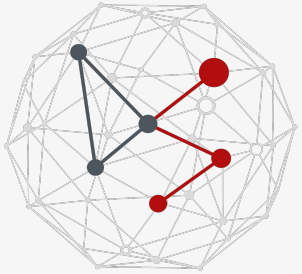
DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

ProtTrans embedding

Master of Science in Data Science

Damiano Piovesan

Protein Language Models Are Everywhere!

DIPARTIMENTO MATEMATICA

Large language models generate functional protein sequences across diverse families

Ali Madani, Ben Krause, Eric R. Gr
Luis Olmos Jr., Caiming Xiong, Zacha

Nature Biotechnology **41**, 1099–110

96k Accesses | **243** Citations | **11**

Designing proteins with language models

Jeffrey A. Ruffolo & Ali Madani

Nature Biotechnology **42**, 200–202 (2024) | Cite this article

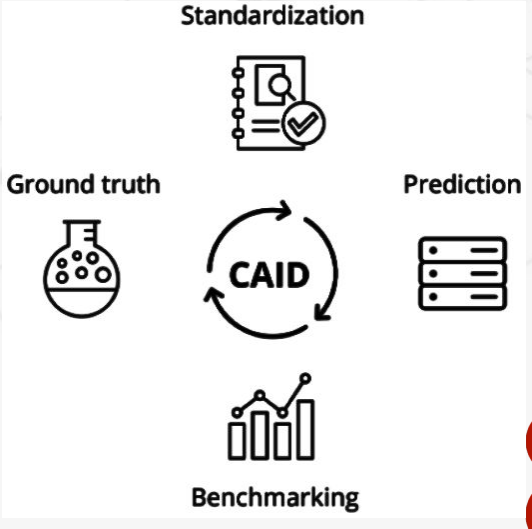
14k Accesses | **8** Citations | **100** Altmetric | Metrics

Google DeepMind

Isomorphic Laboratories

Meta

ROSTLAB.



Predicting Protein Properties

DIPARTIMENTO MATEMATICA

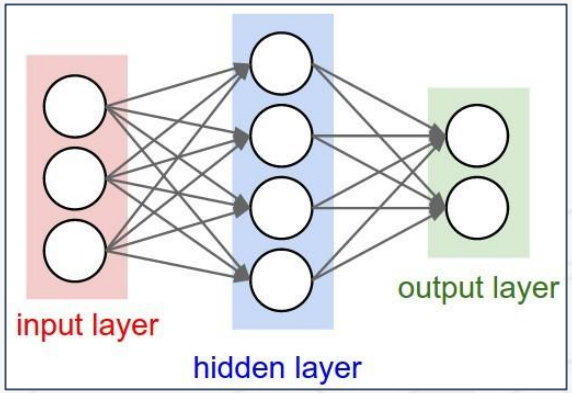
How do we represent protein sequence?

…VTMMGRFLQ…

One Hot encoding ⇒ Too simple

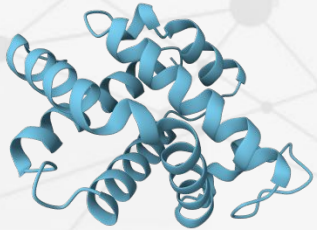
MSA ⇒ Not available for all proteins

Neural Network



RNN, CNN, FNN, LSTM,
k-nn, SVM, …

Prediction Task



Structure Prediction
Function Prediction
Mutation effect predictions
…

DIPARTIMENTO MATEMATICA

Protein Language Models (pLMs)

- Protein sequences are constrained to adopt particular 3D structures optimized for accomplishing particular functions
- These constraints mirror the rules of grammar and meaning in Natural Language Processing (NLP)
- Entire **protein sequence** as a **sentence**
- **Amino acids** as single **words**

