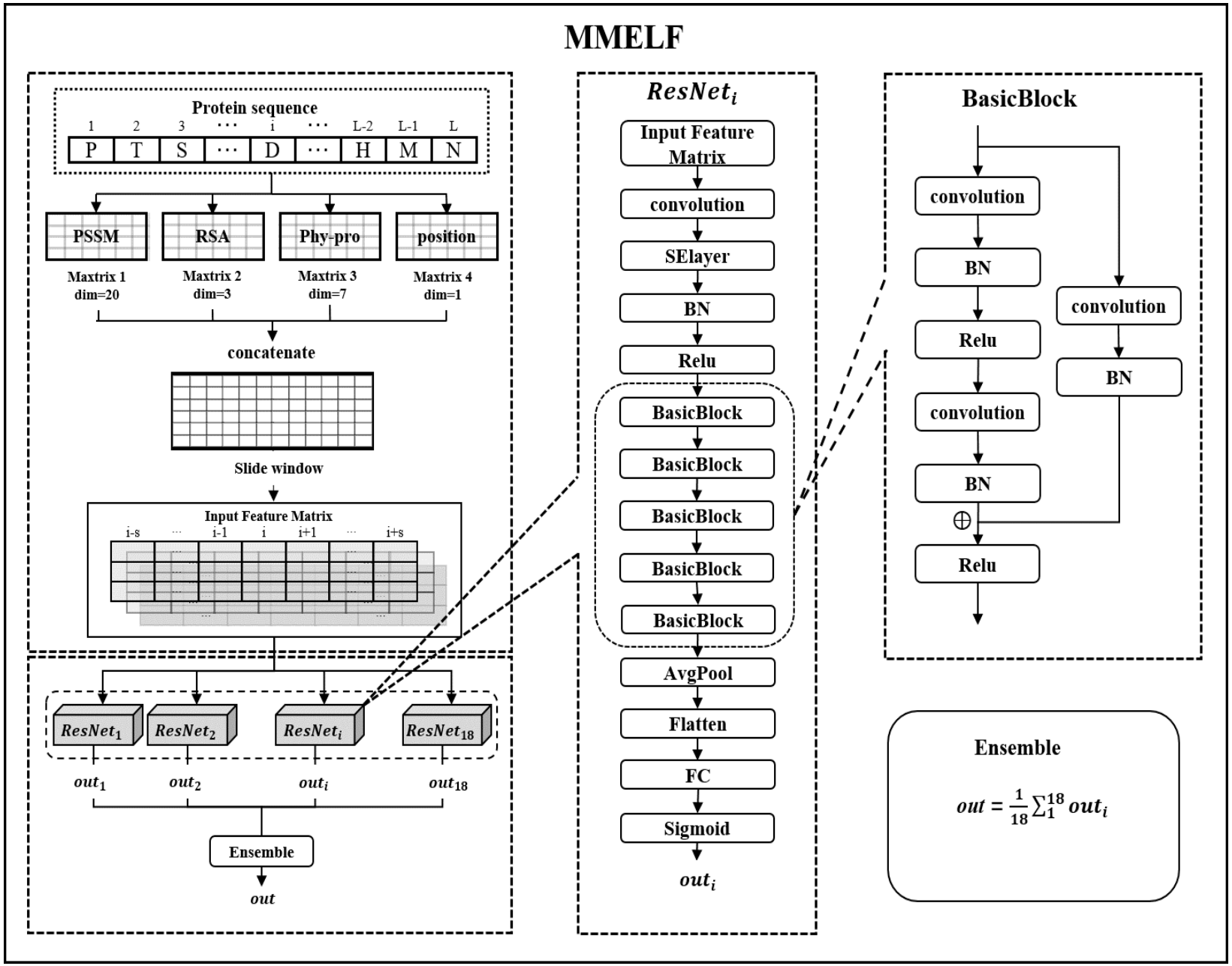
Prediction of Protein-protein Interaction Site

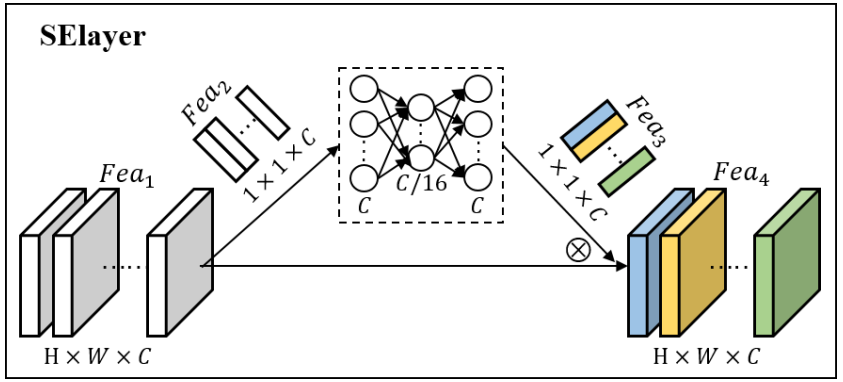
汇报人：董明

Dataset

Dataset	proteins	Total	Binding	Non-binding	% Rate of Binding to Tatal
Training set	9000	3624662	385533	3239129	10.6%
Validation set	841	333426	35582	297844	10.7%
Dset_72	72	18140	1923	16217	10.6%
Dset_164	164	33681	6,096	27585	18.1%
Dset_186	186	36219	5517	30702	15.2%
Dset_355	355	95940	11467	84473	12.0%
Dset_448	448	116500	15810	100690	13.6%

