VirProBERT: Language Model for Virus Host Prediction

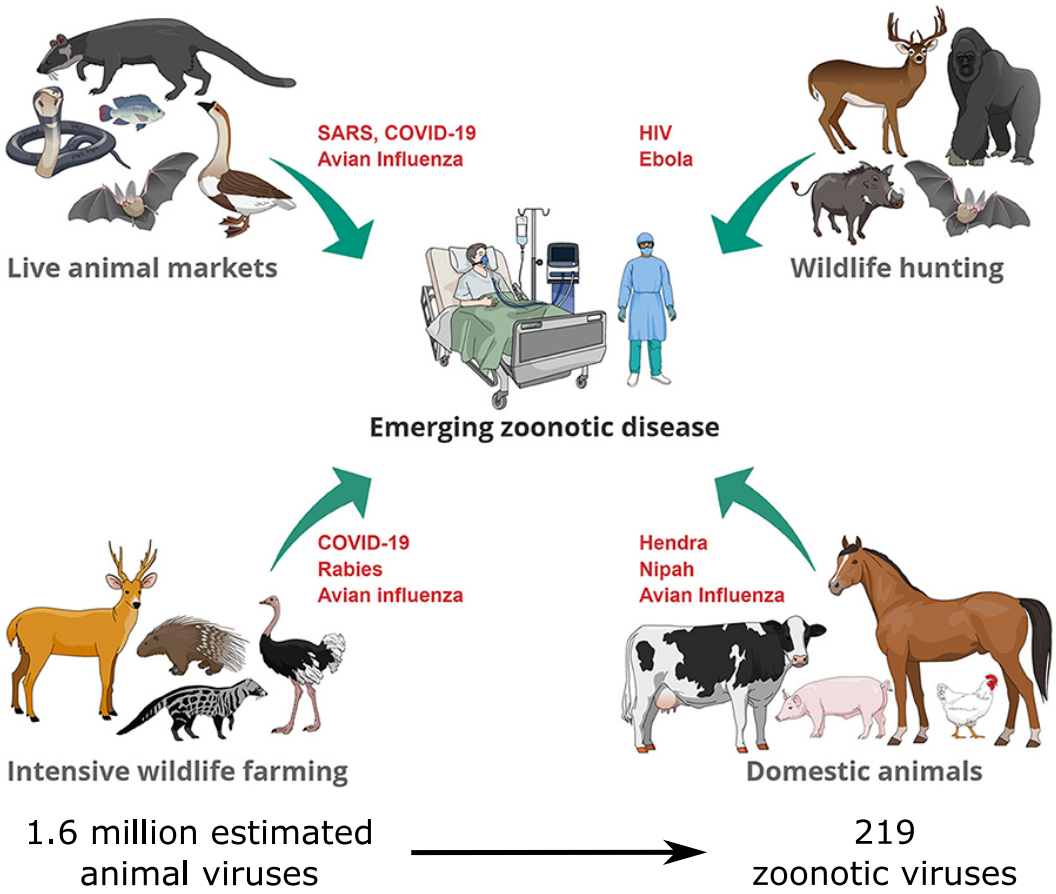
Department of Computer Science
Virginia Tech

CS 5854: Computational Systems Biology
February 19, 24, 26, 2025
(Slides created by Blessy Antony)

Outline

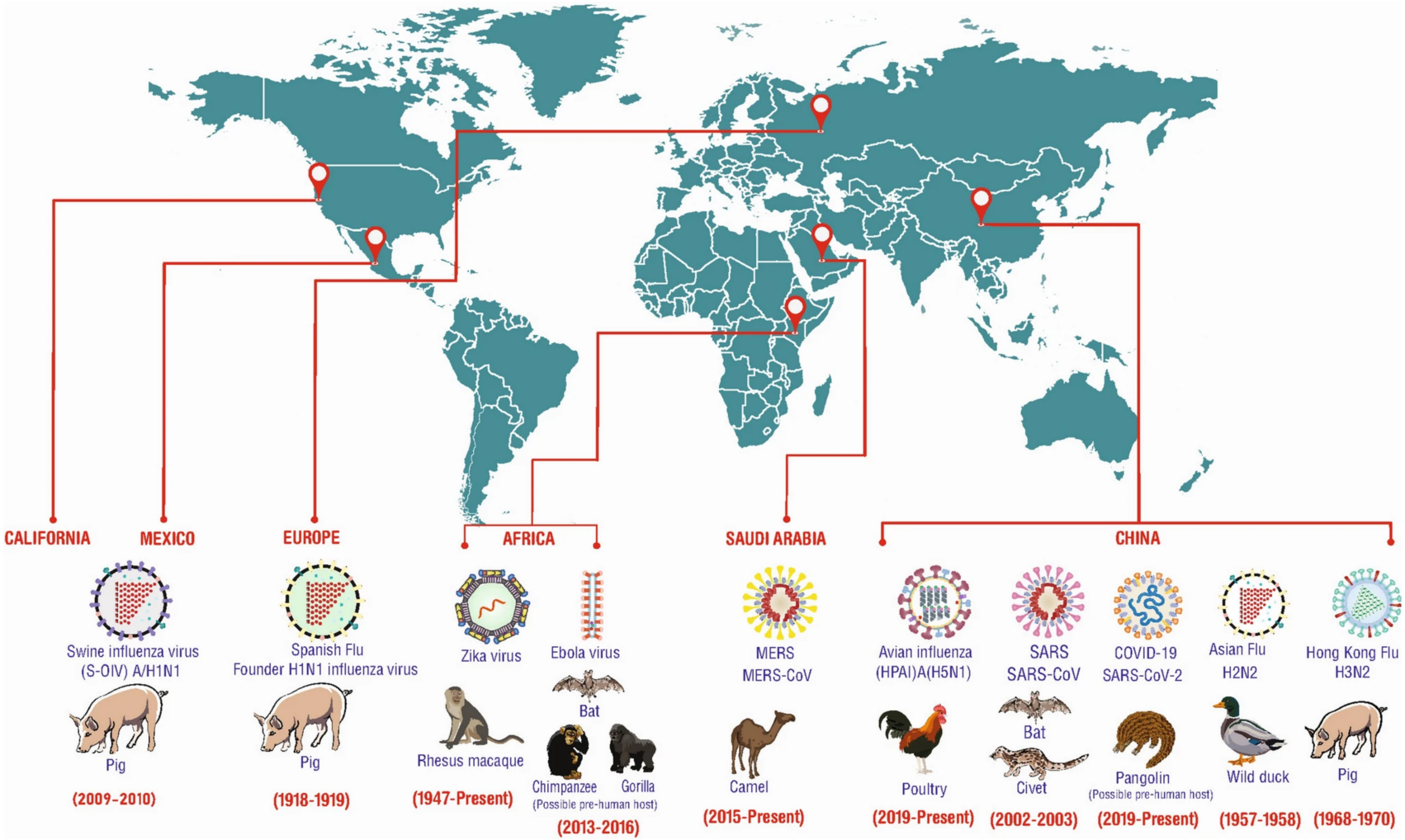
Outline

Zoonosis



I. Magouras et al., "Emerging Zoonotic Diseases: Should We Rethink the Animal–Human Interface?", *Frontiers in Veterinary Science*, 2020.

Zoonosis Examples



Mishra, J. *et al.*, "Linkages between environmental issues and zoonotic diseases: with reference to COVID-19 pandemic.", *Environmental Sustainability*, March 2021.

COVID-19 Pandemic

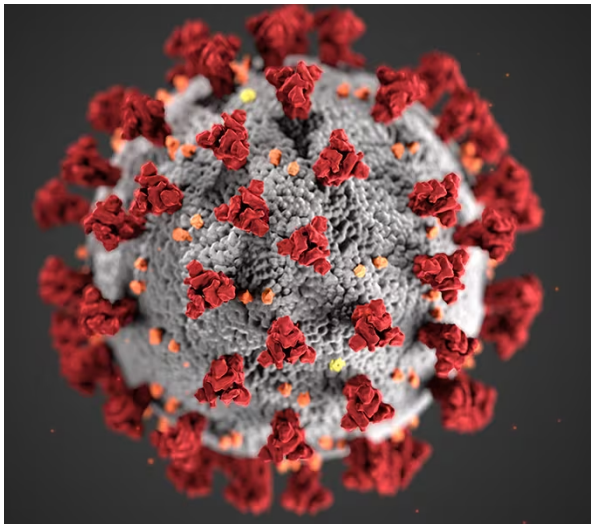
reason.org/policy-brief/covid-19-lockdown-problems-and-alternative-strategies-to-reopening-the-economy

COVID-19 pandemic

774 million infections

7 million deaths*

reuters.com/world/americas/paho-says-40-last-weeks-covid-19-deaths-were-americas-2021-05-12

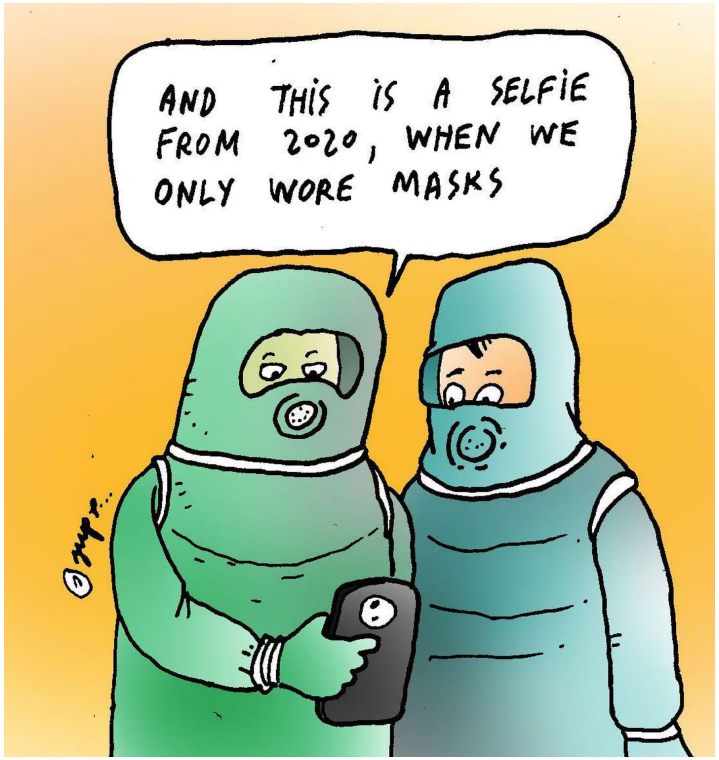


cdc.gov/museum/timeline/covid19.html

SARS-CoV-2

*WHO COVID-19 dashboard (accessed on Feb 11, 2024).

Next Pandemic?

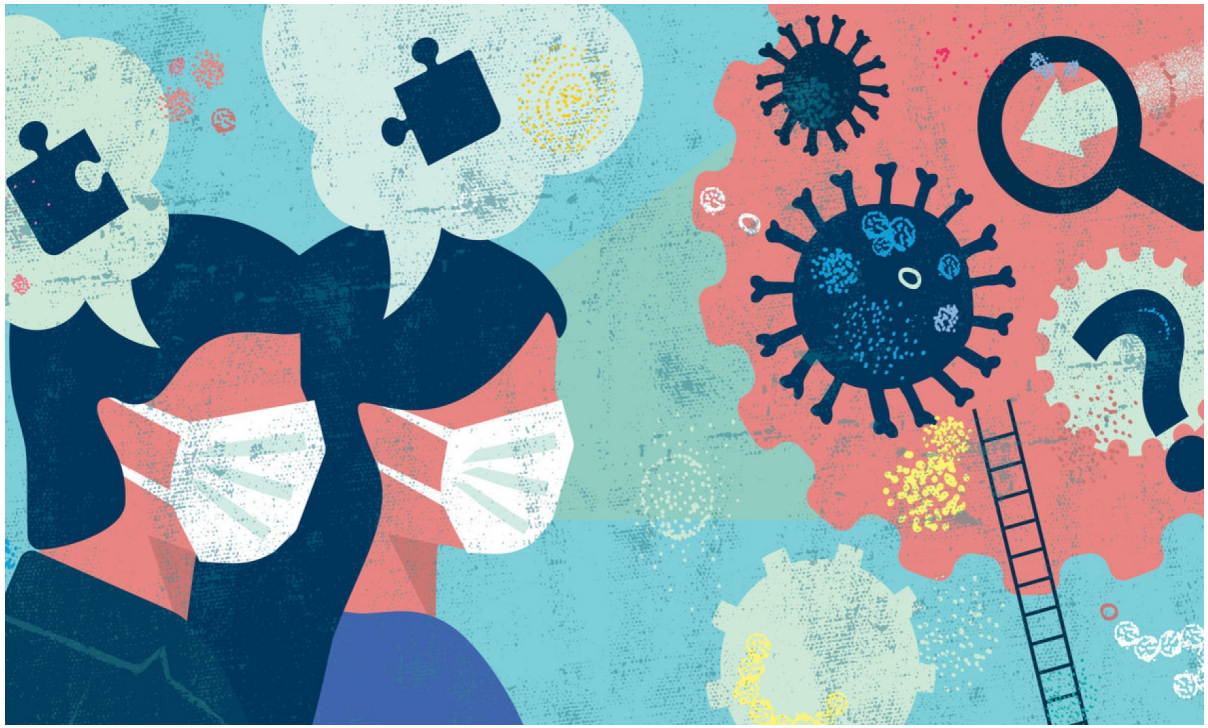


Jup, Ziarul de Iasi, Romania

The number of new infectious diseases with epidemic potential has increased nearly four-fold over the past six decades.

Gruber K., "Preventing zoonotic pandemics: are we there yet?", *The Lancet Microbe.*, 2021.