Natural Language vs. Protein Sequence

DIPARTIMENTO MATEMATICA

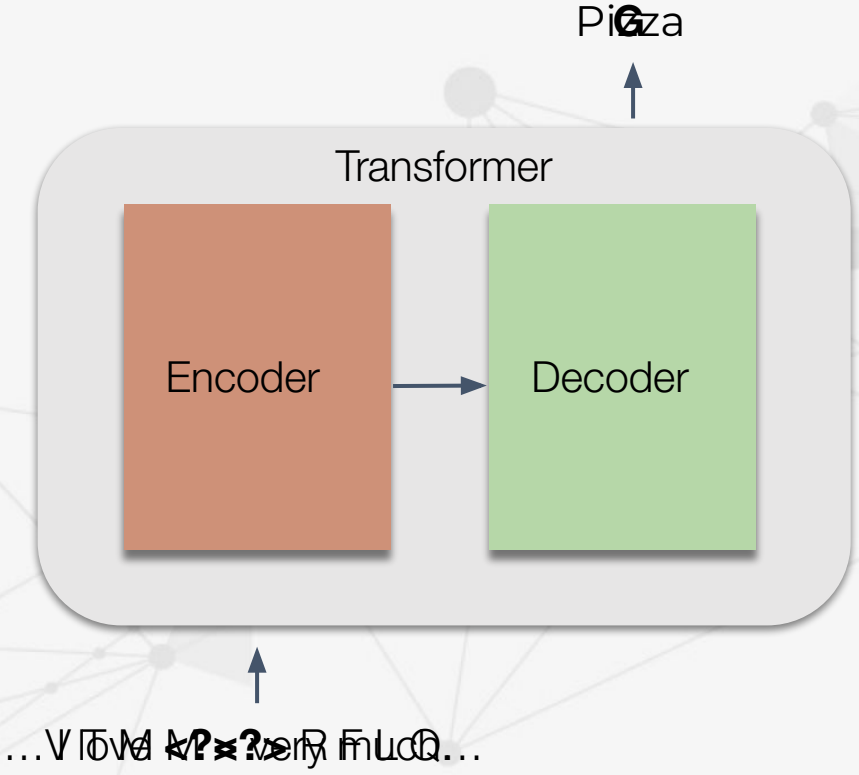
English Language	Protein Sequence
Alphabet : { A , B , C, D, … , Z}	**Amino Acids:** { M, P, Q, R, … , Y}
Words: Today, Weather, Rainy, …	**Domains/Motifs:** IPR042577, Triple Helix, Parallel Beta-helix, …
Sentence: Today the weather is gloomy.	**Sequence:** MPFNGTHNKFKLNYKPEEEY…

Protein Language Models

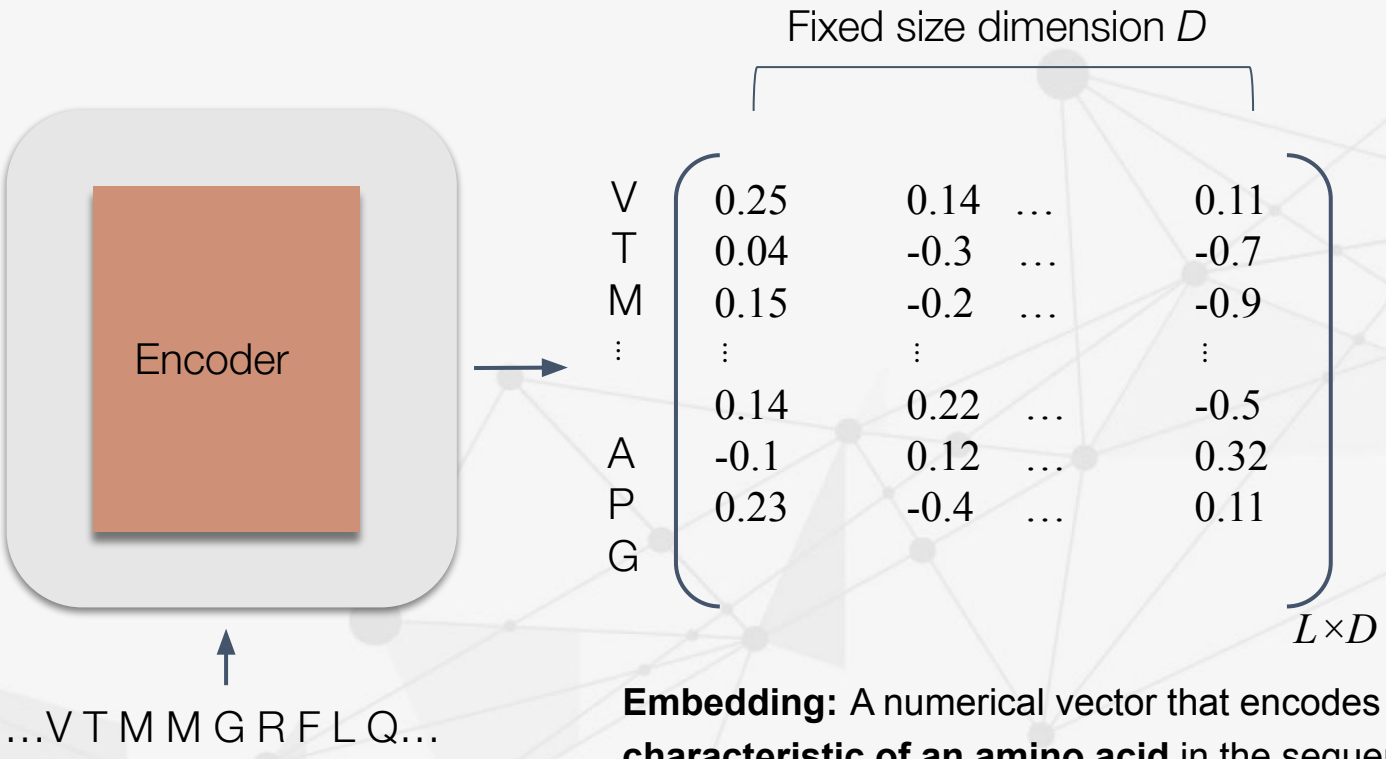
- **Language Model** generates text based on patterns learned from large datasets of human language.

