Using deep learning to annotate the protein universe

Irina Ponamareva

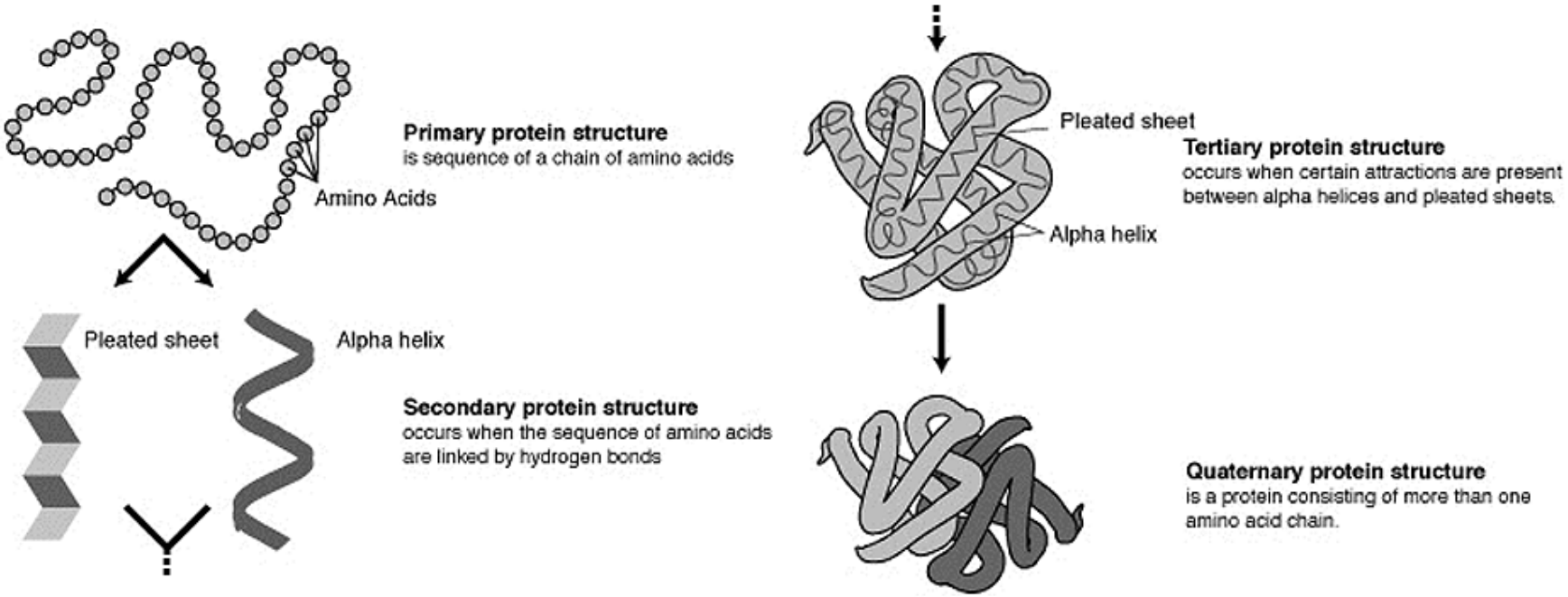
7 September 2022

Problem overview

- We have A LOT of data for protein sequences
- There is a need to assign a function to a protein sequence
- Can we do it using deep learning?

Protein structure

- Primary
- Secondary
- Tetriary
- Quaternary



Proteins and protein domains

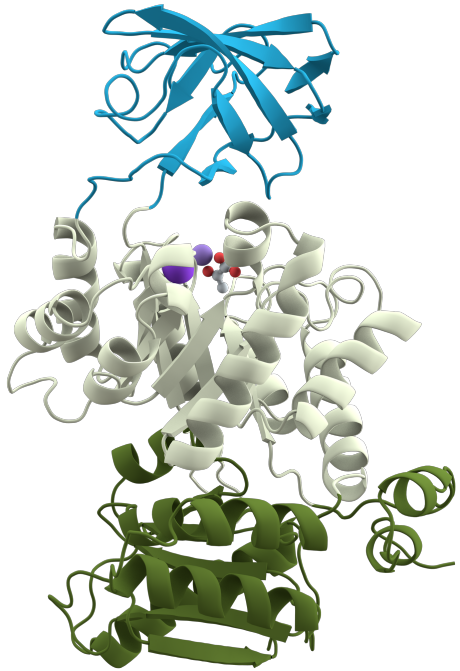
What are protein domains?