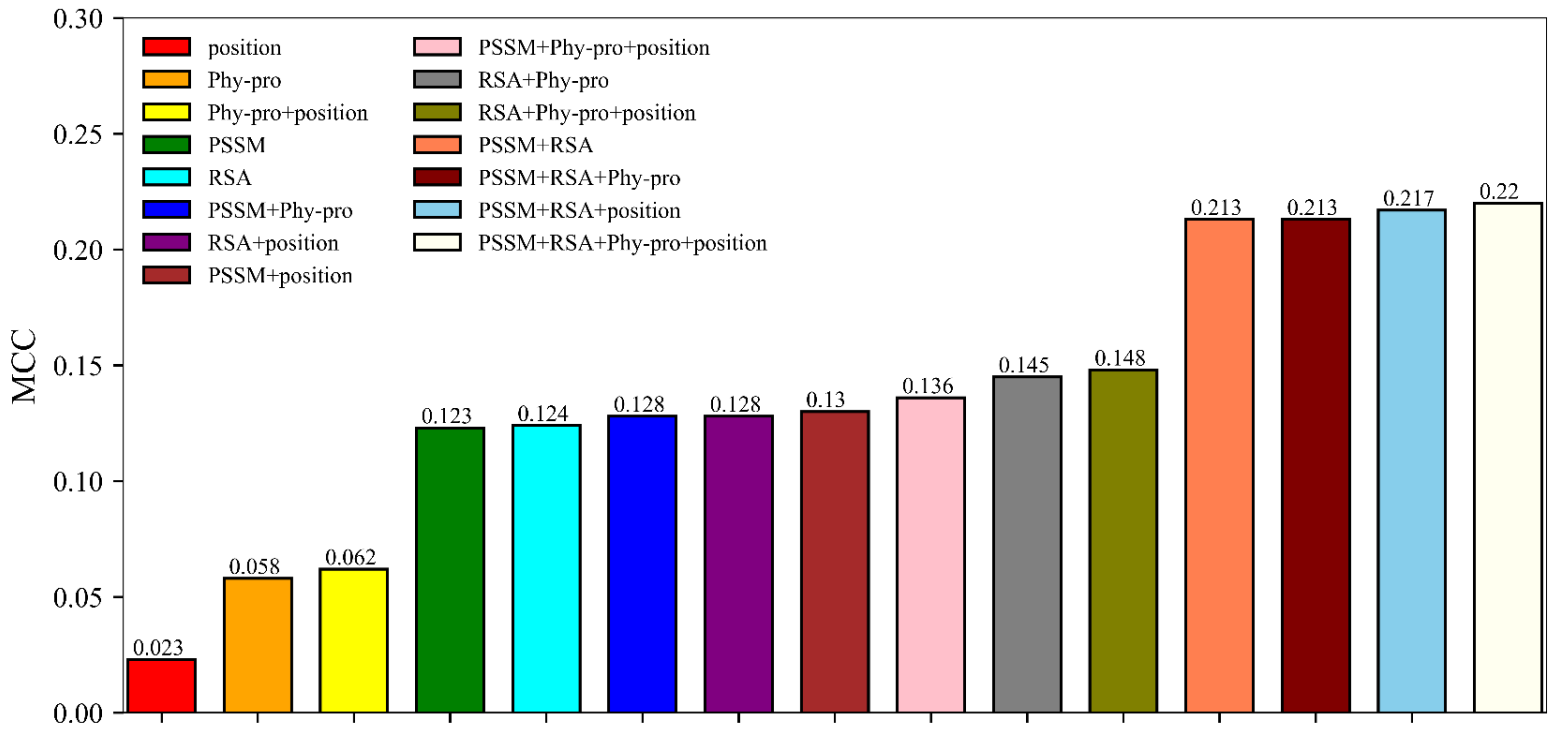
MMELF


MMELF
Protein sequence
1 2 3 ··· i ··· L-2 L-1 L
P T S ··· D ··· H M N
PSSM
RSA
Phy-pro
position
Maxtrix 1 dim=20
Maxtrix 2 dim=3
Maxtrix 3 dim=7
Maxtrix 4 dim=1
concatenate
Slide window
Input Feature Matrix
i-s ··· i-1 i i+1 ··· i+s
ResNet1
ResNet2
ResNeti
ResNet18
out1
out2
outi
out18
Ensemble
out
ResNeti
Input Feature Matrix
convolution
SElayer
BN
Relu
BasicBlock
BasicBlock
BasicBlock
BasicBlock
BasicBlock
AvgPool
Flatten
FC
Sigmoid
outi
BasicBlock
convolution
BN
Relu
convolution
BN
convolution
BN
Relu
Ensemble
out = 1/18 ∑1^18 outi

