MUNI
FI

# UDSMProt: universal deep sequence models for protein classification

Tomáš Pavlík

# In this presentation…

- Motivation and background
- Used algorithms
- Results
- Discussion

# Our goal

| | | |
|---|---|---|
| Cellular Component | trans-Golgi network | Source:UniProtKB |
| Molecular Function | amyloid-beta binding | Source:UniProtKB |
| Biological Process | positive regulation of amyloid-beta formation | Source:UniProtKB |
| Biological Process | regulation of proteolysis | Source:UniProtKB |

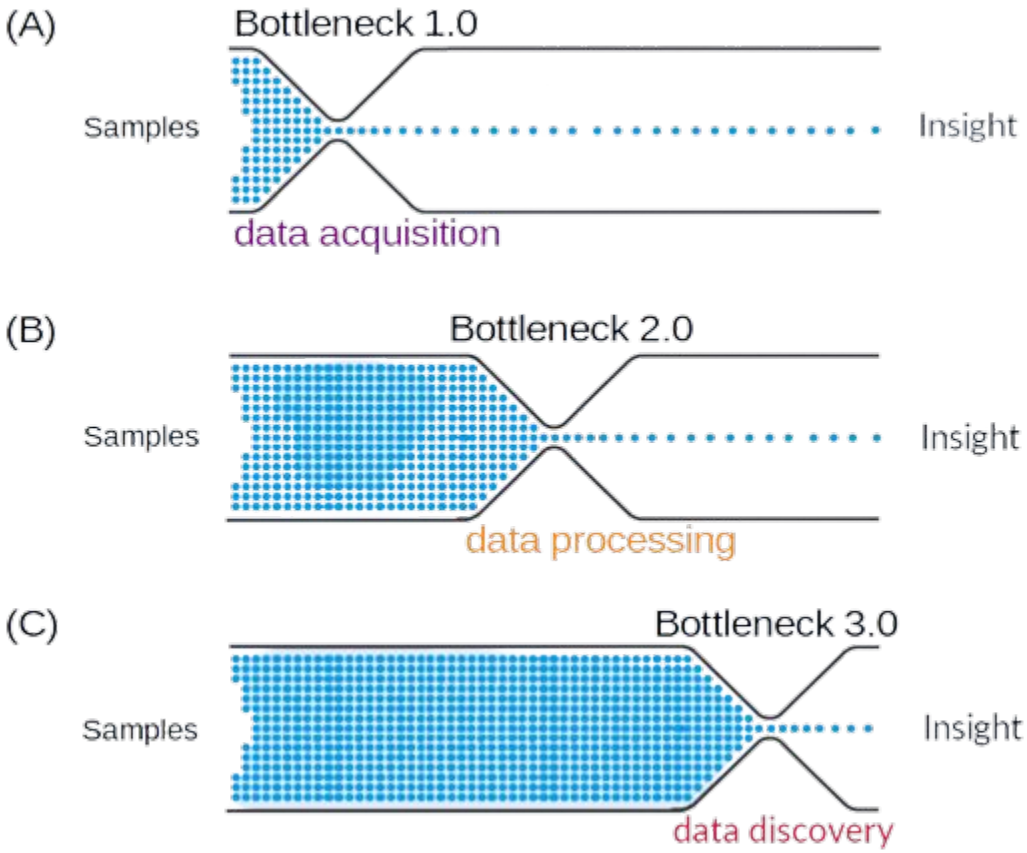
**SQ SEQUENCE** 854 AA; 97802 MW; D7B4A3A95E2E8C3B CRC64;

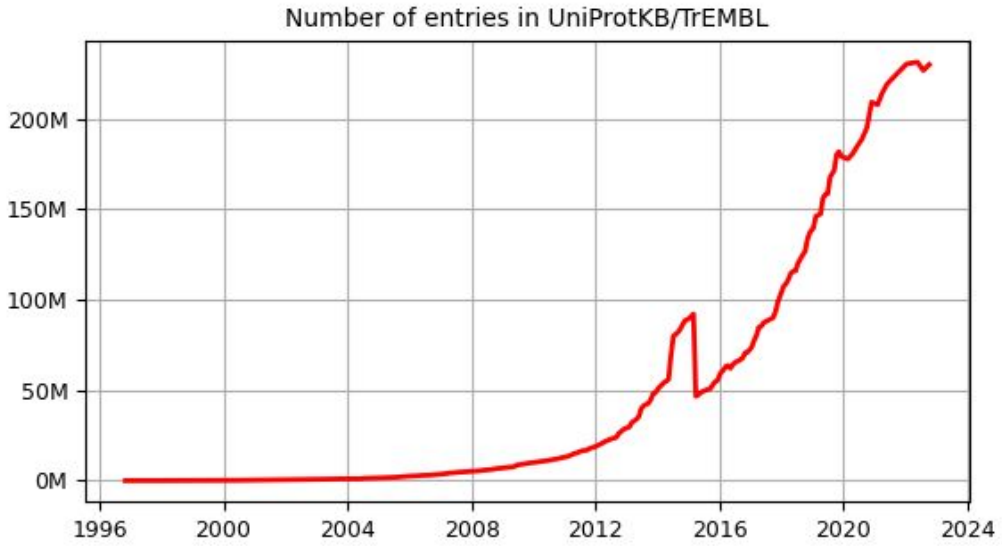
| | | | |
|---|---|---|---|
| CLDCEKMAAL | HCALYCGQGA | QFLEAQIIQW | ISENVSACHS |
| SNMDKLLPHS | SVLTWNTEIP | GITLVTEDIA | LPLMKVLSFK |
| HYNNSVVRRE | WHNLISEEKT | GKRRSAAYVR | NILDNAVKVI |
| SHQRLLMGLM | VSELKDHFLR | HLQGVEKKKI | EQMVLDYISK |
| WVLHFNSRGS | AAEFAVFHIM | TRILEATNSL | FLPLPPGFHT |

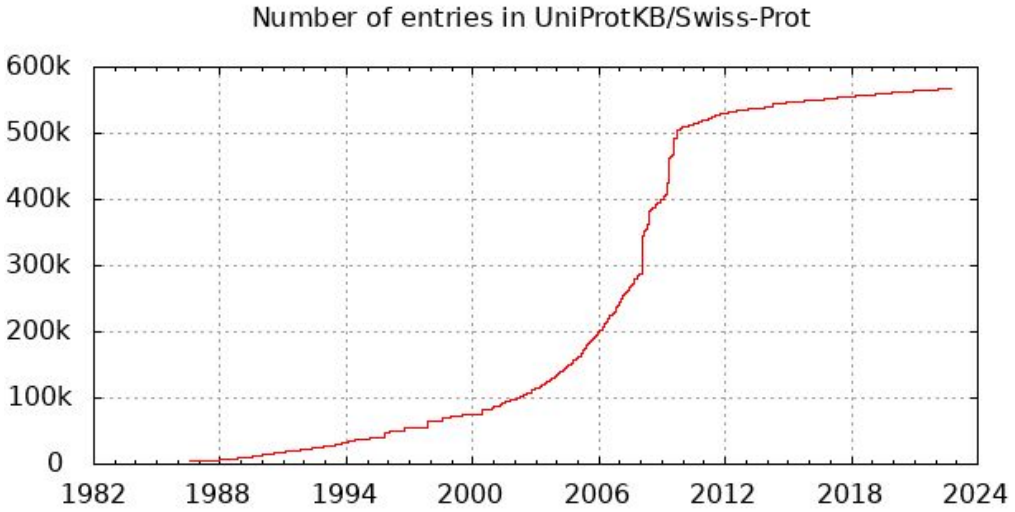
post-translational modifications, interactions, expression…

https://www.uniprot.org/uniprotkb/A4D1B5/entry

# The bottleneck



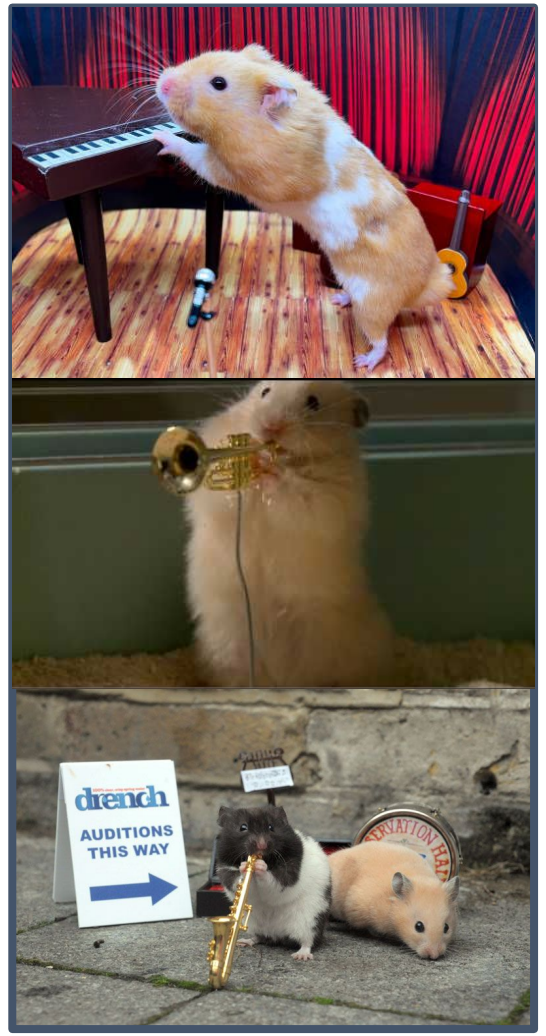




https://bigomics.ch/blog/the-current-bottleneck-in-omics-data-analysis/
https://www.ebi.ac.uk/uniprot/TrEMBLstats
https://web.expasy.org/docs/relnotes/relstat.html

# The way we used to analyse

| | |
|---|---|
| Primary sequence similarity | O(n) |
| Hand-crafted features | inaccurate |
| Experiments | manpower |
| … | |

# Out-of-the-box idea?


Google stock images

# The language of genome

Dvanásti sokoli, sokolovia Tatier,
akoby ich bola mala jedna mater;
jedna mater mala, v mlieku kúpavala,
zlatým povojníčkom bola povíjala.

To sa chlapci, to sa jak oltárne sviece,
keď idú po háji, celý sa trbliece.

**SQ SEQUENCE** 854 AA; 97802 MW;

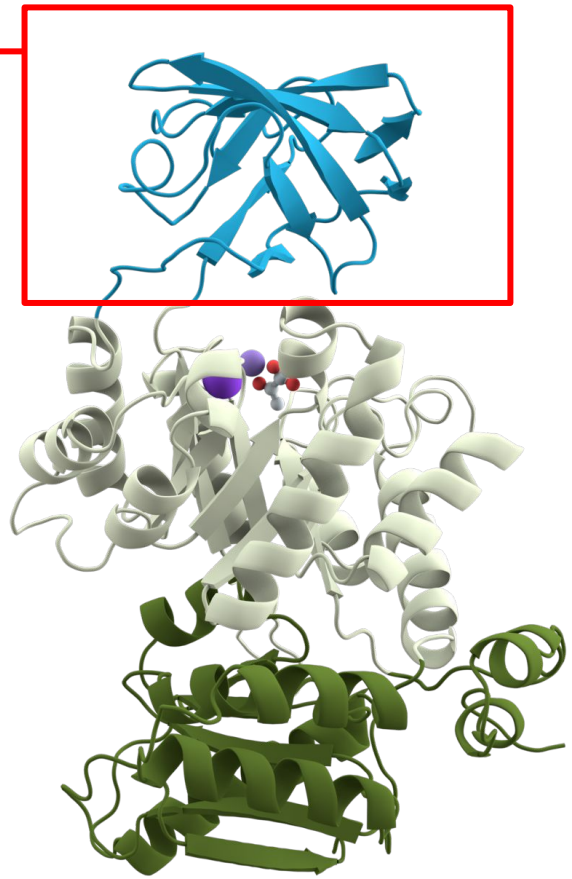
| | | |
|---|---|---|
| CLDCEKMAAL | HCALYCGQGA | QFLEAQIIQW |
| SNMDKLLPHS | SVLTWNTEIP | GITLVTEDIA |
| HYNNSVVRRE | WHNLISEEKT | GKRRSAAYVR |
| SHQRLLMGLM | VSELKDHFLR | HLQGVEKKKI |
| WVLHFNSRGS | AAEFAVFHIM | TRILEATNSL |

https://www.gjar-po.sk/~natalia.krajnakova2b/skola/r1/jan%20botto-smrt%20janosikova.pdf

# The language of genome

Dvanásti sokoli, sokolovia Tatier,
akoby ich bola mala jedna mater;
jedna mater mala, v mlieku kúpavala,
zlatým povojníčkom bola povíjala.

To sa chlapci, to sa jak oltárne sviece,
keď idú po háji, celý sa trbliece.