- "Building blocks"of proteins: structurally, functionally conserved parts of a sequence

Sequence homology

- Sequences are homologous if they share ancestry in the evolutional development, i.e. 'similarity'

Domain family: group of homologous domains



Proteins and protein domains

Domains lie in the protein sequence:

Domain family can be represented by alignment:

Seed sequence alignment for PF13841

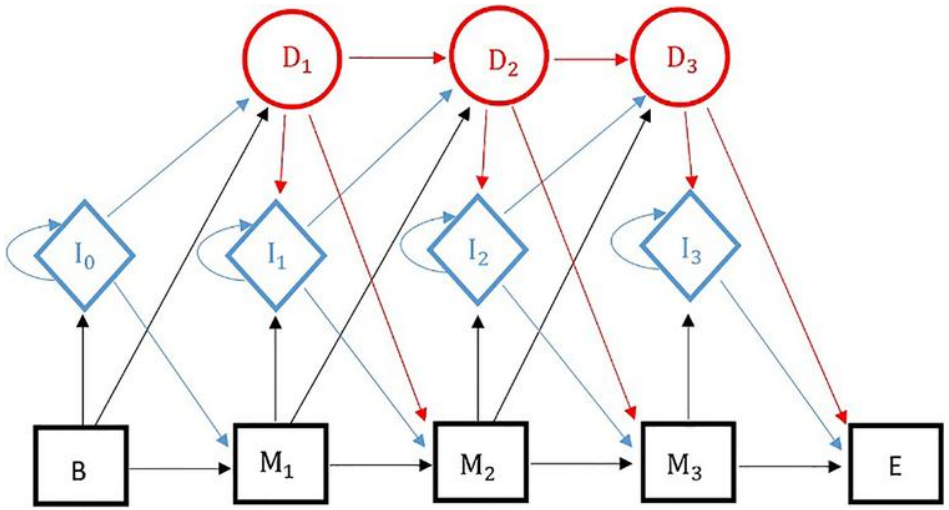
```
GLL12_CHICK/24-55        SCNHD..RGLCRVGNCNPGEYLAKY....C.FEPVILC.C
D131A_PANTR/28-58        ECPSE.YYYHCRL.KCNADEHAIRY....C..ADFSIC.C
Q30KS3_CANLF/28-57       TCSLE..YLNCRM.KCNLDEHAIKY....C..ADWTIC.C
G1LI52_AILME/28-57       TCSLE..GYTCRM.KCNADEHAIRY....C..TDWTIC.C
DFB43_RAT/28-57          ECFSE..YRHCRM.KCKANEYAIRY....C..ADWTIC.C
DFB43_MOUSE/28-56        DCSKH...RHCRM.KCKANEYAVRY....C..EDWTIC.C
DFB15_RAT/25-54          KCSRI..NGRCTE.SCLKNEELIAL....C..QKNLKC.C
D106A_PANTR/25-54        KCNKL..KGTCKN.NCGKNEELIAL....C..QKSLKC.C
A7S075_NEMVE/43-73       QCSEQ.FHGECMK.SCGDLPQMVGL....C..PSGTIC.C
Q1RLJ7_PIG/29-59         RCSKM..YGQCRT.RCYKIEKQIDI....C.YSPSKIC.C
DB114_HUMAN/28-58        RCTKR..YGRCKR.DCLESEKQIDI....C.SLPRKIC.C
I3L8Z9_PIG/21-50         KCWSA..LGRCRT.TCQESEVFHIL....C..RDATMC.C
A7LMA0_BOVIN/26-55       RCWNG..QGACRA.YCTKYEAYMHL....C..SDATLC.C
DFB25_MOUSE/26-55        RCWNG..QGACRT.FCTRQETFMHL....C..PDASLC.C
DB135_PANTR/36-66        SCWRL..QGTCRP.KCLKNETISYF....G..VILYICGC
DB135_HUMAN/36-65        SCWRL..QGTCRP.KCLKNEQYRIL....C..DTIHLC.C
G1LI49_AILME/36-65       TCWRT..KGVCKK.SCGKSEIYHIF....C..DSAHLC.C
```

