大设施-zBiosynth使用文档

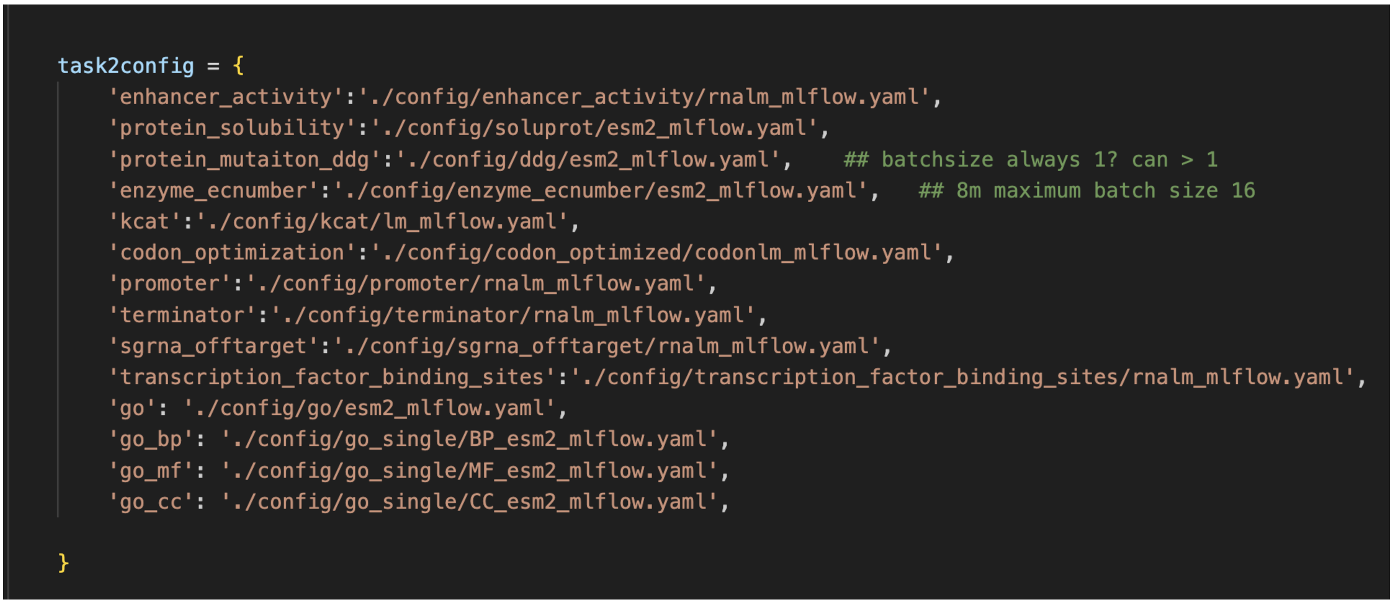
git地址：https://github.com/fuxuliu/zBioSynth/tree/dev_gary，分支：dev_gary

项目结构

- zBioSynth
  - zbiosynth
    - core (训练引擎/logger)
      - engine.py
      - .....
    - data (不同类型数据的预处理脚本)
      - protein_sequence.py
      - nucleotide_sequence.py
      - .....
    - datasets (不同任务训练所需要的dataset class)
      - codon_optimized.py
      - kcat.py
      - protein_solubility.py
      - ......
    - layers ()
    - metrics (评估指标)
    - mlflow_pyfunc_models (打包mlflow model)
      - mlflow_py_models.py (mlflow model class封装)
      - mlflow_signatures.py（记录不同任务对应的mlflow model所需的输入与其输出)
    - models (模型定义)
      - esm2.py
      - codon_models.py

      - rna_lm.py
      - .....
    - tasks (不同类型任务的定义)
      - codon_optimization.py
      - kcat_prediction.py
      - protein_solubility_prediction.py
      - .....
    - transforms
    - utils (工具)
      - comm.py
      - tools.py
  - MLproject (mlflow project)
  - run_single_pltform_mlflow.py (启动训练任务)
  - submit_run.py (平台提交训练任务)
  - zbiosynth.yml (所依赖的conda环境)
  - setup.py (安装脚本)

已支持任务



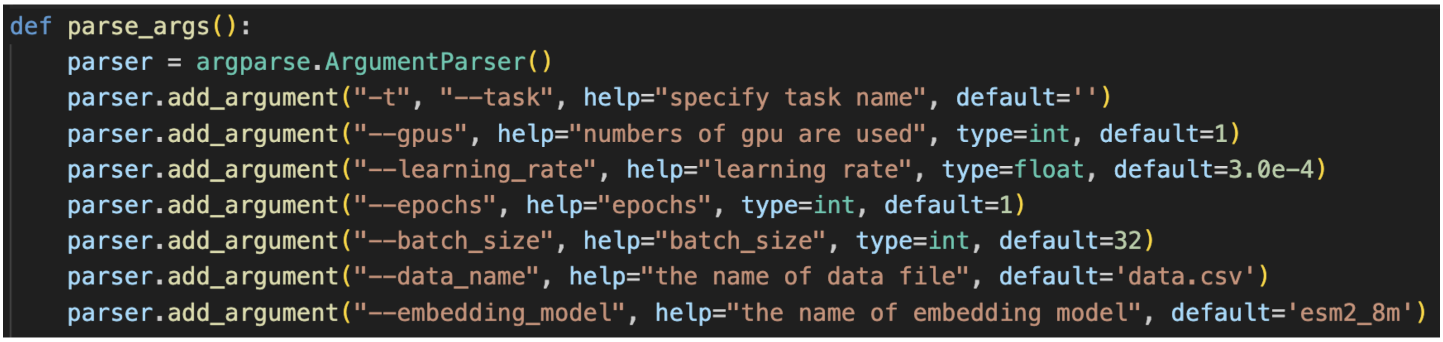
```
task2config = {
    'enhancer_activity':'./config/enhancer_activity/rnalm_mlflow.yaml',
    'protein_solubility':'./config/soluprot/esm2_mlflow.yaml',
    'protein_mutaiton_ddg':'./config/ddg/esm2_mlflow.yaml',    ## batchsize always 1? can > 1
    'enzyme_ecnumber':'./config/enzyme_ecnumber/esm2_mlflow.yaml',   ## 8m maximum batch size 16
    'kcat':'./config/kcat/lm_mlflow.yaml',
    'codon_optimization':'./config/codon_optimized/codonlm_mlflow.yaml',
    'promoter':'./config/promoter/rnalm_mlflow.yaml',
    'terminator':'./config/terminator/rnalm_mlflow.yaml',
    'sgrna_offtarget':'./config/sgrna_offtarget/rnalm_mlflow.yaml',
    'transcription_factor_binding_sites':'./config/transcription_factor_binding_sites/rnalm_mlflow.yaml',
    'go': './config/go/esm2_mlflow.yaml',
    'go_bp': './config/go_single/BP_esm2_mlflow.yaml',
    'go_mf': './config/go_single/MF_esm2_mlflow.yaml',
    'go_cc': './config/go_single/CC_esm2_mlflow.yaml',

}
```