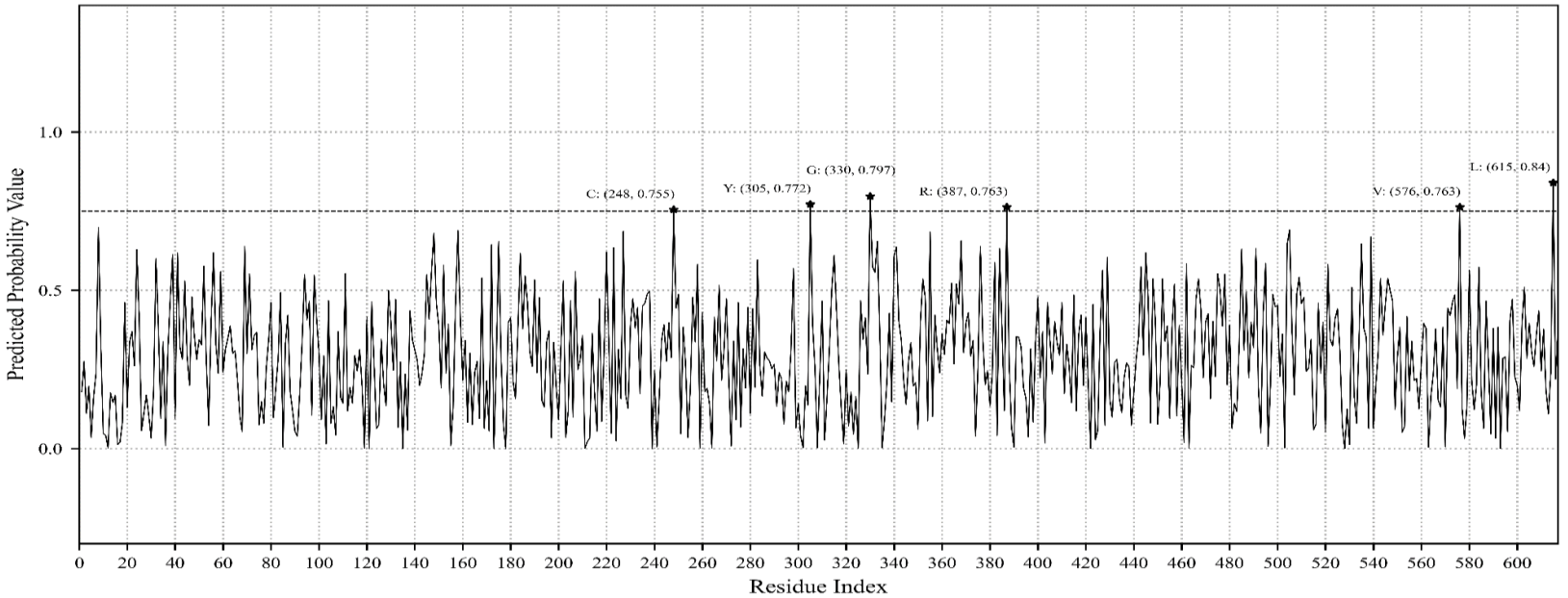

SElayer
Fea1
Fea2
1×1×C
C
C/16
C
Fea3
1×1×C
Fea4
H × W × C
H × W × C

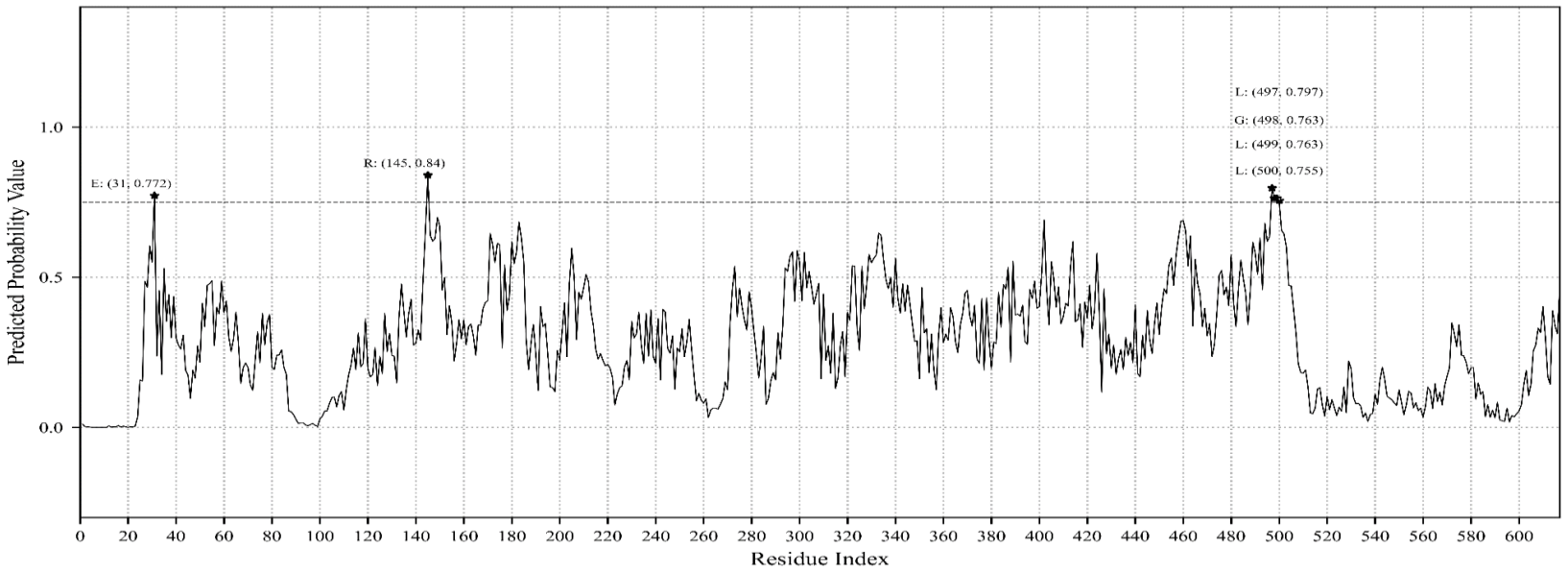
验证不同特征组合的效果

Window size = 1

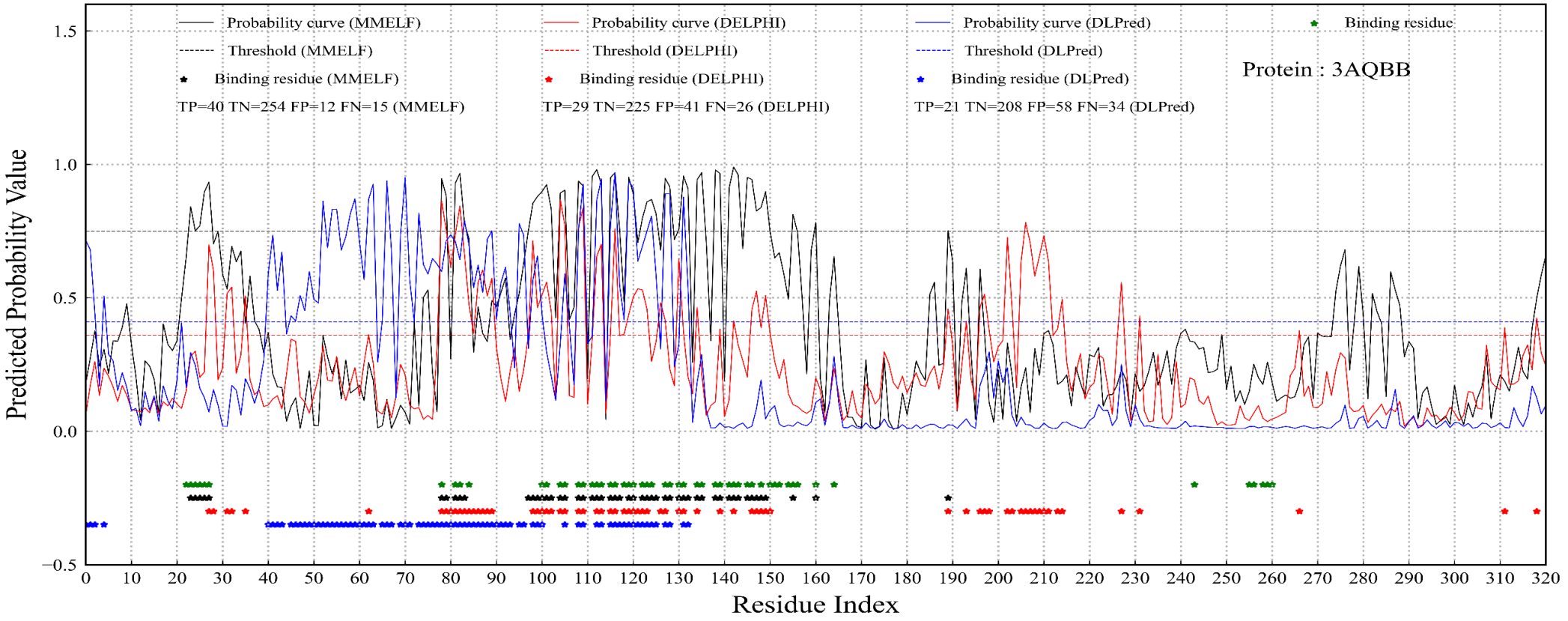


预测结果波动大的问题


Predicted Probability Value
1.0
0.5
0.0
C: (248, 0.755)
Y: (305, 0.772)
G: (330, 0.797)
R: (387, 0.763)
V: (576, 0.763)
L: (615, 0.84)
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540 560 580 600
Residue Index

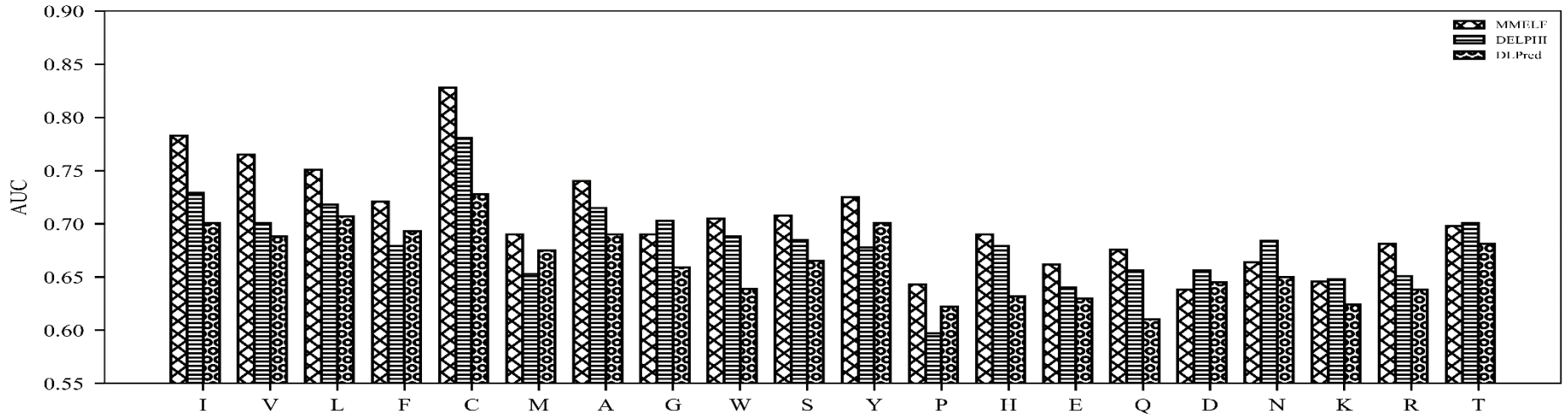

Predicted Probability Value
1.0
0.5
0.0
E: (31, 0.772)
R: (145, 0.84)
L: (497, 0.797)
G: (498, 0.763)