Pfam – a database of proteins and protein domains. Current release (v35): 19,632 families

Existing approaches for function annotation

Main idea: searching for similar annotated sequences

- Profile Hidden Markov Model (pHMM) based — build a probabilistic model of the sequence. **This is a backbone of Pfam**



- Position-specific scoring matrix based

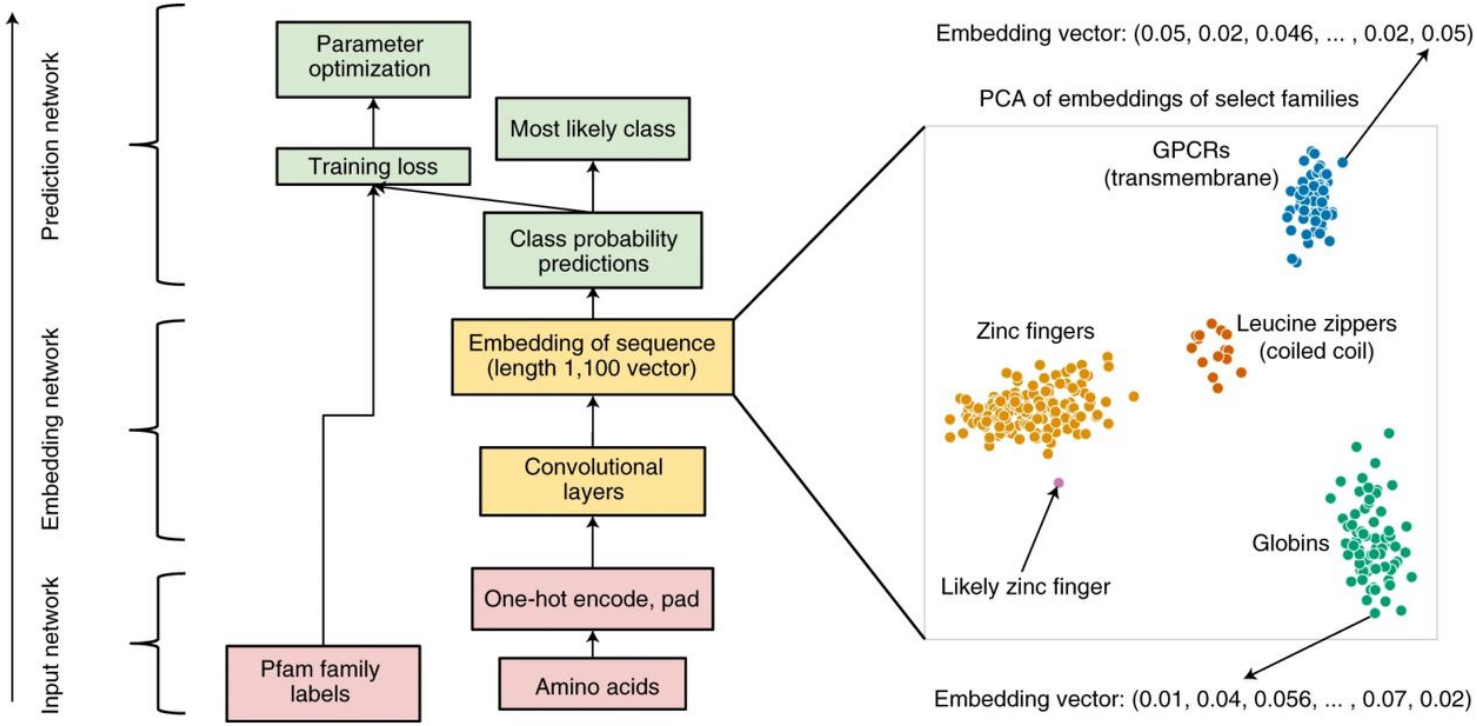
Why pHMMs are not enough?