输入的task name必须是以上的task name，否则训练时找不到对应的task，task会自动找对应的config.yaml

  - enhancer_activity -- dna序列的增强子活性预测 (回归)

  - protein_solubility -- 蛋白质序列的溶解性二分类预测 (二分类)
  - protein_mutaiton_ddg -- 蛋白质氨基酸突变的DDG预测 (回归)
  - enzyme_ecnumber -- 酶EC number的预测 (多分类，后面会改回多标签分类)
  - kcat -- 酶转化效率Kcat预测 (回归)
  - codon_optimization -- 密码子优化 (多分类)
  - promoter -- 启动子二分类预测 (二分类)
  - terminator -- 终止子二分类预测 (二分类)
  - sgrna_offtarget -- sgrna 是否脱靶预测 (二分类)
  - transcription_factor_binding_sites -- 转录因子结合位点预测 (二分类)
  - go -- 蛋白质序列的go term预测 (3 x 多标签分类)
  - go_bp -- 蛋白质序列BP 类别的go term预测 (多标签分类)
  - go_mf -- 蛋白质序列MF 类别的go term预测 (多标签分类)
  - go_cc -- 蛋白质序列 类别的go term预测 (多标签分类)

训练入口暴露的参数（目前）



```python
def parse_args():
    parser = argparse.ArgumentParser()
    parser.add_argument("-t", "--task", help="specify task name", default='')
    parser.add_argument("--gpus", help="numbers of gpu are used", type=int, default=1)
    parser.add_argument("--learning_rate", help="learning rate", type=float, default=3.0e-4)
    parser.add_argument("--epochs", help="epochs", type=int, default=1)
    parser.add_argument("--batch_size", help="batch_size", type=int, default=32)
    parser.add_argument("--data_name", help="the name of data file", default='data.csv')
    parser.add_argument("--embedding_model", help="the name of embedding model", default='esm2_8m')
```

- --task 需要输入任务的名称
- --gpus 训练时需要使用的GPU数量，若=0，则使用cpu进行训练，若>1，则自动切换至多卡进行训练
- -- learning_rate
- -- epochs
- --batch_size
- --data_name 数据集csv文件的名称
- --embedding_model 预训练模型的名称
  - 预训练权重路径：**/share/liufuxu/zBioSynth/resources/pretrained_weights**， 目前支持以下预训练模型

- 如果是kcat这个任务，需要同时输入蛋白质和小分子的预训练模型名称，使用","隔开，并且蛋白质模型在前，如esm2_8m,smole-bert
- 蛋白质序列: "esm2_8m", "esm2_35m", "esm2_150m", "esm2_650m", "esm2_3B", "esm2_15B"
- 核酸序列："rnalm_8m", "rnalm_35m", "rnalm_150m", "rnalm_650m"
- 密码子序列："codonlm_8m", "codonlm_35m", "codonlm_650m"
- 小分子smile序列："molt5-base", "molt5-small", "smole-bert"

使用案例(本地)

- 在本地跑的时候，数据集文件需要放在目录下的data目录下
- 安装zbiosynth，在目录下运行 python setup.py install
- git里面已经存放了一些任务所需的训练集，较大的没有上传，可在这个路径下获取 **/share/liufuxu/zBioSynth/dataset/data**
- **每个task的config.yaml 里都有一个默认的 model_path:/share/liufuxu/zBioSynth/resources/pretrained_weights 本地跑的时候，自动在这个路径下寻找预训练模型。**

密码子优化--codon_optimization

数据格式

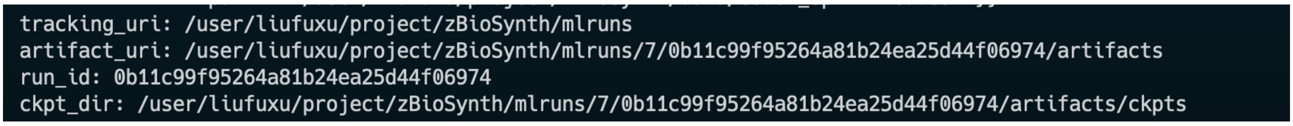
csv with columns:

- **prot_seq，蛋白质序列**
- **dna_seq，优化好的密码子序列/dna序列**
- **split，train/valid/test split**