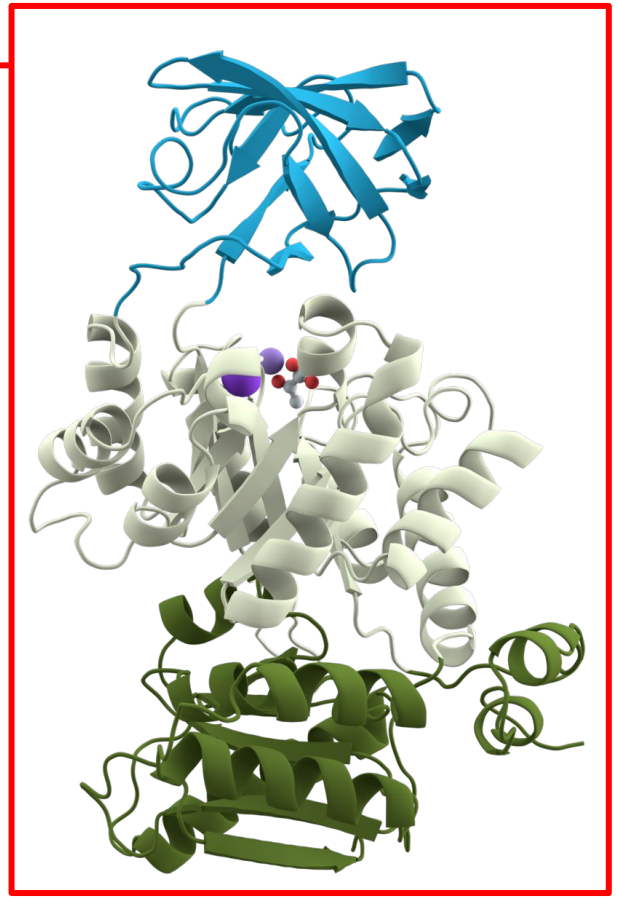


https://en.wikipedia.org/wiki/Protein_domain#/media/File:Pyruvate_kinase_protein_domains.png

# The language of genome

Dvanásti sokoli, sokolovia Tatier,
akoby ich bola mala jedna mater;
jedna mater mala, v mlieku kúpavala,
zlatým povojníčkom bola povíjala.

To sa chlapci, to sa jak oltárne sviece,
keď idú po háji, celý sa trbliece.



https://en.wikipedia.org/wiki/Protein_domain#/media/File:Pyruvate_kinase_protein_domains.png

# The language of genome

global protein classification tasks → text classification tasks

protein annotation tasks → text annotation

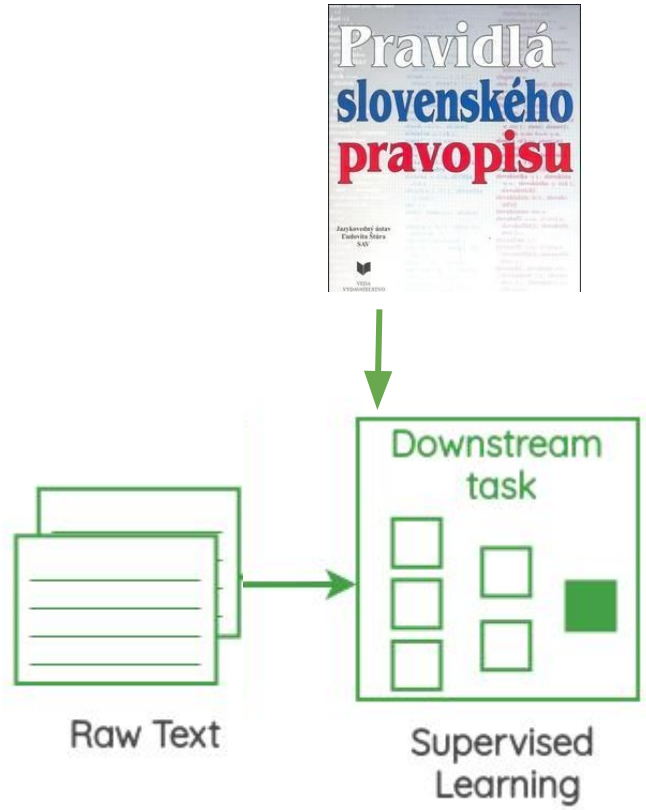
https://www.gjar-po.sk/~natalia.krajnakova2b/skola/r1/jan%20botto-smrt%20janosikova.pdf

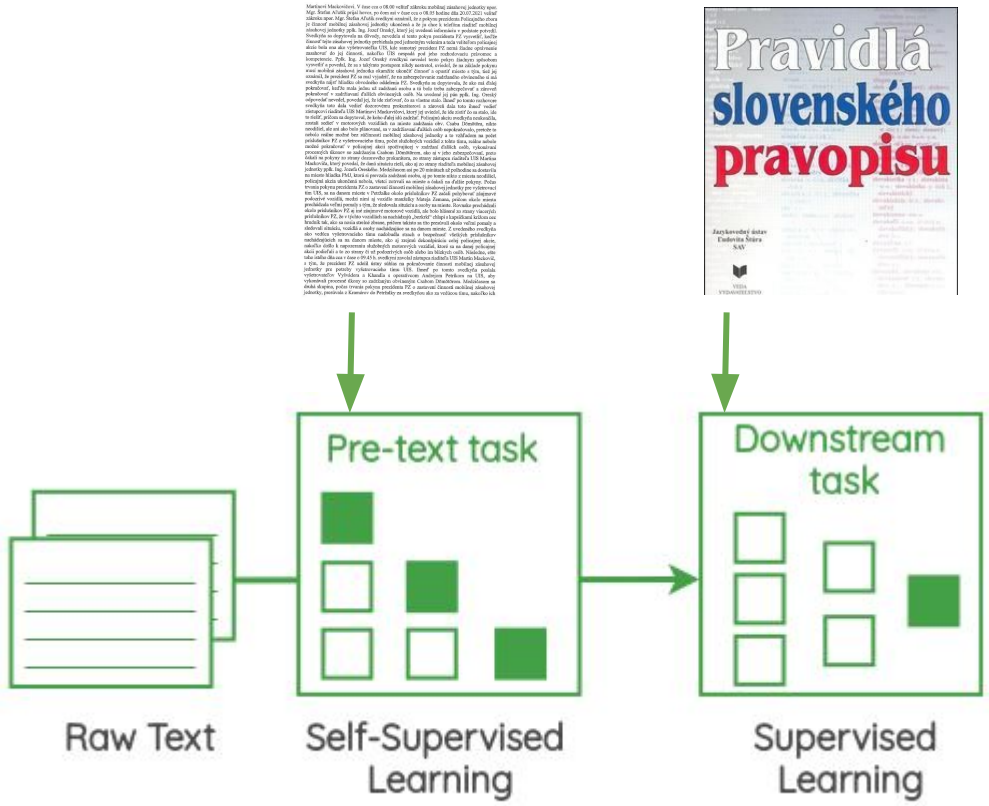
# The language of genome

**grammar?**

https://www.gjar-po.sk/~natalia.krajnakova2b/skola/r1/jan%20botto-smrt%20janosikova.pdf

# The language of genome



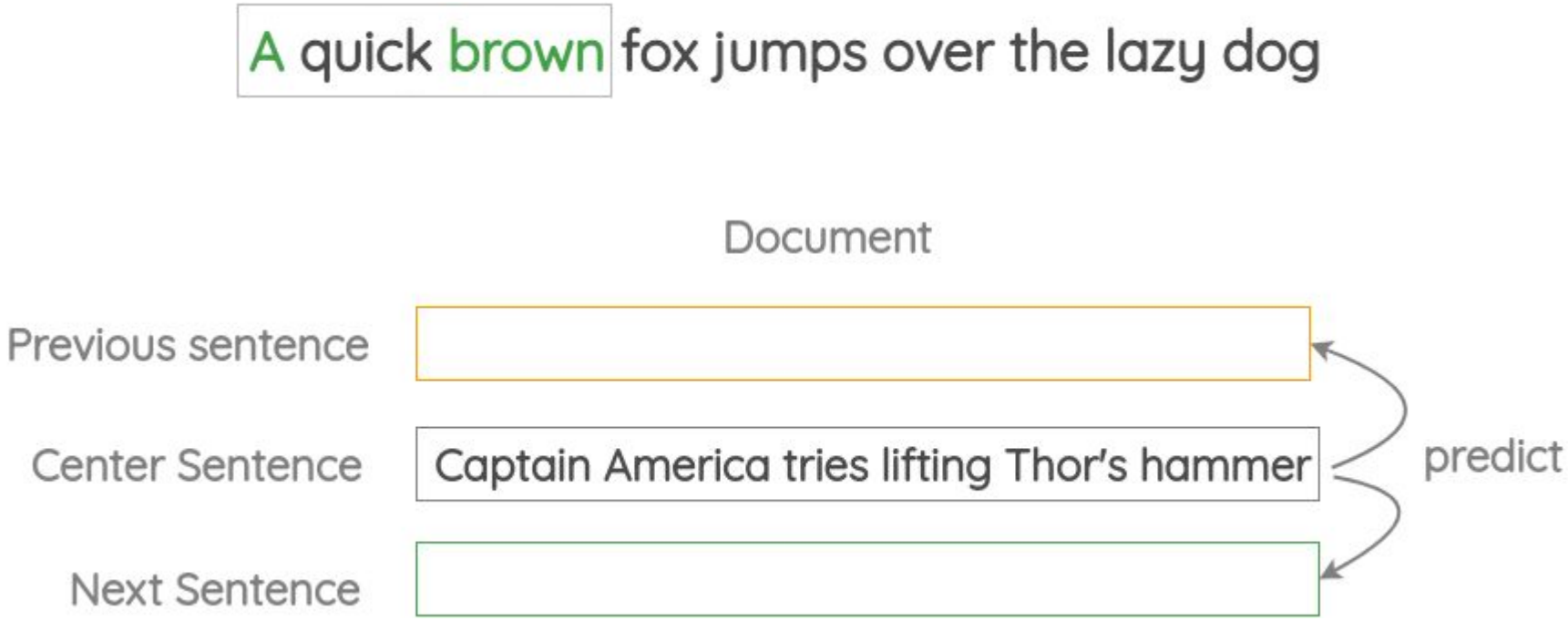

https://amitness.com/2020/05/self-supervised-learning-nlp/

# NLP | Self-supervised learning


A quick brown fox jumps over the lazy dog

https://amitness.com/2020/05/self-supervised-learning-nlp/

# NLP | Self-supervised learning




https://amitness.com/2020/05/self-supervised-learning-nlp/

# NLP | Self-supervised learning

Nothing is _____

https://amitness.com/2020/05/self-supervised-learning-nlp/

# UDSMProt

Nils Strodthoff
Patrick Wagner
Markus Wenzel
Wojciech Samek

Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin 10587, Germany

# UDSMProt

Universal deep sequence model

pre-training: Swiss-Prot

fine-tuning: specific tasks

Benchmarked tasks:

