Deep Learning in Protein Sequence Analysis

Gregory Butler

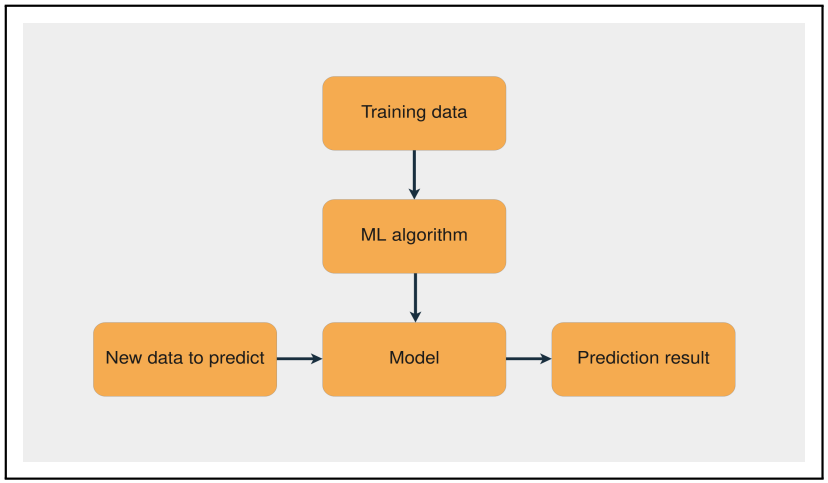
Department of Computer Science & Software Engineering
Concordia University, Montréal, Canada

23 November 2023 — University of Waikato
01 December 2023 — University of New South Wales

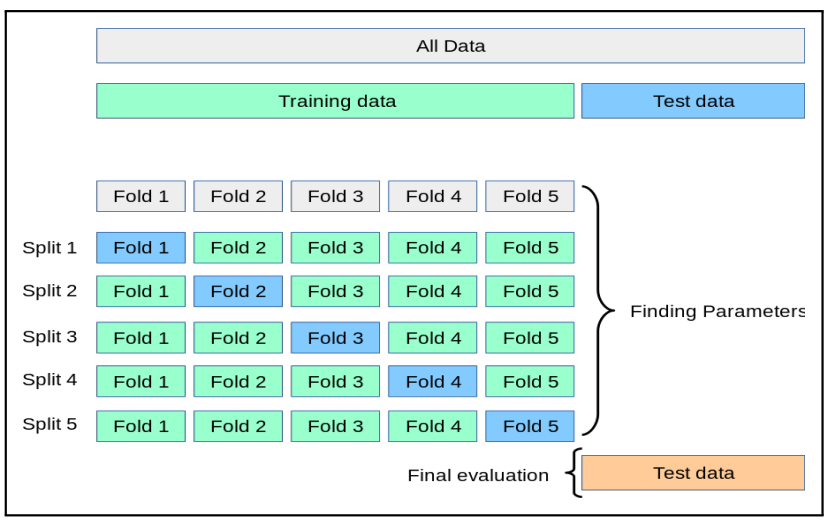
Outline

Supervised Machine Learning — Traditional

Workflow



Cross-Validation



Supervised Learning Dataset

dataset of **known** datapoints with
features and *labels* for
training + validation + independent test

Training

Optimisation of *parameters* for Model
Minimise, across whole training set,
loss between actual label & predicted label

Validation

Use *validation set* to compare Models
Select
best choice of ML algorithm and/or
best *hyper-parameters* for ML algorithm
Cross-Validation
Provides $mean \pm sd$ for selection
Hence, significance of Model differences

Independent Test

To gauge final Model on "new" data
independent of training and validation data

External Validation

Evaluate on real-world data
eg compare result of human experts

ML Concerns

Criteria for Choice of Model

- Whether the model meets the **goal**
- How much **pre-processing** the model needs
  including the time required to *train* the model
- How **accurate** is the model
  in general, how well it performs during evaluation
- How **explainable** is the model
  explainable method; explainable prediction
- How **fast** is the model in making predictions
- How **scalable** is the model

ML Concerns

Data

- Amount
- Quality: *noise* in features and labels
- Imbalance
- Bias

Feature Engineering

Independence of Test Set

data leakage due to unseen *confounding factors*
leak = *evolution* of proteins: set *percent identity (pid)* threshold

pid 60, 40, 20 commonly used

Overfitting

Which Performance Metric

Neural Networks — Traditional Feed Forward

Architecture

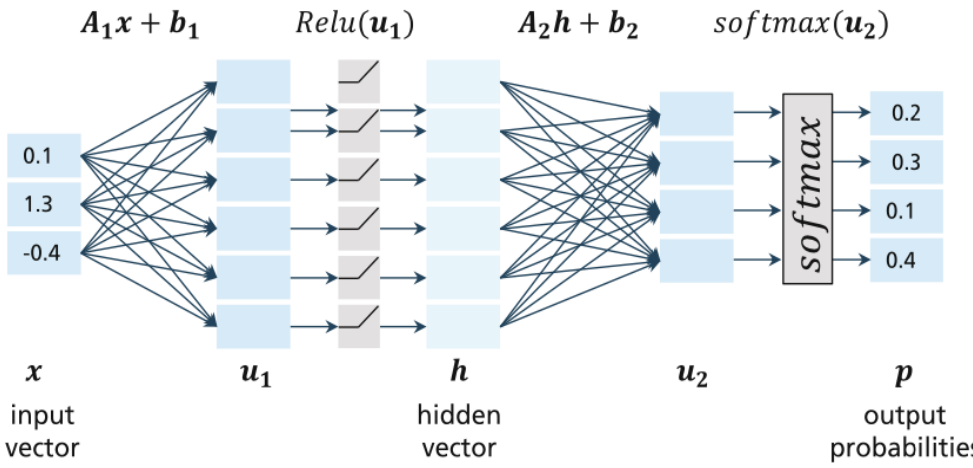


Fig. 1.1 A neural network for classification transforms the input by layers with affine transformations and nonlinear activation functions, e.g. ReLU. The final layer usually is a logistic classifier

Terminology

Activation function combines previous layer and edge weights

Sigmoid $f(x) = 1/(1 + e^{-x})$, Tanh $f(x) = 2 \times sigmoid(2x) - 1$

Rectified Linear Unit (ReLU) $f(x) = 0$ if $x \leq 0$ else x

loss function

optimise by *gradient descent* and *back-propagation*

Deep Neural Networks

Deep — More than Three Hidden Layers

Types

FFNN — Feed Forward NN aka multilayer perceptron (MLP)

CNN — Convolution NN for computer vision

RNN — Recurrent NN for sequential & time-series data

LSTM — Long Short-Term Memory RNN capturing long-term dependencies

Concern: Computation Resources

Require new optimisation algorithms; GPU implementations

Theano: Yoshua Bengio Python library 2010;

released v1.0.0 2017/11/15; now **PyTorch**

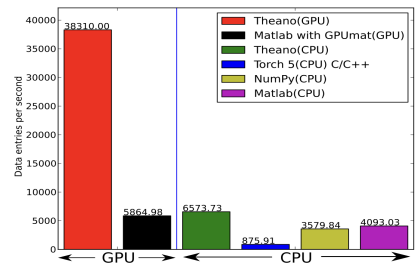
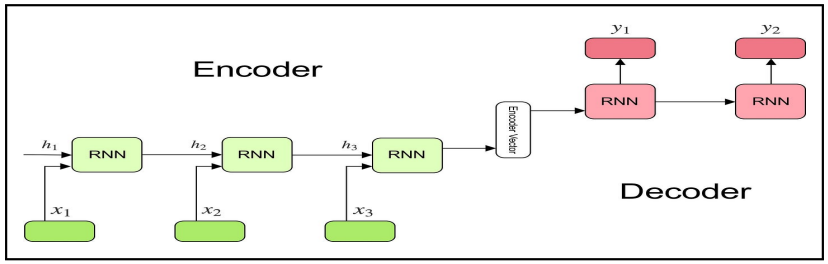


Figure 5: *Fitting a multi-layer perceptron to simulated data with various implementations of stochastic gradient descent. These models have 784 inputs, 500 hidden units, a 10-way classification, and are trained 60 examples at a time.*

J. Bergstra et al. *Theano: A CPU and GPU Math Expression Compiler*. Proc. of the Python for Scientific Computing Conference (SciPy) 2010.