L: (499, 0.763)
L: (500, 0.755)
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540 560 580 600
Residue Index

修改CASE图

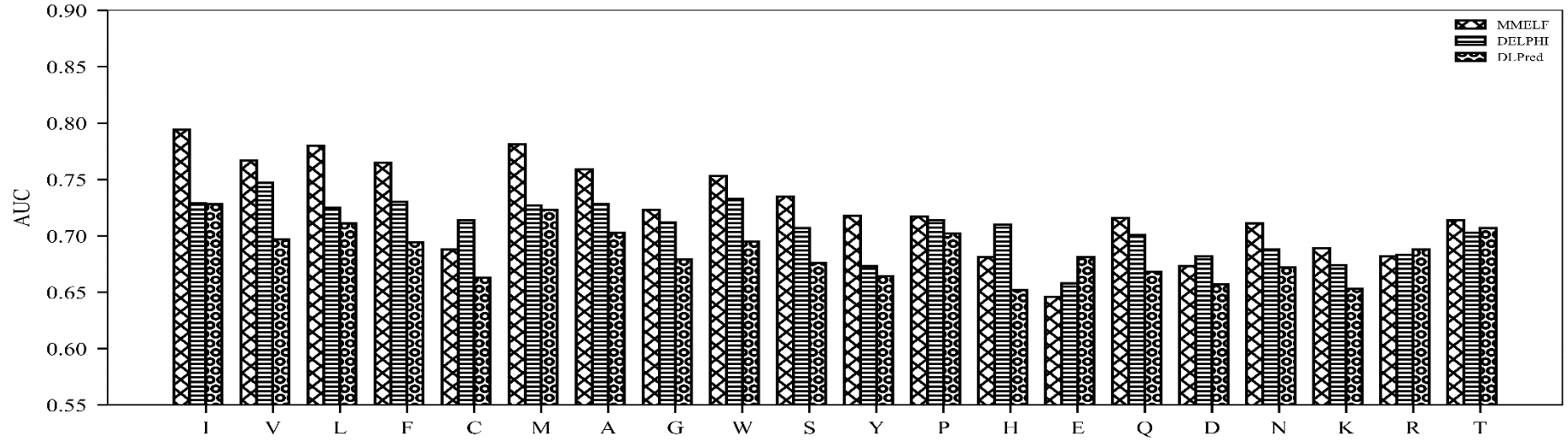

Probability curve (MMELF)
Threshold (MMELF)
Binding residue (MMELF)
TP=40 TN=254 FP=12 FN=15 (MMELF)
Probability curve (DELPHI)
Threshold (DELPHI)
Binding residue (DELPHI)
TP=29 TN=225 FP=41 FN=26 (DELPHI)
Probability curve (DLPred)
Threshold (DLPred)
Binding residue (DLPred)
TP=21 TN=208 FP=58 FN=34 (DLPred)
Binding residue
Protein : 3AQBB
Predicted Probability Value
Residue Index
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
310
320

