Automated Protein Function Description

Meet Barot

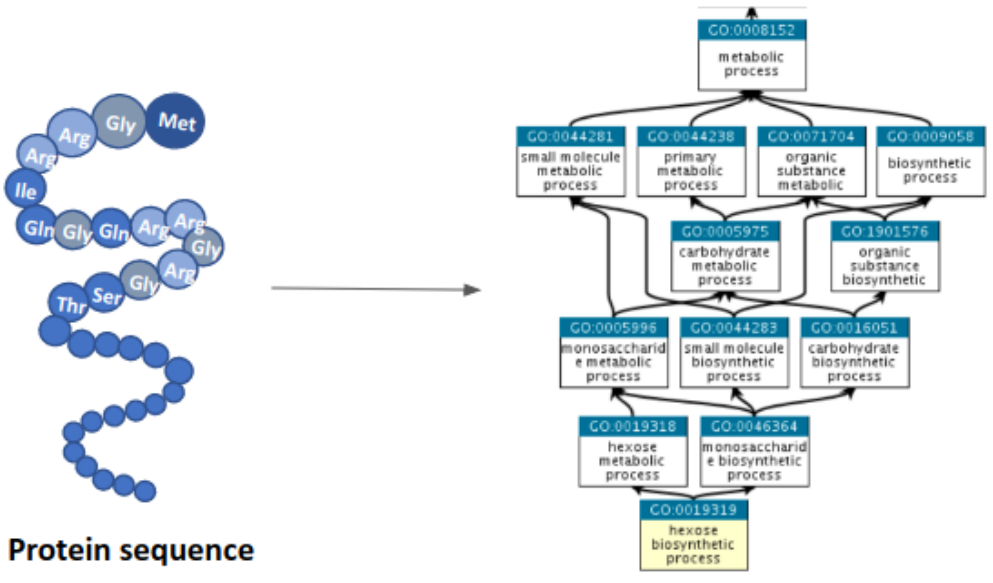
April 4, 2022

Roadmap

1. Protein function prediction $\rightarrow$ protein function description
2. Motivation
3. Evaluation metrics
4. Results
5. Performance issues
6. Experiments to do

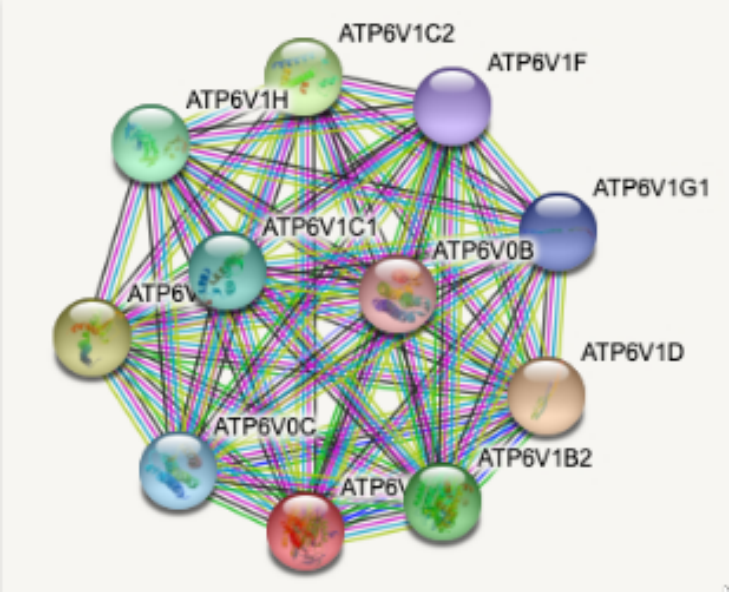
Protein function prediction as it is

Supervised multilabel problem, where sequences are mapped to labels organized into a hierarchy, i.e. 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.

Current methods for function discovery

- Many methods exist for function prediction, but most do not consider the problem of discovering novel functions.

