جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

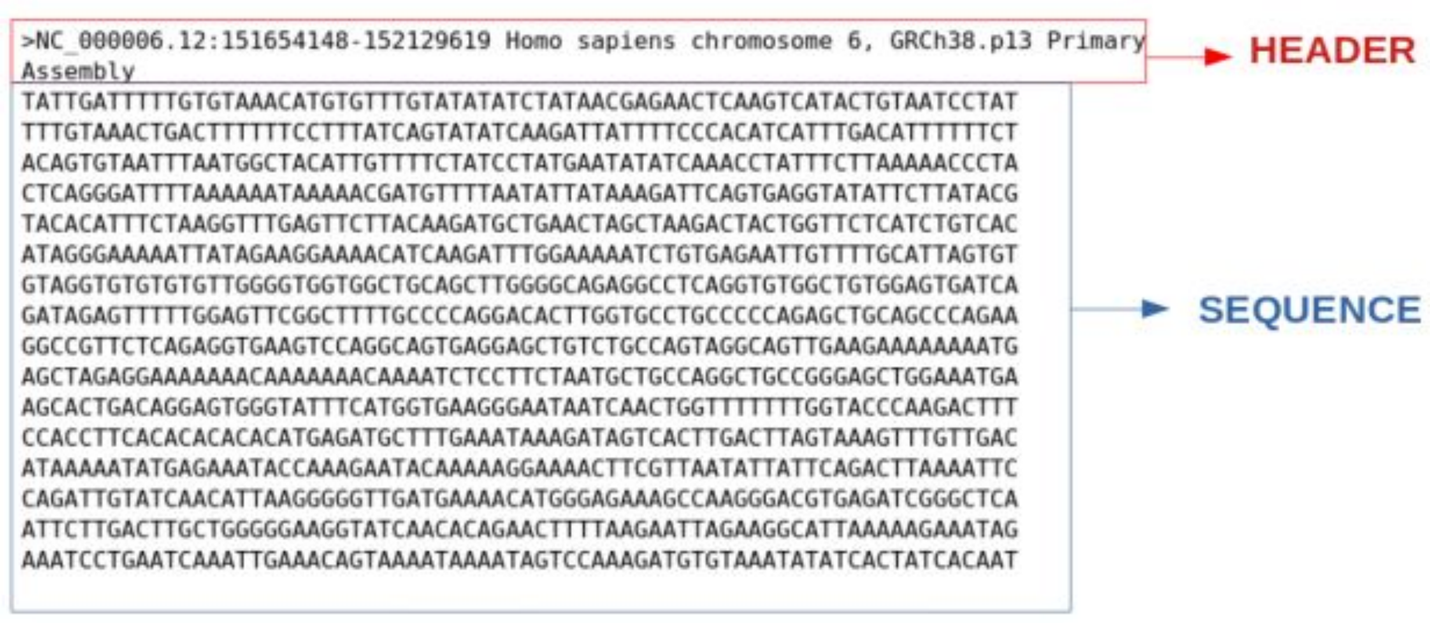
أكاديمية كاوست
KAUST ACADEMY

Applications of AI in Medicine & Biology

King Abdullah University of Science and Technology (KAUST)
KAUST Academy

Data Representation in Biology

- Sequences:
  - FASTA: DNA/RNA/Protein sequences



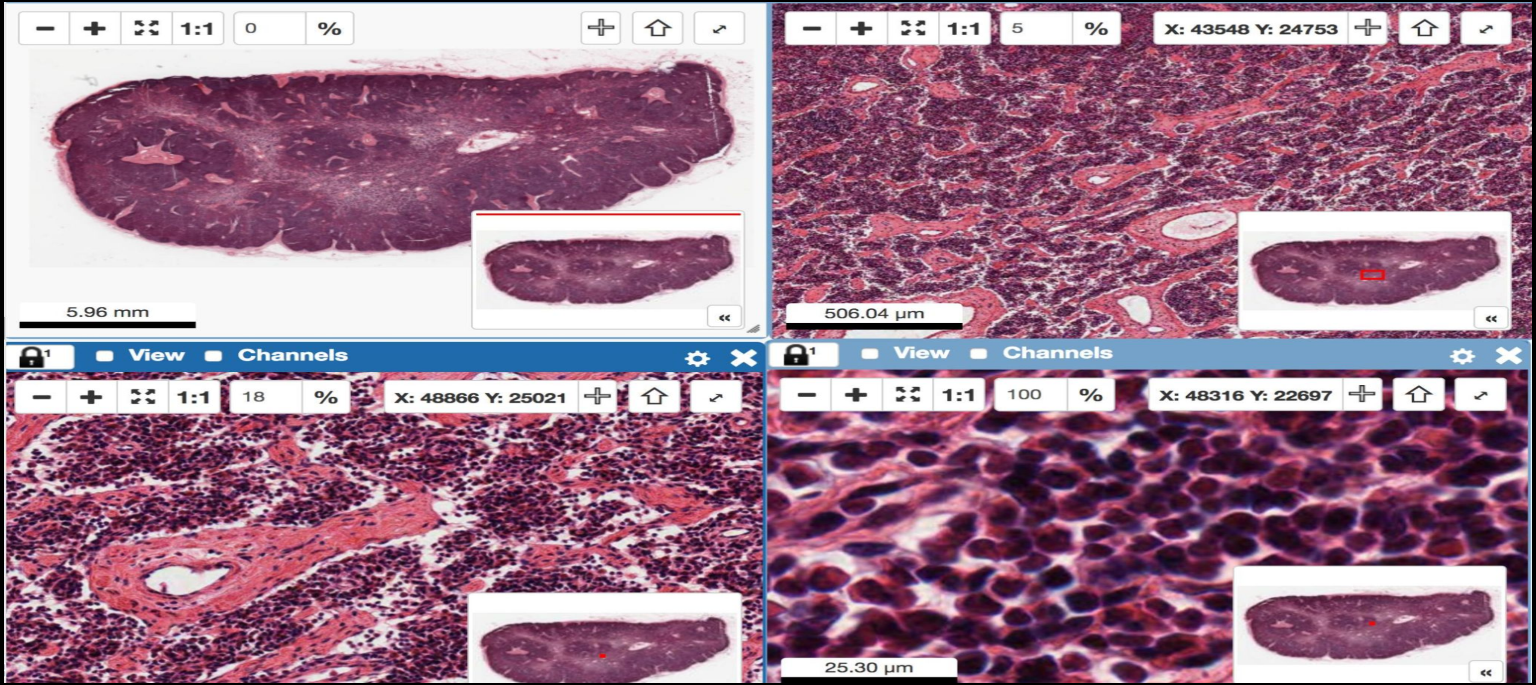
```
>NC_000006.12:151654148-152129619 Homo sapiens chromosome 6, GRCh38.p13 Primary
Assembly
TATTGATTTTTGTGTAAACATGTGTTTGTATATATCTATAACGAGAACTCAAGTCATACTGTAATCCTAT
TTTGTAAACTGACTTTTTTCCTTTATCAGTATATCAAGATTATTTTCCCACATCATTTGACATTTTTTCT
ACAGTGTAATTTAATGGCTACATTGTTTTCTATCCTATGAATATATCAAACCTATTTCTTAAAAACCCTA
CTCAGGGATTTTAAAAAATAAAAACGATGTTTTAATATTATAAAGATTCAGTGAGGTATATTCTTATACG
TACACATTTCTAAGGTTTGAGTTCTTACAAGATGCTGAACTAGCTAAGACTACTGGTTCTCATCTGTCAC
ATAGGGAAAAATTATAGAAGGAAAACATCAAGATTTGGAAAAATCTGTGAGAATTGTTTTGCATTAGTGT
GTAGGTGTGTGTGTTGGGGTGGTGGCTGCAGCTTGGGGCAGAGGCCTCAGGTGTGGCTGTGGAGTGATCA
GATAGAGTTTTTGGAGTTCGGCTTTTGCCCCAGGACACTTGGTGCCTGCCCCCAGAGCTGCAGCCCAGAA
GGCCGTTCTCAGAGGTGAAGTCCAGGCAGTGAGGAGCTGTCTGCCAGTAGGCAGTTGAAGAAAAAAAATG
AGCTAGAGGAAAAAAACAAAAAAACAAAATCTCCTTCTAATGCTGCCAGGCTGCCGGGAGCTGGAAATGA
AGCACTGACAGGAGTGGGTATTTCATGGTGAAGGGAATAATCAACTGGTTTTTTTGGTACCCAAGACTTT
CCACCTTCACACACACACATGAGATGCTTTGAAATAAAGATAGTCACTTGACTTAGTAAAGTTTGTTGAC
ATAAAAATATGAGAAATACCAAAGAATACAAAAAGGAAAACTTCGTTAATATTATTCAGACTTAAAATTC
CAGATTGTATCAACATTAAGGGGGTTGATGAAAACATGGGAGAAAGCCAAGGGACGTGAGATCGGGCTCA
ATTCTTGACTTGCTGGGGGAAGGTATCAACACAGAACTTTTAAGAATTAGAAGGCATTAAAAAGAAATAG
AAATCCTGAATCAAATTGAAACAGTAAAATAAAATAGTCCAAAGATGTGTAAATATATCACTATCACAAT
```