- Some domain families might be too complicated for a single model
- Clans: groups of related families; other tools for determining family similarity
- A lot of manual curation for each model!

ProtCNN and ProtENN

- ProtCNN: convolutional model for protein sequences
- ProtENN: ensemble, majority voting
- Maxwell L. Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter,D. Sculley, Mark A. DePristo, and Lucy J. Colwell. Using deep learning to annotatethe protein universe, 2019.
- Dataset: protein sequences with domains, each contains one domain. Reminder: $\sim$ 19.000 classes

Model architecture



- Input: Amino acid sequence (20-letter alphabet)
- Output: class label ($\sim$ 19,000 classes)

Results: random split and clustered split

Table 1 | Model performance on the random split of Pfam-seed

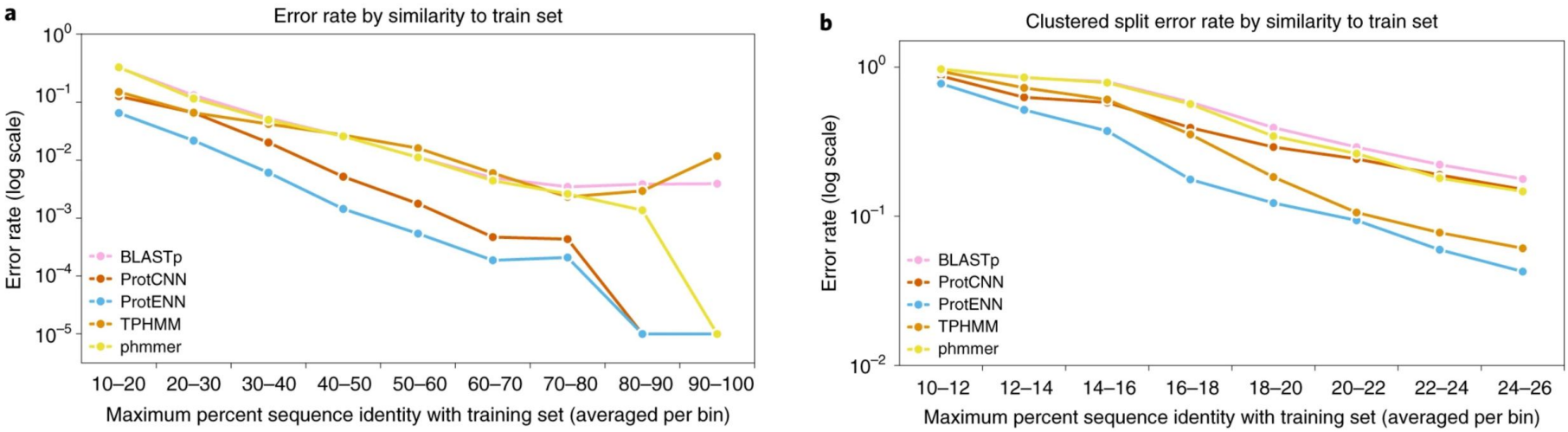
Model	Error rate	Number of errors
TPHMM	1.414%	1,784
phmmer	1.531%	1,932
BLASTp	1.654%	2,087
k-mer	9.994%	12,610
ProtCNN	0.495%	625
ProtENN	**0.162%**	**205**

