# Automated Protein Function Description

Meet Barot, PhD Student, Center for Data Science, NYU

April 29, 2022

# Intro

- Machine learning for protein function prediction

[1] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. "deepNF: deep network fusion for protein function prediction". In: *Bioinformatics* 34.22 (2018), pp. 3873–3881.

[2] Meet Barot et al. "NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity". In: *Bioinformatics* 37.16 (Feb. 2021), pp. 2414–2422. DOI: 10.1093/bioinformatics/btab098.

# Intro

- Machine learning for protein function prediction
- Data: amino acid sequence, networks

[1] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. "deepNF: deep network fusion for protein function prediction". In: *Bioinformatics* 34.22 (2018), pp. 3873–3881.

[2] Meet Barot et al. "NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity". In: *Bioinformatics* 37.16 (Feb. 2021), pp. 2414–2422. DOI: 10.1093/bioinformatics/btab098.

# Intro

- Machine learning for protein function prediction
- Data: amino acid sequence, networks
- deepNF[1]: integrating different types of edges in protein networks for a single organism

[1] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. "deepNF: deep network fusion for protein function prediction". In: *Bioinformatics* 34.22 (2018), pp. 3873–3881.

[2] Meet Barot et al. "NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity". In: *Bioinformatics* 37.16 (Feb. 2021), pp. 2414–2422. DOI: 10.1093/bioinformatics/btab098.

# Intro

- Machine learning for protein function prediction
- Data: amino acid sequence, networks
- deepNF[1]: integrating different types of edges in protein networks for a single organism
- NetQuilt[2]: integrating PPI networks of multiple organisms

[1] Vladimir Gligorijević, Meet Barot, and Richard Bonneau. "deepNF: deep network fusion for protein function prediction". In: *Bioinformatics* 34.22 (2018), pp. 3873–3881.

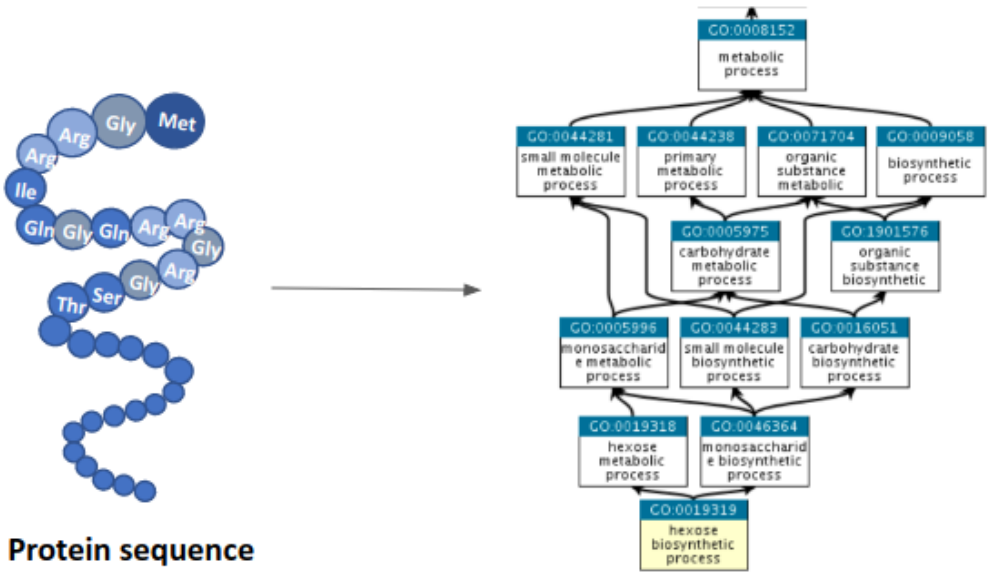
[2] Meet Barot et al. "NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity". In: *Bioinformatics* 37.16 (Feb. 2021), pp. 2414–2422. DOI: 10.1093/bioinformatics/btab098.