Data Representation in Biology

- Sequences:
  - FASTA: DNA/RNA/Protein sequences
- ⇒ NLP Problem

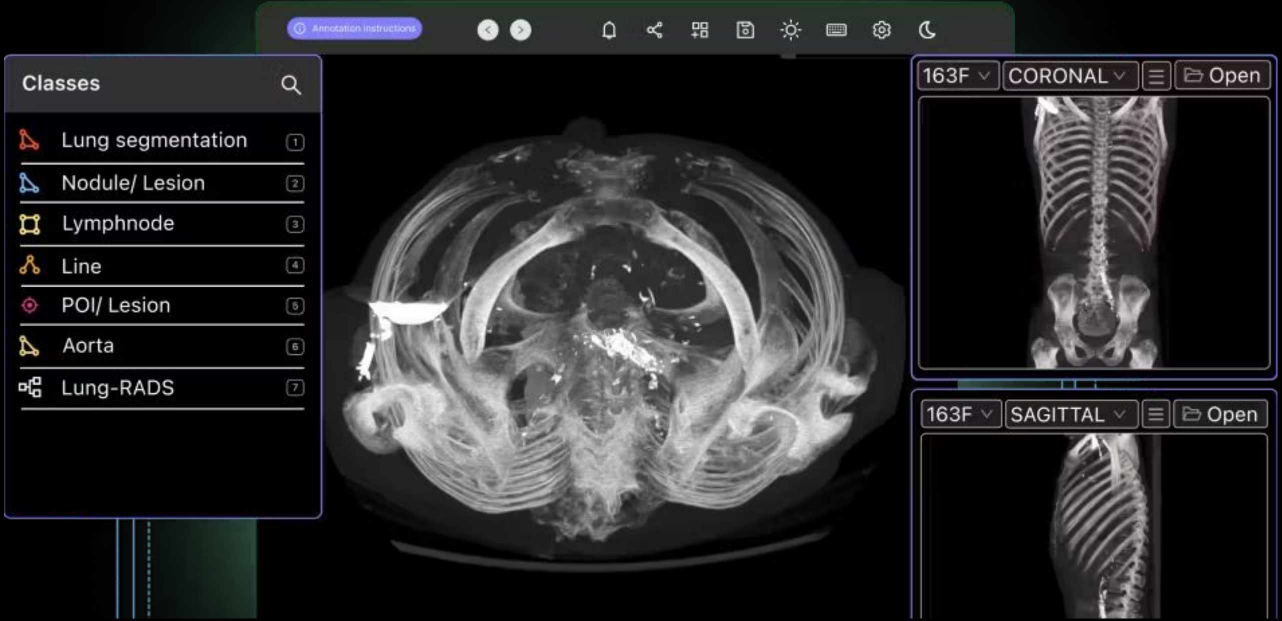
Data Representation in Biology

- Images:
    - TIFF: microscopic images.



Data Representation in Biology

- Images:
  - DICOM: radiological images (e.g. X-Ray, MRI,...).



Data Representation in Biology

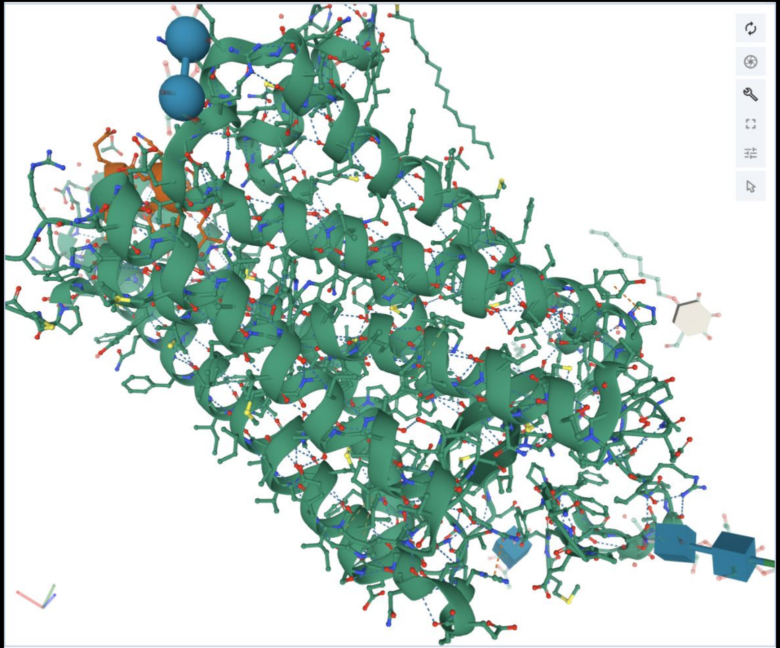
- Images:
  - TIFF: microscopic images (2D/3D).
  - DICOM: radiological images (e.g. X-Ray, MRI,...) (2D/3D).
  - JPEG/PNG after pre-processing.

