工作汇报

姓名：陈开心

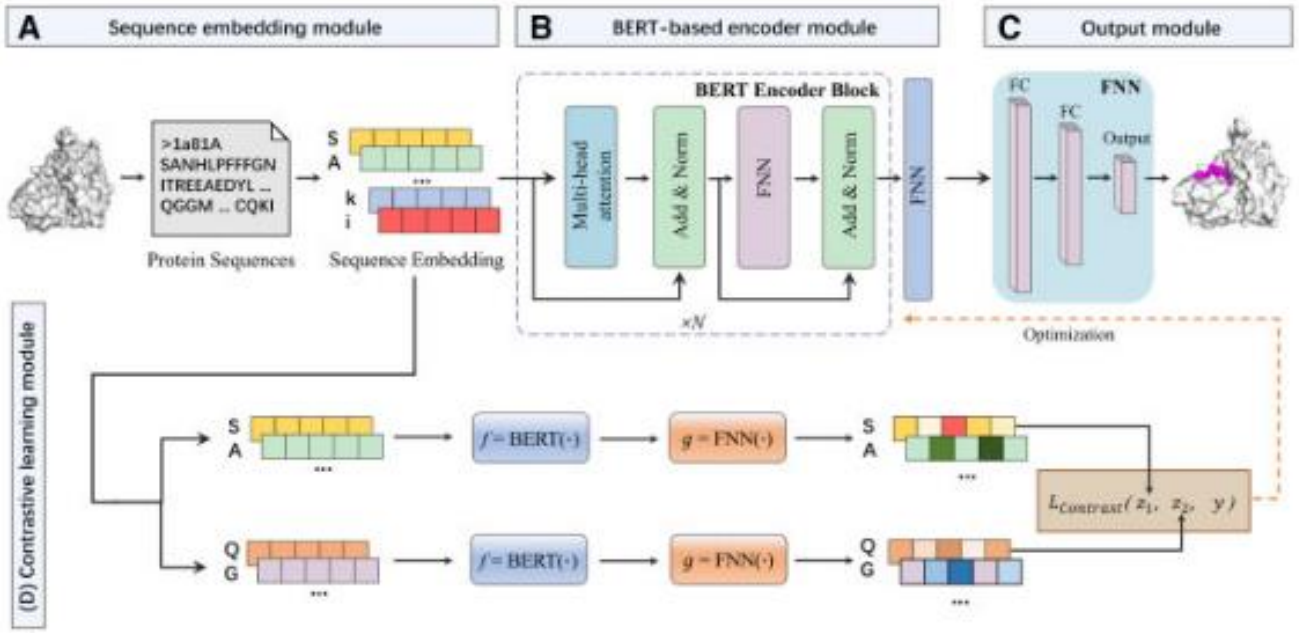
日期：2023.03.16

Dataset

Dataset	Trainset	Testset
Dataset1	1154 proteins	125 protein
	276822 residues	30870 residues
Negative/Positive	261789/15033	29154/1716

Dataset	Trainset	Testset
Dataset2	640 proteins	639 proteins
	157362 residues	150330 residues
Negative/Positive	149103/8259	141840/8490

PepBCL



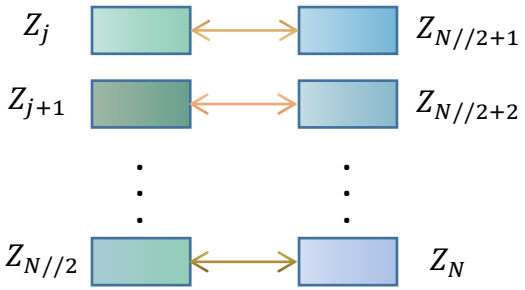
- 一、通过预先训练好的语言模型获取蛋白质序列残基级表征
- 二、设计了一个对比学习模块

Contrastive learning module in PepBCL

算法：一个批次的对比损失

输入：**X**：输入蛋白质序列，**Label**：输入蛋白质序列的标签，**Embed**：序列嵌入模块，**Encode**：基于预训练模型的编码模块，**Cross**：交叉注意力表征捕捉模块，**F**：降低表征维度，**M**：批量大小，**N**：一个批量残基的数量，**Z**：一个批量中残基的表征，**Y**：一个批量中残基的标签,**L**:一个批量的对比损失值

for i = 1, ..., M **do**
  $X_{Embed,i}$ = Embed(X_i)
  $X_{Encode,i}$ = Encode($X_{Embed,i}$)
  $X_{Cross,i}$ = Cross($X_{Encode,i}$)
  $X_{FNN,i}$ = F($X_{Cross,i}$)
  Z = concate(Z, $X_{F,i}$)
  Y = concate(Y, $Label_i$)
end
for j = 1, ..., N//2 **do**
  z_j = Z[j]
  $Z_{N//2+j}$ = Z[N//2+j]
    y = Y[j]^Y[N//2+j]
    L = L + $L_{contrast}$(z_j, $Z_{N//2+j}$, y)
end



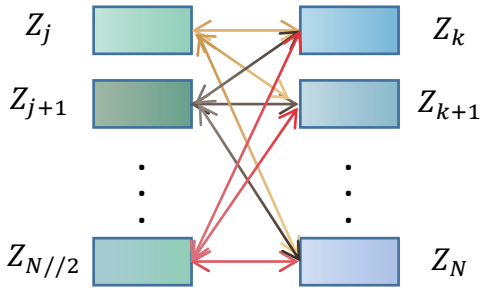
一个批量中的对比次数为N

Contrastive learning module in PepBMP

算法：一个批次的对比损失

输入：**X**：输入蛋白质序列，**Label**：输入蛋白质序列的标签，**Embed**：序列嵌入模块，**Encode**：基于预训练模型的编码模块，**Cross**：交叉注意力表征捕捉模块，**F**：降低表征维度，**M**：批量大小，**N**：一个批量残基的数量，**Z**：一个批量中残基的表征，**Y**：一个批量中残基的标签,**L**:一个批量的对比损失值

for i = 1, ..., M **do**
  $X_{Embed,i}$ = Embed(X_i)
  $X_{Encode,i}$ = Encode($X_{Embed,i}$)
  $X_{Cross,i}$ = Cross($X_{Encode,i}$)
  $X_{FNN,i}$ = F($X_{Cross,i}$)
  Z = concate(Z, $X_{F,i}$)
  Y = concate(Y, $Label_i$)
end
for j = 1, ..., N//2 **do**
  for k = N//2 + 1, ..., N **do**
    z_j = Z[j]
    z_k = Z[k]
    y = Y[j]^Y[k]
    L = L + $L_{contrast}$(z_j, z_k, y)
end



