多视角知识融合的蛋白质功能预测

南京农业大学 人工智能学院
汇报人：朱一亨
2024年08月19日

研究背景

- 现有的蛋白质功能预测方法分类

1. 基于模板匹配的方法
2. 基于统计机器学习的方法
3. 基于深度学习的方法

3.1. 手工设计特征表示 （One-hot encoding、PSSM）

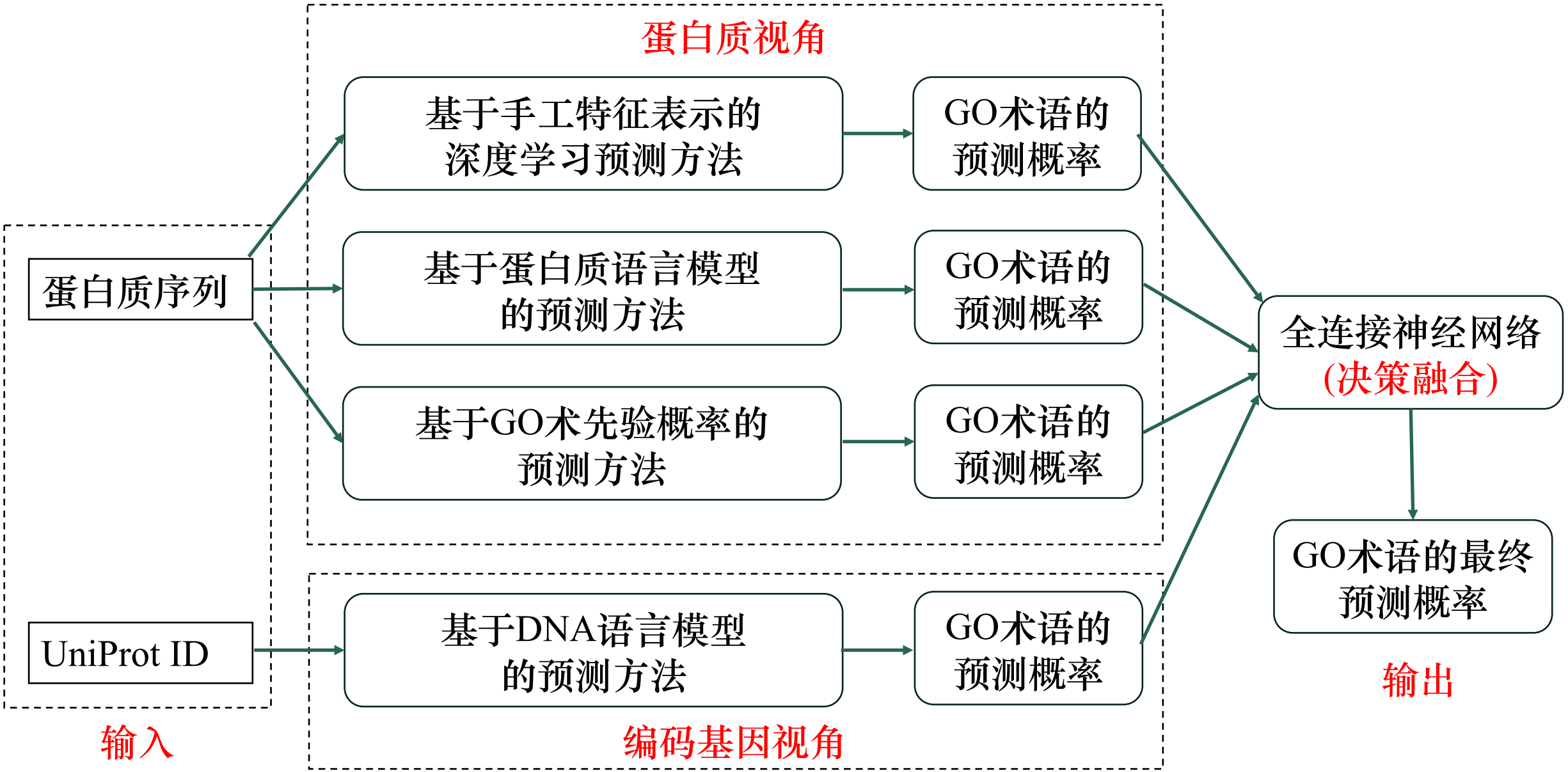
3.2. 大语言模型特征表示 （ESM、ProtTrans） （2022年后成为主流方法）