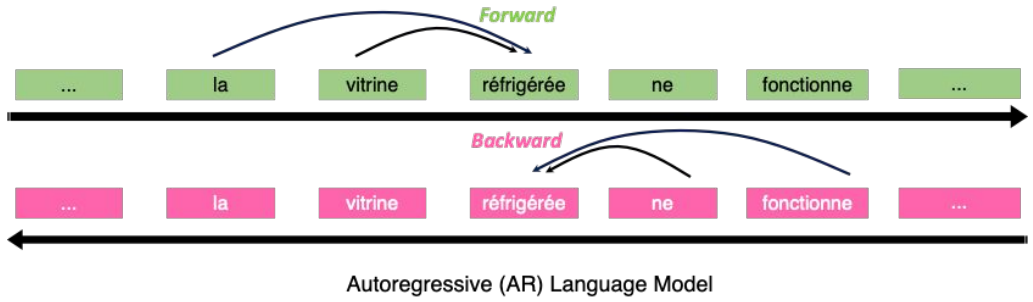
enzyme class prediction

gene ontology prediction

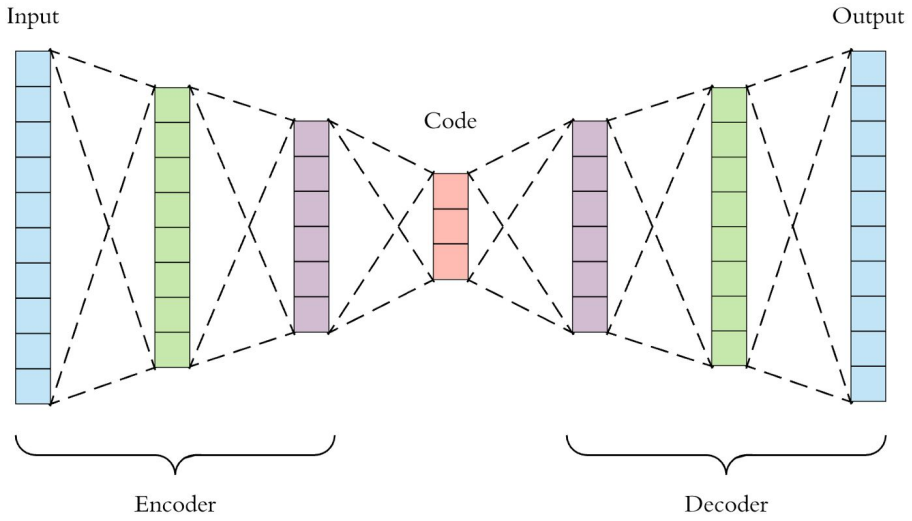
remote homology, fold detection

# Self-supervised pre-training | How-to?

## Autoregressive language modeling




## Autoencoding



# Self-supervised pre-training | How-to?

**Autoregressive language modeling**

less resources
faster
less data needed

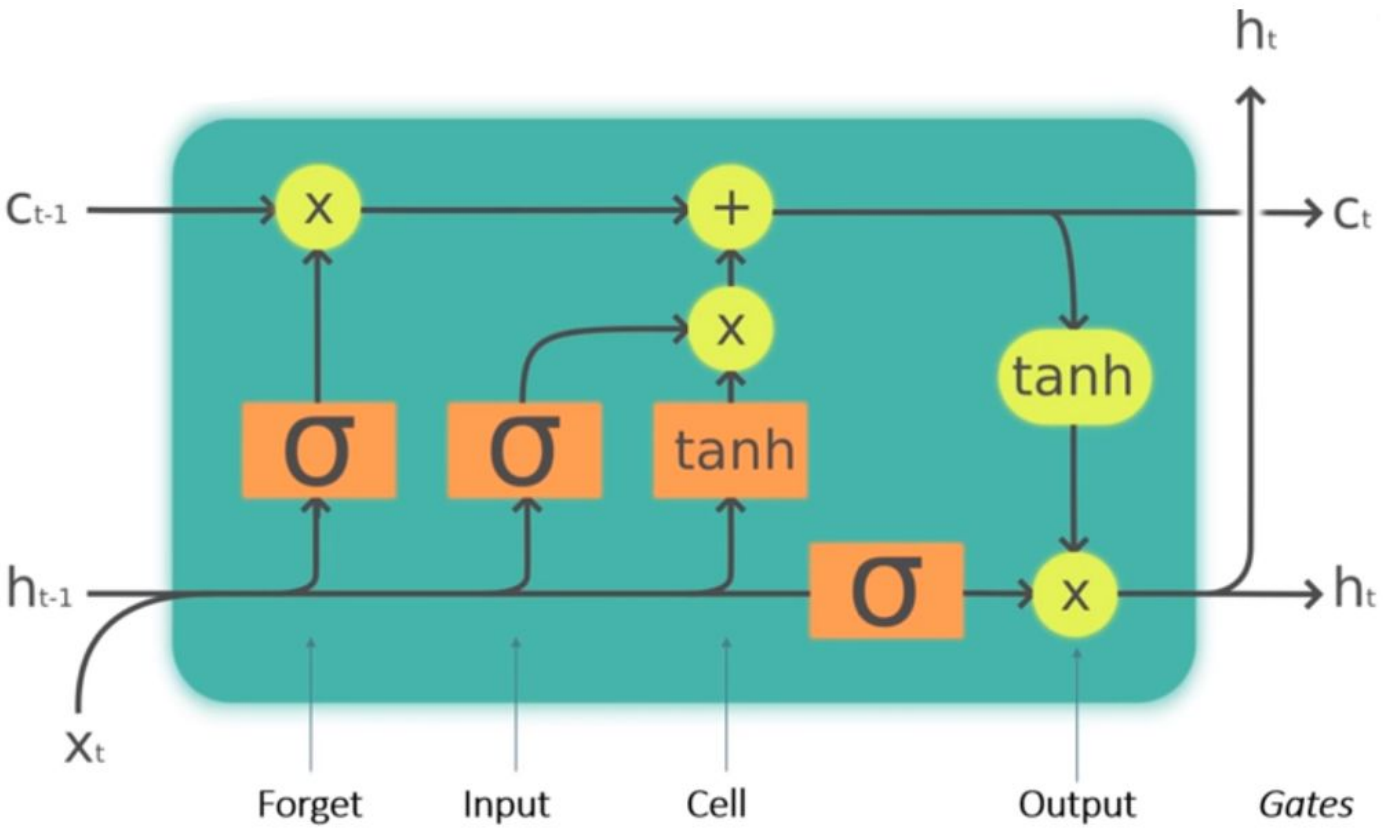
… less accurate

**Autoencoding**

way more resources
significantly slower
markedly more data

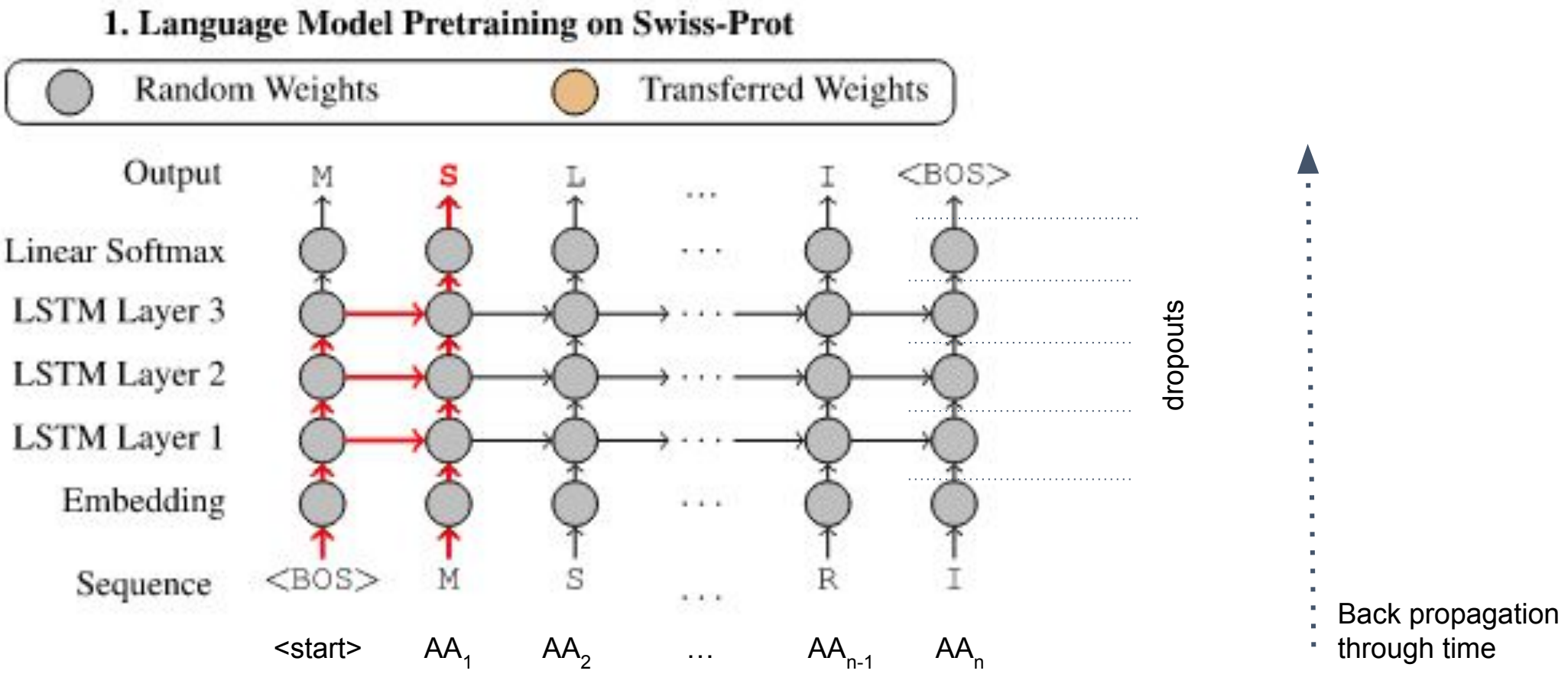
better! but not by much

# The model




google images

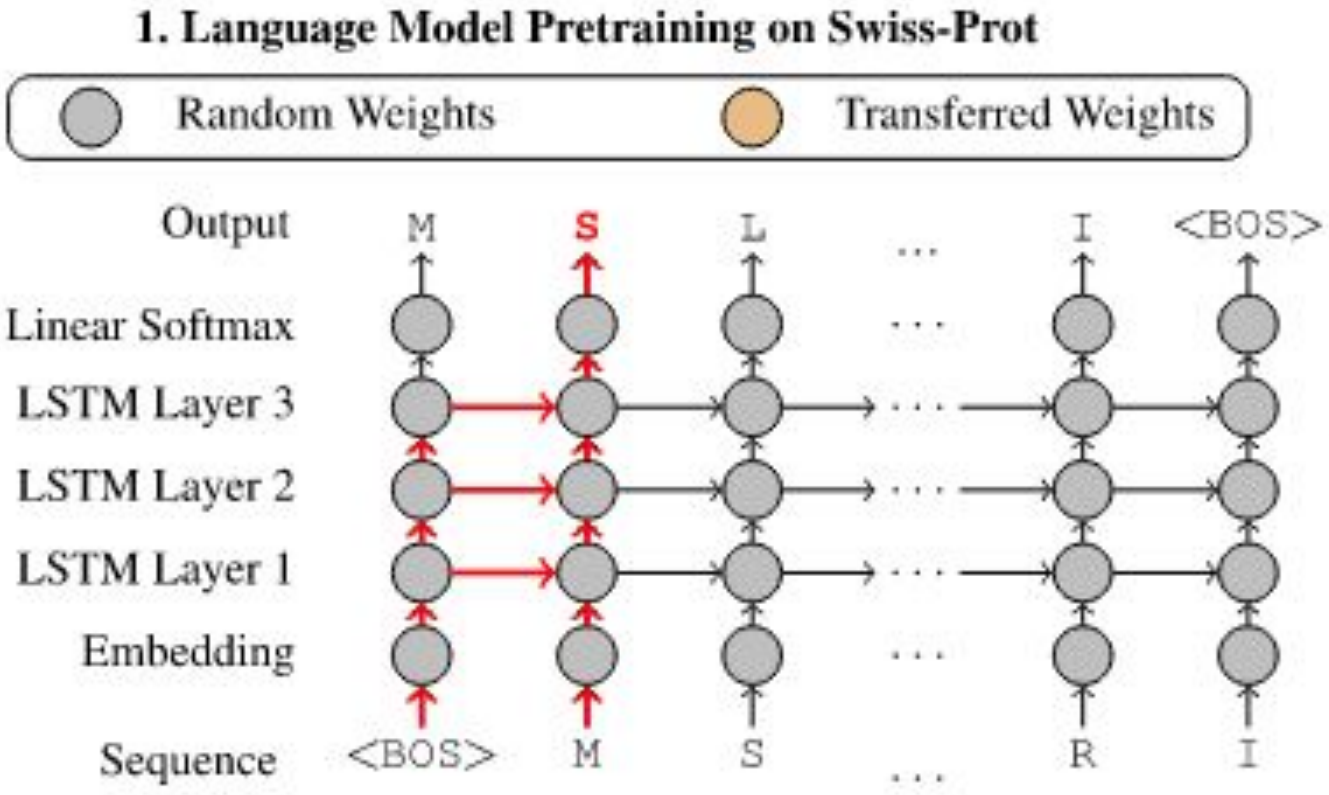
# … simplified


1. Language Model Pretraining on Swiss-Prot
Random Weights
Transferred Weights
Output
Linear Softmax
LSTM Layer 3
LSTM Layer 2
LSTM Layer 1
Embedding
Sequence
M
S
L
I
<BOS>
<BOS>
M
S
R
I
dropouts
<start>
AA1
AA2
…
AAn-1
AAn
Back propagation
through time

# … simplified

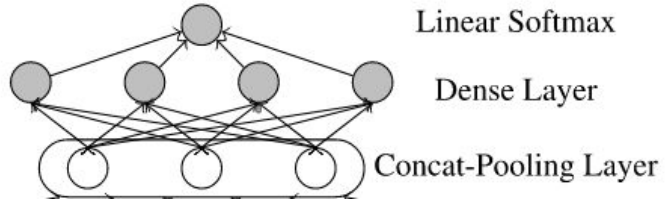
| Parameter | Value |
|---|---|
| Joint parameters | |
| Number of hidden units | 1150 |
| Number of layers | 3 |
| Embedding dimension | 400 |
| Backpropagation through time (bptt) | 70 |
| Gradient clipping | 0.25 |
| Weight decay | 1e-7 |
| Language-model-specific parameters | |
| Dropout $(p_o, p_h, p_i, p_e, p_w)$ | 0.5 · (0.25,0.1,0.2,0.02,0.15) |
| Classifier-specific parameters | |
| Dropout $(p_o, p_h, p_i, p_e, p_w)$ | 0.5 · (0.4,0.2,0.6,0.1,0.5) |
| Max. length (context size) | 1024 |
| Number of hidden units (head) | 50 |