- **Protein Language Model learns the patterns of protein sequence.**

Self-Supervised Learning



Protein Language Models



Embedding: A numerical vector that encodes the **characteristic of an amino acid** in the sequence

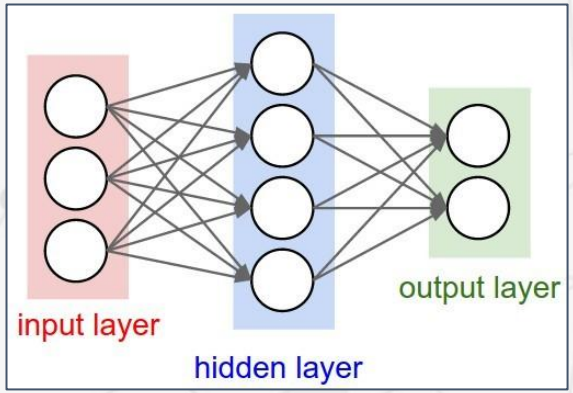
Predicting Protein Properties

DIPARTIMENTO MATEMATICA

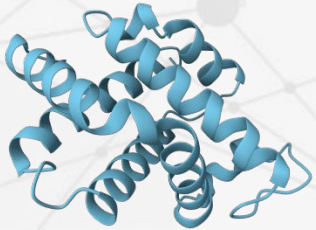
How do we represent protein sequence?

…VTMMGRFLQ…

Neural Network



Prediction Task



Structure Prediction
Function Prediction
Mutation effect predictions

Predicting Protein Properties

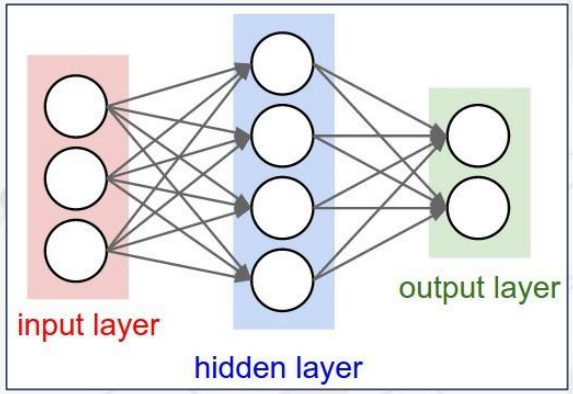
DIPARTIMENTO MATEMATICA

Embeddings

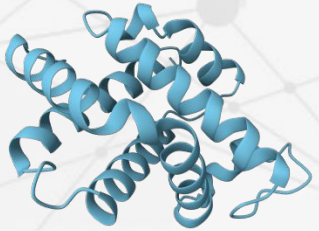
V	0.25	0.14	…	0.11
T	0.04	-0.3	…	-0.7
M	0.15	-0.2	…	-0.9
⋮	⋮	⋮		⋮
A	0.14	0.22	…	-0.5
P	-0.1	0.12	…	0.32
G	0.23	-0.4	…	0.11

$L \times D$

Neural Network



Prediction Task



Structure Prediction
Function Prediction
Mutation effect predictions

Transfer Learning

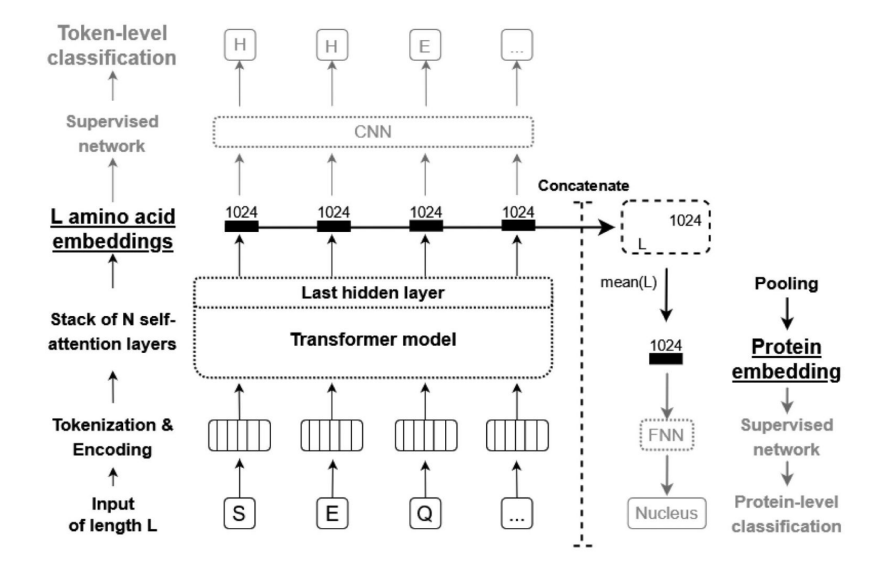
Transfer-learning

- Limitation in annotations do not constrain LMs
- Self-supervised training exclusively relies on the sequential order of input
- After training we can extract embeddings
- Transfer-learning refers to the idea of using embeddings as input for other trained supervised models
- Computationally demanding LM pre-training followed by a less demanding inference

