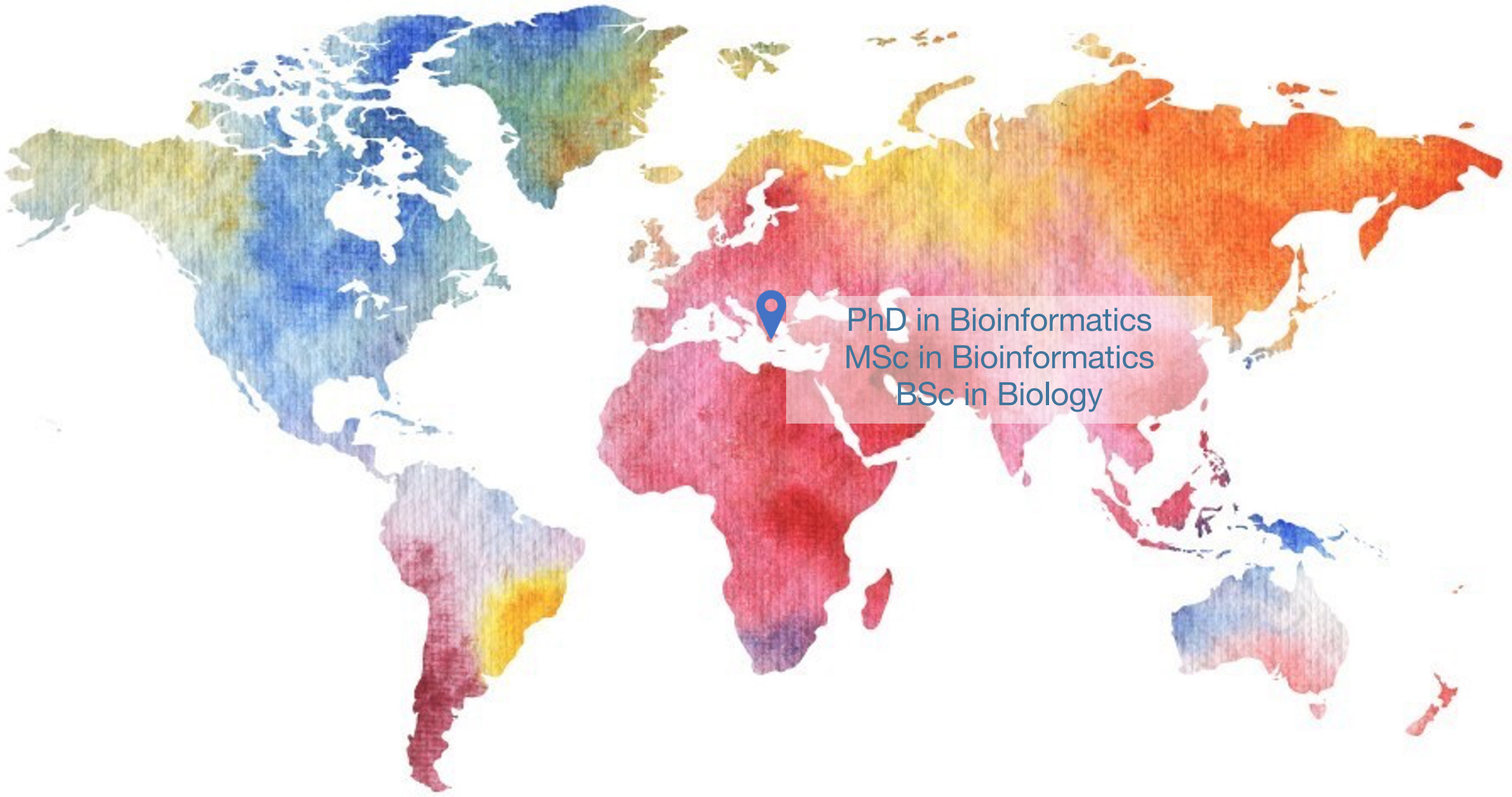
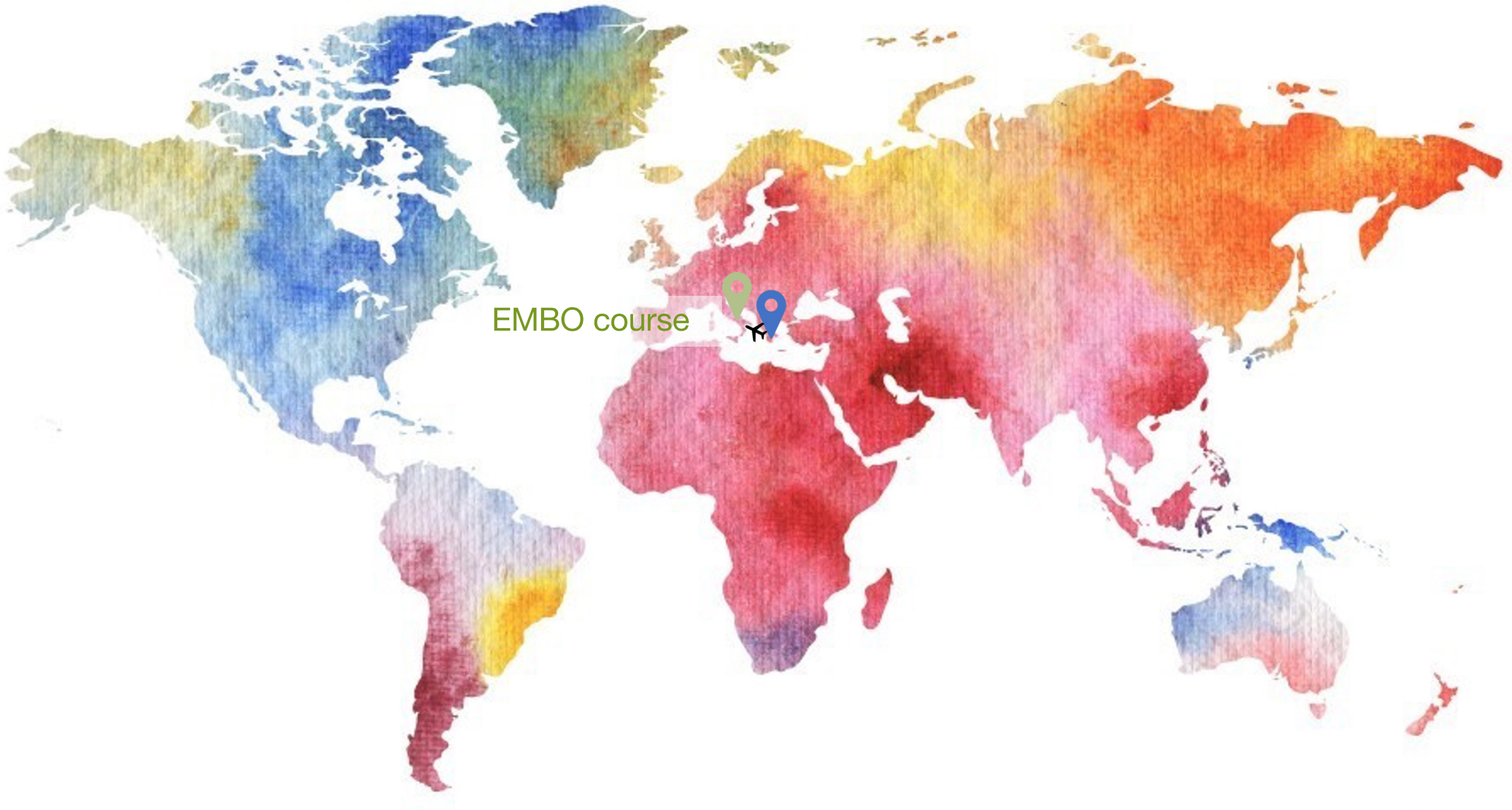
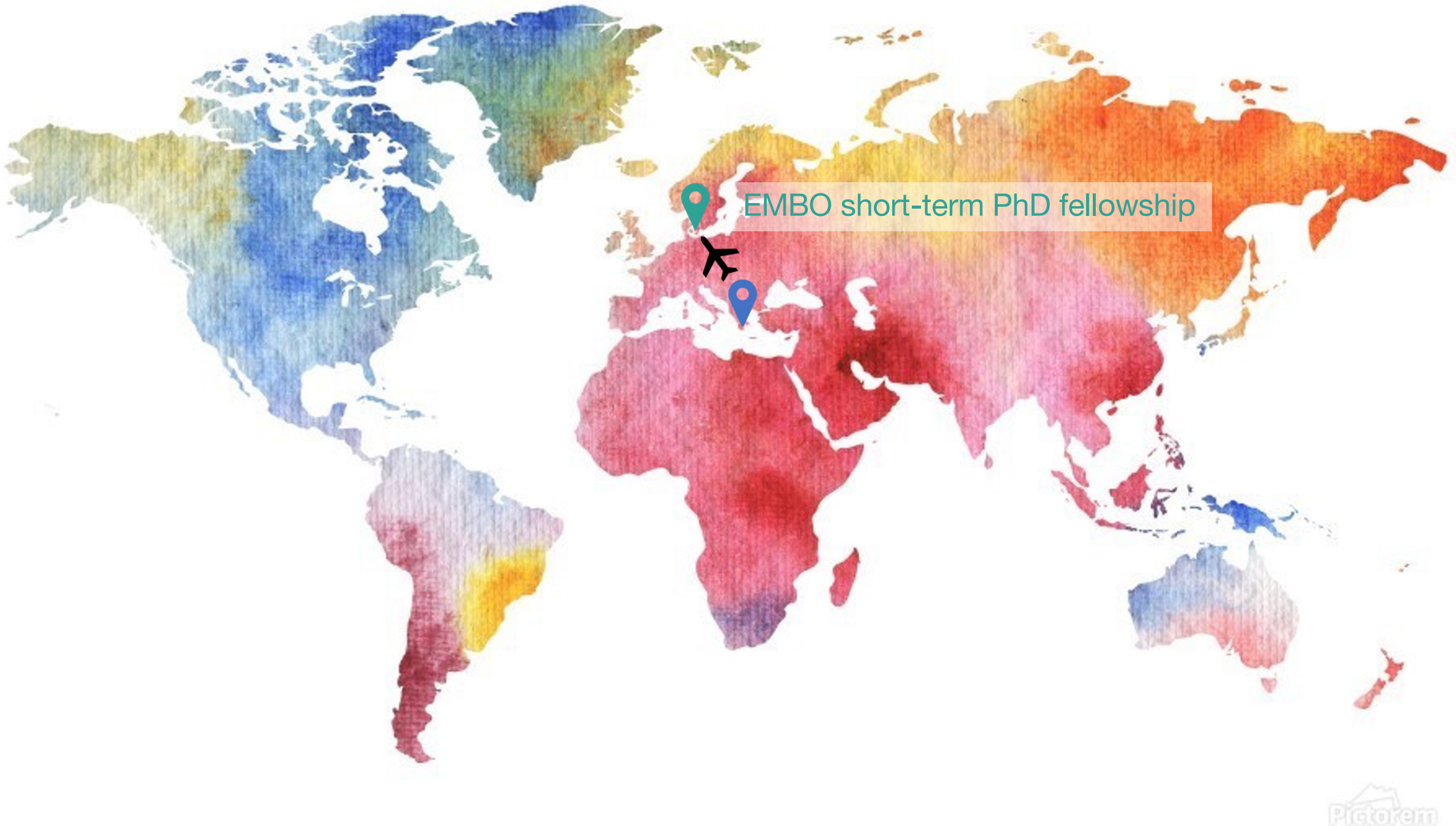
Extracting protein-protein interactions from the literature with deep learning-based text mining