一个批量中的对比次数为N^2

在复合模态编码器模块收集批量大小的表示矩阵，这样就可以获得足够的残基水平数据用于对比学习

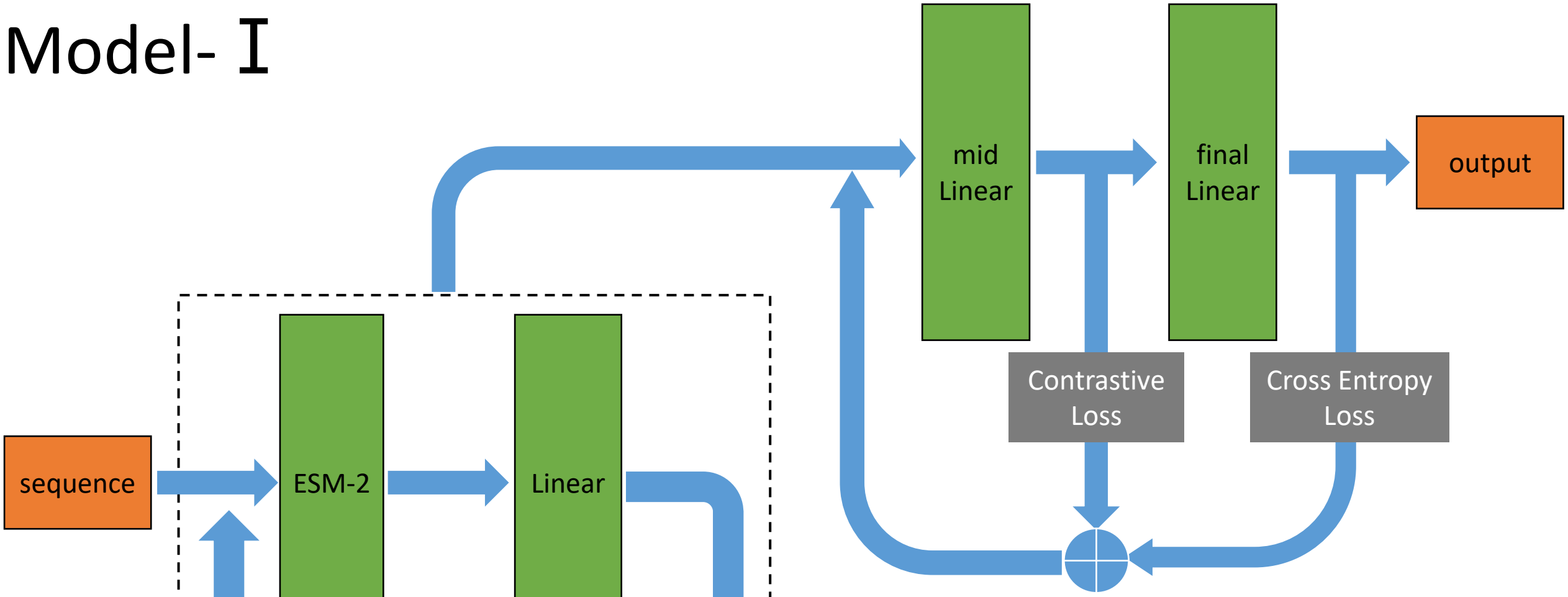
no contrast VS contrast VS new contrast

Dataset1上的结果

Contrast loss	ACC	Pre	Sen	Spe	F1	AUC	MCC
No							
Contrast	0.950	**0.60**	0.319	0.988	0.417	0.83	0.416
New contrast	0.950	0.58	0.347	0.985	0.435	**0.832**	**0.426**