现有方法的不足和挑战

- 目前大部分方法只采用大语言模型抽取蛋白质的特征表示，完全抛弃了传统的手工特征表示方法。
  - DeepGO-SE，Nature Machine Intelligence (ESM2)，2024
  - SPROF-GO，BIB (ProtTrans)， 2023
  - ATGO，PLOS CB (ESM-1b)，2023

- 现有的蛋白质功能预测方法只注重从蛋白质自身挖掘知识，忽略了编码基因中的知识。

Multi-View Knowledge Fusion for GO Prediction (MVK-GO)



数据集构建

(1) 从蛋白质功能注释数据库GOA中，下载全部134778个蛋白质

(2) 选择在UniProt 数据库中状态为“Reviewed”的80653个蛋白质

(3) Training dataset: 70212 proteins, before 2020-06-30

(4) Validation dataset: 974 proteins, 2020-07-01 between 2021-06-30

(5) Test dataset: 1522 proteins, 2021-07-01 between 2023-06-30

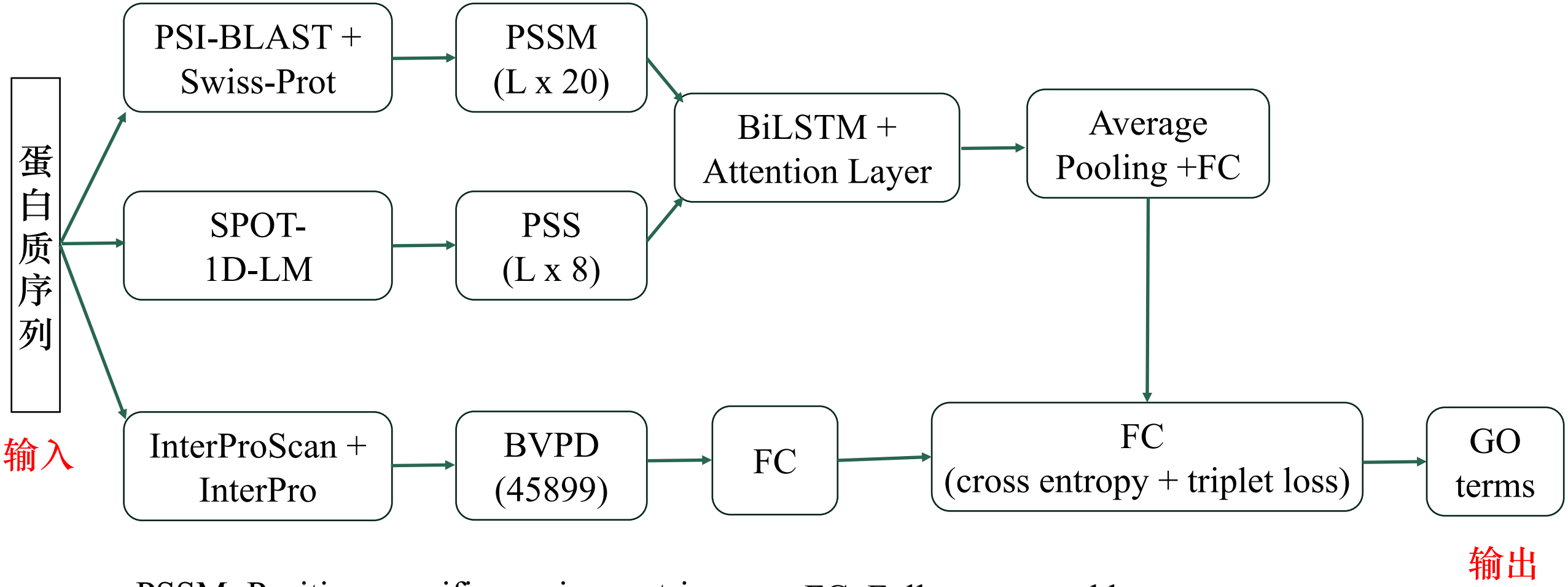
(6) CD-HIT: (sequence identity<30%, training, validation, test datasets)