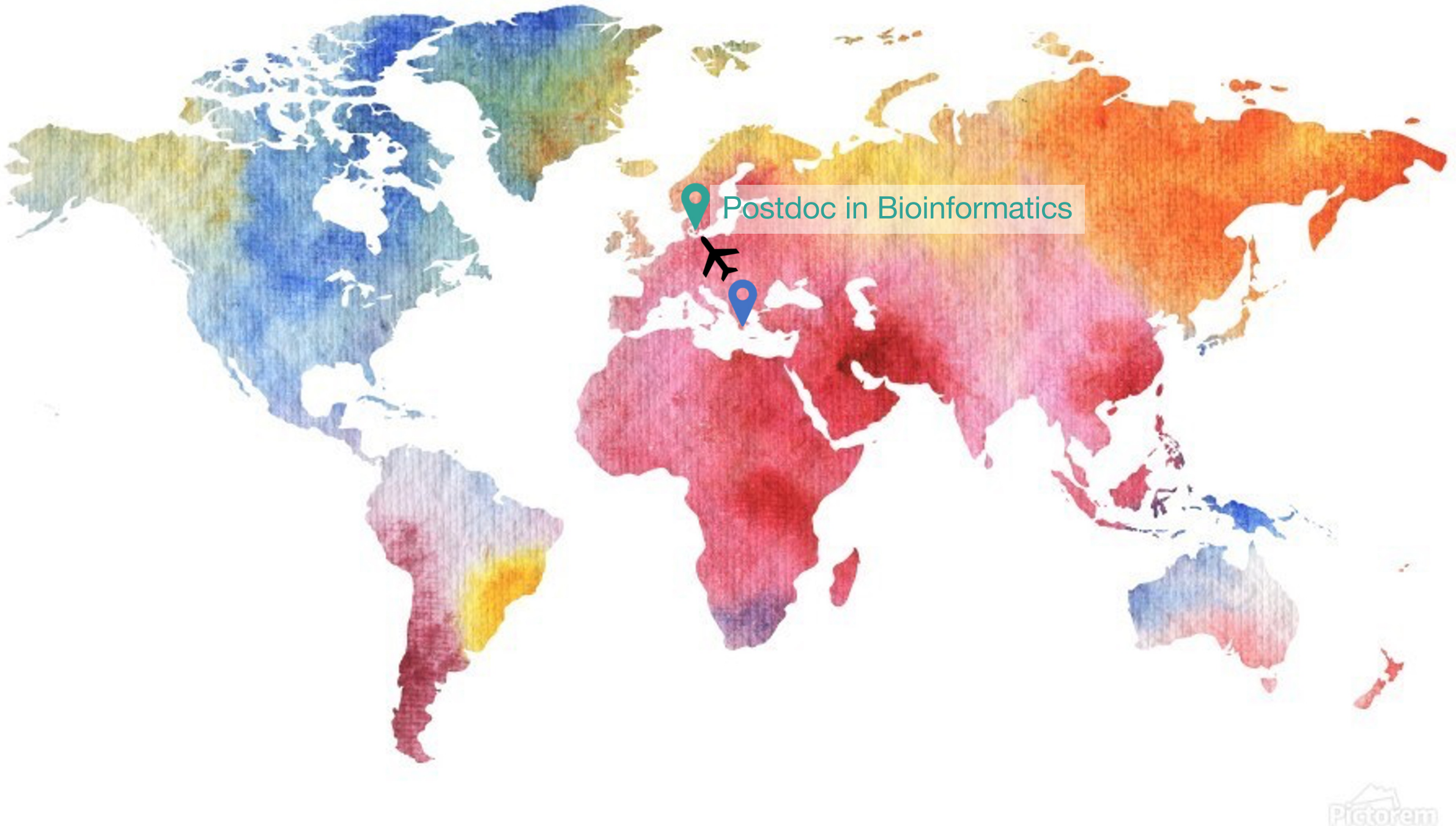
Katerina Nastou
Cellular Network Biology Group
Novo Nordisk Foundation Center for Protein Research
Machine Learning course, 27-04-2023

UNIVERSITY OF COPENHAGEN


PhD in Bioinformatics
MSc in Bioinformatics
BSc in Biology


EMBO course

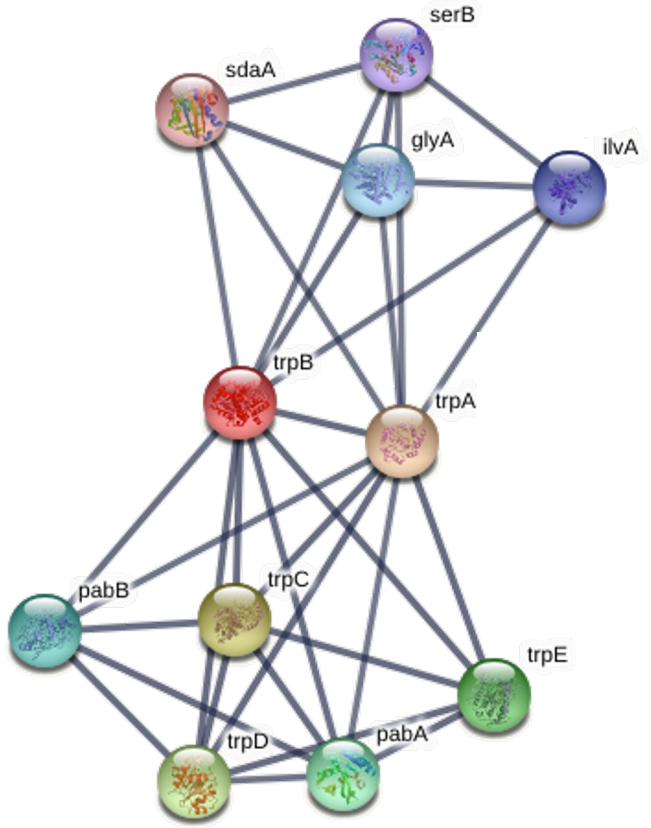

EMBO short-term PhD fellowship


Postdoc in Bioinformatics


MSCA Postdoctoral Fellowship


Deep learning-based text mining
Network biology
Protein function prediction
Database design and development

STRING database



Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. (2022). *The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic Acids Research, Jan 6;51(D1):D638-D646.