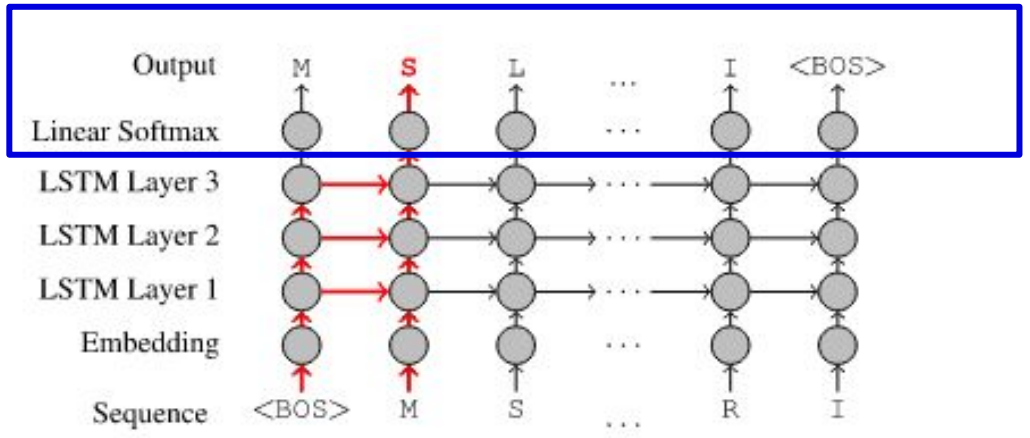
# … simplified




swissprot

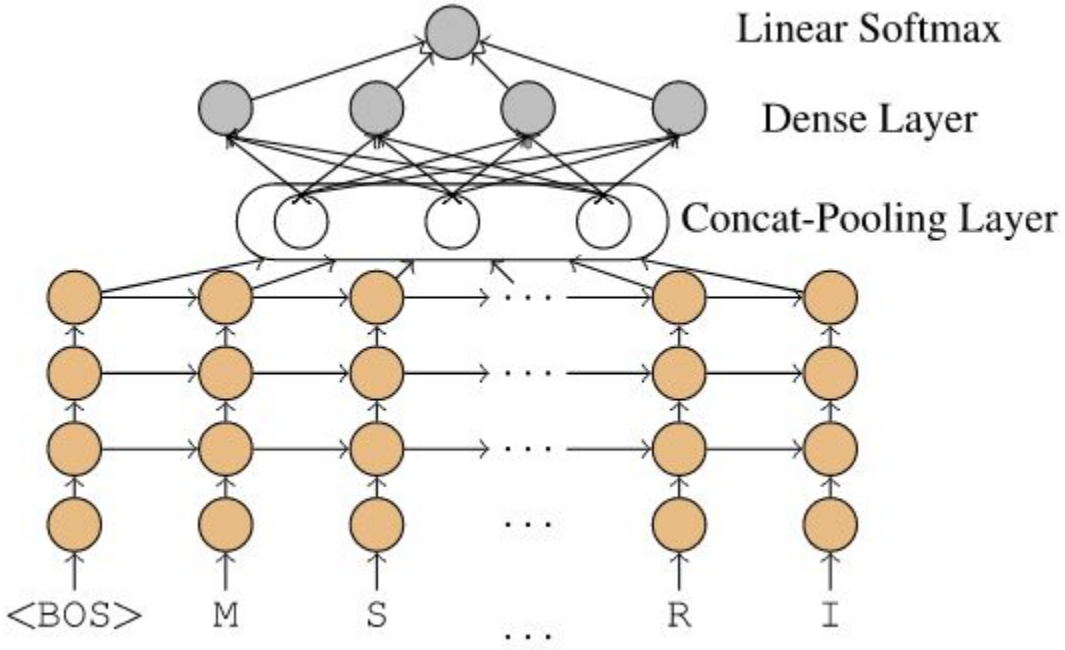
# … simplified




google stock

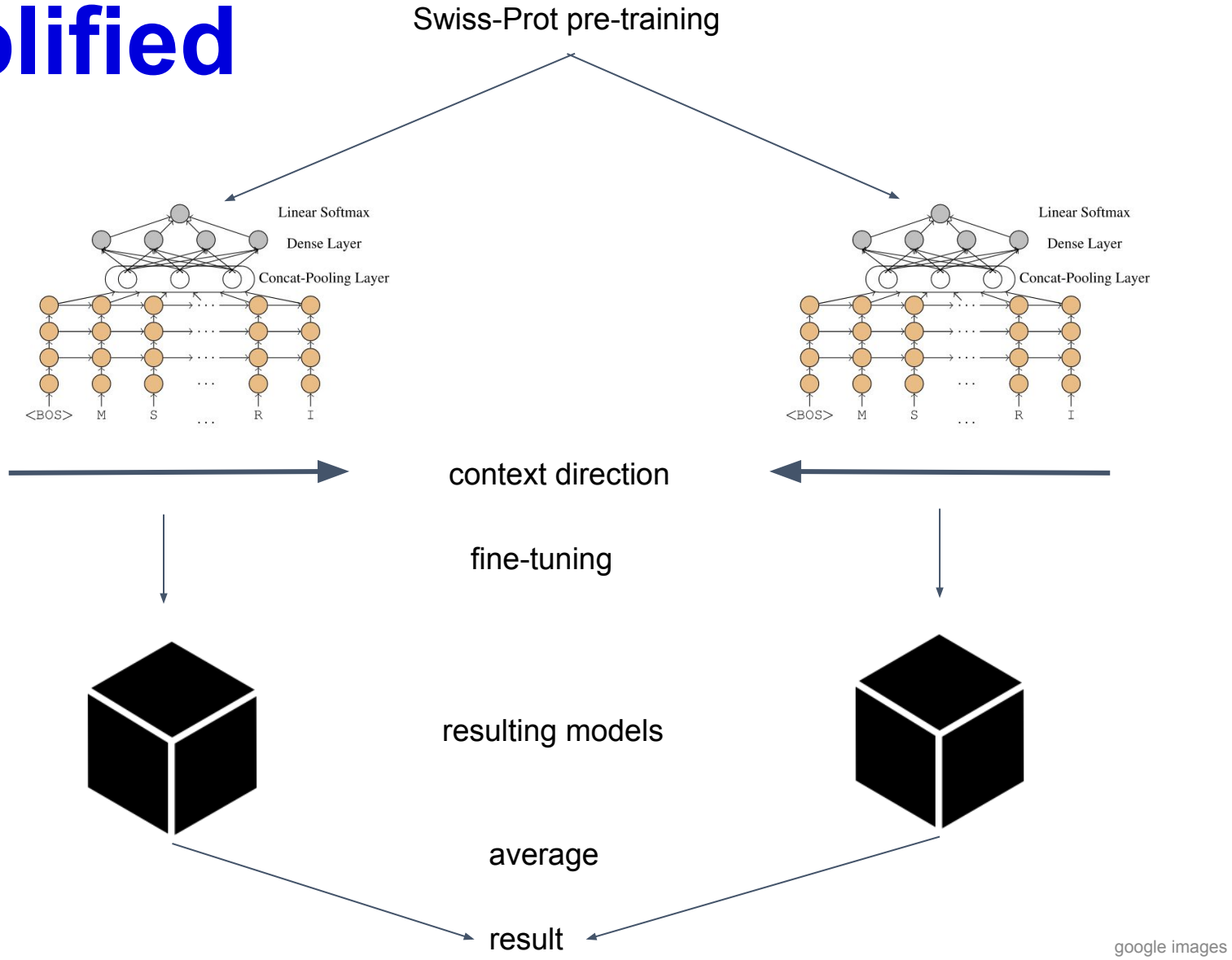
# … simplified




Unfreeze
Decrease learning rate

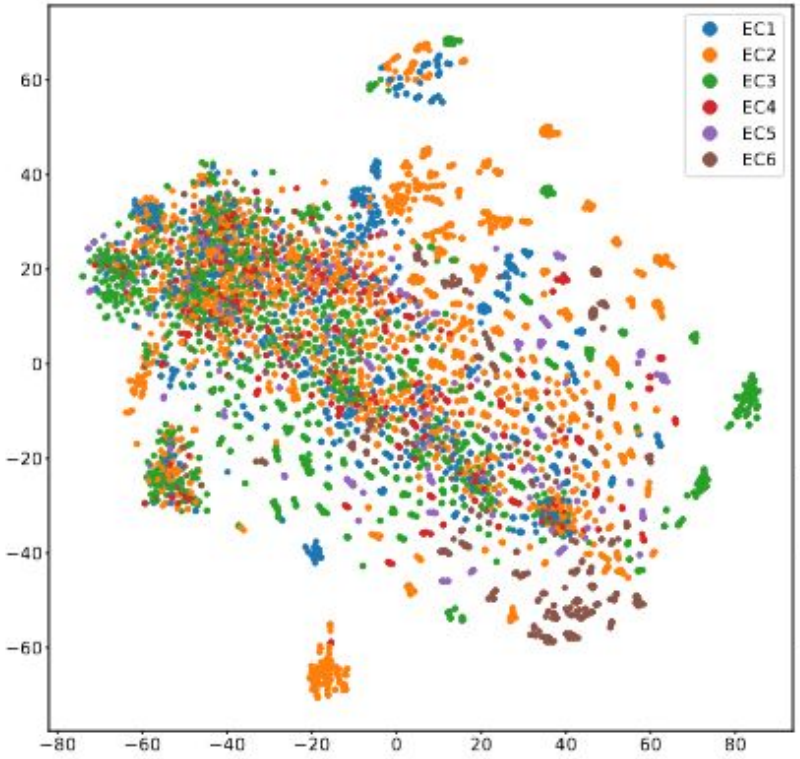
task-specific data

# … simplified

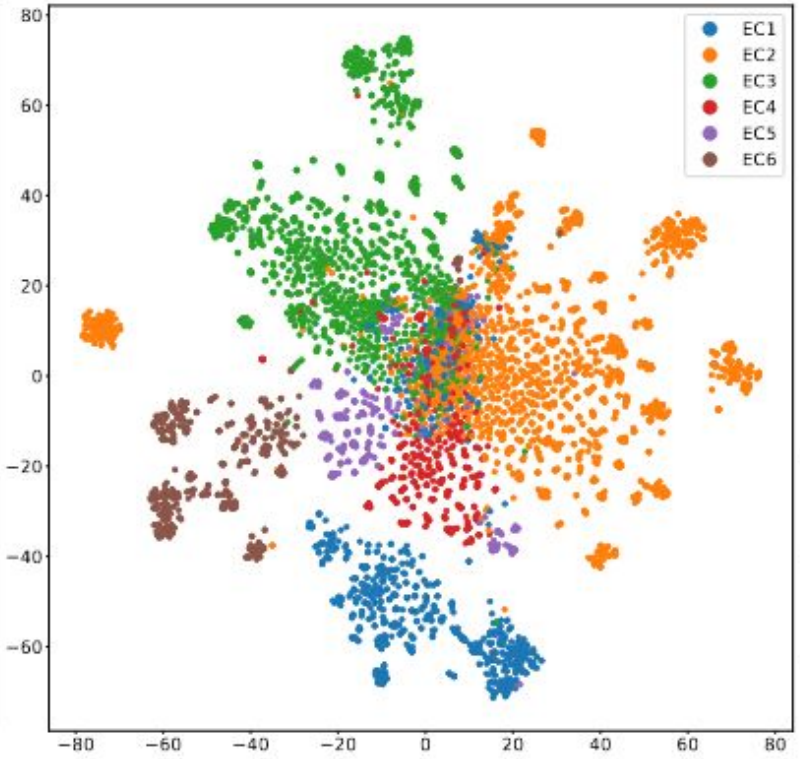



google images

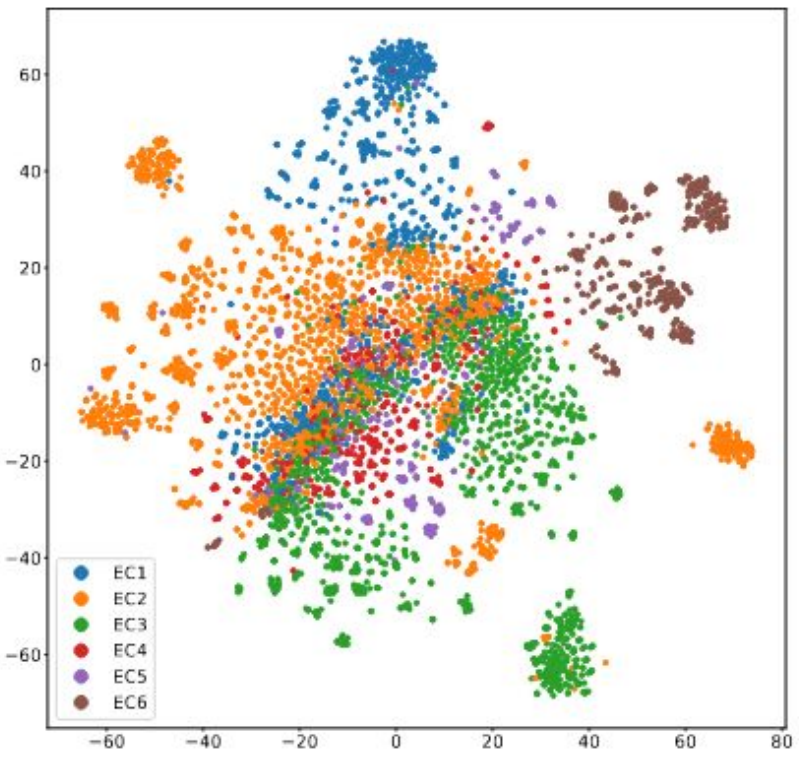
# … simplified




(a) After language model pretraining.



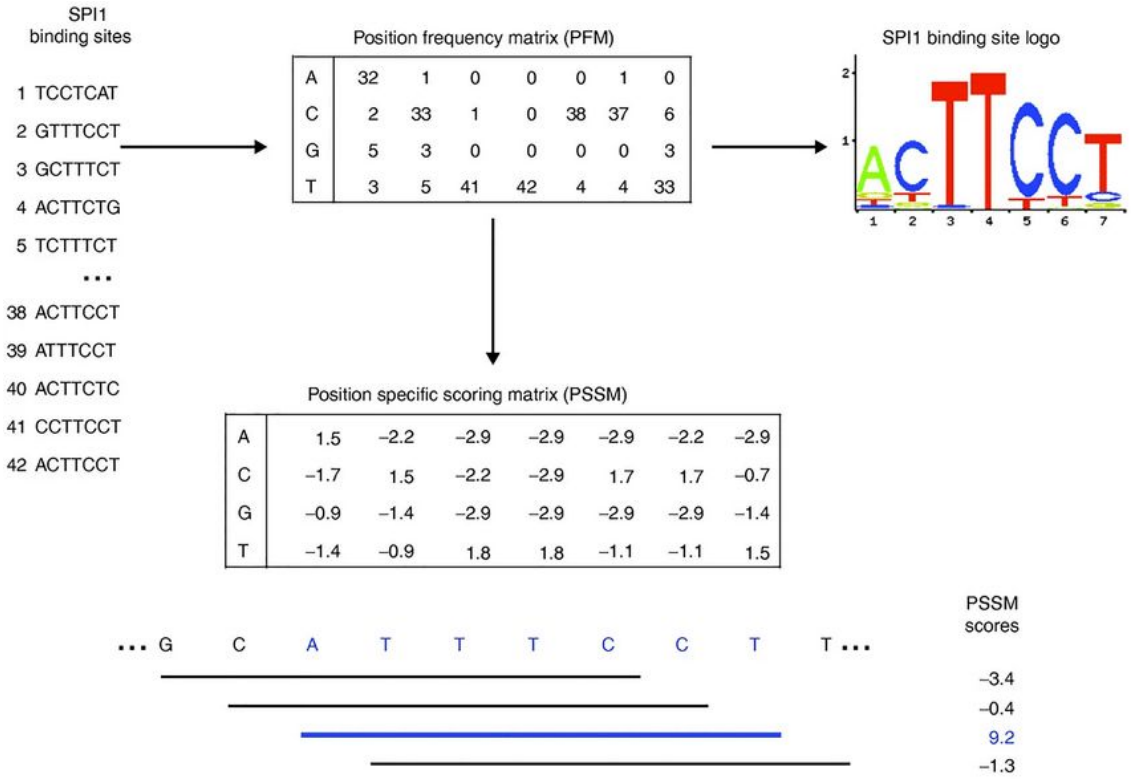

(b) After language model pretraining and finetuning.