注：在多预训练模型、SelfDoc框架下的结果

Model- I



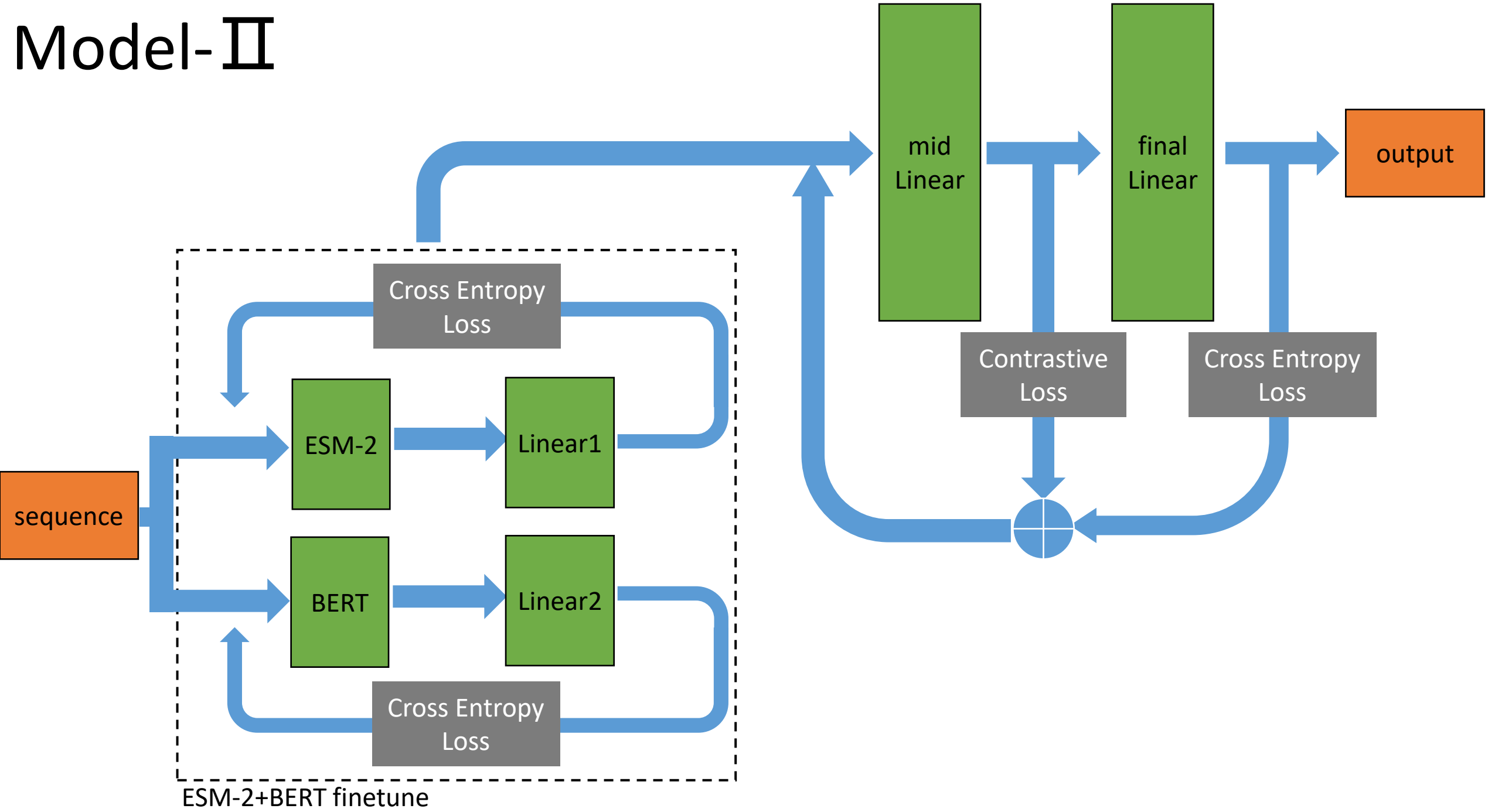
Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives[a,b,1,2], Joshua Meier[a,1], Tom Sercu[a,1], Siddharth Goyal[a,1], Zeming Lin[b], Jason Liu[a], Demi Guo[c,3], Myle Ott[a], C. Lawrence Zitnick[a], Jerry Ma[d,e,3], and Rob Fergus[b]

[a]Facebook AI Research, New York, NY 10003; [b]Department of Computer Science, New York University, New York, NY 10012; [c]Harvard University, Cambridge, MA 02138; [d]Booth School of Business, University of Chicago, Chicago, IL 60637; and [e]Yale Law School, New Haven, CT 06511

Edited by David T. Jones, University College London, London, United Kingdom, and accepted by Editorial Board Member William H. Press December 16, 2020 (received for review August 6, 2020)

