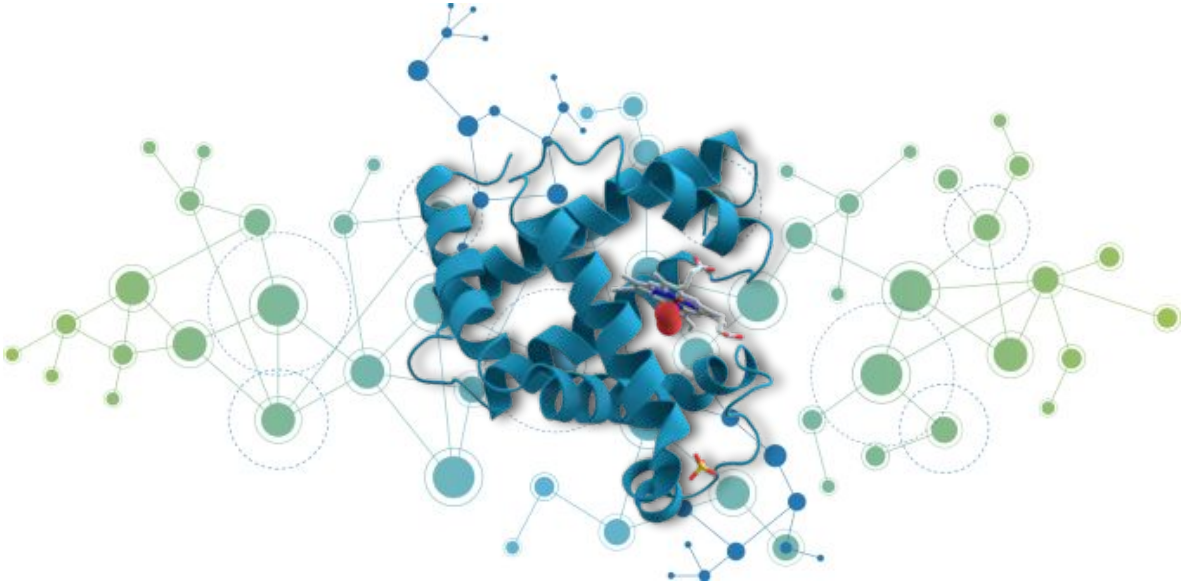
A Brief Talk on the Importance of

Natural Language Processing in Biomedicine



Simone Conia

Reading Group @ Sapienza NLP

September 16th, 2020

SAPIENZA
UNIVERSITÀ DI ROMA

Language models are taking NLP to the next level

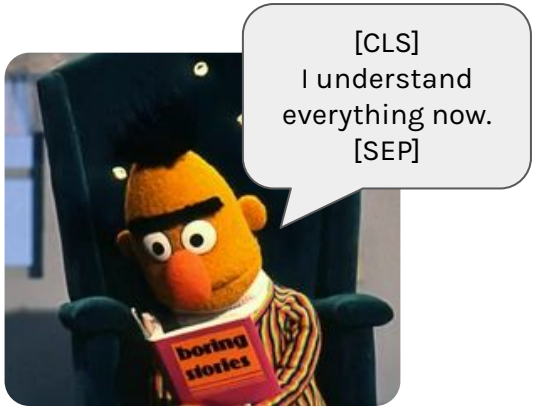
NLP is witnessing an unprecedented growth thanks to the expressiveness and wide availability of (pretrained) language models such as ELMo [1], BERT [2] and GPT [3].

[1] Deep contextualized word representations, Peters et al., 2018
[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al., 2018
[3] Improving Language Understanding by Generative Pre-Training, Radford et al., 2018

Language models are taking NLP to the next level

NLP is witnessing an unprecedented growth thanks to the expressiveness and wide availability of (pretrained) language models such as ELMo [1], BERT [2] and GPT [3].

Language models encode information that has been proven to be fundamental in a wide array of "language understanding" tasks such as question answering and natural language inference.



[1] Deep contextualized word representations, Peters et al., 2018
[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al., 2018
[3] Improving Language Understanding by Generative Pre-Training, Radford et al., 2018

Language models are here to stay

Research on language models is not slowing down!

Larger!
G-Shard [4] with 1 trillion parameters.

[4] GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, Lepikhin et al., 2020
[5] Reformer: The Efficient Transformer, Kitaev et al., 2020
[6] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et al., 2019
[7] REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., 2020

Language models are here to stay

Research on language models is not slowing down!

Larger!
G-Shard [4] with 1 trillion parameters.

More efficient!
Reformer [5] with LSH attention.

[4] GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, Lepikhin et al., 2020
[5] Reformer: The Efficient Transformer, Kitaev et al., 2020
[6] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et al., 2019
[7] REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., 2020

Language models are here to stay

Research on language models is not slowing down!

Larger!
G-Shard [4] with 1 trillion parameters.

More efficient!
Reformer [5] with LSH attention.

Multitask!
T5 [6] is trained on 5 “understanding” tasks.

[4] GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, Lepikhin et al., 2020
[5] Reformer: The Efficient Transformer, Kitaev et al., 2020
[6] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et al., 2019
[7] REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., 2020

Language models are here to stay

Research on language models is not slowing down!

Larger!
G-Shard [4] with 1 trillion parameters.

More efficient!
Reformer [5] with LSH attention.

Multitask!
T5 [6] is trained on 5 “understanding” tasks.

Knowledge!
REALM [7] exploits Wikipedia articles.

[4] GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding, Lepikhin et al., 2020
[5] Reformer: The Efficient Transformer, Kitaev et al., 2020
[6] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et al., 2019
[7] REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., 2020

NLP + Biomedicine

How can language models help research and researchers in Biomedicine?

NLP + Biomedicine

How can language models help research and researchers in Biomedicine?

Automatically find names of people, places, and organizations in text across many languages.

Text Mining
for biomedical documents

NLP + Biomedicine

How can language models help research and researchers in Biomedicine?

Automatically find names of people, places, and organizations in text across many languages.

Text Mining
for biomedical documents

Healthcare
for personalized treatments

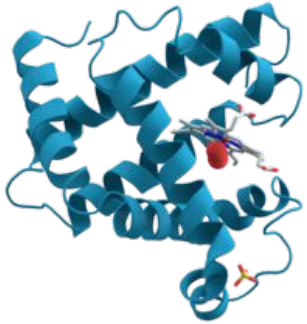
NLP + Biomedicine

How can language models help research and researchers in Biomedicine?

Automatically find names of people, places, and organizations in text across many languages.

Text Mining
for biomedical documents

Healthcare
for personalized treatments



Research
on molecules, diseases and drugs

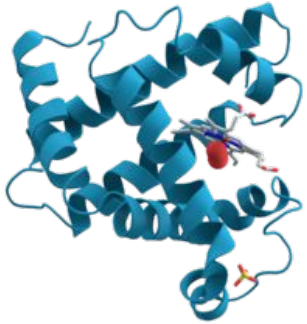
NLP + Biomedicine

How can language models help research and researchers in Biomedicine?

Automatically find names of people, places, and organizations in text across many languages.

Text Mining
for biomedical documents

Healthcare
for personalized treatments



Research
on molecules, diseases and drugs

from traditional to innovative applications!

Preliminaries: a step back to language modeling

A Language model simply tells us the probability of a sentence.

Preliminaries: a step back to language modeling

A Language model simply tells us the probability of a sentence.

On cat is table the the	LOW probability
The cat is on the table	HIGH probability

Preliminaries: a step back to language modeling

A Language model simply tells us the probability of a sentence.

On cat is table the the	LOW probability
The cat is on the table	HIGH probability

Therefore a language model must "understand" sentences.* Notice that syntax alone is often not enough.

The table is on the cat	LOW probability

* This is an over-simplification. See Bender and Koller (2020) on why current language models do not and cannot really understand sentences.

Preliminaries: a step back to language modeling

A Language model simply tells us the probability of a sentence.

On cat is table the the	LOW probability
The cat is on the table	HIGH probability

Therefore a language model must "understand" sentences.* Notice that syntax alone is often not enough.

The table is on the cat	LOW probability

Idea: do not build a new system for each task, **adapt an existing language model!**

* This is an over-simplification. See Bender and Koller (2020) on why current language models do not and cannot really understand sentences.

Preliminaries: what is BERT?

For this talk, we only need to know that BERT is a recently proposed (masked) language model that learns to guess missing words on a huge amount of text.

Preliminaries: what is BERT?

For this talk, we only need to know that BERT is a recently proposed (masked) language model that learns to guess missing words on a huge amount of text.