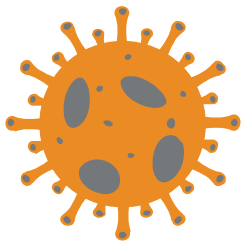
Goal



Why Can't COVID-19 Be Eradicated and Other Lingering Questions, Johns Hopkins, Bloomberg School of Public Health, 2022

How do we accurately foresee the next pandemic and proactively minimize its impact?

SARS-CoV-2

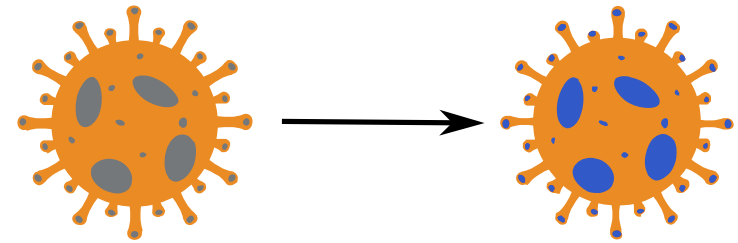


Animal X

Bats?

Holmes EC., "COVID-19—lessons for zoonotic disease.", *Science.*, March 2022.

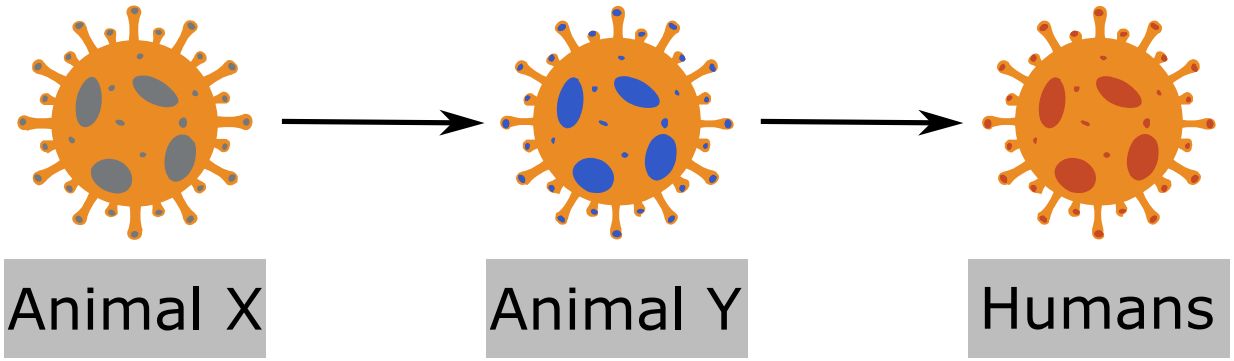
SARS-CoV-2



Pangolins or Raccoon dogs?

Holmes EC., "COVID-19—lessons for zoonotic disease.", *Science.*, March 2022.

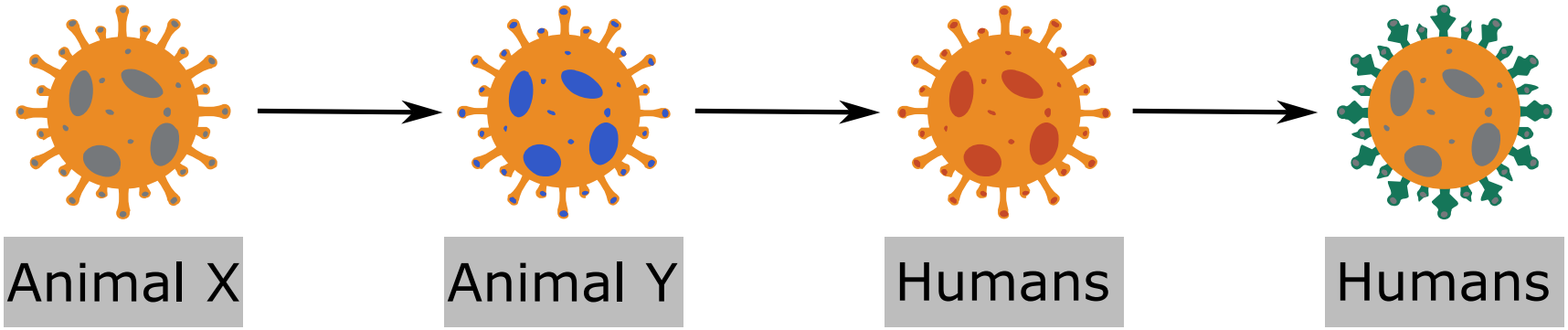
SARS-CoV-2



Zoonotic mutation resulting in human infection

Holmes EC., "COVID-19—lessons for zoonotic disease.", *Science.*, March 2022.

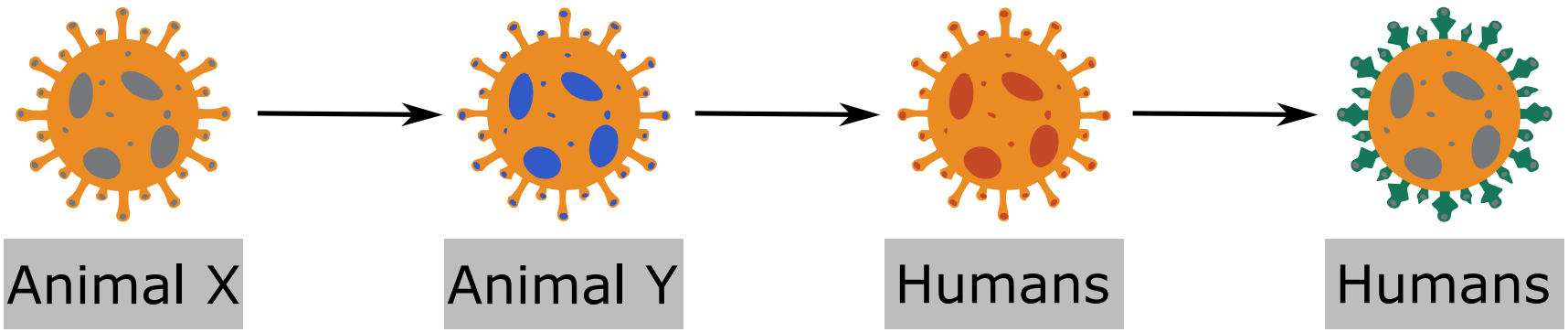
SARS-CoV-2



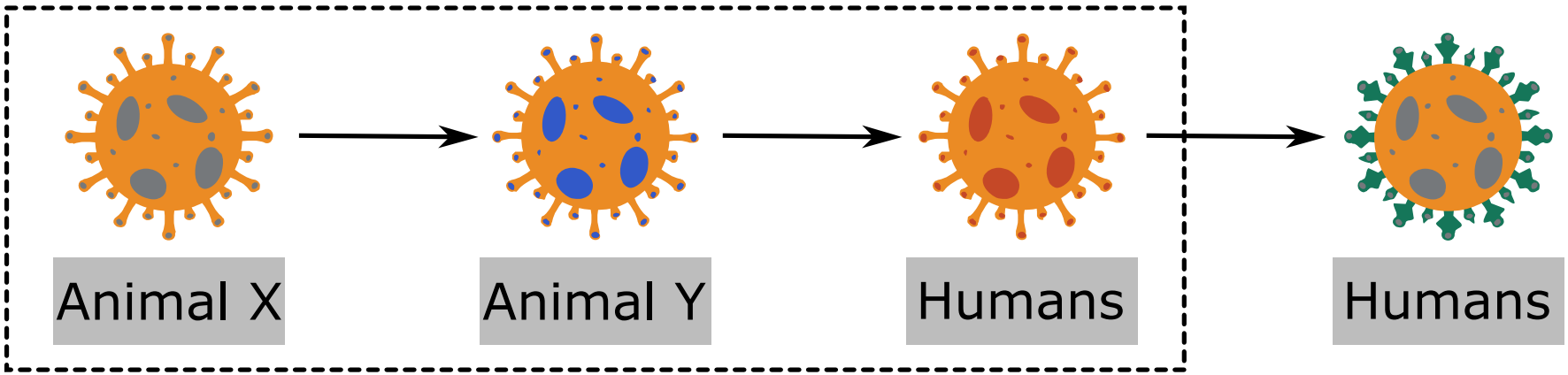
Mutations leading to enhanced replication and transmission

Holmes EC., "COVID-19—lessons for zoonotic disease.", *Science.*, March 2022.

SARS-CoV-2

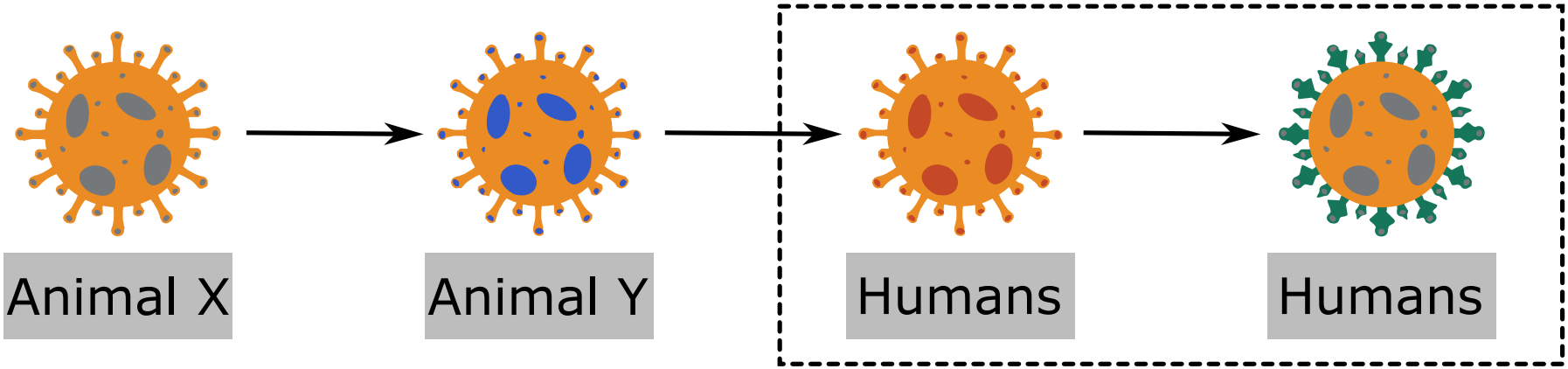


SARS-CoV-2



Identify the potential hosts of a virus observed in nature

SARS-CoV-2



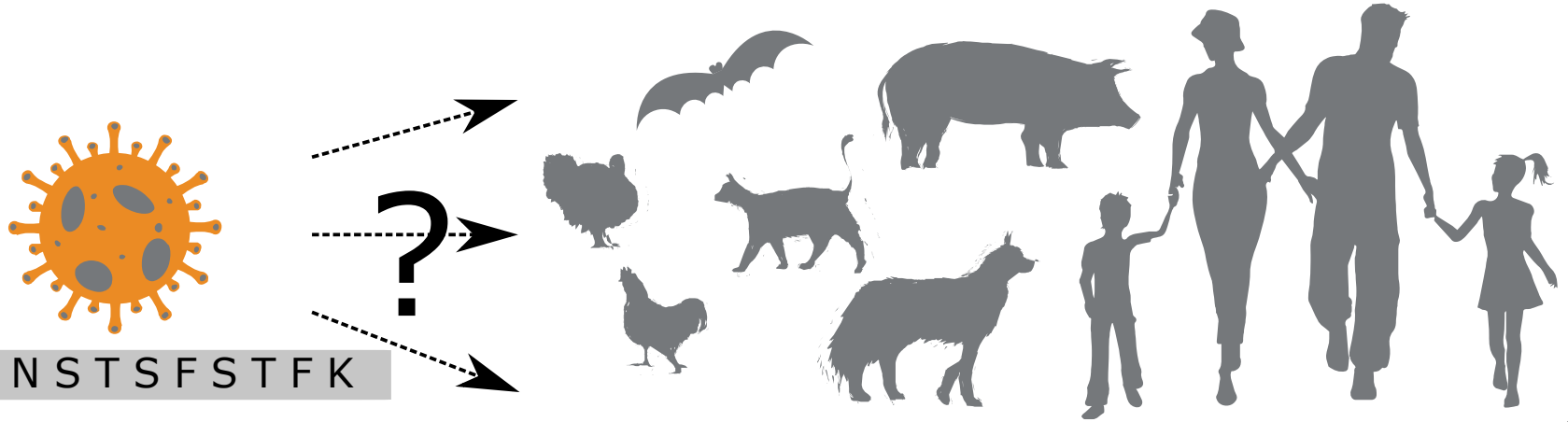
- Proactive outbreak detection, preparedness, and response.
- Targeted R&D for vaccines and therapeutics.

Outline

VirProBERT: Viral Protein Language Model for Virus-Host Prediction

Host Prediction

Given a viral protein sequence, what is the infected host?



Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?
- Multiclass classification: Which hosts will a given virus infect?

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?
- Multiclass classification: Which hosts will a given virus infect?

Studies done so far identify hosts using protein sequences of —

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?
- Multiclass classification: Which hosts will a given virus infect?

Studies done so far identify hosts using protein sequences of —

- Only one specific protein [1].

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?
- Multiclass classification: Which hosts will a given virus infect?

Studies done so far identify hosts using protein sequences of —

- Only one specific protein [1].
- Only one virus of interest [2].

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

Related Work on Host Prediction

- Binary classification: Will a given virus infect humans or not?
- Multiclass classification: Which hosts will a given virus infect?

Studies done so far identify hosts using protein sequences of —

- Only one specific protein [1].
- Only one virus of interest [2].

Unsolved Problem: Predict host of an arbitrary protein sequence in an arbitrary virus.

Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.
Grange ZL et al., "Ranking the risk of animal-to-human spillover for newly discovered viruses", *Proceedings of the National Academy of Sciences*, 2021.
Becker DJ et al., "Optimising predictive models to prioritise viral discovery in zoonotic reservoirs", *The Lancet Microbe*, 2022.