统计模型在每种氨基酸上的表现，使用AUC值进行评估。
对比方法：DELPHI, DLPred （这两个方法最近）数据集Dset164, Dset_186

Dset164



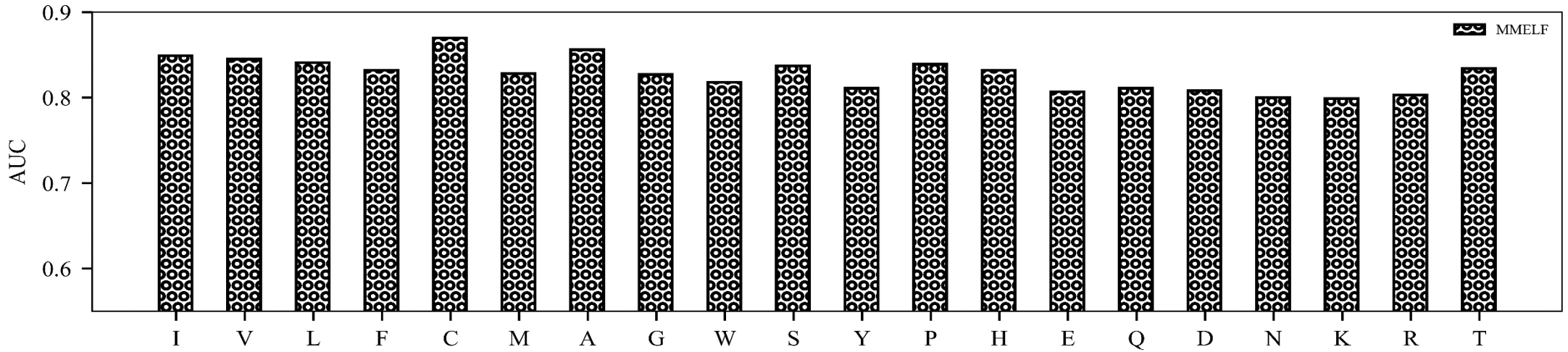
Dset_186



统计验证集每种残基的AUC值

模型数=18 的条件下，求出所有残基的预测值，然后根据残基类型分别统计

Validation set																			
Residue type	I	V	L	F	C	M	A	G	W	S	Y	P	H	E	Q	D	N	K	R	T
AUC	0.849	0.845	0.841	0.832	0.87	0.828	0.856	0.827	0.818	0.837	0.811	0.839	0.832	0.807	0.811	0.808	0.8	0.799	0.803	0.834
number of residues	18036	21657	31537	12746	4828	7791	26429	22318	3769	23379	10079	16985	7755	23683	13700	18755	13999	19991	17721	18268