STRING database

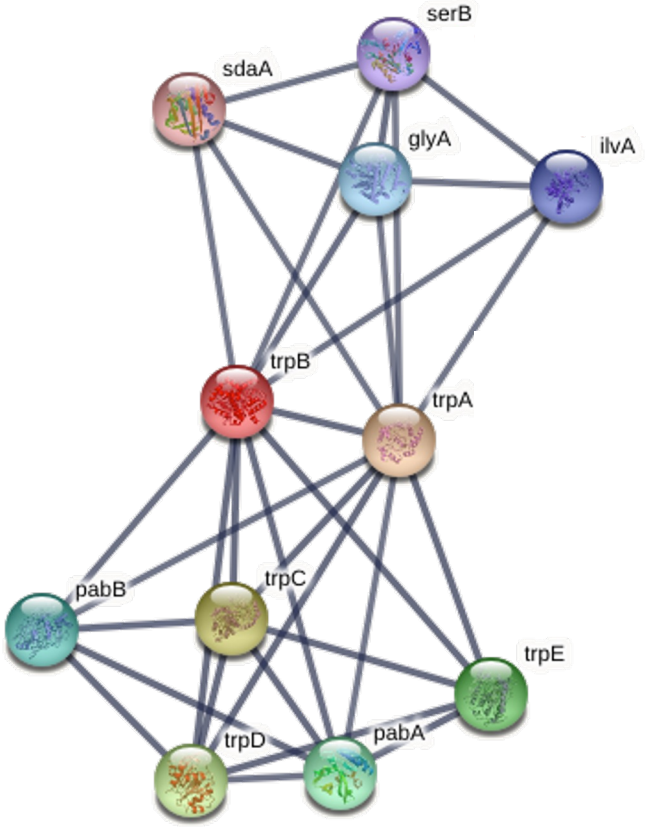
STRING

co-expression

interaction databases

genomic context

text mining



Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. (2022). *The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic Acids Research, Jan 6;51(D1):D638-D646.

~100,000 users monthly

Text Mining

Scientific literature

PubMed


NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
PubMed.gov
Search

PubMed

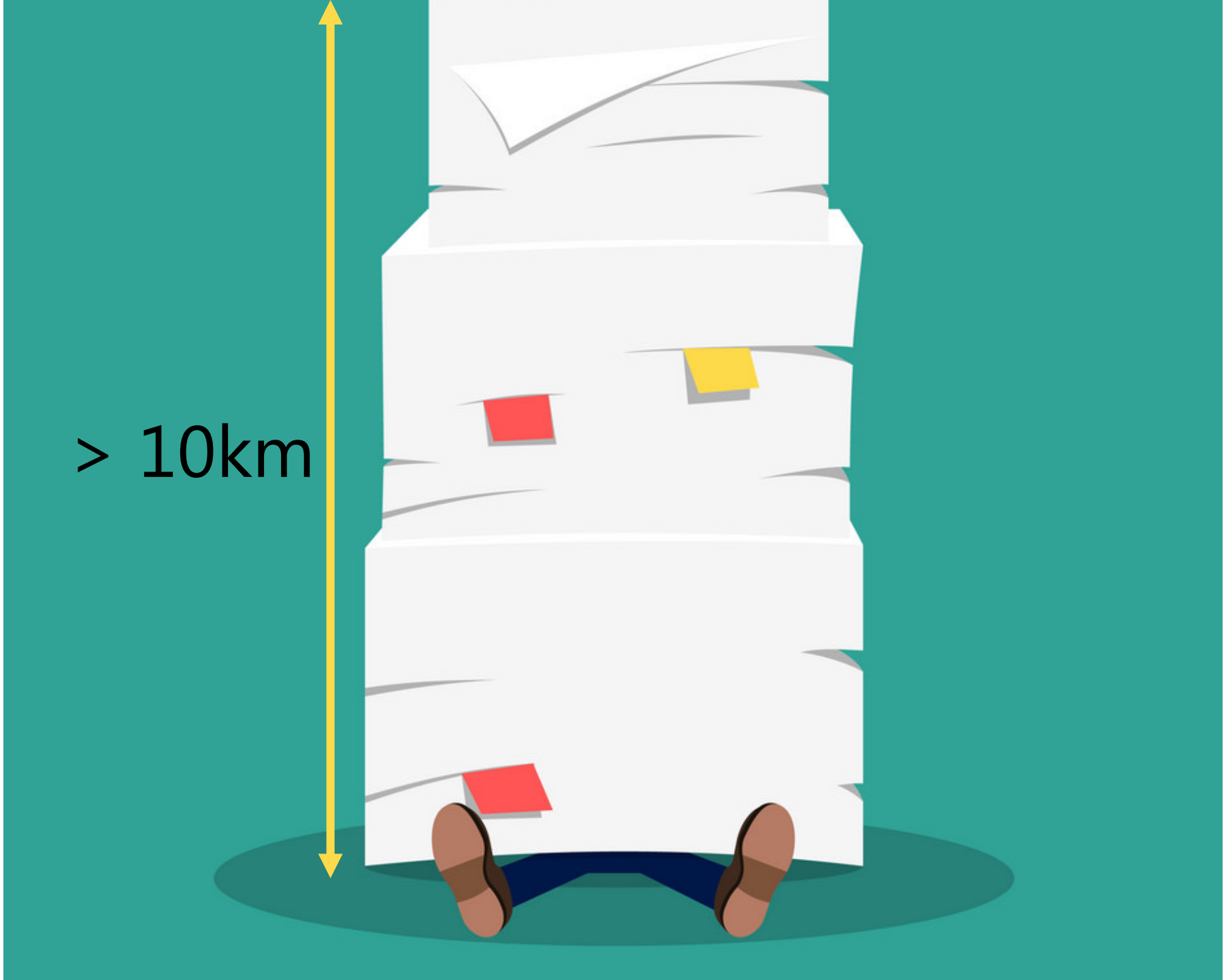
>34 million articles

abstracts only

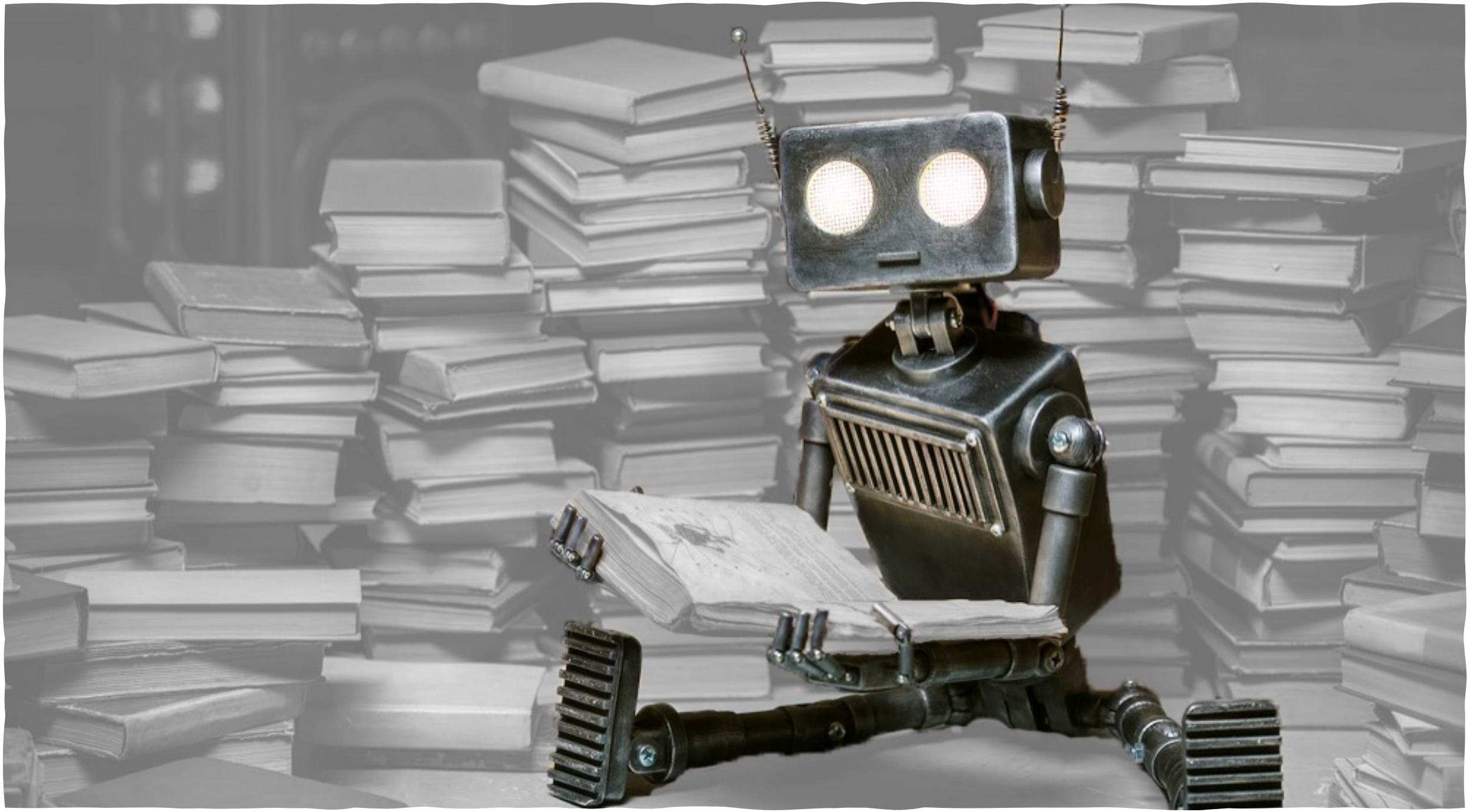
PubMed

~5 pages each

Print in standard 80gr A4 paper


> 10km


> 10km
Too much to read!