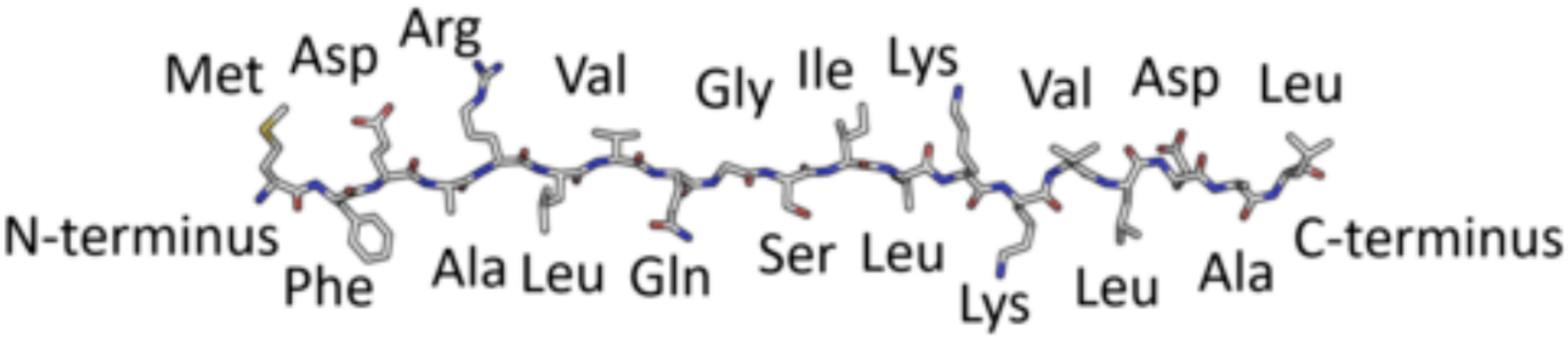
Approach

Learn the properties of viral protein sequences
using natural language processing models

Neil Thomas et al., "Can We Learn the Language of Proteins", *Berkeley Artificial Intelligence Research*, 2019.
Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.

Approach

Learn the properties of viral protein sequences
using natural language processing models




```
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT
GKLPVPWPTL VTTFSYGVQC FSRYPDHMKQ HDFFKSAMPE GYVQERTIFF
KDDGNYKTRA EVKFEGDTLV NRIELKGIDF KEDGNILGHK LEYNYNSHNV
YIMADKQKNG IKVNFKIRHN IEDGSVQLAD HYQQNTPIGD GPVLLPDNHY
LSTQSALSKD PNEKRDHMVL LEFVTAAGIT HGMDELYK
```

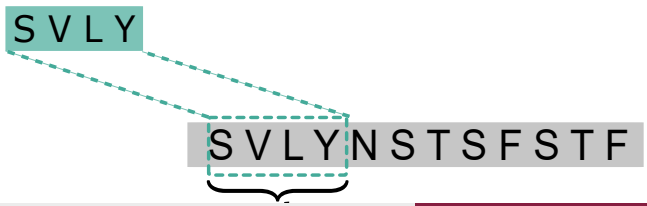
Neil Thomas et al., "Can We Learn the Language of Proteins", *Berkeley Artificial Intelligence Research*, 2019.
Hie et al., "Learning the language of viral evolution and escape", *Science*, 2021.