启动命令

```
python run_single_pltform_mlflow.py \
--task codon_optimization \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 4 \
--gpus 1 \
--data_name codon_optimized.csv \
--embedding_model codonlm_8m
```

mlflow model infer



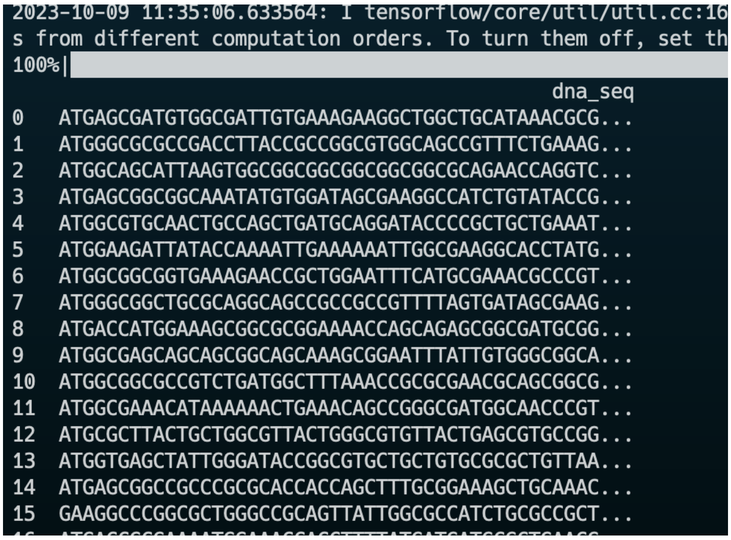
```
tracking_uri: /user/liufuxu/project/zBioSynth/mlruns
artifact_uri: /user/liufuxu/project/zBioSynth/mlruns/7/0b11c99f95264a81b24ea25d44f06974/artifacts
run_id: 0b11c99f95264a81b24ea25d44f06974
ckpt_dir: /user/liufuxu/project/zBioSynth/mlruns/7/0b11c99f95264a81b24ea25d44f06974/artifacts/ckpts
```

替换上面的路径

```
pymodel_mlflow =
mlflow.pyfunc.load_model('/user/liufuxu/project/zBioSynth/mlruns/7/0b11c9
```

启动

```
python mlflow_infer_debug/codon_optimization.py
```



```
2023-10-09 11:35:06.633564: I tensorflow/core/util/util.cc:16
s from different computation orders. To turn them off, set th
100%|
                                                 dna_seq
0   ATGAGCGATGTGGCGATTGTGAAAGAAGGCTGGCTGCATAAACGCG...
1   ATGGGCGCGCCGACCTTACCGCCGGCGTGGCAGCCGTTTCTGAAAG...
2   ATGGCAGCATTAAGTGGCGGCGGCGGCGGCGGCGCAGAACCAGGTC...
3   ATGAGCGGCGGCAAATATGTGGATAGCGAAGGCCATCTGTATACCG...
4   ATGGCGTGCAACTGCCAGCTGATGCAGGATACCCCGCTGCTGAAAT...
5   ATGGAAGATTATACCAAAATTGAAAAAATTGGCGAAGGCACCTATG...
6   ATGGCGGCGGTGAAAGAACCGCTGGAATTTCATGCGAAACGCCCGT...
7   ATGGGCGGCTGCGCAGGCAGCCGCCGCCGTTTTAGTGATAGCGAAG...
8   ATGACCATGGAAAGCGGCGCGGAAAACCAGCAGAGCGGCGATGCGG...
9   ATGGCGAGCAGCAGCGGCAGCAAAGCGGAATTTATTGTGGGCGGCA...
10  ATGGCGGCGCCGTCTGATGGCTTTAAACCGCGCGAACGCAGCGGCG...
11  ATGGCGAAACATAAAAAACTGAAACAGCCGGGCGATGGCAACCCGT...
12  ATGCGCTTACTGCTGGCGTTACTGGGCGTGTTACTGAGCGTGCCGG...
13  ATGGTGAGCTATTGGGATACCGGCGTGCTGCTGTGCGCGCTGTTAA...
14  ATGAGCGGCCGCCCGCGCACCACCAGCTTTGCGGAAAGCTGCAAAC...
15  GAAGGCCCGGCGCTGGGCCGCAGTTATTGGCGCCATCTGCGCCGCT...
```

sgRNA offtarget预测--sgrna_offtarget

数据格式

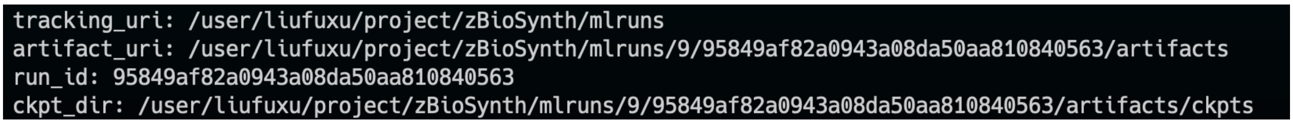
csv with columns:

- **sgrna_seq，sgrna序列**
- **dna_seq，靶点dna序列**
- **label, 标签0/1**
- **split**

启动命令

```
python run_single_pltform_mlflow.py \
--task sgrna_offtarget \
--learning_rate 3e-4 \
--epochs 2 \
--batch_size 64 \
--gpus 1 \
--data_name sgrna_offtarget_v2.csv \
--embedding_model rnalm_8m
```

mlflow model infer



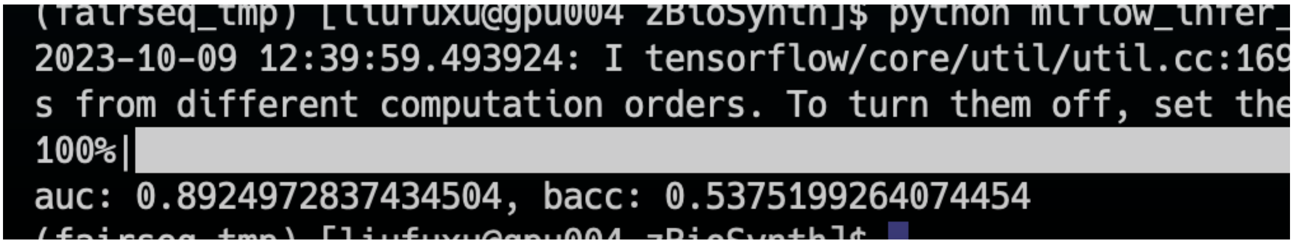
```
tracking_uri: /user/liufuxu/project/zBioSynth/mlruns
artifact_uri: /user/liufuxu/project/zBioSynth/mlruns/9/95849af82a0943a08da50aa810840563/artifacts
run_id: 95849af82a0943a08da50aa810840563
ckpt_dir: /user/liufuxu/project/zBioSynth/mlruns/9/95849af82a0943a08da50aa810840563/artifacts/ckpts
```