Correctly guessing words requires BERT to "**smurf**" from the context of a sentence.

As a consequence, BERT learns good word representations that can be **smurfed** in other tasks.

* The Smurfs often replace words with "smurf", but their sentences are always understandable from the context.

Preliminaries: what is BERT?

For this talk, we only need to know that BERT is a recently proposed (masked) language model that learns to guess missing words on a huge amount of text.

Correctly guessing words requires BERT to "**smurf**" from the context of a sentence.

As a consequence, BERT learns good word representations that can be **smurfed** in other tasks.

That's all we need to know for now about BERT!

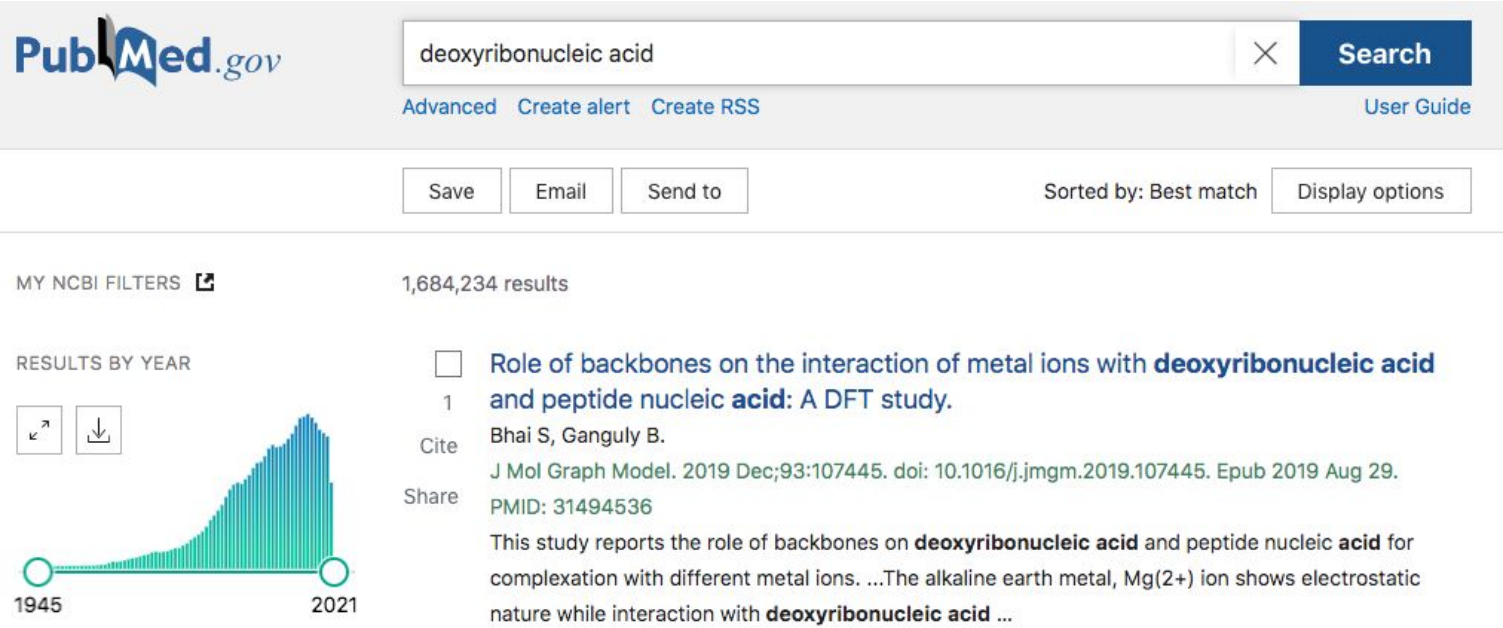
* The Smurfs often replace words with "smurf", but their sentences are always understandable from the context.

Text Analysis and Mining in Biomedicine

Text Analysis/Mining is the process of extracting meaningful/relevant information from unstructured text data (e.g. books, websites, reviews, articles, etc.).

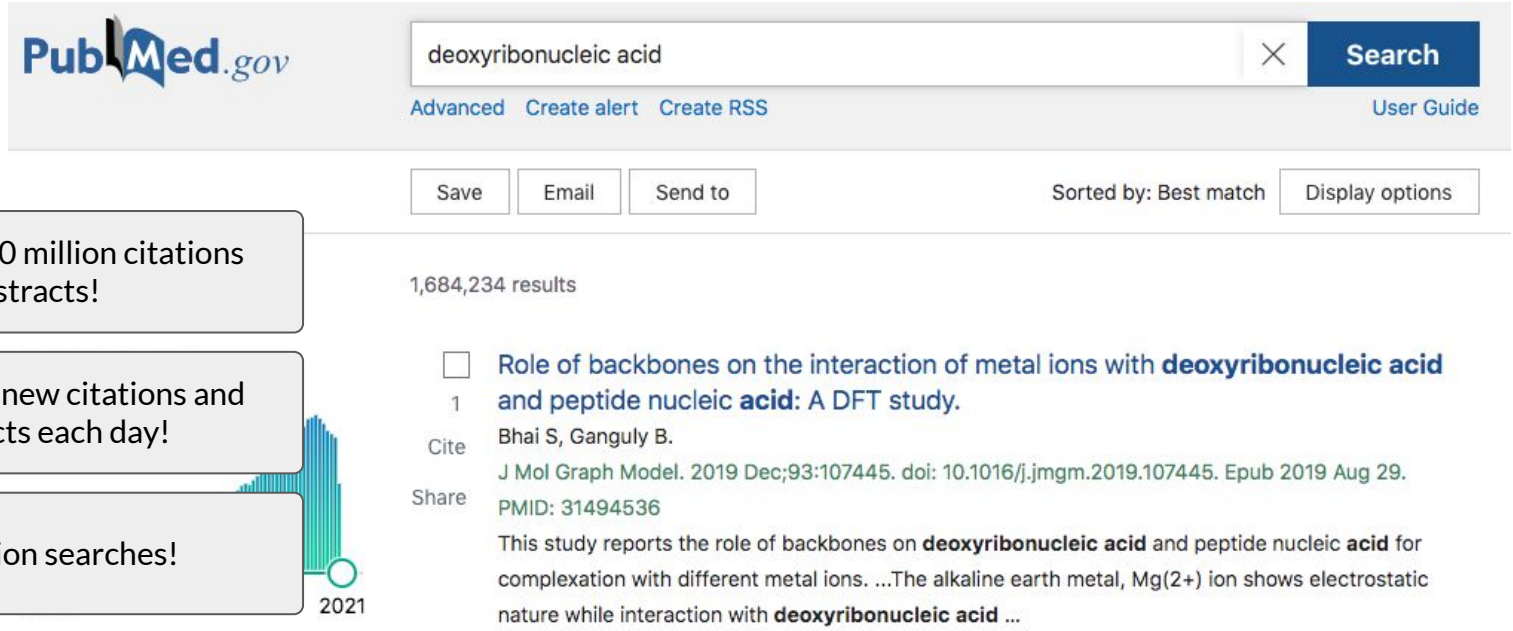
Text Analysis and Mining in Biomedicine

Text Analysis/Mining is the process of extracting meaningful/relevant information from unstructured text data (e.g. books, websites, reviews, articles, etc.).



Text Analysis and Mining in Biomedicine

Text Analysis/Mining is the process of extracting meaningful/relevant information from unstructured text data (e.g. books, websites, reviews, articles, etc.).