(c) After training from scratch.

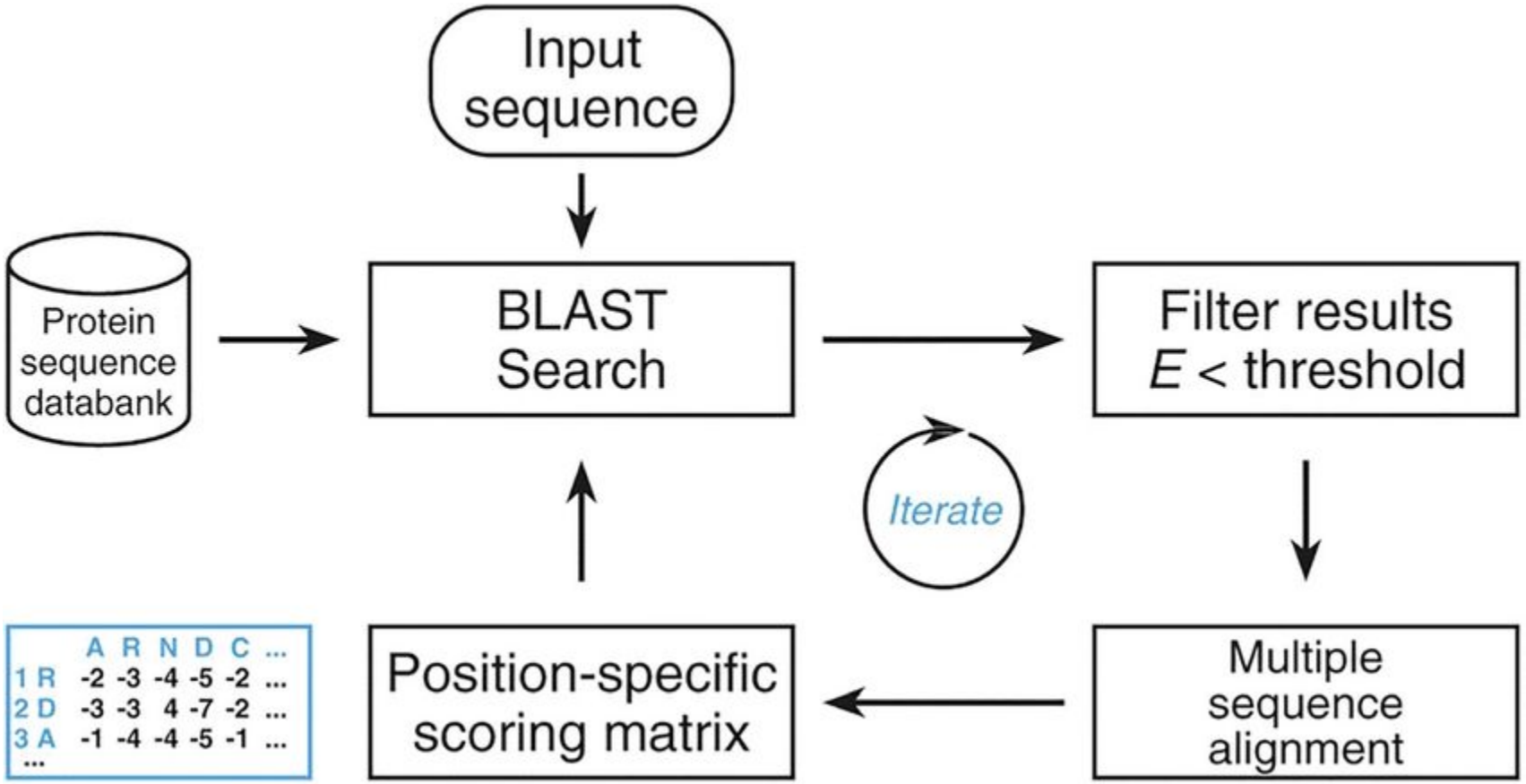
# The baseline model

- Best state-of-the-art methods
- PSSMs (motifs)




https://en.wikipedia.org/wiki/Position_weight_matrix
https://www.researchgate.net/publication/345667774_A_survey_on_deep_learning_in_DNARNA_motif_mining

# The baseline model




https://www.researchgate.net/publication/354444464_Protein_active_site_prediction_for_early_drug_discovery_and_designing

# Results

Tasks:

- Language modelling
- Enzyme class prediction
- Gene ontology
- Remote homology and fold detection

# Language modeling

The problem of similarity threshold

Next character accuracy on

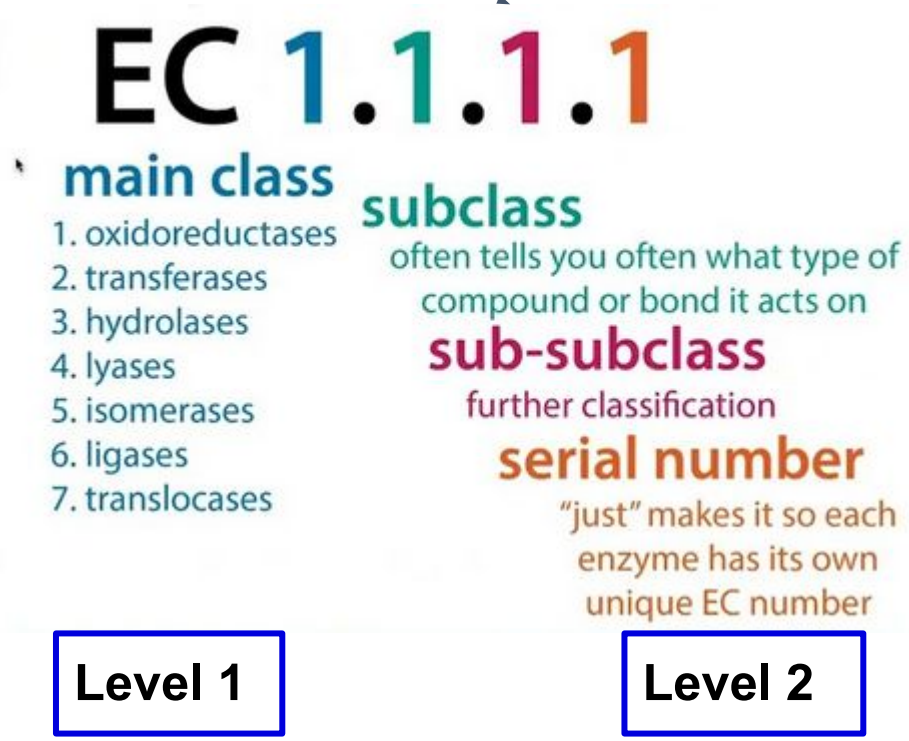
| | |
|---|---|
| random split: | .409 |
| cluster-based split: | .244 |
| random guessing: | .04 |

… still not a problem

# Enzyme class prediction

Level 0

not an enzyme




Level 1

Level 2

https://www.youtube.com/watch?v=8yO1XEzoVIE

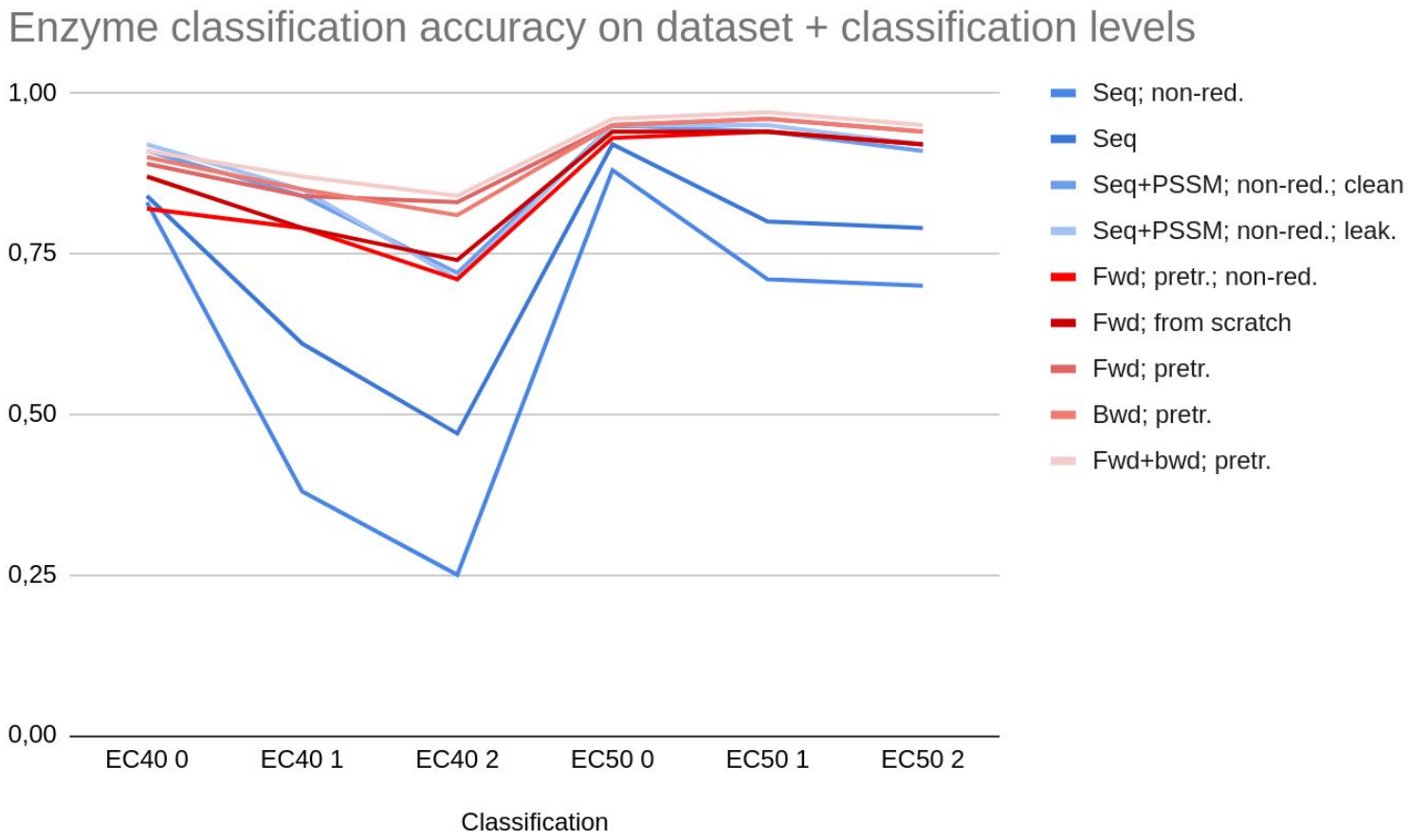
# Enzyme class prediction

| | Level | EC40 | | | EC50 | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 0 | 1 | 2 |
| Baseline | Seq; non-red. | 0.83 | 0.38 | 0.25 | 0.88 | 0.71 | 0.70 |
| | Seq | 0.84 | 0.61 | 0.47 | 0.92 | 0.80 | 0.79 |
| | Seq+PSSM; non-red.; clean | 0.91 | 0.84 | 0.72 | 0.95 | 0.94 | 0.91 |
| | Seq+PSSM; non-red.; leak. | **0.92** | 0.85 | 0.71 | 0.95 | 0.95 | 0.92 |
| *UDSMProt* | Fwd; pretr.; non-red. | 0.82 | 0.79 | 0.71 | 0.93 | 0.94 | 0.92 |
| | Fwd; from scratch | 0.87 | 0.79 | 0.74 | 0.94 | 0.94 | 0.92 |
| | Fwd; pretr. | 0.89 | 0.84 | 0.83 | 0.95 | 0.96 | 0.94 |
| | Bwd; pretr. | 0.90 | 0.85 | 0.81 | 0.95 | 0.96 | 0.94 |
| | Fwd+bwd; pretr. | 0.91 | **0.87** | **0.84** | **0.96** | **0.97** | **0.95** |

*Note*: The best-performing classifiers are marked in bold face.

Fwd/bwd, training in forward/backward direction; seq, raw sequence as input; non-red, training on non-redundant sequences, i.e. representatives only; pretr., using language model pre-training; leak., leakage PSSM features computed on the full dataset.