STRING text-mining channel before v11.5:
Statistical analysis of co-occurrence in documents + Natural Language Processing

Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C., Jensen L.J. *STRING v9.1: protein-protein interaction networks, with increased coverage and integration*. Nucleic Acids Res. 2013

STRING text-mining channel:

Statistical analysis of co-occurrence in documents +

Natural Language Processing

Functional associations

Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C., Jensen L.J. *STRING v9.1: protein-protein interaction networks, with increased coverage and integration*. Nucleic Acids Res. 2013

STRING text-mining channel before v11.5:

Statistical analysis of co-occurrence in documents +

Natural Language Processing

physical and regulatory interactions

Franceschini A., Szklarczyk D., Frankild S., Kuhn M., Simonovic M., Roth A., Lin J., Minguez P., Bork P., von Mering C., Jensen L.J. *STRING v9.1: protein-protein interaction networks, with increased coverage and integration*. Nucleic Acids Res. 2013

STRING text-mining channel before v11.5 :

Statistical analysis of co-occurrence in documents +

Natural Language Processing

physical and regulatory interactions

STRING text-mining channel before v11.5 :

Statistical analysis of co-occurrence in documents +

Natural Language Processing

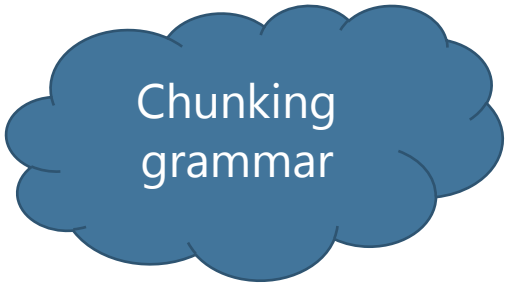
physical and regulatory interactions

STRING text-mining channel before v11.5 :

Statistical analysis of co-occurrence in documents +

Natural Language Processing

physical and regulatory interactions



STRING text-mining channel before v11.5 :

Statistical analysis of co-occurrence in documents +

Natural Language Processing

physical and regulatory interactions

🔥 Deep Learning-based Language Representation Models 🔥

BERT

Devlin, J., Chang, M., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv, (2018).

What is BERT?

BERT: Bidirectional Encoder Representations from Transformers

BERT: Bidirectional Encoder Representations from Transformers

Language modelling system, pre-trained with unlabeled data, then fine-tuned

BERT: Bidirectional Encoder Representations from Transformers

Language modelling system,
pre-trained with unlabeled
data, then fine-tuned

BERT is not made to tackle a specific task or answer a specific problem but it gives the best, the most efficient and the most flexible representation for words and sequences possible

BERT: Bidirectional Encoder Representations from Transformers

Language modelling system,
pre-trained with unlabeled
data, then fine-tuning

Add a small layer to the model made for our specific task and train the whole thing very lightly
(really fast, because the core of the model has been already trained)

BERT: Bidirectional Encoder Representations from Transformers

Defines the powerful neural network architecture of BERT, that is based on a self-attention mechanism

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. (2017) *Attention is All you Need*. arXiv. 5998--6008.

BERT: Bidirectional Encoder Representations from Transformers

Has full access to left and right **context** when dealing with a word (masked)

Defines the training process

Before that in NLP we had access only to the left or the right part of the sentence or in pseudo-bidirectional models we have separate access to left and right and then concatenate both

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

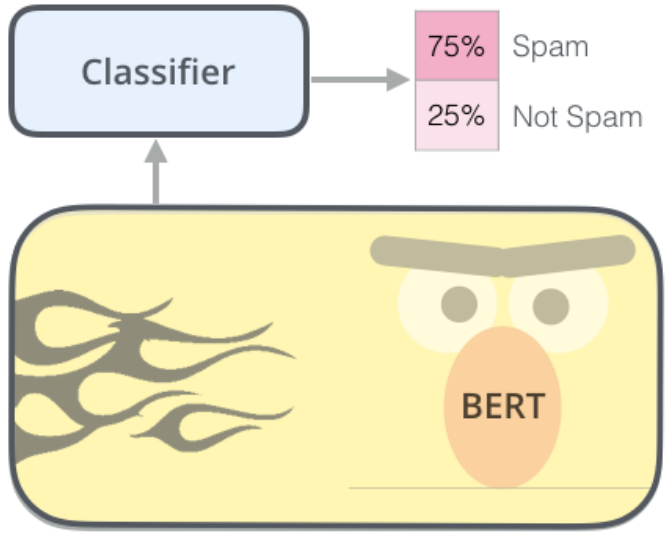
Objective:

Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model: (pre-trained in step #1)



Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

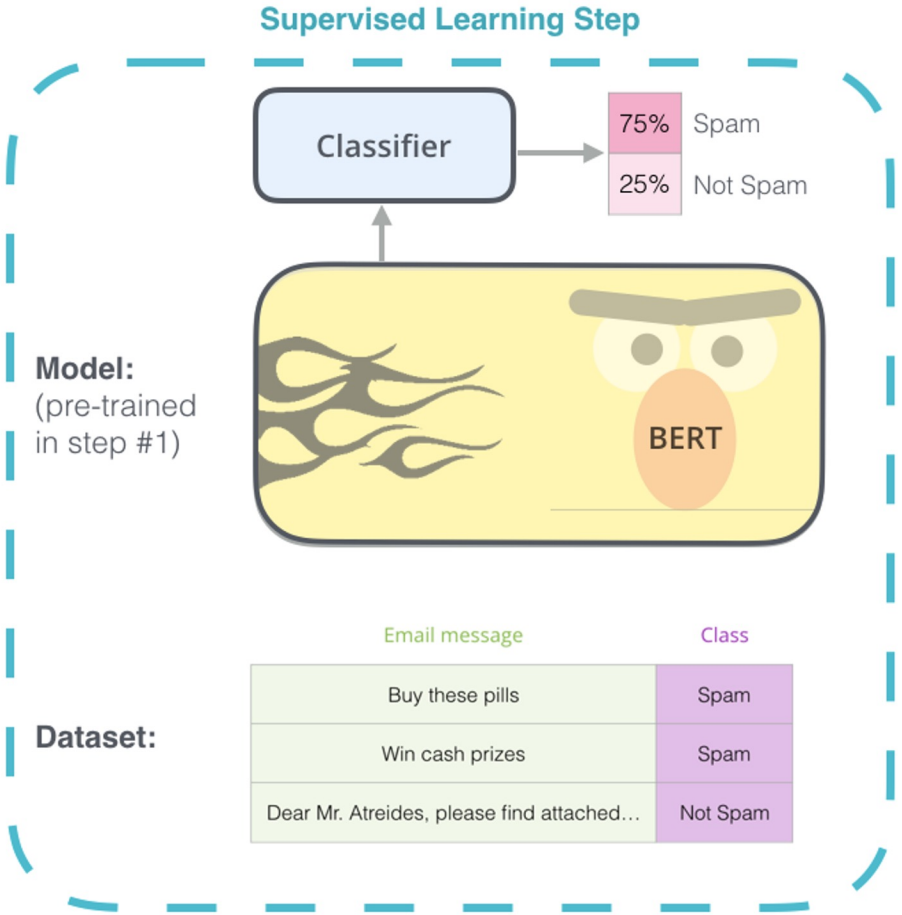
J. Alammar, *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning),* 2019 http://jalammar.github.io/illustrated-bert/

BERT: Bidirectional Encoder Representations from Transformers

During **supervised training** words of interest are **masked** and the model learns the label for the masked word based on the **context** around them

Example	Label
When **[MASK]** becomes mutated, it loses its function, resulting in abnormal cell proliferation and tumor progression.	gene
Patients suffering from **[MASK]** have a low white blood cell count (WBC).	disease

2 - Supervised training on a specific task with a labeled dataset.



Domain-specific deep-learning models

Domain-specific deep-learning models

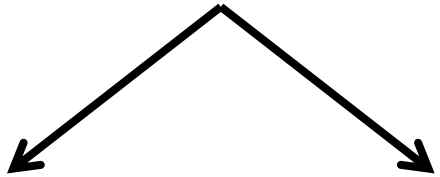
BioBERT, SciBERT, BlueBERT (NCBI-BERT),
BioMegatron, RoBERTa-bio and more...

Major challenge in using deep learning for Relation Extraction

Obtaining enough amounts of labelled text for training

Major challenge in using deep learning for Relation Extraction

Obtaining enough amounts of labelled text for training

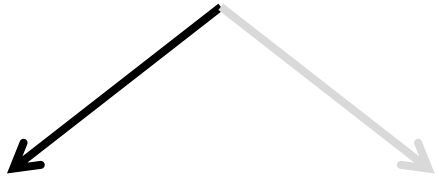


Manual annotation

Distant supervision

Major challenge in using deep learning for Relation Extraction

Obtaining enough amounts of labelled text for training



Manual annotation

Distant supervision

Manual Annotation

+good quality

+more detailed annotations

-manual

Development of guidelines for manual annotation

https://katnastou.github.io/stringdb-typed-relation-annotation-docs

Documentation for Typed Relation Annotation

General guidelines

- Annotations should be made according to the annotator's best understanding of the **author's intended meaning in context**. For example, relations expressed using ambiguous verbs such as **"associate"** that express complex formation in some contexts but not others should be annotated if and only if the annotator interprets the authors as intending to describe complex formation. The annotators should only use the text excerpt they have available to make this judgement, (e.g.)
- Annotators should treat named entities as being **masked**, i.e. they shouldn't annotate relationships between entities just based on their names, when they would be unable to make the same annotations for two other entities.