Deep Neural Networks — Transformers

Encoder/Decoder



Attention Mechanisms to Capture Context

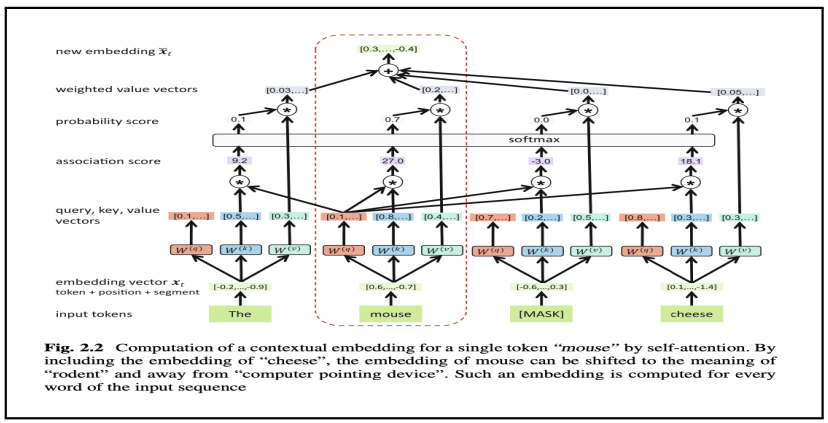
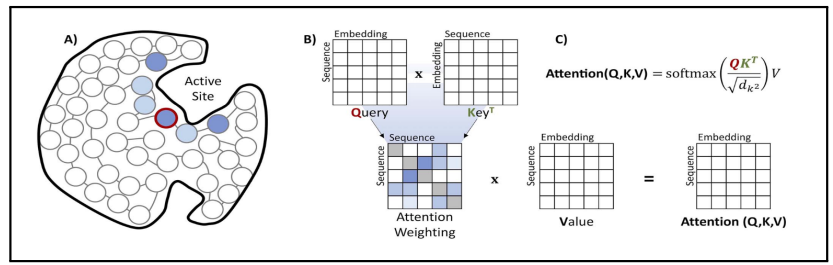


Fig. 2.2 Computation of a contextual embedding for a single token *"mouse"* by self-attention. By including the embedding of "cheese", the embedding of mouse can be shifted to the meaning of "rodent" and away from "computer pointing device". Such an embedding is computed for every word of the input sequence



BERT Encoder: attention, but no RNN

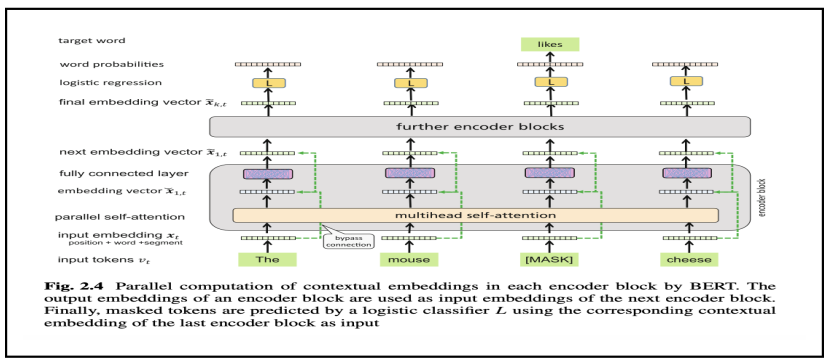


Fig. 2.4 Parallel computation of contextual embeddings in each encoder block by BERT. The output embeddings of an encoder block are used as input embeddings of the next encoder block. Finally, masked tokens are predicted by a logistic classifier L using the corresponding contextual embedding of the last encoder block as input

Deep Neural Networks — Transformers

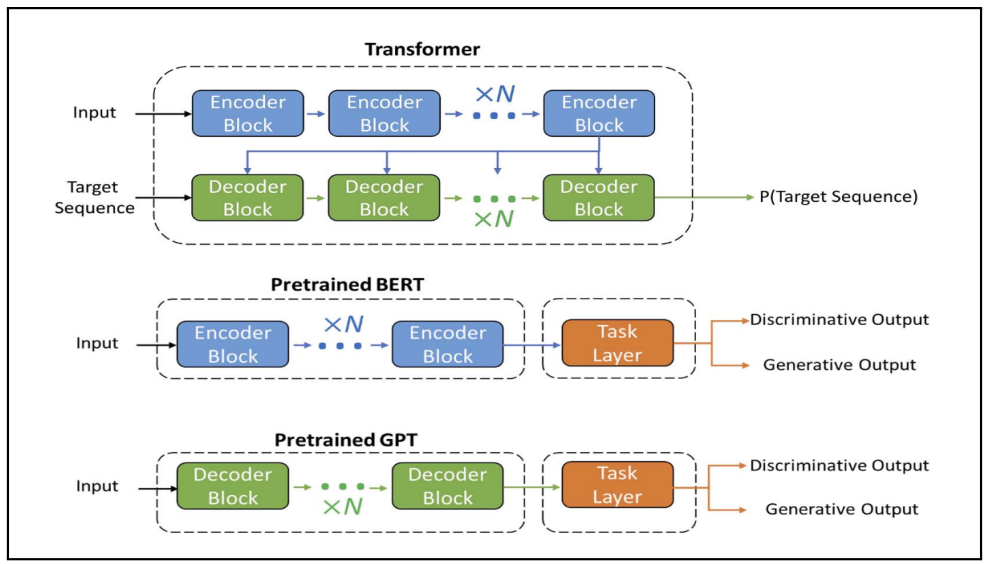
BERT

Bidirectional Encoder Representation from Transformers

GPT

Generative Pre-Training

Transformers

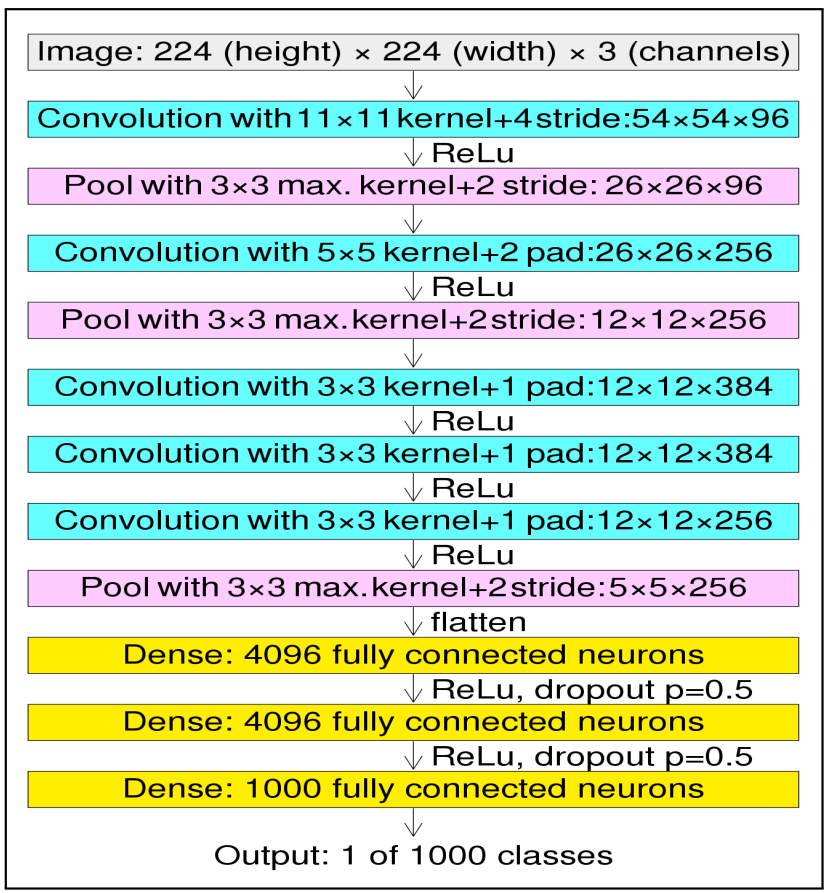


Deep Neural Networks

Breakthrough Moment for DL — 2012

`AlexNet` wins *ImageNet* 2012 Challenge on 2012-09-30
Achieved ***top-5 error*** of 15.3% vs runner-up 26.2%

AlexNet



ImageNet

14M+ images hand-annotated
20K+ categories of objects in images

ImageNet Challenge

2010+ ImageNet Large Scale Visual Recognition Challenge
Dataset is ImageNet subset
1000 non-overlapping categories
1000 approx. images per category
1.2M training images
50K validation images
150K test images

AlexNet

Convolution Neural Network (CNN)
ReLU activation function
Multiple GPUs for training

Alex Krizhevsky, Ilya Sutskever, G.E. Hinton (2017-05-24). *ImageNet classification with deep convolutional neural networks*. Communications of the ACM **60 (6)**: 84–90. `doi:10.1145/3065386`

AlphaFold Breakthrough for Protein Structure Prediction

AlphaFold

2018: `AlphaFold` top-ranked in CASP13 (Critical Assessment of Techniques for Protein Structure Prediction)
CNN, supervised learning with 29K proteins+structures from PDB

2020: `AlphaFold2` top-ranked in CASP14
RMSD between the Cα atoms: 0.96Å vs 2.83Å for runner-up
Transformer, triangle attention mechanism, MSA information
Training 7d on 128 TPU v3 cores; Fine-tuning 4d