Data Representation in Biology

- Images:
  - TIFF: microscopic images.
  - DICOM: radiological images (e.g. X-Ray, MRI,...).
  - JPEG/PNG after pre-processing.
- ⇒ Vision Problem

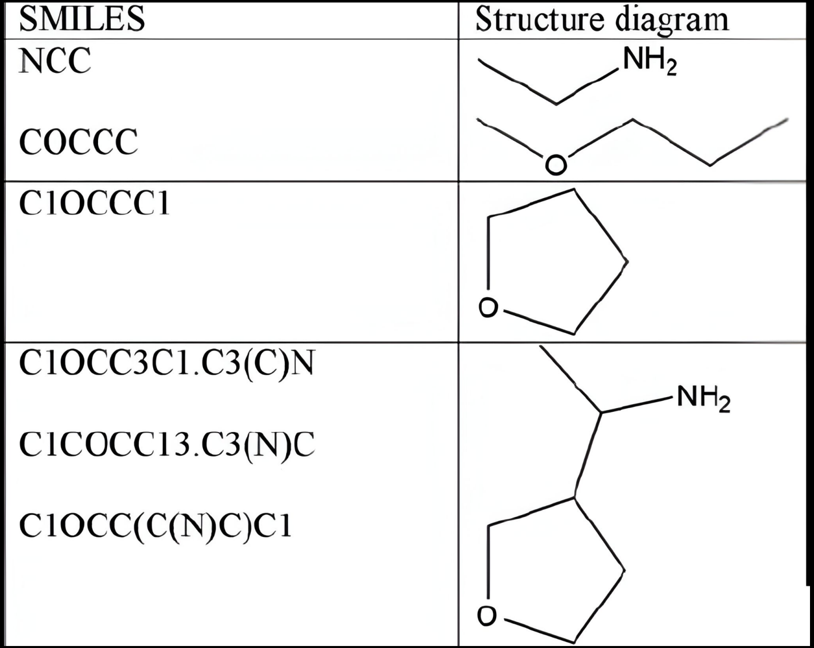
Data Representation in Biology

- 3D-Structures:
  - PDB: 3D coordinates per atom.



Data Representation in Biology

- Molecular Strings:
  - SMILES: representing chemical structures as text strings.



SMILES	Structure diagram
NCC COCCC	NH_2 O
C1OCCC1	O
C1OCC3C1.C3(C)N C1COCC13.C3(N)C C1OCC(C(N)C)C1	NH_2 O

Data Representation in Biology

- 3D-Structures:
  - PDB: 3d coordinates per atom.
- Molecular Strings:
  - SMILES: representing chemical structures as text strings.
- ⇒ Graphs / NLP Problem

Data Representation in Biology

- Let’s have a look at some problems…

1) Human Protein Atlas Image Classification

- A given protein can be in one, several, or different subcellular compartments depending on cell type and conditions

1) Human Protein Atlas Image Classification

- A given protein can be in one, several, or different subcellular compartments depending on cell type and conditions
- Goal: Develop model capable of classifying the subcellular compartments that have proteins using microscope images.

1) Human Protein Atlas Image Classification

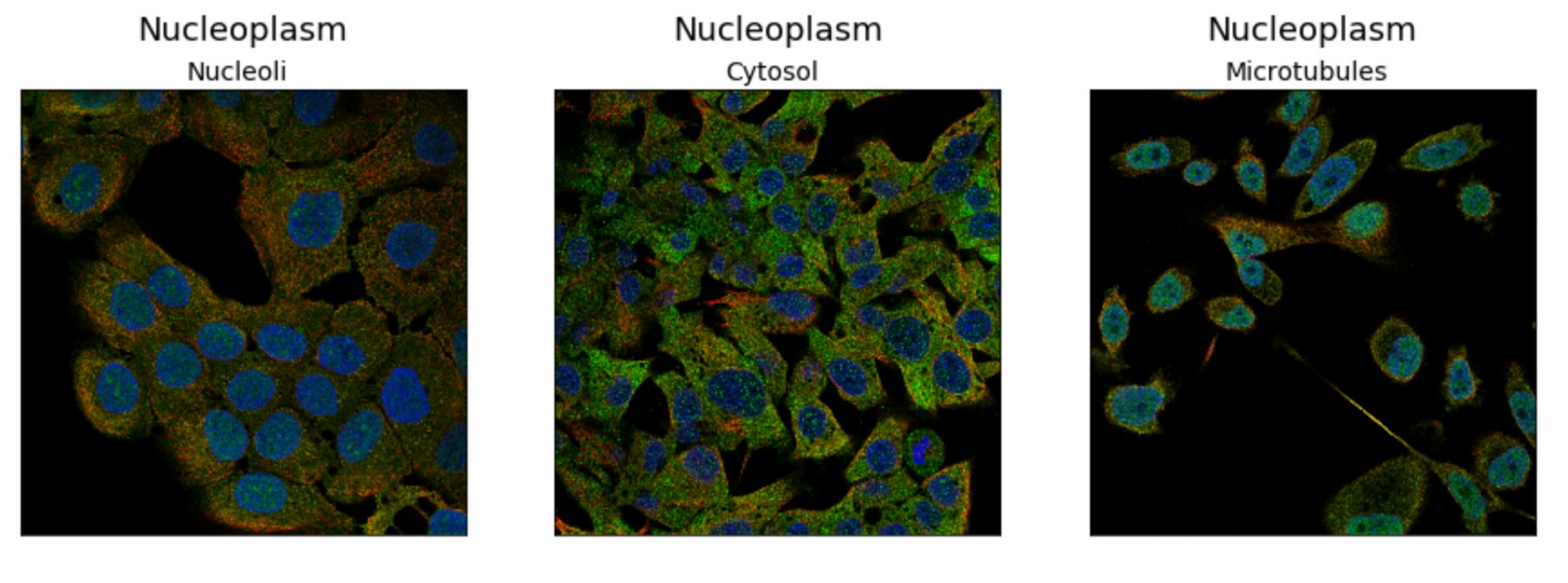
- A given protein can be in one, several, or different subcellular compartments depending on cell type and conditions
- Goal: Develop model capable of classifying the subcellular compartments that have proteins using microscope images.
- This model will be used to build a tool integrated with a smart-microscopy system to identify a protein's location(s) from images.

1) Human Protein Atlas Image Classification

- A given protein can be in one, several, or different subcellular compartments depending on cell type and conditions
- Goal: Develop model capable of classifying the subcellular compartments that have proteins using microscope images.
- This model will be used to build a tool integrated with a smart-microscopy system to identify a protein's location(s) from images.
- Data Modality: A mix of 2048x2048 and 3072x3072 2D TIFF images.

