# CELL-E 2: Translating Proteins to Pictures and Back with a Bidirectional Text-to-Image Transformer

**Emaad Khwaja**

***UC Berkeley - UCSF Bioengineering Graduate Program***

***UC Berkeley EECS***

# Goal

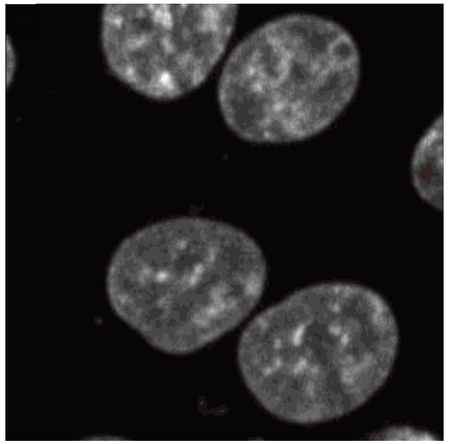
# Predict protein localization *in silico*.

*Protein Sequence*

MVCFRLFPVPGSGLVLV
CLVLGAVRSYALELNLT
DSENATCLYAKWQMNF
TVRYETTNKTYKTVTI
SDHGTVTYNGSICGDD
QNGPKIAVQFGPGFSWI
ANFTKAASTYSIDSVSF
SYNTGDNTTFPDAEDK
GILTVDELLAIRIPLND...

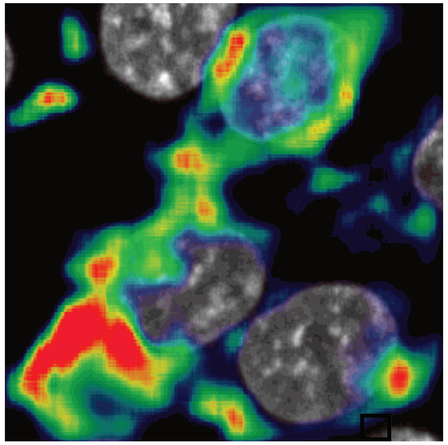
+

*Nucleus Image*



→

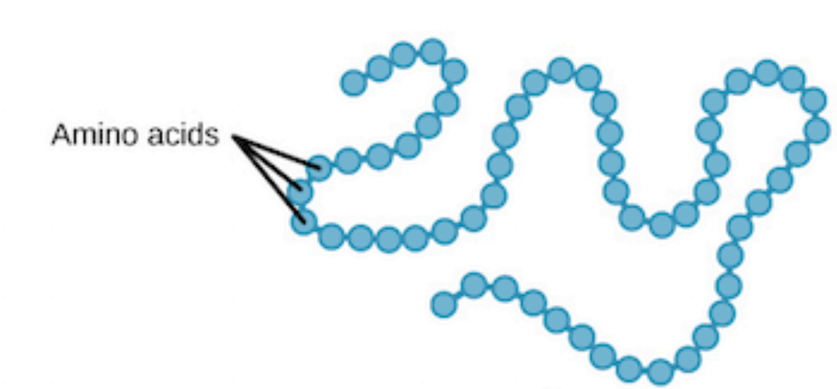
*Predicted Localization*



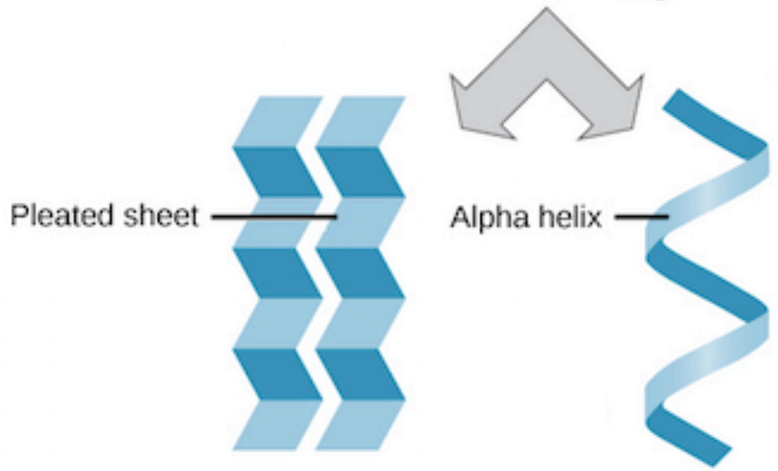
# Why?

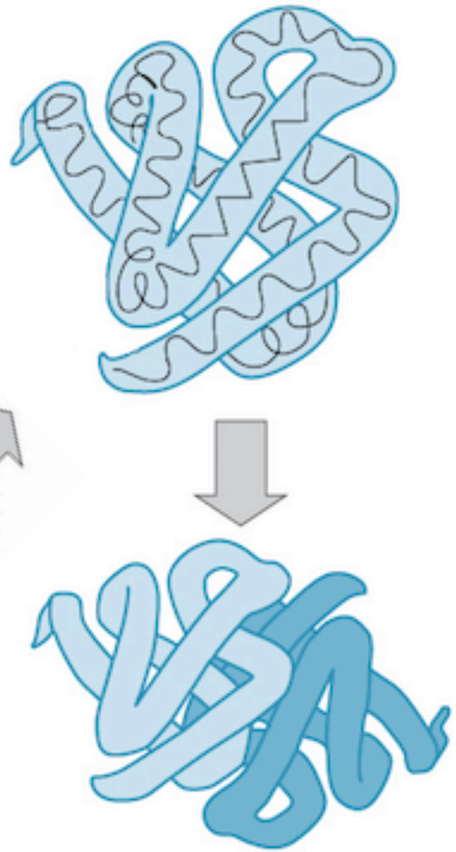
# Proteins dictate biological function.

**Primary protein structure**
sequence of a chain of animo acids




**Secondary protein structure**
hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern



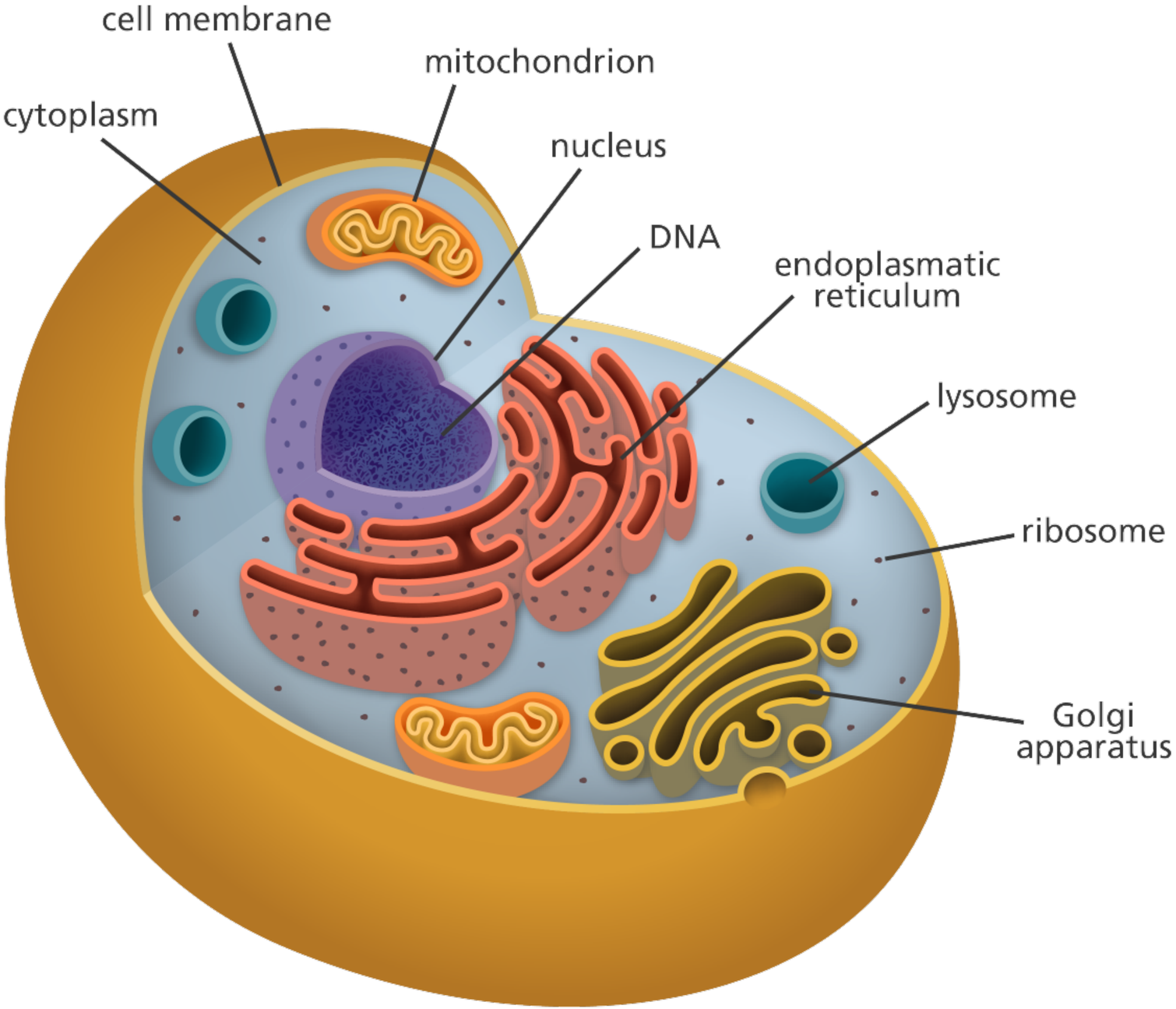



**Tertiary protein structure**
three-dimensional folding pattern of a protein due to side chain interactions

**Quaternary protein structure**
protein consisting of more than one amino acid chain

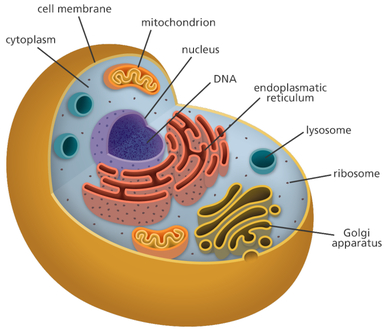
*Source: Khan Academy*

Predicting localization allows us to gain fundamental understanding of biology.




*Source: Genome Research Limited*

# Correct localization is necessary for function.




# Mislocalization can result in dysfunction and disease.

# CELL-E

*Protein Sequence*

*Nucleus Image*

*Protein Threshold Image*

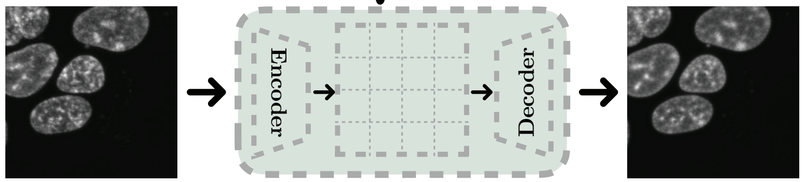
MSSASGLRRGHPAGGEEN...

BERT Tokenizer

Start | 16 | 22 | 22 | 5 | ... | End | Pad

Pre-trained TAPE Embedding

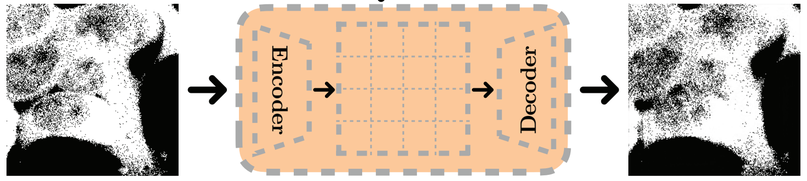
VQGAN




97 | 175 | 117 | 73 | 204 | ... | 231 | 231

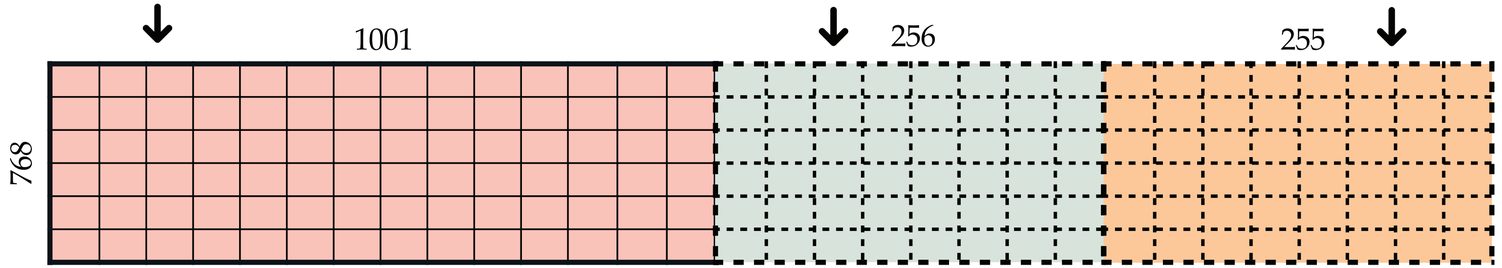
Learned Nucleus Embedding

VQGAN




812 | 574 | 802 | 612 | 804 | ... | 932 | 1023 | 712

Learned Threshold Embedding




Transformer

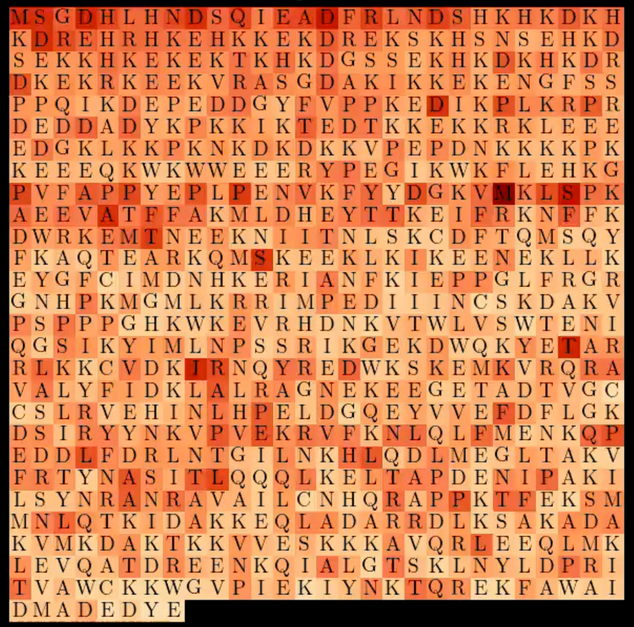
16 | 22 | 22 | 5 | ... | End | Pad

97 | 175 | 117 | 73 | 204 | ... | 231 | 231

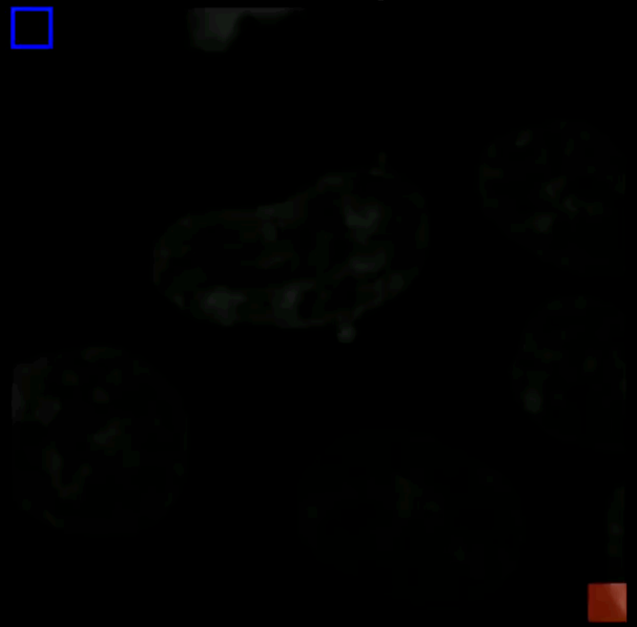
812 | 574 | 802 | 612 | 804 | ... | 932 | 1023 | 712