Table 2 | Model performance on the clustered split of Pfam-seed

Model	Error rate	Number of errors
Top Pick HMM	18.1%	3,844
phmmer	32.6%	6,942
BLASTp	35.9%	7,639
ProtCNN	27.6%	5,882
ProtENN	**12.2%**	**2,590**

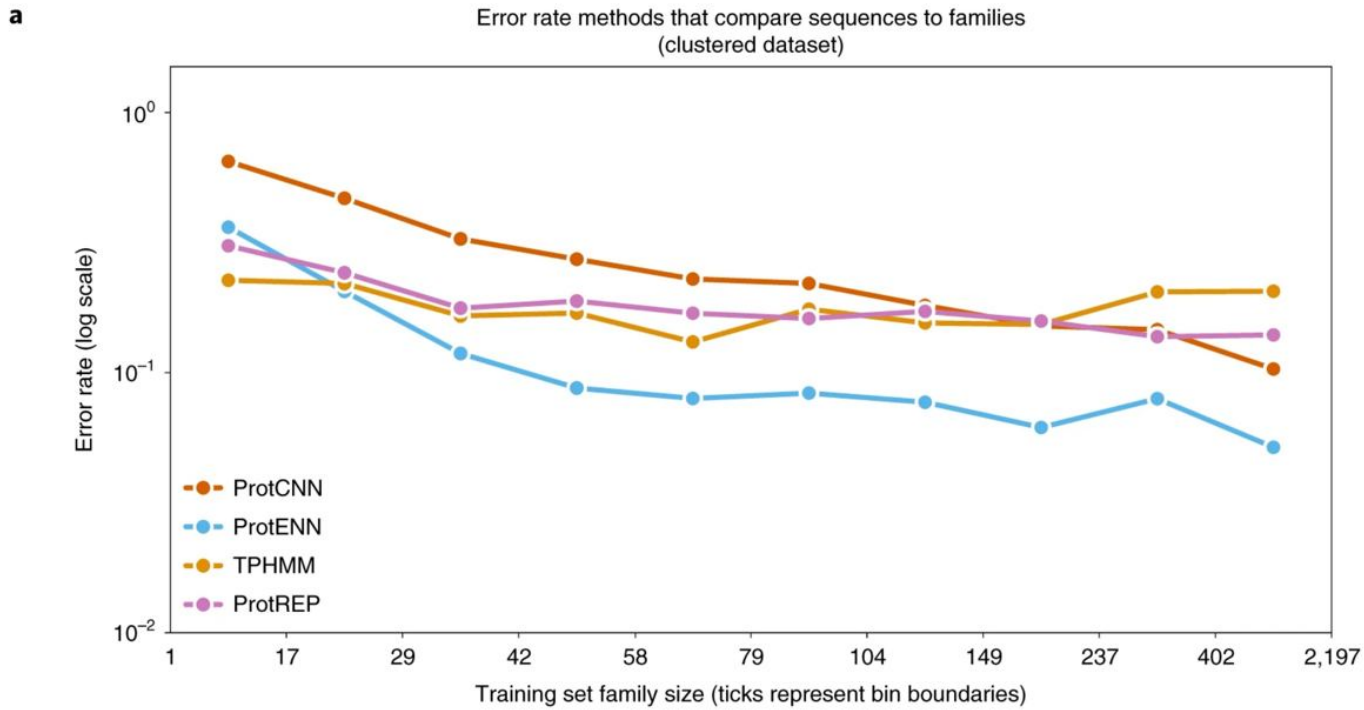
- Test size: 126,171 (random split), 21,293 (clustered split)
- Outperforming state-of-art pHMM and PSSM methods