现有主流蛋白质功能预测方法的性能比较

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Validation Dataset (974 proteins)	BLAST	0.627	0.409	0.502	7.47	23.14	8.26	0.397	0.240	0.306
	ESM2 + FC	0.679	0.430	0.618	7.09	22.59	7.34	0.627	0.351	0.635
	ProtTrans + FC	0.678	0.415	**0.639**	7.23	22.80	7.18	0.632	0.347	0.590
	ATGO	0.689	0.415	0.604	6.86	24.63	7.49	0.650	0.362	0.641
	MVKGO	**0.701**	**0.439**	0.634	**6.76**	**21.90**	**7.14**	**0.707**	**0.388**	**0.650**
Test Dataset (1522 proteins)	BLAST	0.645	0.395	0.495	7.89	25.11	8.67	0.377	0.227	0.274
	ESM2 + FC	0.687	0.427	0.617	7.33	23.96	7.67	0.616	0.346	0.618
	ProtTrans + FC	0.680	0.424	0.623	7.58	23.95	7.64	0.621	0.355	0.581
	ATGO	0.691	0.424	0.607	7.24	23.99	7.87	0.658	0.361	0.625
	MVKGO	**0.706**	**0.446**	**0.630**	**7.00**	**23.53**	**7.62**	**0.710**	**0.381**	**0.641**

基于手工特征表示的深度学习预测方法

Hand-craft feature-based deep learning method for GO prediction (HCFGO)



PSSM: Position-specific scoring matrix

PSS: Predicted secondary structure

FC: Fully connected layer

BVPD: Binary vector for protein family

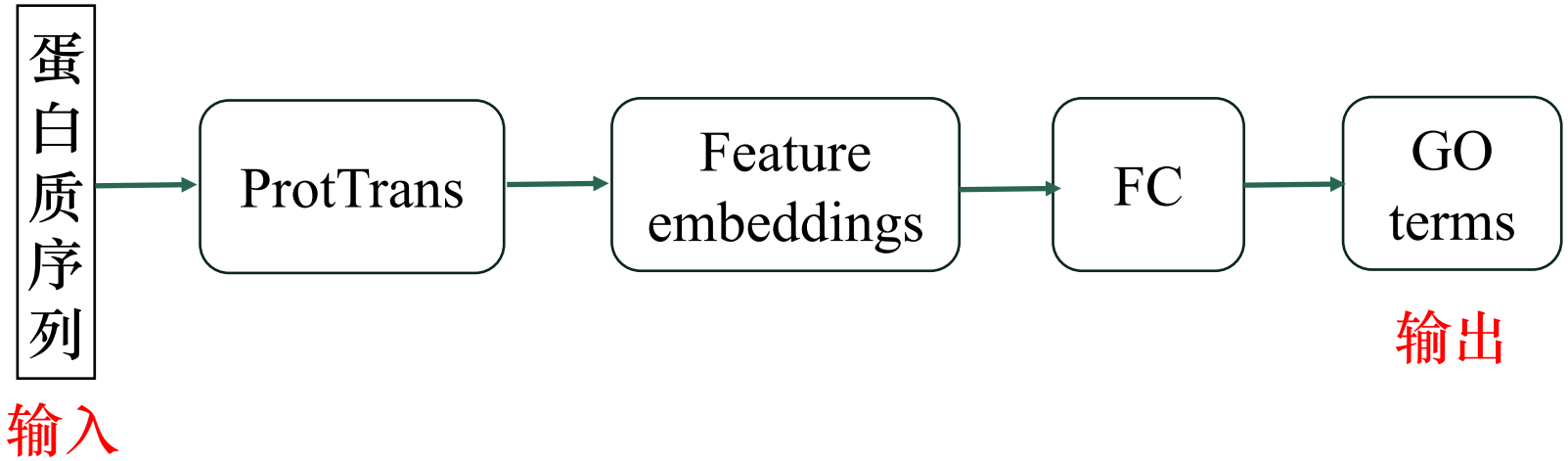
HCFGO消融实验

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Validation Dataset (974 proteins)	PSSM	0.607		0.559	8.15		8.06	0.515		0.521
	PSS									
	InterPro	0.662	0.402	0.575	7.36	22.95	8.00	0.606	0.309	0.547
	PSSM + PSSM + InterPro (C)	0.667	0.407	**0.594**	7.20	22.69	**7.54**	0.607	0.329	0.564
	PSSM + PSSM + InterPro (CT)	**0.675**	**0.410**	0.592	**7.01**	**22.67**	7.56	**0.623**	**0.349**	**0.568**
Test Dataset (1522 proteins)	PSSM	0.611		0.537	8.45		8.70	0.525		0.490
	PSS									
	InterPro	0.664	0.389	0.570	7.63	24.48	8.33	0.615	0.285	0.515
	PSSM + PSSM + InterPro (C)	0.679	0.403	0.578	7.41	24.03	**8.11**	0.619	0.320	0.535
	PSSM + PSSM + InterPro (CT)	**0.682**	**0.412**	**0.580**	**7.23**	**23.91**	8.14	**0.630**	**0.340**	**0.539**

C: Cross-Entropy　　CT: Cross-Entropy + Triplet Loss　　PSSM + PSSM + InterPro (CT) =HCFGO