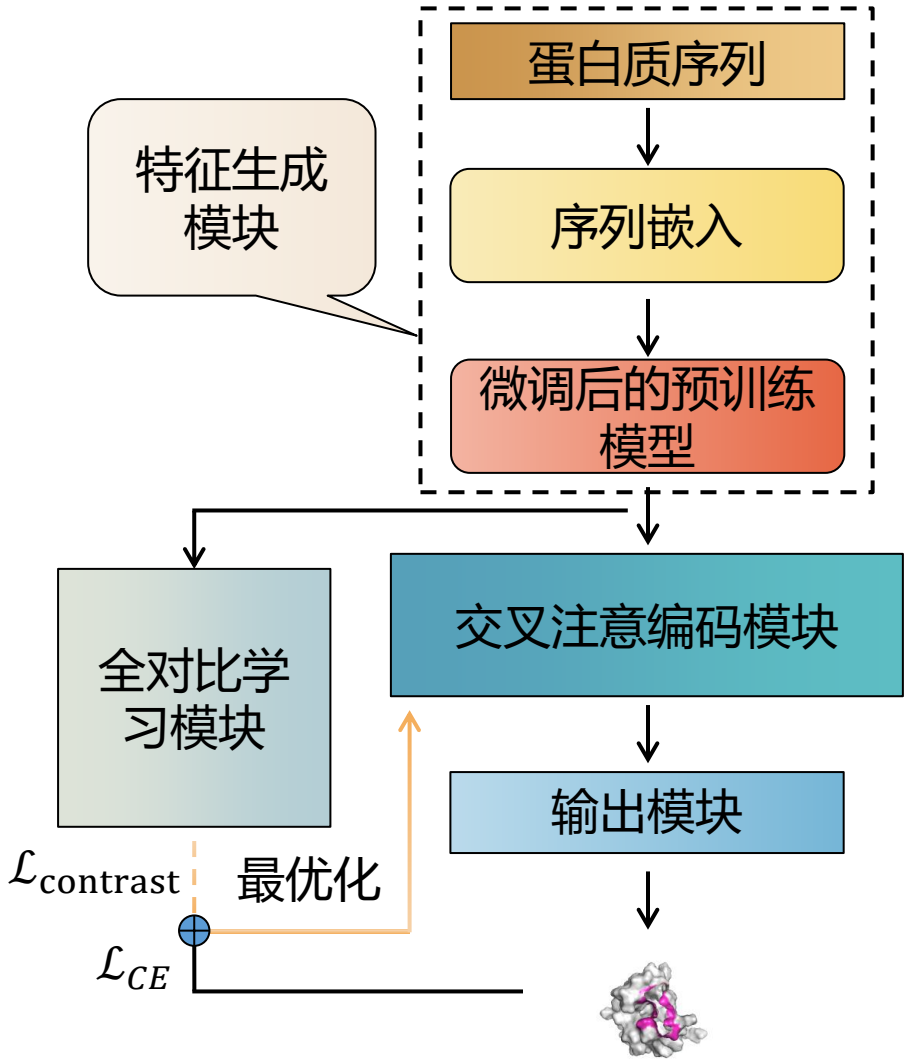
Model-Ⅱ

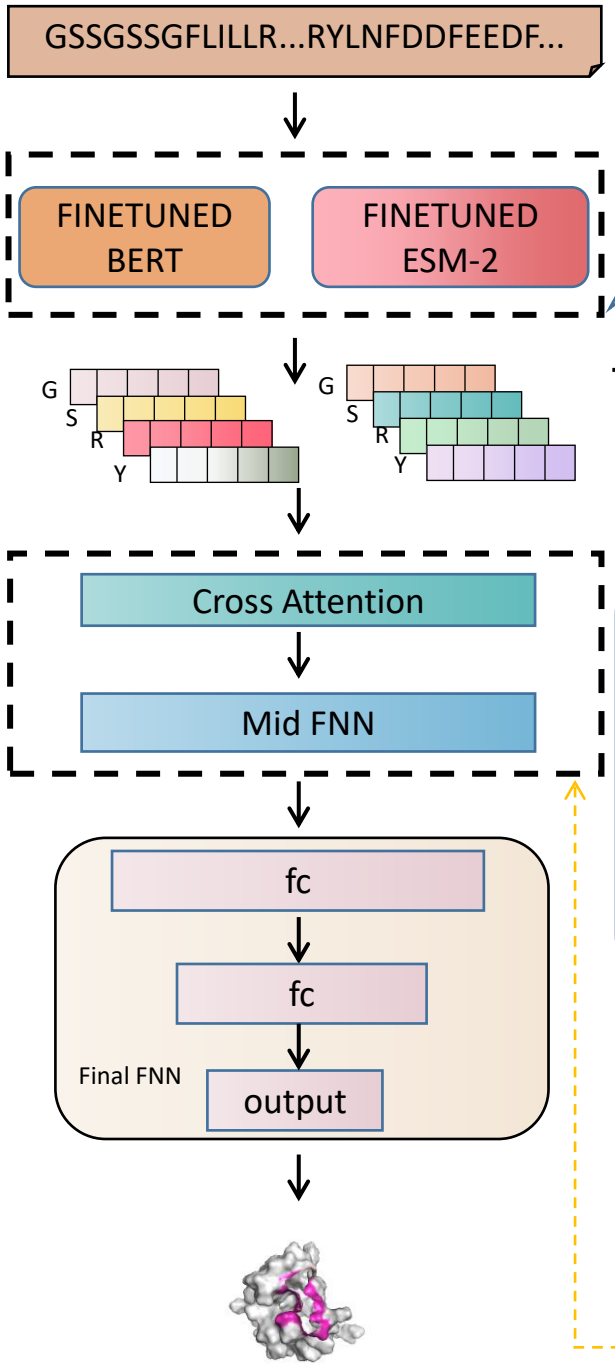

mid Linear
final Linear
output
Cross Entropy Loss
Contrastive Loss
Cross Entropy Loss
ESM-2
Linear1
sequence
BERT
Linear2
Cross Entropy Loss
ESM-2+BERT finetune

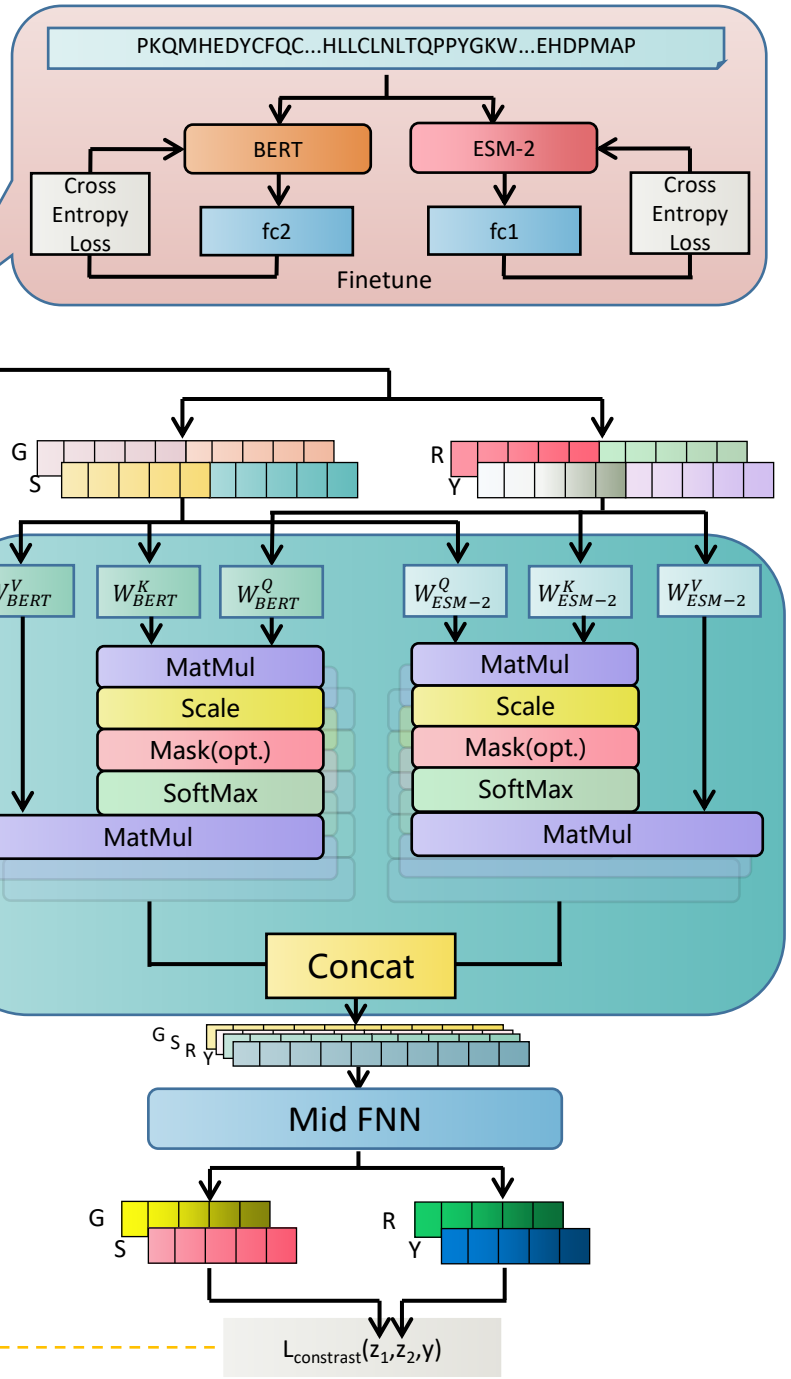
Dataset1上的结果

Pretrain model	ACC	Pre	Sen	Spe	F1	AUC	MCC
BERT	0.939	0.43	0.279	0.978	0.338	0.781	0.315
ESM-2	0.939	0.45	0.426	0.970	0.439	0.827	0.407
ESM-2+BERT	0.943	**0.480**	0.405	0.975	0.441	**0.830**	**0.413**