# Enzyme class prediction




| Dataset | Best | Accuracy |
|---|---|---|
| EC40 0 | **Baseline** | 0,92 |
| EC40 1 | **UDSMProt** | 0,87 |
| EC40 2 | **UDSMProt** | 0,84 |
| EC50 0 | **UDSMProt** | 0,96 |
| EC50 1 | **UDSMProt** | 0,97 |
| EC50 2 | **UDSMProt** | 0,95 |

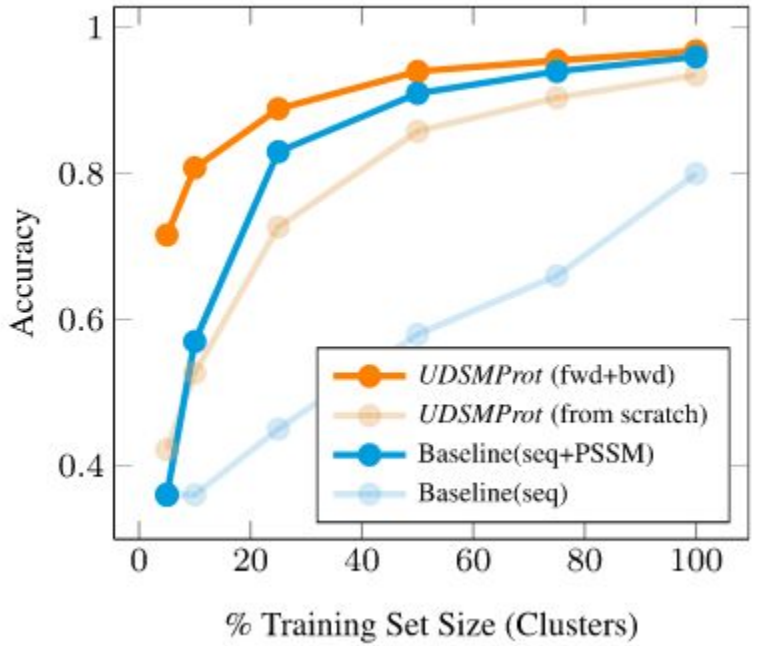
# Enzyme class prediction

| | Level | DEEPre (acc.) | | | ECPred (mean $F_1$) | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 0 | 1 |
| | *ECPred* | — | — | — | 0.96 | **0.96** |
| | *DEEPre* (seq+PSSM) | 0.88 | 0.82 | 0.43 | — | — |
| | Baseline[a] (seq+PSSM) | **0.91** | **0.84** | 0.59 | 0.97 | 0.94 |
| *UDSMProt*[a] | Fwd; pretr. | 0.86 | 0.81 | 0.75 | 0.95 | 0.93 |
| | Bwd; pretr. | 0.86 | 0.83 | 0.73 | 0.97 | 0.93 |
| | Fwd+bwd; pretr. | 0.87 | **0.84** | **0.78** | 0.97 | 0.94 |
| | Fwd; pretr.; red. | — | — | — | 0.97 | 0.95 |
| | Bwd; pretr.; red. | — | — | — | 0.97 | 0.95 |
| | Fwd+bwd; pretr.; red. | — | — | — | **0.98** | 0.95 |

*Note*: Results on the *DEEPre* dataset were evaluated using 5-fold cross-validation.

Fwd/bwd, training in forward/backward direction; seq, raw sequence as input; pretr., using language model pre-training.

# Training dataset size



# Gene Ontology prediction

Cellular Components (CCOs)

Molecular Functions (MFOs)

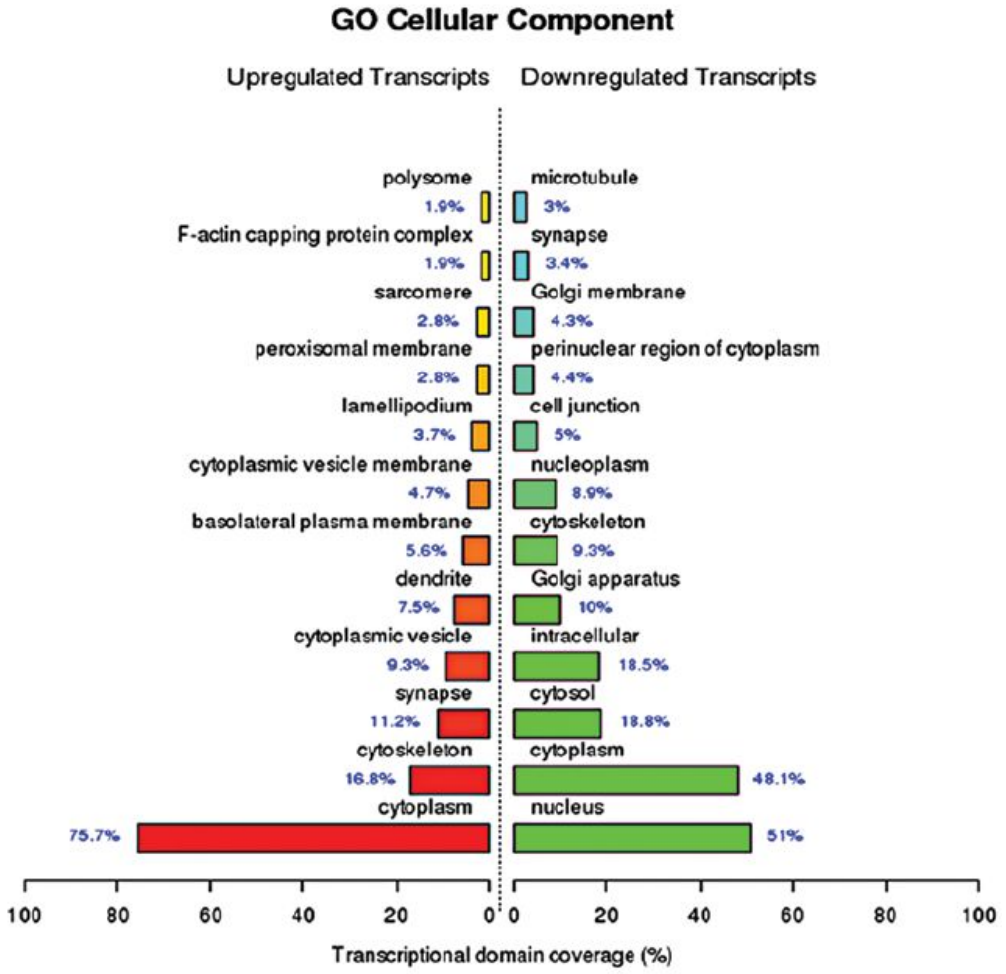
Biological Processes (BPOs)

# Gene Ontology prediction

Cellular Components (CCOs)

Molecular Functions (MFOs)

Biological Processes (BPOs)



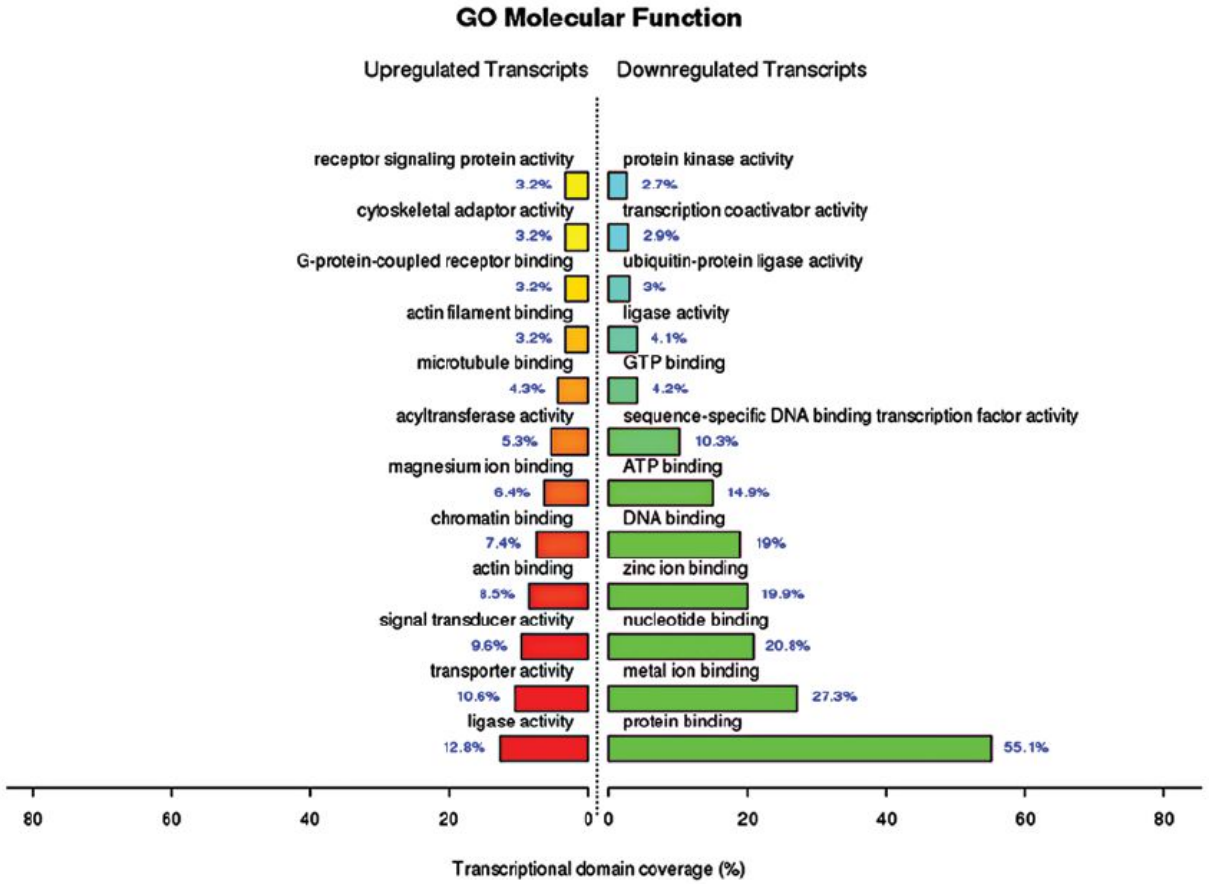

https://www.researchgate.net/publication/261187744_Expression_analysis_of_serum_microRNAs_in_idiopathic_pulmonary_fibrosis

# Gene Ontology prediction

Cellular Components (CCOs)

Molecular Functions (MFOs)

Biological Processes (BPOs)



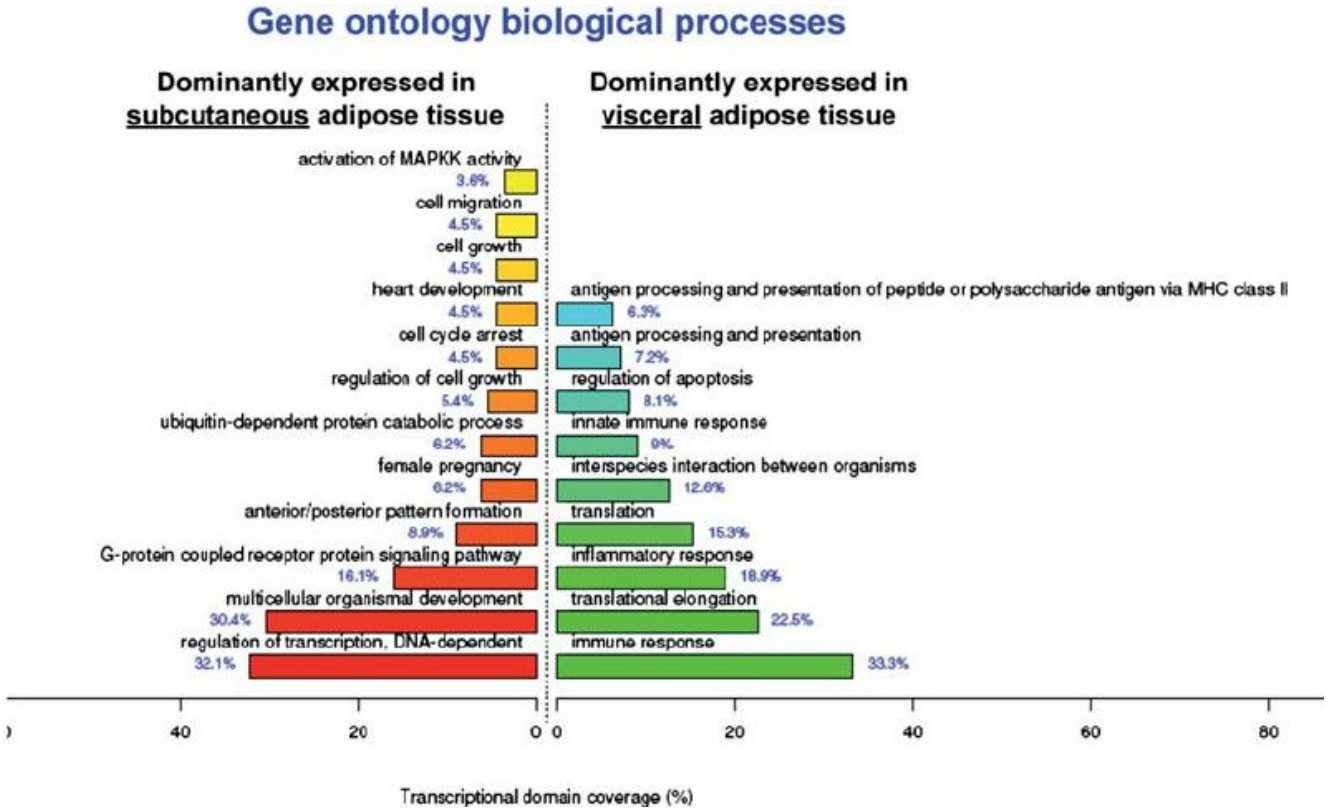

https://www.researchgate.net/publication/261187744_Expression_analysis_of_serum_microRNAs_in_idiopathic_pulmonary_fibrosis

# Gene Ontology prediction

Cellular Components (CCOs)

Molecular Functions (MFOs)

Biological Processes (BPOs)

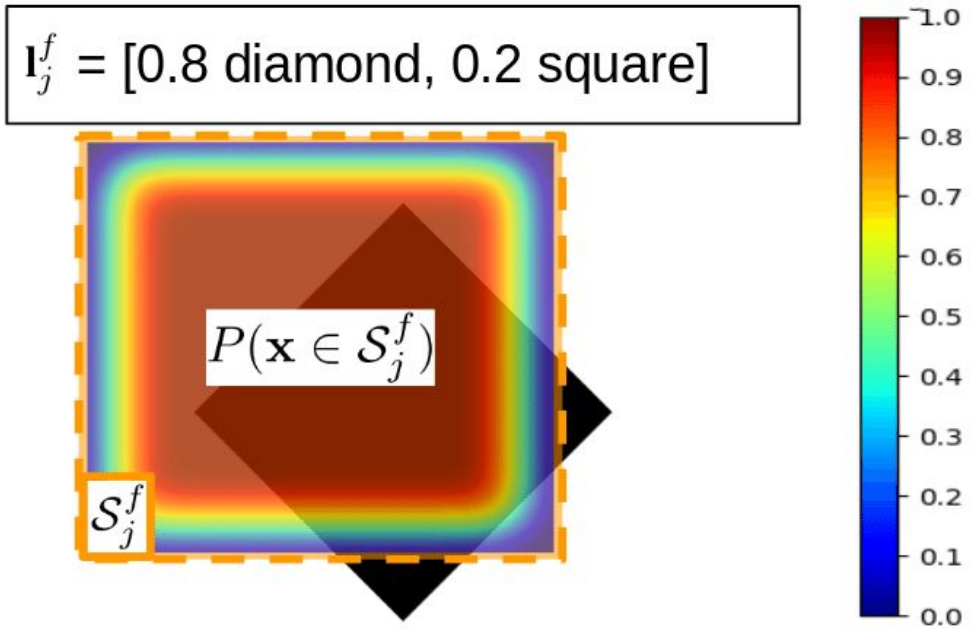



https://www.researchgate.net/publication/261187744_Expression_analysis_of_serum_microRNAs_in_idiopathic_pulmonary_fibrosis