基于蛋白质语言模型的预测方法

Protein language model-based method for GO prediction (PLMGO)



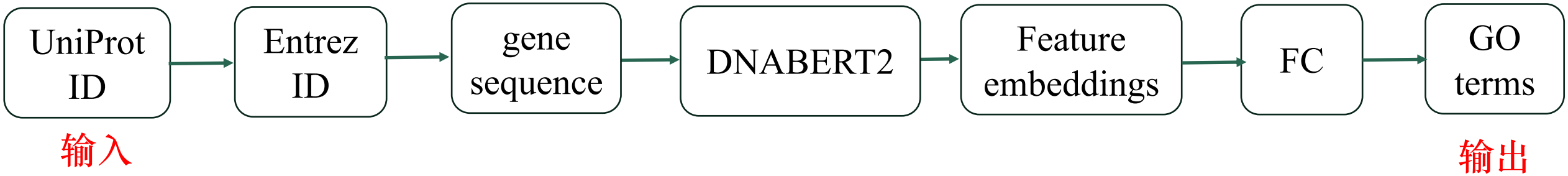
基于GO术先验概率的预测方法

Navie Bayes-based method for GO prediction (Naive)

$$S(G_i, P_j) = \frac{N_{G_i}}{N_D}$$

基于DNA语言模型的预测方法

DNA language model-based method for GO prediction (DLMGO)



MVK-GO消融实验

PDN: PLMGO + DLMGO + Naïve

HDN: HCFGO + DLMGO + Naive

HPN: HCFGO + PLMGO + Naive

HPD: HCFGO + PLMGO + DLMGO

MKVGO: HCFGO + PLMGO + DMLGO + Naive

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Validation Dataset (974 proteins)	HCFGO	0.675	0.410	0.592	7.01	22.67	7.56	0.623	0.349	0.568
	PLMGO	0.678	0.415	**0.639**	7.23	22.80	7.18	0.632	0.347	0.590
	DLMGO	0.294	0.232	0.403	11.18	25.42	8.29	0.219	0.124	0.318
	Naïve	0.380	0.237	0.474	11.00	25.67	8.64	0.171	0.130	0.352
	PDN	0.678	0.414	0.631	7.23	22.40	7.17	0.669	0.355	0.647
	HDN	0.675	0.418	0.594	7.00	22.45	7.51	0.663	0.357	0.585
	HPN	0.696	0.434	0.637	6.85	22.04	7.17	0.699	0.380	0.618
	HPD	0.691	0.435	0.632	6.86	21.96	7.15	0.698	0.386	**0.654**
	MKVGO	**0.701**	**0.439**	0.634	**6.76**	**21.90**	**7.14**	**0.707**	**0.388**	0.650

MVK-GO消融实验

PDN: PLMGO + DLMGO + Naïve

HDN: HCFGO + DLMGO + Naive

HPN: HCFGO + PLMGO + Naive

HPD: HCFGO + PLMGO + DLMGO

MKVGO: HCFGO + PLMGO + DMLGO + Naive

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Test Dataset (1522 proteins)	HCFGO	0.682	0.412	0.580	7.23	23.91	8.14	0.630	0.340	0.539
	PLMGO	0.680	0.424	0.623	7.58	23.95	7.64	0.621	0.355	0.581
	DLMGO	0.319	0.252	0.450	11.77	26.97	8.99	0.219	0.142	0.390
	Naïve	0.367	0.234	0.470	11.79	27.16	9.04	0.174	0.129	0.342
	PDN	0.682	0.433	0.627	7.53	23.74	7.63	0.664	0.358	0.638
	HDN	0.684	0.422	0.582	7.19	23.78	8.12	0.679	0.347	0.577
	HPN	0.704	0.439	0.624	**6.99**	23.36	7.63	0.706	0.374	0.607
	HPD	0.700	0.441	0.632	7.03	23.28	7.63	0.708	**0.382**	0.641
	MKVGO	**0.706**	**0.446**	**0.630**	7.00	**23.25**	**7.62**	**0.710**	0.381	**0.641**