Model-Ⅲ


蛋白质序列
特征生成模块
序列嵌入
微调后的预训练模型
全对比学习模块
交叉注意编码模块
输出模块
$\mathcal{L}_{\text{contrast}}$
最优化
$\mathcal{L}_{CE}$


GSSGSSGFLILLR...RYLNFDDFEEDF...
FINETUNED BERT
FINETUNED ESM-2
G
S
R
Y
Cross Attention
Mid FNN
fc
fc
Final FNN
output


PKQMHEDYCFQC...HLLCLNLTQPPYGKW...EHDPMAP
BERT
ESM-2
Cross Entropy Loss
fc2
fc1
Cross Entropy Loss
Finetune
W^V_{BERT}
W^K_{BERT}
W^Q_{BERT}
W^Q_{ESM-2}
W^K_{ESM-2}
W^V_{ESM-2}
MatMul
Scale
Mask(opt.)
SoftMax
MatMul
Concat
Mid FNN
$L_{constrast}(z_1,z_2,y)$

Result

Dataset1上的结果

Network	ACC	Pre	Sen	Spe	F1	AUC	MCC
Fc	0.943	0.480	0.405	0.975	0.441	**0.83**	0.413
Self-atten	0.948	**0.550**	0.358	0.983	0.434	**0.83**	**0.419**
Corss-atten	0.945	0.510	0.390	0.978	0.442	**0.83**	0.418

Dataset2上的结果

Network	ACC	Pre	Sen	Spe	F1	AUC	MCC
Fc							
Self-atten							
Corss-atten	0.944	0.51	0.260	0.985	0.343	0.795	0.337

SelfDoc

SelfDoc: Self-Supervised Document Representation Learning

Peizhao Li[1]*, Jiuxiang Gu[2], Jason Kuen[2], Vlad I. Morariu[2], Handong Zhao[2],
Rajiv Jain[2], Varun Manjunatha[2], Hongfu Liu[1]
[1]Brandeis University, [2]Adobe Research
{peizhaoli,hongfuliu}@brandeis.edu
{jigu,kuen,morariu,hazhao,rajijain,vmanjuna}@adobe.com

Feature

D = $\{p_1, \dots, p_N\}$
一个文档D由N个文档对象方案（文本块、标题、列表、表格和插图）组成，
p_i=$\{x^i_{pos} \in \mathbb{R}^4, x^i_{visn} \in \mathbb{R}^{d_{visn}}, x^i_{lang} \in \mathbb{R}^{d_{lang}}\}$
每个对象p_i由2维坐标x^i_{pos}、视觉特征x^i_{visn}、文本句子嵌入特征x^i_{lang}组成，
视觉特征和文本句子嵌入特征的维度分别为d_{visn}和d_{lang}.

Input preprocessing

合并位置信息，将输入特征映射到隐藏状态：$H^0_T = \{h^1_T, \dots, h^N_T\}$, $H^0_V = \{h^1_V, \dots, h^N_V\}$

$$h^i_T = W_T x^i_{lang} + W_P x^i_{pos}, h^i_V = W_V x^i_{visn} + W_P x^i_{pos}$$

其中矩阵$W_T \in \mathbb{R}^{d_h \times d_{lang}}$, $W_T \in \mathbb{R}^{d_h \times d_{visn}}$, $W_P \in \mathbb{R}^{d_h \times 4}$将特征投影到$d_h$维。

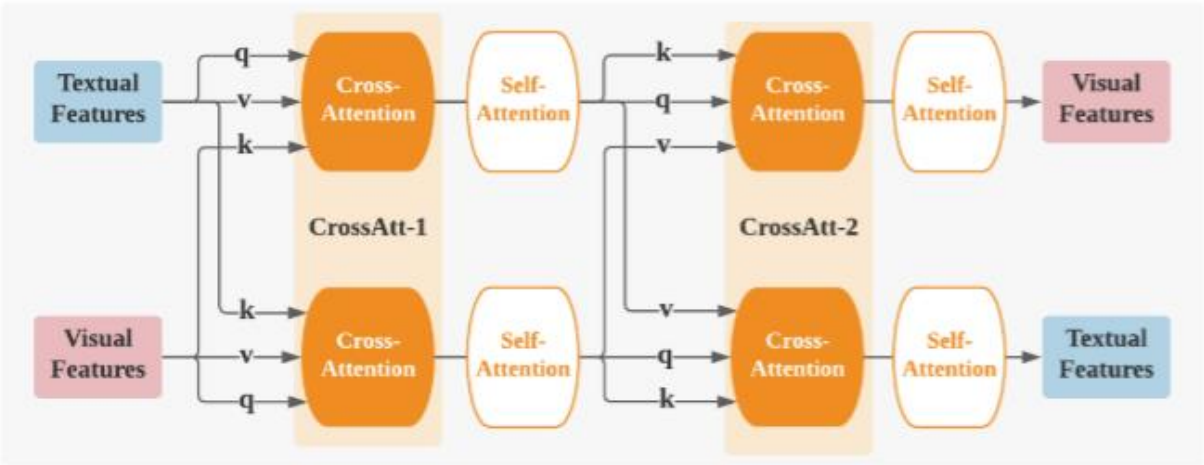
SelfDoc

单模态编码器包含多头注意、前馈层、残差连接和归一化层
对于多层单模态编码器中，令$H^l=\{h_1,\dots,h_N\}$为第l层编码特征，则H^0是合并了位置信息的输入特征向量。
输出特征有：

$$H_{att}^l = LN\left(f_{SelfAtt}\left(H^l\right)+H^l\right)$$
$$H^{l+1} = LN\left(f_{FF}\left(H_{att}^l\right)+H_{att}^l\right)$$