Complex formation

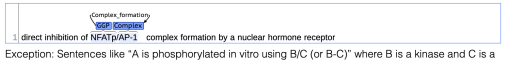
Undirected binary relation associating two proteins that form a complex. Annotated for any statement implying the existence of a complex, including statements explicitly discussing the dissociation of a complex. Relevant ontology terms:

- GO:0065003 (**protein-containing complex assembly**): The aggregation, arrangement and bonding together of a set of macromolecules to form a protein-containing complex.
- GO:0032984 (**protein-containing complex disassembly**): The disaggregation of a protein-containing macromolecular complex into its constituent components.
- GO:0032991 (**protein-containing complex**): A stable assembly of two or more macromolecules, i.e. proteins, nucleic acids, carbohydrates or lipids, in which at least one component is a protein and the constituent parts function together.

Note that by contrast to the scope of GO:0032991 (protein-containing complex) and related terms, the annotated complex formation relation is restricted to cases where both of the associated constituents are *proteins*, *protein complexes*, *protein families*, *groups of proteins* or *chemicals*.

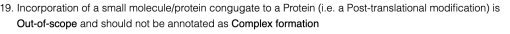
Detailed guidelines

1. Complex formation relations can be annotated between two different protein mentions, but also between the same mentions, when the masked entities could be viewed as two different entities, e.g.. However, statements such as "homodimerization of A" **are not annotated** as *Complex formation*
2. Complexes of more than two proteins are annotated by creating **all binary relations** between the components
3. Nominalized expressions ("interaction of A and B", "A/B interaction", "A:B complex") and noun phrases with **any surface word** that can be understood as implying the existence of a complex ("A/B complex", "A/B heterodimer") are **annotated** as expressing complex formation relations. However, **in the absence of any such word**, text such as "A/B" is not annotated. The text A-B will be annotated based on the understanding of the annotator from the entire context (abstract or paragraph) and not based on former biological knowledge.



Exception: Sentences like "A is phosphorylated in vitro using B/C (or B-C)" where B is a kinase and C is a cyclin have been annotated as "B_complex_formation_C". These can be reverted or we can check the error rate in this specific subproblem.

4. Relations **should not be interpreted as combinations**, on the contrary each annotated relation should be **valid on each own** (e.g. *"A positively regulates the proteolytic degradation of B and that leads to the rapid* 18716324
15. If protein A acts as an **effector** for protein B then annotate **B>Regulates>A**. Example from 22159419
16. If **transfection** with a protein leads to it obtaining a relationship with another protein entity, then that relationship should be annotated. Example from 10613896
17. **Chemicals** COVALENTLY bound to other entities are not annotated as **Complex formation**, since complex formation is non-covalent.
18. Orientation of *Protein A* relatively to *Protein B* is not enough cue to annotate **Complex formation** e.g. from 19858358



19. Incorporation of a small molecule/protein congugate to a Protein (i.e. a Post-translational modification) is **Out-of-scope** and should not be annotated as **Complex formation**
20. Proteoforms (e.g. proteins with PTMs, or isoforms), should receive annotations as if they were the main isoform/unmodified protein. e.g. from 19015234

---

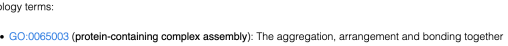
21. **Complex formation** should be annotated when a **Chemical** binds to any other entity (**Protein**, **Family** or **Complex**) unless it is clearly stated that the bond is covalent (either by the fact that it is a PTM or covalently bound is mentioned in the text).

Negation and speculation

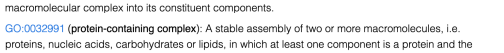
1. Statements explicitly **denying** the formation of a complex (e.g. "A does not bind B") are **not annotated** in any way. However, if the negated statement is qualified with conditions in a way that implies that the proteins would normally form a complex, the statement is annotated as if the negation were absent (e.g. *"When A is phosphorylated, it fails to form a complex with B"*).
2. Statements expressed **speculatively** or with **hedging** expressions (e.g. *"may form a complex"*) are **annotated** identically to affirmative statements (in effect, **speculation and hedging are ignored**).

depletion of B", should be annotated as *"A regulation_of_proteolysis B"* and *"A negative_regulation B"* and not the combination of relations *"A positive_regulation B"* and *"A regulation_of_proteolysis B"*)

5. **Co-immunoprecipitation** can be used as an indicator of complex formation between two protein mentions
6. *"A regulation_of_proteolysis B"* does not necessarily imply *"A negative regulation of B"*, so this should be annotated with care
7. Post-translational modifications should **not** receive a binding annotation unless binding is clearly mentioned in context. PTMs imply transient interactions which will not be present in physical interaction databases, so they shouldn't be annotated as such. For an example of a corner case see Specific examples
8. The following are generally understood as implying *Complex formation*:
   - consitutive association
   - stable association
9. The following are generally understood as **NOT** implying *Complex formation*:
   - synergize
   - stabilize
10. If **part of a protein/complex** has the ability to **form a complex**, then the ability of the entire protein/complex to do the same can be extrapolated from that.
11. Subcellular localization is not annotated for *Complex formation* even if the structure is made of proteins.
12. When an entity is a substrate of another entity then the relation connecting them is **Catalysis of protein modification** e.g from 18312697



13. If a cyclin-kinase complex phosphorylates protein A then **Catalysis of phosphorylation** should be added both for the *kinase* **AND** the *cyclin* e.g from 11416152



14. Synthetic lethal interactions are genetic and thus are **NOT** annotated as Complex formation. Example from

Regulation

The initial idea behind using the general term *Regulation* is that we want to get in there as much as possible in terms of directionality for the edges. So, in order to do that, we will have to be a little bit more flexible with the hierarchy to include something very general that would allow us to have directionality, even in cases where we don't know the type of effect A has on B, but we know it is upstream. Also, we would have to be a bit more flexible with what we annotate in general and even in cases where we are not 100% sure, to add an annotation as long as we are pretty certain it is what the authors mean). The relevant GO term is Regulation of biological process. However, some more specific rules had to be added for this class for annotation consistency

1. **Protein X can do sth to Protein Y, in a Chemical Z dependent manner**: Z>regulates>X
2. **Protein X responds to Chemical Y**: Y>regulates>X
3. **Changes in protein level of protein A lead to changes in the molecular function of protein B**: A>regulates>B
4. **Protein A levels are regulated in a protein-B dependent manner**: B>Regulates>A (As they can be regulated through expression, i.e. positively or through degradation, i.e. negatively)
5. **Protein A participates in the assembly of Complex B**: A>Regulates>B
6. **Protein A structure is regulated by Protein B**: B>Regulates>A (not sure if the structure is regulated positively or negatively)
7. **Protein A and protein B gain a function upon forming a complex and this function affects another entity C**: A>regulates C AND B>regulates>C
8. **Targeting of protein A to a specific subcellular location is mediated/regulated/blocked/allowed (including secretion) by protein B**: B>regulates>A
9. **Protein A mediates the activation/inhibition of protein B**: A>Regulates>B

lo annotation:

1. **Protein A induces a PTM of protein B**: This is not annotated, unless changes in the levels/structure of protein B related to the PTM are described in the sentence, where Positive/Negative regulation should be annotated.

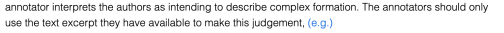
2. The same rule applies if the **PTM** is a **small protein (de)conjugation**. So there shouldn't be any extra annotations regarding regulations of protein levels for *Catalysis of ubiquitination*, *Catalysis of SUMOylation* or *Catalysis of NEDDylation*, unless clearly stated in a sentence.
3. **Proteasome** is the only complex that performs protein degradation, so **Regulation of degradation** is **NOT** annotated in that case (as we are annotating Proteins who regulate, not perform, the degradation process).

Positive regulation:

1. **Protein A is essential/required/needed in the assembly of Complex B**: A> Positively regulates>B
2. The word **chaperone** can be used as a hint to annotate Positive regulation
3. **Protein A is necessary for Protein B to perform a function**: A> Positive Regulation>B
4. **Protein A function is restored in the presence of Protein B**: B>Positively Regulates>A
5. **Protein A structure is maintained by Protein B**: B>Positively Regulates>A
6. **Protein A directly activates/inhibits protein B**: A>Positively/Negatively Regulates>B
7. **Protein A levels increase/decrease because of protein B**: B>Positively/Negatively Regulates>A

Named Entity annotation rules

1. Entity name mentions like *ubiquitin* or reporter genes (e.g. *GFP*) which are *GGPs* but are blacklisted by tagger, will be assigned the **blacklisted** attribute
2. Histones:
   - Tag *H2, H3* etc. when they appear standalone
   - Include *histone* in the span when it appears with one of the names (e.g. *histone H3*)
   - Tag *histone* as **Protein family or group** when it appears standalone.
   - We could then either go discontinuous or decomposed for mentions such as *histones H2A and H3*.
   - Methylated histones are also tagged as **GGP** even though tagger will not detect them (e.g. 25002145)
3. *Amino acid residues* should not be annotated as **chemical** when they are part of a polypeptide chain
4. *Glycosylphosphatidylinosiol* (GPI) should not be annotated as **chemical** as it cannot be a standalone chemical
5. Determiners like *the* should not be included in the entity span of **GGP**, **Protein-containing complex** and **Protein family or group**
6. Mutants of specific proteins will receive **GGP** annotations and an *Entity Attribute*: **Mutant**
7. In order for the annotated text to be as close as possible to the ideal NE annotation produced by the NER system, cases where only part-of **mutant names** are standalone entities, only these mentions should be annotated, e.g. in the following example from 18039934 **sam35** and **NOT** *sam35-2* is annotated as a ggp