See also: `RosettaFold`, `ESMFold`, `ColabFold`, `OpenFold`

Alpha Protein Structure Database (`alphafold.ebi.ac.uk`)

200M+ million entries; broad coverage of UniProt
Predicted Aligned Error (PAE) for each entry

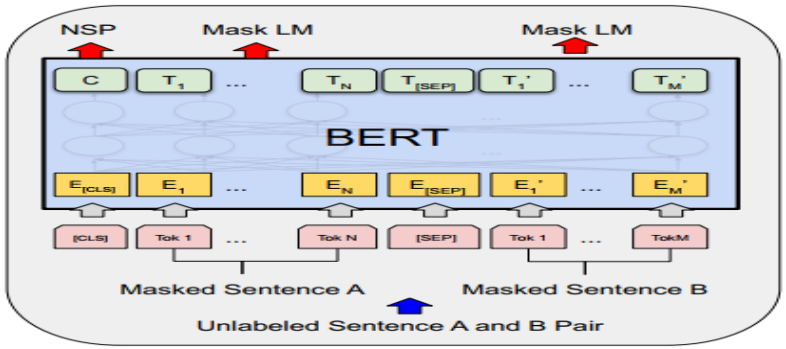
John Jumper et al. *Highly accurate protein structure prediction with AlphaFold*. Nature **596**, 583–589 (2021). `https://doi.org/10.1038/s41586-021-03819-2`

Large Language Models — NLP

Pre-training — Self-Supervised

Task independent; Large corpus of text; Large computation

MLM: Masked Language Modeling; NSP: Next Sentence Prediction



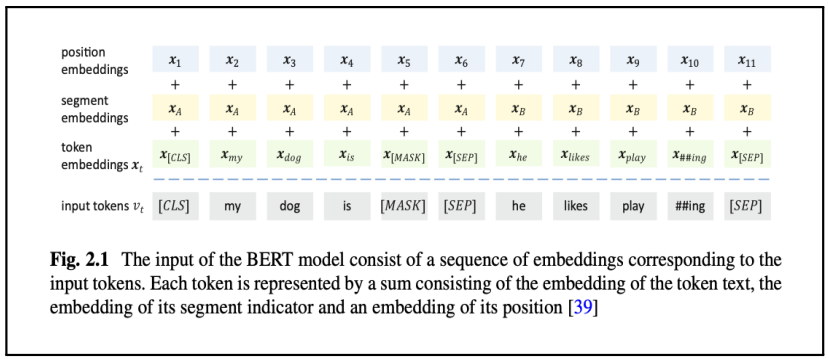
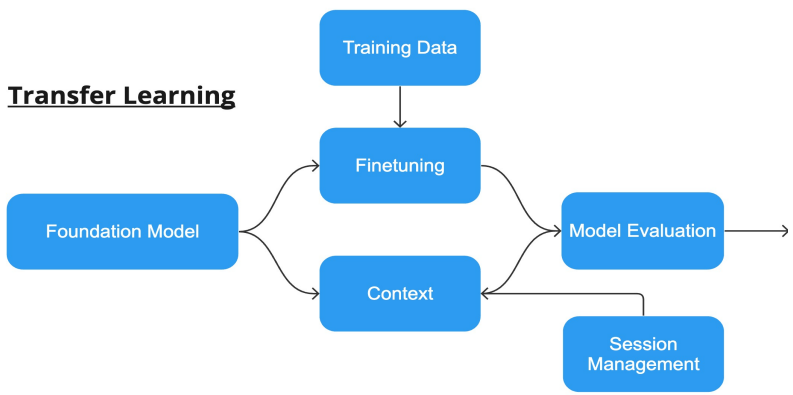


Fig. 2.1 The input of the BERT model consist of a sequence of embeddings corresponding to the input tokens. Each token is represented by a sum consisting of the embedding of the token text, the embedding of its segment indicator and an embedding of its position [39]

Fine-Tuning — Supervised

Downstream task-specific

Foundation Models & Transfer Learning



Protein Sequence Representation

Traditional

amino acid composition vector

k-mer and skipped *k*-mer

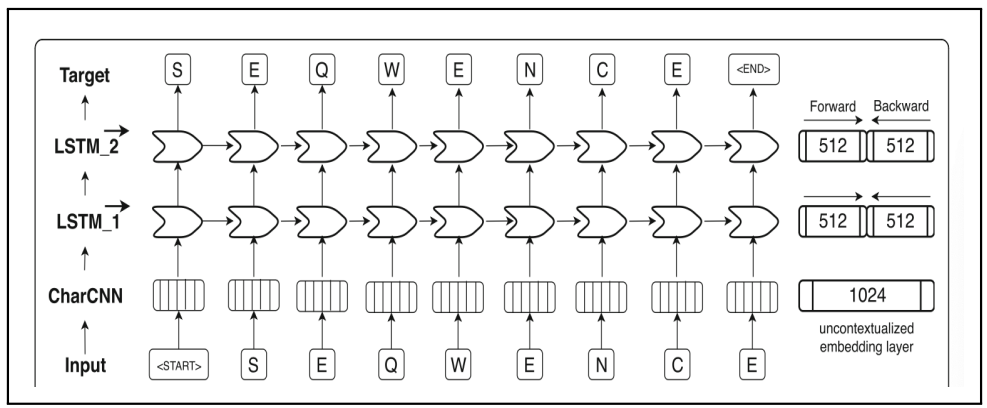
many physiochemical & surface accessibility "features"

Post NLP Deep Learning

sequence as text: each amino acid as "word"

(truncated) sequence as image: AA as 20-dim one-hot encoding

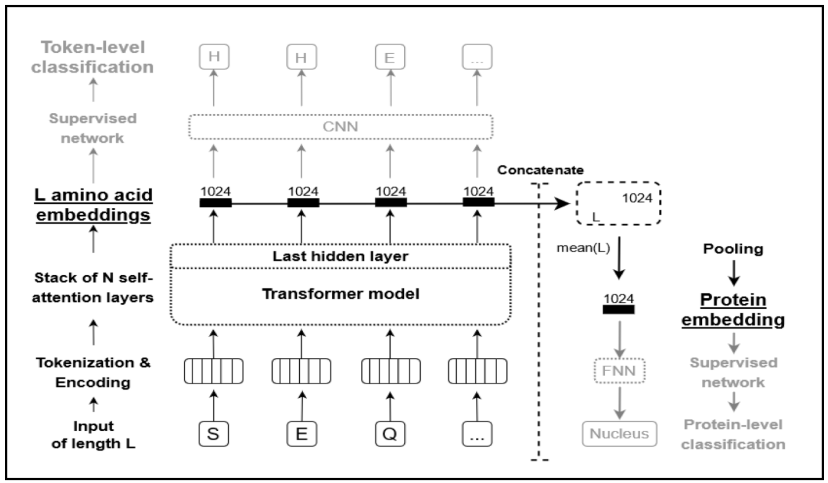
SeqVec (2019): RostLab based on ELMo (LSTM)



Z. Chen et al. *iFeature: a python package and web server for features extraction and selection from protein and peptide sequences*. Bioinformatics, 2018, **34** (14) pp. 2499–2502, doi:10.1093/bioinformatics/bty140. https://ifeature.erc.monash.edu

M. Heinzinger et al. *Modeling aspects of the language of life through transfer-learning protein sequences*. BMC Bioinformatics 20, 723 (2019). https://doi.org/10.1186/s12859-019-3220-8

Protein Language Model (PLM): MLM pre-training



Training Sets

UniRef100: 216M proteins, 80B AA

BFD: 2.1B proteins, 393B+ AA

Tasks

sec. structure; localization Q10, Q2

role of MSA (evolution info)

ProtT5 — *best performer* w/o MSA

Secondary Structure

Dataset	*CASP12*	*NEW364*
DeepProtVec	62.9	64.7
ProtTXL*	71.5	72.8
ProtTXL-BFD*	71.7	72.2
DeepSeqVec	73.0	76.0
ProtXLNet*	73.7	77.3
ProtElectra*	73,9	78,1
ProtAlbert*	74.6	78.5
ProtBert*	75.0	80.1
ProtBert-BFD*	75.8	81.1
ESM-1b	76.9	82.6
ProtT5-XXL-BFD*	77.7	81.6
ProtT5-XL-BFD*	77.5	82.0
ProtT5-XXL-U50*	79.2	83.3
ProtT5-XL-U50*	81.4	**84.8**
NetSurfP-2.0	**82.0**	84.3