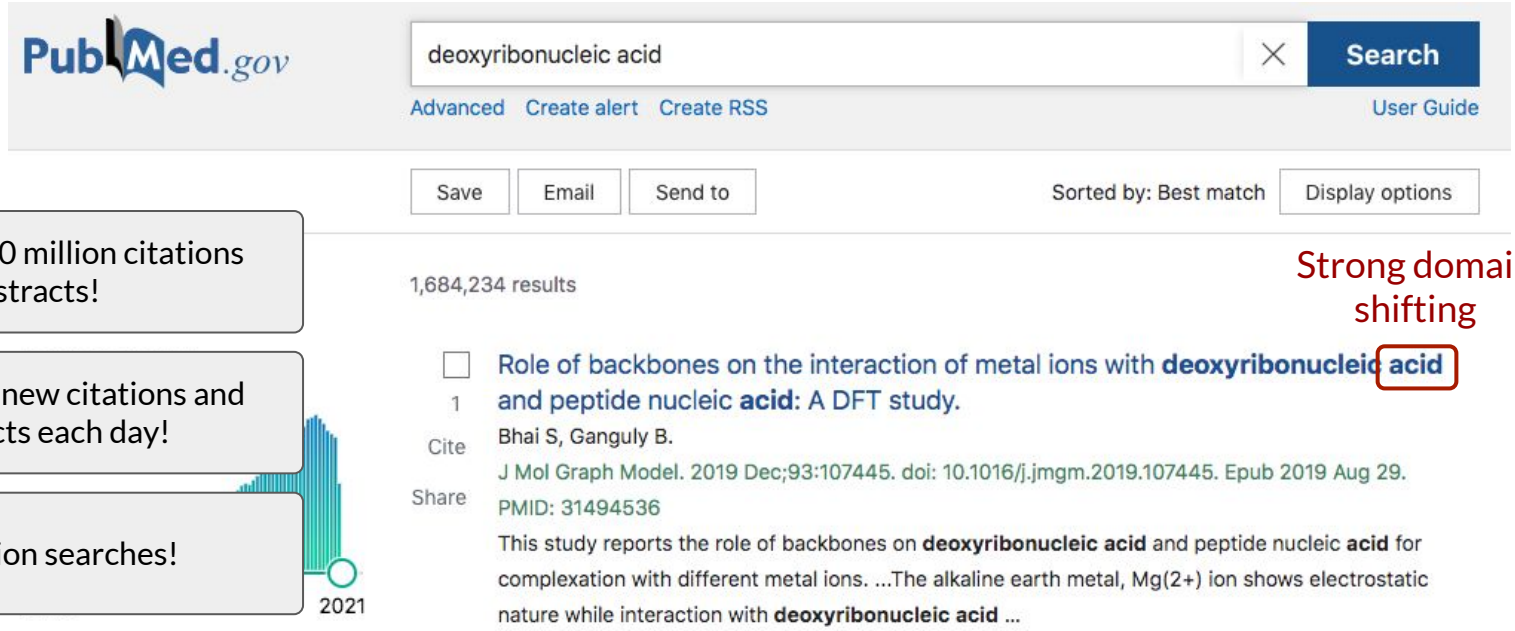
Over 30 million citations and abstracts!

3000+ new citations and abstracts each day!

3.3 billion searches!

Text Analysis and Mining in Biomedicine

Text Analysis/Mining is the process of extracting meaningful/relevant information from unstructured text data (e.g. books, websites, reviews, articles, etc.).



Over 30 million citations and abstracts!

3000+ new citations and abstracts each day!

3.3 billion searches!

Strong domain shifting

BioBERT: specializing BERT for biomedical text mining

BERT is extremely easy to specialize as long as domain-specific text data is available in large quantities.

BioBERT: specializing BERT for biomedical text mining

BERT is extremely easy to specialize as long as domain-specific text data is available in large quantities.

BERT
pretrained on Wikipedia
and BookCorpus

PubMed
abstracts and full
text articles

BioBERT

[8] BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Lee et al., 2019

BioBERT: zero supervision, no feature engineering

BioBERT is a leap forward in biomedical text mining!

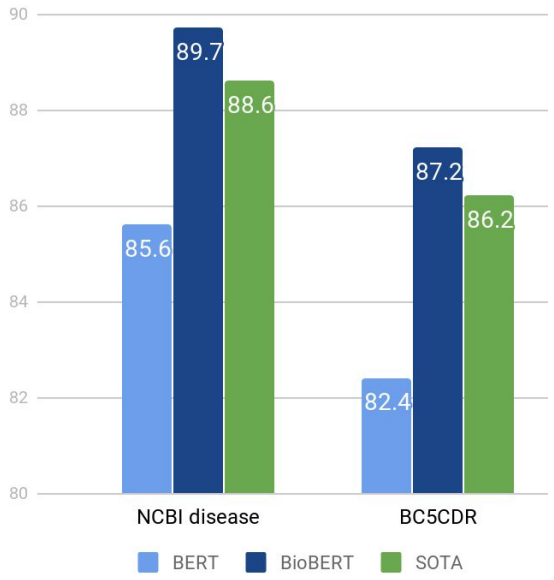
BioBERT

Compared to previous approaches:

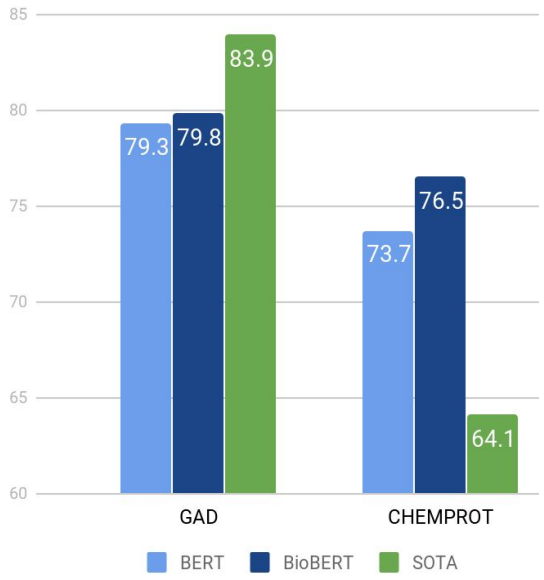
- No complex **preprocessing** strategies
- No need for specialized **task-specific architectures**
- No need for large amounts of **labeled** data
- No **biomedical expertise** required

BioBERT: evaluation on biomedical tasks

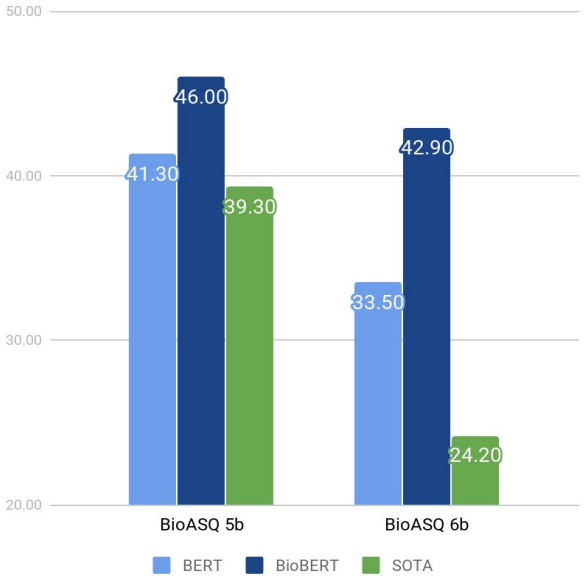
Named Entity Recognition
(F1 Score)



Relation Extraction
(F1 Score)

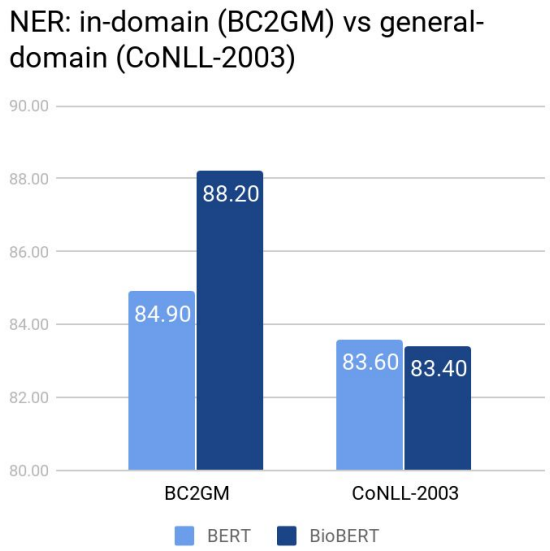


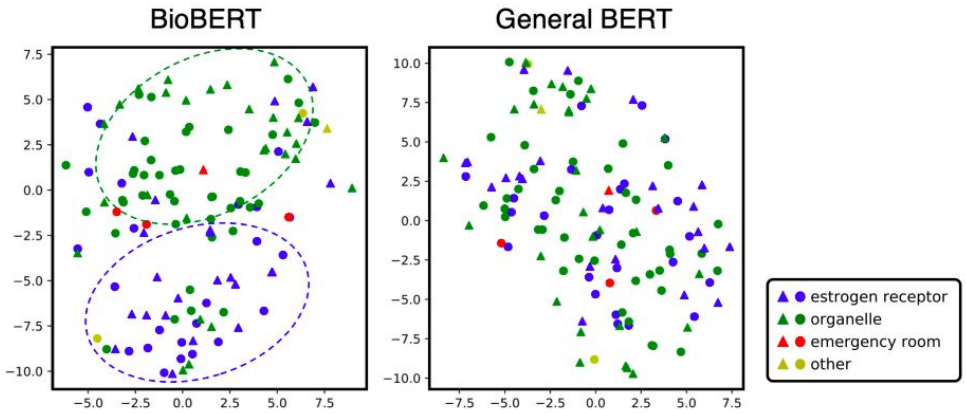
Question Answering
(Accuracy)



BioBERT: in-domain and general-domain generalization

Jin et al. (2019) show that the drop in performance of BioBERT on general-domain data (CoNLL-2003) is negligible.