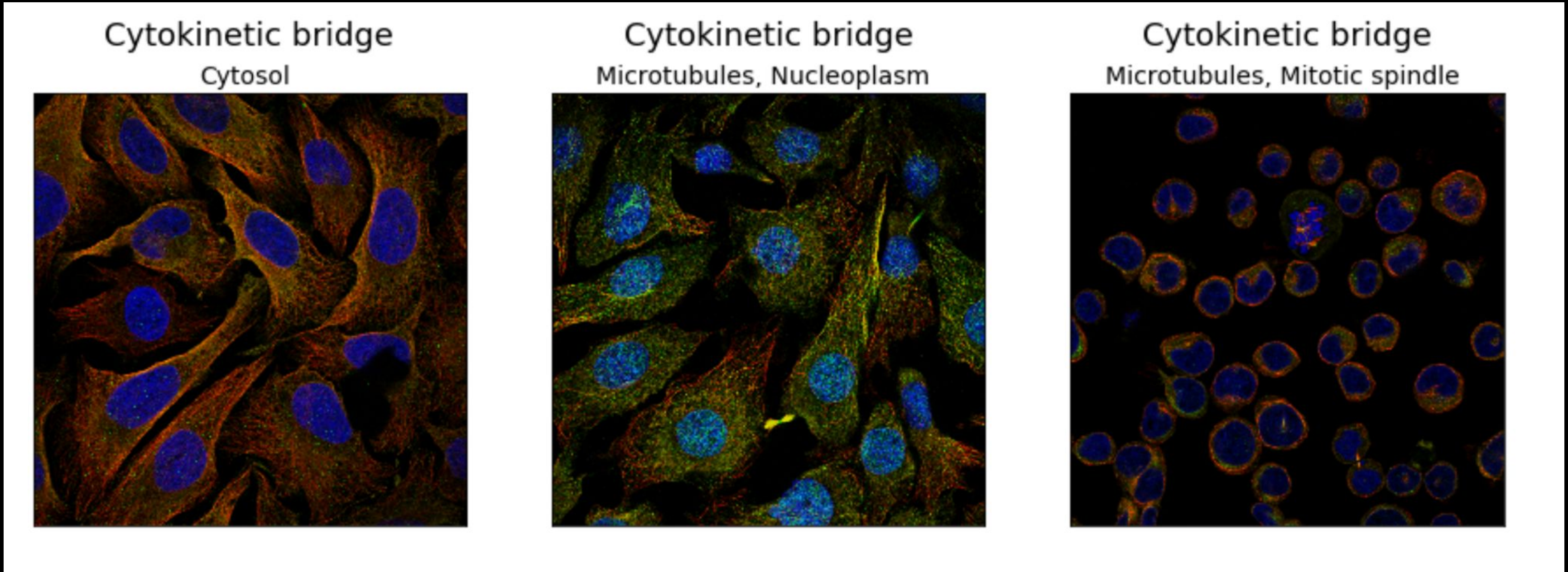
1) Human Protein Atlas Image Classification

Green = the protein of interest; blue/red/yellow are constant reference markers (nucleus, microtubules, ER) to help you judge where the green signal sits.



1) Human Protein Atlas Image Classification

Green = the protein of interest; blue/red/yellow are constant reference markers (nucleus, microtubules, ER) to help you judge where the green signal sits.



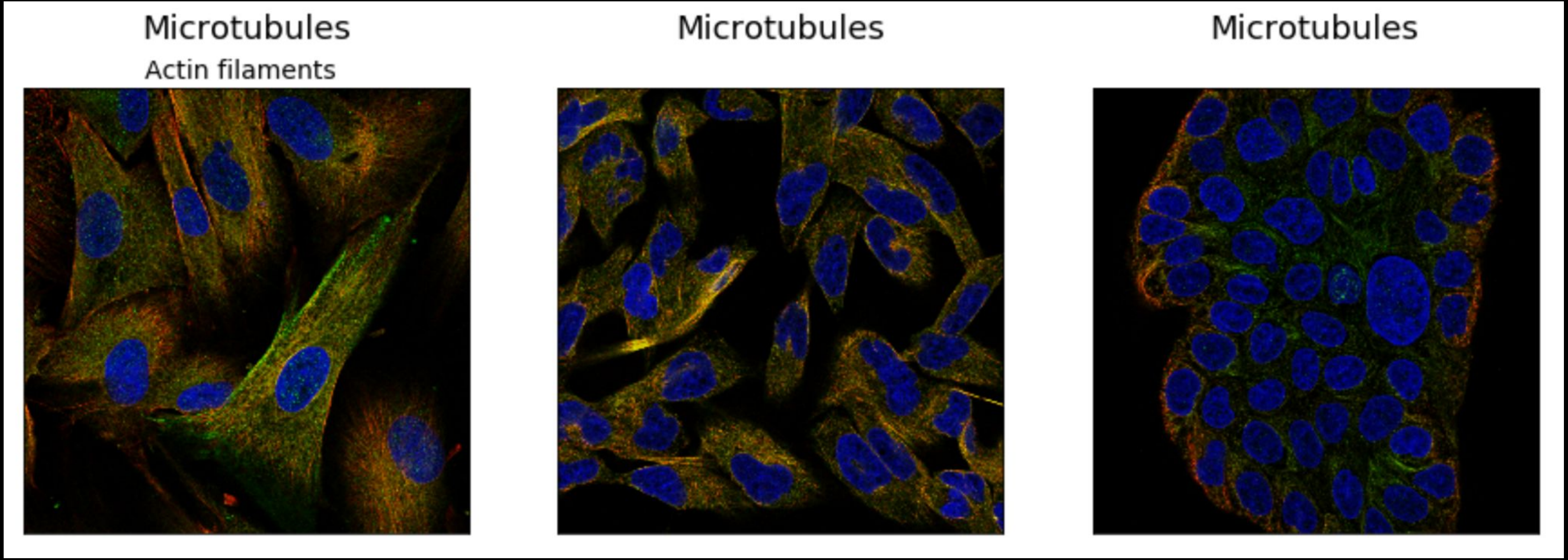
Cytokinetic bridge
Cytosol

Cytokinetic bridge
Microtubules, Nucleoplasm

Cytokinetic bridge
Microtubules, Mitotic spindle

1) Human Protein Atlas Image Classification

Green = the protein of interest; blue/red/yellow are constant reference markers (nucleus, microtubules, ER) to help you judge where the green signal sits.