An exception is when **mutant names are a single word**, and then they are annotated as one mutant entity e.g. **rex1Delta** in the following sentence from 16100378

8. Named entities that are part of antibodies should be annotated as the corresponding NE type and should receive a *Note: antibody*, e.g. from 20214800 ***TBD: antibodies that are one entity mention with "anti" included: http://ann.turkunlp.org:8088/index.xhtml#/string-relation-corpus/physical-interaction-dbs-full-texts-02/15520228_40***
9. rRNAs and tRNAs are currently annotated as **GGP** with **noncoding** attribute e.g. from 16100378 & from 19239245
10. **Fusion proteins** should be treated as two entities for the purposes of annotation and during the creation of the training dataset. These should get an *Entity Attribute*: **Fusion**. The reporter protein in fusion should get a note: **not tagged by tagger** if it is not detect by tagger. E.g. in this document **NRIF3** will receive an *Entity Attribute*: **Fusion** and **Gal4** will receive an *Entity Attribute*: **Fusion** and a *Note: not tagged by tagger* 11713274

11. *Domains* and other *protein regions* should **NOT** be annotated as *GGP*
12. *FLAG* and *6xHis* are polypeptide protein tags and should receive an *OOS* annotation, or should not be annotated at all.
13. *ATP* and *ADP* are annotated as **OOS**
14. *GTP* and *GDP* are annotated as **Chemicals** due to their signalling function
15. **Palmitate**, **Myristate**, **Acetate** and **ADP-ribose** should receive a **Chemical** annotation. When they are mentioned as [3H]chemical, only the chemical is annotated (e.g. **myristate** in [3H]myristate)
16. **prenyl** is a functional group and thus **NOT** annotated as chemical. Similarly **geranylgeranyl** and **farnesyl** are also protein groups and are **NOT annotated as Chemical**.

pecific rules for complexes/families and plural form annotations

- If a term is in Gene Ontology and is assigned a Protein-containing complex annotation then it is considered a Complex in this annotation effort,
- If a term is found in Gene ontology but it is NOT a **protein-containing complex**, then it will **NOT** be considered a *Complex* in this effort
- If a term is not at all present in Gene Ontology then other resources in the field will be used to decide whether it should be considered a *Complex* or not (e.g. Complex Portal, Reactome).
- There is no clear distinction in Gene Ontology between small (e.g. NF-kappaB) and large (e.g. Nuclear Pore) complexes. However, we have decided that very large complexes like the nuclear pore and the replication fork will **NOT** receive a complex annotation, as they fall out of scope of what is a complex for STRING.
- For cases where it is difficult to distinguish *family* from *domain* mentions, the field type in Pfam could be used to aid in making a decision (if available)

1. **Complexes** (**Protein-containing complex**) and protein **Families** (**Protein family or group**) are annotated if

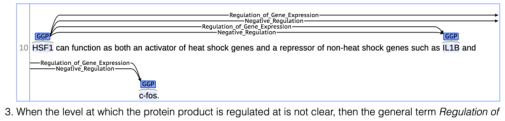
there is data verifying the existence of them. As a general rule for **Complexes** Gene Ontology - Cellular Component or another database of Complexes is enough to annotate an entity as Complex. For protein **Families** information is gathered from InterPro or Wikipedia or specific publications for very rare cases.

2. The words *"complex"*, *"family"* and *"group"* should **not** be part of the entity annotations.
3. Plural forms should be marked as "plural" in the annotation notes only if the protein(s) mentioned do not form a complex, do not belong in the same family or cannot be considered as forming a group.
4. When it comes to annotating enzyme families, extra attention should be taken when **enzyme X activity** is mentioned in text. In the cases, where what is discussed is the actual function of performing the catalysis, extra caution should be taken to annotate these as **Out-of-scope**.
5. The following complexes/families/groups should get the corresponding annotations (Exception: if in specific organisms a family is only represented by one protein member and in that organism the names of the family and the protein entity **coincide**, then the context should be used to adress how that mention should be annotated).
   - 2-oxo acid decarboxylase: Complex
   - acetyl-CoA carboxylase: Complex
   - acetyltransferase, acyltransferase: Too broad, **DO NOT annotate as family**
   - Actin: Family
   - adenylyl cyclase: Family
   - ADP ribosylation factors – ARF: Family
   - ADP-ribosyltransferase, ART: Family, ARTC
   - Akt, Protein Kinase B: Family
   - AMPK: Complex
   - AP-1: Complex **or** Family depending on the context
   - AP-2: Complex
   - Asialoglycoprotein receptor: Subfamily of lectins: Family
   - Arf-like small GTPases, Arl: Family
   - B-cell antigen receptor (BCR): Complex
   - Bone morphogenic protein BMP: **Not a family**
   - Calcium-independent phospholipases A2, iPLA2s: Family
   - Cytosolic phospholipases A2, cPLA2: Family
   - carboxylase: Family
   - casein: Family
   - casein kinase 1: Family
   - casein kinase 2: Family
   - caspase: Family
   - Cav beta: Family?
   - caveolin: Family
   - CELO: chicken embryo lethal orphan (CELO) virus: **NOT a family**
   - chaperones: OOS (group of proteins, **NOT a family**)
   - Cohesin: Complex
   - collagen: Complex
   - Condensin: Complex
   - CREB/ATF: Protein family
   - cullin: Family
   - Cullin-RING ligases (CRL), cullin-based ubiquitin ligases, Cullin-
   - cyclic nucleotide phosphodiesterase, 3',5'-cyclic-nucleotide pho
   - Cystatin: Family
   - DHHC palmitoyltransferase: Family
   - dihydrolipoyl acyltransferase:Complex
   - dihydrolipoyl dehydrogenase: Complex
   - DnaJ-family chaperone: Family
   - Dynamin: Family
   - EGR: Protein family or group
   - eIF2, eukaryotic initiation factor-2: Complex
   - ErbB, epidermal growth factor (EGF)-like growth factors: Family
   - Ets: Protein family or group
   - F-box , F-box protein, FBP: Family
   - Farnesyltransferase: Complex
   - Fe-protein: Complex
   - Fetuin: Family
   - Formin: Family
   - Frizzled: Family
   - Furin-like convertase: Family maybe?
   - G protein, Guanine nucleotide binding protein: Family
   - G-protein alpha: Family
   - GATA: Protein family or group
   - GDI (guanosine diphosphate dissociation inhibitor): Family
   - GEF, Guanine nucleotide exchange factor: **Not a Family**
   - Geranylgeranyltransferase I: Complex
   - Geranylgeranyltransferase II: Complex
   - [glutamate receptor: Not a family, metabotropic glutamate receptor & ionotropic glutamate receptor are families
   - Glutamine synthetase: **NOT** a Family 3 classes structurally but not sequence related
   - glycoprotein hormone receptor: Family](https://www.ebi.ac.uk/interpro/entry/InterPro/IPR002131/)
   - glycosyltransferases: **Not a family**, there are 65 families of those enzymes
   - GPCRs, G-protein coupled receptors: Family
   - Green fluorescent protein, GFP: Family
   - GTPase: Family
   - hemoproteins: **NOT a family**
   - Histatin: Family
   - histone: Family
   - Histone deacetylase (HDAC): Family
   - High mobility group proteins, HMG: Family
   - Huntington Interacting Protein, HIP: **Not a family**
   - Interleukins: Look here
   - Ikappa: Family
   - [IgG: Family?]
   - inositol phosphatase: Family
   - Integrin-related receptors: Family
   - IκB kinase (IKK)-related kinases, IKK: Family
   - IκB kinase: Complex
   - kappa B: Protein-containing complex
   - laminin: Family
   - lectin: Family
   - LPT: Protein-containing complex, maybe
   - MAPK, ERK: Family, Family
   - mucin: Family
   - myosin: Family
   - myosin heavy chain: Family
   - myosin light chain: Family
   - myosin phosphatase: **NOT a family**
   - myosin phosphatase targetting subunit (MYPT): Family (read description)
   - N-glycosylase apurinic/apyrimidinic (AP) lyase: Family
   - N-type calcium channel: Family
   - NADPH oxidase (NOX): Family
   - Neurokinin receptor, Tachykinin receptor: Family
   - NF-kappaB, NF-kappaB/Rel: Protein-containing complex
   - NFAT: Protein family or group
   - nitric oxide synthase: Family
   - nitrogenase: Family
   - NMDA receptor: Family or Complex? (I now have it as Family)
   - nuclear receptor: Family, Evolution
   - nuclear pore: **Not a Complex** (the structure is too big)
   - nucleosome: Complex
   - Oxoglutarate dehydrogenase: Complex
   - p66: Family
   - Palmitoyl-transferase: **Not a family**
   - Ras GTPase, Ras/Rap-related GTPase, Ras, Ras/Rab: Family
   - Rap GTP-binding proteins, Rap
   - PDGF, Platelet-derived growth factor: Family
   - Phospholipases A2: **NOT a family**
   - phosphorylase: Family (not phosphorylase a)
   - PI3K, phosphatidylinositol 3'-OH kinase, phosphatidylinositol 3-kinase: Family
   - PIPLC, phosphatidylinositol-specific phospholipase C: Family
   - poly (ADP-ribose) polymerase: Family
   - Proline-rich protein, PRP: Family
   - Proprotein convertase PPC PC: Family
   - Proteasome (in all its forms, e.g. 26S proteasome): Complex
   - protein inhibitor of activated STAT (PIAS): Family
   - Protein Kinase B: Family
   - Protein Kinase C PKC: Family
   - proteoglycans: **NOT a family**
   - puryvate dehydrogenase: Complex
   - Rab: Family
   - Rab geranylgeranyl transferase: Complex
   - Rac: Family
   - Rel: Protein family or group
   - Rep (Rab escort protein): Family
   - replication fork: **Not a Complex** (the structure is too big)
   - RHL, rat hepatic lectins: Family
   - Rho (GTPases): Family
   - ribosomal protein S6 kinase: Family
   - ROP GTPases, Rho of plants, ROP: Family
   - Secreted phospholipases A2, SPLA2: Family
   - selectin: Family
   - Septin: Family
   - serine protease: **NOT a family**
   - sirtuin: Family
   - small GTP-binding protein: Family
   - Small GTPase: Family
   - small heat shock proteins, small HSPs, sHSP: Family
   - SMC: Protein-containing complex
   - SNARE: Family
   - SOS, son of sevenless: Family
   - sphingosine 1-phosphate receptor S1P receptors: Class of G
   - Src: Family
   - steroid hormone receptor: **NOT a family**
   - SUMO: Family
   - SUMO E2 conjugating enzyme: Family (only one protein, ann
   - SUMO E3 ligase: **DO NOT annotate as family** To date, three u suggested to have SUMO E3 ligase activity
   - Synaptotagmin: Family
   - tetraspanin: Family
   - TFIID: Protein-containing complex
   - TNF: Protein family OR TNF: GGP
   - TNFR: Protein family or group
   - transducin: Complex
   - Triosephosphate isomerase: Family
   - trypsin: Family
   - Tyrosylprotein sulfotransferase: **NOT a family**
   - The four families of ubiquitin ligases should be annotated as follows:
     1. HECT E3 ligase: Family
     2. U-box E3 ligase: Family
     3. PHD-finger E3 ligase: Family
     4. RING-finger E3 ligase: Family, members of which are:
        1. SKP1–CUL1–F-box-protein (SCF and all its forms with the F-box protein defined): Complex
        2. anaphase-promoting complex (cyclosome, APC/C): Complex
     5. RING-finger E3 ligase: Family, members of which are:
        - Ubiquitin-activating enzyme E1: Family
        - Ubiquitin-conjugating enzyme E2: Family
        - Vascular endothelial growth factor, VEGF: Family
        - Vesicle-associated membrane protein, VAMP: Family
        - Voltage-gated calcium channels, calcium channel: Family
        - Wnt: Family
        - WRKY: Family
        - YCK: Family
   - Annotations should be applied to all variants: e.g. **NF kappaB**, **NF-kappaB**, **NFkappaB** should all be marked as **Protein-containing complex**
   - Entities that still need resolution on whether they should be annotated as **Protein Family** or not:
     - cysteine protease: http://localhost:9878/index.xhtml#/exhaustive-ptm-corpus-cleaned/15769255
     - Estrogen Receptor (ER): Family
   - GDP/GTP exchange proteins
   - MEK - MKK - MAPKK