# Gene Ontology prediction

| | Methods | $F_{max}$ | | | $S_{min}$ | | | AUPR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MFO | BPO | CCO | MFO | BPO | CCO | MFO | BPO | CCO |
| Single | *Naive* | 0.306 | 0.318 | 0.605 | 12.105 | 38.890 | 9.646 | 0.150 | 0.219 | 0.512 |
| | *DiamondScore* | 0.548 | 0.439 | 0.621 | 8.736 | 34.060 | 7.997 | 0.362 | 0.240 | 0.363 |
| | *DeepGO* | 0.449 | 0.398 | 0.667 | 10.722 | 35.085 | 7.861 | 0.409 | 0.328 | 0.696 |
| | *DeepGOCNN* | 0.409 | 0.383 | 0.663 | 11.296 | 36.451 | 8.642 | 0.350 | 0.316 | 0.688 |
| Ensemble | *DeepText2GO* | **0.627** | 0.441 | 0.694 | 5.240 | 17.713 | **4.531** | **0.605** | 0.336 | **0.729** |
| | *GOLabeler* | 0.580 | 0.370 | 0.687 | **5.077** | **15.177** | 5.518 | 0.546 | 0.225 | 0.700 |
| | *DeepGOPlus* | 0.585 | 0.474 | **0.699** | 8.824 | 33.576 | 7.693 | 0.536 | 0.407 | 0.726 |
| *UDSMProt*[a] | Fwd; from scratch | 0.418 | 0.303 | 0.655 | 14.906 | 47.208 | 12.929 | 0.304 | 0.284 | 0.612 |
| | Fwd; pretr. | 0.465 | 0.404 | 0.683 | 10.578 | 36.667 | 8.210 | 0.406 | 0.345 | 0.695 |
| | Bwd; pretr. | 0.465 | 0.403 | 0.664 | 10.802 | 36.361 | 8.210 | 0.414 | 0.348 | 0.685 |
| | Fwd+bwd; pretr. | 0.481 | 0.411 | 0.682 | 10.505 | 36.147 | 8.244 | 0.472 | 0.356 | 0.704 |
| | Bwd+bwd; pretr. + *DiamondScore* | 0.582 | **0.475** | 0.697 | 8.787 | 33.615 | 7.618 | 0.548 | **0.422** | 0.728 |

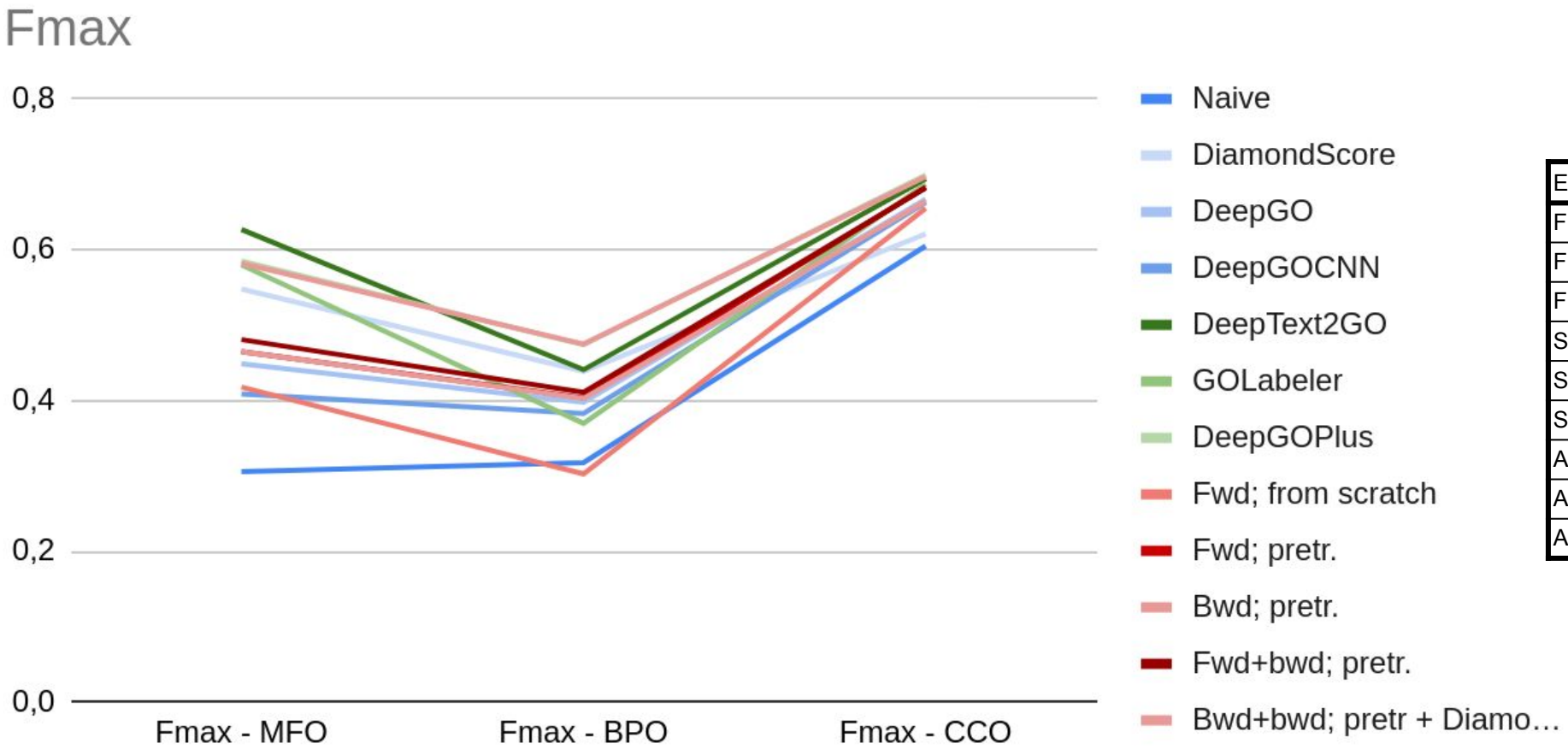
# $F_{max}$, $S_{min}$, AUPR





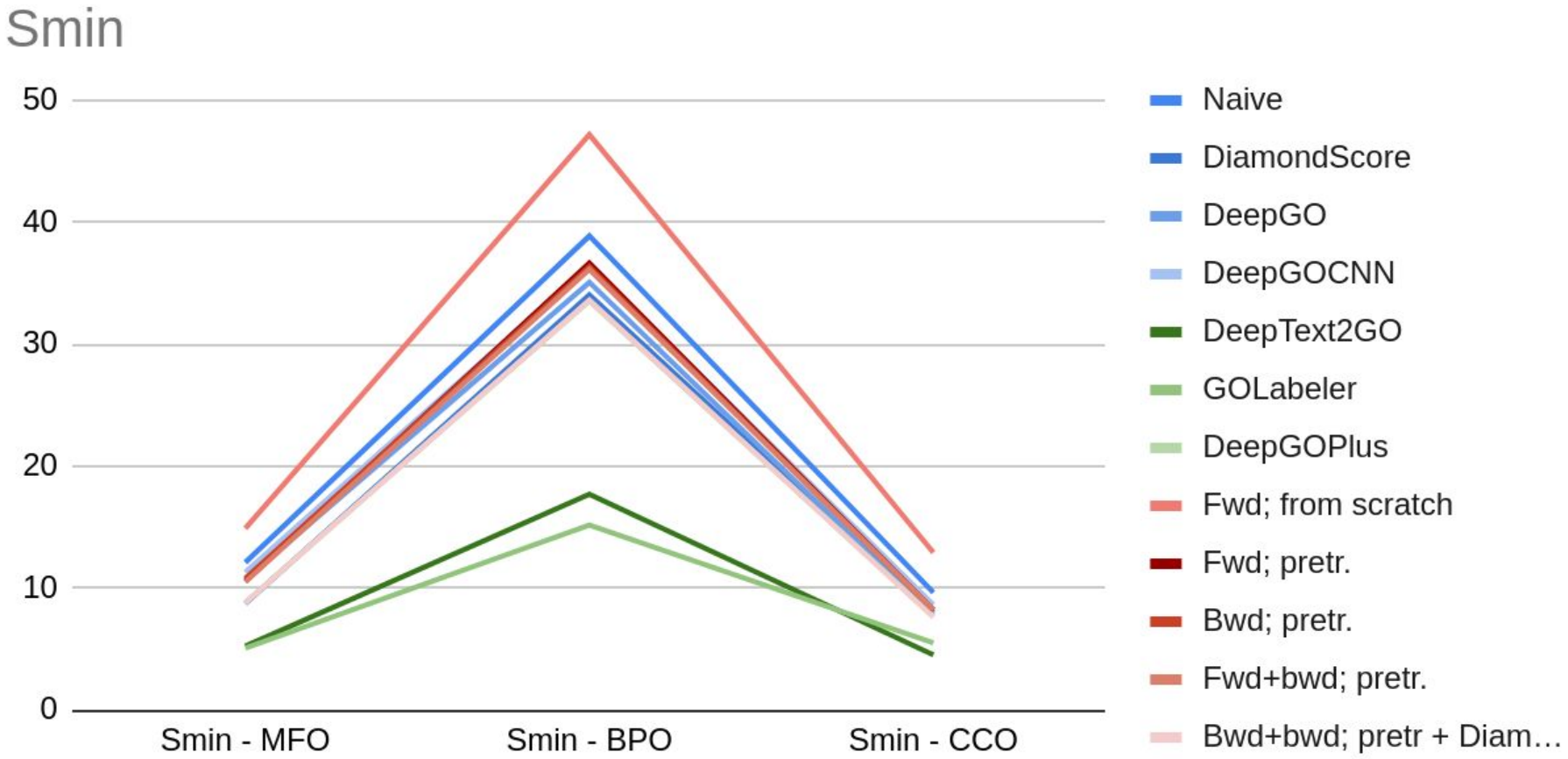
https://www.researchgate.net/publication/261187744_Expression_analysis_of_serum_microRNAs_in_idiopathic_pulmonary_fibrosis
https://www.pinterest.co.uk/pin/692991461400478556/

# Gene Ontology prediction



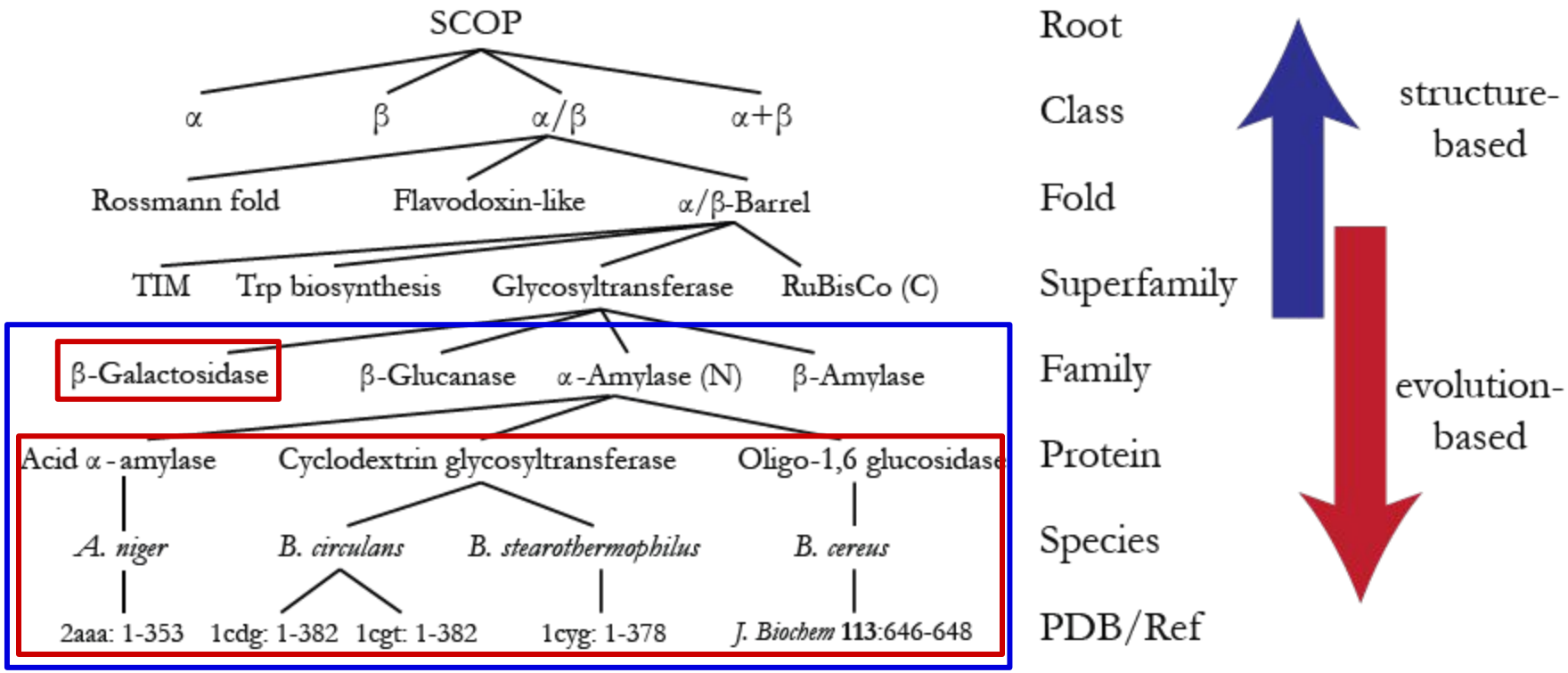

| Evaluation | Classification | Best Model |
|---|---|---|
| Fmax | MFO | **DiamondScore** |
| Fmax | BPO | **DiamondScore** |
| Fmax | CCO | **UDSMProt** |
| Smin | MFO | **DiamondScore** |
| Smin | BPO | **DiamondScore** |
| Smin | CCO | **DeepGO** |
| AUPR | MFO | **UDSMProt** |
| AUPR | BPO | **UDSMProt** |
| AUPR | CCO | **UDSMProt** |