[1] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms". In: *bioRxiv* (2022). DOI: 10.1101/2022.01.14.476325. eprint: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.

Current methods for function discovery

- Many methods exist for function prediction, but most do not consider the problem of discovering novel functions.
- Clustering-based methods are not able to give much information about the new functional categories that they predict

[1] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms". In: *bioRxiv* (2022). DOI: 10.1101/2022.01.14.476325. eprint: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.

Current methods for function discovery

- Many methods exist for function prediction, but most do not consider the problem of discovering novel functions.
- Clustering-based methods are not able to give much information about the new functional categories that they predict
- DeepGOZero[1] predicts for terms not included in the training set, but this is limited to terms with ontological relations with known terms.

[1] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms". In: *bioRxiv* (2022). DOI: 10.1101/2022.01.14.476325. eprint: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.

Proposed model

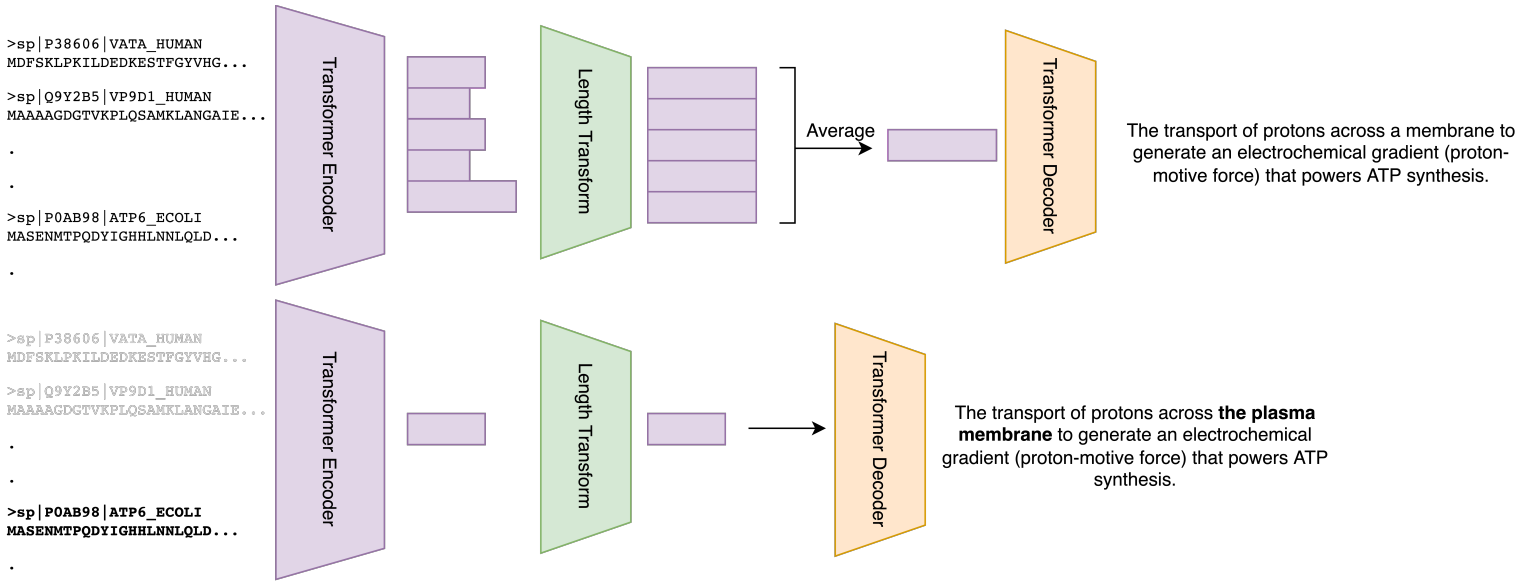


Figure: Transformer encoder-decoder model. Estimates $P(d|S)$ where S is a set of protein sequences and d is a given description.