**DNA Topoisomerase I**
**Relative Attention Weights**

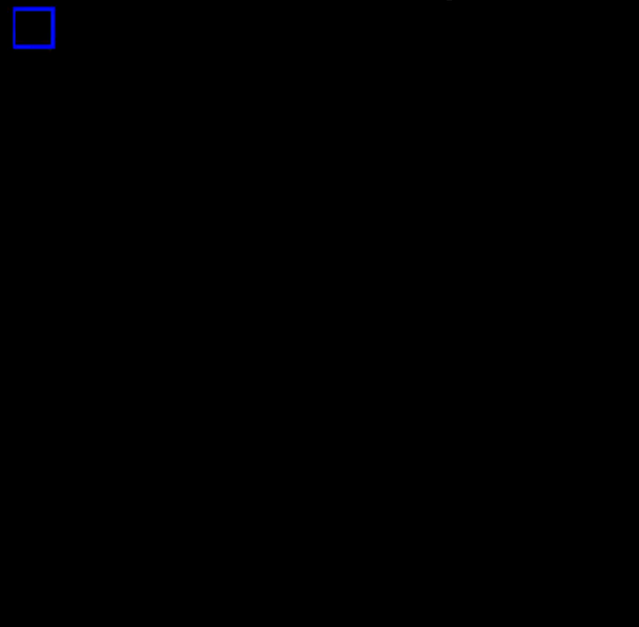
*Sequence*



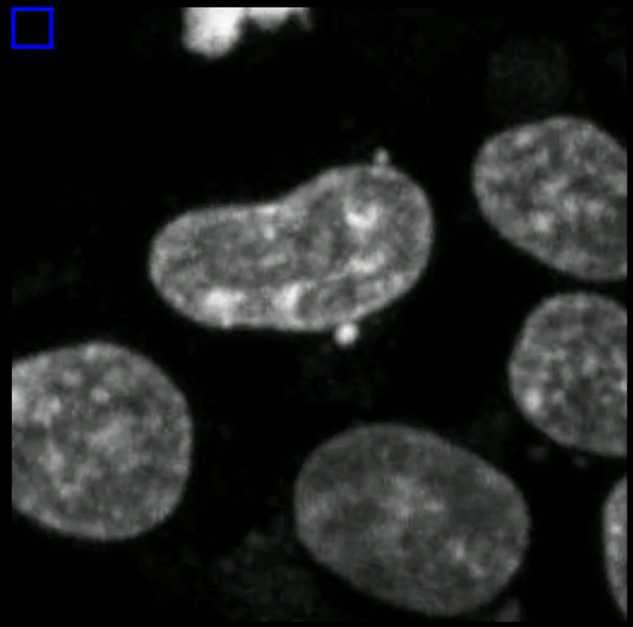

*Nucleus Image*



*Predicted Threshold Image*



*Predicted PDF*



# CELL-E 2

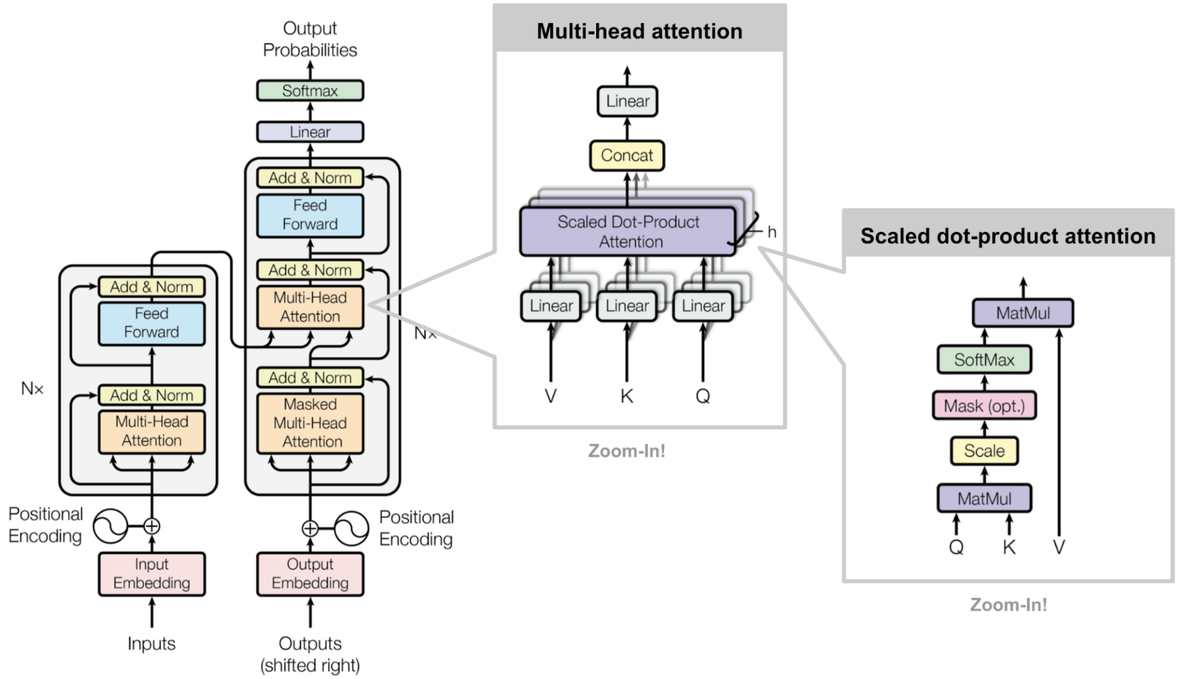
# Drawbacks

- **Small dataset**
  - **Not many proteins in OpenCell**
- Unidirectional
  - Can do text-to-image but cant tell you what text should be there with a localization pattern
- Autoregressive generation is slow

THE HUMAN PROTEIN ATLAS

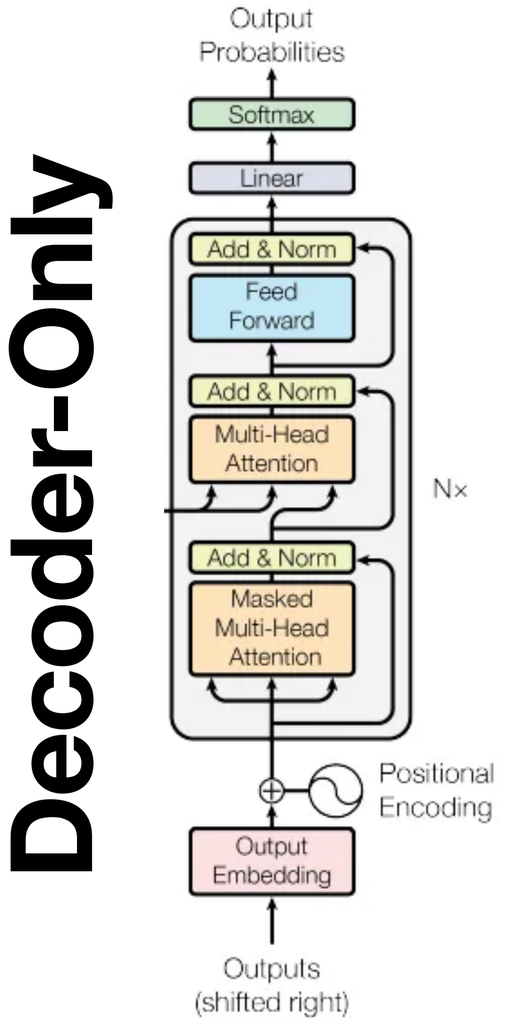
# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- **Unidirectional**
  - **Can do text-to-image but cant tell you what text should be there with a localization pattern**
- Autoregressive generation is slow



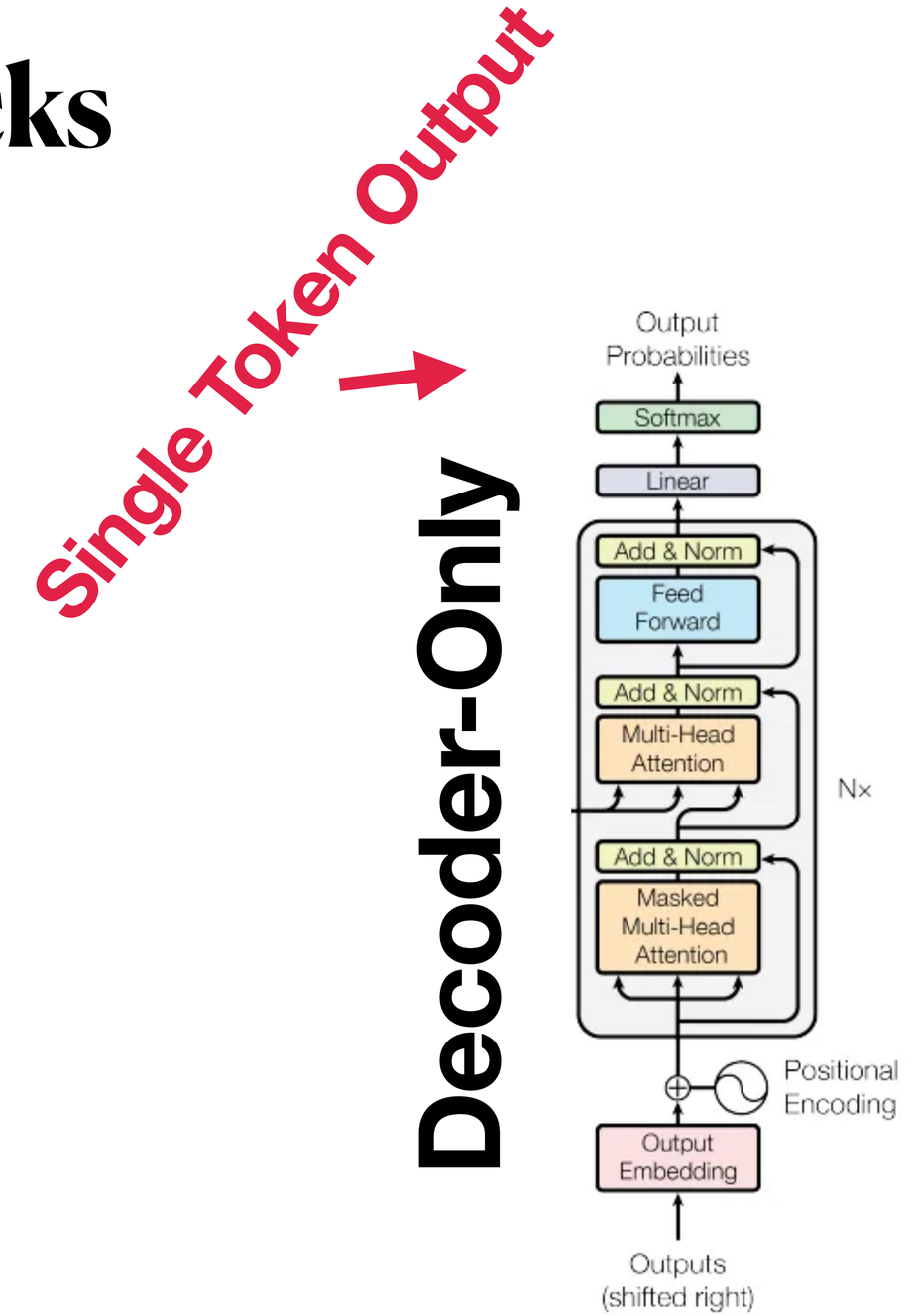
# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- **Unidirectional**
  - **Can do text-to-image but cant tell you what text should be there with a localization pattern**
- Autoregressive generation is slow



# Drawbacks

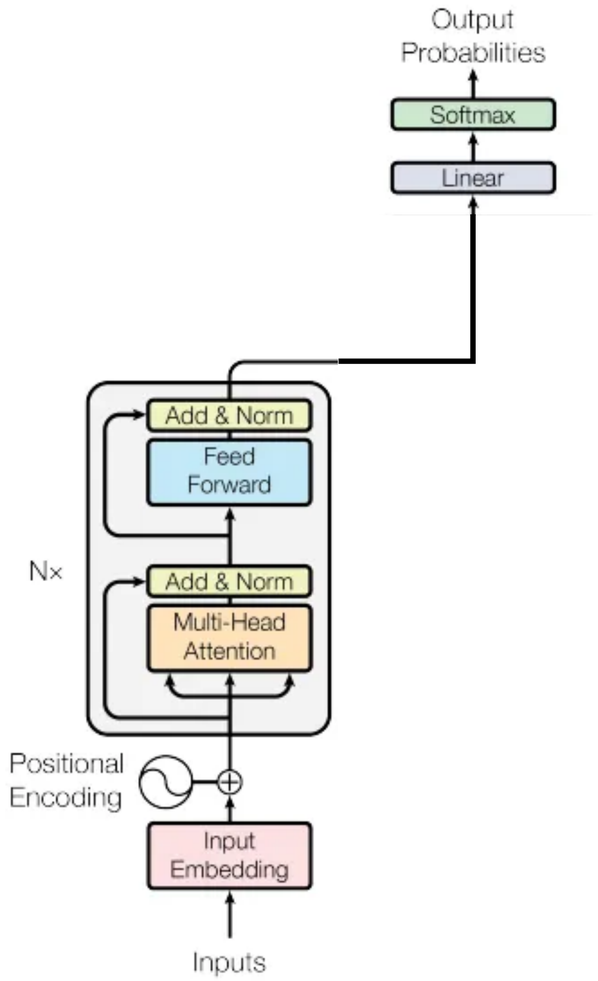
- Small dataset
  - Not many proteins in OpenCell
- **Unidirectional**
  - **Can do text-to-image but cant tell you what text should be there with a localization pattern**
- Autoregressive generation is slow



# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- **Unidirectional**
  - **Can do text-to-image but cant tell you what text should be there with a localization pattern**
- Autoregressive generation is slow

**Encoder-Only**

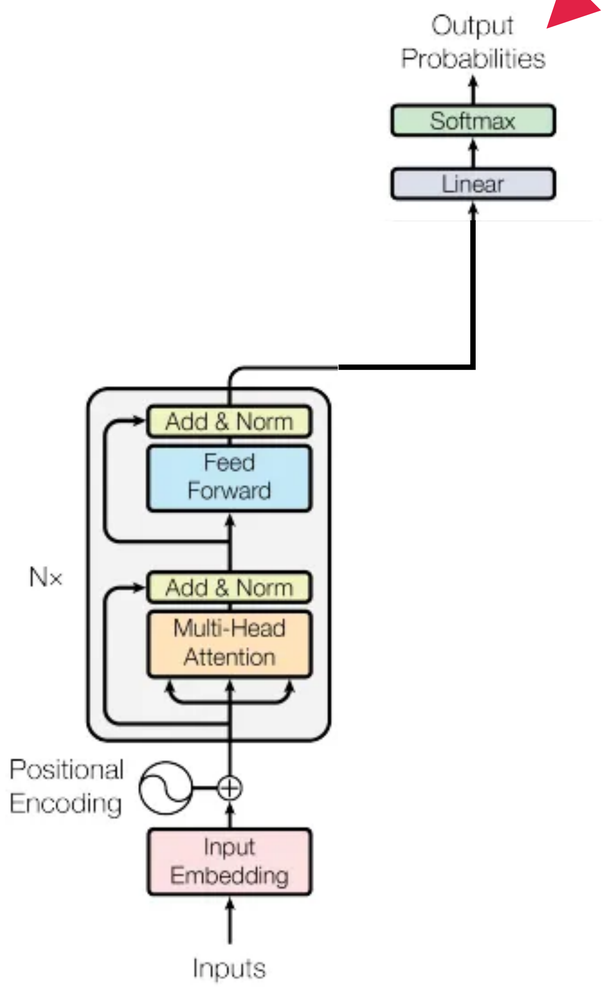


# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- **Unidirectional**
  - **Can do text-to-image but cant tell you what text should be there with a localization pattern**
- Autoregressive generation is slow

**Full Sequence Output**

**Encoder-Only**

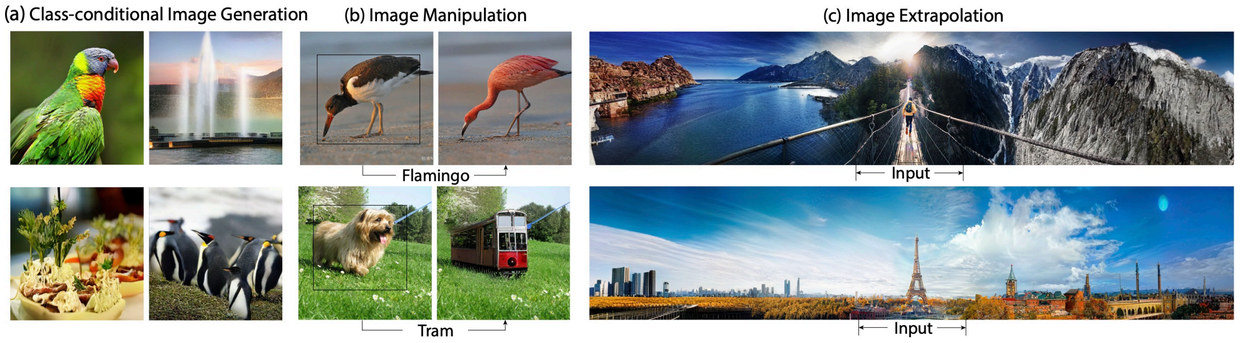


# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- Unidirectional
  - Can do text-to-image but cant tell you what text should be there with a localization pattern
- **Autoregressive generation is slow**

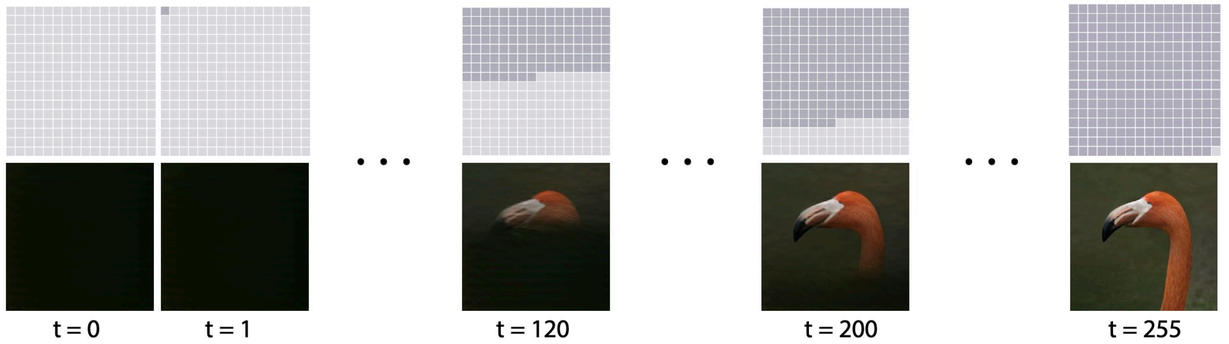
**MaskGIT: Masked Generative Image Transformer**

Huiwen Chang Han Zhang Lu Jiang Ce Liu* William T. Freeman
Google Research



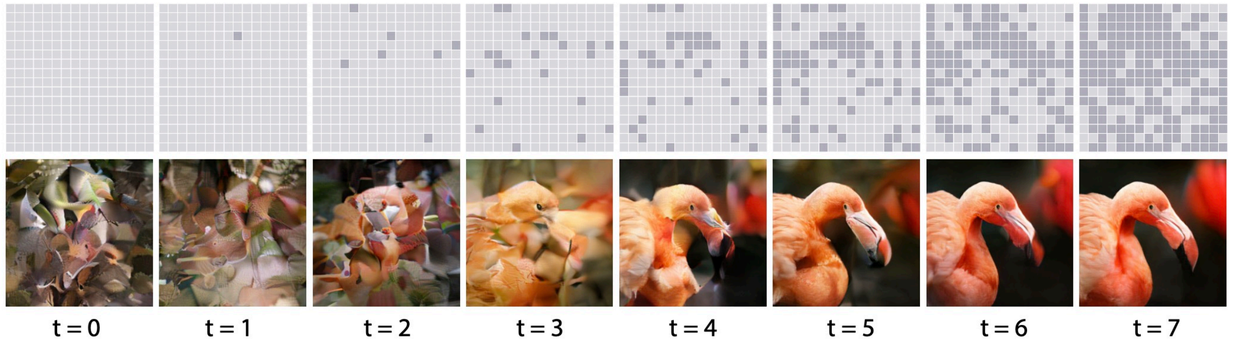
# Drawbacks

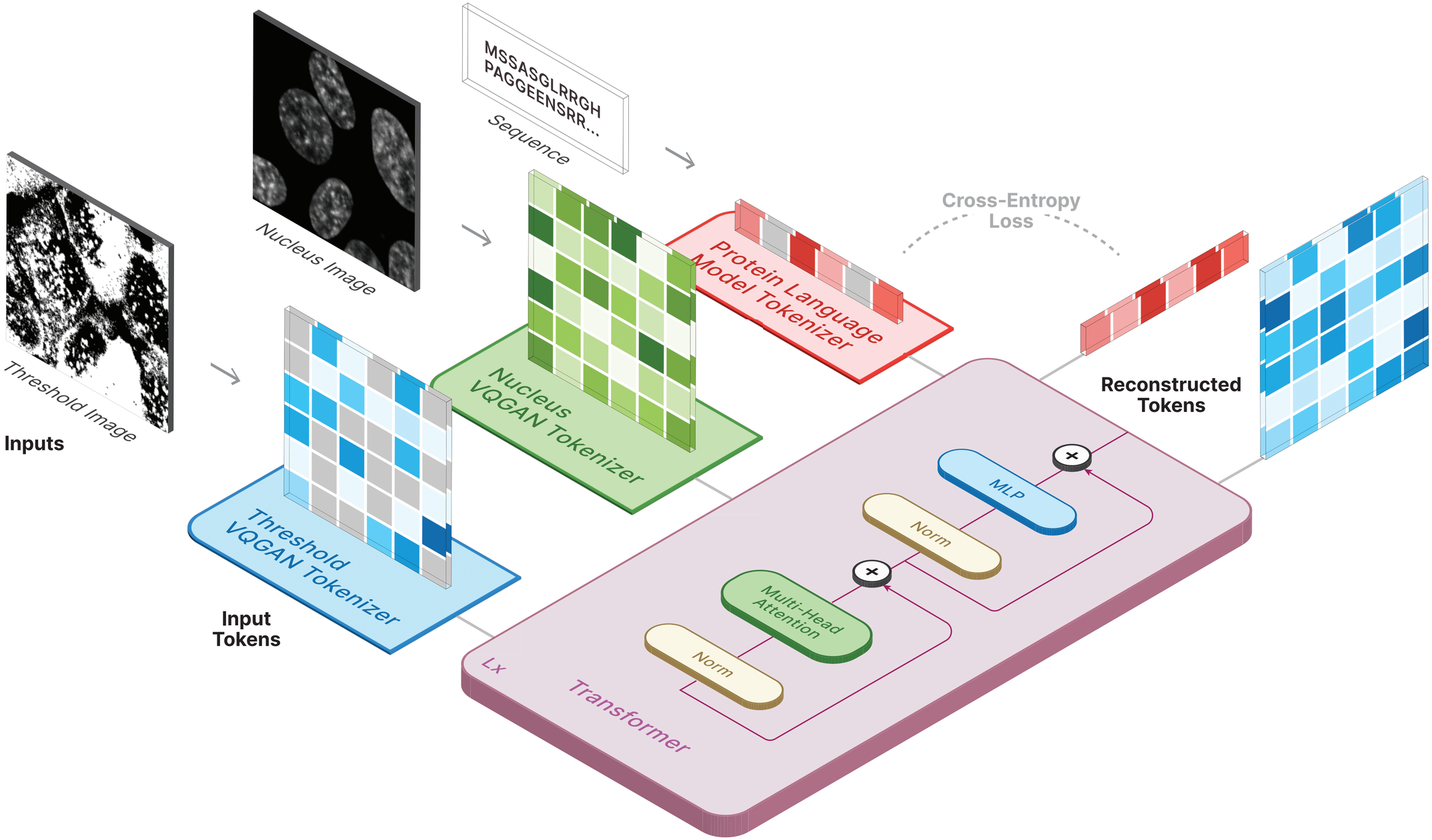
- Small dataset
  - Not many proteins in OpenCell
- Unidirectional
  - Can do text-to-image but cant tell you what text should be there with a localization pattern
- **Autoregressive generation is slow**

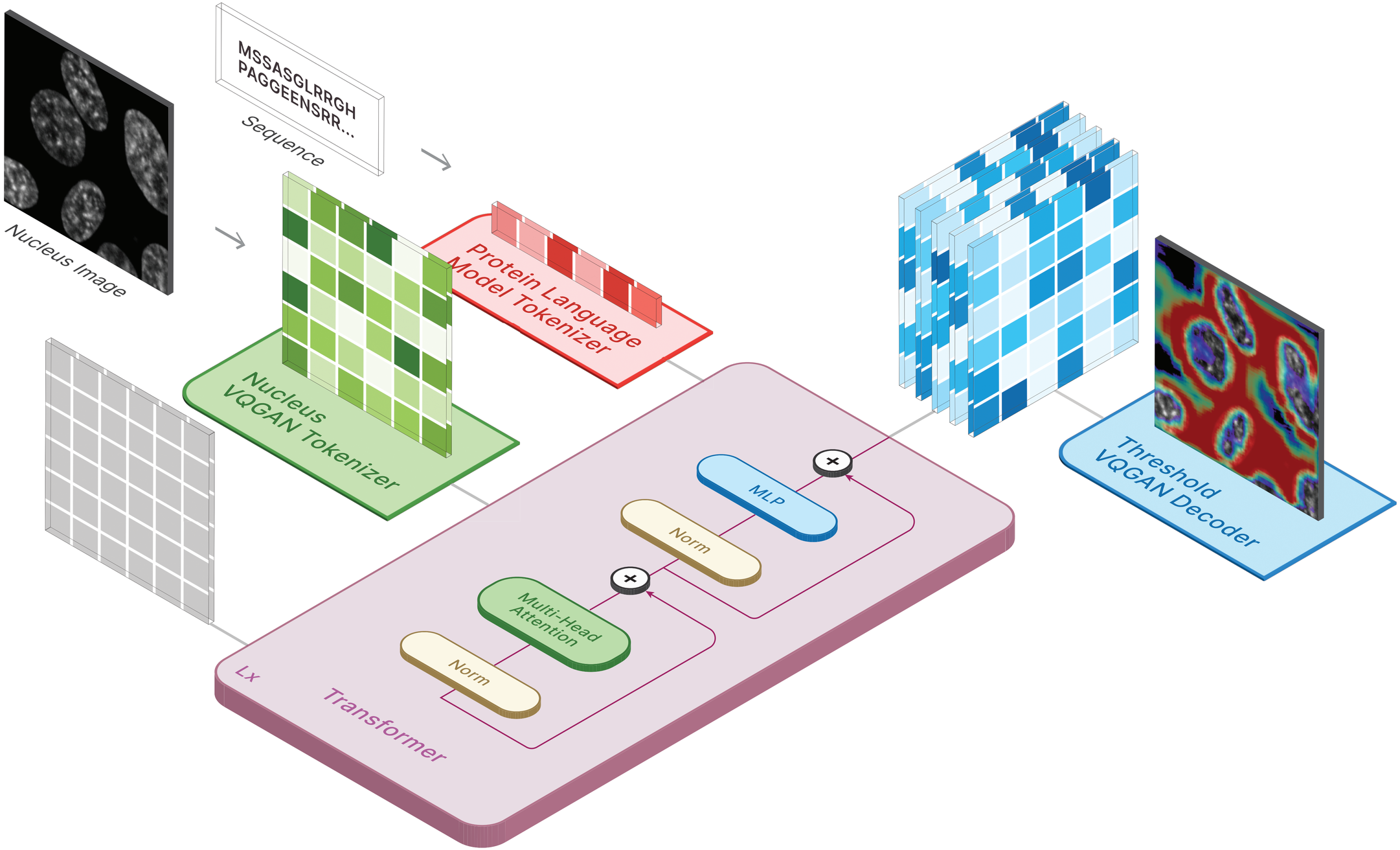


# Drawbacks

- Small dataset
  - Not many proteins in OpenCell
- Unidirectional
  - Can do text-to-image but cant tell you what text should be there with a localization pattern
- **Autoregressive generation is slow**