DIPARTIMENTO MATEMATICA

Transfer-learning

- pLMs are trained to predict masked amino acids in already known sequences
- Embedding can be used as input for supervised training of per-residue/per-token and pre-protein/pooling prediction tasks



Per-protein pooling

Pooling strategies

- Min, max, mean, concatenation
- Mean works better (10%)

Data

- Protein databases contain several orders of magnitude more tokens than corpora used in NLP
- Google's Billion Word data set contains 829m tokens
- Interpreting protein domains as words would just cut the number of tokens by a factor of 100

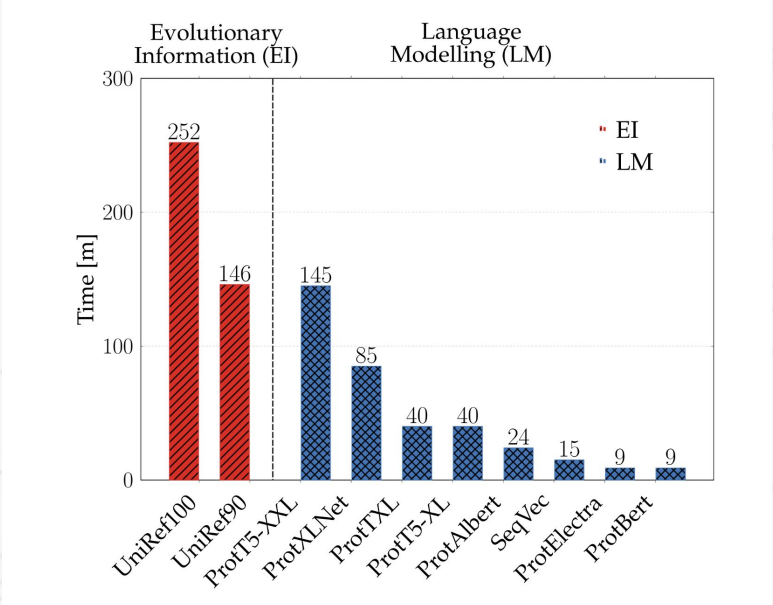
Data LM	*UniRef50*	*UniRef100*	*BFD*
Number proteins [in m]	45	216	2,122
Number of amino acids [in b]	14	88	393
Disk space [in GB]	26	150	572

Units: number of proteins in millions (m), of amino acids in billions (b), and of disk space in GB (uncompressed storage as text).

BFD, Big Fantastic Database

Fast prediction

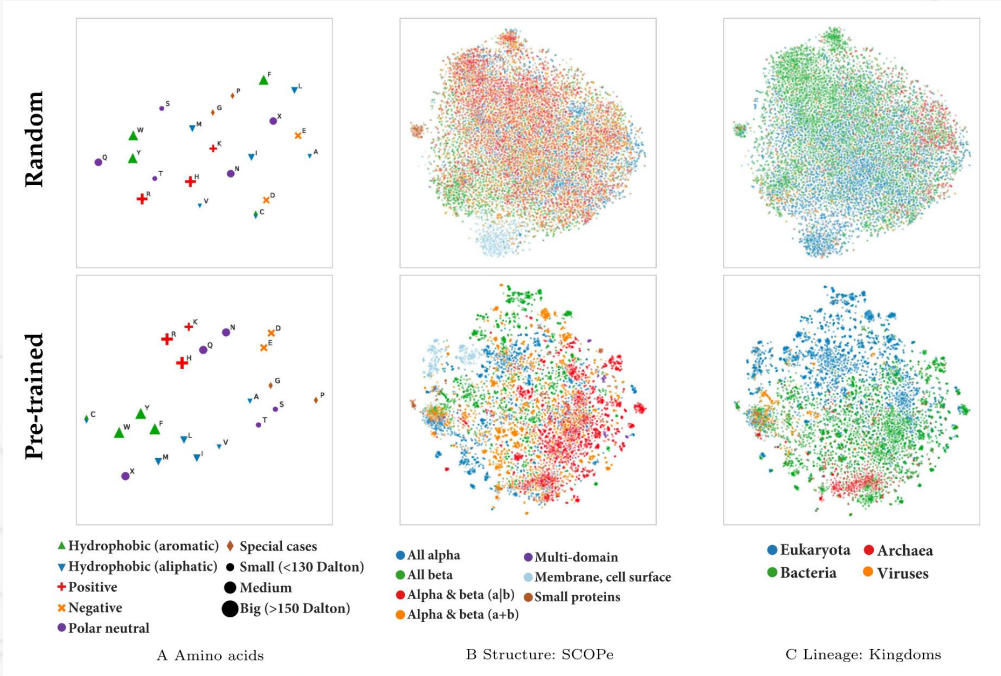
- Calculating evolutionary information takes at least 4-6 times longer to process the same input