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, including 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, including zero-shot classification
4. Implementation

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, including zero-shot classification
4. Implementation
   - Transformer encoder-decoder model

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, including zero-shot classification
4. Implementation
   - Transformer encoder-decoder model
   - Length transform

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, including zero-shot classification
4. Implementation
   - Transformer encoder-decoder model
   - Length transform
   - Generation using beam search

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 NLP in translation, summarization, etc.

How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.
  - This is not a solved problem in NLP in translation, summarization, etc.
  - Automated measures of translation quality are heuristics based on n-gram similarity

How do you evaluate this model?

- Evaluating *generated* descriptions can really only be done manually.
  - This is not a solved problem in NLP in translation, summarization, etc.
  - Automated measures of translation quality are heuristics based on n-gram similarity
  - 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 NLP in translation, summarization, etc.
  - Automated measures of translation quality are heuristics based on n-gram similarity
  - 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

Three things that we care about for generated descriptions:

1. Correctness

Evaluation

Three things that we care about for generated descriptions:

1. Correctness
2. Specificity

Evaluation

Three things that we care about for generated descriptions:

1. Correctness
2. Specificity
3. Robustness

Attribute 1: Annotation correctness.

Descriptions of GO terms that annotate the entire sequence set should be scored higher than terms that do not.

Attribute 1: Annotation correctness.

Let D_S be the set of GO term descriptions associated with sequence set S.

Attribute 1: Annotation correctness.

Let D_S be the set of GO term descriptions associated with sequence set S.

$$P(d \in D_S|S) > P(d \notin D_S|S)$$

Attribute 1: Annotation correctness.

Let D_S be the set of GO term descriptions associated with sequence set S.

$$P(d \in D_S|S) > P(d \notin D_S|S)$$

We can calculate the average number of times a correct description outranks an incorrect one:

$$\frac{1}{|D_S| * |D_S^c|} \sum_{d_i \in D_S, d_j \notin D_S} \mathbb{1}(P(d_i|S) > P(d_j|S))$$

where D_S^c is the complement of D_S and $\mathbb{1}$ is the indicator function.

Attribute 2: Specificity preference.

Among correct terms, the model should score child terms higher than their parent terms.

Attribute 2: Specificity preference.

Let $A(d)$ denote the description of a direct parent of the GO term described by d.

Attribute 2: Specificity preference.

Let $A(d)$ denote the description of a direct parent of the GO term described by d.We want:

$$P(d \in D_S|S) > P(A(d) \in D_S|S)$$

Attribute 2: Specificity preference.

Let $A(d)$ denote the description of a direct parent of the GO term described by d.We want:

$$P(d \in D_S|S) > P(A(d) \in D_S|S)$$

We can calculate the average number of times a correct child term outranks its parent term(s):

$$\frac{1}{|D_S|} \sum_{d_i \in D_S} \mathbb{1}(P(d_i|S) > P(A(d_i)|S))$$

Attribute 3: Annotation robustness.

Any pair of sequence sets that have the same GO descriptions in common should produce scores with the same rankings for those GO descriptions.

Attribute 3: Annotation robustness.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$, and let $R(X)$ be a ranking function that gives the rankings of probabilities in X.

Attribute 3: Annotation robustness.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$, and let $R(X)$ be a ranking function that gives the rankings of probabilities in X.

$$R_d(P(d \in D_{S_i}|S_i)) = R_d(P(d \in D_{S_i}|S_j))$$

Attribute 3: Annotation robustness.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$, and let $R(X)$ be a ranking function that gives the rankings of probabilities in X.

$$R_d(P(d \in D_{S_i}|S_i)) = R_d(P(d \in D_{S_i}|S_j))$$

We can calculate the average Spearman's rank correlation of the rankings for all sequence sets' correct descriptions.