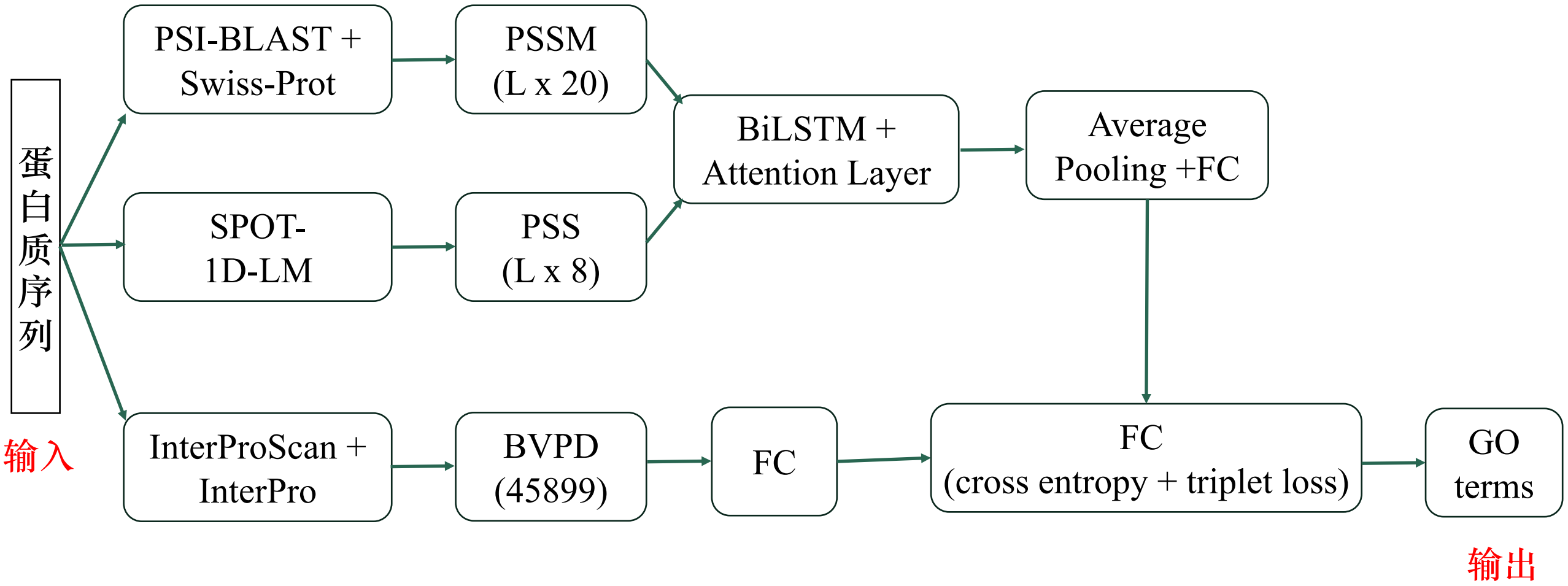
为什么选择ProtTrans 而不选择ESM系列?

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Validation Dataset (974 proteins)	MVK-GO (ProtTrans)	0.701	0.439	**0.634**	**6.76**	**21.90**	**7.14**	0.707	0.388	**0.650**
	MVK-GO (ESM2 + FC)	0.697	**0.441**	0.622	6.81	21.72	7.23	**0.714**	**0.390**	0.645
	MVK-GO (ATGO)	**0.703**	0.432	0.611	6.63	21.96	7.38	0.697	0.386	0.649
Test Dataset (1522 proteins)	MVK-GO (ProtTrans)	**0.706**	**0.446**	**0.630**	7.00	23.53	**7.62**	**0.710**	**0.381**	**0.641**
	MVK-GO (ESM2 + FC)	0.705	0.440	0.616	**6.97**	23.30	7.67	0.708	0.378	0.634
	MVK-GO (ATGO)	0.699	0.435	0.609	7.05	**23.39**	7.85	0.706	0.375	0.638

为什么不采用复杂的神经网络处理蛋白质语言模型的特征表示?

Dataset	Method	F_{max}			S_{min}			AUPR		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
Validation Dataset	ESM2 + FC	**0.679**			**7.09**			**0.627**		
(974 proteins)	ESM2 + BiLSTM	0.649			7.53			0.591		
Test Dataset	ESM2 + FC	**0.687**			**7.33**			0.616		
(1522 proteins)	ESM2 + BiLSTM	0.664			7.68			**0.617**		

为什么采用决策融合而不采用特征融合?



我们可以得到什么结论?

- 针对传统的手工特征表示方法，只有设计复杂的深度神经网络，仍然可以取得较好的性能。

- 基于手工特征表示的预测方法（HCFGO）能够与基于蛋白质/DNA大语言模型的预测方法（PLMGO和DLMGO）相互补充，进一步提升预测性能。

下一步的工作计划？

- 采用更好预测性能的PSSM
- 在CAFA5数据集上测试性能
- 在HCFGO中考虑融入结构数据