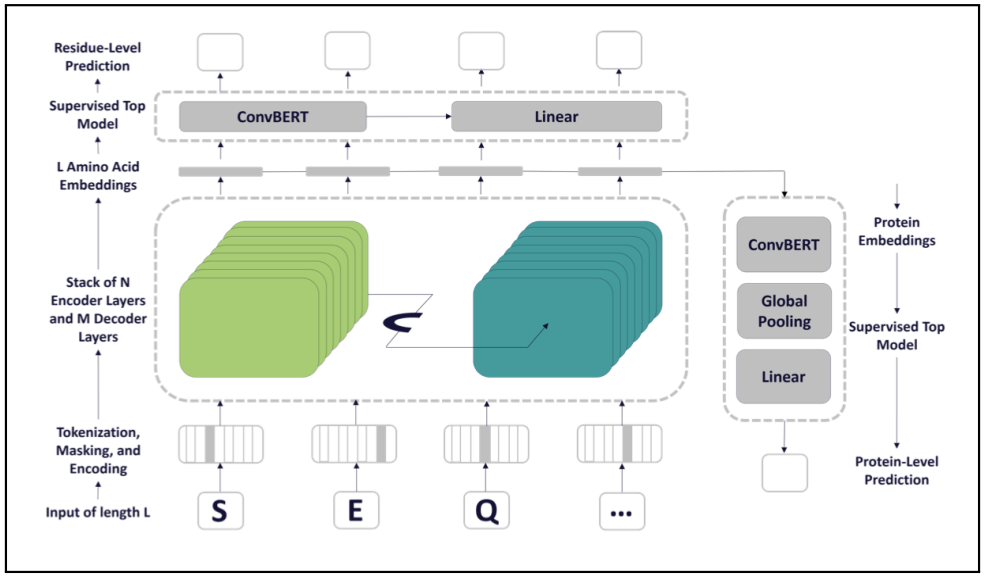
A. Elnaggar et al. *ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning*. IEEE Trans PAMI , vol. 44, no. 10, pp. 7112-7127, 1 Oct. 2022, doi:10.1109/TPAMI.2021.3095381

Open Source PLM

Hyperparameter	ProtTXL		ProtBert		ProtXLNet	ProtAlbert	ProtElectra	ProtT5-XL		ProtT5-XXL	
Dataset	BFD100	UniRef100	BFD100	UniRef100	UniRef100	UniRef100	UniRef100	UniRef50	BFD100	UniRef50	BFD100
Number of Layers	32	30	30		30	12	30	24		24	
Hidden Layers Size	1024		1024		1024	4096	1024	1024		1024	
Hidden Layers Intermediate Size	4096		4096		4096	16384	4096	16384		65536	
Number of Heads	14	16	16		16	64	16	32		128	
Positional Encoding Limits	-		40K		-	40K	40K	-		-	
Dropout	0.15		0.0		0.1	0.0	0.0	0.1		0.1	0.0
Target Length	512		512/2048		512	512/2048	512/1024	512		512	
Memory Length	512		-		384	-	-	-		-	
Masking Probability	-		15%		-	15%	25%	15%		15%	
Local Batch Size	8	5	32/6	30/5	2	21/2	18/7	8	4	8	4
Global Batch Size	44928	22464	32768/6144	15360/2560	1024	10752/1024	9216/3584	2048	4096	2048	4096
Optimizer	Lamb		Lamb		Adam	Lamb	Lamb	AdaFactor		AdaFactor	
Learning Rate	0.0005	0.002	0.002		0.00001	0.002	0.002	0.01		0.01	
Weight Decay	0.0	0.01	0.01		0.01	0.01	0.01	0.0		0.0	
Training Steps	40.7K	31.3K	800K/200K	300K/100K	847K	150K/150K	400K/400K	991K	1.2M	343K	920K
Warm-up Steps	13.6K	5.5K	140K/20K	40K/0K	20K	40K/5K	40K/40K	10K		10K	
Mixed Precision	FP16 Model Weight Fp32 Master Weight		None		None	None	None	None		None	
Number of Parameters	562M	409M	420M		409M	224M	420M	3B		11B	
System	Summit	Summit	TPU Pod		TPU Pod	TPU Pod	TPU Pod	TPU Pod		TPU Pod	
Number of Nodes	936		128	64	64	64	64	32	128	32	128
Number of GPUs/TPUs	5616		1024	512	512	512	512	256	1024	256	1024

A. Elnaggar et al. *ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning*. IEEE Trans PAMI , vol. 44, no. 10, pp. 7112-7127, 1 Oct. 2022, doi:10.1109/TPAMI.2021.3095381

Ankh — PLM of Choice



Task	Dataset	Ankh	Ankh_base	ProtT5-XL-U50	ESM-1b	ESM-2 (650M)	ESM-2 (3B)	ESM-2 (15B)
SSP	*CASP12* [41]	**83.8**±3%	80.8±4%	83.4±4%	79.6±4%	82.3±4%	83.3±4%	83.2±3%
	CASP14 [42]	**77.6**±3%	76.8±3%	74.1±3%	75.1±4%	77.0±3%	76.8±3%	76.8±4%
CP	*ProteinNet L/1* [34]	**49.0**±8%	43.2±8%	44.7±8%	30.0±6%	29.6±6%	30.7±6%	33.3±6%
	ProteinNet L/5	**73.2**±11%	66.6±11%	69.2±11%	50.1±10%	50.2±10%	52.7±10%	54.7±10%
	CASP14 L/1	**30.2**±8%	28.8±7%	26.9±7%	24.6±6%	25.0±6%	24.8±7%	25.9±7%
	CASP14 L/5	**50.7**±11%	48.0±11%	42.4±14%	40.0±11%	38.4±13%	41.9±14%	40.4±15%
EAT		71.7±6%	**74.8**±6%	71.0±6%	64.5±7%	55.5±7%	65.6±6%	65.4±7%
FolP		**61.1**±4%	58.8±4%	57.6±4%	57.6±4%	56.3±4%	60.5±4%	56.7±4%
FluP		**0.62**±0.004	0.61±0.004	0.58±0.004	0.5±0.005	0.48±0.005	0.48±0.005	0.55±0.004
SolP		**76.4**±2%	74.2±2%	74.4±2%	67.3±2%	75.0±2%	74.9±2%	60.4±2%
GB1P		0.84±0.008	**0.85**±0.008	0.78±0.01	0.81±0.009	0.82±0.009	0.81±0.009	0.57±0.02
LocP		**83.2**±2%	81.4±2%	**83.2**±2%	80.0±2%	81.8±2%	82.4±2%	81.8±2%

Summary of the benchmarking results involving, *Ankh* and *Ankh_base*, with ProtT5-XL-U50, ESM-1b, ESM-2 (650M), ESM-2 (3B), ESM-2 (15B): We report the Spearman Correlation score for the regression tasks and accuracy scores for all classification tasks, except for contact prediction where we report the L/1 and L/5 precision. In EAT, the score reported is the mean of the accuracy scores of the four annotations (Class, Architecture, Topology and Homologous super-family). Task Abbreviations: SSP: Secondary Structure Prediction; CP: Contact Prediction; EAT: Embedding-based Annotation Transfer; FolP: Fold Prediction; FluP: Fluorescence Prediction; SolP: Solubility Prediction; GB1P: GB1 Fitness Prediction ;LocP: Localization Prediction

Ahmed Elnaggar et al. *Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling*. biorxiv 2023 `https://doi.org/10.48550/arXiv.2301.06568`

Benchmark Tasks for PLMs

PEER: Protein sEquence undERstanding (MILA)