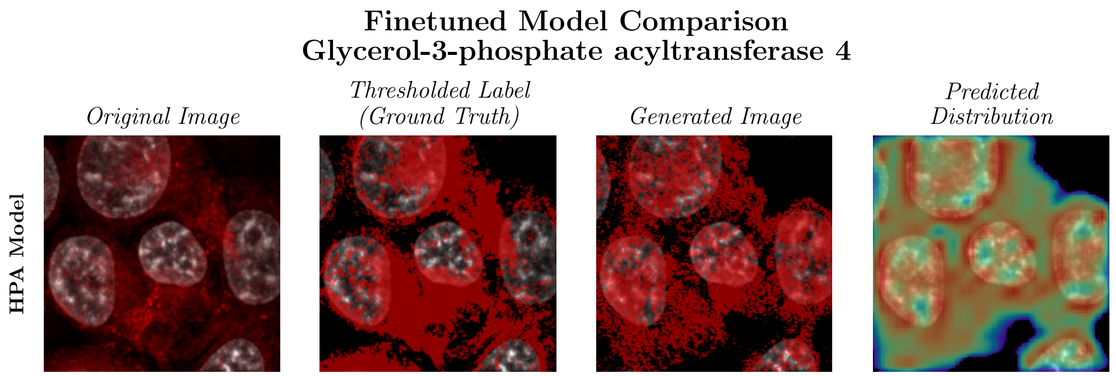

MSSASGLRRGH
PAGGEENSRR...
Sequence
Nucleus Image
Threshold Image
Inputs
Protein Language
Model Tokenizer
Nucleus
VQGAN Tokenizer
Threshold
VQGAN Tokenizer
Input
Tokens
Cross-Entropy
Loss
Reconstructed
Tokens
MLP
Norm
Multi-Head
Attention
Norm
Lx
Transformer

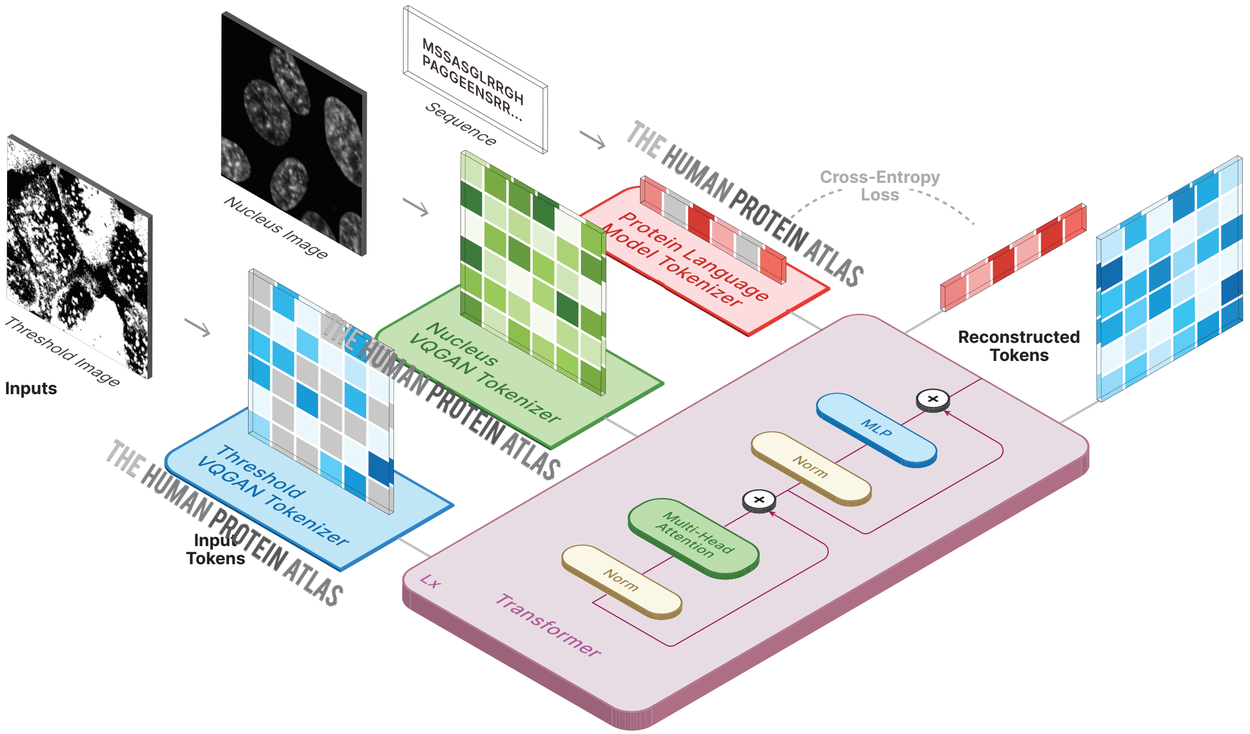

Nucleus Image
MSSASGLRRGH
PAGGEENSRR...
Sequence
Protein Language
Model Tokenizer
Nucleus
VQGAN Tokenizer
Threshold
VQGAN Decoder
MLP
Norm
Multi-Head
Attention
Norm
Lx
Transformer

Original Image
Thresholded Label (Ground Truth)
Generated Image
Predicted Distribution
T-complex protein 1 subunit theta
Chromosome-associated kinesin KIF4A
Phosphatidylinositol-binding clathrin assembly protein
Proteasome assembly chaperone 2

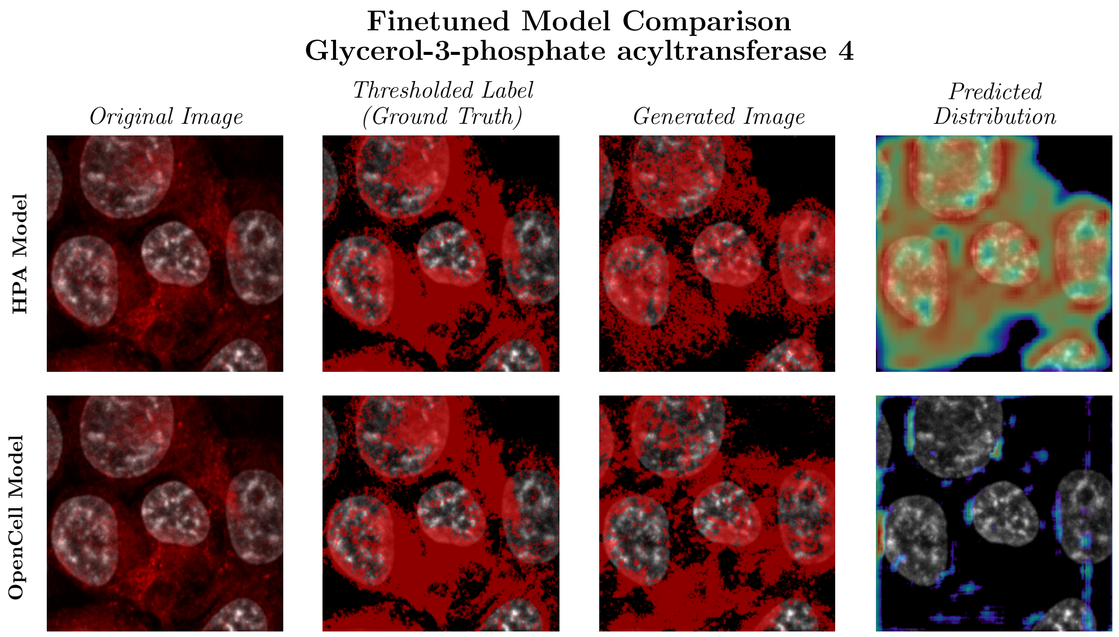
Original Image
Thresholded Label (Ground Truth)
Generated Image
Predicted Distribution
Eukaryotic translation initiation factor 5
Proteasome assembly chaperone 3
Phosphatidylinositol 4-phosphate 3-kinase C2 domain-containing subunit alpha
Probable ATP-dependent RNA helicase DDX5
Proteasome activator complex subunit 4
ADP-ribosylation factor-like protein 2-binding protein