自注意函数定义如下：

$$f_{SelfAtt}\left(H^l\right) = softmax\left(\frac{q\left(H^l\right)k\left(H^l\right)^T}{\sqrt{d_k}}\right)v\left(H^l\right)$$



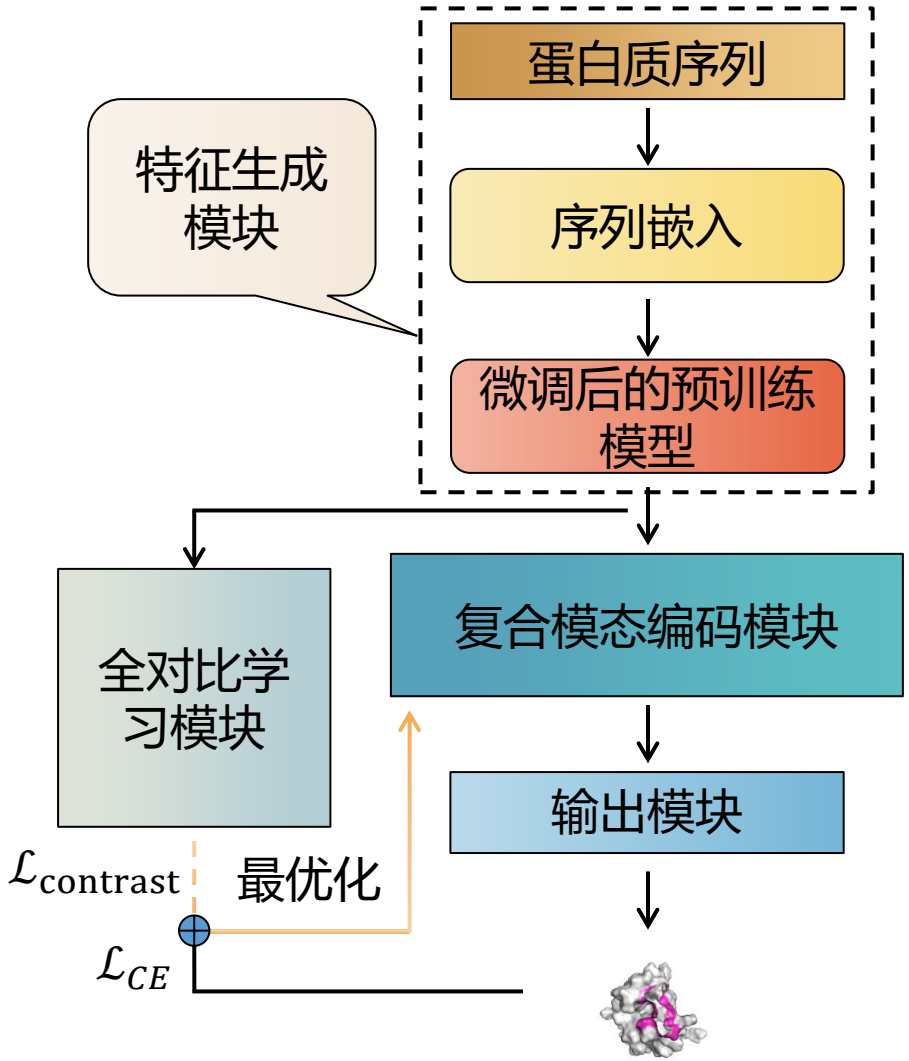
跨模态编码器类似于单模态编码器
只是把自注意函数替换成$f_{CrossAtt1}(\cdot)$或者$f_{CrossAtt2}(\cdot)$
H_T^l和H_V^l中添加下标T、V表示情态，分别是文本和视觉的中间表征
第一层识别语言和视觉信息之间的一致性，有：

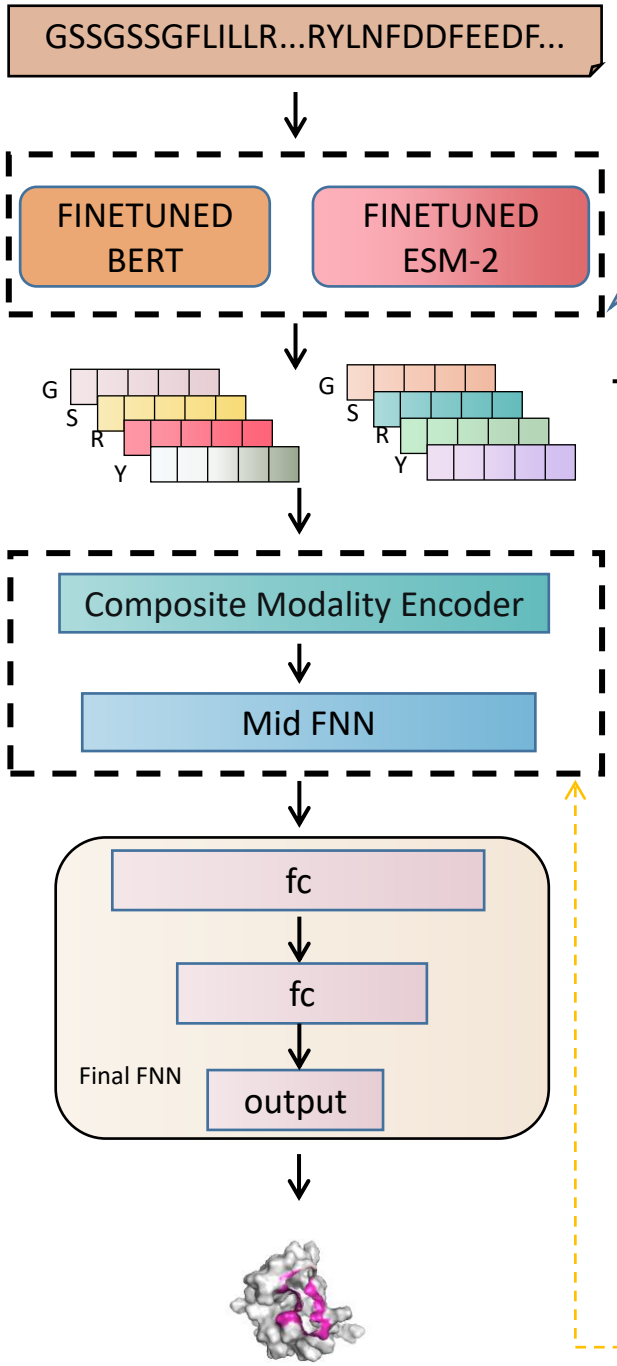
$$f_{CrossAtt1}\left(H_T^l\right) = softmax\left(\frac{q\left(H_T^l\right)k\left(H_V^l\right)^T}{\sqrt{d_k}}\right)v\left(H_T^l\right)$$
$$f_{CrossAtt1}\left(H_V^l\right) = softmax\left(\frac{q\left(H_V^l\right)k\left(H_T^l\right)^T}{\sqrt{d_k}}\right)v\left(H_V^l\right)$$