17 tasks; single task learning & multi-task learning

Task (Acronym)	Task Category	Data Source	#Protein	Seq. len.	#Train/Validation/Test	Metric
Function Prediction						
GB1 fitness prediction (GB1)	Protein-wise Reg.	FLIP [16]	8,733	$378.6_{(0.9)}$	381/43/8,309	Spearman's ρ
AAV fitness prediction (AAV)	Protein-wise Reg.	FLIP [16]	82,583	$1033.0_{(3.4)}$	28,626/3,181/50,776	Spearman's ρ
Thermostability prediction (Thermo)	Protein-wise Reg.	FLIP [16]	7,158	$880.6_{(974.2)}$	5,149/643/1,366	Spearman's ρ
Fluorescence prediction (Flu)	Protein-wise Reg.	Sarkisyan's dataset [71]	54,025	$343.3_{(1.3)}$	21,446/5,362/27,217	Spearman's ρ
Stability prediction (Sta)	Protein-wise Reg.	Rocklin's dataset [66]	68,934	$66.6_{(5.2)}$	53,571/2,512/12,851	Spearman's ρ
β-lactamase activity prediction (β-lac)	Protein-wise Reg.	Envision [25]	5,198	$396.1_{(0.7)}$	4,158/520/520	Spearman's ρ
Solubility prediction (Sol)	Protein-wise Cls.	DeepSol [39]	71,419	$424.1_{(225.9)}$	62,478/6,942/1,999	Acc
Localization Prediction						
Subcellular localization prediction (Sub)	Protein-wise Cls.	DeepLoc [2]	13,961	$665.3_{(395.3)}$	8,945/2,248/2,768	Acc
Binary localization prediction (Bin)	Protein-wise Cls.	DeepLoc [2]	8,634	$636.5_{(396.5)}$	5,161/1,727/1,746	Acc
Structure Prediction						
Contact prediction (Cont)	Residue-pair Cls.	ProteinNet [3]	25,563	$320.0_{(275.2)}$	25,299/224/40	L/5 precision
Fold classification (Fold)	Protein-wise Cls.	DeepSF [31]	13,766	$235.4_{(155.1)}$	12,312/736/718	Acc
Secondary structure prediction (SSP)	Residue-wise Cls.	NetSurfP-2.0 [41]	11,361	$360.5_{(229.3)}$	8,678/2,170/513	Acc
Protein-Protein Interaction Prediction						
Yeast PPI prediction (Yst)	Protein-pair Cls.	Guo's dataset [26]	1,707	$726.3_{(432.0)}$	1,668/131/373	Acc
Human PPI prediction (Hum)	Protein-pair Cls.	Pan's dataset [59]	5,553	$727.7_{(438.2)}$	6,844/277/227	Acc
PPI affinity prediction (Aff)	Protein-pair Reg.	SKEMPI [56]	627	$304.9_{(193.8)}$	2,127/212/343	RMSE
Protein-Ligand Interaction Prediction						
Affinity prediction on PDBbind (PDB)	Protein-ligand Reg.	PDBbind [49]	10,607	$414.9_{(234.3)}$	16,436/937/285	RMSE
Affinity prediction on BindingDB (BDB)	Protein-ligand Reg.	BindingDB [47]	1,006	$799.8_{(417.0)}$	7,900/878/5,230	RMSE

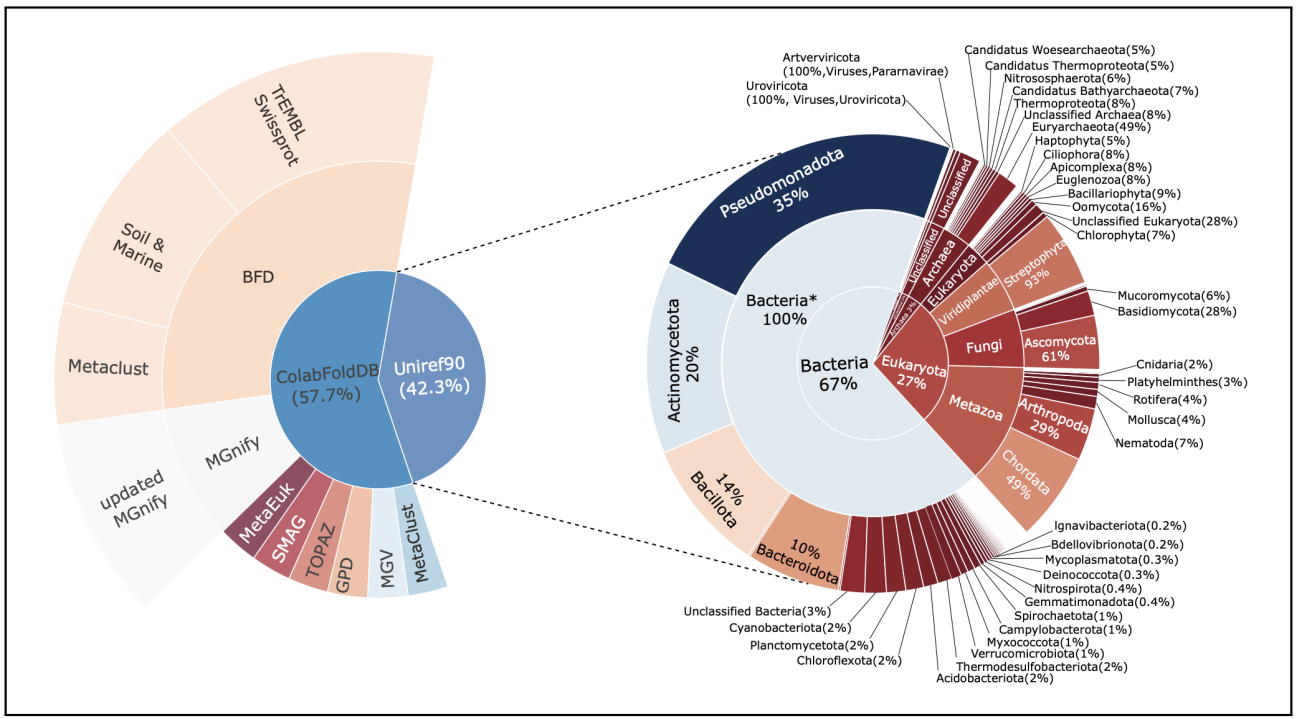
Minghao Xu et al (2022). *PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding*. Proc. NeurIPS 2022 Track on Datasets and Benchmarks.
Henriette Capel et al. *ProteinGLUE multi-task benchmark suite for self-supervised protein modeling*. Sci Rep 12, 16047 (2022). `https://doi.org/10.1038/s41598-022-19608-4`

xTrimoPGLM PLM (proprietary BioMap Research)

XTrimoPGLM

100B parameters; 1T training tokens; SOTA in 13/15 tasks

Training Set



Training: MLM+GLM

Bo Chen et al. *xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein*. biorxiv 2023 doi:https://doi.org/10.1101/2023.07.05.547496

DL in Protein Sequence Analysis — SOTA

DeepLoc

DeepLoc (2017): CNN+attention+RNN
DeepLoc 2.0 (2022): ESM-1b/ProtT5+attention
high-quality model ProtT5-XL-Uniref50 (3B parameters)
high-throughput model 33-layer ESM-1b (650M parameters)

DeepGO (2018) & DeepGOPlus (2020): CNN, GO aware

DeepEC (2019): CNN

DeepTMHMM (2022): ESM-1b+LSTM+CRF

SignalP 6.0 (2022): ProtBERT pre-trained on UniRef100

Vineet Thumuluri et al. *DeepLoc 2.0: multi-label subcellular localization prediction using protein language models*. Nucleic Acids Research, Vol 50 (W1), 5 July 2022, pp. W228–W234, https://doi.org/10.1093/nar/gkac278
Kulmanov M, Hoehndorf R. *DeepGOPlus: improved protein function prediction from sequence*. Bioinformatics. 2021 May 23;37(8):1187. doi:10.1093/bioinformatics/btaa763
Jae Yong Ryu et al. *Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers*. PNAS June 20, 2019 116 (28) 13996-14001 https://doi.org/10.1073/pnas.1821905116
Jeppe Hallgren et al. *DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks*. biorxiv 2022 doi:https://doi.org/10.1101/2022.04.08.487609
F. Teufel et al. *SignalP 6.0 predicts all five types of signal peptides using protein language models*. Nat Biotechnol 40, 1023–1025 (2022). https://doi.org/10.1038/s41587-021-01156-3

TooT-Suite Project for Protein Sequence Analysis

Proposal in 2017 before advent of PLMs

Aim — Initial

Predict and classify transmembrane transport proteins
Tasks: Discriminate membrane proteins, transport proteins
Tasks: Predict SC (substrate class) & SS (specific substrate)
Apply to proteomes and meta-proteomes

Aim — After PLM successes

Investigate DL for these tasks
Broaden benchmark tasks for PLMs
Are there protein-specific (not NLP) pre-training methods?

Conclusions

PLM Foundation Models superceding other DL methods
Ankh is open-source PLM of choice
No single one-size-fits-all for task-specific transfer learning

Trade-Offs

PLM size
  versus
cost of pre-training and fine-tuning
  and
classification throughput

Open Question

Is fine-tuning both task-specific component and PLM worth the computation cost?

DL in bioinformatics is only just beginning!

Thank You!

Questions, Please?