DIPARTIMENTO MATEMATICA

Evaluation of embeddings

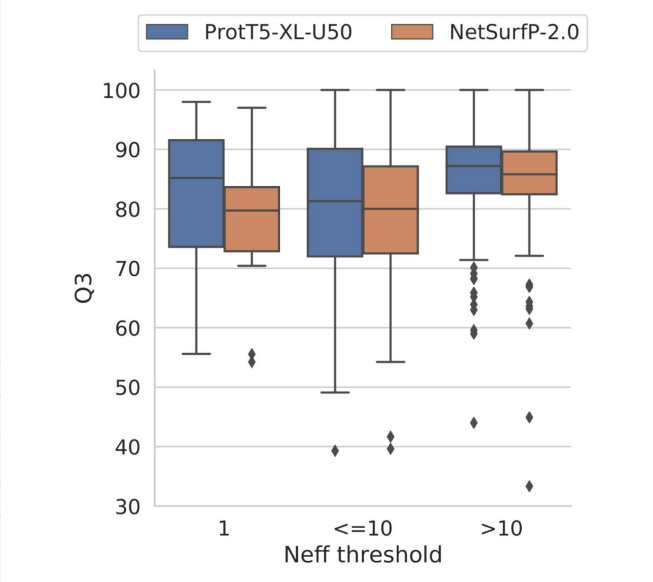
- Project the high-dimensional representations of embeddings down to two dimensions using **t-SNE** (t-distributed stochastic neighbor embedding)
- Compare with annotations (labels)
  - Structural class (SCOPe)
  - The three major domains of life (archaea, bacteria, eukarya) plus viruses
  - Biophysical features of amino acids



DIPARTIMENTO MATEMATICA

Per-residue secondary structure prediction

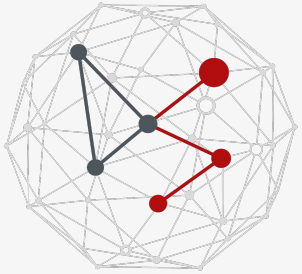
- Best performance training on BFD and refining on UniRef50
- Work well also for small families, those with less evolutionary information
- More pre-training steps (more samples), better than bigger models
- Larger models absorb information faster, but they see fewer samples in the same amount of computing power



Neff, Number of Effective Sequences. Number of proteins in MSA after clustering at 62% identity

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

FoldSeek

Master of Science in Data Science

Damiano Piovesan

Similarity search

- The most widely used approach to protein annotation and analysis is based on sequence similarity search (e.g. BLAST)
- However, many proteins cannot be annotated because detecting distant evolutionary relationships from sequences alone remain challenging
- Detecting similarity by 3D superposition offers higher sensitivity
- However, current tools are much too slow to cope with today's scale of structure databases
    - AlphaFoldDB, 214 million structures
    - ESM Atlas, 617 million metagenomic structures

DIPARTIMENTO MATEMATICA

Similarity search

- TM-align, 1 query Vs 100 million protein structures takes a month on 1 CPU core
- All-vs-all 10 millenia on 1,000-core cluster
- Sequence searching is 4-5 orders of magnitude faster
- All-vs-all of 100 million sequences would take 1 week on the same cluster with MMseqs2