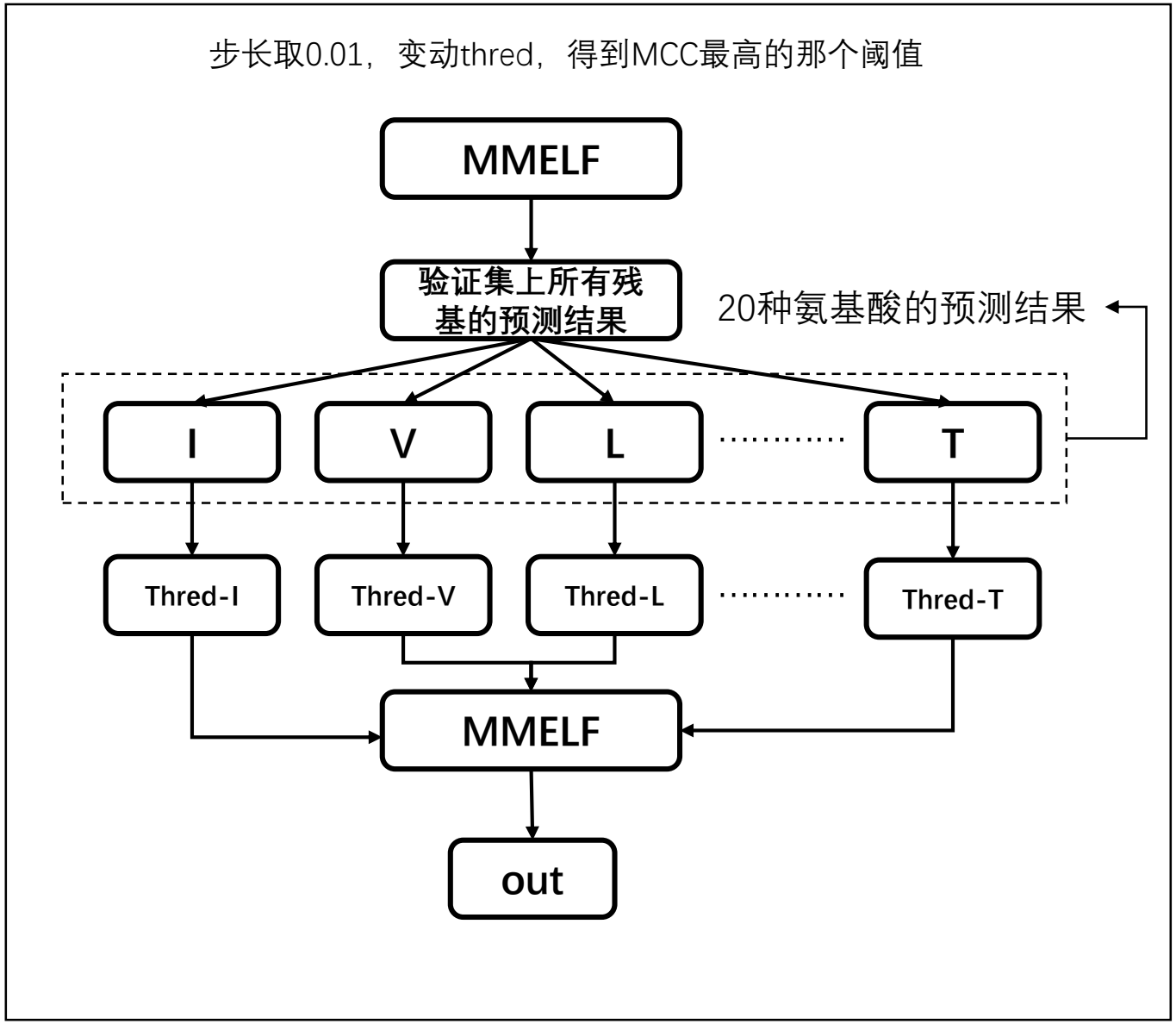
统计18个ResNet的每种残基的AUC

Validation set & AUC value																			
Residue type	I	V	L	F	C	M	A	G	W	S	Y	P	H	E	Q	D	N	K	R	T
ResNet 1	**0.827**	0.822	0.813	0.808	0.841	0.804	0.83	0.798	0.789	0.812	0.782	0.812	0.805	**0.787**	0.785	0.779	0.77	0.768	0.77	**0.81**
ResNet 2	0.824	**0.823**	0.817	**0.81**	**0.844**	0.804	**0.837**	**0.807**	0.783	0.815	0.782	0.809	0.808	0.784	0.787	**0.783**	0.772	**0.773**	0.777	0.808
ResNet 3	0.819	**0.823**	**0.822**	0.806	0.841	**0.806**	0.832	0.802	0.786	0.812	0.784	**0.813**	0.798	0.781	0.779	0.78	0.77	**0.773**	0.775	**0.81**
ResNet 4	0.818	0.821	0.815	0.805	0.843	0.799	0.827	0.791	0.788	0.812	0.777	0.812	0.805	0.78	0.787	0.779	**0.774**	0.768	**0.778**	0.804
ResNet 5	0.816	0.819	0.814	0.805	0.834	0.803	0.83	0.797	0.79	**0.817**	**0.785**	0.812	0.804	0.777	**0.789**	0.78	0.771	0.77	0.775	0.806
ResNet 6	0.82	0.819	0.815	0.805	0.837	0.8	0.83	0.804	0.785	0.808	0.784	0.808	0.806	0.773	0.777	0.776	**0.774**	0.766	0.771	0.807
ResNet 7	0.811	0.806	0.803	0.798	0.84	0.797	0.818	0.784	0.79	0.801	0.771	0.806	**0.809**	0.777	0.784	0.775	0.767	0.76	0.772	0.797
ResNet 8	0.814	0.815	0.811	0.804	0.832	0.804	0.826	0.795	**0.797**	0.808	0.781	0.807	0.8	0.771	0.778	0.777	0.769	0.765	0.773	0.8
ResNet 9	0.802	0.803	0.801	0.788	0.828	0.778	0.812	0.784	0.775	0.799	0.772	0.804	0.797	0.767	0.773	0.766	0.757	0.749	0.764	0.8
ResNet 10	0.815	0.815	0.809	0.799	0.834	0.792	0.819	0.801	0.781	0.805	0.78	0.803	0.804	0.769	0.774	0.77	0.772	0.76	0.772	0.801
ResNet 11	0.815	0.815	0.807	0.799	0.825	0.792	0.824	0.788	0.778	0.803	0.775	0.802	0.784	0.763	0.773	0.773	0.768	0.769	0.759	0.796
ResNet 12	0.807	0.796	0.801	0.791	0.822	0.784	0.815	0.784	0.78	0.798	0.768	0.806	0.801	0.775	0.777	0.77	0.757	0.758	0.768	0.794
ResNet 13	0.816	0.807	0.805	0.796	0.822	0.787	0.823	0.792	0.78	0.804	0.772	0.801	0.8	0.773	0.775	0.77	0.767	0.76	0.771	0.796
ResNet 14	0.809	0.804	0.802	0.789	0.834	0.793	0.817	0.796	0.779	0.797	0.773	0.8	0.789	0.769	0.773	0.765	0.751	0.755	0.768	0.798
ResNet 15	0.812	0.81	0.806	0.8	0.831	0.793	0823	0.793	0.795	0.802	0.775	0.808	0.797	0.771	0.778	0.775	0.765	0.77	0.766	0.8
ResNet 16	0.812	0.815	0.807	0.794	0.836	0.804	0.826	0.794	0.781	0.798	0.771	0.799	0.789	0.772	0.77	0.769	0.752	0.757	0.765	0.801
ResNet 17	0.809	0.806	0.802	0.795	0.824	0.789	0.819	0.788	0.776	0.793	0.762	0.794	0.789	0.76	0.764	0.765	0.763	0.749	0.761	0.791
ResNet 18	0.804	0.796	0.794	0.785	0.83	0.786	0.806	0.77	0.779	0.788	0.763	0.792	0.788	0.765	0.775	0.756	0.751	0.752	0.755	0.782