Outline

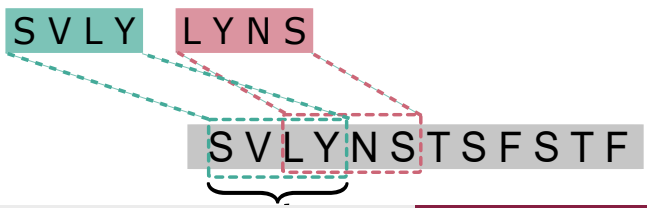
VirProBERT Architecture

S V L Y N S T S F S T F

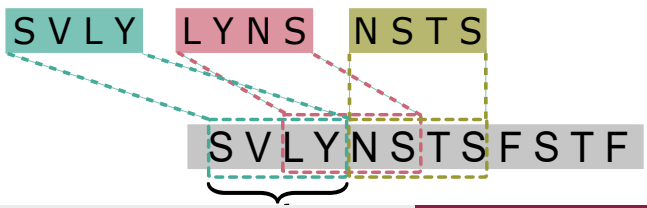
VirProBERT Architecture



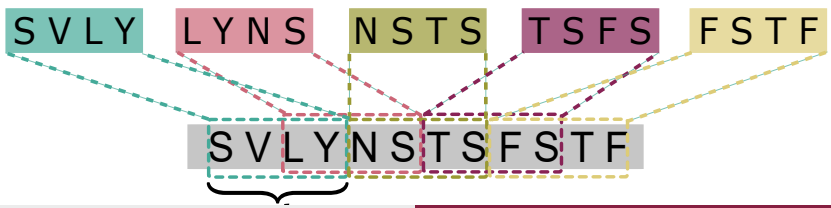
VirProBERT Architecture



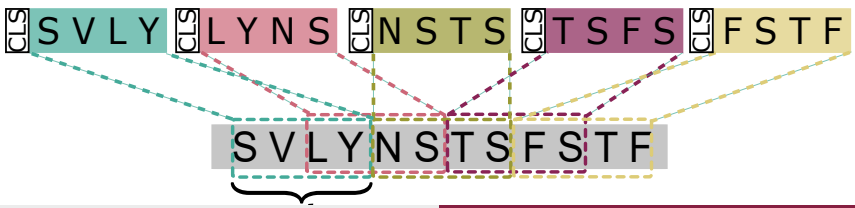
VirProBERT Architecture



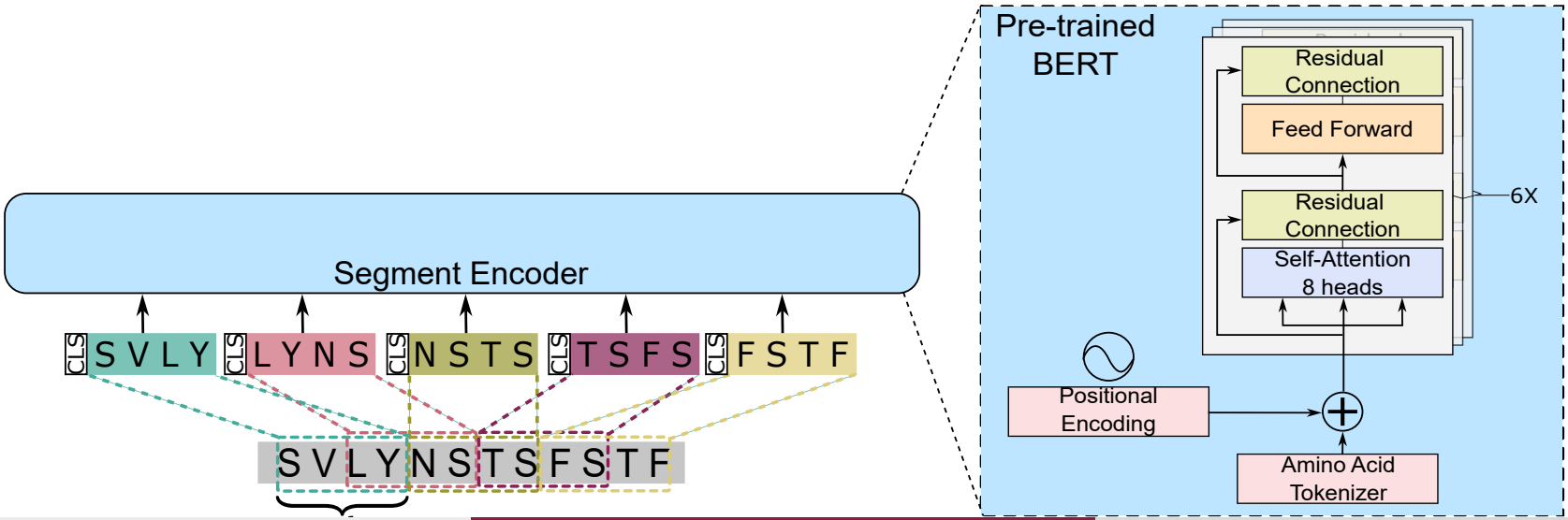
VirProBERT Architecture



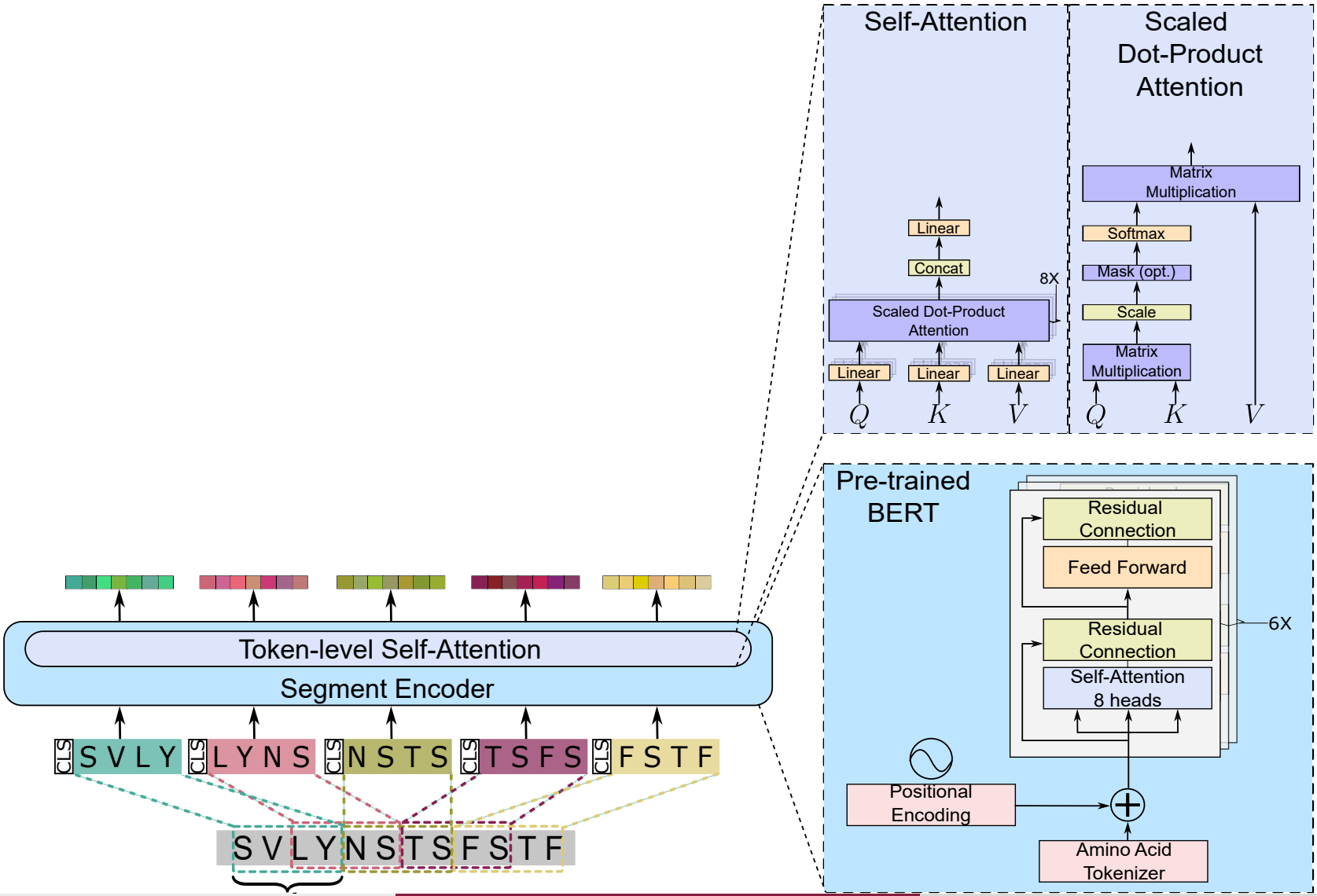
VirProBERT Architecture



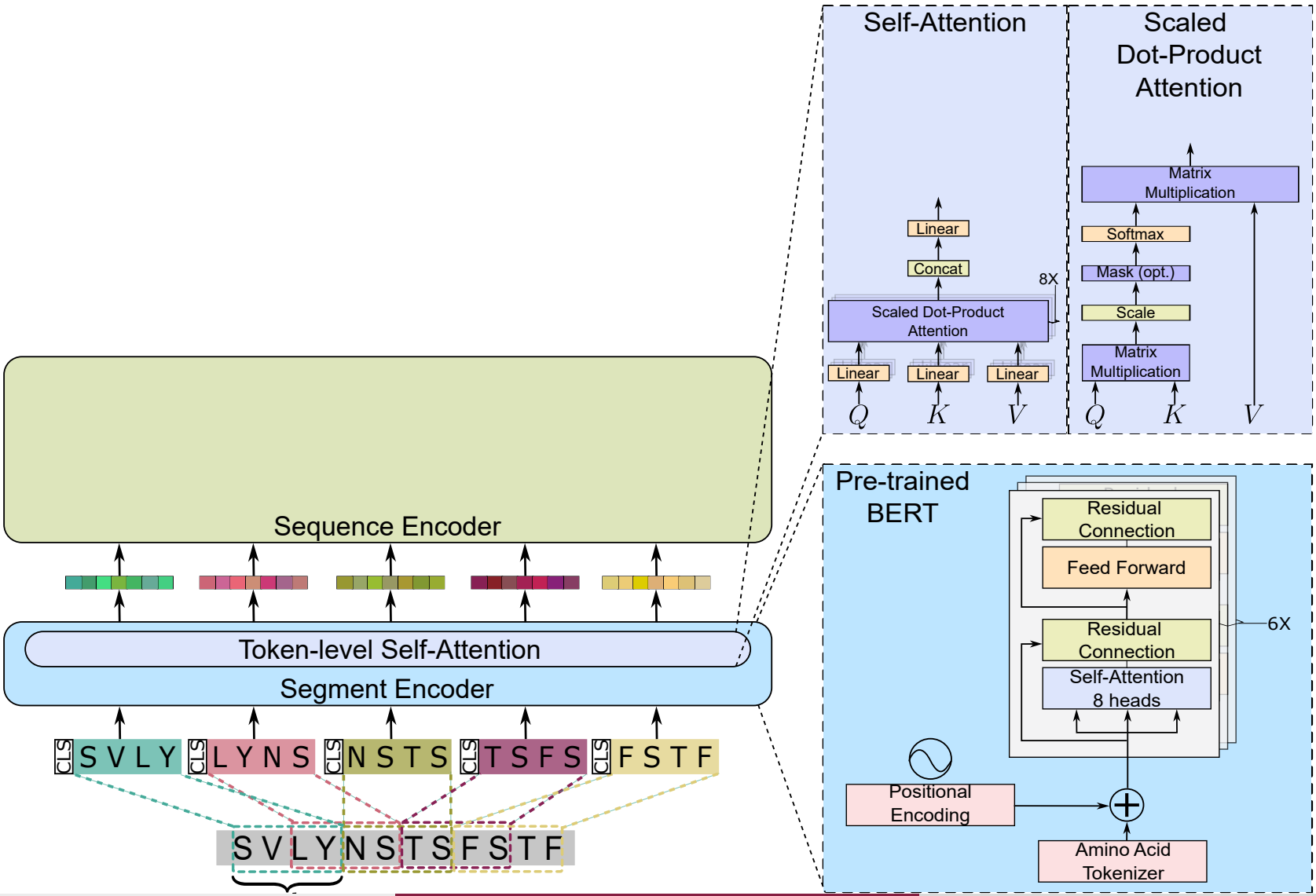
VirProBERT Architecture



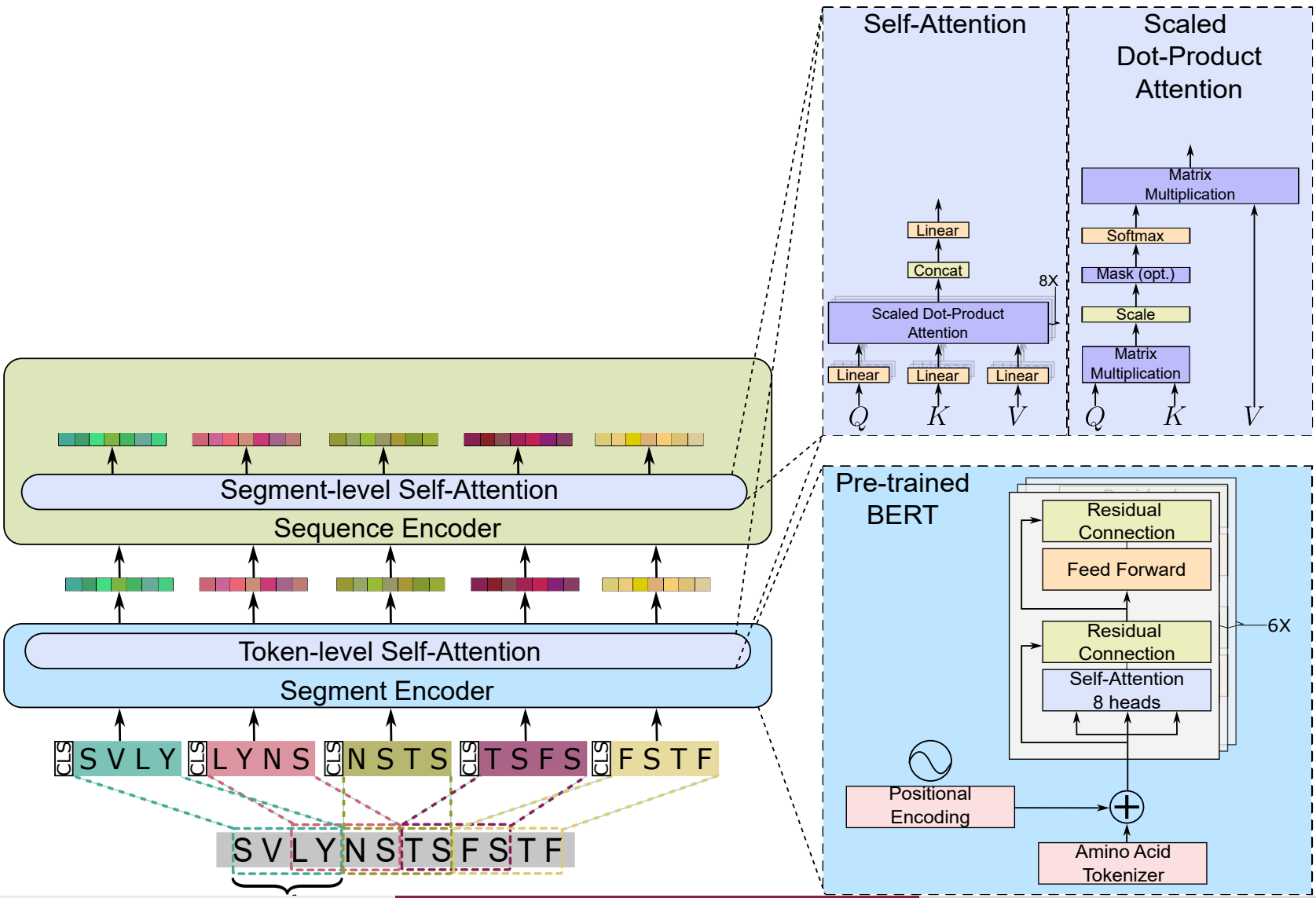
VirProBERT Architecture



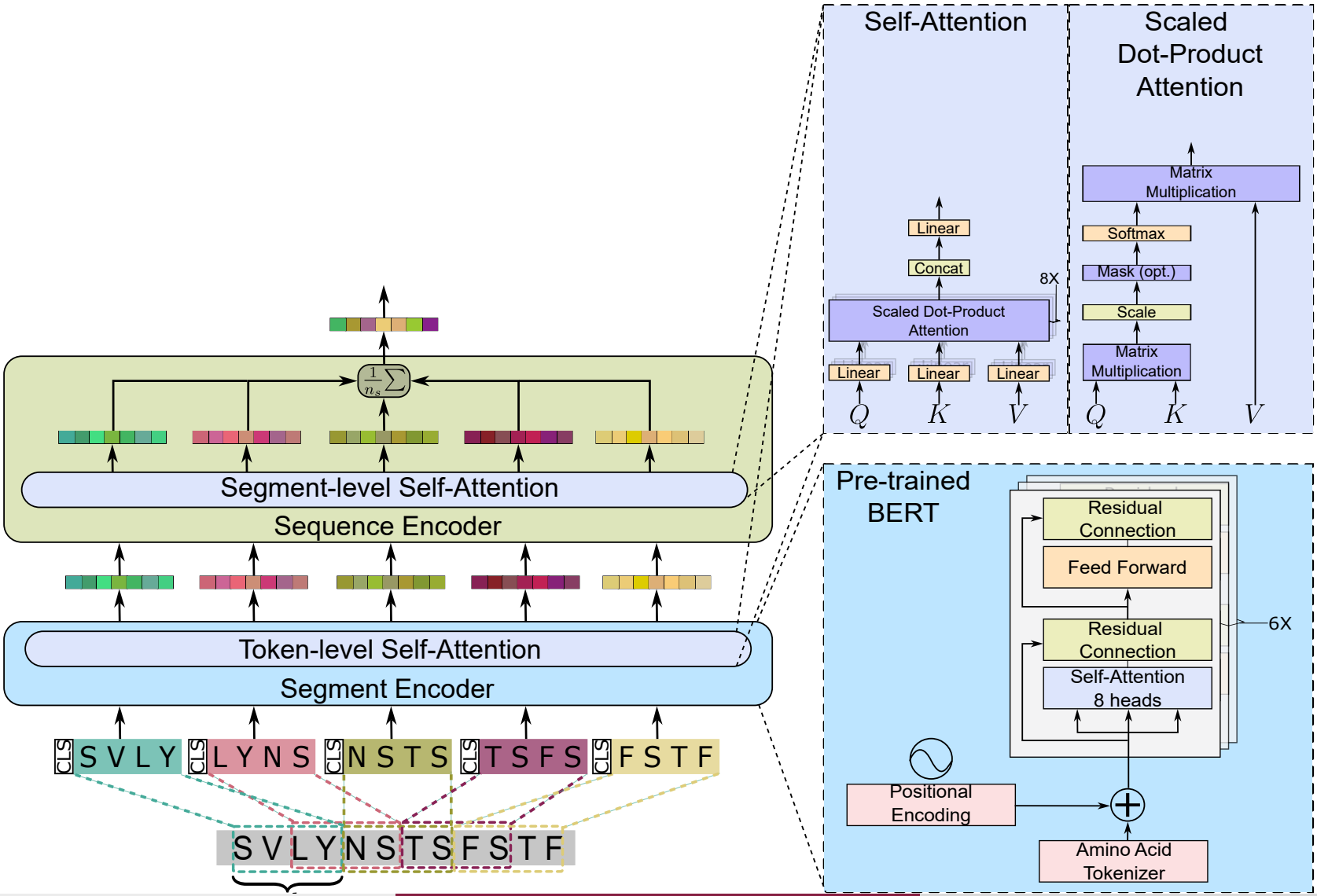
VirProBERT Architecture



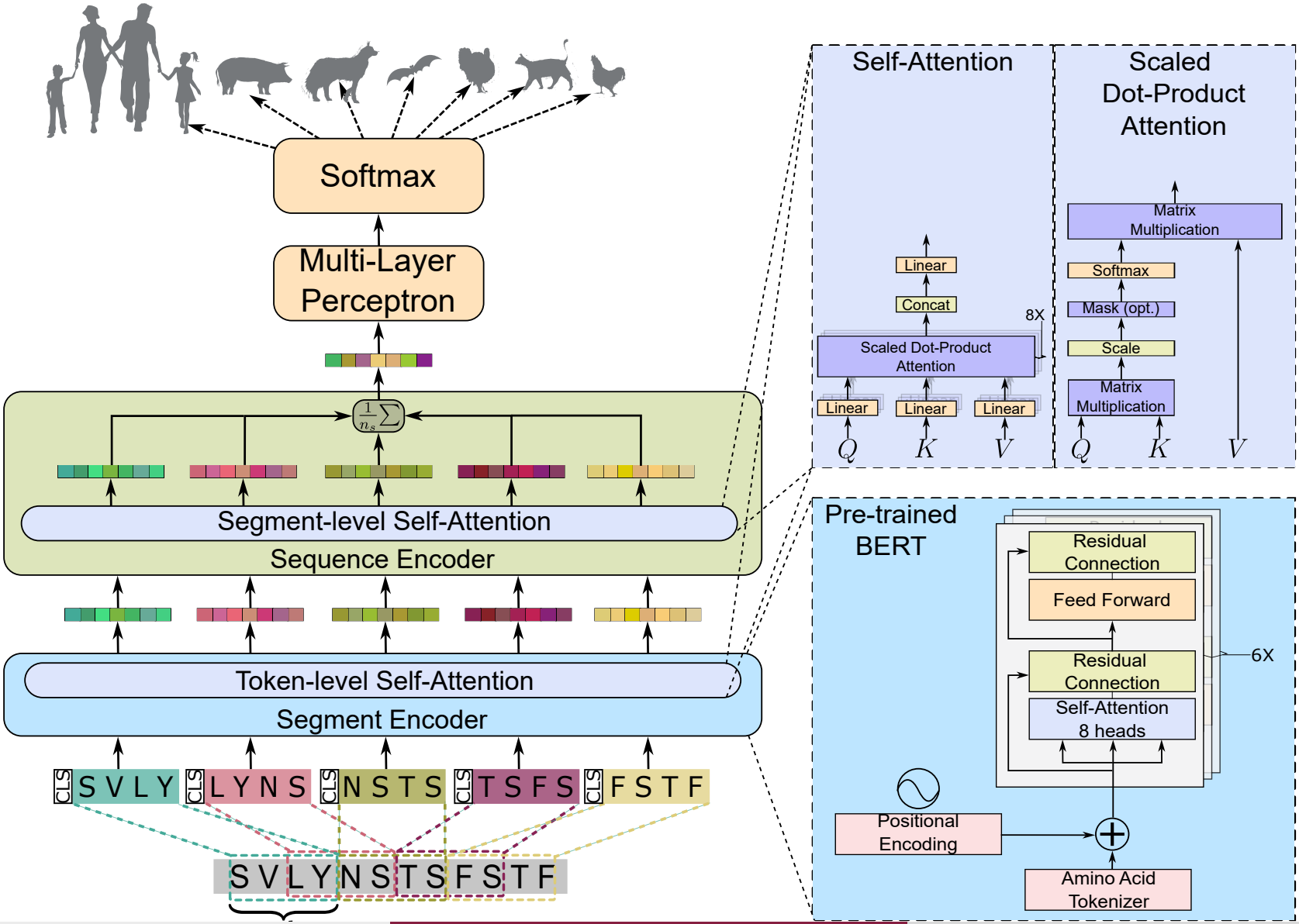
VirProBERT Architecture



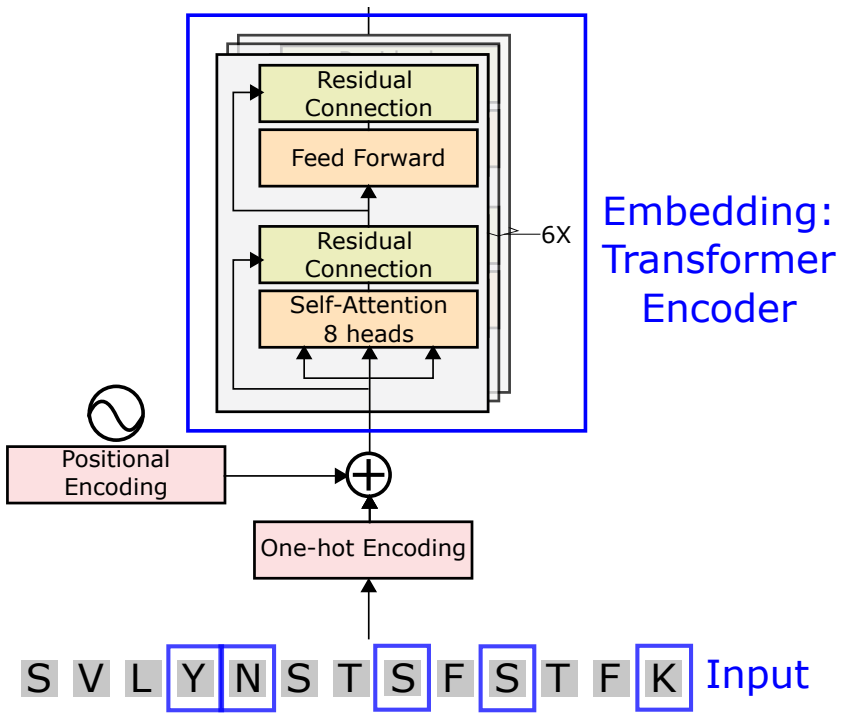
VirProBERT Architecture



VirProBERT Architecture

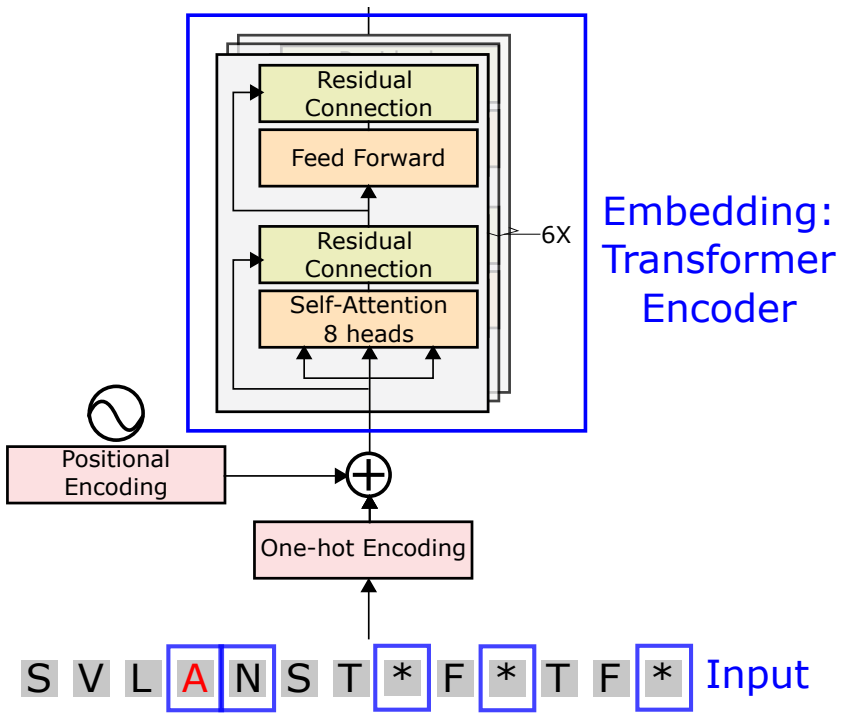


Masked Language Modeling



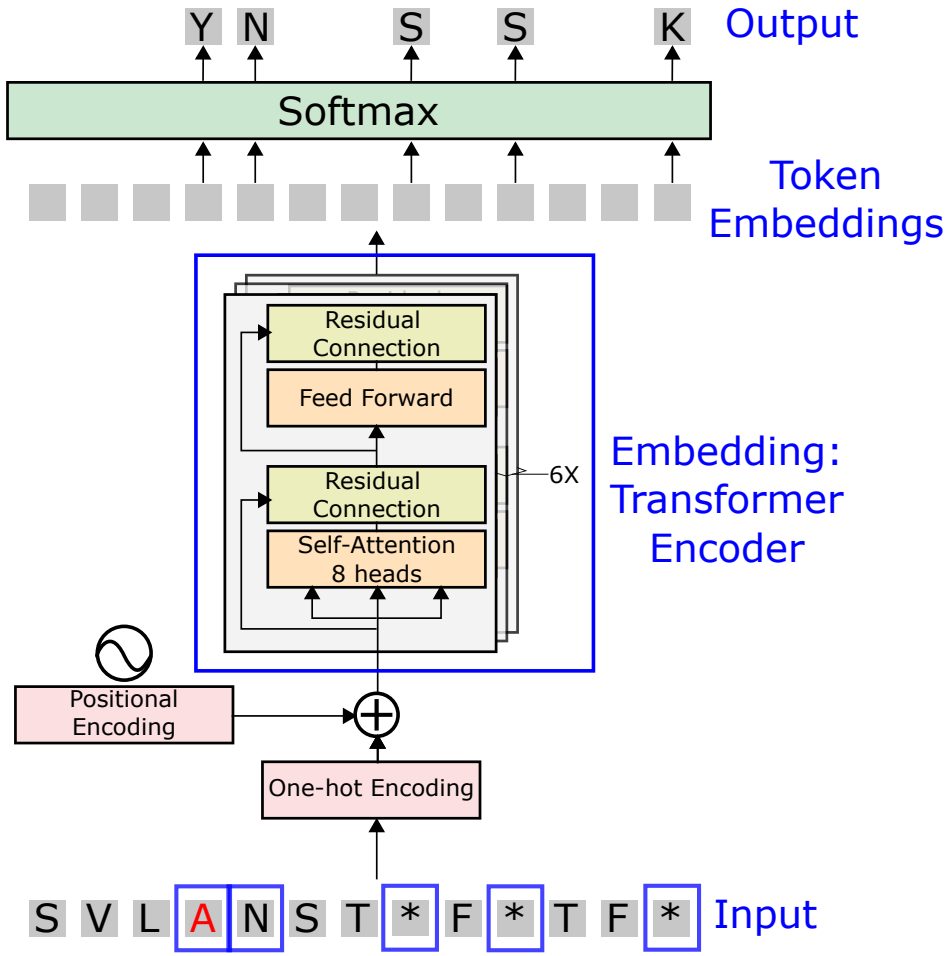
Pre-train the **Segment Encoder** using **Masked Language Modeling** with all viral protein sequences.

Masked Language Modeling



Pre-train the **Segment Encoder** using **Masked Language Modeling** with all viral protein sequences.

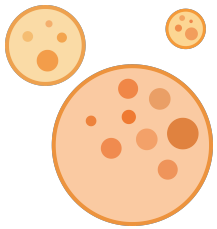
Masked Language Modeling



Pre-train the **Segment Encoder** using **Masked Language Modeling** with all viral protein sequences.