# Roadmap

1. Protein function prediction $\rightarrow$ protein function description
2. Motivation
3. Model
4. Evaluation metrics
5. Results

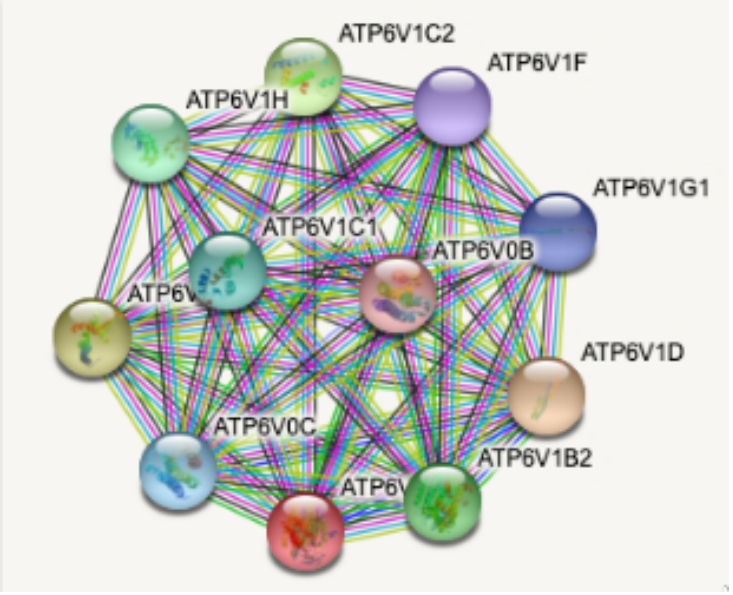
# Protein function prediction as it is

Supervised multilabel problem, where sequences are mapped to labels organized into a hierarchy, e.g. the Gene Ontology



# Protein function prediction as it **should be**

Given a set of proteins, describe their common function.




$\rightarrow$

"Proton-transporting atpase activity, rotational mechanism."

# Motivation for protein function description

Why make a model that describes the common functions of a set of proteins in natural language?

# Motivation for protein function description

- Why use sets of proteins?

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.
- Why use natural language?

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.
- Why use natural language?
  - We can avoid having our predictions be limited to a pre-defined set of functions

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.
- Why use natural language?
  - We can avoid having our predictions be limited to a pre-defined set of functions
  - With language, the model can compose new functions out of the same pieces that we use to explain the world to each other.

# Motivation for protein function description

- Why use sets of proteins?
  - A function description is our abstraction of the common property of a group of proteins.
  - We discover functions by understanding that a group of proteins do something in common.
- Why use natural language?
  - We can avoid having our predictions be limited to a pre-defined set of functions
  - With language, the model can compose new functions out of the same pieces that we use to explain the world to each other.

To discover new categories of protein function with info to guide experimental design to test for them, we need a model that generates functional descriptions.