[9] Probing Biomedical Embeddings from Language Models, Jin et al., 2019

Growing interest in biomedicine: MEDIQA 2019 @ ACL

The growing interest in NLP applied to medicine is also proved by the organization of several workshops and shared tasks.

MEDIQA [10] proposed 3 shared tasks:

1. Natural Language Inference
2. Question Entailment Recognition
3. Question Answering

72 participants!

[10] Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering, Abacha et al., 2019

Growing interest in biomedicine: W-NUT 2020 @ EMNLP

The growing interest in NLP applied to medicine is also proved by the organization of several workshops and shared tasks.

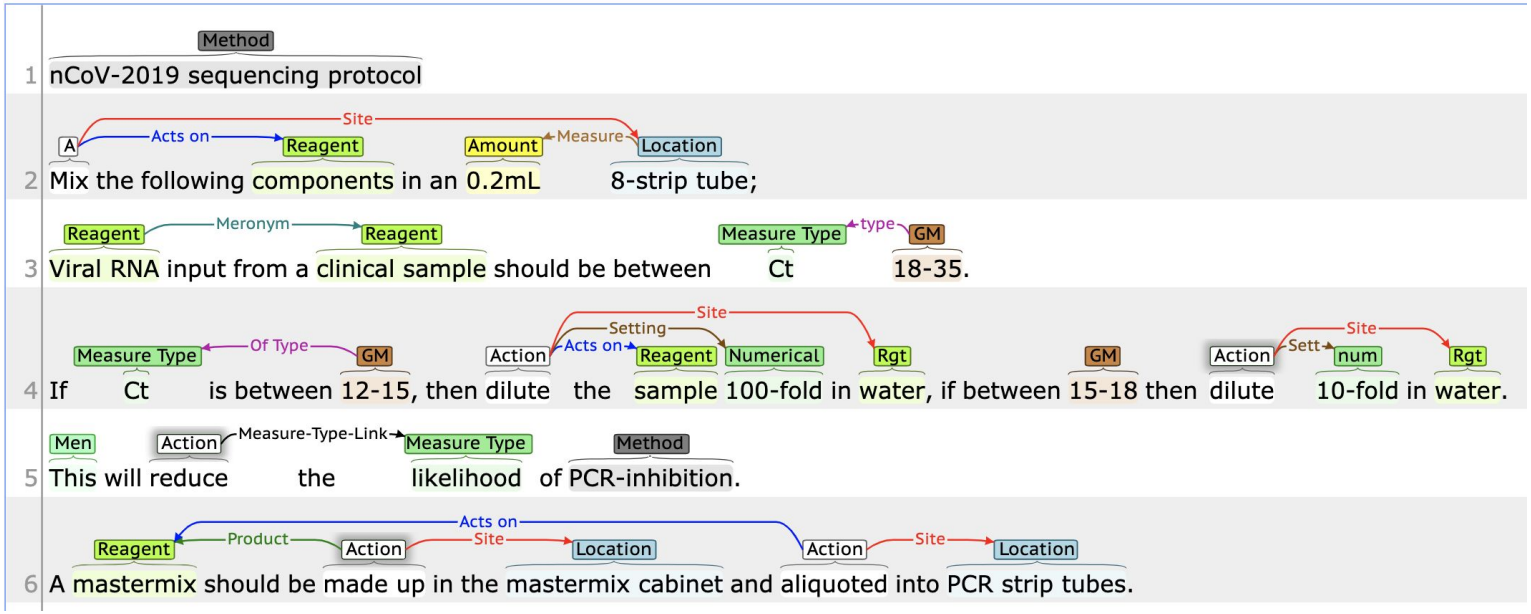
W-NUT [11] is proposing 3 shared tasks:

1. **Entity and Relation Recognition over wet-lab protocols**
2. Identification of Informative COVID-19 English Tweets
3. COVID-19 Event Extraction from Twitter

[11] http://noisy-text.github.io/2020/

Entity and Relation Recognition over wet-lab protocols

Lab protocols specify steps in performing a lab procedure. They are noisy, dense, and domain-specific. System entries are invited for event recognition and relation extraction over these lab protocols.



ClinicalBERT: specializing BERT on clinical notes

BERT is also finding application in healthcare thanks to its flexibility.

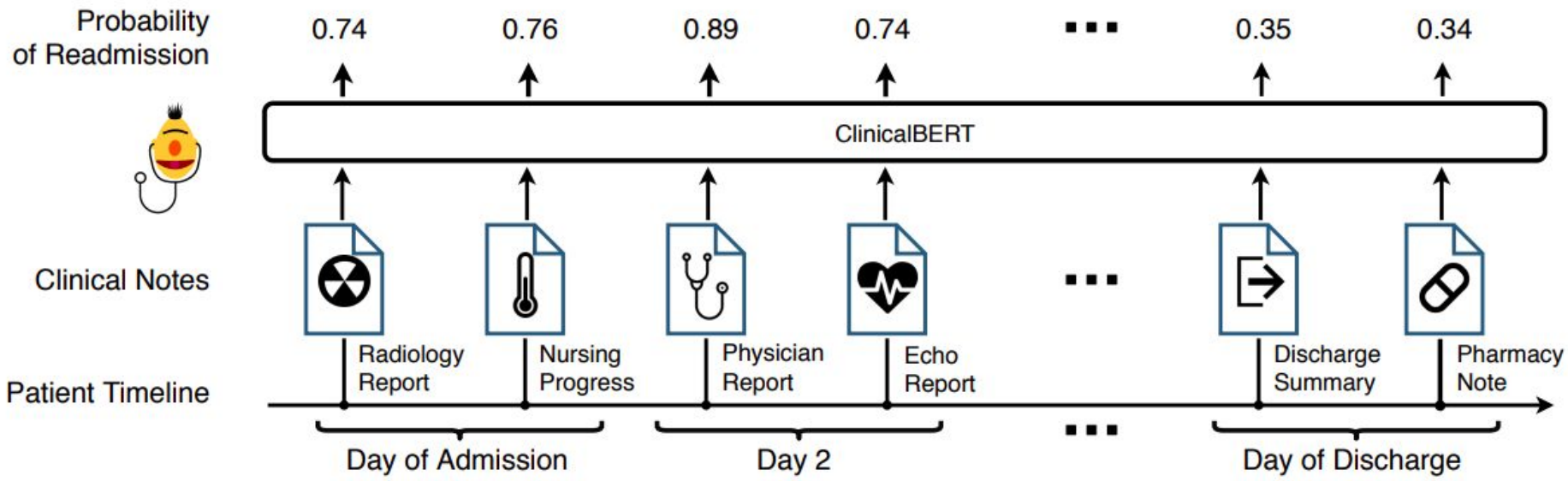
BERT
pretrained on Wikipedia
and BookCorpus

Clinical notes
about patient
medical histories

ClinicalBERT

[12] ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, Huang et al., 2019

ClinicalBERT: predicting hospital readmissions



ClinicalBERT: avoiding wastes and improving patients' QoL

While the idea behind ClinicalBERT is relatively simple, it can have a great impact for **hospitals**:

- Readmissions cause an estimated annual (avoidable) cost of **17 billion dollars**.
- For each readmission, doctors and nurses must **reread dozens of clinical notes**.

ClinicalBERT: avoiding wastes and improving patients' QoL

While the idea behind ClinicalBERT is relatively simple, it can have a great impact for **hospitals**:

- Readmissions cause an estimated annual (avoidable) cost of **17 billion dollars**.
- For each readmission, doctors and nurses must **reread dozens of clinical notes**.

But also for **patients**:

- Readmissions may cause **traumas** or require **longer treatments**.
- Increased **healthcare-related expenses**.