替换上面的路径

```
pymodel_mlflow =
mlflow.pyfunc.load_model('/user/liufuxu/project/zBioSynth/mlruns/9/958
```

启动

```
python mlflow_infer_debug/sgrna_offtarget.py
```



```
(fairseq_tmp) [liufuxu@gpu004 zBioSynth]$ python mlflow_infer_
2023-10-09 12:39:59.493924: I tensorflow/core/util/util.cc:169
s from different computation orders. To turn them off, set the
100%|
auc: 0.8924972837434504, bacc: 0.5375199264074454
(fairseq_tmp) [liufuxu@gpu004 zBioSynth]$
```

Kcat预测--kcat

数据格式

csv with columns:

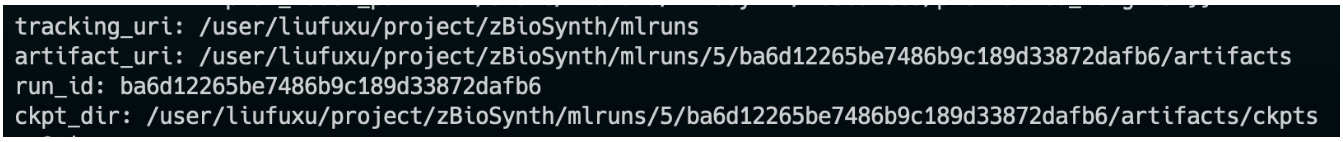
- **seq, 蛋白质/酶序列**
- **smile, 小分子smile序列**
- **kcat, 标签**

- split

启动命令

```
python run_single_pltform_mlflow.py \
--task kcat \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 8 \
--gpus 1 \
--data_name kcat.csv \
--embedding_model esm2_8m,smole-bert
```

mlflow model infer



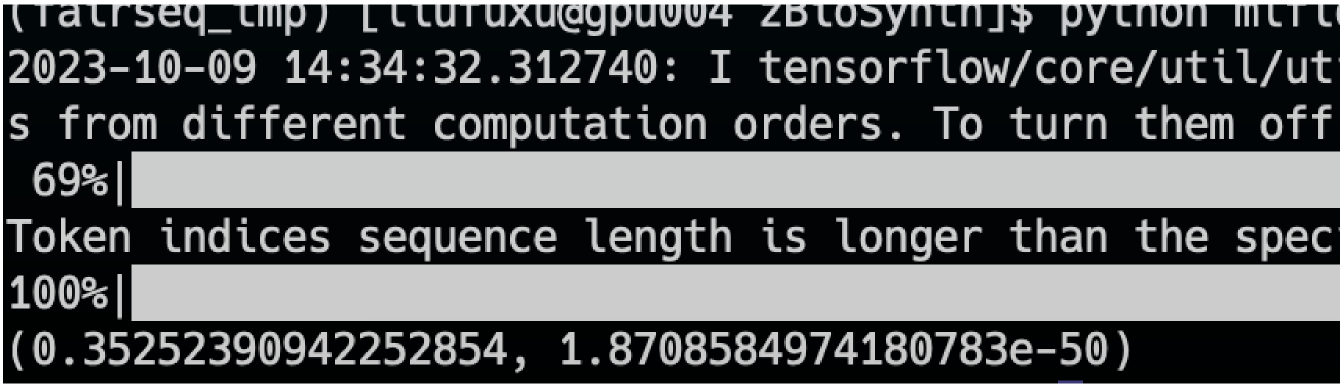
```
tracking_uri: /user/liufuxu/project/zBioSynth/mlruns
artifact_uri: /user/liufuxu/project/zBioSynth/mlruns/5/ba6d12265be7486b9c189d33872dafb6/artifacts
run_id: ba6d12265be7486b9c189d33872dafb6
ckpt_dir: /user/liufuxu/project/zBioSynth/mlruns/5/ba6d12265be7486b9c189d33872dafb6/artifacts/ckpts
```

替换上面的路径

```
pymodel_mlflow =
mlflow.pyfunc.load_model('/user/liufuxu/project/zBioSynth/mlruns/5/ba6d122

```

启动

```
python mlflow_infer_debug/kcat.py
```



```
(fairseq_tmp) [liufuxu@gpu004 zBioSynth]$ python mlfl
2023-10-09 14:34:32.312740: I tensorflow/core/util/ut
s from different computation orders. To turn them off
 69%|
Token indices sequence length is longer than the spec
100%|
(0.35252390942252854, 1.8708584974180783e-50)
```

protein solubility预测--protein_solubility

数据格式

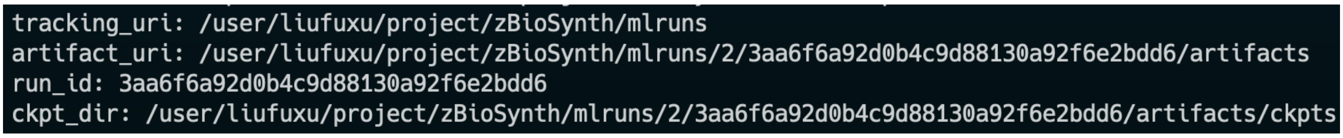
csv with columns:

- **seq, 蛋白质序列**
- **solubility, 0/1标签**
- **split**

启动命令

```
python run_single_pltform_mlflow.py \
--task protein_solubility \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 32 \
--gpus 1 \
--data_name soluprot.csv \
--embedding_model esm2_8m
```

mlflow model infer



```
tracking_uri: /user/liufuxu/project/zBioSynth/mlruns
artifact_uri: /user/liufuxu/project/zBioSynth/mlruns/2/3aa6f6a92d0b4c9d88130a92f6e2bdd6/artifacts
run_id: 3aa6f6a92d0b4c9d88130a92f6e2bdd6
ckpt_dir: /user/liufuxu/project/zBioSynth/mlruns/2/3aa6f6a92d0b4c9d88130a92f6e2bdd6/artifacts/ckpts
```