Example TooT-BERT-M — Membrane Proteins

Aim

Discriminate membrane proteins from non-membrane

Evaluate PLM ProtBERT-BFD for this task; Logistic Regression

Compare frozen vs fine-tuned approach

Dataset

TABLE I: DS-M: Membrane dataset

Class	Training	Test	Total
Membrane	7,945	883	8,828
Nonmembrane	8,157	907	9,064
Total	16,102	1,790	17,892

TABLE II: Functional types of membrane proteins

Type	Count	Percentage
Transporter	2224	25%
Receptor	1123	13%
Enzyme	2878	33%
Other	2603	29%

TABLE III: Structural types of the membrane proteins

Type	Count	Percentage
Single-pass	2684	36%
Multi-pass	2877	39%
Lipid-anchor	460	6%
GPI-anchor	218	3%
Peripheral	1175	16%

Hyperparameters

TABLE IV: Fine-tuning BERT hyperparameters

Hyperparamer	Value
training epochs	10
training batch size	1
evaluation batch size	32
warmup steps	1000
weight decay	0.01
gradient accumulation steps	64

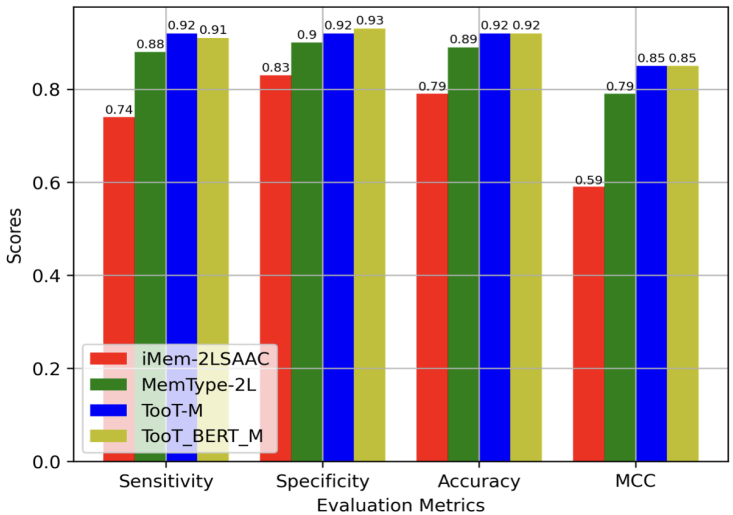
H. Ghazikhani & G. Butler. *TooT-BERT-M: Discriminating Membrane Proteins from Non-Membrane Proteins using a BERT Representation of Protein Primary Sequences*. CIBCB, 2022.
doi:10.1109/CIBCB55180.2022.9863026

Example TooT-BERT-M — Membrane Proteins

TooT-BERT-M is State-of-the-Art (SOTA)?

TABLE VII: Comparison with other methods

Method	Sen(%)	Spc(%)	Acc(%)	MCC
iMem-2LSAAC	74.52	83.90	79.27	0.59
MemType-2L	88.67	90.19	89.44	0.79
TooT-M	**92.41**	92.5	**92.46**	**0.85**
TooT-BERT-M	91.28	**93.61**	**92.46**	**0.85**



Frozen vs Fine-Tuned

TABLE V: BERT representation comparison

Rep	Sen(%)	Spc(%)	Acc(%)	MCC
Frozen	91.18	83.47	87.37	0.7492
Fine-tuned	91.28	93.61	92.46	0.8493

Example TooT-BERT-T-CNN — Transport Proteins

Aim

Discriminate transport proteins from non-transport
Evaluate PLM ProtBERT-BFD for this task
Compare traditonal ML with CNN

Aim — Fine-Tuning PLMs & Catastrophic Forgetting

"Catastrophic forgetting refers to the phenomenon where a model, when exposed to new data, tends to forget previously acquired knowledge"'
MembraneBERT adds knowledge of membrane vs non-membrane

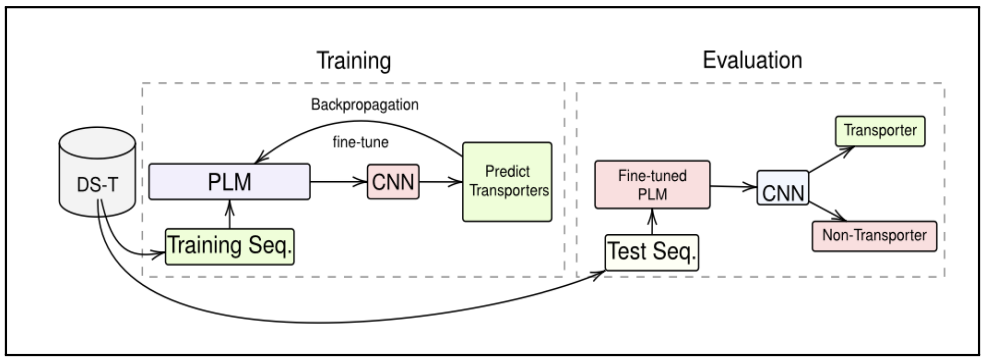
H. Ghazikhani & G. Butler. *Enhanced identification of membrane transport proteins: a hybrid approach combining ProtBERT-BFD and convolutional neural networks*. Journal of Integrative Bioinformatics 20 (2) 2023. https://doi.org/10.1515/jib-2022-0055

Example TooT-BERT-T-CNN

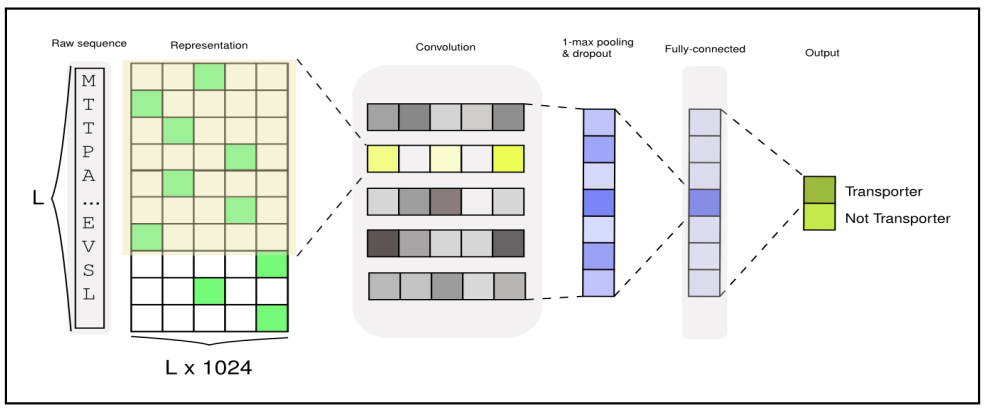
Dataset

Class	Training	Test	Total
Transporter	780	120	900
Non-transporter	600	60	660
Total	1380	180	1560

Workflow



CNN



Example TooT-BERT-T-CNN

TooT-BERT-T-CNN is SOTA

Classifier	Representer	Sen	Spc	Acc	MCC
TooT-T [6]	Traditional[a]	94.17	88.33	92.22	0.8200
TooT-BERT-T [7]	ProtBERT-BFD	**95.83**	90.00	93.89	0.8620
TooT-BERT-CNN-T	ProtBERT-BFD	95.00	**95.00**	**95.00**	**0.8894**

CNN outperforms traditional ML

Classifier	Representation	CV				Independent			
		Sen	Spc	Acc	MCC	Sen	Spc	Acc	MCC
kNN	ProtBERT-BFD	97.02 ± 2.79	97.10 ± 2.78	97.06 ± 2.65	0.9405 ± 0.0537	93.33	88.33	92.20	0.8250
	ProtBERT	91.21 ± 2.37	64.25 ± 2.79	79.49 ± 1.95	0.5857 ± 0.0422	95.00	60.00	83.89	0.6265
	MembraneBERT	98.00 ± 3.54	96.79 ± 5.08	97.47 ± 4.20	0.9485 ± 0.0857	85.83	88.33	86.67	0.7172
RF	ProtBERT-BFD	95.84 ± 3.13	97.11 ± 3.04	96.38 ± 3.08	0.9276 ± 0.0619	94.17	88.33	92.22	0.8250
	ProtBERT	88.40 ± 3.38	76.91 ± 4.40	83.31 ± 2.42	0.6635 ± 0.0493	89.17	78.33	83.89	0.6750
	MembraneBERT	97.82 ± 3.68	96.88 ± 5.10	97.43 ± 4.29	0.9473 ± 0.0877	85.00	90.00	86.67	0.7073
SVM	ProtBERT-BFD	94.05 ± 2.80	86.10 ± 2.68	90.59 ± 2.50	0.7999 ± 0.0506	**100.00**	90.00	92.78	0.8369
	ProtBERT	85.69 ± 2.69	53.97 ± 2.80	71.90 ± 1.64	0.4186 ± 0.0360	**100.00**	86.67	90.00	0.7771
	MembraneBERT	97.65 ± 3.64	96.68 ± 4.81	97.23 ± 4.13	0.9439 ± 0.0838	85.00	91.67	85.00	0.6930
LR	ProtBERT-BFD	96.79 ± 3.27	97.33 ± 2.91	97.03 ± 3.05	0.9400 ± 0.0617	95.83	90.00	93.89	0.8620
	ProtBERT	90.64 ± 2.42	82.33 ± 2.95	87.03 ± 2.02	0.7358 ± 0.0410	92.50	80.00	88.33	0.7347
	MembraneBERT	98.08 ± 3.53	97.00 ± 5.18	97.61 ± 4.25	0.9513 ± 0.0866	86.67	85.00	86.11	0.6989
FFNN	ProtBERT-BFD	92.13 ± 7.08	91.79 ± 6.98	91.79 ± 6.98	0.7924 ± 0.0586	92.50	90.00	90.00	0.8043
	ProtBERT	85.95 ± 6.79	78.44 ± 7.51	82.37 ± 2.29	0.6480 ± 0.0402	**100.00**	50.00	87.22	0.7414
	MembraneBERT	95.37 ± 5.49	94.60 ± 6.73	95.43 ± 4.74	0.9073 ± 0.0936	60.00	28.33	85.00	0.6832
CNN	ProtBERT-BFD	85.64 ± 7.25	95.33 ± 3.85	89.85 ± 3.57	0.8072 ± 0.0642	95.00	**95.00**	**95.00**	**0.8894**
	ProtBERT	95.00 ± 3.58	81.16 ± 1.47	88.98 ± 4.95	0.7855 ± 0.0943	95.00	90.00	93.33	0.8500
	MembraneBERT	98.71 ± 0.90	97.83 ± 1.25	98.33 ± 0.71	0.9662 ± 0.0157	90.83	91.66	91.11	0.8070