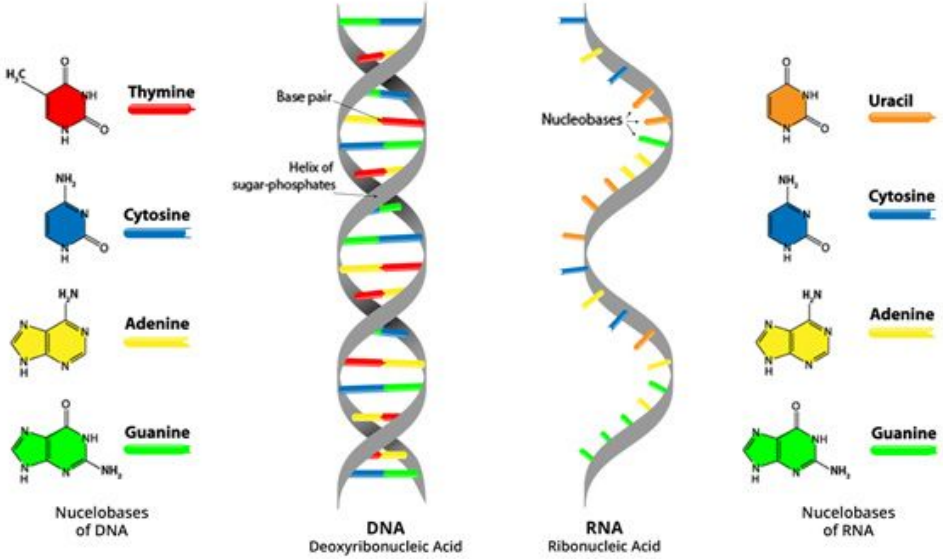
New frontiers: self-supervised NLP for biomedical research

Many fundamental molecules in our bodies can be represented as sequences.

New frontiers: self-supervised NLP for biomedical research

Many fundamental molecules in our bodies can be represented as sequences.

Can we apply NLP techniques to DNA modelling?



New frontiers: self-supervised NLP for biomedical research

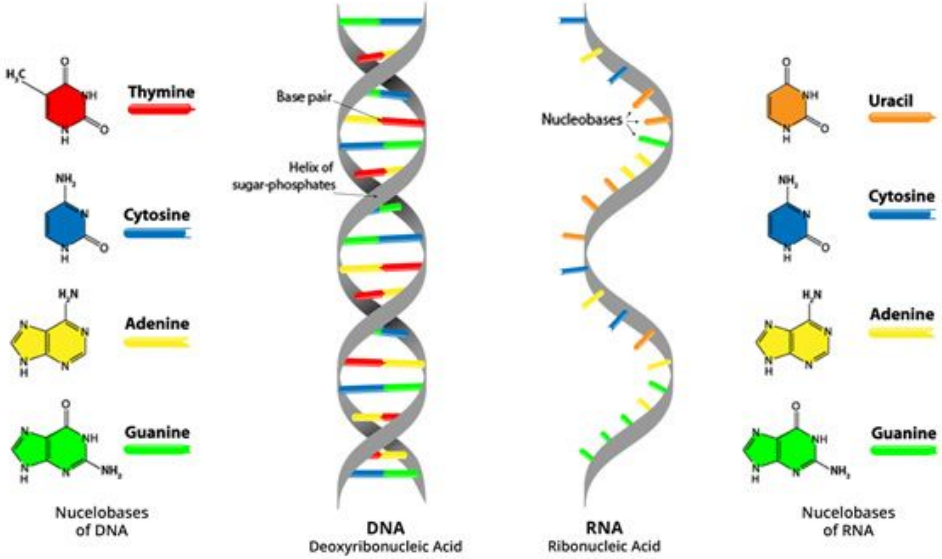
Many fundamental molecules in our bodies can be represented as sequences.

Can we apply NLP techniques to DNA modelling?

Yes!

- It's a sequence.
- Only 4 nucleosides.

But...



New frontiers: self-supervised NLP for biomedical research

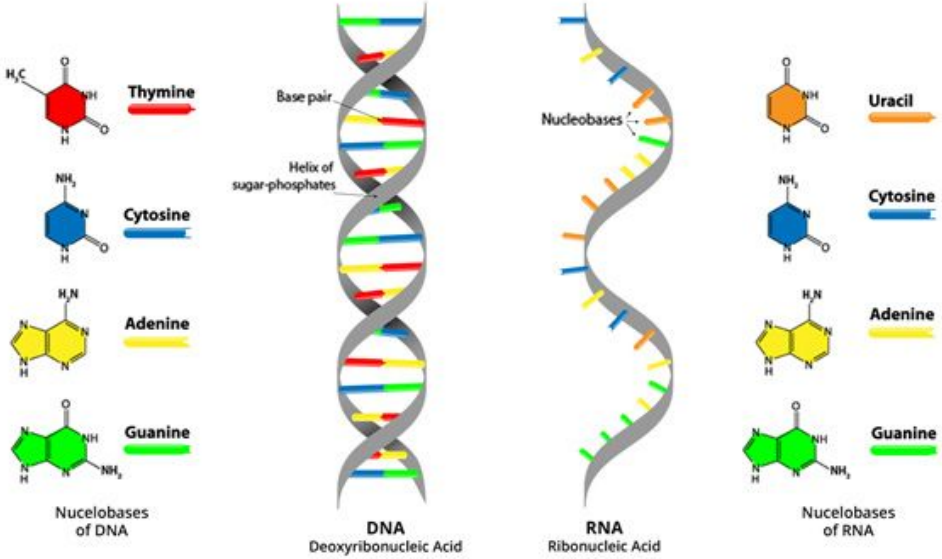
Many fundamental molecules in our bodies can be represented as sequences.

Can we apply NLP techniques to DNA modelling?

Yes!

- It's a sequence.
- Only 4 nucleosides.

But...



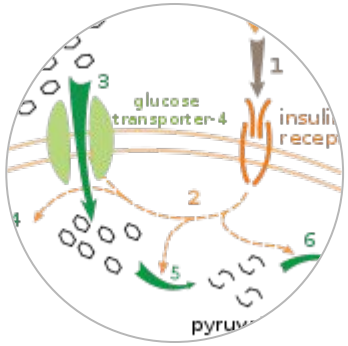
No...

- Too long sequences.

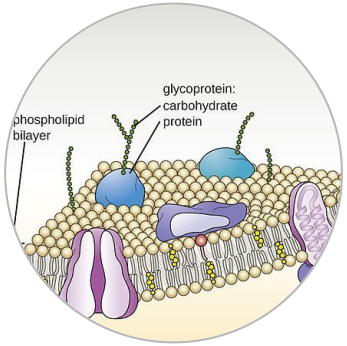
A human DNA molecule has up to 250 million nucleosides...

Proteins: a brief introduction

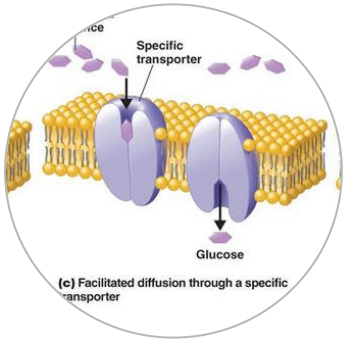
Proteins are biomolecules that play a fundamental role within our organisms.



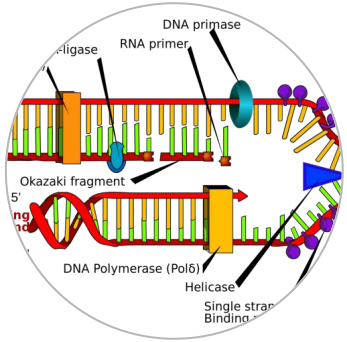
Metabolism
to convert food to energy.



Cell structure
is often provided by proteins.



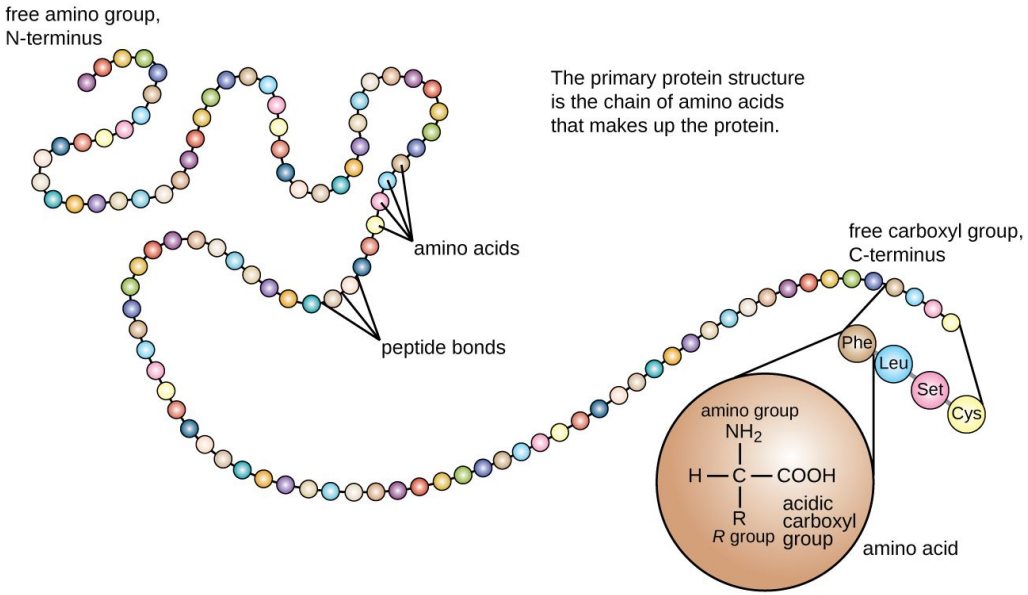
Transport
of other molecules.