# Proposed model

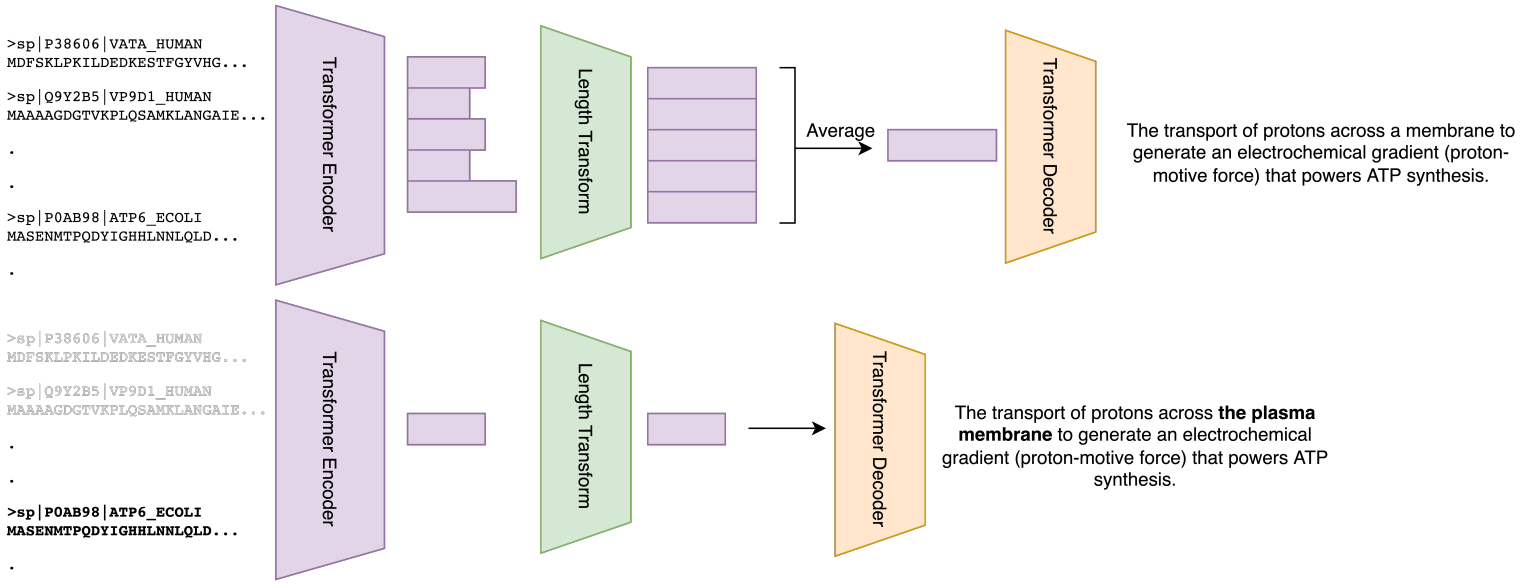



Figure: Transformer encoder-decoder model. Estimates $P(d|S)$ where S is a set of protein sequences and d is a given description.

# Key aspects of method

1. Input: Protein sets to describe, invariant to order

# Key aspects of method

1. Input: Protein sets to describe, invariant to order
2. Autoregressive generation of descriptions

# Key aspects of method

1. Input: Protein sets to describe, invariant to order
2. Autoregressive generation of descriptions
3. Evaluation

# Key aspects of method

1. Input: Protein sets to describe, invariant to order
2. Autoregressive generation of descriptions
3. Evaluation
   - Description Generation: difficult to evaluate

# Key aspects of method

1. Input: Protein sets to describe, invariant to order
2. Autoregressive generation of descriptions
3. Evaluation
   - Description Generation: difficult to evaluate
   - Scoring descriptions of unseen categories ("zero-shot" classification)

# Key aspects of method

1. Input: Protein sets to describe, invariant to order
2. Autoregressive generation of descriptions
3. Evaluation
   - Description Generation: difficult to evaluate
   - Scoring descriptions of unseen categories ("zero-shot" classification)

# How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.

# How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.
  - This is not a solved problem in natural language processing.

# How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.
  - This is not a solved problem in natural language processing.
  - Most work relies on manual ranking of generated text samples in terms of their similarity to a gold standard.

# How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.
  - This is not a solved problem in natural language processing.
  - Most work relies on manual ranking of generated text samples in terms of their similarity to a gold standard.
- We can, however, evaluate the scoring that the model assigns to pairs of sequence sets and their known descriptions.

# Evaluation metrics

Three things that we care about for generated descriptions:

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one
2. Specificity: Among correct descriptions, we want descriptions that are more specific to be ranked higher than more general ones.

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one
2. Specificity: Among correct descriptions, we want descriptions that are more specific to be ranked higher than more general ones.
   - Average number of times a correct child term outranks its parent term(s)

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one
2. Specificity: Among correct descriptions, we want descriptions that are more specific to be ranked higher than more general ones.
   - Average number of times a correct child term outranks its parent term(s)
3. Robustness: For two sets of proteins that have the same common functions among each set, we want the correct descriptions' rankings to be the same for both sets.

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one
2. Specificity: Among correct descriptions, we want descriptions that are more specific to be ranked higher than more general ones.
   - Average number of times a correct child term outranks its parent term(s)
3. Robustness: For two sets of proteins that have the same common functions among each set, we want the correct descriptions' rankings to be the same for both sets.
   - Average Spearman's rank correlation of the sequence sets' correct terms

# Evaluation metrics

Three things that we care about for generated descriptions:

1. Correctness: We want accurate descriptions about the protein set's function to be ranked higher than inaccurate ones.
   - Average number of times a correct description outranks an incorrect one
2. Specificity: Among correct descriptions, we want descriptions that are more specific to be ranked higher than more general ones.
   - Average number of times a correct child term outranks its parent term(s)
3. Robustness: For two sets of proteins that have the same common functions among each set, we want the correct descriptions' rankings to be the same for both sets.
   - Average Spearman's rank correlation of the sequence sets' correct terms