替换上面的路径

```
pymodel_mlflow =
mlflow.pyfunc.load_model('/user/liufuxu/project/zBioSynth/mlruns/2/3aa6f6a

```

启动

```
python mlflow_infer_debug/soluprot.py
```

Enhancer activity预测--enhancer_activity

数据格式

csv with columns:

- **seq, 核酸序列**
- **activity, 强度**
- **split**

启动命令

```
python run_single_pltform_mlflow.py \
--task enhancer_activity \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 32 \
--gpus 1 \
--data_name rice_enhancer.csv \
--embedding_model rnalm_8m
```

mlflow model infer

启动

```
python mlflow_infer_debug/soluprot.py
```

蛋白突变ddg预测--protein_mutaiton_ddg

数据格式

csv with columns:

- **wt_seq, 野生型蛋白质序列**
- **mut_seq, 突变型蛋白质序列**
- **position, 突变位置，从1开始**
- **ddg, 标签**
- **split**

启动命令

```
python run_single_pltform_mlflow.py \
--task protein_mutaiton_ddg \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 8 \
--gpus 1 \
--data_name ddg.csv \
--embedding_model esm2_8m
```

mlflow model infer

启动

```
python mlflow_infer_debug/ddg.py
```

promoter/non-promoter预测--promoter

数据格式

csv with columns:

- **seq, 核酸序列**
- **label, 标签**
- **split**

启动命令

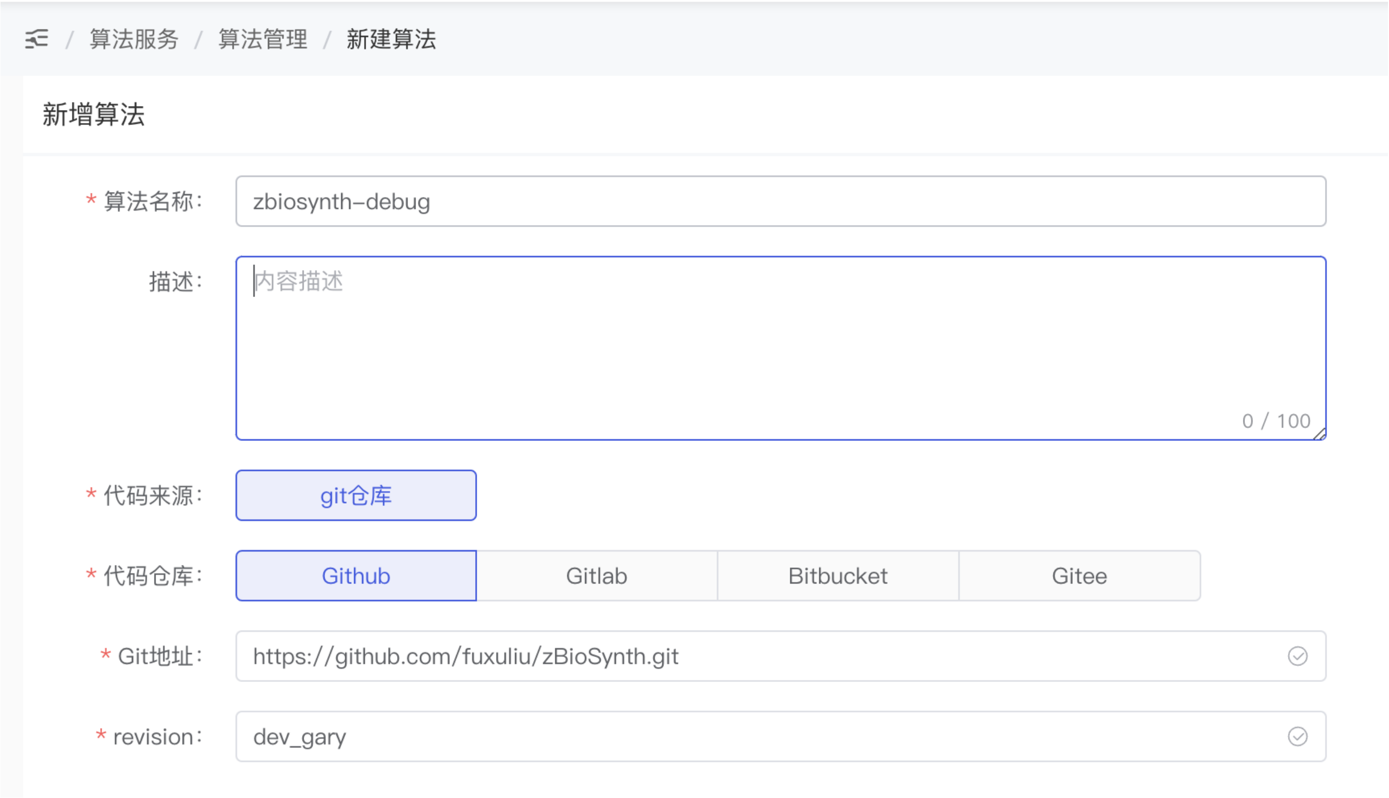
```
python run_single_pltform_mlflow.py \
--task promoter \
--learning_rate 3e-4 \
--epochs 5 \
--batch_size 32 \
--gpus 1 \
--data_name promoter.csv \
--embedding_model rnalm_8m
```

mlflow model infer

启动

```
python mlflow_infer_debug/promoter.py
```

使用案例(平台)