Dataset

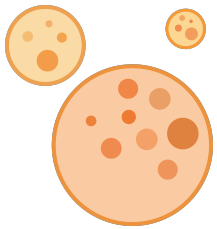
UniRef90



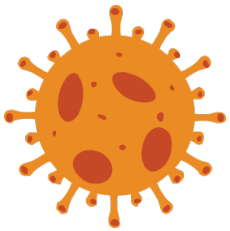
Protein sequences with atleast 90% sequence similarity

Dataset

UniRef90 → Viruses



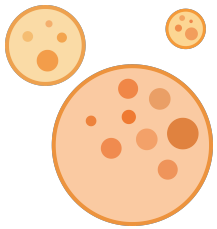
Protein sequences with atleast 90% sequence similarity



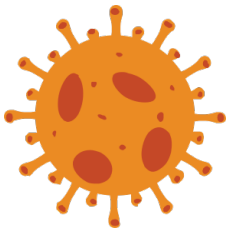
Retain sequences from clusters of viral proteins

Dataset

UniRef90 → Viruses



Protein sequences with atleast 90% sequence similarity



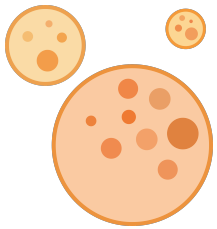
Retain sequences from clusters of viral proteins

Pre-training Dataset

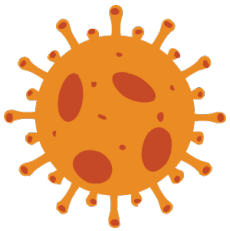
1.2 million sequences

Dataset

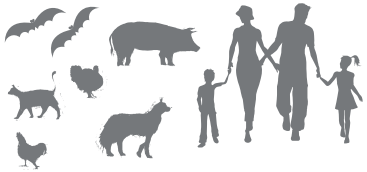
UniRef90 → Viruses → Vertebrates



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



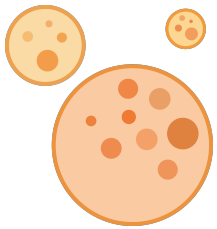
Retain clusters where the host is a vertebrate

Pre-training Dataset

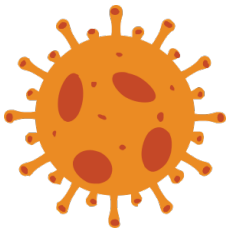
1.2 million sequences

Dataset

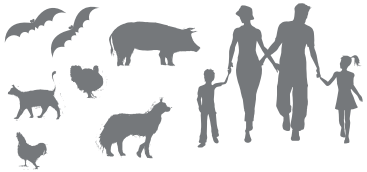
UniRef90 → Viruses → Vertebrates



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Pre-training Dataset

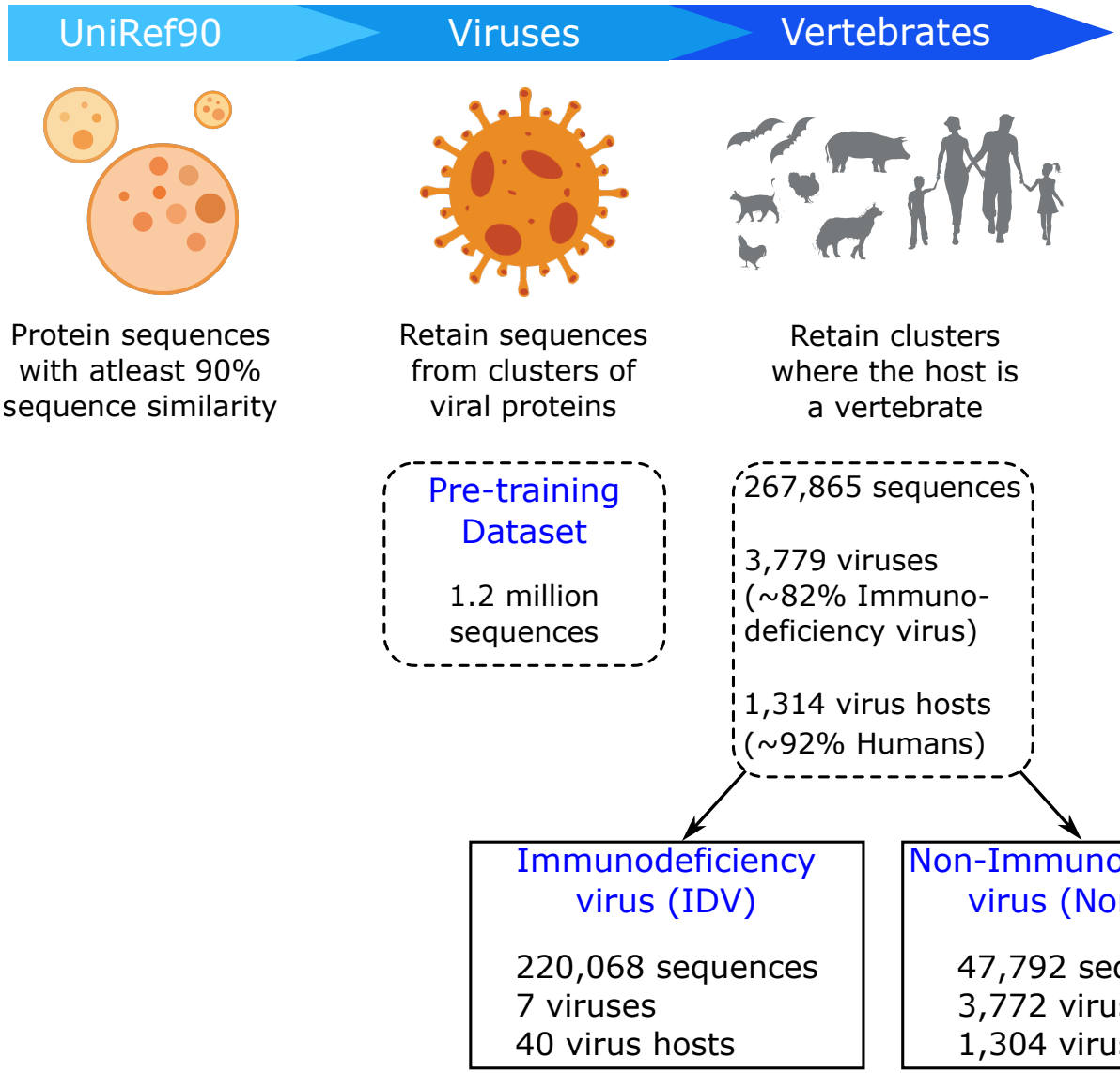
1.2 million sequences

267,865 sequences

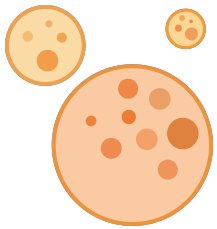
3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