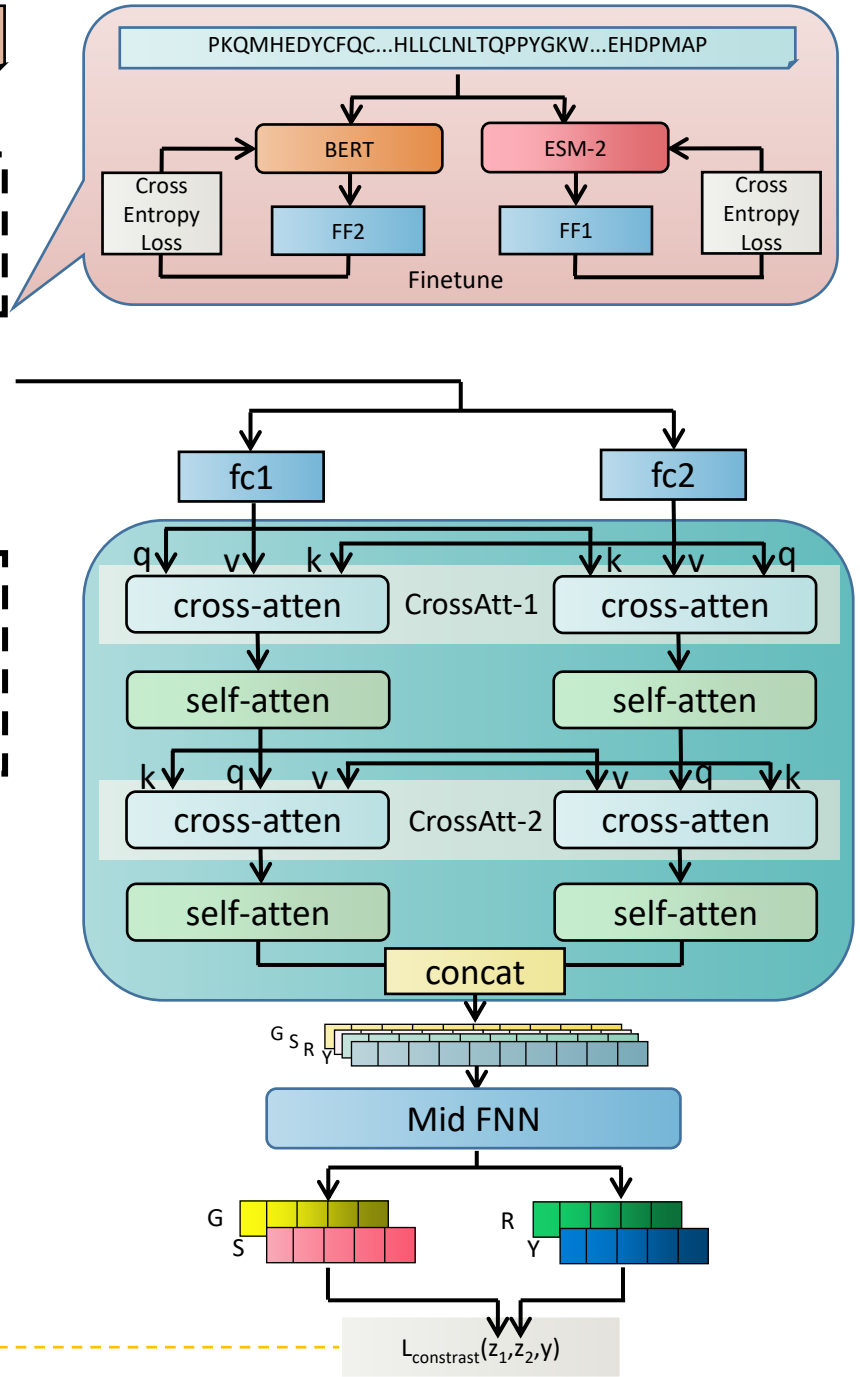
第二层是发现从一个模态到另一个模态的内在关系的操作，有：

$$f_{CrossAtt2}\left(H_T^l\right) = softmax\left(\frac{q\left(H_V^l\right)k\left(H_V^l\right)^T}{\sqrt{d_k}}\right)v\left(H_T^l\right)$$
$$f_{CrossAtt2}\left(H_V^l\right) = softmax\left(\frac{q\left(H_T^l\right)k\left(H_T^l\right)^T}{\sqrt{d_k}}\right)v\left(H_V^l\right)$$

Model-IV


蛋白质序列
特征生成模块
序列嵌入
微调后的预训练模型
全对比学习模块
复合模态编码模块
输出模块
$\mathcal{L}_{contrast}$
最优化
$\mathcal{L}_{CE}$


GSSGSSGFLILLR...RYLNFDDFEEDF...
FINETUNED BERT
FINETUNED ESM-2
G
S
R
Y
Composite Modality Encoder
Mid FNN
fc
fc
Final FNN
output


PKQMHEDYCFQC...HLLCLNLTQPPYGKW...EHDPMAP
BERT
ESM-2
Cross Entropy Loss
FF2
FF1
Cross Entropy Loss
Finetune
fc1
fc2
q
v
k
cross-atten
CrossAtt-1
self-atten
CrossAtt-2
concat
Mid FNN
$L_{constrast}(z_1,z_2,y)$