Results by sequence similarity



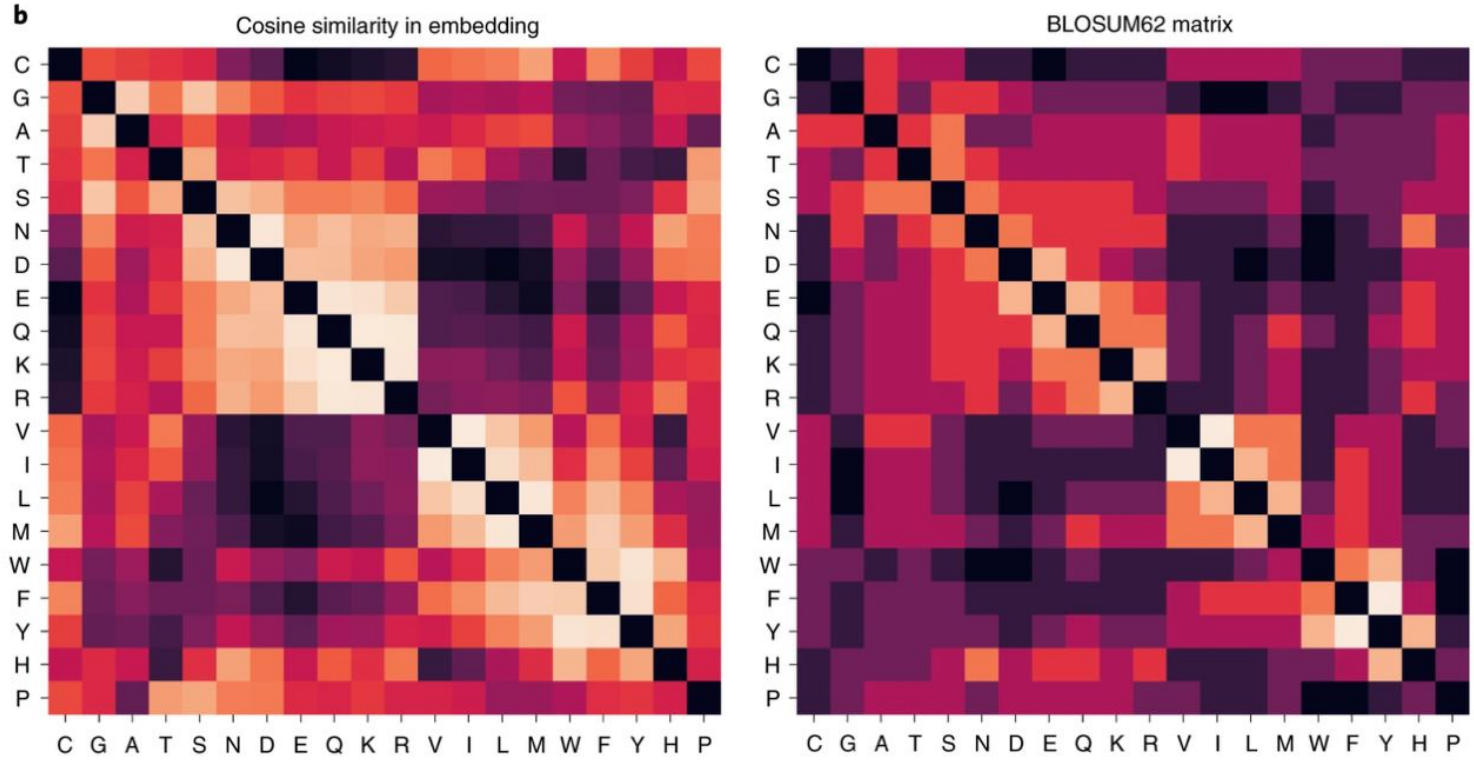
- Random split: ProtENN and ProtCNN make significantly fewer errors than alignment-based methods for sequences with sequence identity less than 90%. ProtENN outperforms these methods even at the lowest (10-20%) similarity rate
- For the clustered split, where all sequence identities are 25%, ProtENN is significantly more accurate for all bins: expanding the coverage!