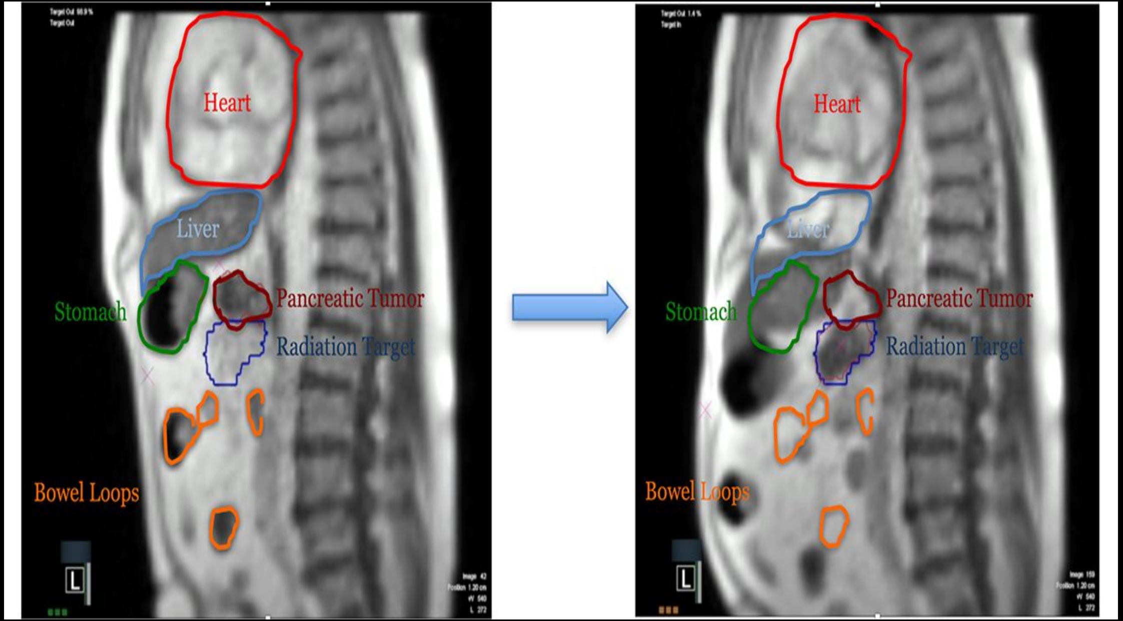
1) Human Protein Atlas Image Classification

⇒ Multilabel classification problem (28 subcellular components).

⇒ Very simple vision problem.

2) UW-Madison GI Tract Image Segmentation

- Radiation oncologists must manually segment the position of the stomach and intestines in order to adjust the direction of the x-ray beams to increase the dose delivery to the tumor and avoid the stomach and intestines.



Normal

Breath Hold

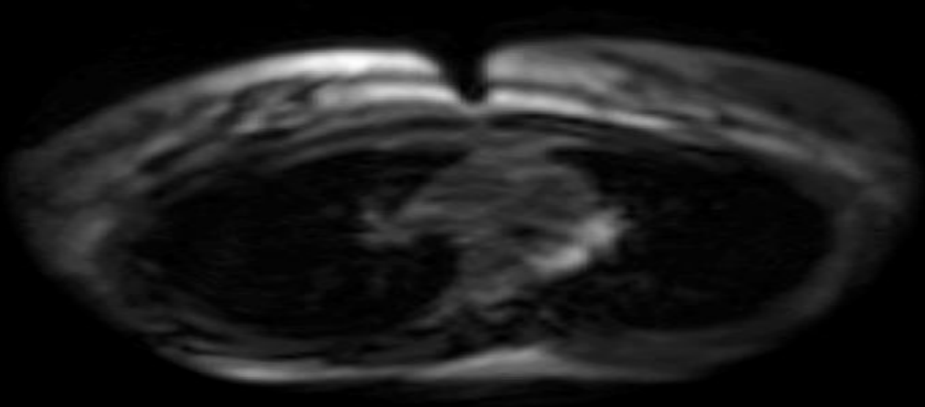
2) UW-Madison GI Tract Image Segmentation

- Radiation oncologists must manually segment the position of the stomach and intestines in order to adjust the direction of the x-ray beams to increase the dose delivery to the tumor and avoid the stomach and intestines.
- Goal: Create a model to automatically segment the stomach and intestines on MRI scans.

2) UW-Madison GI Tract Image Segmentation

- Radiation oncologists must manually segment the position of the stomach and intestines in order to adjust the direction of the x-ray beams to increase the dose delivery to the tumor and avoid the stomach and intestines.
- Goal: Create a model to automatically segment the stomach and intestines on MRI scans.
- The MRI scans are from actual cancer patients who had 1-5 MRI scans on separate days during their radiation treatment.

2) UW-Madison GI Tract Image Segmentation



link

2) UW-Madison GI Tract Image Segmentation

- Data Modality: Preprocessed PNG 2D slices (3D Images)

2) UW-Madison GI Tract Image Segmentation

- Data Modality: Preprocessed PNG 2D slices (3D Images)
- Task: 2D / 2.5D / 3D Segmentation.

2) UW-Madison GI Tract Image Segmentation

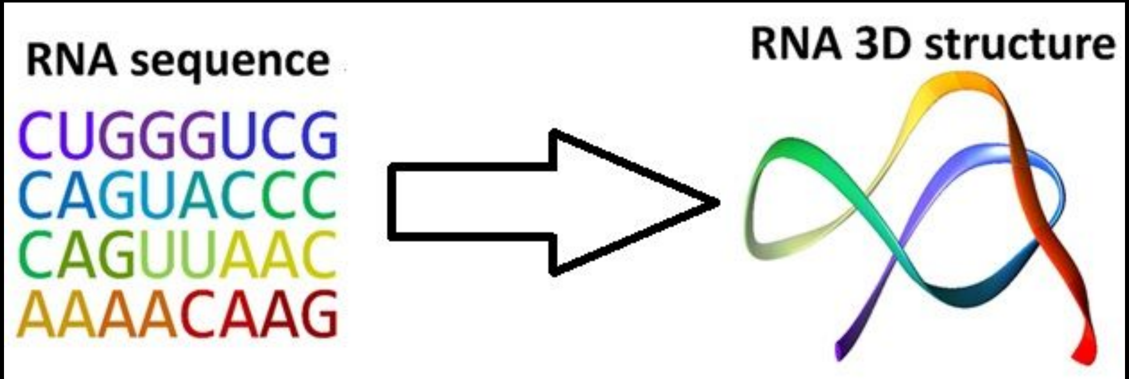
- Data Modality: Preprocessed PNG 2D slices (3D Images)
- Task: 2D / 2.5D / 3D Segmentation.
- Models: 2D / 3D Unet

2) UW-Madison GI Tract Image Segmentation

- Data Modality: Preprocessed PNG 2D slices (3D Images)
- Task: 2D / 2.5D / 3D Segmentation.
- Models: 2D / 3D Unet
- FYI: Using a detection model first to pick the interesting regions, then segmenting them worked better here👀.