Specific Examples discussed

1. Instances of binding and phosphorylation should be seperately annotated as two events when binding is clearly mentioned in text.

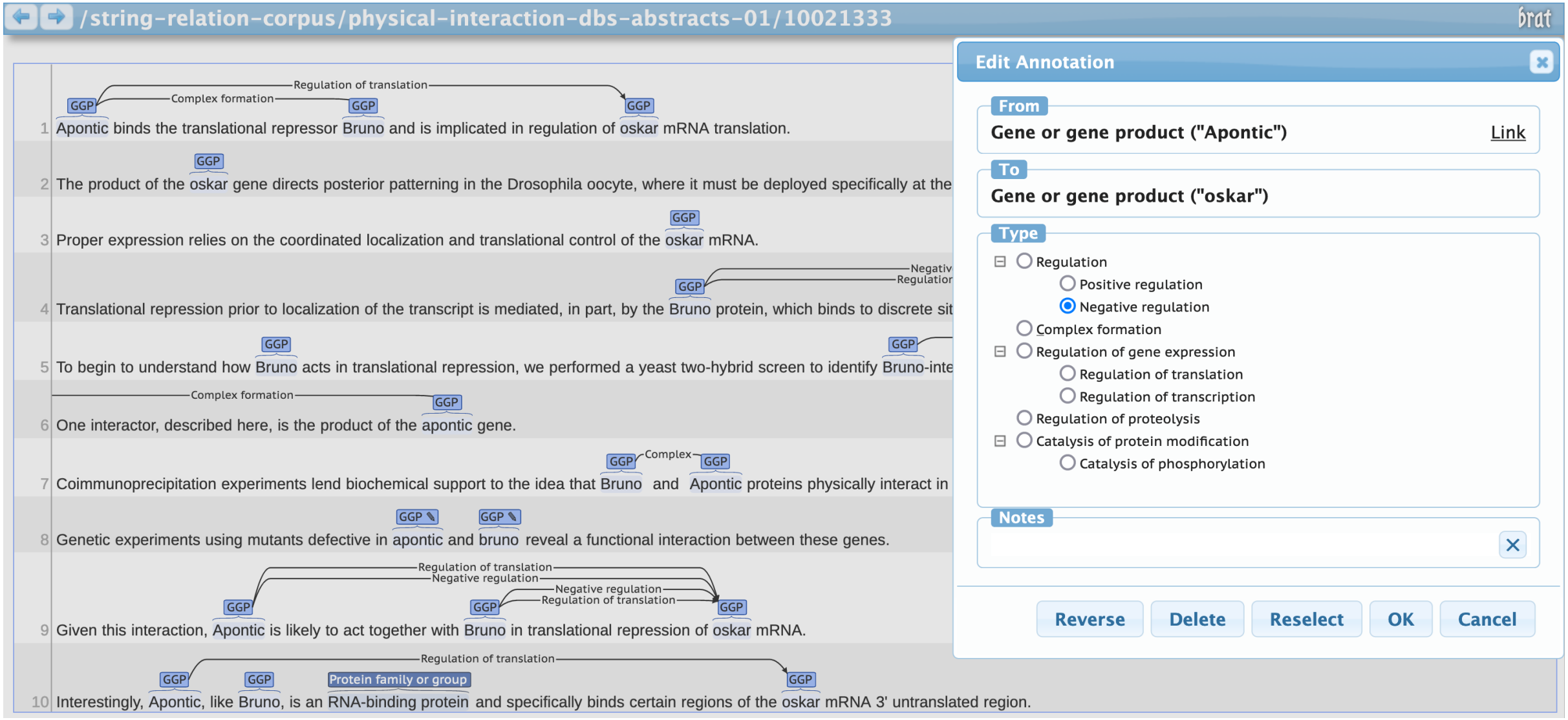
2. When *"regulation of expression"* is mentioned in text, if the annotator suspects that the intended meaning of the authors is **Regulation of Transcription** based on the context of the document they have available, then the annotator should annotate **Regulation of Gene Expression**, with a **Note: "Potentially Regulation of Transcription"**.



3. When the level at which the protein product is regulated at is not clear, then the general term *Regulation of Gene Expression* should be used.

And there is more...

https://brat.nlplab.org/



P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, J. Tsujii, *BRAT: a web-based tool for NLP-assisted text annotation.* Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012

Identify relationships based on context

What relationships?

v11.5: Identify **physical interactions** based on context

... interaction of A and B ...

... A-B complex ...

Identify physical interactions based on context

... A binds to B ...

... A/B heterodimer ...

... A interacts with B ...

Where to get textual information from?

Physical interactions (BioGRID, IntAct, MINT)

Previous BioNLP shared tasks annotation datasets

Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, et al. *The BioGRID interaction database: 2019 update*. Nucleic Acids Research, 2019
Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases*. Nucleic Acids Research, 2014
Pyysalo S, Ohta T, Rak R, Rowley A, Chun HW, Jung SJ, et al. *Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013*. BMC Bioinformatics 2015
Kim JD, Pyysalo S, *BioNLP Shared Task,* Encyclopedia of Systems Biology, 2013

Corpus (version 1)

Physical interactions (BioGRID, IntAct, MINT)

400 documents (300 abstracts + 100 full-text paragraphs)

Previous BioNLP shared tasks annotation datasets

137 documents enriched in complex formation events

Training dataset (version 1)

Physical interactions (BioGRID, IntAct, MINT)
400 documents (300 abstracts + 100 full-text paragraphs)

Previous BioNLP shared tasks annotation datasets
137 documents enriched in complex formation events

Total no of relationships:
Positive (complex formation): 1607
Negative (Not a complex): 10392

Training dataset (version 1)

Total no of relationships:
Positive (complex formation): 1607
Negative (Not a complex): 10392

Shuffle data &
reduce redundancy

Positive (complex formation): 1298
Negative (Not a complex): 6385

Training dataset (version 1)

Total no of relationships:
Positive (complex formation): 1607
Negative (Not a complex): 10392

Shuffle data &
reduce redundancy

Labelled dataset
to fine tune
BioBERT model

Positive (complex formation): 1298
Negative (Not a complex): 6385

Training dataset (version 1)

Total no of relationships:
Positive (complex formation): 1607
Negative (Not a complex): 10392

Shuffle data &
reduce redundancy

Labelled dataset
to fine tune
BioBERT model

Positive (complex formation): 1298
Negative (Not a complex): 6385

Even less
depending
on the max
seq len

Test various hyperparameters
(grid search)

Models: **BioBERT_base, BioBERT_large**
Batch size: 32, 64, 256, 512, 1024, 2048
Learning rate: 5E-6, 1E-5, 2E-5
Epochs: 5, 10, 20
Maximum sequence length: 96 (large), 256 (base)

Test various hyperparameters

Best combination

Batch size: 32, 64, 256, 512, **1024**, 2048
Learning rate: **5E-6**, 1E-5, 2E-5
Epochs: 5, 10, **20**
Maximum sequence length: 96 (large), **256 (base)**

Test various hyperparameters

Best combination

Batch size: 32, 64, 256, 512, **1024**, 2048

Learning rate: **5E-6**, 1E-5, 2E-5

Epochs: 5, 10, **20**

Maximum sequence length: 96 (large), **256 (base)**

F1-score: 83.1%

Training dataset (version 1)

Total no of relationships:
Positive (complex formation): 1607
Negative (Not a complex): 10392

Shuffle data &
reduce redundancy

The culprit?