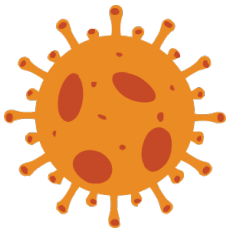
Dataset


UniRef90
Viruses
Vertebrates
Protein sequences with atleast 90% sequence similarity
Retain sequences from clusters of viral proteins
Retain clusters where the host is a vertebrate
Pre-training Dataset
1.2 million sequences
267,865 sequences
3,779 viruses (~82% Immuno-deficiency virus)
1,314 virus hosts (~92% Humans)
Immunodeficiency virus (IDV)
220,068 sequences
7 viruses
40 virus hosts
Non-Immunodeficiency virus (Non-IDV)
47,792 sequences
3,772 viruses
1,304 virus hosts

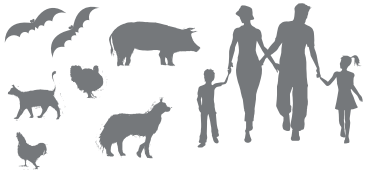
Dataset



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Hosts with atleast 1% prevalence

Pre-training Dataset

1.2 million sequences

267,865 sequences

3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

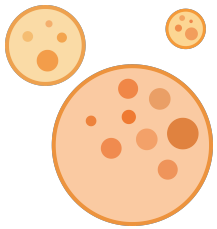
Immunodeficiency virus (IDV)

220,068 sequences
7 viruses
40 virus hosts

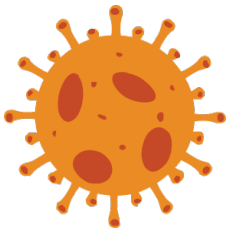
Non-Immunodeficiency virus (Non-IDV)

47,792 sequences
3,772 viruses
1,304 virus hosts

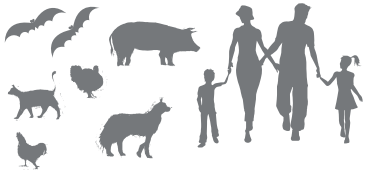
Dataset



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Pre-training Dataset

1.2 million sequences

267,865 sequences

3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

Hosts with atleast 1% prevalence

Hosts with prevalence < 1%

Rare Hosts

16,074 sequences
143 virus hosts

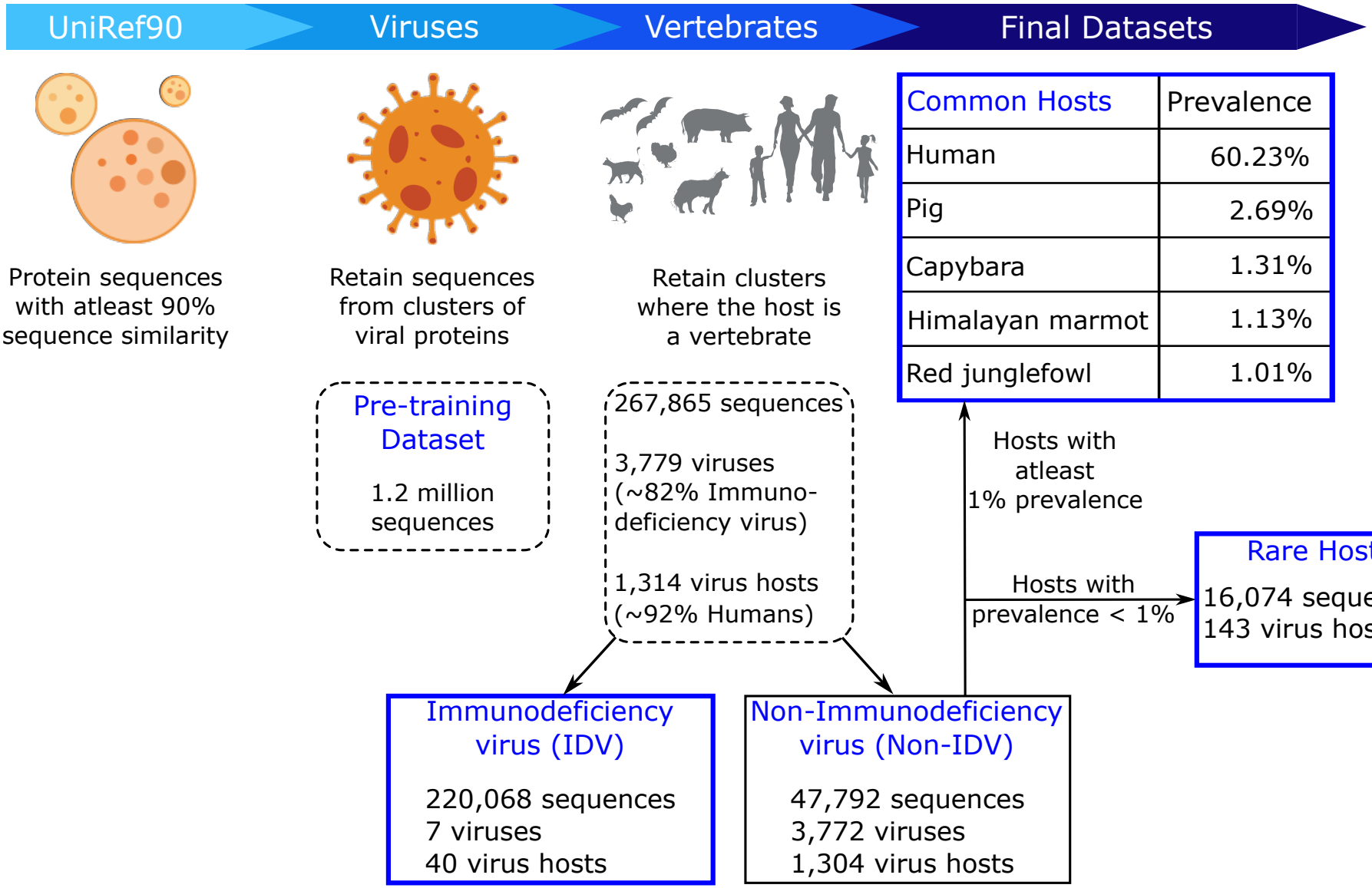
Immunodeficiency virus (IDV)

220,068 sequences
7 viruses
40 virus hosts

Non-Immunodeficiency virus (Non-IDV)

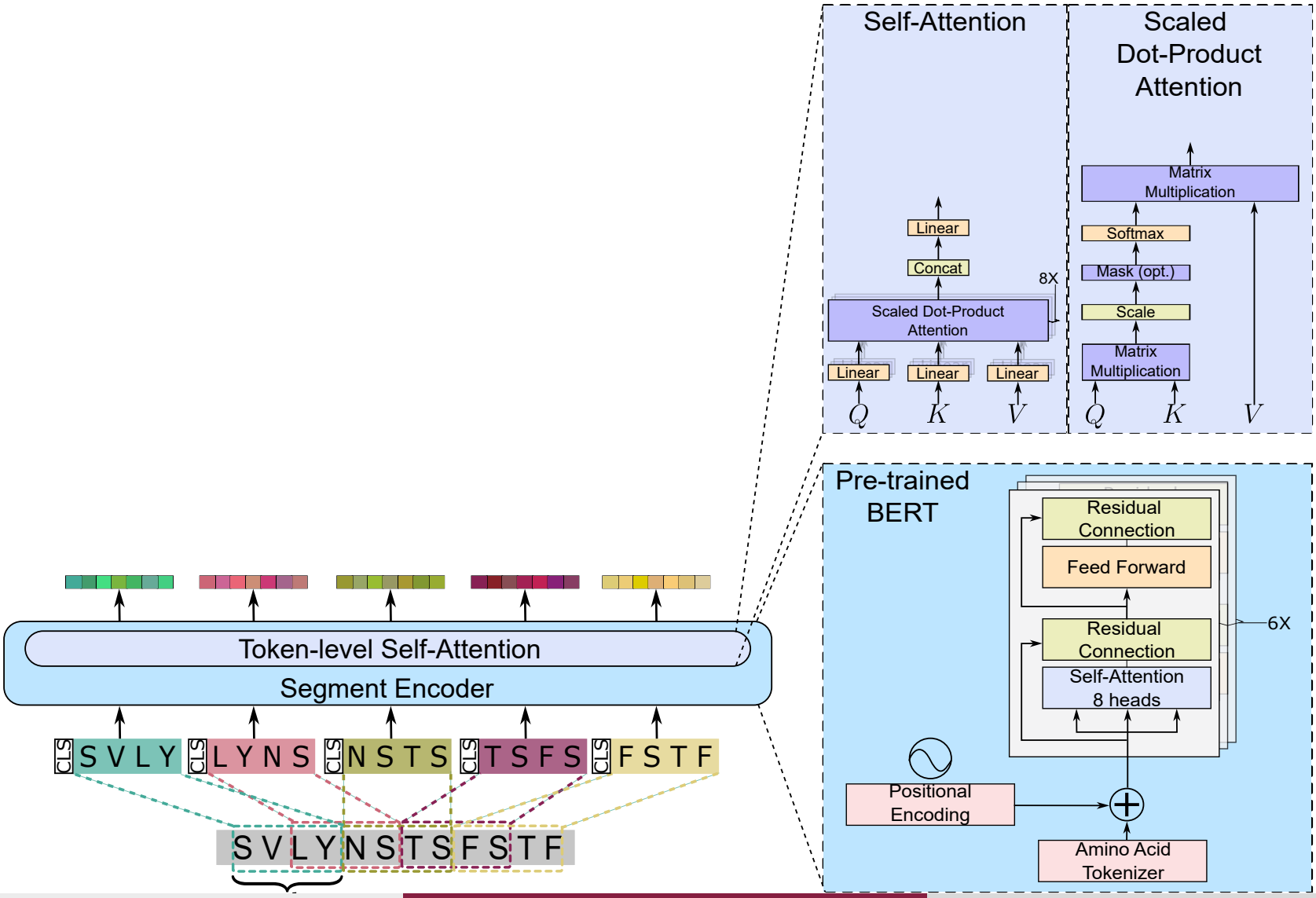
47,792 sequences
3,772 viruses
1,304 virus hosts

Dataset



Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Pre-training



Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.

Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.
- Problem caused by long sequences.

Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.
- Problem caused by long sequences.
  - Overlapping segments of length 256.
  - Stride=64.

Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.
- Problem caused by long sequences.
  - Overlapping segments of length 256.
  - Stride=64.
- Input: Masked protein segments.

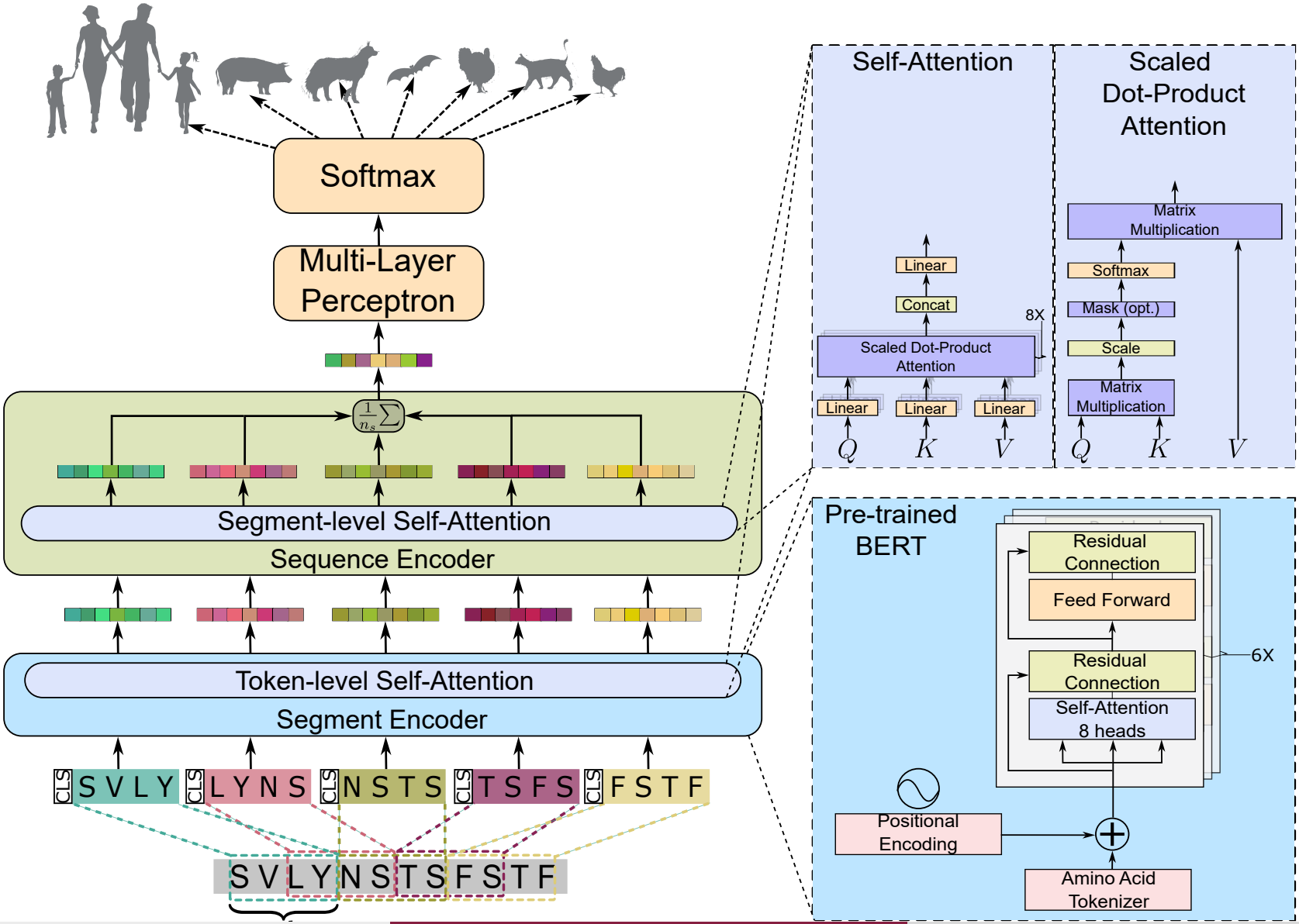
Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.
- Problem caused by long sequences.
  - Overlapping segments of length 256.
  - Stride=64.
- Input: Masked protein segments.
- Output: Predictions of amino acids for the masked positions.