Results: smallest families



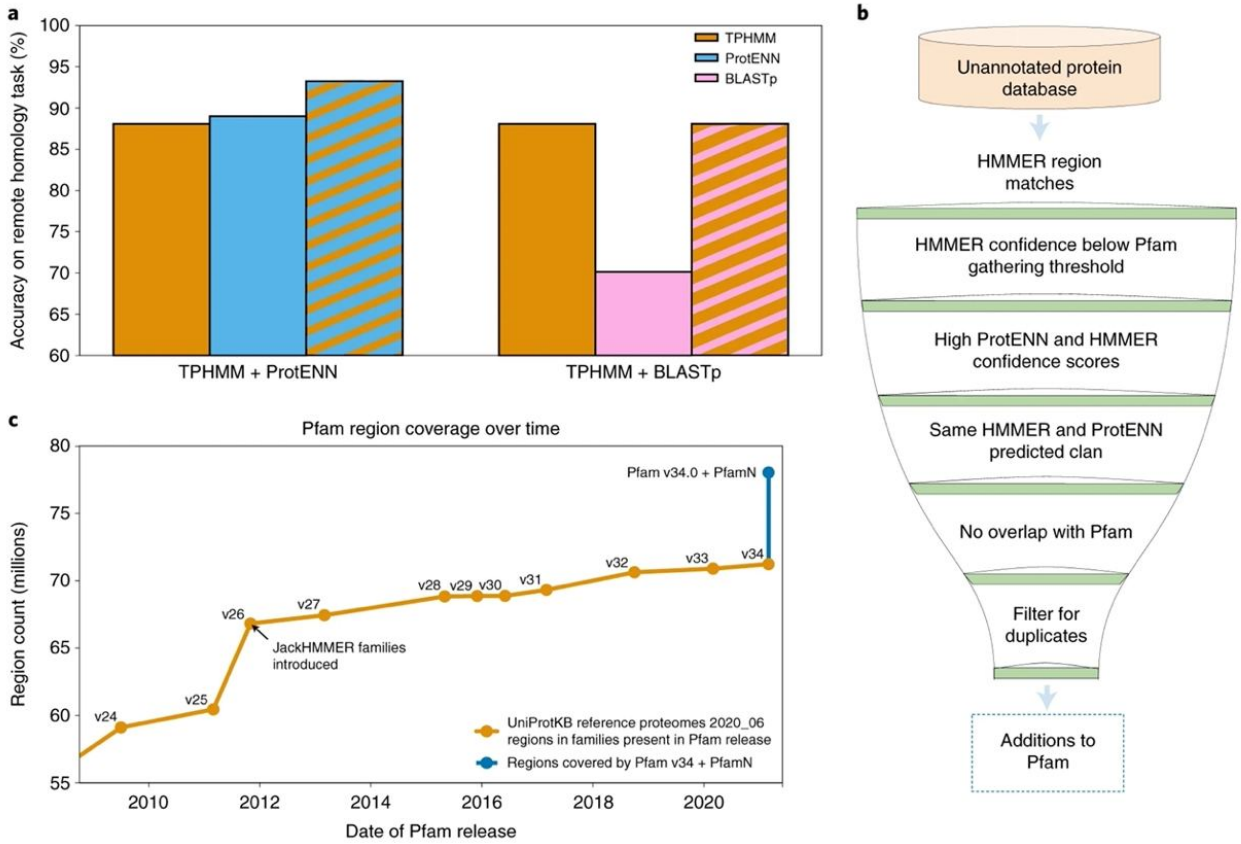
- We adapt one-shot approach and use the nearest neighbor for prediction (ProtREP)
- For the smallest families, same computational cost as ProtCNN provides higher accuracy

Results: learned embedding



- The amino acid embedding learned by ProtCNN from unaligned sequence data reflects the overall structure of the BLOSUM62 matrix

Results: remote homology task



- We can learn information complementary to pHMMs (clan level)

What else

- The dataset is too simple: we need to **detect** domains
- Per-residue predictions: for multi-domain proteins, nested domains...
- Novel family discovery
- Clustering: defining relationships between different families