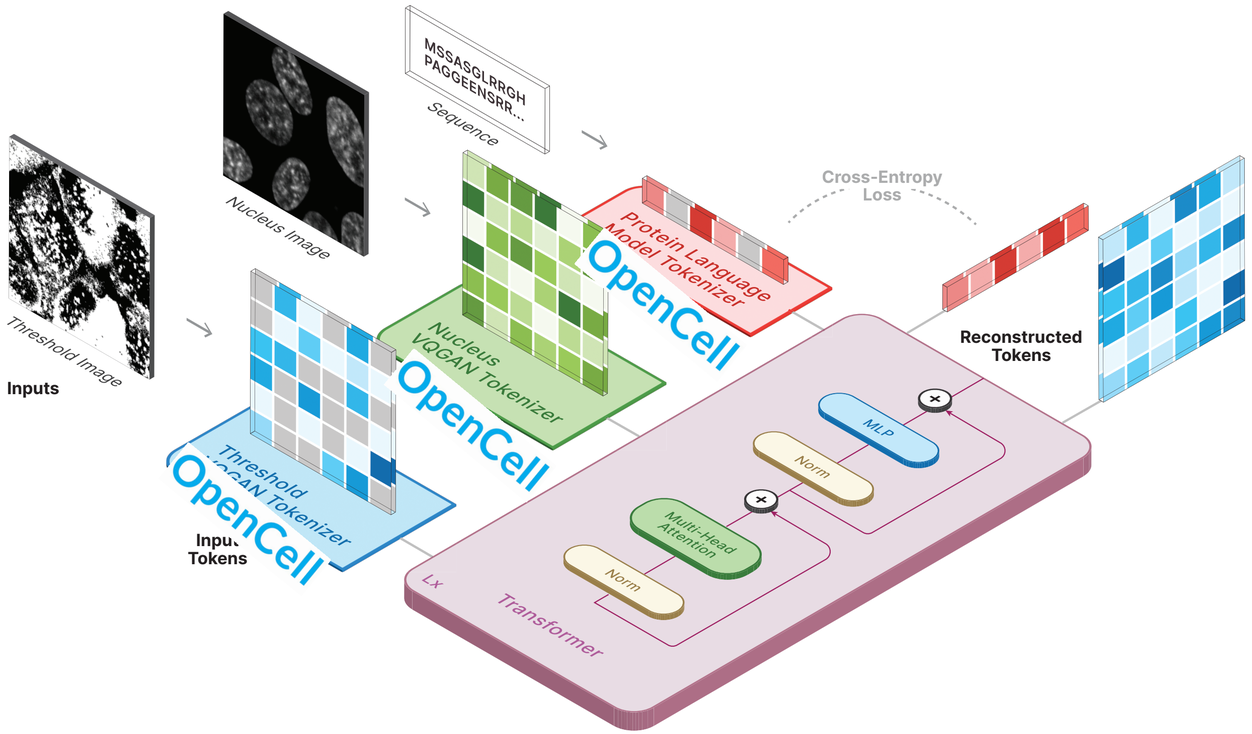
# OpenCell Finetuning



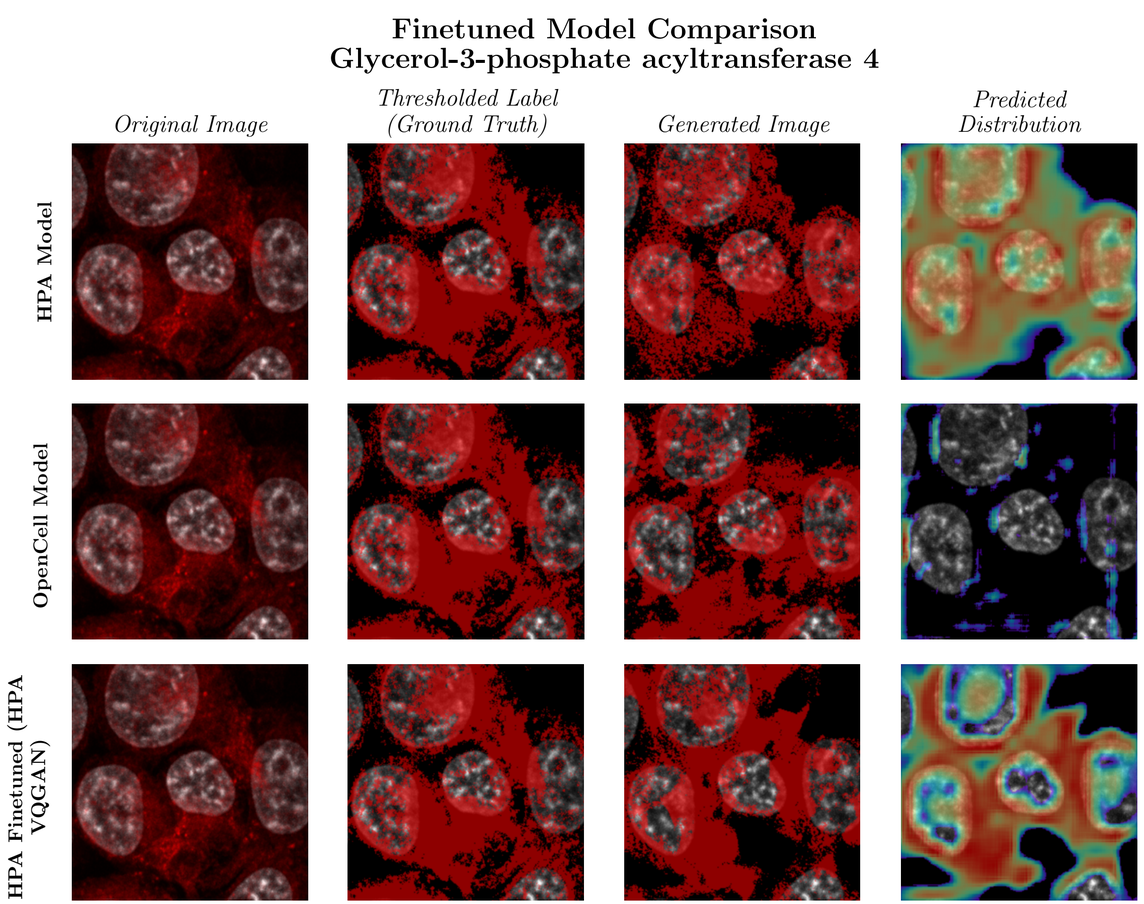



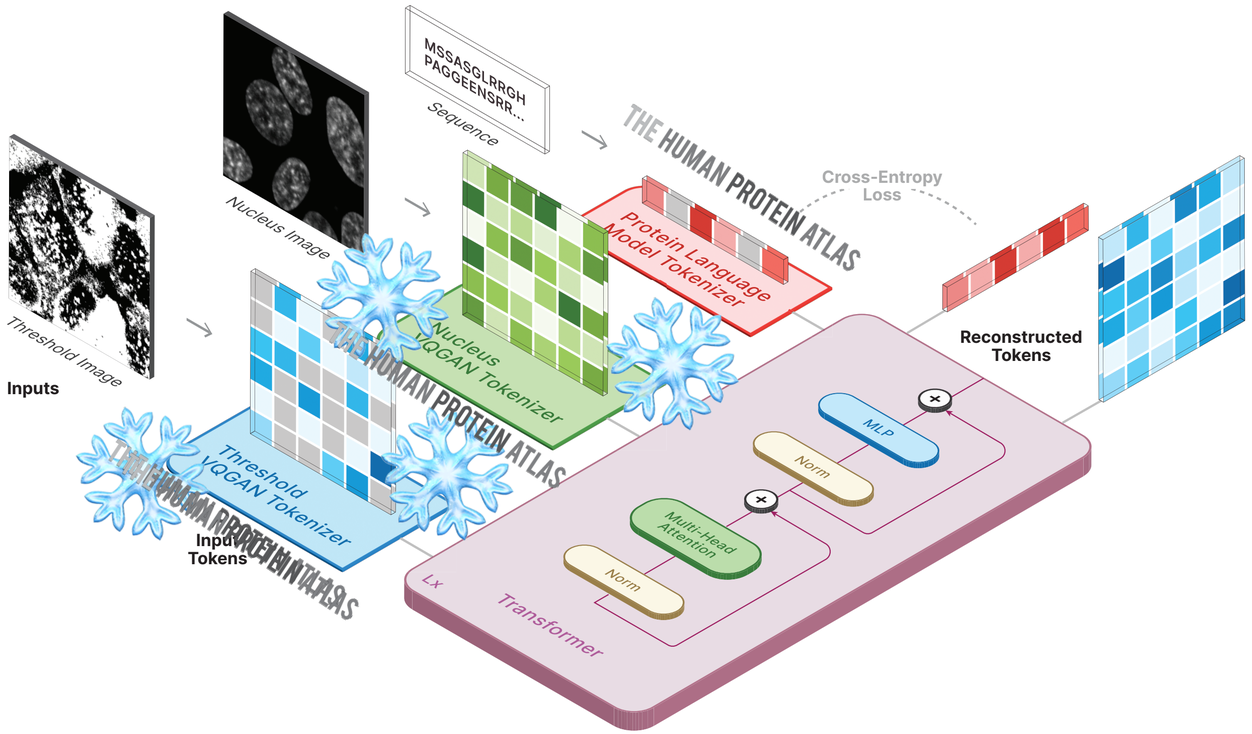
# OpenCell Finetuning

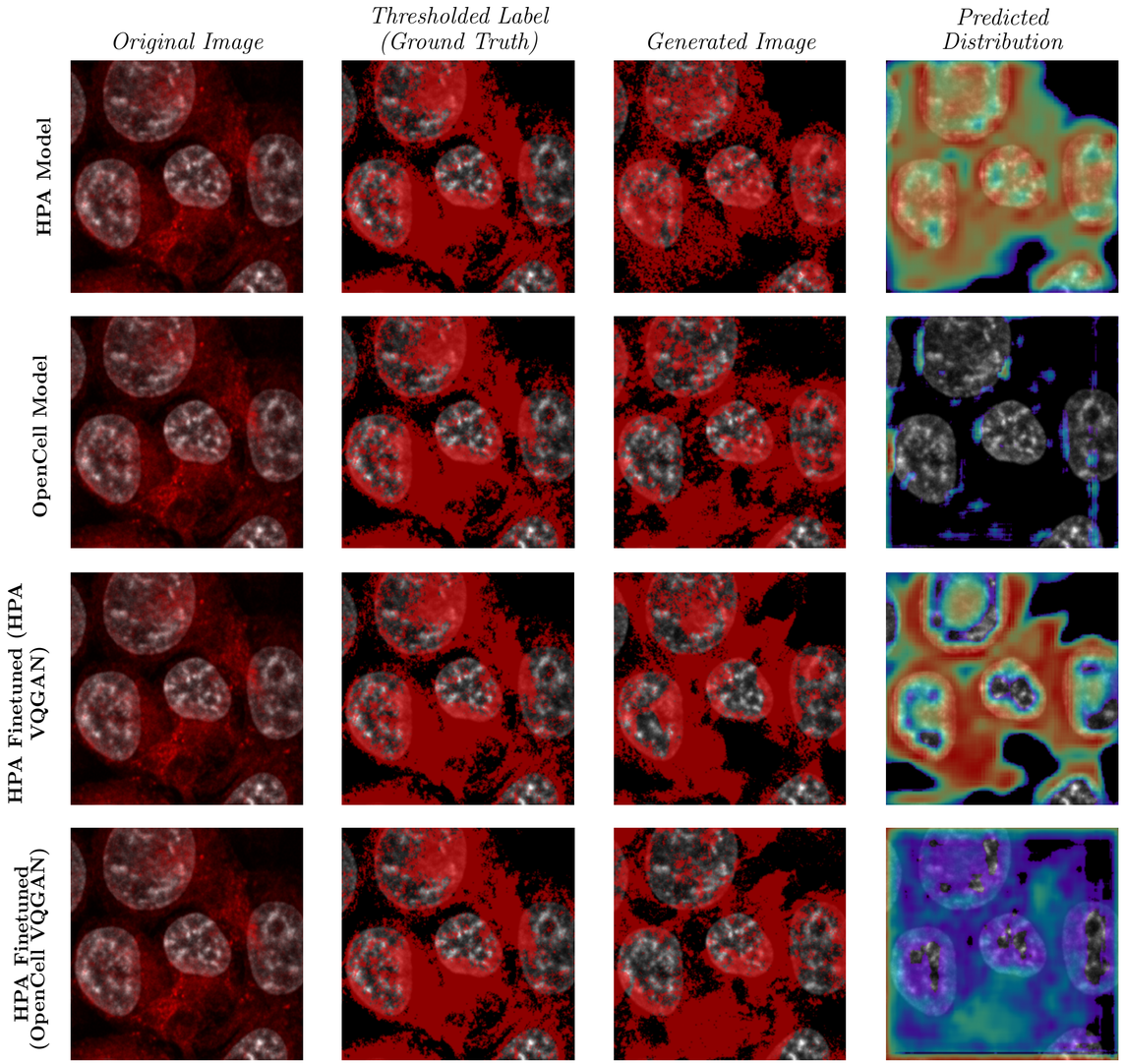


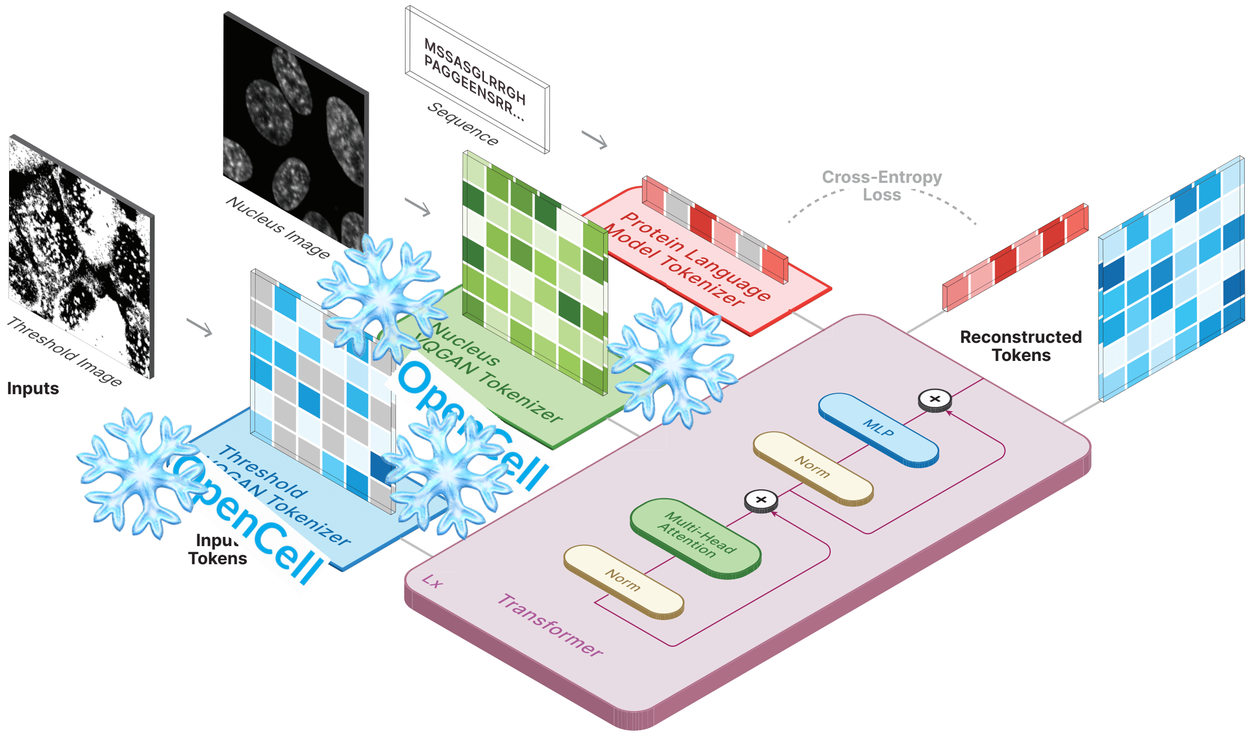
# OpenCell Finetuning



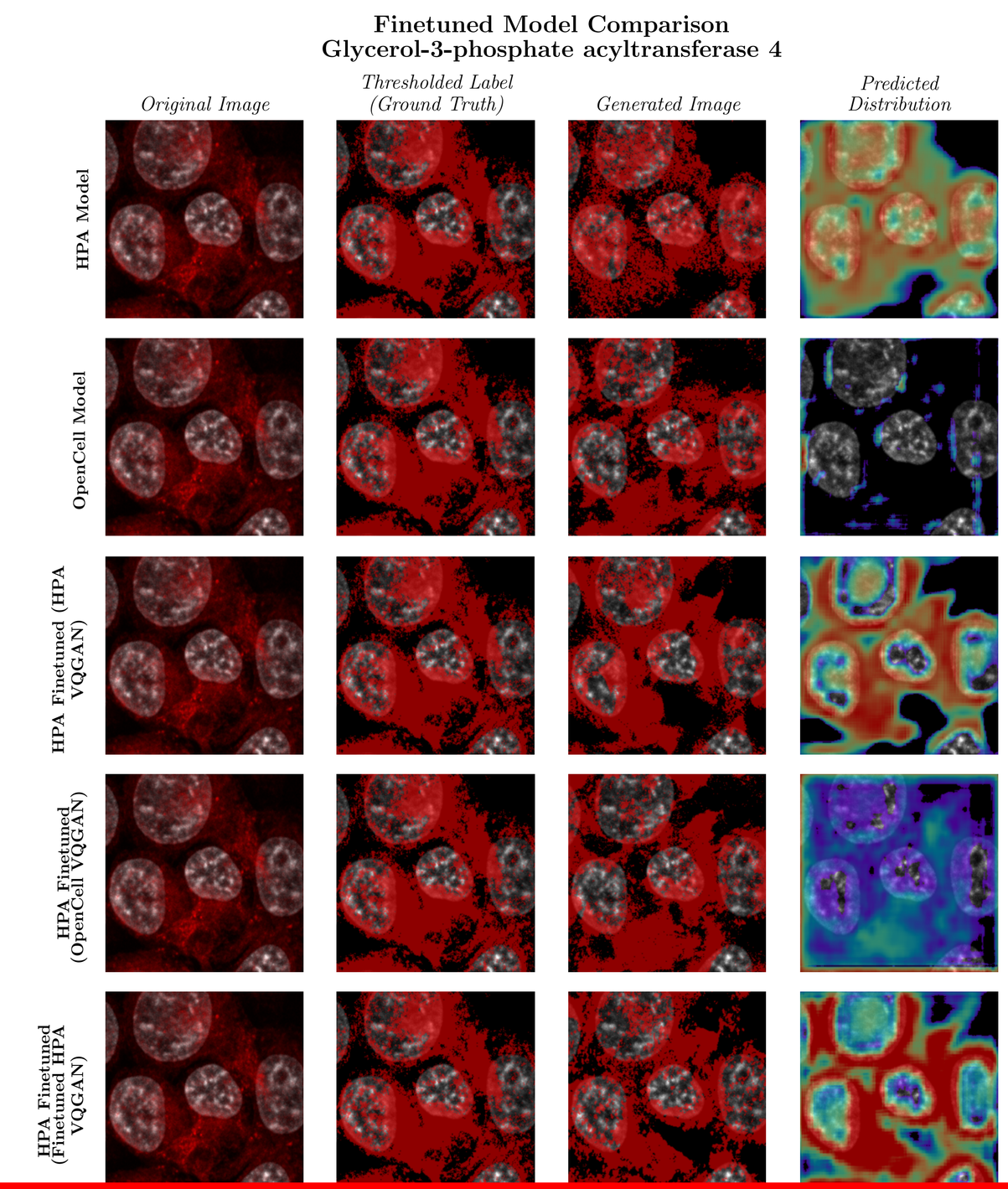



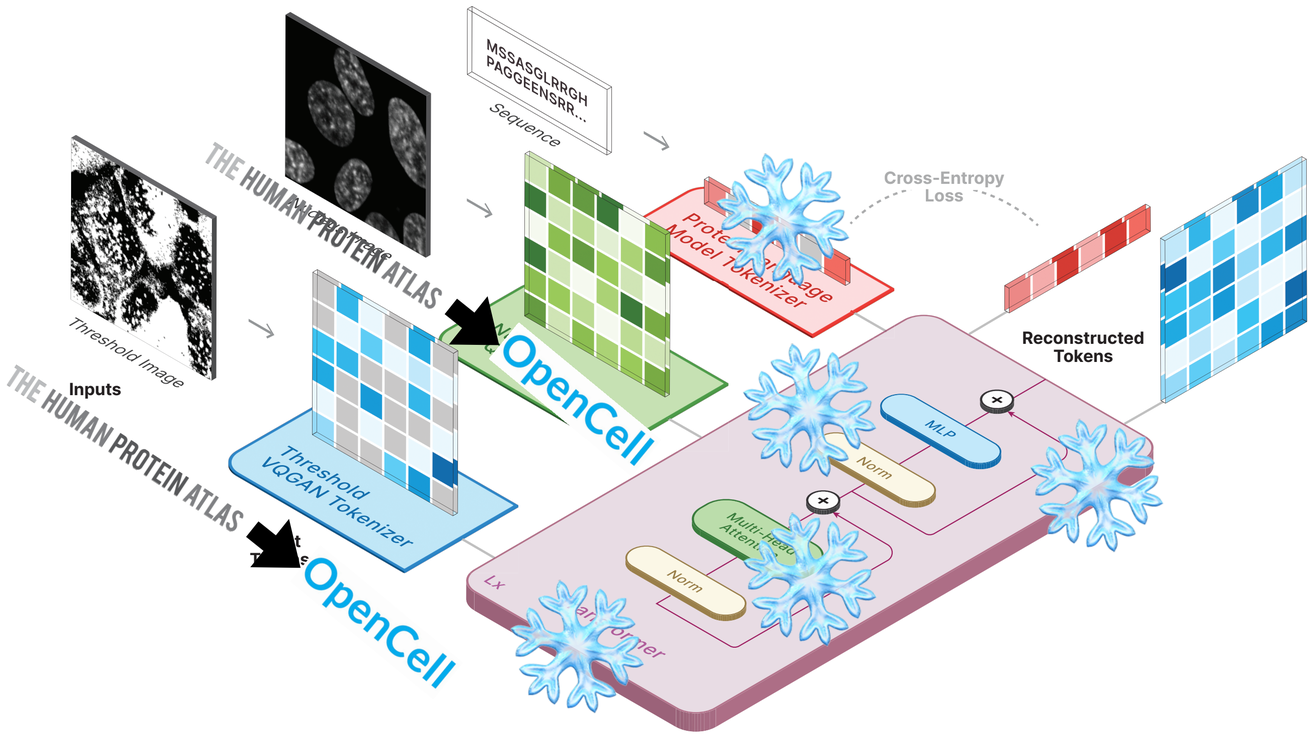



# OpenCell Finetuning

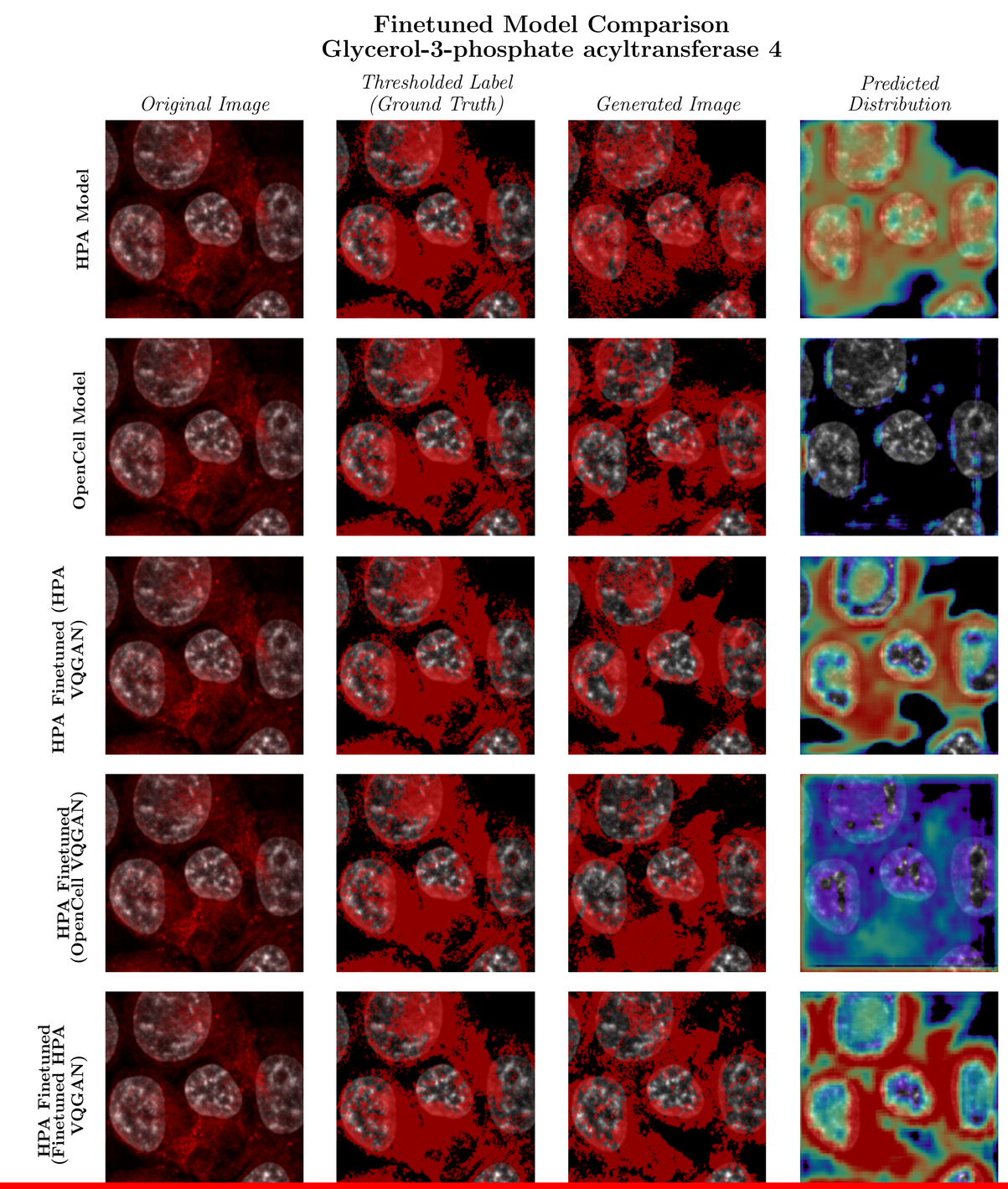


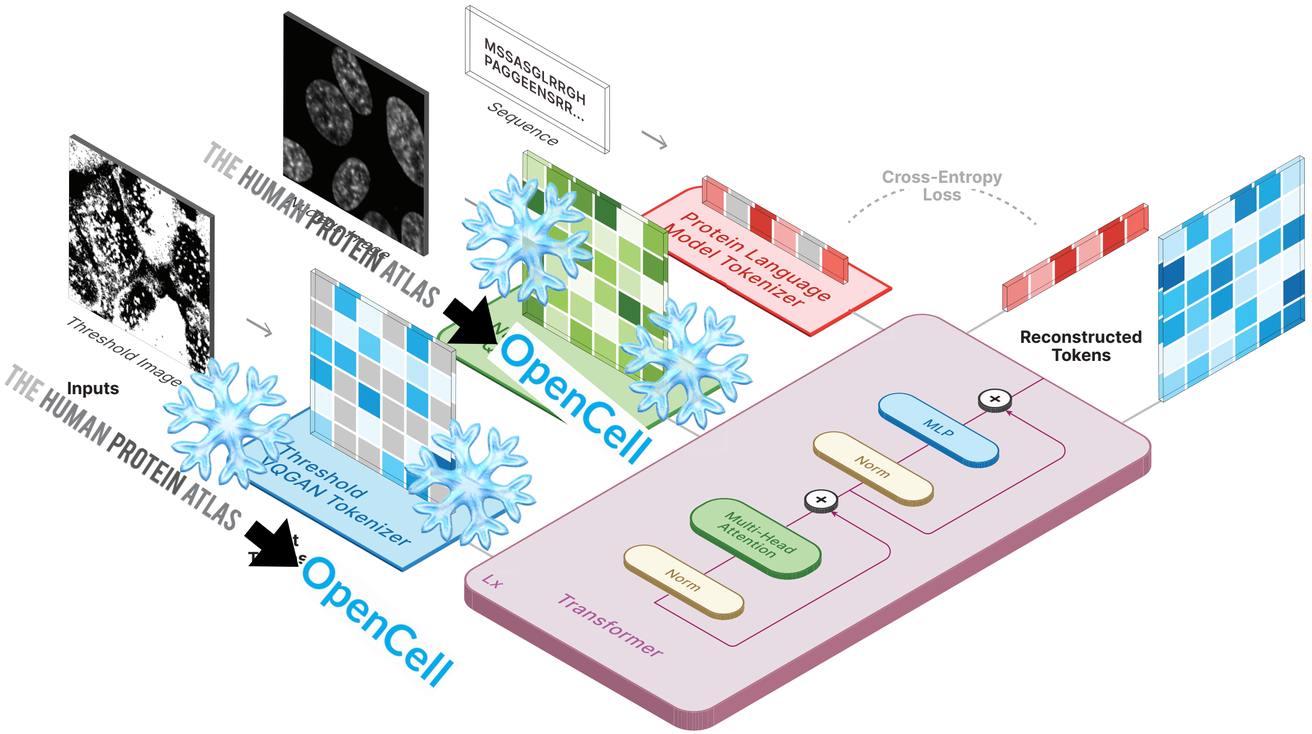
# OpenCell Finetuning



# OpenCell Finetuning



# CELL-E vs CELL-E 2

## On OpenCell Validation Set

| | Nucleus % MAE | Pixel Image MAE | PDF MAE | SSIM | FID | IS |
|---|---|---|---|---|---|---|
| **CELL-E** | .0347 ± .0294 | .3671 ± .1117 | .3653 ± .1008 | **.2060 ± .1846** | **10.5555** | 2.4762 ± .0866 |
