请进行标注，并且注上自己的改进/创新部分
    - 在报告中注明算法在ArchiveII数据集上的表现
    - 标注小组同学分工情况
3. Presentation用到的PPT

7. 评分规则

- Presentation（20%）
- 在未公开测试集(包含少量假结数据)上的测试准确率（20%）（测试集在截止时间前一周发布）
- 报告（40%）
- 代码（20%）
- 如果在假结数据上的表现优异，可以获得一定的bonus points

关于group project 1，有问题可联系：tuyanlun@sjtu.edu.cn

关于group project 2，有问题可联系：wangf98@sjtu.edu.cn