Replication
of DNA and RNA molecules.

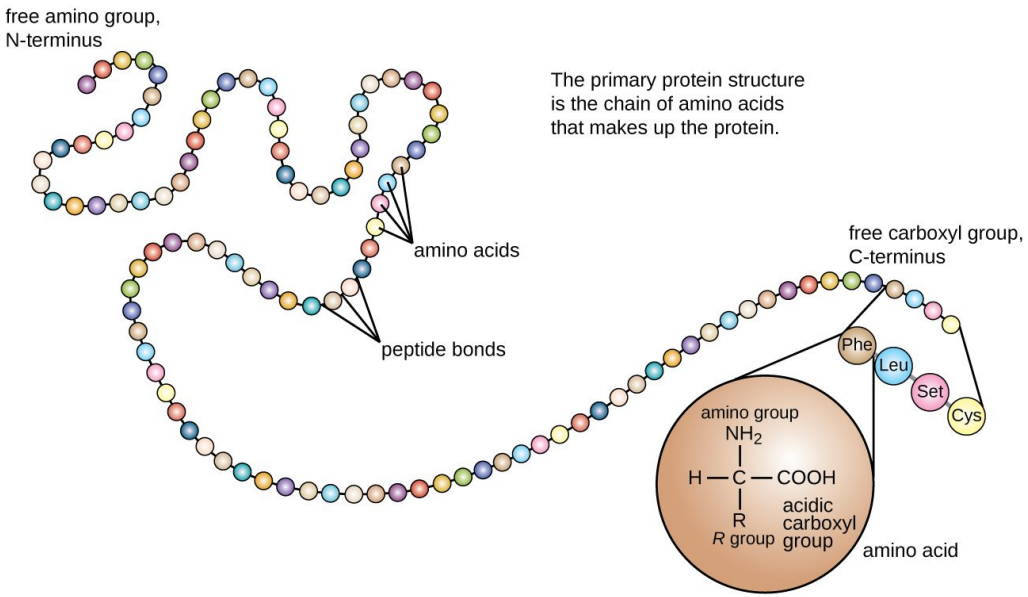
Proteins: sequences of amino acids

In their simplest form, proteins can be seen as sequences of smaller molecules called amino acids.



Proteins: sequences of amino acids

In their simplest form, proteins can be seen as sequences of smaller molecules called amino acids.

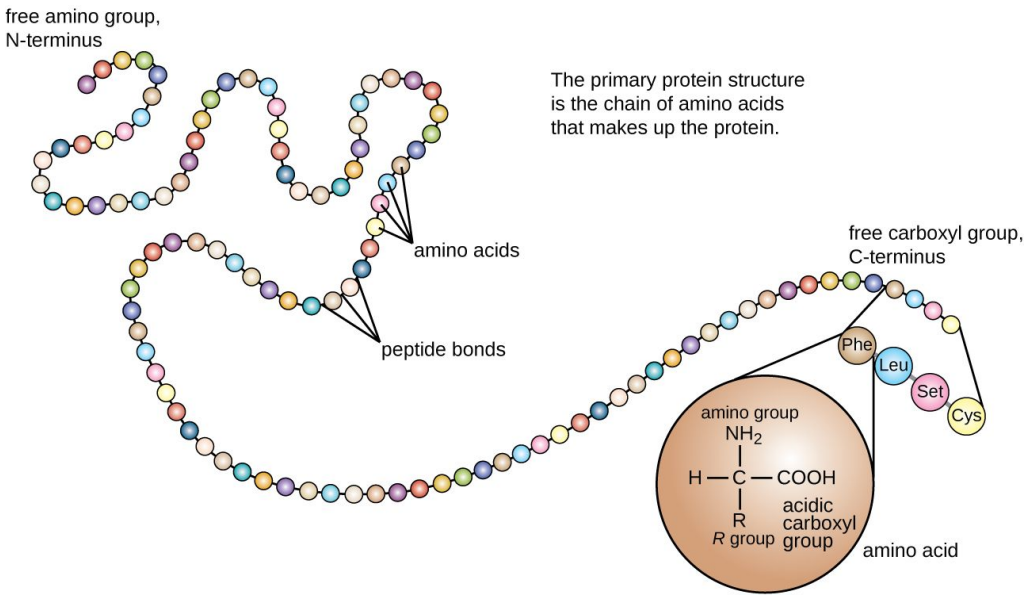


Good news:

- Sequences are usually no longer than a few hundreds.
- A sequence is believed to completely characterize a protein's behavior.

Proteins: sequences of amino acids

In their simplest form, proteins can be seen as sequences of smaller molecules called amino acids.



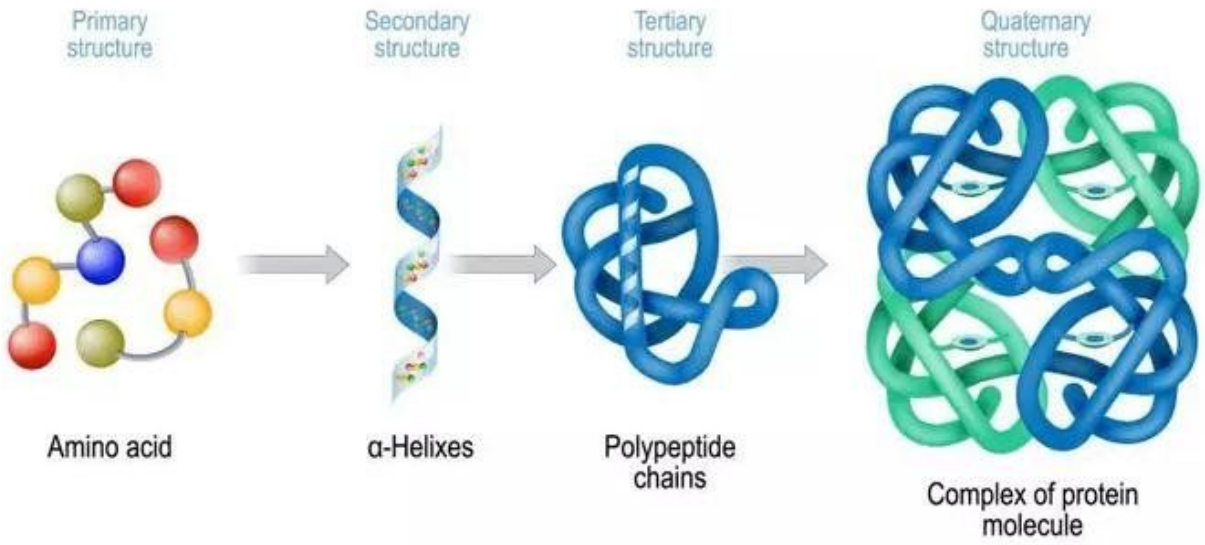
Good news:

- Sequences are usually no longer than a few hundreds.
- A sequence is believed to completely characterize a protein's behavior.

Bad news:

- Amino acid interactions are not linear, i.e., they may interact even if they are separated by hundreds of other amino acids.