# Gene Ontology prediction




| Evaluation | Classification | Best Model |
|---|---|---|
| Fmax | MFO | **DiamondScore** |
| Fmax | BPO | **DiamondScore** |
| Fmax | CCO | **UDSMProt** |
| Smin | MFO | **DiamondScore** |
| Smin | BPO | **DiamondScore** |
| Smin | CCO | **DeepGO** |
| AUPR | MFO | **UDSMProt** |
| AUPR | BPO | **UDSMProt** |
| AUPR | CCO | **UDSMProt** |

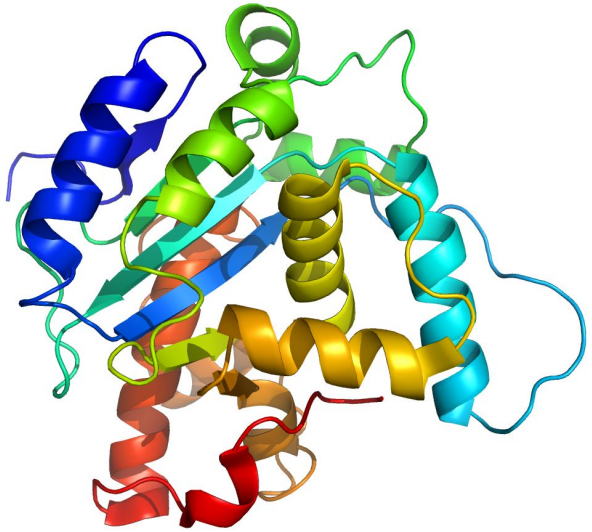
# Remote Homology and Fold detection



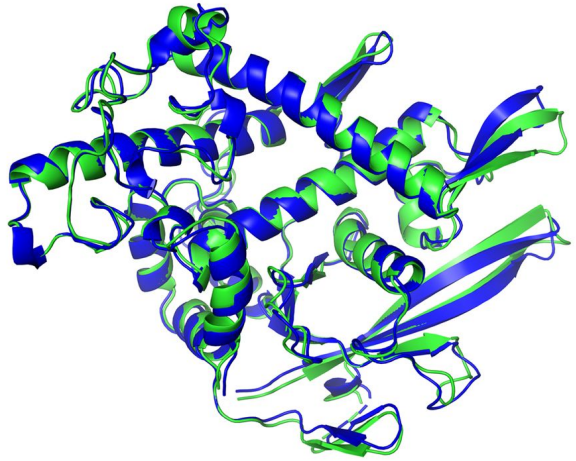

**Homologous** **Remote Homologous**

https://scop.berkeley.edu/help/ver=1.69

# Remote Homology and Fold detection



Our protein



Reference protein

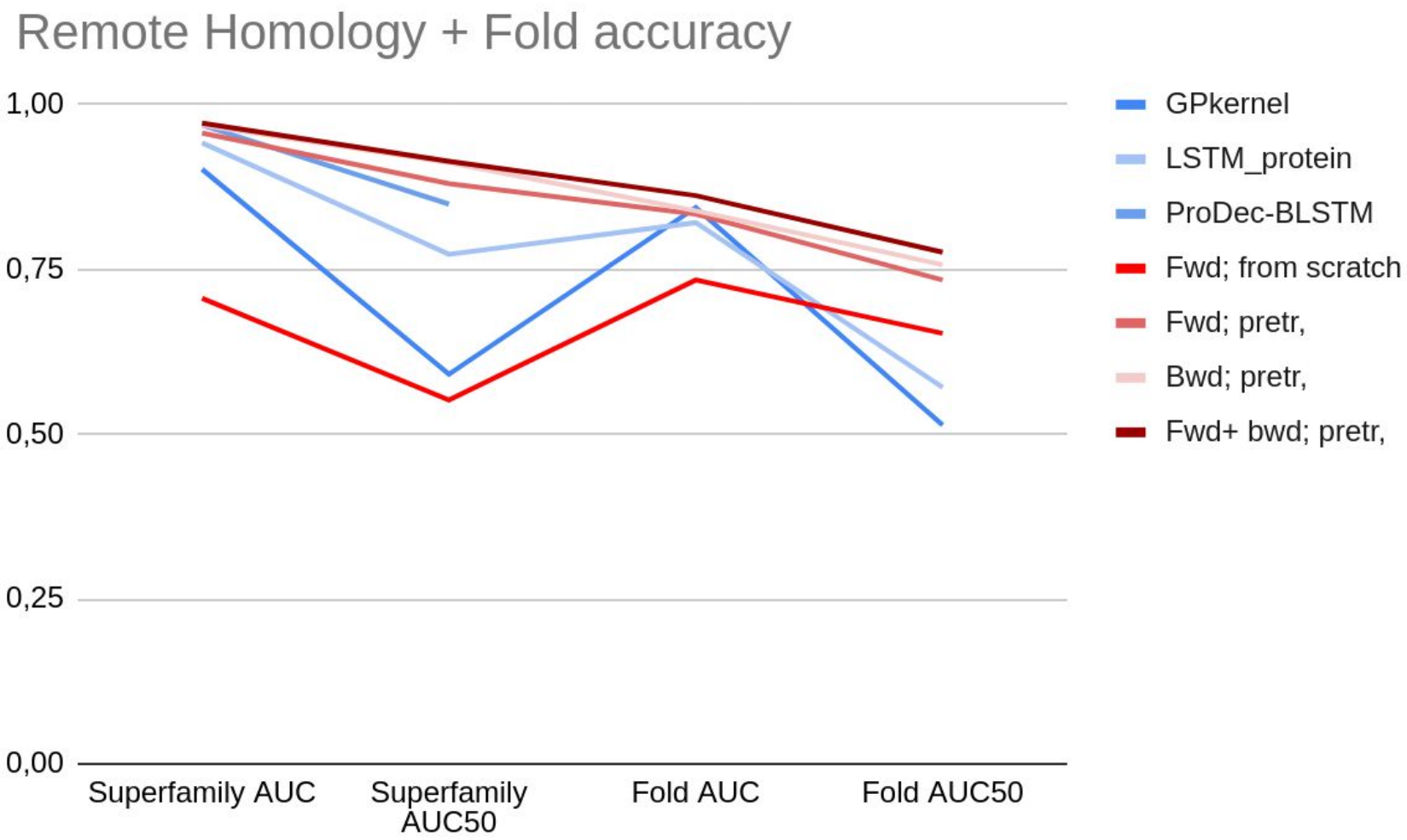
**Same superfamily?**
**Same fold?**

https://www.science.org/content/article/game-has-changed-ai-triumphs-solving-protein-structures
https://frontlinegenomics.com/alphafold-2-protein-structure-prediction-software-for-all/

# Remote Homology and Fold detection

| | Methods | Superfamily level | | Fold level | |
|---|---|---|---|---|---|
| | | AUC | $AUC_{50}$ | AUC | $AUC_{50}$ |
| | *GPkernel* | 0.902 | 0.591 | 0.844 | 0.514 |
| | *LSTM_protein* | 0.942 | 0.773 | 0.821 | 0.571 |
| | *ProDec-BLSTM* | 0.969 | 0.849 | — | — |
| *UDSMProt*[a] | Fwd; from scratch | 0.706 | 0.552 | 0.734 | 0.653 |
| | Fwd; pretr. | 0.957 | 0.880 | 0.834 | 0.734 |
| | Bwd; pretr. | 0.969 | 0.912 | 0.839 | 0.757 |
| | Fwd+bwd; pretr. | **0.972** | **0.914** | **0.862** | **0.776** |

# Remote Homology and Fold detection

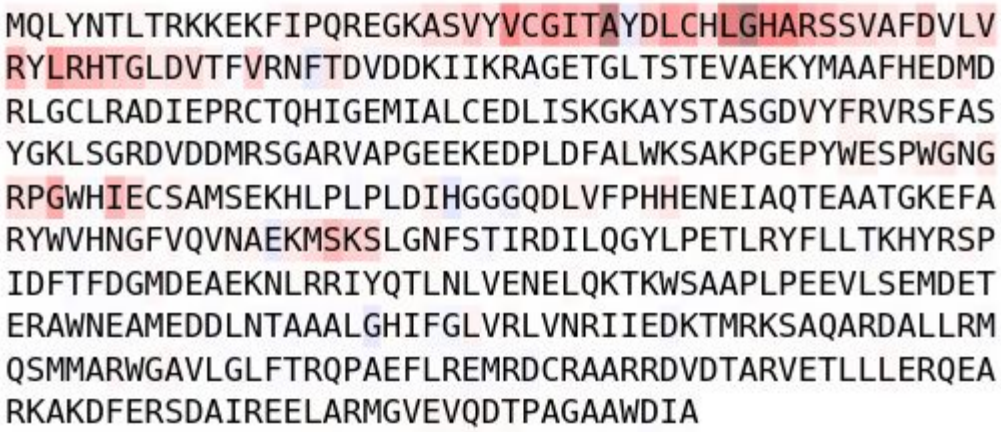


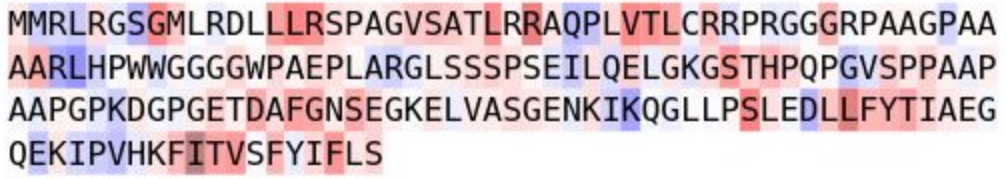

| Task | Best |
|---|---|
| Superfamily AUC | **UDSMProt** |
| Superfamily AUC50 | **UDSMProt** |
| Fold AUC | **UDSMProt** |
| Fold AUC50 | **UDSMProt** |

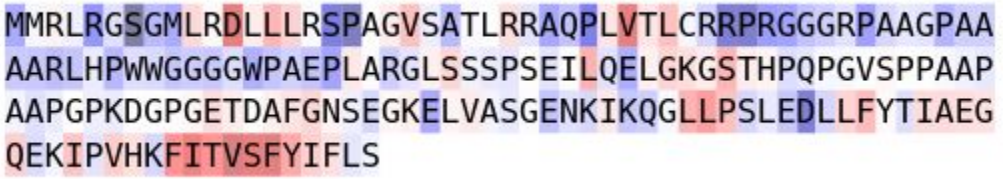
# Explainable ML

MKLNVDGLLVYFPYDYIYPEQFSYMLELKRTLDAKGHGVLEMPSGTGKTV
SLLALIVAYQRAFPLEVTKLIYCSRTVPEIEKVIEELRKLLSFYEQQEGE
KLPFLGLALSSRKNLCIHPEVTPLRFGKDVDGKCHSLTASYVRAQYQQDA
SLPHCRFYEEFDAHGRQVPLPAGIYNLDDLKALGQRQGWCPYFLARYSIL
HANVVVYSYHYLLDPKIADLVSKELARKAVVVFDEAHNIDNVCIDSMSVN
LTRRTLDRCQSNLDTLQKTVLRIKETDEQRLRDEYRRLVEGLREASAARE
TDAHLANPVLPDEVLQEAVPGSIRTAEHFLGFLRRLLEYVKWRLRVQHVV
QESPPAFLSGLAQRVCIQRKPLRFCAERLRSLLHTLEIADLADFSPLTLL
ANFATLVSTYAKGFTIIIEPFDDRTPTIANPILHFSCMDASLAIKPVFER
FQSVIITSGTLSPLDIYPKILDFHPVTMATFTMTLARVCLCPMIIGRGND
QVAISSKFETREDIAVIRNYGNLLLEMSAVVPDGIVAFFTSYQYMESTVA
SWYEQGILENIQRNKLLFIETQDGAETSVALEKYQEACENGRGAILLSVA
RGKVSEGIDFVHHYGRAVIMFGVPYVYTQSRILKARLEYLRDQFQIREND
FLTFDAMRHAAQCVGRAIRGKTDYGLMVFADKRFARADKRGKLPRWIQEH
LTDSNLNLTVDEGVQVAKYFLRQMAQPFHREDQLGLSLLSLEQLQSEETL
RRVEQIAQQL

# Explainable ML



# Authors' notes

Problem-specific architectures are less important than training

Redundant sequences are important

Bidirectional context is important

Leveraging large amounts of unlabeled data shows promising results

# Thank you for your attention!

MUNI
FI

# UDSMProt: universal deep sequence models for protein classification

Tomáš Pavlík