Evidence of Catastrophic Forgetting

See MembraneBERT in table above

Example TooT-BERT-ICAT

Aim

Predict specific substrates for inorganic ion transporters
Evaluate PLM ProtBERT-BFD for this task
Compare Logistic Regression with FFNN
Compare frozen vs fine-tuned approach
Does transfer learning handle the small dataset?

TABLE I: Four extracted datasets

Dataset	Size	Trainset	Testset	# Classes
UniProt-ICAT-100	4,112	3,289	823	12
UniProt-ICAT-60	1,429	1,143	286	11
SwissProt-ICAT-100	2,140	1,712	428	11
SwissProt-ICAT-60	1,098	878	220	11

The number of sequences and classes in each of the four datasets. Each dataset has been divided into a training set and test set with 80%-20% ratio randomly and stratified.

TABLE II: Size of substrate classes for each dataset

Class	CHEBI	Substrate	UP-ICAT-100	UP-ICAT-60	SP-ICAT-10	SP-ICAT-60
0	CHEBI:24636	proton	1307	600	883	484
1	CHEBI:29108	calcium(2+)	868	210	350	154
2	CHEBI:29103	potassium(1+)	726	193	299	146
3	CHEBI:17996	chloride	458	129	177	75
4	CHEBI:29101	sodium(1+)	429	123	235	99
5	CHEBI:16189	sulfate	98	49	38	27
6	CHEBI:29105	zinc(2+)	80	51	67	47
7	CHEBI:28938	ammonium	55	28	36	23
8	CHEBI:17632	nitrate	28	18	27	18
9	CHEBI:29033	iron(2+)	24	16	16	13
10	CHEBI:35780	phosphate ion	22	12	12	12
11	CHEBI:49552	copper(1+)	17	*9*	*7*	*6*
Total			4,112	1,429	2,140	1,098

S. Ataei & G. Butler. *Predicting the specific substrate for transmembrane transport proteins using BERT language model*. CIBCB, 2022. doi:10.1109/CIBCB55180.2022.9863051

Example TooT-BERT-ICAT

Results

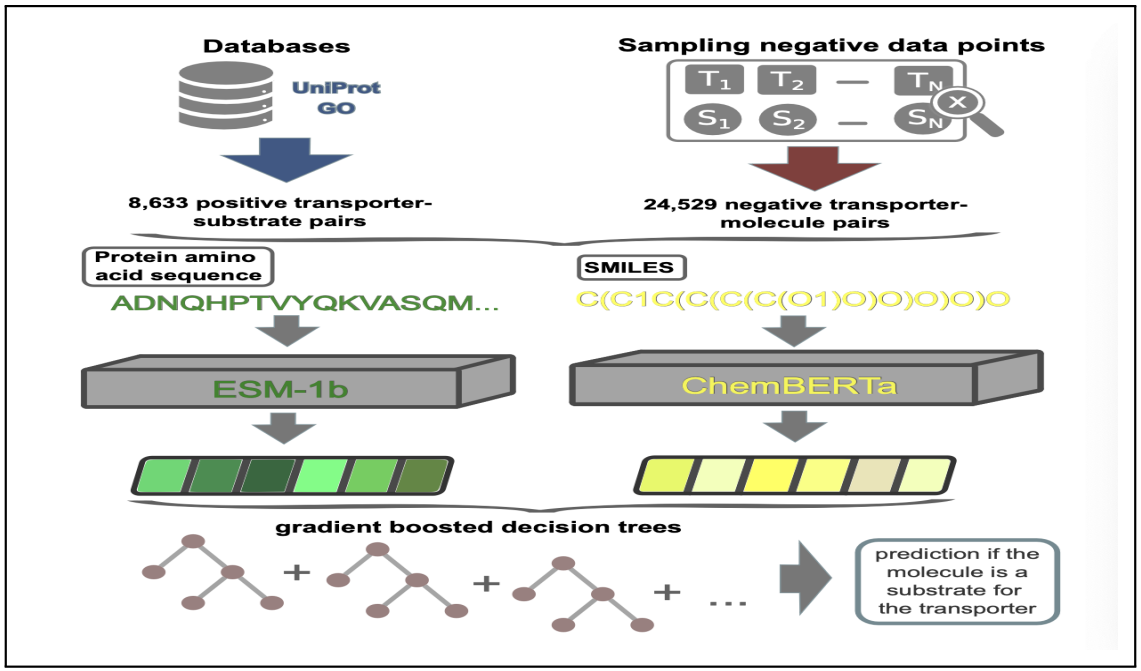
TABLE IV: Independent testset results comparison for Logistic Regression(LR), Feed-forward Neural Networks (FNN), and Fine-tuned BERT (FTB)

Dataset	# Classes	Accuracy			Precision			Recall			F1-score			MCC		
		LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB
UniProt-ICAT-100	12	0.975	0.985	**0.993**	0.852	0.911	**0.959**	0.630	0.832	**0.881**	0.680	0.858	**0.913**	0.813	0.889	**0.948**
UniProt-ICAT-60	11	0.952	0.963	0.982	0.734	0.794	0.899	0.415	0.597	0.868	0.442	0.664	0.903	0.640	0.726	0.867
SwissProt-ICAT-100	11	0.971	0.981	0.991	0.839	0.895	0.951	0.509	0.748	0.879	0.548	0.783	0.907	0.785	0.861	0.936
SwissProt-ICAT-60	11	0.950	0.956	0.979	0.723	0.759	0.886	0.425	0.595	0.757	0.465	0.598	0.802	0.616	0.682	0.850

TABLE V: Detailed results for classification of UniProt-ICAT-100 using Fine-tuned ProtBERT

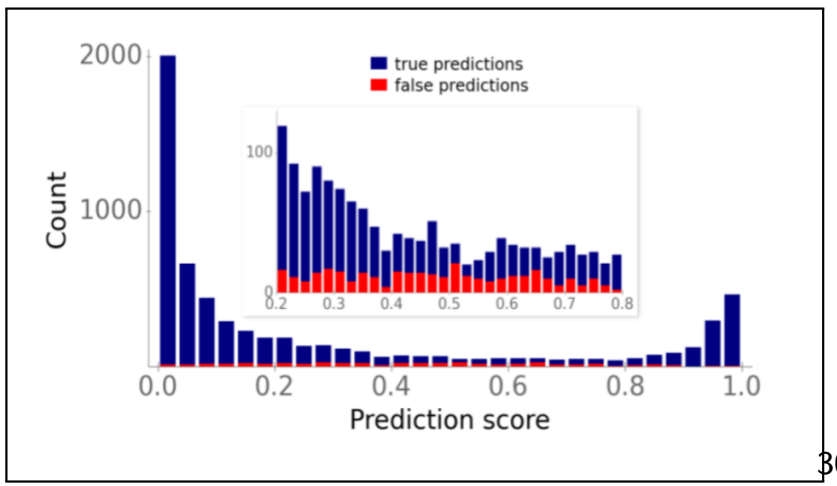
Substrate	Trainset	Validation	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	209	261	245	17	16	545	0.960	0.935	0.939	0.937	0.908
calcium(2+)	555	139	174	157	12	17	637	0.965	0.929	0.902	0.916	0.893
potassium(1+)	465	116	145	136	15	9	663	0.971	0.901	0.938	0.919	0.901
chloride	293	73	92	86	9	6	722	0.982	0.905	0.935	0.920	0.910
sodium(1+)	274	69	86	75	14	11	723	0.970	0.843	0.872	0.857	0.840
sulfate	62	16	20	14	4	6	799	0.988	0.778	0.700	0.737	0.732
zinc(2+)	51	13	16	14	0	2	807	0.998	**1.000**	0.875	0.933	0.934
ammonium	35	9	11	11	1	0	811	0.999	0.917	**1.000**	0.957	0.957
nitrate	18	4	6	4	0	2	817	0.998	**1.000**	0.667	0.800	0.816
iron(2+)	15	4	5	2	0	3	818	0.996	**1.000**	0.400	0.571	0.631
phosphate ion	15	3	4	3	1	1	818	0.998	0.750	0.750	0.750	0.749
copper(1+)	11	3	3	3	0	0	820	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