Pre-training

- Self-supervised learning: Masked Language Modeling.
- Dataset: 1.2 million viral protein sequences.
- Problem caused by long sequences.
  - Overlapping segments of length 256.
  - Stride=64.
- Input: Masked protein segments.
- Output: Predictions of amino acids for the masked positions.
- Dataset split: 90% for training, 10% for validation.

Fine-tuning


Softmax
Multi-Layer Perceptron
Segment-level Self-Attention
Sequence Encoder
Token-level Self-Attention
Segment Encoder
CLS S V L Y
CLS L Y N S
CLS N S T S
CLS T S F S
CLS F S T F
S V L Y N S T S F S T F
Self-Attention
Linear
Concat
8X
Scaled Dot-Product Attention
Linear
Linear
Linear
Q
K
V
Scaled Dot-Product Attention
Matrix Multiplication
Softmax
Mask (opt.)
Scale
Matrix Multiplication
Q
K
V
Pre-trained BERT
Residual Connection
Feed Forward
Residual Connection
Self-Attention 8 heads
6X
Positional Encoding
Amino Acid Tokenizer

Fine-tuning

- Supervised learning: Multi-class classification to predict hosts.

Fine-tuning

- Supervised learning: Multi-class classification to predict hosts.
- Dataset: Non-IV dataset - common classes only (31,718 sequences)

Host	**Prevalence**
Human	90.79%
Pig	4.04%
Capybara	1.96%
Himalayan Marmot	1.70%
Red junglefowl	1.51%

Fine-tuning

- Supervised learning: Multi-class classification to predict hosts.
- Dataset: Non-IV dataset - common classes only (31,718 sequences)

Host	**Prevalence**
Human	90.79%
Pig	4.04%
Capybara	1.96%
Himalayan Marmot	1.70%
Red junglefowl	1.51%

- Input: Protein sequence.
- Output: Probabilities for each of the five classes.
- Dataset split: 80% training, 10% validation, 10% testing.

Loss Function for Binary Classification

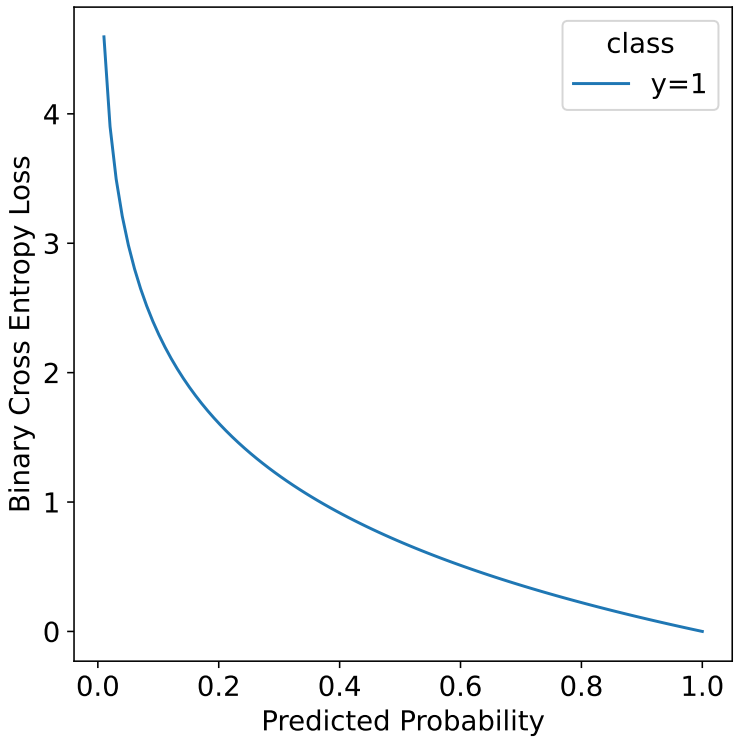
For any sample i,
$p^{(i)} =$ predicted probability
$y^{(i)} = 0$ or 1 is the true class

$$BCE\left(y^{(i)}, p^{(i)}\right) = -\left(y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log \left(1 - p^{(i)}\right)\right)$$

Loss Function for Binary Classification

For any sample i,
$p^{(i)}$ = predicted probability
$y^{(i)} = 0$ or 1 is the true class

$$BCE\left(y^{(i)}, p^{(i)}\right) = -\left(y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right)\right)$$

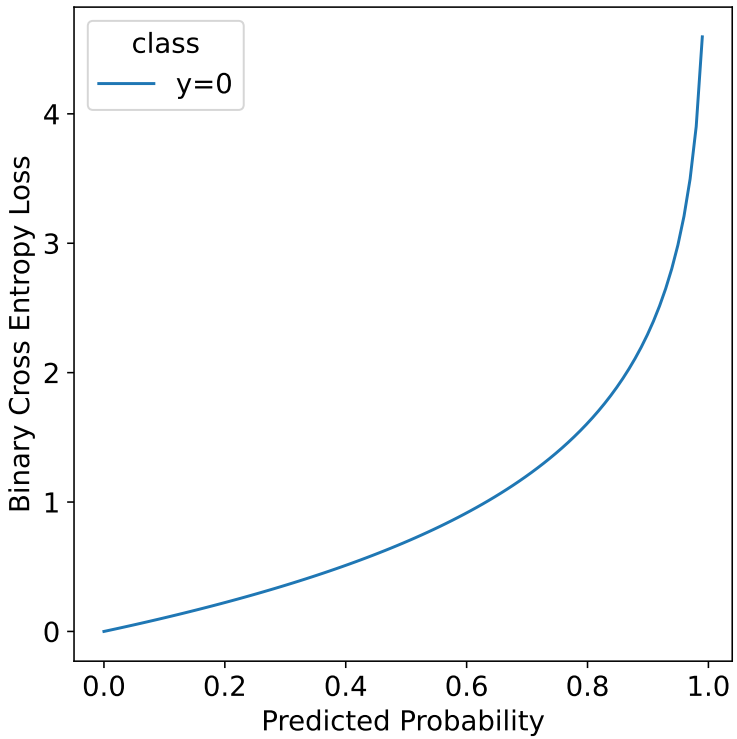


If $y^{(i)} = 1$, then
$BCE(y^{(i)}, p^{(i)}) = -\log p^{(i)}$

Loss Function for Binary Classification

For any sample i,
$p^{(i)}$ = predicted probability
$y^{(i)} = 0$ or 1 is the true class

$$BCE\left(y^{(i)}, p^{(i)}\right) = -\left(y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right)\right)$$

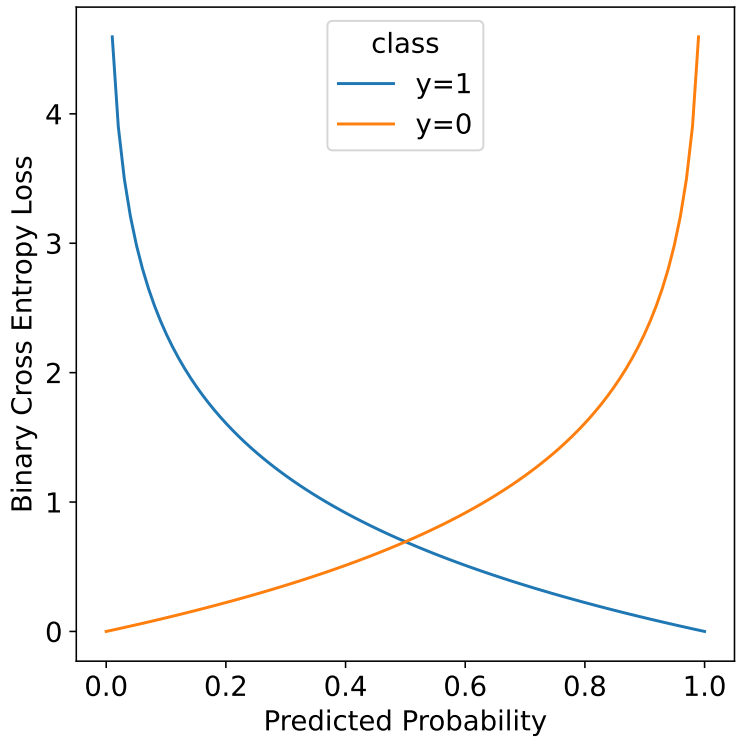


If $y^{(i)} = 0$, then
$BCE(y^{(i)}, p^{(i)}) = -\log\left(1 - p^{(i)}\right)$

Loss Function for Binary Classification

For any sample i,
$p^{(i)}$ = predicted probability
$y^{(i)} = 0$ or 1 is the true class

$$BCE\left(y^{(i)}, p^{(i)}\right) = -\left(y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right)\right)$$



Cross Entropy for Multi-class Classification

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

Focal Loss

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

Lin et. al., "Focal Loss for Dense Object Detection", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Focal Loss

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

$$FL(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} \left(1 - p_c^{(i)}\right)^{\gamma} y_c^{(i)} \log p_c^{(i)}$$

$\gamma \geq 0$: Focusing parameter

Lin et. al., "Focal Loss for Dense Object Detection", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Focal Loss

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

$$FL(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} \left(1 - p_c^{(i)}\right)^{\gamma} y_c^{(i)} \log p_c^{(i)}$$

$\gamma \geq 0$: Focusing parameter
$p_c^{(i)} \to 1$: Loss for a well-classified sample is down-weighted (easy sample).

Lin et. al., "Focal Loss for Dense Object Detection", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Focal Loss

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

$$FL(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} \left(1 - p_c^{(i)}\right)^{\gamma} y_c^{(i)} \log p_c^{(i)}$$

$\gamma \geq 0$: Focusing parameter
$p_c^{(i)} \to 1$: Loss for a well-classified sample is down-weighted (easy sample).
$p_c^{(i)} \to 0$: Loss for a mis-classified sample is up-weighted (difficult sample).

Lin et. al., "Focal Loss for Dense Object Detection", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Focal Loss

For any sample i,
C = Set of classes
$p_c^{(i)}$ = predicted probability for class c
$y_c^{(i)}$ = true label (0 or 1 for each class)

$$CE(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} y_c^{(i)} \log p_c^{(i)}$$

$$FL(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} \left(1 - p_c^{(i)}\right)^{\gamma} y_c^{(i)} \log p_c^{(i)}$$

$\gamma \geq 0$: Focusing parameter
$p_c^{(i)} \to 1$: Loss for a well-classified sample is down-weighted (easy sample).
$p_c^{(i)} \to 0$: Loss for a mis-classified sample is up-weighted (difficult sample).
$\gamma = 0$: Focal Loss = Cross Entropy.

Lin et. al., "Focal Loss for Dense Object Detection", *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Focal Loss

$$FL(\mathbf{y}^{(i)}, \mathbf{p}^{(i)}) = -\sum_{c \in C} \alpha_c \left(1 - p_c^{(i)}\right)^{\gamma} y_c^{(i)} \log p_c^{(i)}$$

α_c: weighting factor of class c. Inversely proportional to class frequency

Evaluation: Baseline Models

- Logistic Regression, Random Forest, Support Vector Machine (Radial Basis Function kernel).

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct. 2022.

M. Heinzinger et al., "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, Dec. 2024.

T. Hayes et al., "Simulating 500 million years of evolution with a language model," *Science*, Jan. 2025.

Evaluation: Baseline Models

- Logistic Regression, Random Forest, Support Vector Machine (Radial Basis Function kernel).
- Vision (adapted for text): Convolution Neural Network.

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct. 2022.

M. Heinzinger et al., "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, Dec. 2024.

T. Hayes et al., "Simulating 500 million years of evolution with a language model," *Science*, Jan. 2025.

Evaluation: Baseline Models

- Logistic Regression, Random Forest, Support Vector Machine (Radial Basis Function kernel).
- Vision (adapted for text): Convolution Neural Network.
- Language: Recurrent Neural Network, Long Short-Term Memory.

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct. 2022.

M. Heinzinger et al., "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, Dec. 2024.

T. Hayes et al., "Simulating 500 million years of evolution with a language model," *Science*, Jan. 2025.

Evaluation: Baseline Models

- Logistic Regression, Random Forest, Support Vector Machine (Radial Basis Function kernel).
- Vision (adapted for text): Convolution Neural Network.
- Language: Recurrent Neural Network, Long Short-Term Memory.
- Protein Language Models (pLMs): ProtT5, ProstT5, ESM3.

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct. 2022.

M. Heinzinger et al., "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, Dec. 2024.

T. Hayes et al., "Simulating 500 million years of evolution with a language model," *Science*, Jan. 2025.

Evaluation: Baseline Models

- Logistic Regression, Random Forest, Support Vector Machine (Radial Basis Function kernel).
- Vision (adapted for text): Convolution Neural Network.
- Language: Recurrent Neural Network, Long Short-Term Memory.
- Protein Language Models (pLMs): ProtT5, ProstT5, ESM3.