算法服务 / 算法管理 / 新建算法

新增算法

* 算法名称： zbiosynth-debug

描述： 内容描述 0 / 100

* 代码来源： git仓库

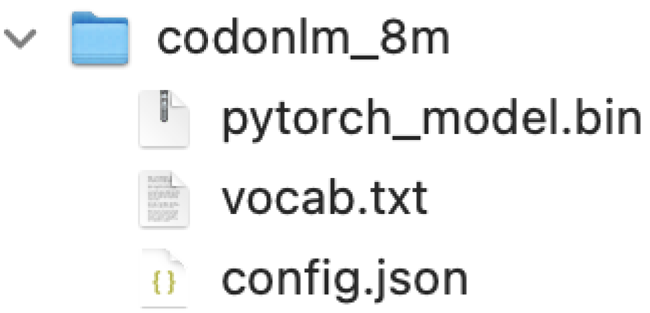
* 代码仓库： Github | Gitlab | Bitbucket | Gitee

* Git地址： https://github.com/fuxuliu/zBioSynth.git

* revision： dev_gary

密码子优化

上传特征集，预训练模型，打包成codon_LM.zip, zip -q -r codon_LM.zip codonlm_8m/

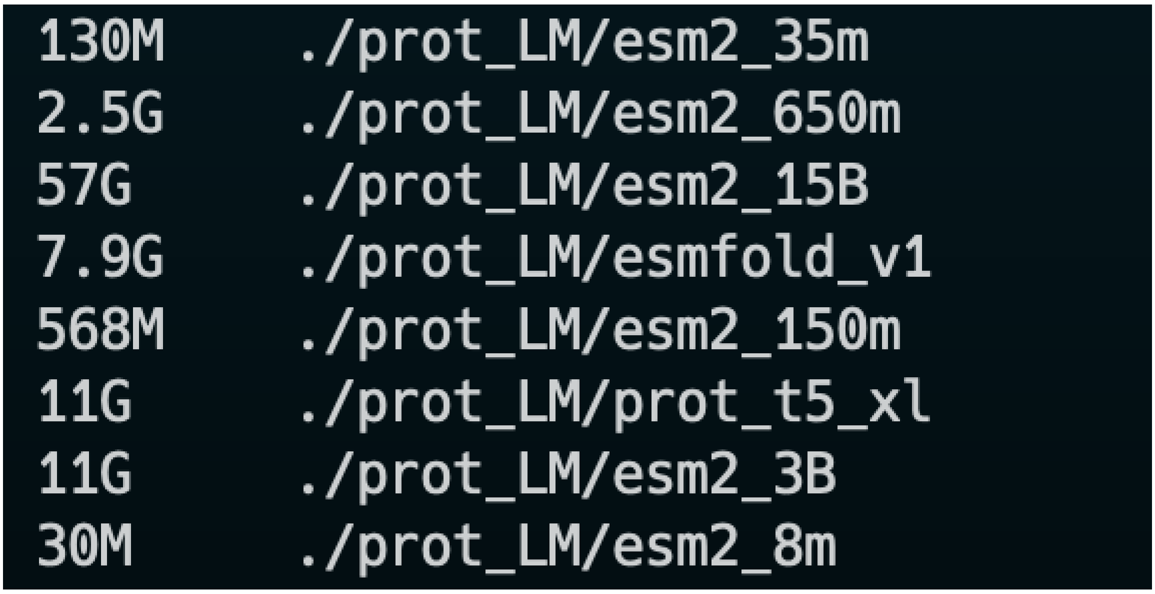


codonlm_8m
- pytorch_model.bin
- vocab.txt
- config.json

预训练模型打包命名规则

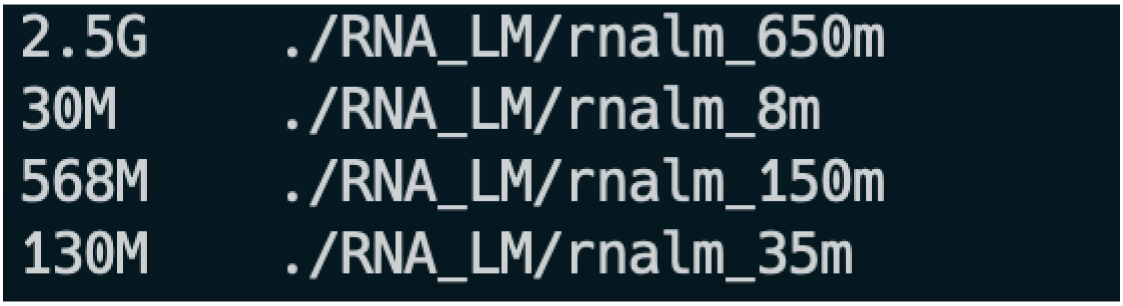
路径：/share/liufuxu/zBioSynth/resources/pretrained_weights

蛋白质序列模型--prot_LM



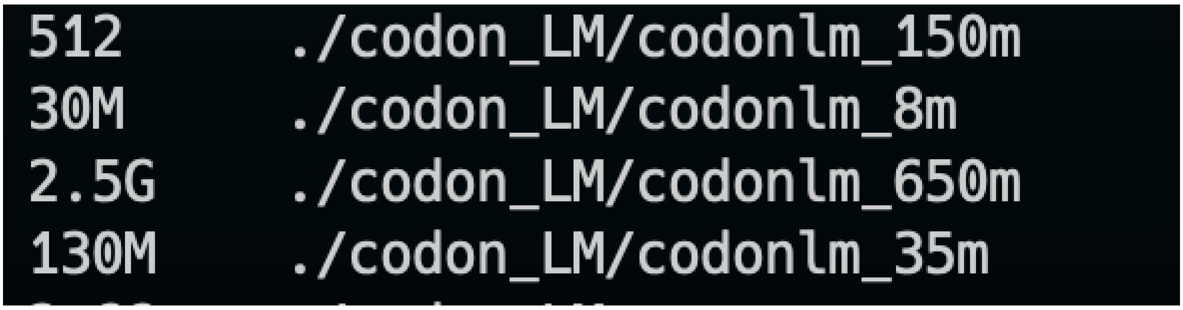
```
130M      ./prot_LM/esm2_35m
2.5G      ./prot_LM/esm2_650m
57G       ./prot_LM/esm2_15B
7.9G      ./prot_LM/esmfold_v1
568M      ./prot_LM/esm2_150m
11G       ./prot_LM/prot_t5_xl
11G       ./prot_LM/esm2_3B
30M       ./prot_LM/esm2_8m
```