DIPARTIMENTO MATEMATICA

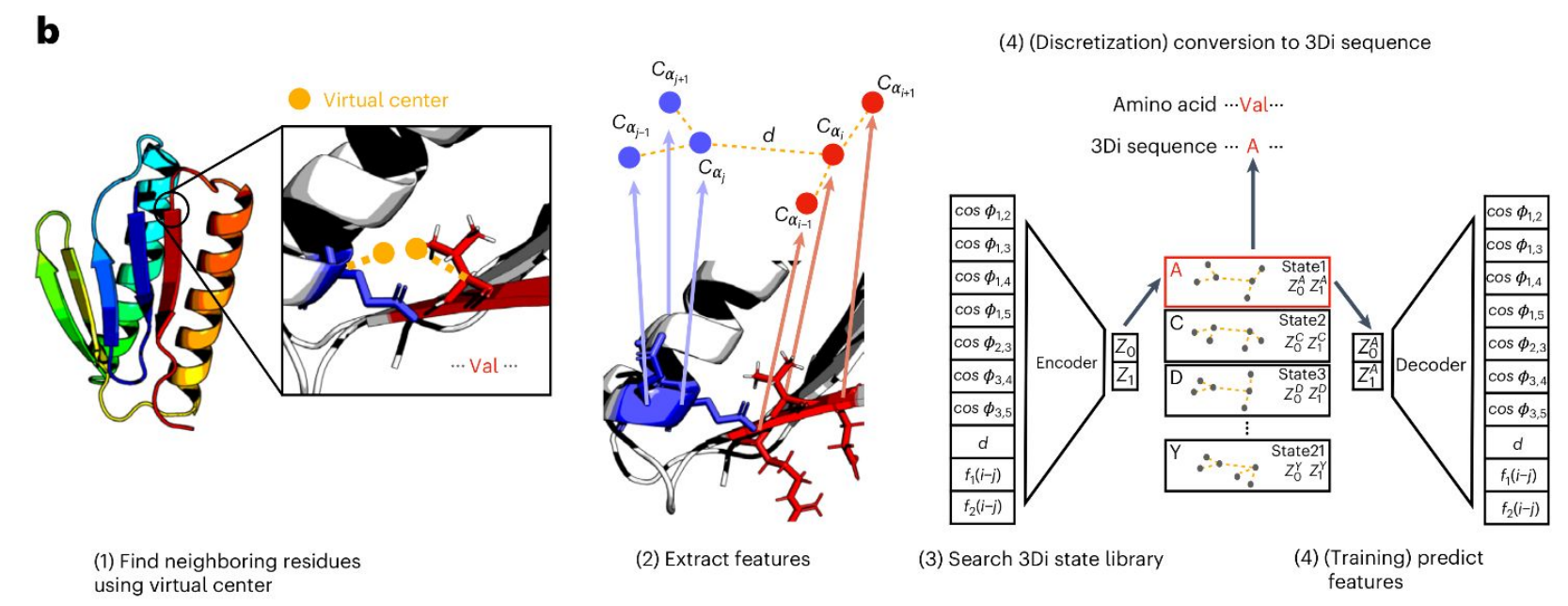
Why structural alignment is so slow?

- Sequence search employ fast and sensitive pre-filter algorithms (e.g. BLAST k-mer)
- Structural similarity scores are non-local. Changing the alignment in one part affects the similarity in all other parts

Structure as sequence

- Describe the amino acid backbone of proteins as sequences over a structural alphabet
- Compare structures using sequence alignments
- FoldSeek does not describe the backbone but, rather, tertiary interactions

Learning the 3Di alphabet

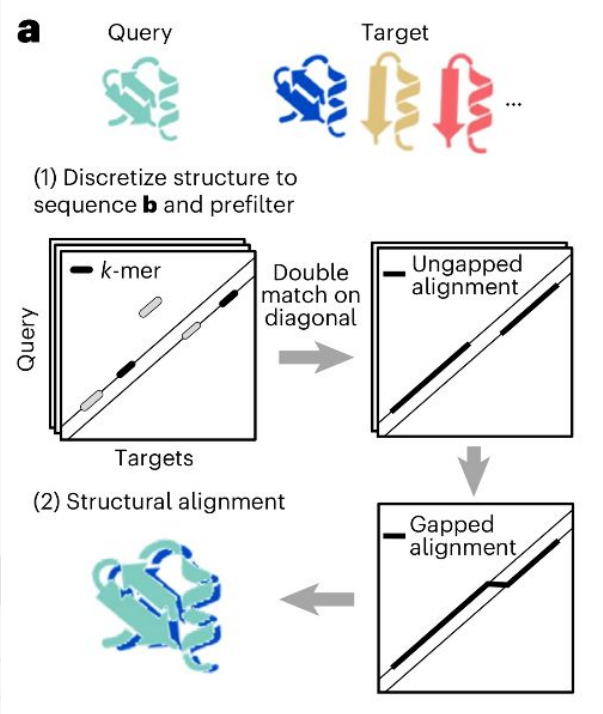


3Di alphabet

20 states of 3D interactions alphabet (3Di) that describes for each residue i the geometric conformation with its **spatially closest residue j**