Attribute 3: Annotation robustness.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$, and let $R(X)$ be a ranking function that gives the rankings of probabilities in X.

$$R_d(P(d \in D_{S_i}|S_i)) = R_d(P(d \in D_{S_i}|S_j))$$

We can calculate the average Spearman's rank correlation of the rankings for all sequence sets' correct descriptions.Let $R_{S_i} = R(P(D_{S_i}|S_i))$:

$$\frac{1}{N * (N-1)} \sum_{S_i, S_j} \frac{\text{cov}(R_{S_i}, R_{S_j})}{\sigma_{R_{S_i}} \sigma_{R_{S_j}}}$$

where N is the total number of sequence sets that have the exact set of GO descriptions D_{S_i}.

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 Performances

Metric	Validation set peformance
Annotation Correctness	0.54
Specificity Preference	0.57
Annotation Robustness	0.88

Training set generation examples

- Prediction: any process that modulates the frequency , rate or extent of calcidiol 1-monooxygenase activity .
- Actual description: any process that modulates the rate , frequency or extent of calcidiol 1-monooxygenase activity . calcidiol 1-monooxygenase activity is catalysis of the reaction calcidiol + nadph + h+ + o2 = calcitriol + nadp+ + h2o .

Training set generation examples

- Prediction: the modification of histone h4 by the addition of the addition of histone h2a .
- Actual description: the process involved in retention of aberrant or improperly formed mrnas , e . g . those that are incorrectly or incompletely spliced or that have incorrectly formed 3 ' -ends , within the nucleus at the site of transcription .

Training set generation examples

- Prediction: the process in which the developmental fate .
- Actual description: the multiplication or reproduction of cells , resulting in the expansion of a cell population that contributes to the shaping of the heart .

Training set generation examples

- Prediction: catalysis of the reaction (+ h2o + phosphate + phosphate .
- Actual description: enables the transfer of a solute or solutes from one side of a membrane to the other according to the reaction atp + h2o + thiamine (out) = adp + h (+) + phosphate + thiamine (in) .

Validation set generation examples

- Prediction: catalysis of the reaction sinapaldehyde + nadph = nadp+ = h+ .
- Actual description: binding to fructose 6-phosphate .

Validation set generation examples

- Prediction: catalysis of the reaction 2-deoxy-d-ribose 5-phosphate = d-glyceraldehyde 3-phosphate + acetaldehyde + acetaldehyde + acetaldehyde ...
- Actual description: catalysis of the reaction n6- (1 , 2-dicarboxyethyl) amp = fumarate + amp .

Validation set generation examples

- Prediction: the chemical reactions and pathways resulting in the formation of asparagine , the fundamental heterocyclic group of asparagine , from simpler precursors , the formation of the formation of the formation of the formation of the multisubunit water-soluble proteins , the formation of the multisubunit water-soluble proteins , the multisubunit water-soluble proteins , the multisubunit water-soluble proteins , the formation of the formation of the formation ...
- Actual description: the chemical reactions and pathways involving of salicylic acid (2-hydroxybenzoic acid) , a derivative of benzoic acid .

Current experiments to do

- Remove restriction of training/testing on functions with 32 to 1280 examples
- Architecture search
- Add oversmoothing regularization[2]
- Train on TREMBL annotations (more than 100 million proteins with some GO annotation, unreviewed)

[2] Ilia Kulikov, Maksim Eremeev, and Kyunghyun Cho. "Characterizing and addressing the issue of oversmoothing in neural autoregressive sequence modeling". In: *ArXiv* abs/2112.08914 (2021).

Vladimir Gligorijevic

Richard Bonneau

Kyunghyun Cho

Contact me at meetbarot@nyu.edu

[1] Ilia Kulikov, Maksim Eremeev, and Kyunghyun Cho. "Characterizing and addressing the issue of oversmoothing in neural autoregressive sequence modeling". In: *ArXiv* abs/2112.08914 (2021).

[2] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms". In: *bioRxiv* (2022). DOI: 10.1101/2022.01.14.476325. eprint: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.