Labelled dataset
to fine tune
BioBERT model

Positive (complex formation): 1298
Negative (Not a complex): 6385

Use Best model

Run Prediction on all pairs of proteins in the literature that are in the same sentence (large scale run)

Benchmarking and Score calibration for physical interactions against **Complex Portal**

Physical interactions mode for version 11.5 of STRING

Check our paper for more details

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., and von Mering, C. (2021). *The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets.* Nucleic Acids Res, 49(D1), D605-612

v12: Identify **physical interactions** based on context (again)

Corpus (version 2)

Physical interactions (BioGRID, IntAct, MINT)

~~400~~ 800 documents (~~300~~ 400 abstracts + ~~100~~ 400 full-text paragraphs)

Previous BioNLP shared tasks annotation datasets

137 documents enriched in complex formation events

Regulatory Interactions (Reactome)
300 documents (abstracts)

Genetic Interactions (BioGRID)
50 documents (abstracts)

Total number of documents in Corpus v2: 1,287

Introduction of new Named Entity types

Gene or Gene Product (Protein → GGP)

Protein Family or Group

Protein-containing Complex

Chemical

Out-of-scope

New training/dev/test split

New train/dev/test split

Document-based and not sentence-based

60% train / 20% dev / 20% test doc split
instead of 70% train / 10% dev / 20% test

avoid data leakage to dev set

60/20/20 split also applies to the complex formation relationships (not only the documents)

Total number of positive (complex formation) relationships in corpus v2

Training set: 2,131

Dev set: 650

Test set: 644

Ignoring equivalent entities

Total number of complex formation relationships in training set?

Total number of complex formation relationships?

Depends on experimental setup

Exp	Training pairs (Positive/Negatives)	Masked entities	Not masked entities	Comment	Positive pair count in train	Negative pair count in train
1-A	**(X, Y) X, Y ∈ {Prot, Prot_BL, Complex, Chemical, Family)**		**Prot, Prot_BL, Complex, Chemical, Family**	maximum training data No masking	2,489	201,673
1-B	**(X, Y) X, Y ∈ {Prot, Prot_BL, Complex, Chemical, Family)**	**Prot, Prot_BL, Complex, Chemical, Family**		maximum training data maximum masking	2,489	201,673
2	**(Prot, Prot)**	**Prot**	**Prot_BL, Complex, Chemical, Family**	minimum training data minimum masking most similar to clean dev	1,961	113,527
3	**(X, Y) X, Y ∈ {Prot, Prot_BL, Complex}**	**Prot, Prot_BL, Complex, Chemical, Family**		more training data compared to Experiment #2 maximum masking	2,117	138,961
4	**(X, Y) X, Y ∈ {Prot, Prot_BL, Complex}**	**Prot, Prot_BL, Complex,**	**Chemical, Family**	more training data compared to Experiment #2 less masking compared to Experiment #3	2,117	138,961

Test A LOT of hyperparameters

Models: **BERT_X**
Batch size: 16
Learning rate: 2e-5 3e-5 5e-5 5e-6 5e-7
Epochs: 2, 3, 5, 8, 10, 12
Maximum sequence length: 64, 128, 256, 512
Random seed index: 0, 1, 2, 3
Optimizer: adam, adam_warmup
Strategy: MARK, MASK

x50

Test various hyperparameters
(final grid search)

Models: BioBERT_base, **RoBERTa-large-PM-M3-Voc-hf**
Experimental setup: 1-A, **1-B (MASK),** 2, 3 , 4
Batch size: 5
Learning rate: **3E-6**, 4E-6, 5E-6
Epochs: 7, 8, 9, 10, **11**, 12
Maximum sequence length: **128**, 144, 160, 176, 192

Mean F1-score: 84.25%
(std = 0.82%)

Not comparable with previous F-score

Best model

Run Prediction on all pairs of proteins in the literature that are in the **same paragraph** (large scale run)

Benchmarking and Score calibration for physical interactions against **Complex Portal** and for functional against KEGG

Physical interactions for version 12 of STRING

Check out the latest paper

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. (2022). *The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic Acids Research, Online ahead of print

v12.5: Identify **physical interactions and typed directed relationships** based on context

What relationships?

Relationship types

```
Regulation
  Positive regulation
  Negative regulation

Complex formation

Regulation of gene expression
  Regulation of transcription
  Regulation of translation

Regulation of degradation

Catalysis of Post-translational modification
    Catalysis of small protein conjugation/removal
        Catalysis of small protein conjugation
            Catalysis of Ubiquitination
            Catalysis of SUMOylation
            Catalysis of Neddylation
            Other catalysis of small protein conjugation
        Catalysis of small protein removal
            Catalysis of Deubiquitination
            Catalysis of DeSUMOylation
            Catalysis of Deneddylation
            Other catalysis of small protein removal
```

- Catalysis of Post-translational modification
  - Catalysis of phosphoryl group conjugation/removal
    - Catalysis of Phosphorylation
    - Catalysis of Dephosphorylation
  - Catalysis of small molecule conjugation/removal (excluding phosphoryl group)
    - Catalysis of small molecule conjugation
      - Catalysis of Methylation
      - Catalysis of Acylation
        - Catalysis of Acetylation
        - Catalysis of Palmitoylation
        - Catalysis of Myristoylation
      - Catalysis of lipidation
        - Catalysis of prenylation
          - Catalysis of farnesylation
          - Catalysis of geranylgeranylation
      - Catalysis of Glycosylation
      - Catalysis of ADP-ribosylation
      - Other catalysis of small molecule conjugation
    - Catalysis of small molecule removal
      - Catalysis of Demethylation
      - Catalysis of Deacylation
        - Catalysis of Deacetylation
        - Catalysis of Depalmitoylation
      - Catalysis of Deglycosylation
      - Other catalysis of small molecule removal

Corpus (version 3)

Physical interactions (BioGRID, IntAct, MINT)

800 documents (400 abstracts + 400 full-text paragraphs)

Previous BioNLP shared tasks annotation datasets

137 documents enriched in complex formation events

Regulatory Interactions (Reactome)

300 documents (abstracts)

Genetic Interactions (BioGRID)

50 documents (abstracts)

Exhaustive PTM corpus

234 documents enriched in post-translational modification events

Total number of documents in Corpus v3: 1,521

Total number of “positive” relationships in corpus v3

Training set: 5,508 (2,362 Complex formation)

Dev set: 1,778 (741 Complex formation)

Test set: 1,735 (720 Complex formation)

Experiments?

Initial experiments: not on all relationship types

Initial experiments: not on all relationship types

Only on those with more than 40 relationships in training set

Setup: from binary to multi-label

Initial experiments with small grid search

Models: **RoBERTa-large-PM-M3-Voc-hf**
Experimental setup: 1-A (MARK)
Batch size: 5
Learning rate: 2e-6, **3E-6**, 5E-6
Epochs: 8, 9, 10, **12**
Maximum sequence length: **128**

Best mean F1-score:
71.88% (std = 0.48%)

Future directions

Corpus v4: Expand corpus to 2,000 documents

(i.e. add 479 more documents)

Corpus v4: Add more documents for PTM relationships

Corpus v5: Expand corpus to 3,000 documents

Corpus v5: Add more full-text excerpts

Run hyperparameter grid searches with final corpus

Different experimental setups

Different experimental setups

Main experimental question: how to treat relationships without enough examples to train/test?

Run predictions on all protein pairs

But before adding to STRING v12.5

New benchmark set for regulatory typed relationships

In the meantime: Create network of directed typed relationships from deep learning-based text mining

Use proteomics datasets to check whether one can infer upstream regulators based on downstream events

Strange Planet
by Nathan Pyle

Thank you for your attention!

JensenLab

Lars Juhl Jensen
Stefano Roncelli
Mikaela Koutrouli
Esmaeil Nourani
Oana Palasca
Rebecca Kirsch
Katerina Nastou
Maud Kerstholt
Dewei Hu
Nadezhda Tsankova Doncheva
Foteini Aktypi

The Novo Nordisk Foundation Center for Protein Research

TurkuNLP lab

Sampo Pyysalo
Farrokh Mehryary

TURKUNLP .ORG

TURUN YLIOPISTO

novo nordisk fonden

ACADEMY OF FINLAND

computerome

CSC

Contact: katerina.nastou@cpr.ku.dk