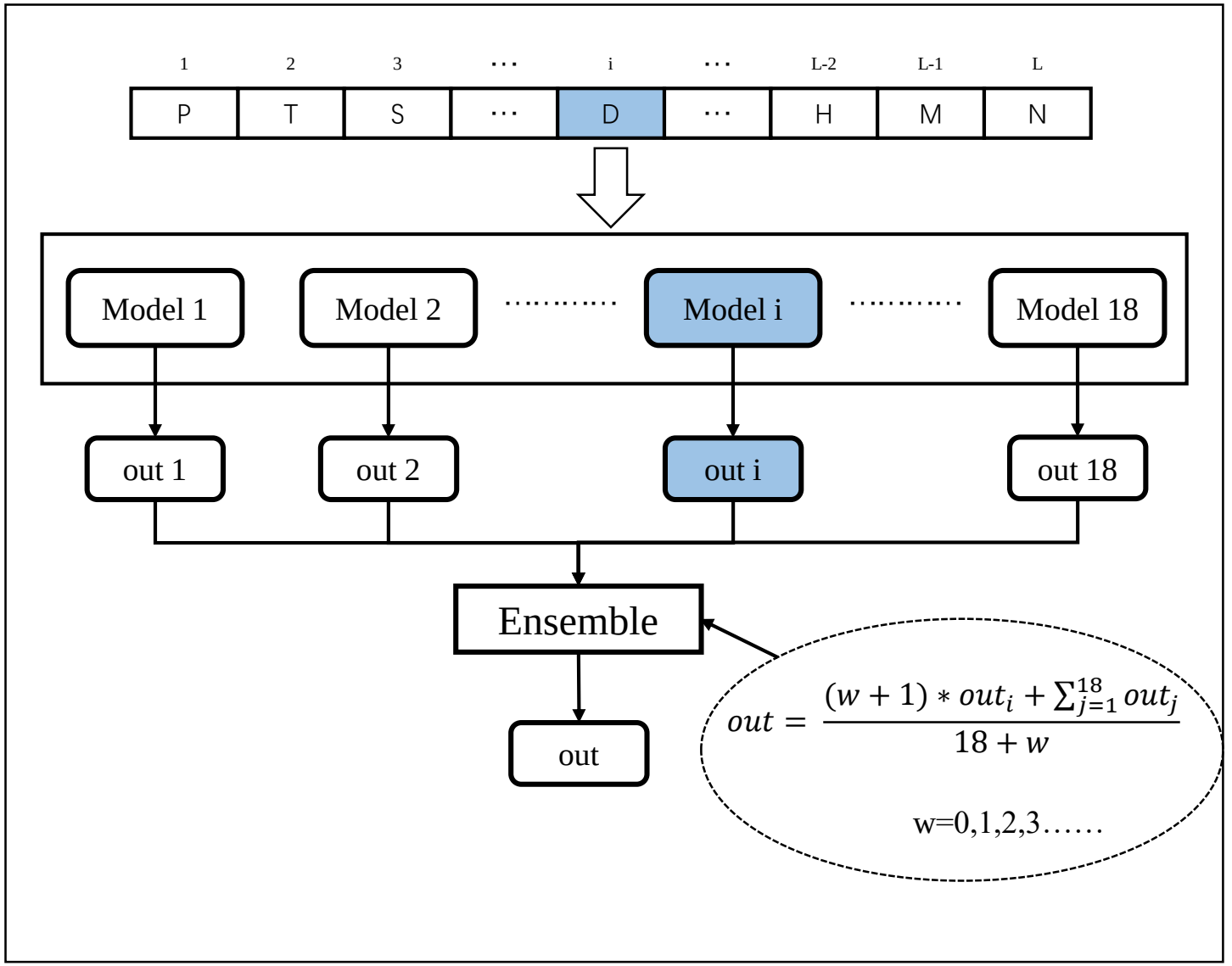
集成方式-细分不同残基类型的分割阈值

动机：原先是在所有类型残基上统计得到的阈值，考虑到模型对于不同氨基酸的敏感度不同，所以看看细分阈值会不会取得更好的结果

先在验证集上得到所有残基的预测结果，然后按氨基酸种类分类，求每种氨基酸的分割阈值，在根据这些阈值去做预测



集成方式-对擅长预测某残基的模型进行加权



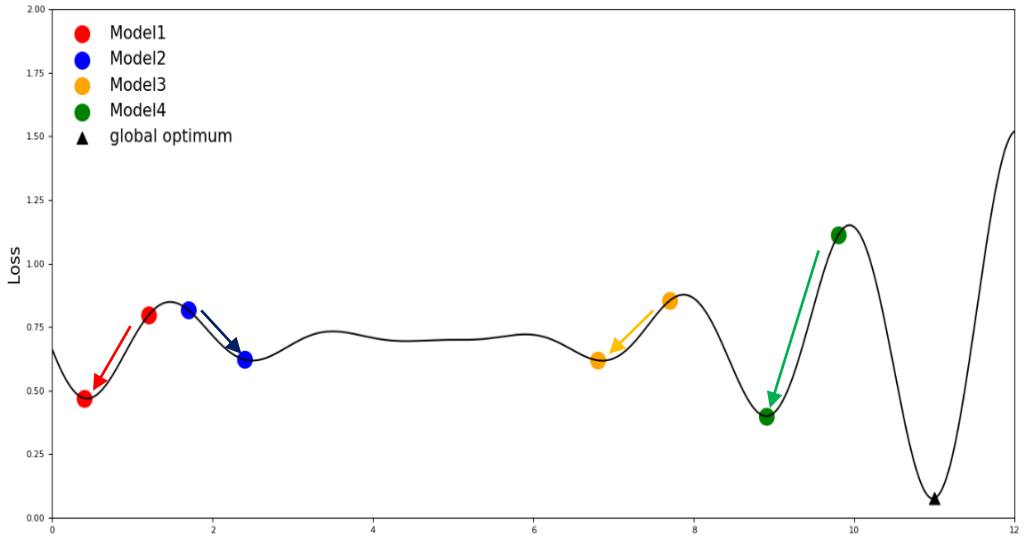
Result

Validation set						
Ensemble	ACC	PRE	SPE	SEN	MCC	F1
Mean-w-0	0.878	0.434	0.927	0.469	0.383	0.451
Mean-w-2	0.881	0.442	0.931	0.457	0.383	0.45
Mean-w-4	0.88	0.439	0.93	0.459	0.382	0.449
Mean-w-6	0.876	0.426	0.923	0.477	0.381	0.45
Mean-w-8	0.887	0.468	0.943	0.42	0.381	0.443
Mean-w-10	0.883	0.451	0.936	0.438	0.38	0.445
Mean-res	0.878	0.431	0.927	0.46	0.377	0.445