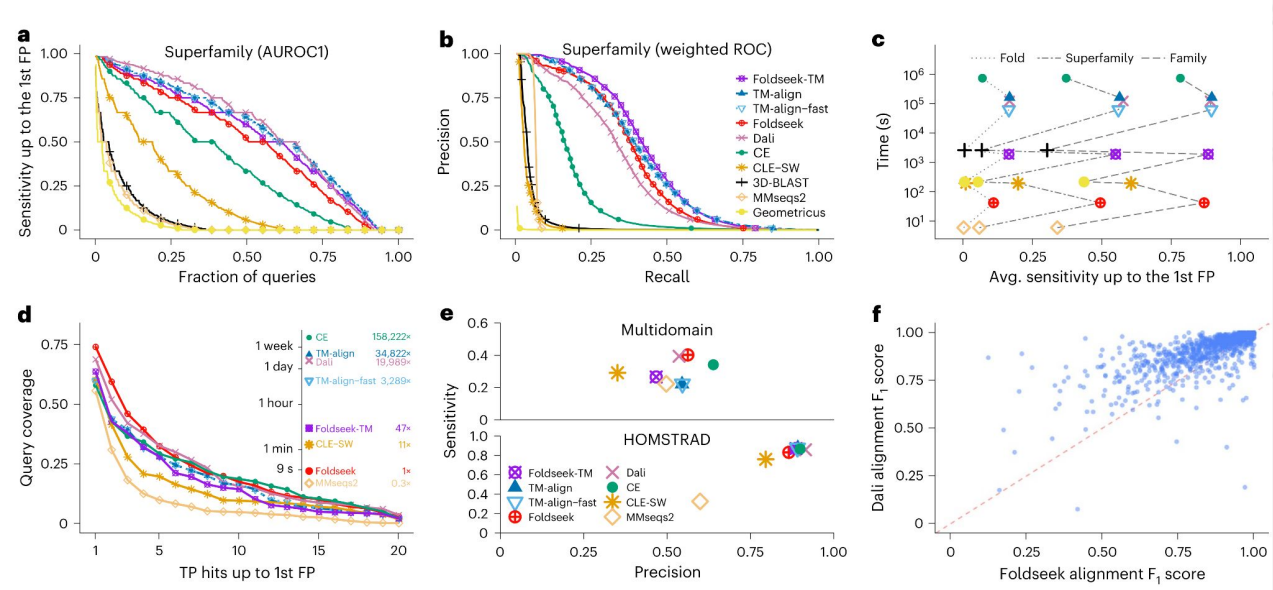
- Weaker dependency between consecutive letters
- More evenly distributed state frequencies (more information densities, less FPs)
- The highest information density is encoded in protein cores and the lowest in non-conserved coil/loop regions, whereas the opposite is true for backbone structural alphabets

Searching

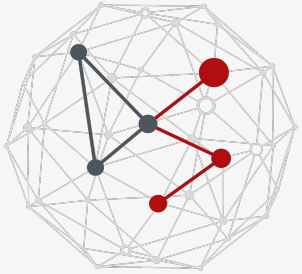

a
Query
Target
...
(1) Discretize structure to sequence b and prefilter
k-mer
Query
Targets
Double match on diagonal
Ungapped alignment
Gapped alignment
(2) Structural alignment

Performance

- 4k-20k times faster than structural aligners when evaluated against SCOPe
- 20K-180K times faster when compared on AlphaFoldDB



1222·2022 800 ANNI
UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

FoldComp

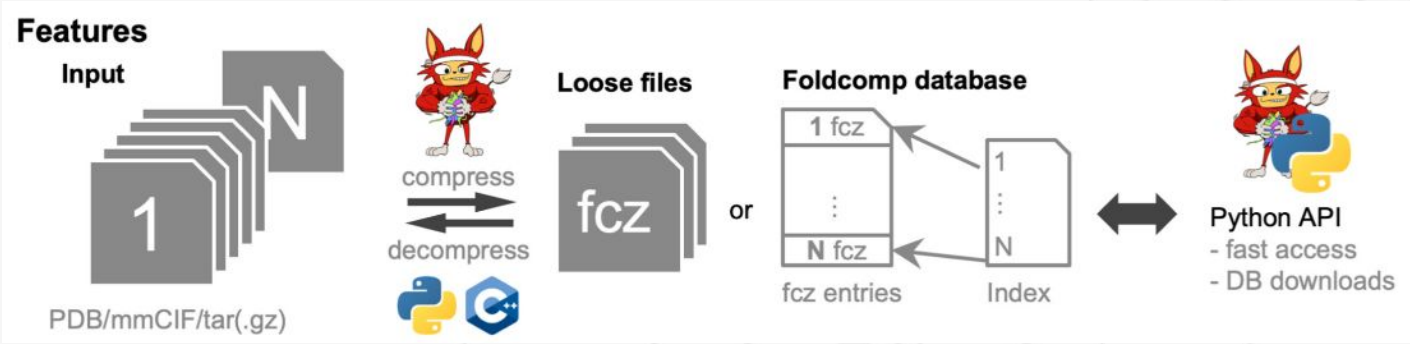
Master of Science in Data Science

Damiano Piovesan

Storage

- AlphaFoldDB 25 TB
- ESM Atlas 15 TB
- PDB / mmCIF, 80-byte for each atom
- Compressed formats
  - Gzip
  - BinaryCIF
  - MMTF, transform 3D into 2D and use PNG-image compression

FoldComp


Features
Input
N
1
PDB/mmCIF/tar(.gz)
compress
decompress
Loose files
fcz
or
Foldcomp database
1 fcz
N fcz
fcz entries
1
N
Index
Python API
- fast access
- DB downloads