# Data

- Uniprot-KB Swiss-Prot (manually annotated and reviewed), 566,996 proteins total
  1. Maximum number of proteins per GO term: 1280
  2. Minmum number of proteins per GO term: 32
  3. Total number of proteins in training set: 316k
  4. Total number of proteins in validation set: 180k
  5. Total number of GO terms in training set: 9053
  6. Total number of GO terms in validation set: 2264

# Results

Table: Model Validation Set Performances

| Metric | Model scores | Term-normalized model scores |
|---|---|---|
| Correctness | 0.57 | 0.83 |
| Specificity | 0.52 | 0.58 |
| Robustness | 0.84 | 0.44 |

# Description generation examples

- Taking a Gene Ontology term that was not included in training

# Description generation examples

- Taking a Gene Ontology term that was not included in training
- Sampling the proteins annotated with that term

# Description generation examples

- Taking a Gene Ontology term that was not included in training
- Sampling the proteins annotated with that term
- Use these proteins as input to the model

# Description generation examples

- Taking a Gene Ontology term that was not included in training
- Sampling the proteins annotated with that term
- Use these proteins as input to the model
- Use beam search over token probabilities to generate tokens sequentially

# Validation set generation examples

GO:0032099: negative regulation of appetite
Actual description:
any process that reduces appetite .
Prediction:
any process that results in a change in state or activity of an organism ( in terms of a result of a leptin stimulus .

# Validation set generation examples

GO:0003725: double-stranded RNA binding
Actual description:
binding to double-stranded rna .
Prediction:
any process that activates or increases the rate or extent of rrna molecule .

# Validation set generation examples

GO:0035434: copper ion transmembrane transport
Actual description:
the directed movement of copper cation across a membrane .
Prediction:
the increase in size ( protons .

# Validation set generation examples

GO:0046992: oxidoreductase activity, acting on X-H and Y-H to form an X-Y bond
Actual description:
catalysis of an oxidation-reduction ( redox ) reaction in which x-h and y-h form x-y .
Prediction:
the formation of methane , the formula ch4 .

# Validation set generation examples

GO:0042023: DNA endoreduplication
Actual description:
regulated re-replication of dna within a single cell cycle , resulting in an increased cell ploidy . an example of this process occurs in the synthesis of drosophila salivary gland cell polytene chromosomes .
Prediction:
any process that modulates the frequency , rate or extent of chromatin organization .

# Future human-assisted evaluation of function discovery

- Human-assisted evaluation will be needed for the descriptions generated for a given set of novel proteins.

[3] Nisan Stiennon et al. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021, Daniel M Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

# Future human-assisted evaluation of function discovery

- Human-assisted evaluation will be needed for the descriptions generated for a given set of novel proteins.
- One possible way of obtaining human feedback would be to ask an expert Gene Ontology curator to choose between two descriptions for a given sequence set that is generated from a trained model.

[3] Nisan Stiennon et al. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021, Daniel M Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

# Future human-assisted evaluation of function discovery

- Human-assisted evaluation will be needed for the descriptions generated for a given set of novel proteins.
- One possible way of obtaining human feedback would be to ask an expert Gene Ontology curator to choose between two descriptions for a given sequence set that is generated from a trained model.
  - This would be expensive, as the task needs to be done by an expert.

[3] Nisan Stiennon et al. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021, Daniel M Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

# Future human-assisted evaluation of function discovery

- Human-assisted evaluation will be needed for the descriptions generated for a given set of novel proteins.
- One possible way of obtaining human feedback would be to ask an expert Gene Ontology curator to choose between two descriptions for a given sequence set that is generated from a trained model.
  - This would be expensive, as the task needs to be done by an expert.
- This feedback could be used to fine-tune the model to produce more accurate, fluid or generally desirable descriptions of proteins, as has been done for document summarization models[3].

[3] Nisan Stiennon et al. "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021, Daniel M Ziegler et al. "Fine-tuning language models from human preferences". In: *arXiv preprint arXiv:1909.08593* (2019).

Vladimir Gligorijevic

Richard Bonneau

Kyunghyun Cho

Contact me at meetbarot@nyu.edu