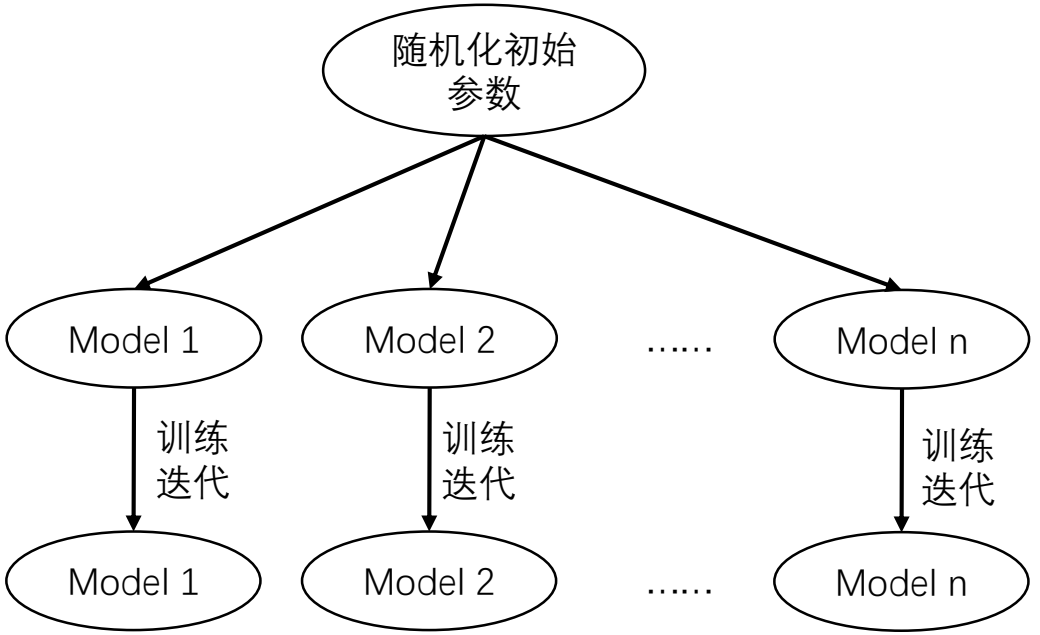
在权重变大的时候，性能反而下降的原因分析:

MMELF本身是利用模型的多样性提升性能的，集成的模型之间相似度低效果会好些，但是通过加大个别模型的权重，可能导致了，个别模型“个性”更加鲜明，从而弱化了模型之间多样性的效果。

论文中对于不同初始随机参数能提升模型性能的解释



由于模型不同的初始化参数，根据梯度下降的特点，它很可能会找一个最近的局部最低点收敛。这样得到的模型一般不是最好的。



这些训练所得的模型具有较高的预测方差，对此统计了这些模型之间的预测相关性以及模型参数之间的差异性

假设：灰色区域是目测目标，当预测点落在灰色区域内，则判定预测正确，反之预测错误。

从图中可以看出当灰色区域变窄的时候，也就是预测要求变高的时候，方差较高的模型的精度将下降

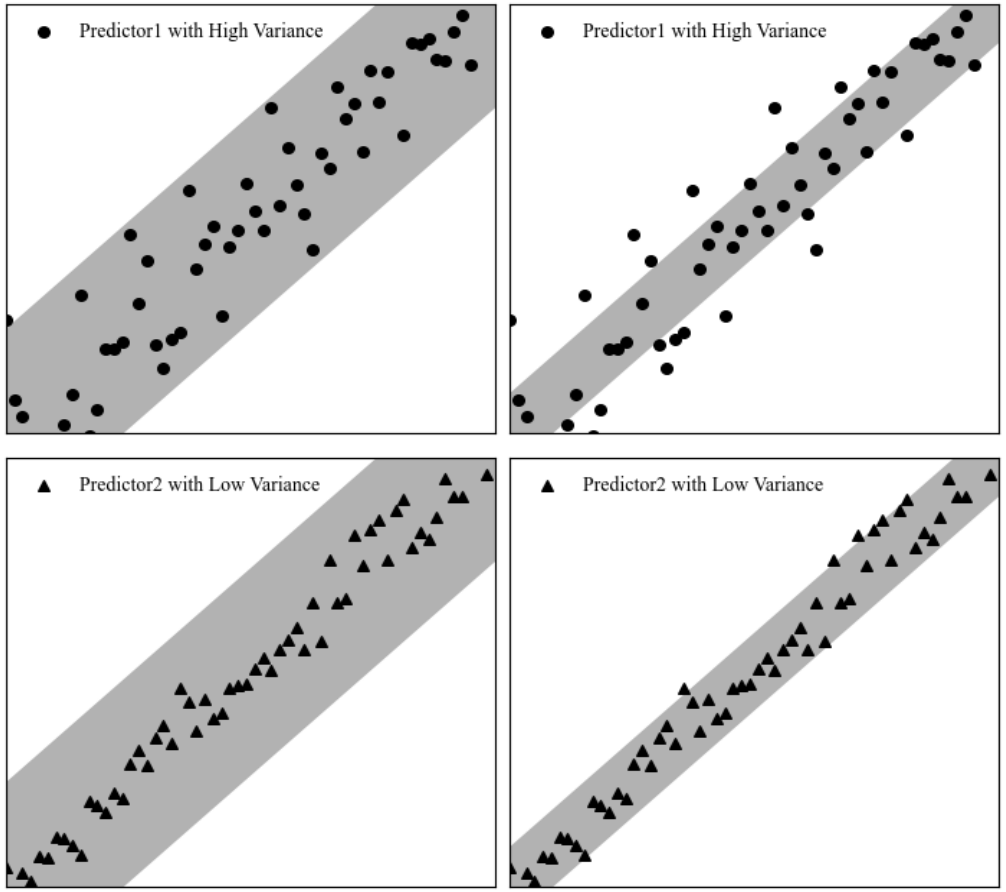
均值法能提高预测精度的依据

X之间完全不独立

$$Var\left(\frac{\sum X_i}{n}\right) = Var(X_i)$$

X之间相互独立

$$Var\left(\frac{\sum X_i}{n}\right) = \frac{Var(X_i)}{n}$$



均值集成不能阻止单个模型陷入局部最优，但多个模型联合起来可以缓解单个模型陷入局部最优的弊端，从而提升精度

Feature Plan

1.完善论文
2.再想一些策略

Result

Dset_72						
Ensemble	ACC	PRE	SPE	SEN	MCC	F1
Mean-all	0.795	0.245	0.837	0.447	0.221	0.317
Mean-res	0.784	0.235	0.823	0.459	0.214	0.311

Dset_164						
Ensemble	ACC	PRE	SPE	SEN	MCC	F1
Mean-all	0.741	0.345	0.799	0.478	0.247	0.401
Mean-res	0.728	0.333	0.779	0.499	0.241	0.399

Dset_186						
Ensemble	ACC	PRE	SPE	SEN	MCC	F1
Mean-all	0.757	0.318	0.800	0.520	0.266	0.395
Mean-res	0.738	0.299	0.775	0.534	0.249	0.383

Dset_355						
Ensemble	ACC	PRE	SPE	SEN	MCC	F1
Mean-all	0.846	0.398	0.886	0.556	0.384	0.464
Mean-res	0.834	0.372	0.870	0.569	0.369	0.450