| **CELL-E 2** | **.0170 ± .0160** | **.3449 ± .1305** | **.3487 ± .1340** | .1881 ± .1541 | 19.2683 | **3.6083 ± .2013** |

Table S8: Speed Comparison

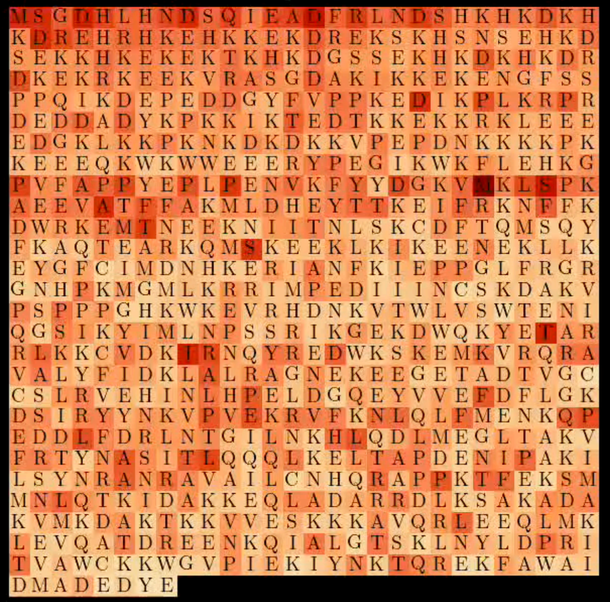
| Model | Hidden Size | Autoregressive | Mean Generation Time (s) |
|---|---|---|---|
| CELL-E (Cached) | 768 | Yes | $18.2740 \pm 0.0451$ |
| CELL-E (Non-Cached) | 768 | Yes | $28.7694 \pm 0.3207$ |
| CELL-E 2 | 480 | Yes | $55.0057 \pm 0.2069$ |
| CELL-E 2 | 640 | Yes | $62.9650 \pm 0.1033$ |
| CELL-E 2 | 1280 | Yes | $74.3698 \pm 0.1788$ |
| CELL-E 2 | 2560 | Yes | $128.9960 \pm 0.3718$ |
| **CELL-E 2** | **480** | **No** | $\mathbf{0.2784 \pm 0.0006}$ |
| CELL-E 2 | 640 | No | $0.3067 \pm 0.0012$ |
| CELL-E 2 | 1280 | No | $0.3249 \pm 0.0011$ |
| CELL-E 2 | 2560 | No | $0.5487 \pm 0.0022$ |