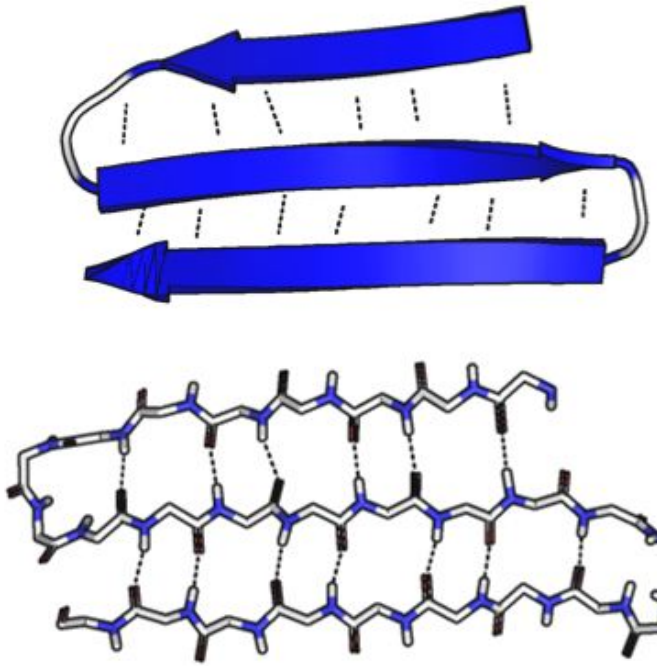
Proteins: structure



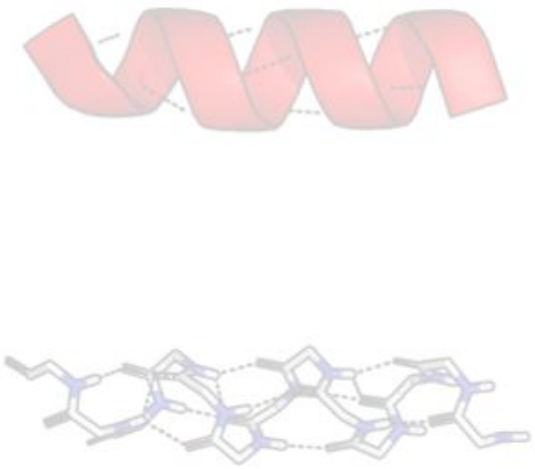
There is also a quinary structure...

Proteins: secondary structure

Secondary



β-Sheet (3 strands)

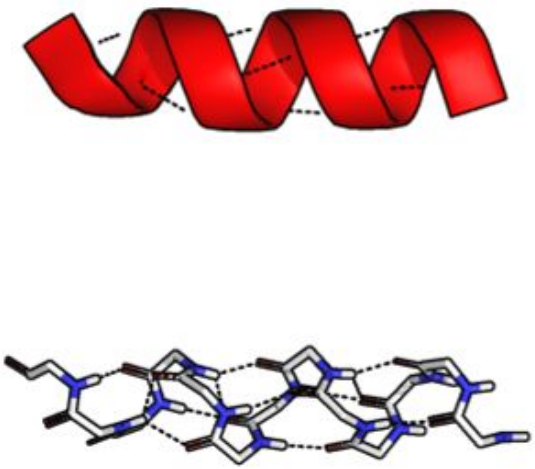


α-helix

Proteins: secondary structure

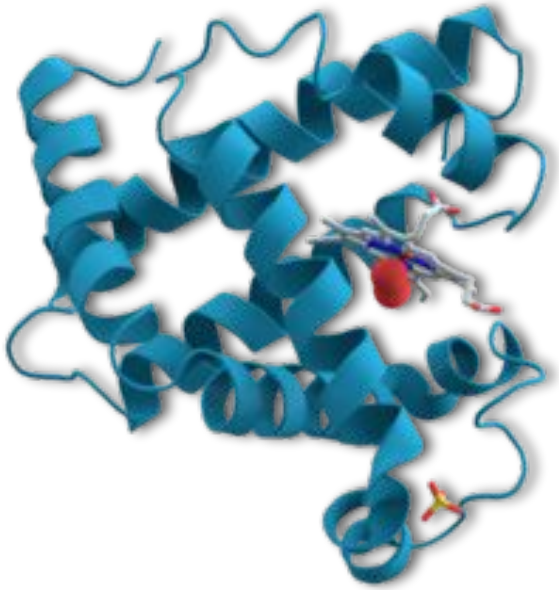
Secondary

β-Sheet (3 strands)



α-helix

Proteins: tertiary structure



Folding, Unfolding and Refolding

As of now, there is no general and accurate folding model, so researchers mostly still rely on experience, intuition and lengthy trials.

Evaluating Protein Transfer Learning with TAPE

Simple idea: train a language model (e.g. BERT) on a large amount of amino acids sequences.

```
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT
GKLPVPWPTL VTTFSYGVQC FSRYPDHMKQ HDFFKSAMPE GYVQERTIFF
KDDGNYKTRA EVKFEGDTLV NRIELKGIDF KEDGNILGHK LEYNYNSHNV
YIMADKQKNG IKVNFKIRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHY
LSTQSALSKD PNEKRDHMVL LEFVTAAGIT HGMDELYK
```

[13] Evaluating Protein Transfer Learning with TAPE, Rao et al., 2019

Evaluating Protein Transfer Learning with TAPE

Simple idea: train a language model (e.g. BERT) on a large amount of amino acids sequences.

```
MSKGEE[redacted]FTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT
GKLPVPW[redacted]TL VTTFSYGVQC [redacted]YPDHMKQ HDFFK[redacted]AMP[redacted] GYVQERTIFF
KDD[redacted]NYKTRA EVK[redacted]EGDT[redacted]V N[redacted]IELKGIDF KEDGNILGHK L[redacted]YNYNSH[redacted]V
YIMADKQ[redacted]NG IKVNFKIRHN I[redacted]D[redacted]SVQLAD [redacted]YQQNTPIGD GPVL[redacted]PDNHY
LSTQSALSKD PN[redacted]DHMVL LEFVTAAGIT HGMDELYK
```

[13] Evaluating Protein Transfer Learning with TAPE, Rao et al., 2019

Evaluating Protein Transfer Learning with TAPE

Simple idea: train a language model (e.g. BERT) on a large amount of amino acids sequences.

Then, further train the pretrained language model on a complex protein task such as secondary and tertiary structure prediction, where labeled data may be scarce.

[13] Evaluating Protein Transfer Learning with TAPE, Rao et al., 2019

Evaluating Protein Transfer Learning with TAPE

Simple idea: train a language model (e.g. BERT) on a large amount of amino acids sequences.

Then, further train the pretrained language model on a complex protein task such as secondary and tertiary structure prediction, where labeled data may be scarce.

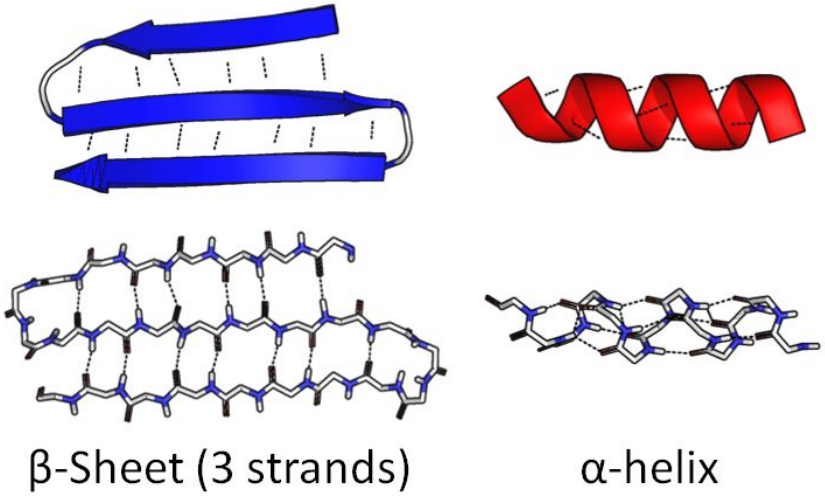
Datasets:

- **Protein Data Bank** (PDB): 160’000 protein sequences.
- **UniParc**: 300’000’000 protein sequences.

[13] Evaluating Protein Transfer Learning with TAPE, Rao et al., 2019

Secondary Structure Prediction

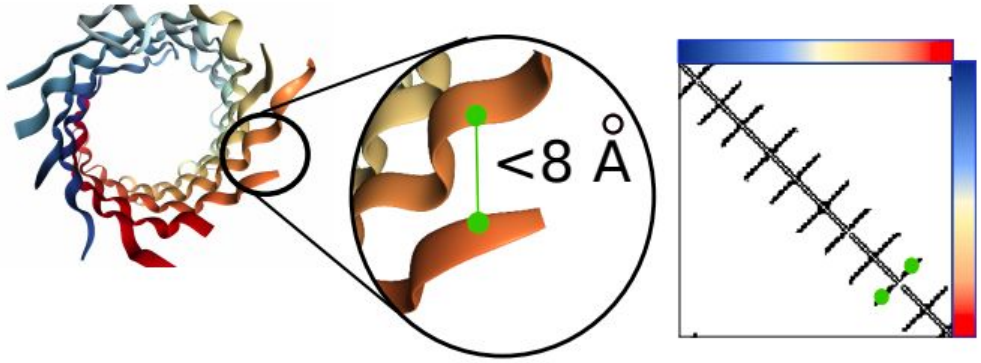
Determine whether an amino acid in the protein belongs to an alpha helix or beta sheet or neither of them.