SPOT using (seq, substrate) pairs

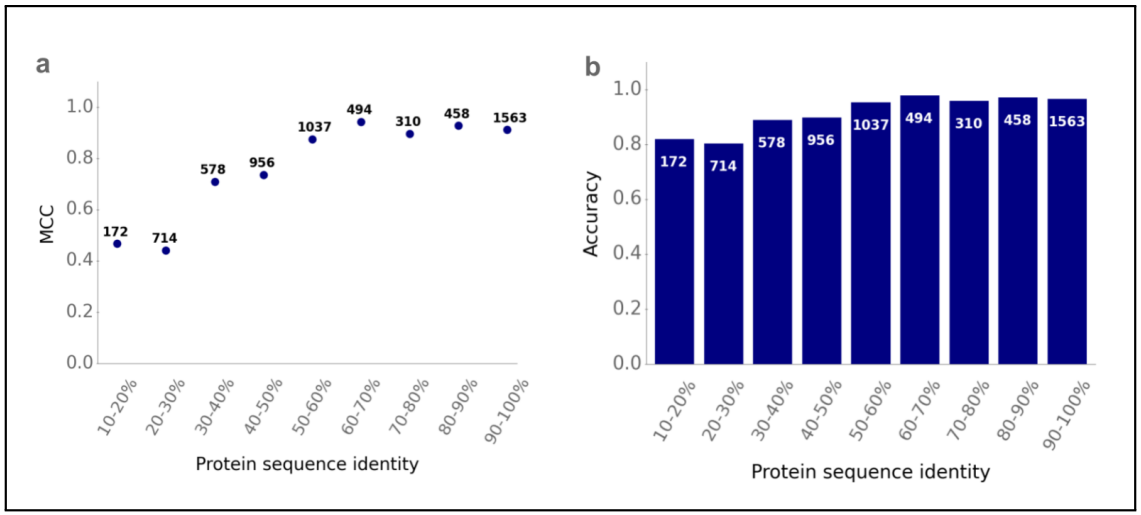


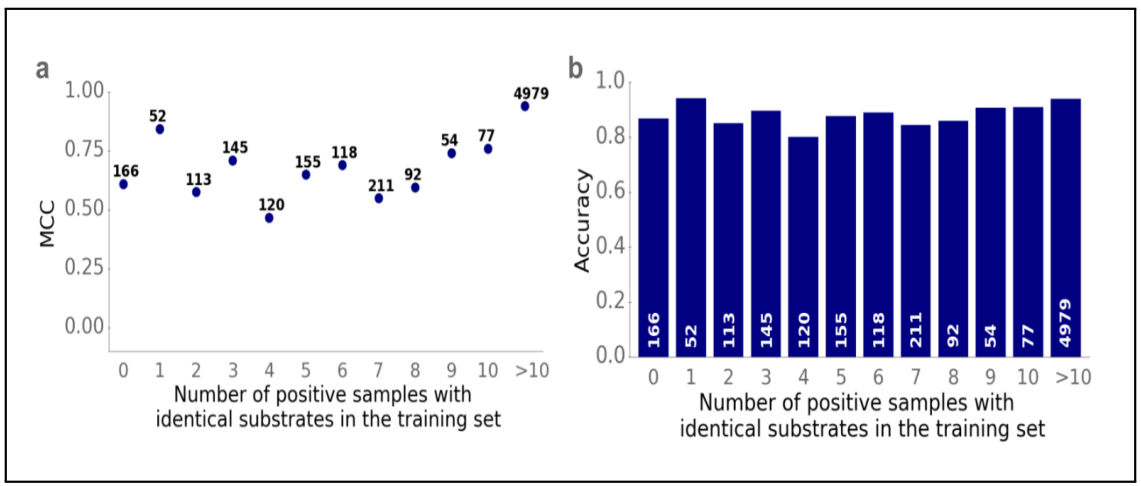
SPOT Discriminating Transporters from Non-Transporters

	Accuracy	ROC-AUC	MCC
ESM-1b + ECFP	91.5%	0.956	0.78
ESM-1b$_{ts}$ + ECFP	90.0%	0.955	0.75
ESM-1b + ChemBERTa	92.4%	0.961	0.80
ESM-1b$_{ts}$ + ChemBERTa	90.6%	0.957	0.76



SPOT — Independence of Test Set?





Alexander Kroll et al (2023). *A general substrate prediction model for transport proteins using machine and deep learning*. biorxiv doi.org/10.1101/2023.10.31.564943

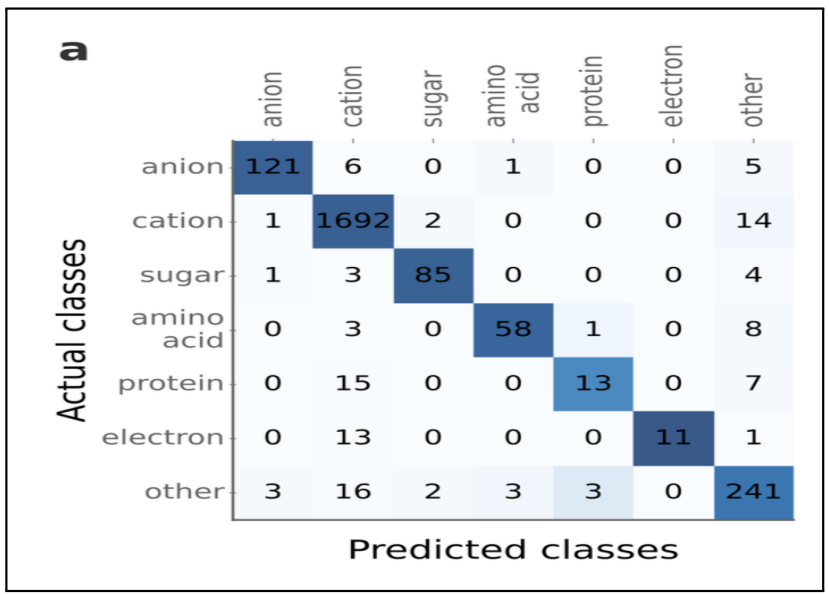
SPOT Predicting Seven Substrate Classes

Classes	No. of training data points	No. of test data points	No. of different ChEBI IDs
cation	6914	1709	87
anion	625	133	59
sugar	350	93	92
amino acid / oligopeptide	282	70	163
protein / mRNA	105	35	10
electron	77	25	1
other	978	268	945

Dataset

Class	Accuracy	MCC
anion	99.27%	0.93
cation	96.87%	0.92
sugar	99.49%	0.93
amino acid / oligopeptide	99.31%	0.88
protein / mRNA	98.89%	0.53
electron	99.4%	0.66
other	97.17%	0.86

Performance



Confusion Matrix

Conclusions

PLM Foundation Models superceding other DL methods
Ankh is open-source PLM of choice
No single one-size-fits-all for task-specific transfer learning

Trade-Offs

PLM size
  versus
cost of pre-training and fine-tuning
  and
classification throughput

Open Question

Is fine-tuning both task-specific component and PLM worth the computation cost?

DL in bioinformatics is only just beginning!

Thank You!

Questions, Please?

Generative AI and Biology

Protein Design

Zhao, J.; Yan, W.; Yang, Y. *DeepTP: A Deep Learning Model for Thermophilic Protein Prediction*. Int. J. Mol. Sci. 2023, 24, 2217. https://doi.org/10.3390/ijms24032217

Madani, A., Krause, B., Greene, E.R. et al. *Large language models generate functional protein sequences across diverse families*. Nat Biotechnol (2023). https://doi.org/10.1038/s41587-022-01618-2

Kroll A, Engqvist MKM, Heckmann D, Lercher MJ (2021) *Deep learning allows genome-scale prediction of Michaelis constants from structural features*. PLoS Biol 19(10): e3001402. https://doi.org/10.1371/journal.pbio.3001402

Kroll, A., Ranjan, S., Engqvist, M.K.M. et al. *A general model to predict small molecule substrates of enzymes based on machine and deep learning*. Nat Commun 14, 2787 (2023). https://doi.org/10.1038/s41467-023-38347-2

Mehrsa Mardikoraem, Zirui Wang, Nathaniel Pascual and Daniel Woldring. *Generative models for protein sequence modeling: recent advances and future directions*. Briefings in Bioinformatics, 2023, 24(6), 1–19 https://doi.org/10.1093/bib/bbad358