# 66x Faster with Better Performance!

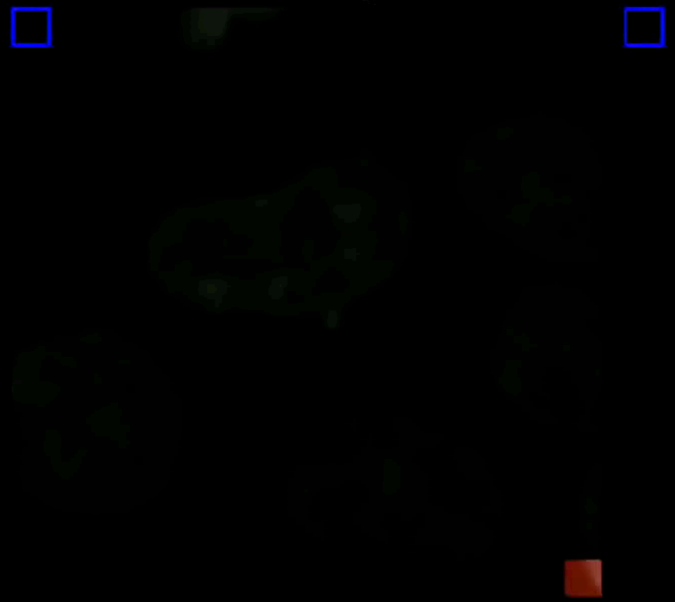
CELL-E 1

DNA Topoisomerase I
Relative Attention Weights

*Sequence*



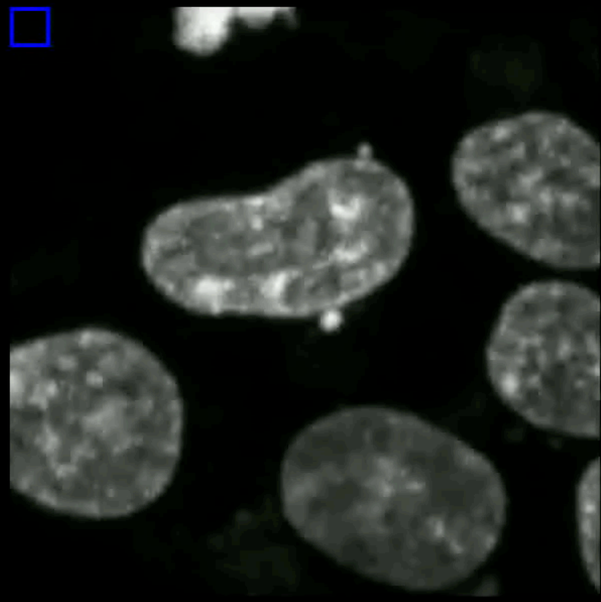
*Nucleus Image*



*Predicted Threshold Image*



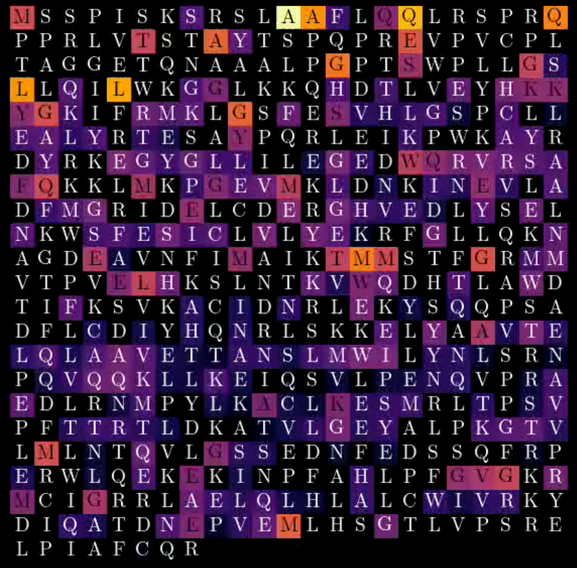
*Predicted PDF*



CELL-E 2

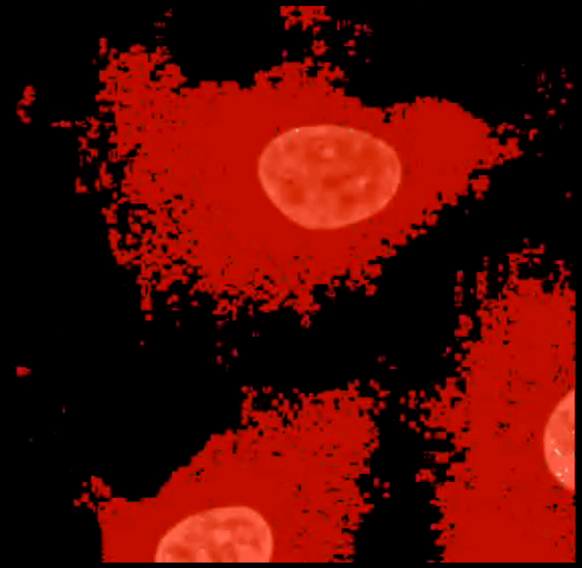
1,25-dihydroxyvitamin D(3) 24-hydroxylase, mitochondrial
Relative Attention Weights

*Sequence*

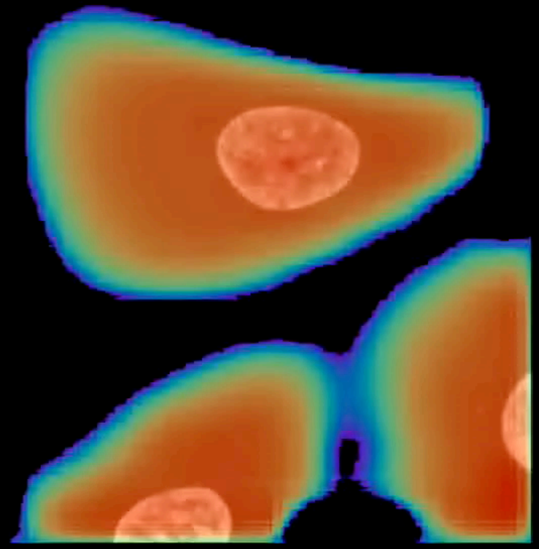


*Nucleus Image*

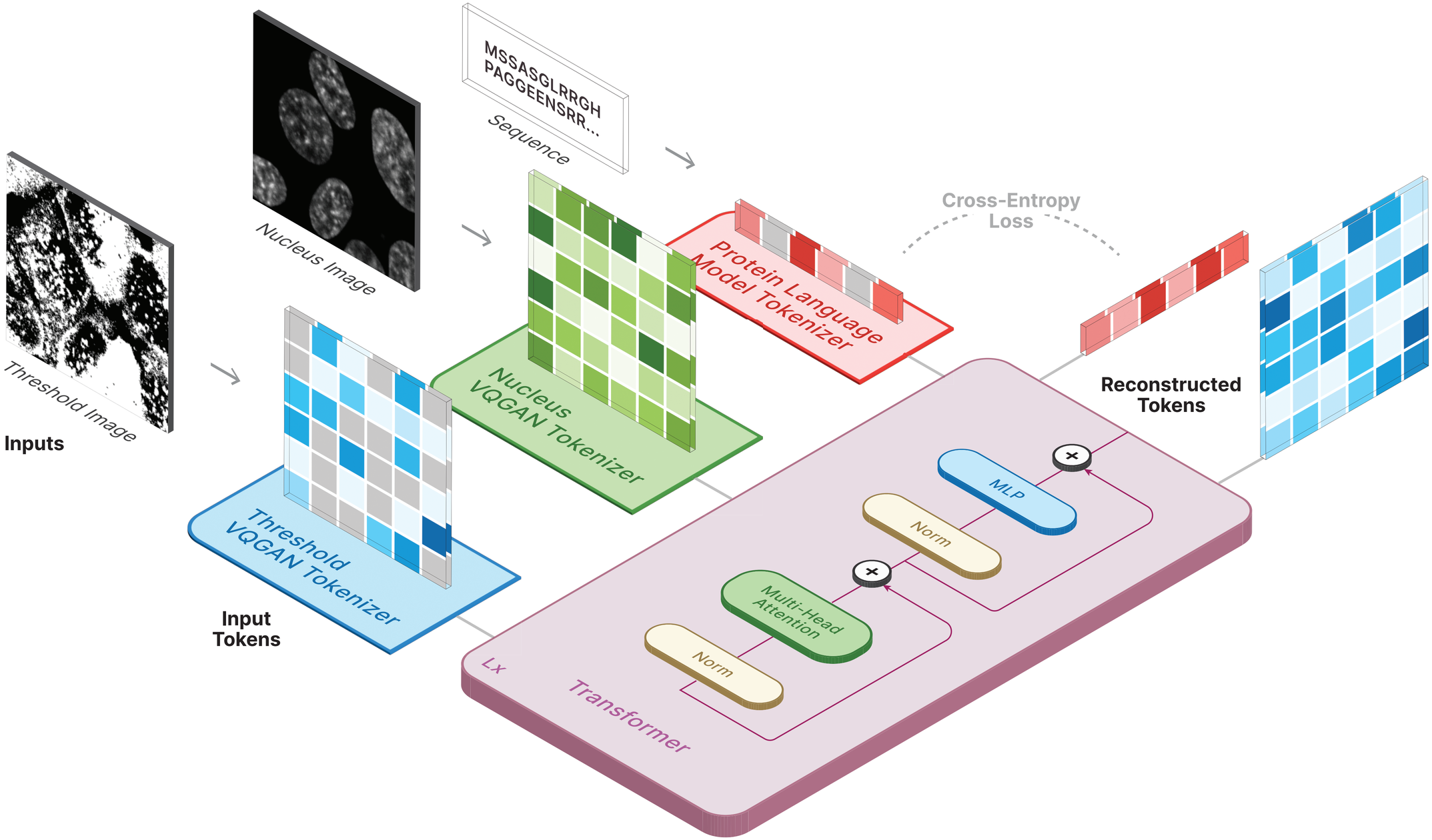
*Predicted Threshold Image*

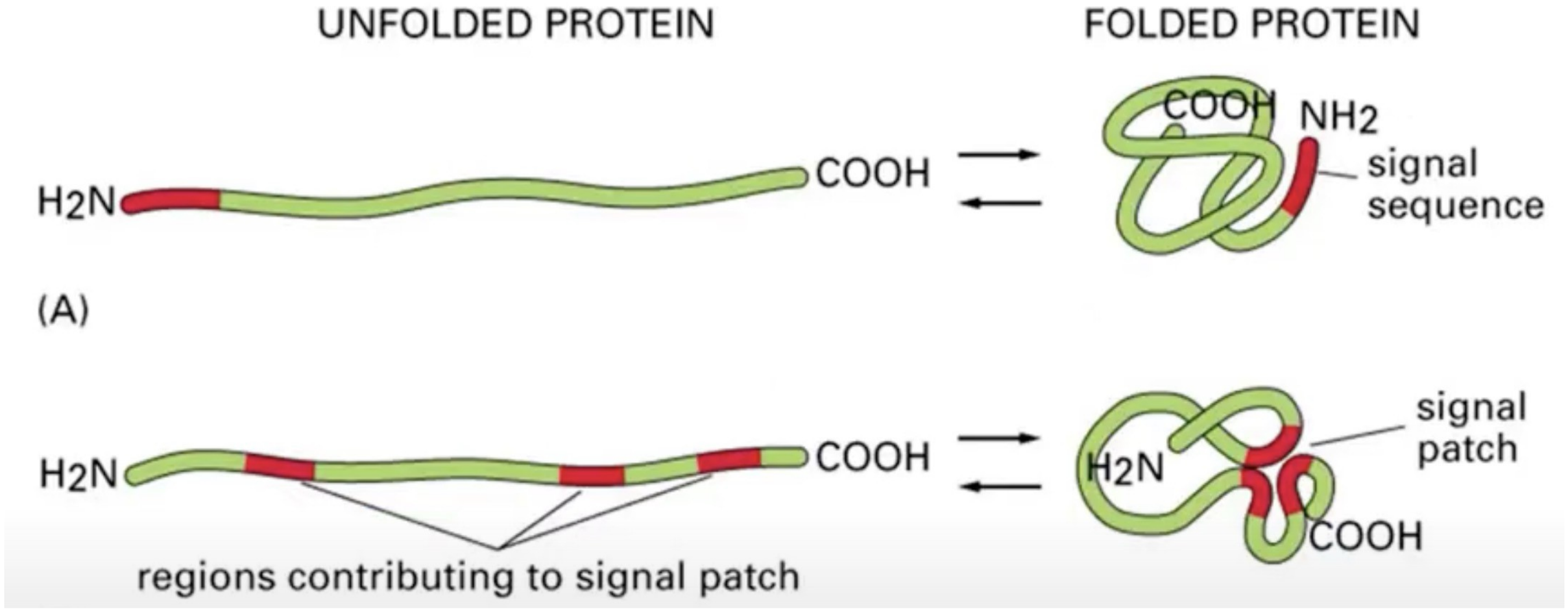


*Predicted PDF*



# *De novo* Protein Design


MSSASGLRRGH
PAGGEENSRR...
Sequence
Nucleus Image
Threshold Image
Inputs
Protein Language
Model Tokenizer
Nucleus
VQGAN Tokenizer
Threshold
VQGAN Tokenizer
Input
Tokens
Cross-Entropy
Loss
Reconstructed
Tokens
MLP
Norm
Multi-Head
Attention
Norm
Lx
Transformer




*Source: Molecular Biology of the Cell, 4th Edition*

## NLSdb sequence search results

Show 10 entries

Copy CSV

Previous 1 Next

| Query | Signal | SignalType | Start | End | ConfNuc | ConfFam | AnnotationType | Origin | Added | Modified |
|---|---|---|---|---|---|---|---|---|---|---|
| query | DKEKRK | NLS | 85 | 90 | 3 | 3 | Potential | In Silico Mutagenesis | 2017-10-11 | 2017-10-11 |
| query | KEKKRK | NLS | 159 | 164 | 6 | 6 | Potential | In Silico Mutagenesis | 2017-10-11 | 2017-10-11 |
| query | KKRKLE | NLS | 161 | 166 | 14 | 13 | By Expert | P02545, P48678, P48679, Q3ZD69 | 2017-10-11 | 2017-10-11 |
| query | PLKRPR | NLS | 135 | 140 | 4 | 3 | Potential | In Silico Mutagenesis | 2017-10-11 | 2017-10-11 |
| All | All | All | | | | | All | All | | |

Showing 1 to 4 of 4 entries

Previous 1 Next

# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

```
MSKGEELFTGVVPILVELDGDVN
GHKFSVSGEGEGDATYGKLTLKF
ICTTGKLPVPWPTLVTTFSYGVQ
CFSRYPDHMKQHDFFKSAMPEGY
VQERTIFFKDDGNYKTRAEVKFE
GDTLVNRIELKGIDFKEDGNILG
HKLEYNYNSHNVYIMADKQKNGI
KVNFKIRHNIEDGSVQLADHYQQ
NTPIGDGPVLLPDNHYLSTQSAL
SKDPNEKRDHMVLLEFVTAAGIT
HGMDELYK
```

# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

<start>MSKGEELFTGVVPILV
ELDGDVNGHKFSVSGEGEGDATY
GKLTLKFICTTGKLPVPWPTLVT
TFSYGVQCFSRYPDHMKQHDFFK
SAMPEGYVQERTIFFKDDGNYKT
RAEVKFEGDTLVNRIELKGIDFK
EDGNILGHKLEYNYNSHNVYIMA
DKQKNGIKVNFKIRHNIEDGSVQ
LADHYQQNTPIGDGPVLLPDNHY
LSTQSALSKDPNEKRDHMVLLEF
VTAAGITHGMDELYK<end>

# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

<start><mask>MSKGEELFTG
VVPILVELDGDVNGHKFSVSGEG
EGDATYGKLTLKFICTTGKLPVP
WPTLVTTFSYGVQCFSRYPDHMK
QHDFFKSAMPEGYVQERTIFFKD
DGNYKTRAEVKFEGDTLVNRIEL
KGIDFKEDGNILGHKLEYNYNSH
NVYIMADKQKNGIKVNFKIRHNI
EDGSVQLADHYQQNTPIGDGPVL
LPDNHYLSTQSALSKDPNEKRDH
MVLLEFVTAAGITHGMDELYK<e
nd>

# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

<start><mask><mask><mask>MSKGEELFTGVVPILVELDGD
VNGHKFSVSGEGEGDATYGKLTL
KFICTTGKLPVPWPTLVTTFSYG
VQCFSRYPDHMKQHDFFKSAMPE
GYVQERTIFFKDDGNYKTRAEVK
FEGDTLVNRIELKGIDFKEDGNI
LGHKLEYNYNSHNVYIMADKQKN
GIKVNFKIRHNIEDGSVQLADHY
QQNTPIGDGPVLLPDNHYLSTQS
ALSKDPNEKRDHMVLLEFVTAAG
ITHGMDELYK<end>