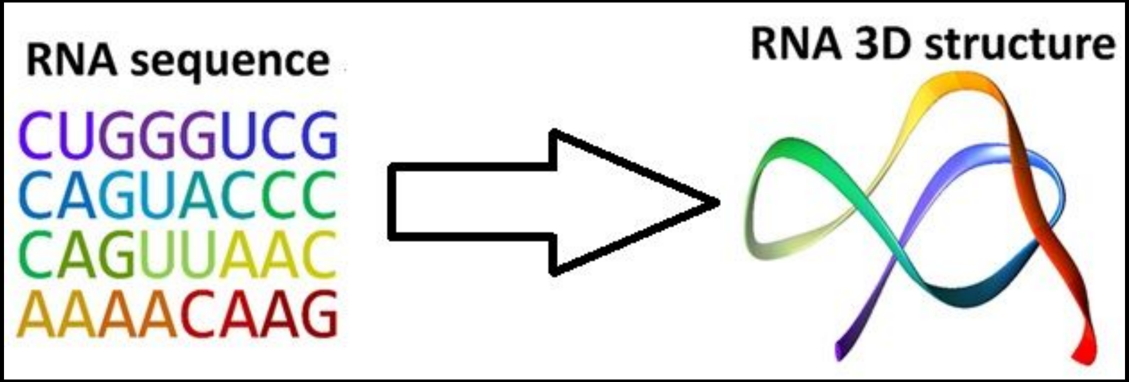
3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule's 3D structure from its sequence.



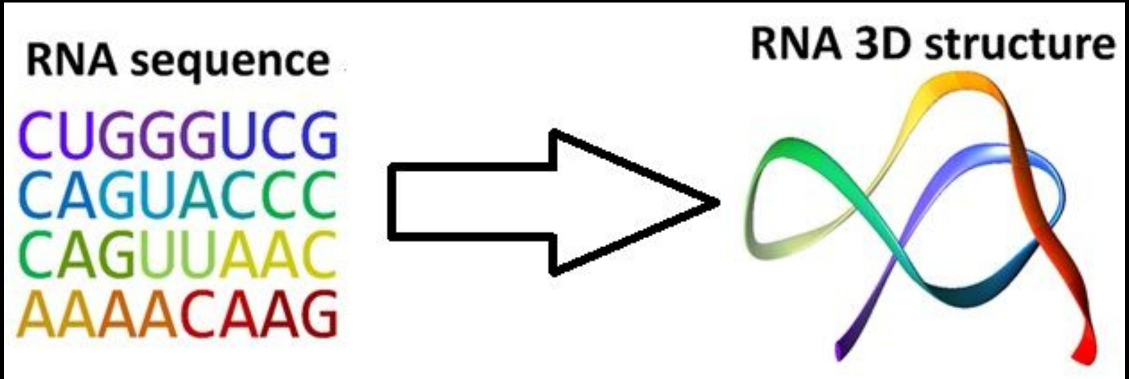
3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule's 3D structure from its sequence.



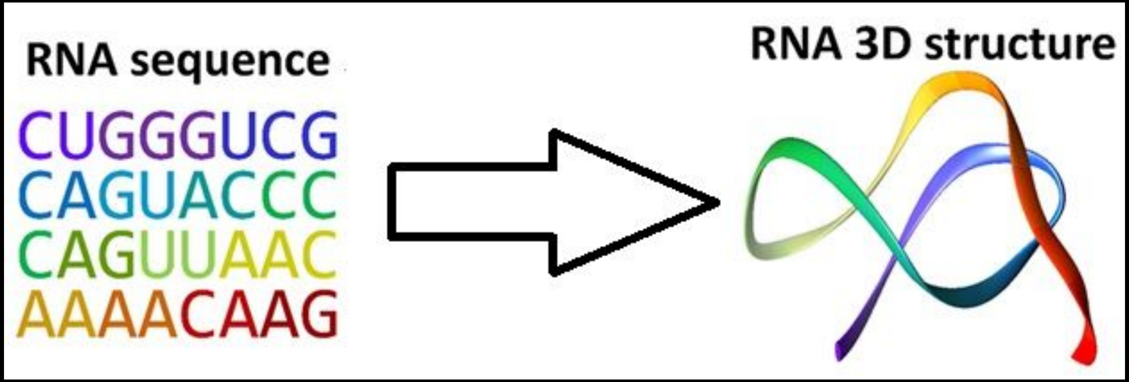
3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule’s 3D structure from its sequence.
- In other words, predict the 3D coordinates of each atom.



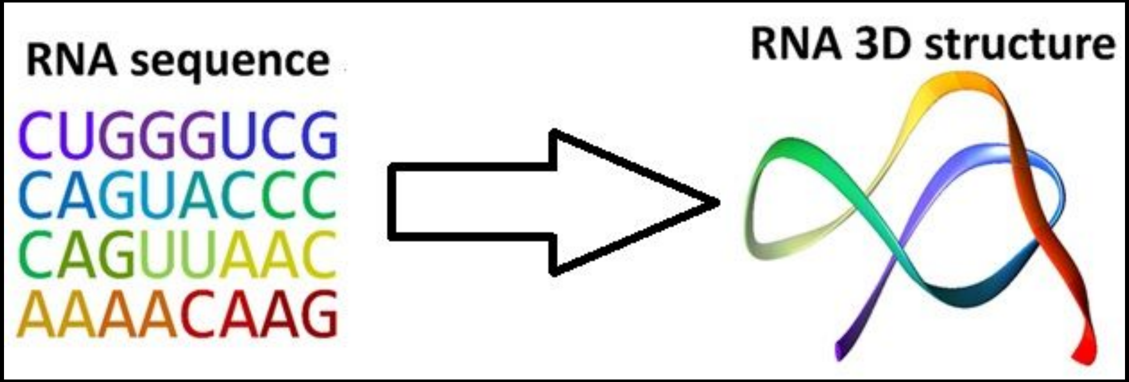
3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule's 3D structure from its sequence.
- In other words, predict the 3D coordinates of each atom.
- Data Modality: Sequence + PDB



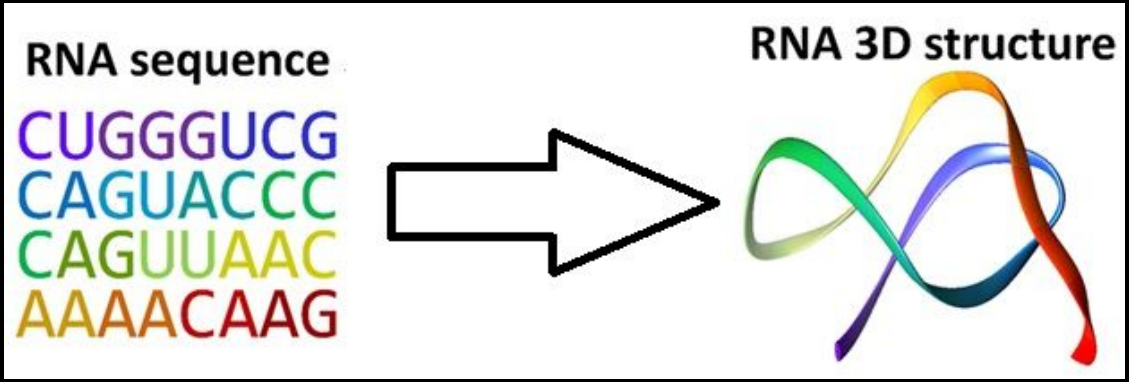
3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule's 3D structure from its sequence.
- In other words, predict the 3D coordinates of each atom.
- Data Modality: Sequence + PDB
- Task: Regression ez



3) Stanford RNA 3D Folding

- Goal: Predict an RNA molecule's 3D structure from its sequence.
- In other words, predict the 3D coordinates of each atom.
- Data Modality: Sequence + PDB
- Task: Regression ez



3) Stanford RNA 3D Folding

- Data Modality: Sequence + PDB ⇒ 3D coordinates per atom.
- Task: Regression on X, Y, Z.
- Model??