Result

Dataset1上的结果

Network	ACC	Pre	Sen	Spe	F1	AUC	MCC
Fc	0.943	0.480	0.405	0.975	0.441	0.83	0.413
Self-atten	0.948	0.550	0.358	0.983	0.434	0.83	0.419
Corss-atten	0.945	0.510	0.390	0.978	0.442	0.83	0.418
SelfDoc	0.950	**0.580**	0.347	0.985	0.435	**0.832**	**0.426**

Dataset2上的结果

Network	ACC	Pre	Sen	Spe	F1	AUC	MCC
Fc							
Self-atten							
Corss-atten	0.944	0.51	0.260	0.985	0.343	0.795	0.337
SelfDoc	0.942	0.48	0.283	0.982	0.357	0.804	0.341

Effect of batch size on experimental results

Dataset1上的结果

Batchsize	ACC	Pre	Sen	Spe	F1	AUC	MCC
2							
4							
6							
8	0.950	0.58	0.347	0.985	0.435	0.832	0.426
10							
12							
14							
16							
18							
20							

Comparisons with other methods

Dataset1上的结果

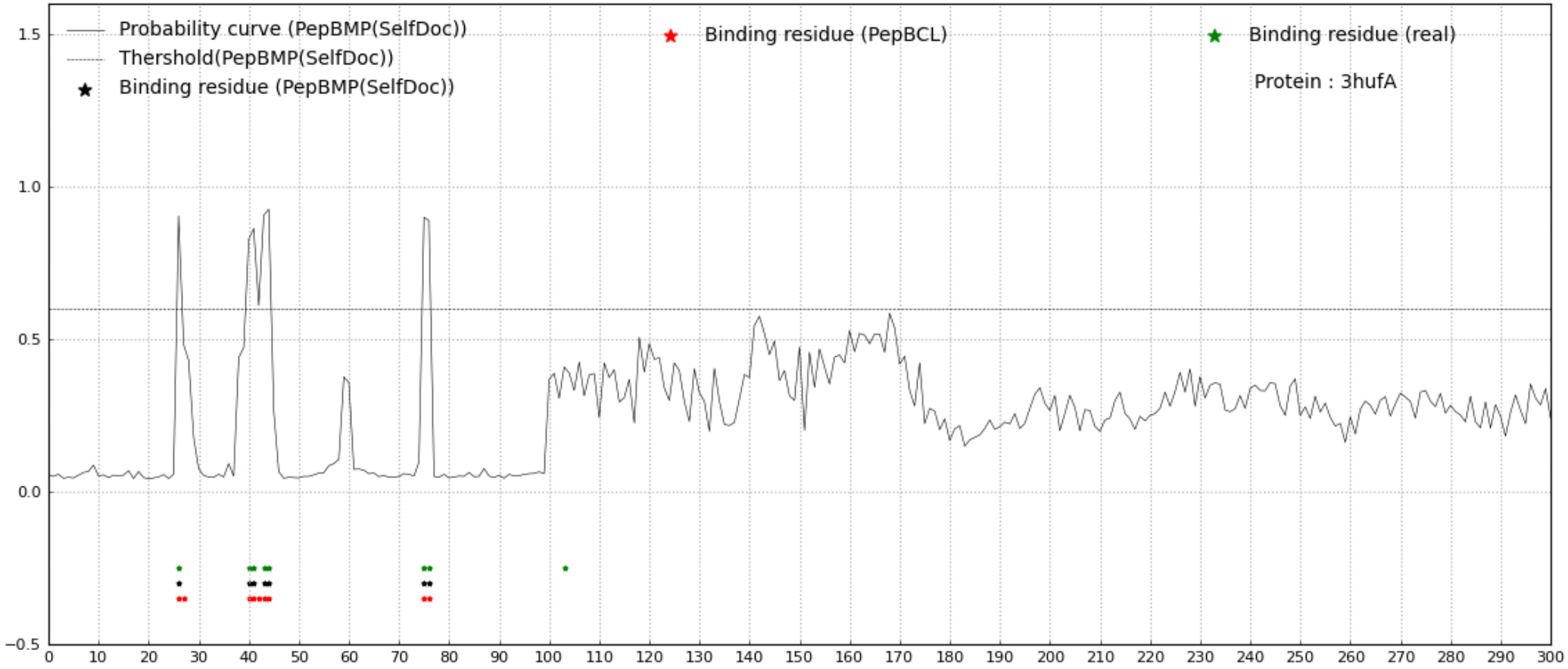
Methods	**ACC**	**Pre**	**Sen**	**Spe**	**F1**	**AUC**	**MCC**
Pepsite*	-	-	0.180	0.970	-	0.610	0.200
Peptimap*	-	-	0.320	0.950	-	0.630	0.270
SPRINT-Seq	-	-	0.210	0.960	-	0.680	0.200
SPRINT-Str*	-	-	0.240	0.980	-	0.780	0.290
PepBind	-	0.469	0.344	-	-	0.793	0.372
Visual	-	-	0.670	0.680	-	0.730	0.170
PepNN-Seq	-	-	-	-	-	0.805	0.278
PepNN-Struct*	-	-	-	-	-	**0.841**	0.321
PepBCL	-	0.540	0.315	0.984	-	0.815	0.385
PepBMP	0.950	**0.58**	0.347	0.985	0.435	0.832	**0.426**

Comparisons with other methods

Dataset2上的结果

Methods	**ACC**	**Pre**	**Sen**	**Spe**	**F1**	**AUC**	**MCC**
PepBind	-	0.450	0.317	-	-	0.767	0.348
PepNN-Seq	-	-	-	-	-	0.792	0.251
PepNN-Struct*	-	-	-	-	-	**0.838**	0.301
PepBCL	-	0.470	0.252	0.983	-	0.804	0.312
PepBMP	0.942	**0.480**	0.283	0.982	0.357	0.804	**0.341**

Case


Probability curve (PepBMP(SelfDoc))
Thershold(PepBMP(SelfDoc))
Binding residue (PepBMP(SelfDoc))
Binding residue (PepBCL)
Binding residue (real)
Protein : 3hufA
1.5
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300