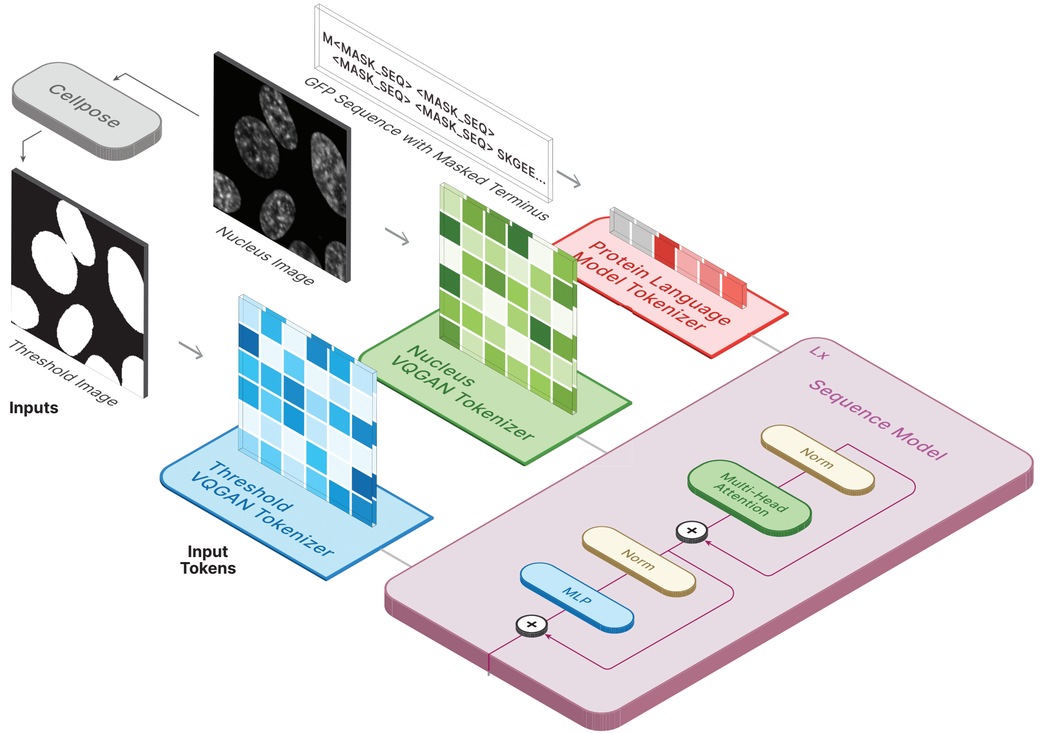
# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

<start>MSKGEELFTGVVPILV
ELDGDVNGHKFSVSGEGEGDATY
GKLTLKFICTTGKLPVPWPTLVT
TFSYGVQCFSRYPDHMKQHDFFK
SAMPEGYVQERTIFFKDDGNYKT
RAEVKFEGDTLVNRIELKGIDFK
EDGNILGHKLEYNYNSHNVYIMA
DKQKNGIKVNFKIRHNIEDGSVQ
LADHYQQNTPIGDGPVLLPDNHY
LSTQSALSKDPNEKRDHMVLLEF
VTAAGITHGMDELYK<mask><m
ask><mask><end>

# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate



# Procedure

- Pre-select number of masked positions
- Generate 300 candidates per terminus with sequence model
- Feed predicted sequence to image model to cross-validate

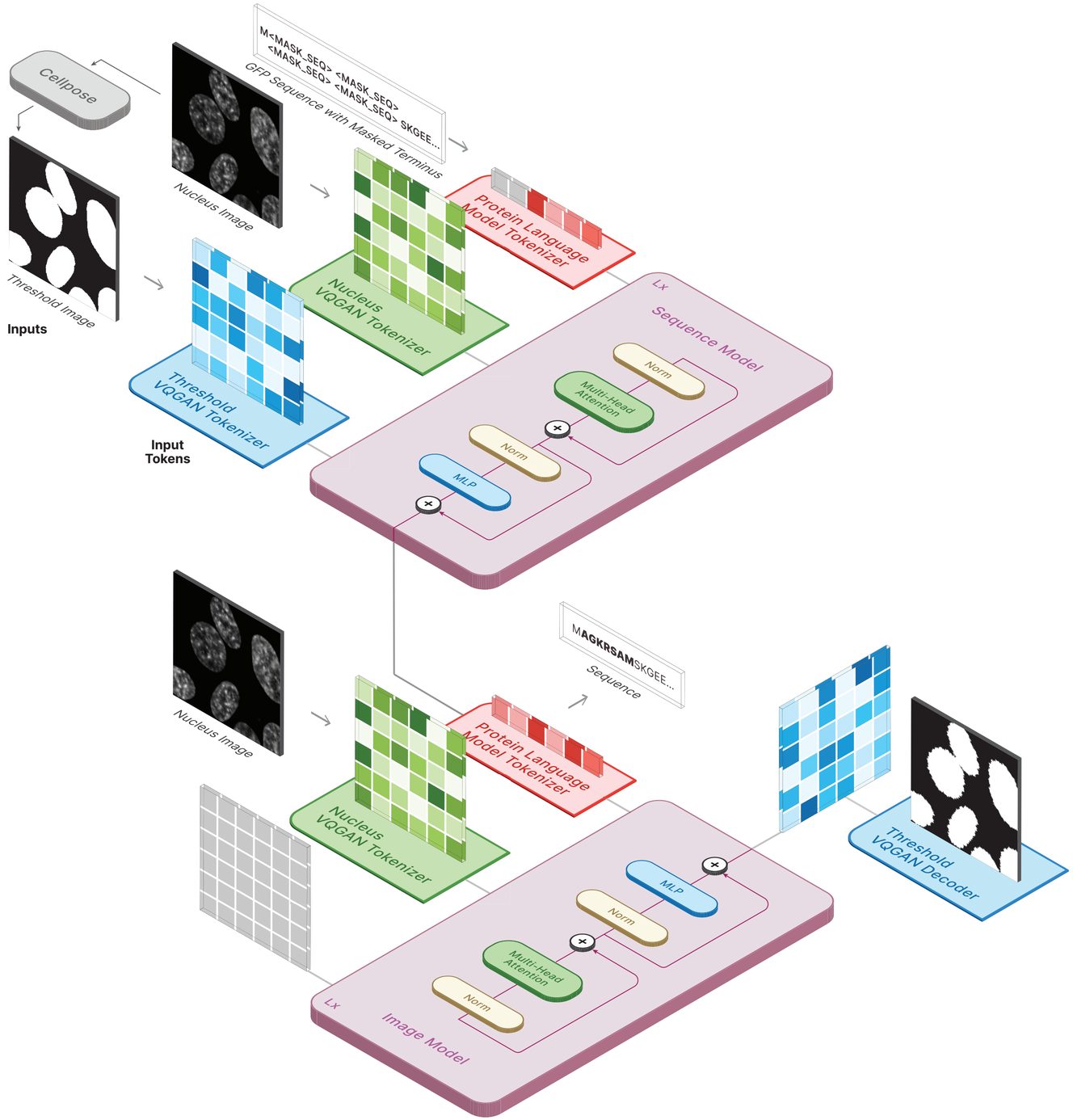


Table S10: NLS candidates sorted by nucleus proportion.

| Terminus | Sequence | Terminus | Sequence |
|---|---|---|---|
| N | RKRRQR | C | SPTAFPSNVIETIRVKRRMEL |
| N | NKRPRKKEK | C | EFRAKYRQMGSRKKKKSGQWSA |
| C | RPKVI | N | KKHKLRSVPDLTELMRMIFLAP |
| C | VLKRAKKD | N | KLLRFAGKSGMMVLLAPHSGKM |
| C | RHKKKKIA | C | IFQADKDQKAHPPAKKAPSELMQ |
| N | HRRKKR | C | KGKVKSIMIPPKSRKSLAKVPLS |
| C | RSQKRK | N | AAGKSFKPRIKKSRMTRDSSETMA |
| N | KCKKKN | C | TGNRIFGETPSWERERKRPGGGQQ |
| N | KGKRFSK | C | NKLQKHSKRQPHKLQAMKLKYPTWE |
| C | AKRLKGK | C | LVFPNRDASIKKPLQNPPQKRRCMIM |
| C | SKKAKKNKM | N | LPKRRRLSRRKKVELEPEYGWEEEVVV |
| C | EEKRPRF | N | TEAPARTAVKKSRAMKGYIARLASSPS |
| N | MKICIT | C | IEKSKGKEAPKSSPPLKQNQRSRKMVK |
| N | AVPAKRARIDG | C | FQVRASPKGKPATKNKLRLLKIRRHRV |
| C | ESHHLPRAKKR | C | LQEGTRTRSQKAQEPKFKKVSGDIPNK |
| N | GKERSYPPISKR | N | SDPNTAQYPWMPPQATKRAAMAAREAE |
| C | KLKKRNRQPEDKK | C | HYKKEKRKRSASPILAEEPVPKCARTLR |
| C | GGKFATGKKKKPKM | C | LDKRKRIKPPKEEQKELMRKMWGPGSSL |
| N | PSKLLRQ | N | GSKKSRTATDSLESRMAMEDVAMGEESE |
| C | QRRKGQKFQT | C | EGSGLVPGNSRKRPEPKKPKKRKKVRRK |
| C | KTCPPKRPVVEW | C | RKKRQAIQAVTMGRIKKKSYEKQWSKFED |
| C | DKEKKRKNDHEK | C | ASTVPAYSRSKAGKVEPKPKQKKTQRNAP |
| N | FRFSC | C | SKQQAEINLKAAKPLETTDISLSKKEKKDM |

# NLS Properties

- Short (5-15 Amino Acids)
- Highly Basic (R and K)
- Clustered

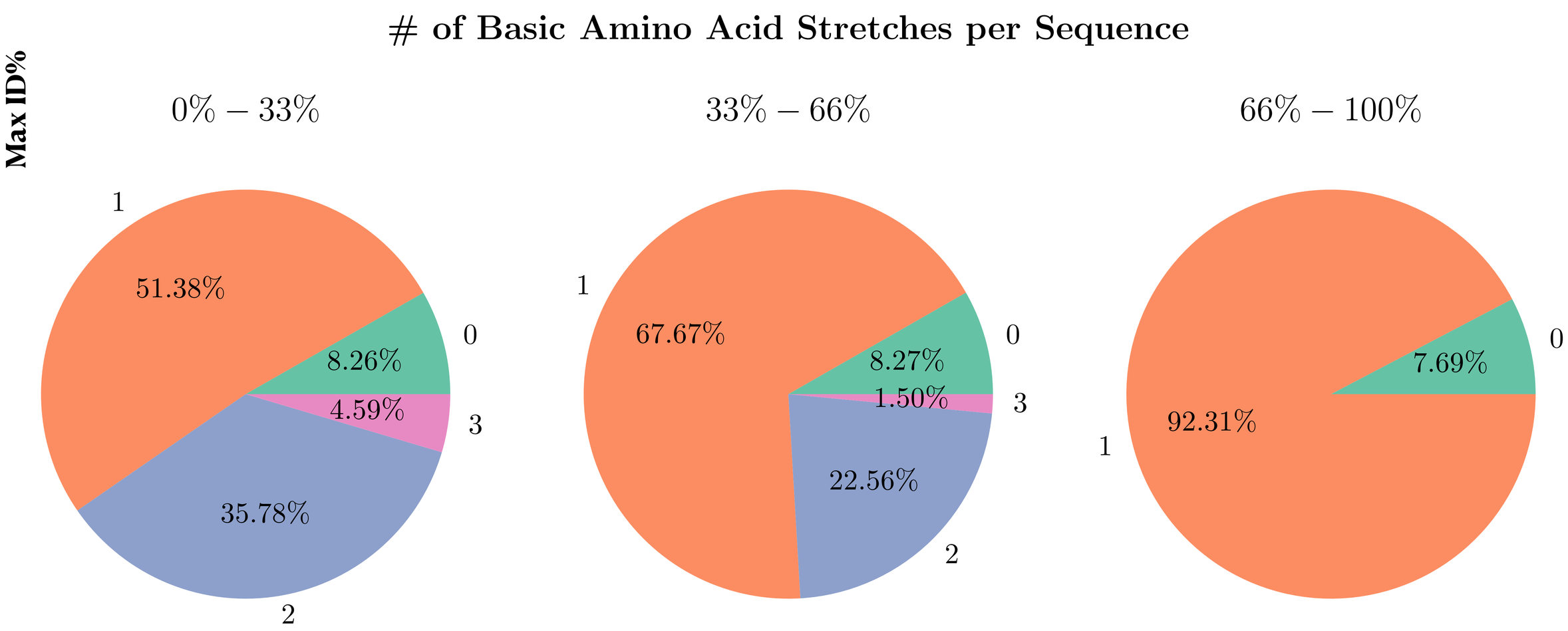

Monopartite NLSs


Bipartite NLSs

| Max ID % | # Sequences | Mean Sequence Length | Mean % R or K |
|---|---|---|---|
| 0% - 33% | 109 | $25.6606 \pm 3.0099$ | $20.6379 \pm 8.6101$ |
| 33% - 66% | 133 | $17.1955 \pm 5.0804$ | $32.0076 \pm 12.8334$ |
| 66% - 100% | 13 | $6.9231 \pm 1.2558$ | $57.5794 \pm 17.9351$ |

Highly Dissimilar

| Max ID % | # Sequences | Mean Sequence Length | Mean % R or K |
|---|---|---|---|
| 0% - 33% | 109 | 25.6606 ± 3.0099 | 20.6379 ± 8.6101 |
| 33% - 66% | 133 | 17.1955 ± 5.0804 | 32.0076 ± 12.8334 |
| 66% - 100% | 13 | 6.9231 ± 1.2558 | 57.5794 ± 17.9351 |


Max ID%
# of Basic Amino Acid Stretches per Sequence
0% – 33%
33% – 66%
66% – 100%
1
51.38%
0
8.26%
4.59%
3
35.78%
2
1
67.67%
0
8.27%
1.50%
3
22.56%
2
0
7.69%
1
92.31%

# DeepLoc 2.0 Prediction:

- 89% Nuclear Localizing Proteins
- 91% contain NLS

# Future Directions

- Leverage more data
- Incorporate structural information
- Validate *de novo* candidates

# Thanks!