核酸序列模型--RNA_LM



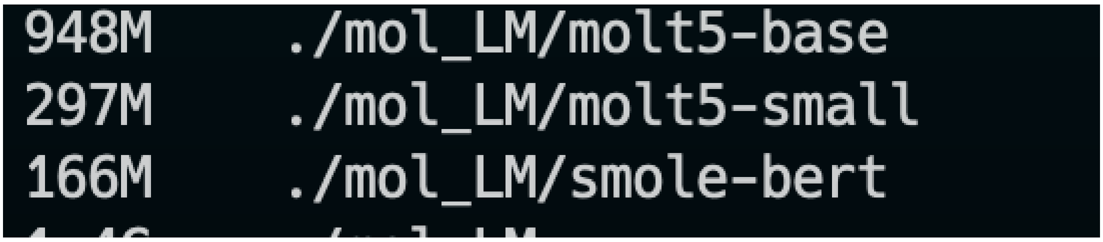
```
2.5G      ./RNA_LM/rnalm_650m
30M       ./RNA_LM/rnalm_8m
568M      ./RNA_LM/rnalm_150m
130M      ./RNA_LM/rnalm_35m
```

密码子序列模型--codon_LM

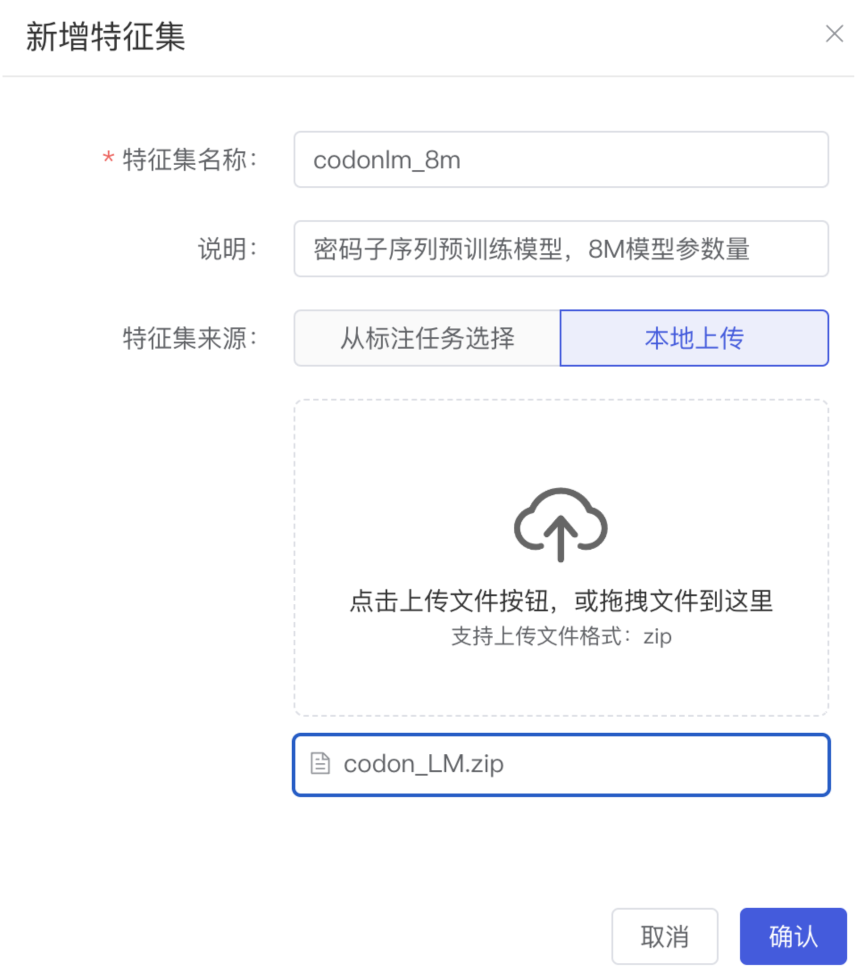


```
512       ./codon_LM/codonlm_150m
30M       ./codon_LM/codonlm_8m
2.5G      ./codon_LM/codonlm_650m
130M      ./codon_LM/codonlm_35m
```

小分子序列模型--mol_LM



```
948M      ./mol_LM/molt5-base
297M      ./mol_LM/molt5-small
166M      ./mol_LM/smole-bert
```