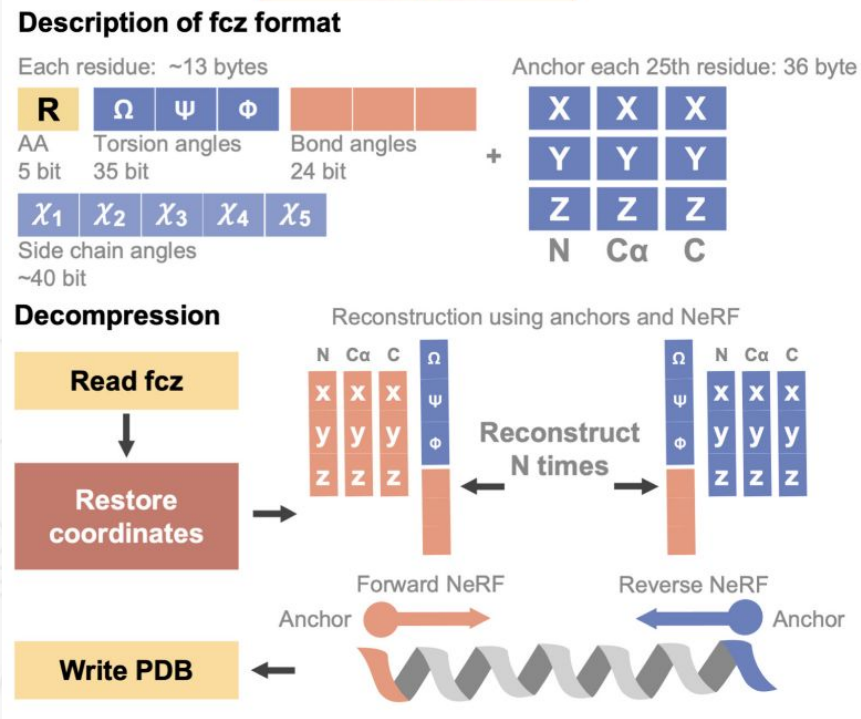
Compression & decompression

Compression

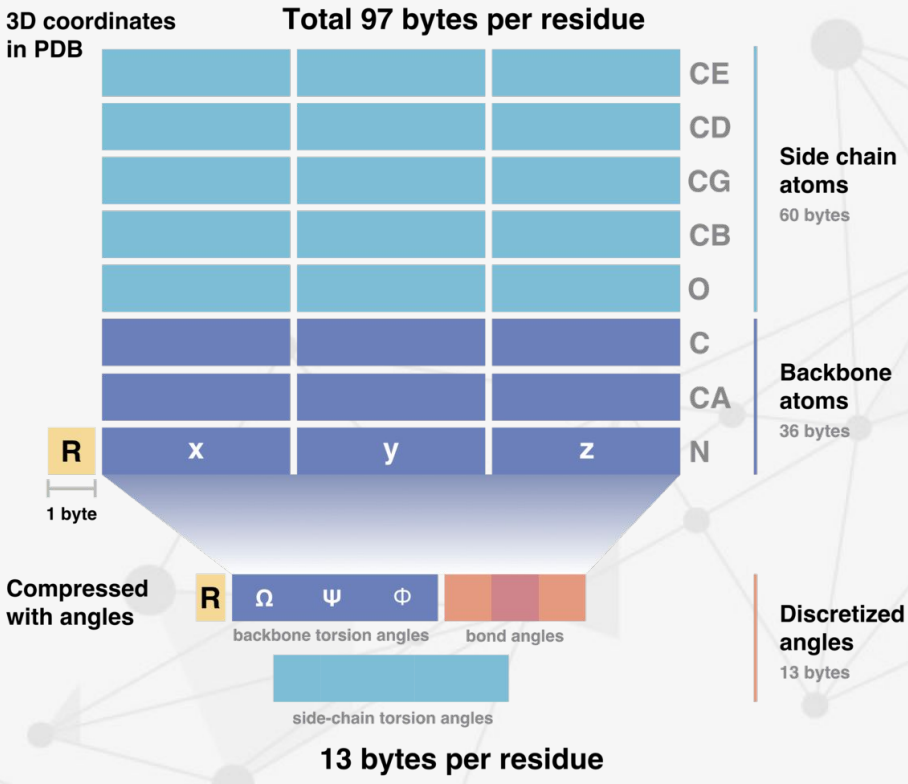
- Store internal coordinates, backbone torsions, bond angles, and an anchor every 25 residues (N, C, C-alpha coordinates)

Decompression

- Extend coordinates from N- to C-terminal and then form C- to N- using **Natural Extension Reference Frames (NeRF)**
- Average the coordinates in between



Encoding


3D coordinates in PDB
Total 97 bytes per residue
CE
CD
CG
CB
O
C
CA
R
x
y
z
N
1 byte
Side chain atoms
60 bytes
Backbone atoms
36 bytes
Compressed with angles
R
Ω
Ψ
Φ
backbone torsion angles
bond angles
side-chain torsion angles
Discretized angles
13 bytes
13 bytes per residue

Performance