3) Stanford RNA 3D Folding

- Data Modality: Sequence + PDB ⇒ 3D coordinates per atom.
- Task: Regression on X, Y, Z.
- Model:
  - ML (Atom / Sequence features + Regressor)
  - Transformers (Sequence + Regression Head)
  - Graph NN (Graphs + Node Regressor)

3) Stanford RNA 3D Folding

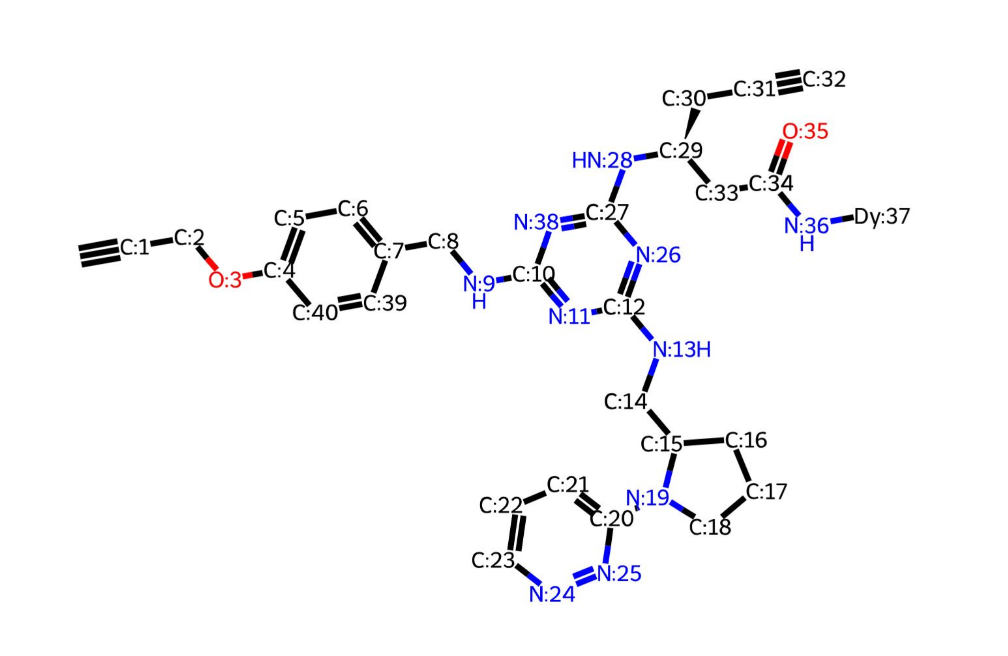
- Data Modality: Sequence + PDB ⇒ 3D coordinates per atom.
- Task: Regression on X, Y, Z.
- Model:
  - ML (Atom / Sequence features + Regressor)
  - Transformers (Sequence + Regression Head)
  - Graph NN (Graphs + Node Regressor)
- E.g. RibonanzaNet (transformer)

3) Stanford RNA 3D Folding

- Data Modality: Sequence + PDB ⇒ 3D coordinates per atom.
- Task: Regression on X, Y, Z.
- Model:
  - ML (Atom / Sequence features + Regressor)
  - Transformers (Sequence + Regression Head)
  - Graph NN (Graphs + Node Regressor)
- Google Deepmind did some good work here though
  - AlphaFold 2: Delivers near-atomic 3D protein structures from sequence alone.
  - AlphaFold 3: Extends prediction to whole complexes—proteins with DNA/RNA, ligands, ions, and modifications

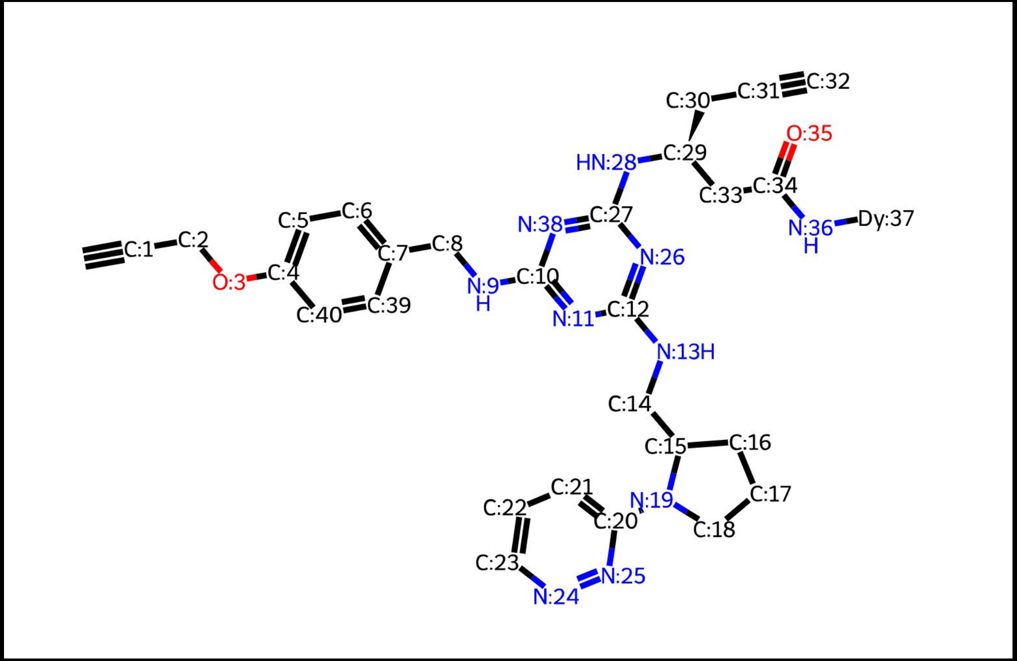
4) NeurIPS 2024 - Predict New Medicines with BELKA

- Knowing the binding affinity of small molecules to specific protein targets is a critical step in drug development.