新增特征集

* 特征集名称：codonlm_8m

说明：密码子序列预训练模型，8M模型参数量

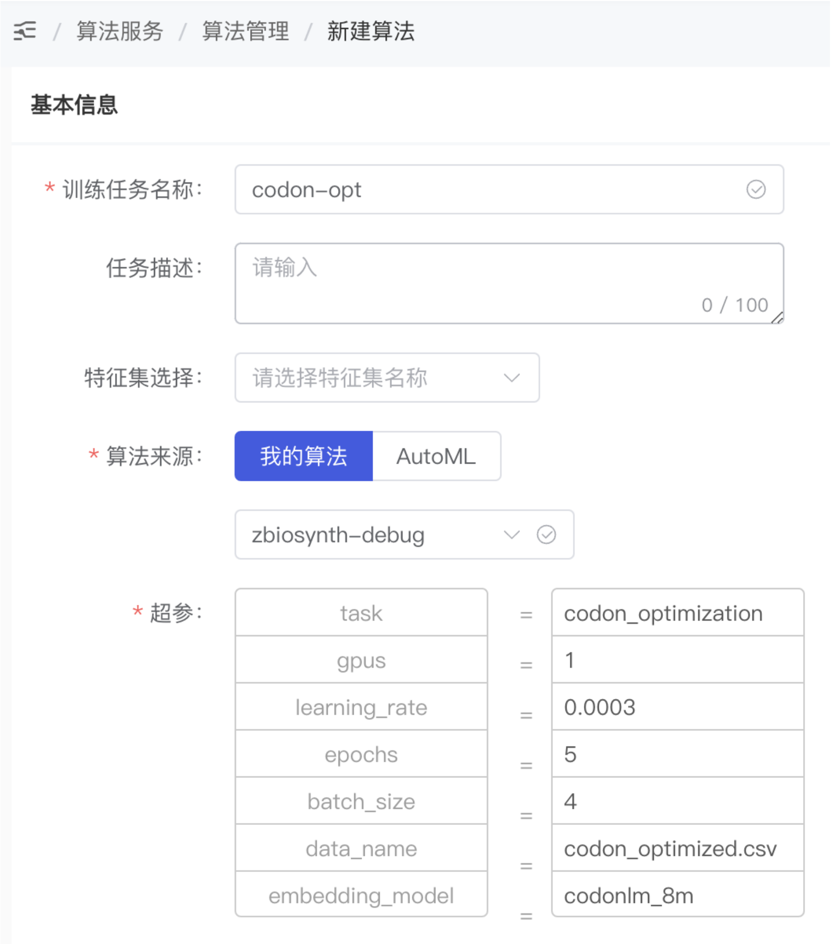
特征集来源：从标注任务选择 [✓ 本地上传]

点击上传文件按钮，或拖拽文件到这里

支持上传文件格式：zip

codon_LM.zip

创建训练作业，记得选择特征集（预训练模型），由于平台暂不支持多个特征集的导入，数据集已传入git，暂时使用git里的数据集进行debug



算法服务 / 算法管理 / 新建算法

基本信息

* 训练任务名称：codon–opt

任务描述：请输入 0 / 100

特征集选择：请选择特征集名称

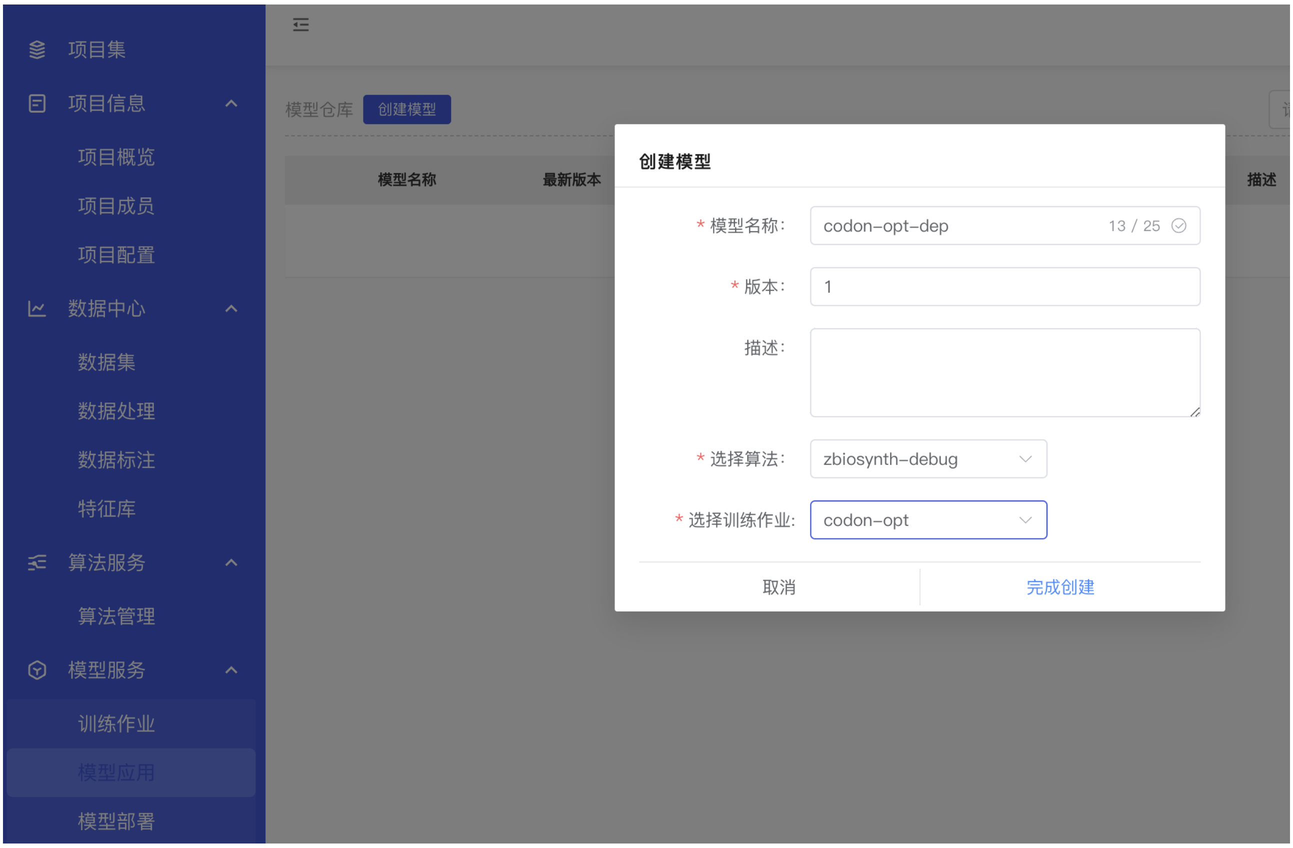
* 算法来源：我的算法 | AutoML

zbiosynth–debug

* 超参：

task	codon_optimization
gpus	1
learning_rate	0.0003
epochs	5
batch_size	4
data_name	codon_optimized.csv
embedding_model	codonlm_8m

创建模型应用，构建模型



创建模型

* 模型名称：codon–opt–dep 13 / 25

* 版本：1

描述：

* 选择算法：zbiosynth–debug

* 选择训练作业：codon–opt

取消 完成创建