Understanding local structure and amino acid interactions.

Contact Prediction

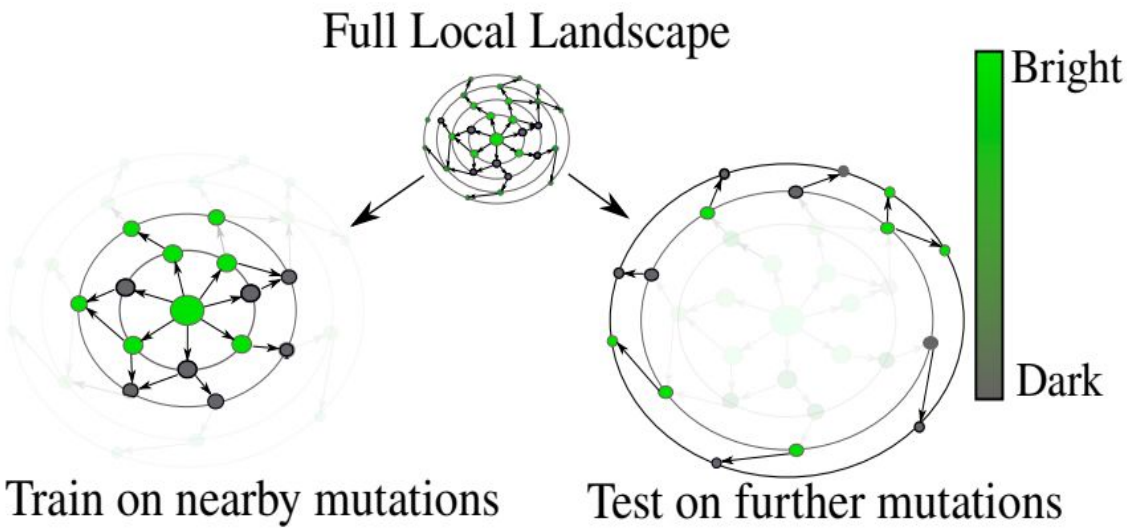
Related to tertiary structure prediction, consists in determining whether two non-consecutive amino acids are 3-dimensionally close in the folded protein.



Understanding global structure and amino acid interactions.

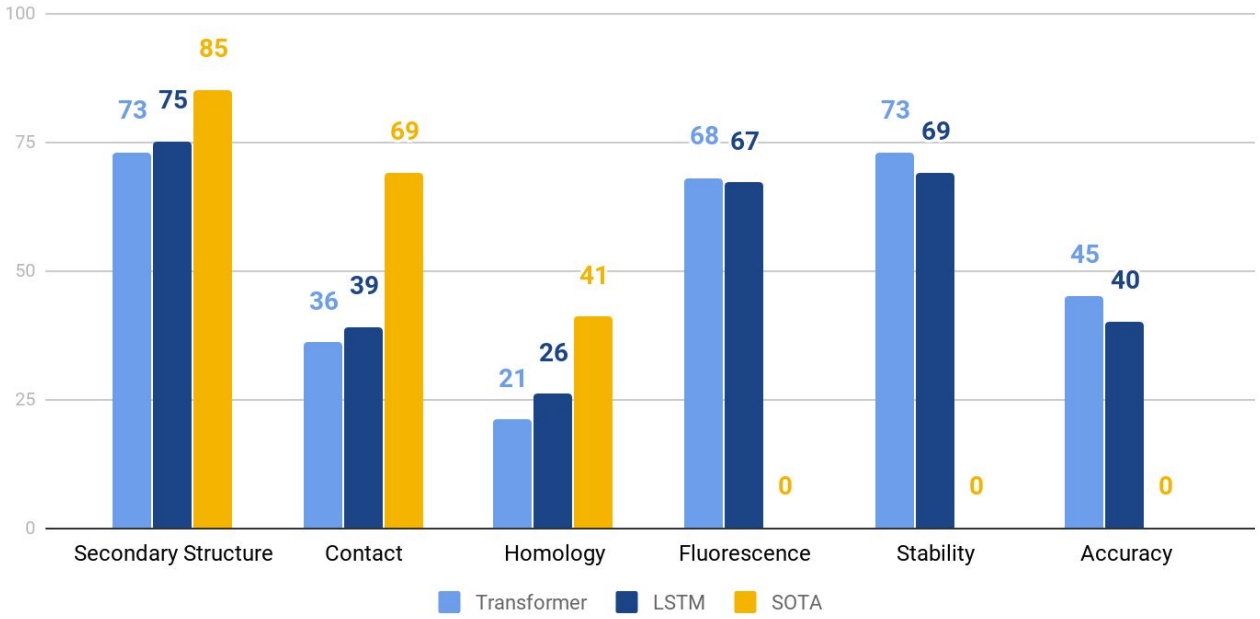
Fluorescence test: genotypes and phenotypes

Related to the overall behavior of a protein, consists in determining whether a mutation of a single amino acid changes the characteristics of the protein.



Results: self-supervision and pretraining work!

Simple self-supervision and language model pretraining show promising results, even though there is still a significant gap to bridge with respect to complex specialized models.



Learning the “language” of proteins

The NeurIPS work of Rao et al. (2019) showed that self-supervised language modeling is a promising direction for future protein models.

- Compared to specialized architectures, **one language model fits multiple tasks**.
- Rao et al. (2019) experiment with **a relatively small model**, “only” 38M parameters.
- **No feature engineering is required** for any of the proposed models.

Conclusion

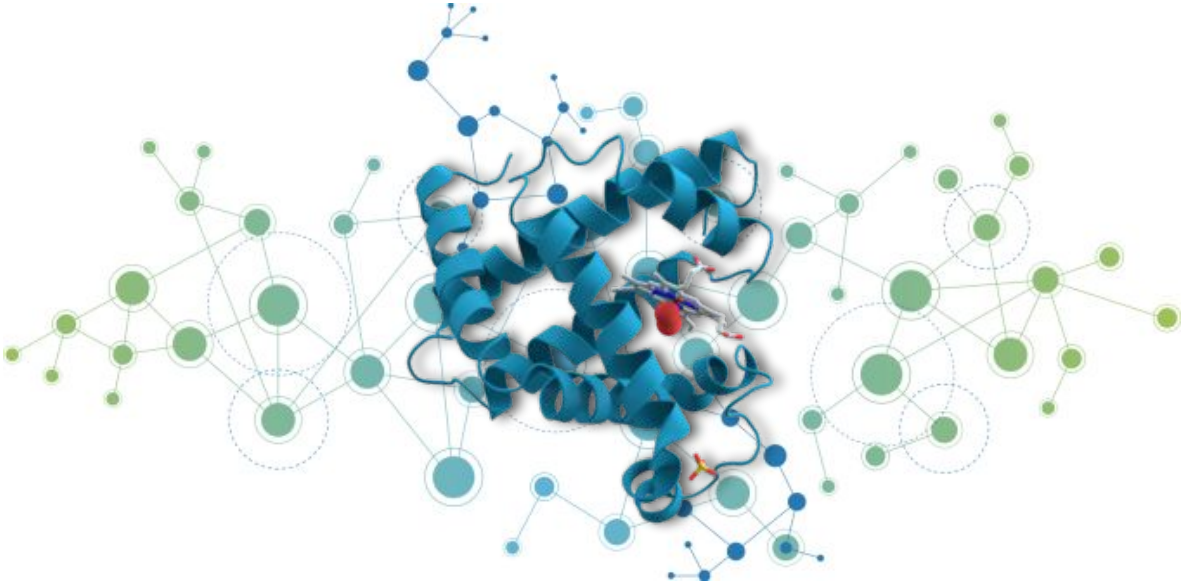
Today we have seen self-supervised language model pretraining for:

- Biomedical text analysis and text mining.
- Real-life clinical situations.
- Basic research on protein modeling.

The future for self-supervised pretraining techniques in language modeling looks brighter and brighter!

Feel free to ask questions,

Thanks for your attention!



Simone Conia

Reading Group @ Sapienza NLP

September 16th, 2020

SAPIENZA
UNIVERSITÀ DI ROMA