4) NeurIPS 2024 - Predict New Medicines with BELKA

- Knowing the binding affinity of small molecules to specific protein targets is a critical step in drug development.
- Goal: Predict which drug-like small molecules (chemicals) will bind to three possible protein targets.



4) NeurIPS 2024 - Predict New Medicines with BELKA

- Data Modality: SMILES + Protein Target Name

4) NeurIPS 2024 - Predict New Medicines with BELKA

- Data Modality: SMILES + Protein Target Name
- Task: Binary Classification (Will bind or not).

4) NeurIPS 2024 - Predict New Medicines with BELKA

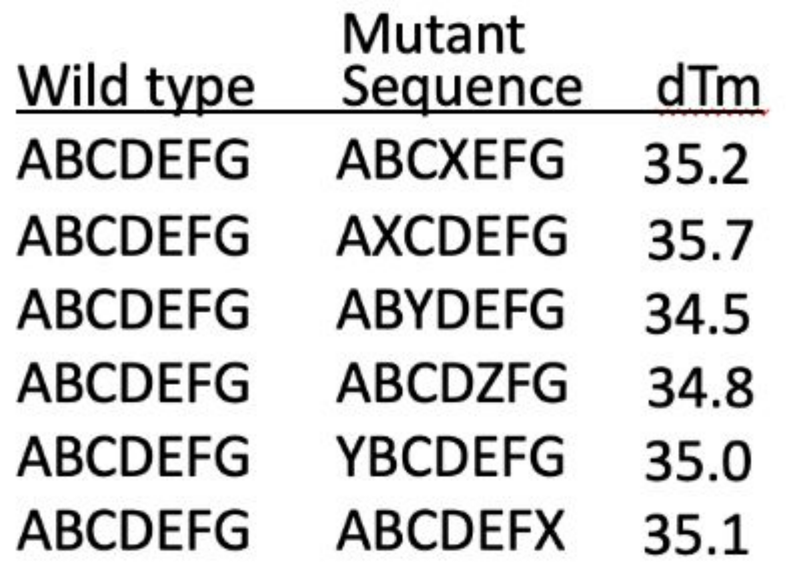
- Data Modality: SMILES + Protein Target Name
- Task: Binary Classification (Will bind or not).
- Model:
  - ML (SMILES features + protein + Classifier)
  - Transformers (SMILES backbone (e.g. ChemBERTa) + Classification Head)
  - Graph NN (Graph Classification + Protein name as a feature within the nodes)

4) NeurIPS 2024 - Predict New Medicines with BELKA

- Data Modality: SMILES + Protein Target Name
- Task: Binary Classification (Will bind or not).
- Model:
  - ML (SMILES features + protein + Classifier)
  - Transformers (SMILES backbone (e.g. ChemBERTa) + Classification Head)
  - Graph NN (Graph Classification + Protein name as a feature within the nodes)
  - 1D CNN

5) Novozymes Enzyme Stability Prediction

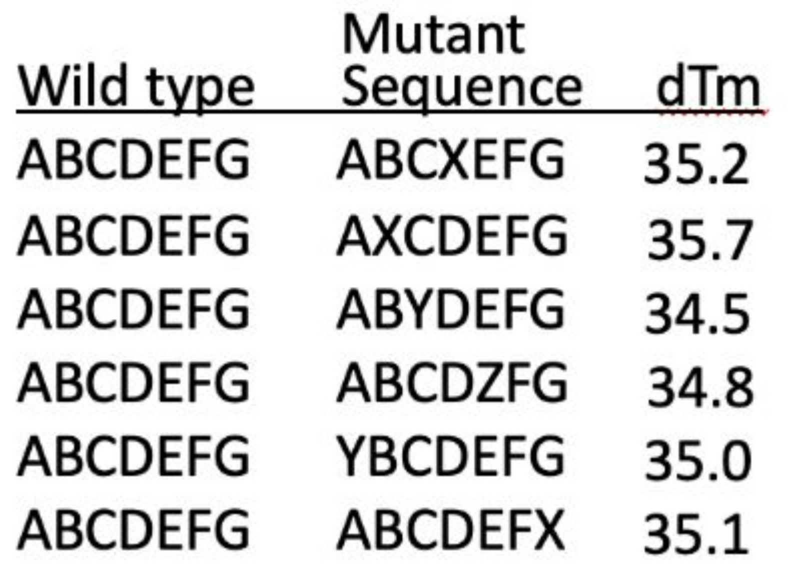
- Predicting protein stability is a fundamental problem in biotechnology.
- Its applications include enzyme engineering for addressing the world's challenges in sustainability, carbon neutrality and more.



Wild type	Mutant Sequence	dTm
ABCDEFG	ABCXEFG	35.2
ABCDEFG	AXCDEFG	35.7
ABCDEFG	ABYDEFG	34.5
ABCDEFG	ABCDZFG	34.8
ABCDEFG	YBCDEFG	35.0
ABCDEFG	ABCDEFX	35.1

5) Novozymes Enzyme Stability Prediction

- Goal: Predict the ranking of protein thermostability (as measured by melting point, tm) after single-point amino acid mutation and deletion.



	Mutant	
Wild type	Sequence	dTm
ABCDEFG	ABCXEFG	35.2
ABCDEFG	AXCDEFG	35.7
ABCDEFG	ABYDEFG	34.5
ABCDEFG	ABCDZFG	34.8
ABCDEFG	YBCDEFG	35.0
ABCDEFG	ABCDEFX	35.1

5) Novozymes Enzyme Stability Prediction

- Data Modality: Sequences (Wildtype + mutant)
- Task: Ranking (how to do it?).

5) Novozymes Enzyme Stability Prediction

- Data Modality: Sequences (Wildtype + mutant)
- Task: Ranking (regression then sort ez).
- Model:

5) Novozymes Enzyme Stability Prediction

- Data Modality: Sequences (Wildtype + mutant)
- Task: Ranking (regression then sort ez).
- Model:
  - ML (TF-IDF features + diff features + Regressor)
  - Transformers (protein backbone (e.g. ProtBert) + Regression Head)
  - Graph NN (Graph Regressor)

5) Novozymes Enzyme Stability Prediction

- Data Modality: Sequences (Wildtype + mutant)
- Task: Ranking (regression then sort ez).
- Model:
  - ML (TF-IDF features + diff features + Regressor)
  - Transformers (protein backbone (e.g. ProtBert) + Regression Head)
  - Graph NN (Graph Regressor)
  - 3D CNN

5) Novozymes Enzyme Stability Prediction

- Data Modality: Sequences (Wildtype + mutant)
- Task: Ranking (regression then sort ez).
- Model:
  - ML (TF-IDF features + diff features + Regressor)
  - Transformers (protein backbone (e.g. ProtBert) + Regression Head)
  - Graph NN (Graph Regressor)
  - 3D CNN ⇒ Convert sequences to PDB then use 3D CNN ⇒ The most powerful idea (ThermoNet link)

Thanks for Attending!

Prepared By: Mohamed Eltayeb