Model Name	# of Params	Size of pre-training dataset
VirProBERT	18.4M	1.2M
ProtT5	1.2B	45M
ProstT5	1.2B	35M
ESM3	98B	3.15B

A. Elnaggar et al., "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Oct. 2022.

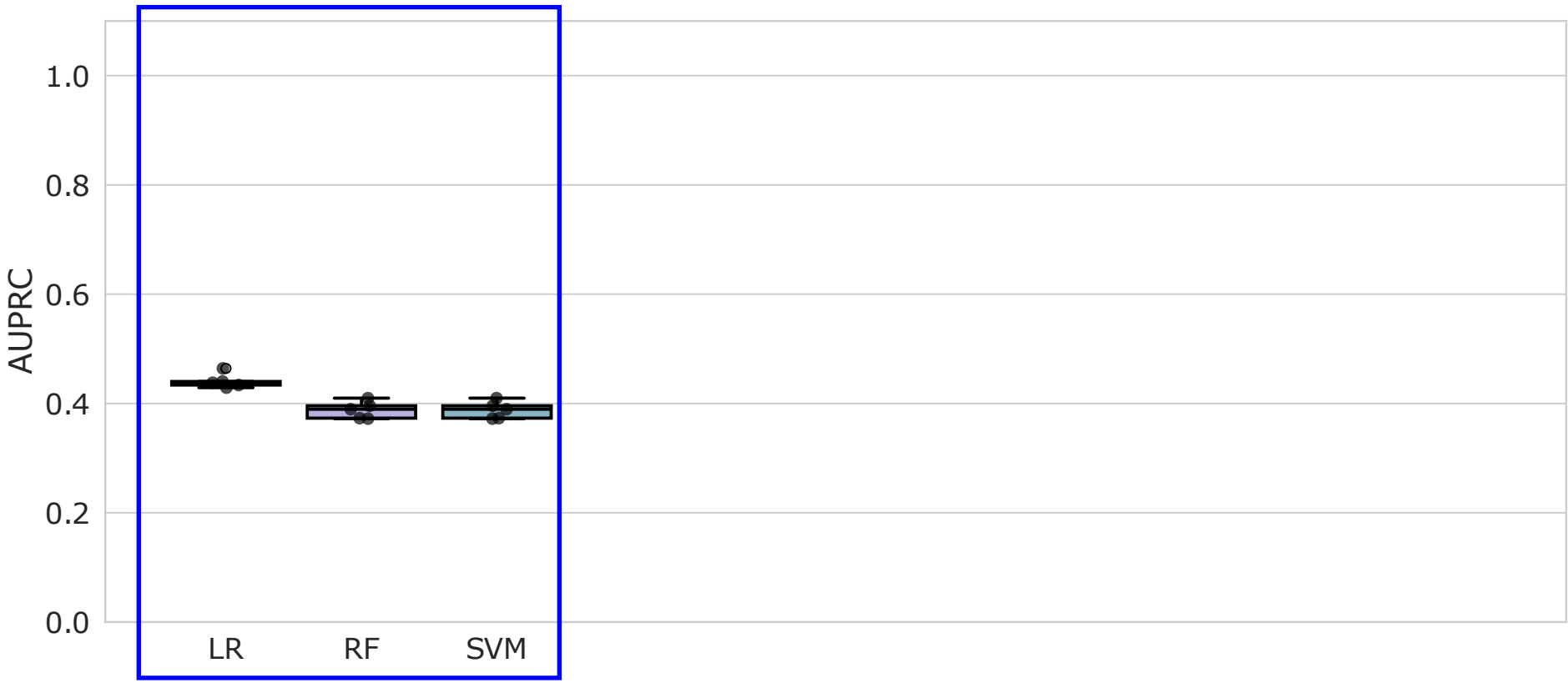
M. Heinzinger et al., "Bilingual language model for protein sequence and structure," *NAR Genomics and Bioinformatics*, Dec. 2024.

T. Hayes et al., "Simulating 500 million years of evolution with a language model," *Science*, Jan. 2025.

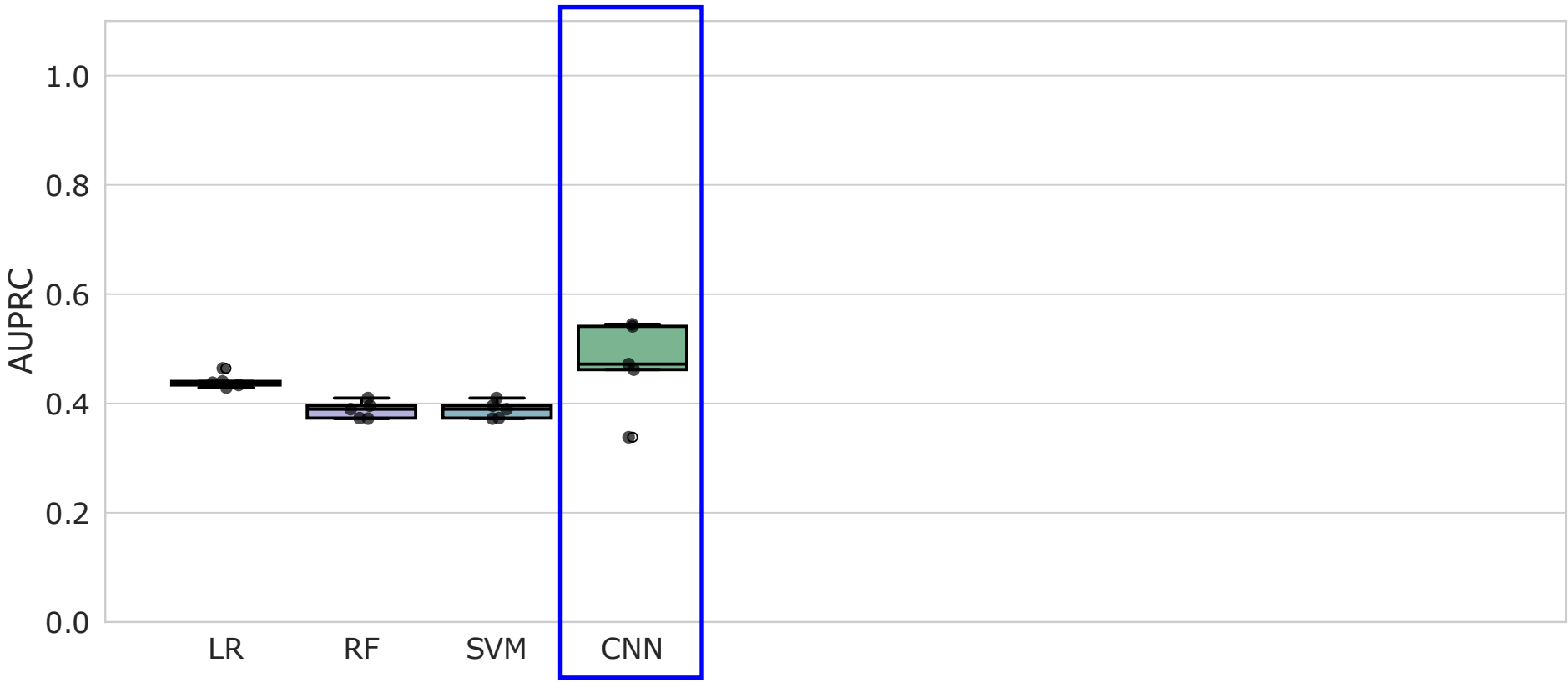
Outline

Predicting Common Hosts: Macro AUPRC

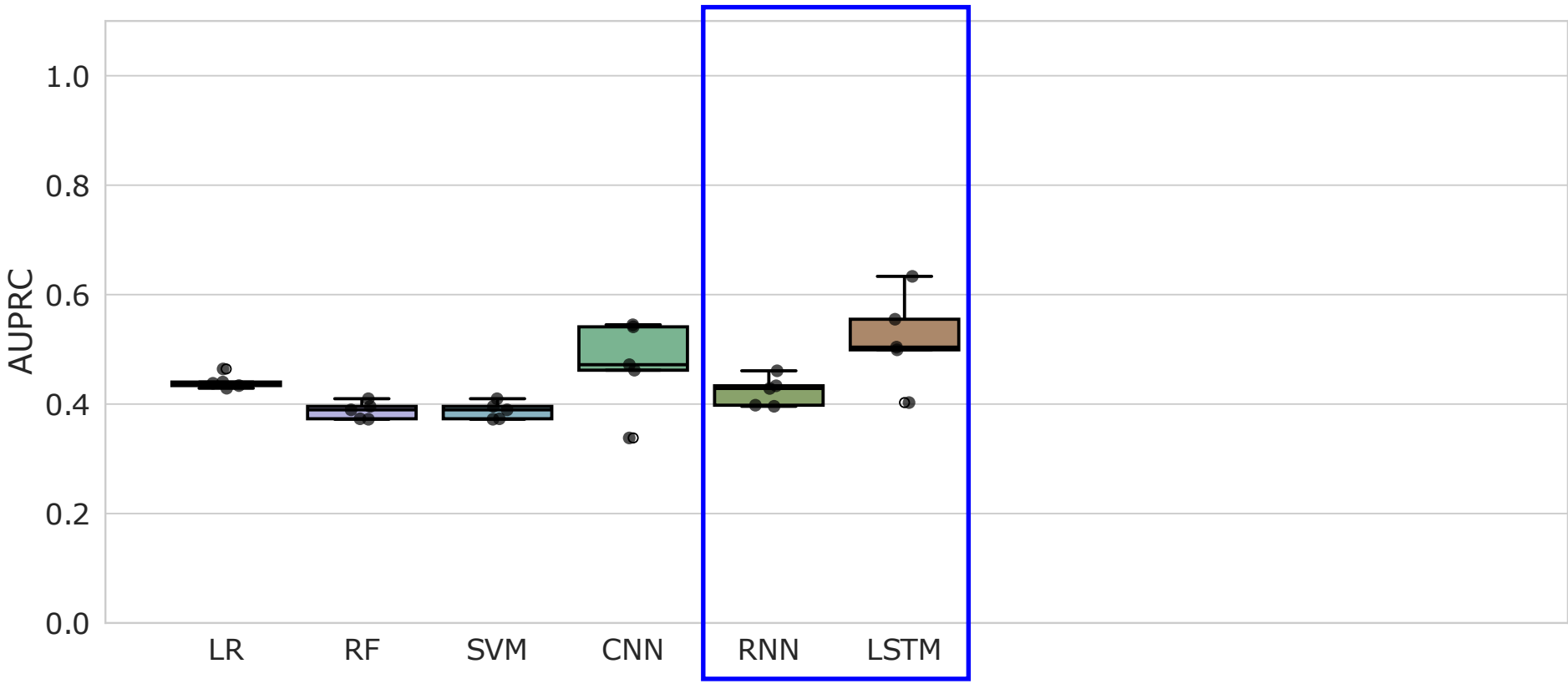
Predicting Common Hosts: Macro AUPRC



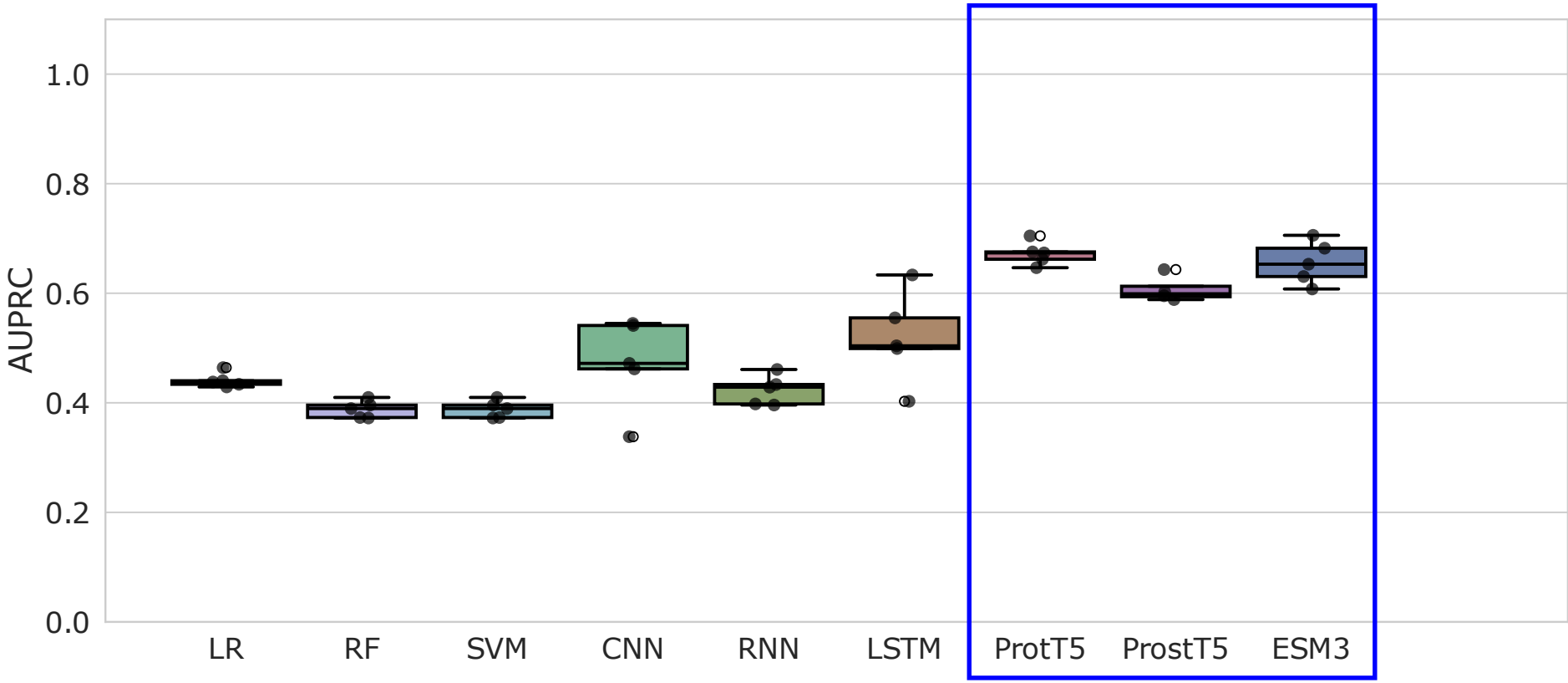
Predicting Common Hosts: Macro AUPRC



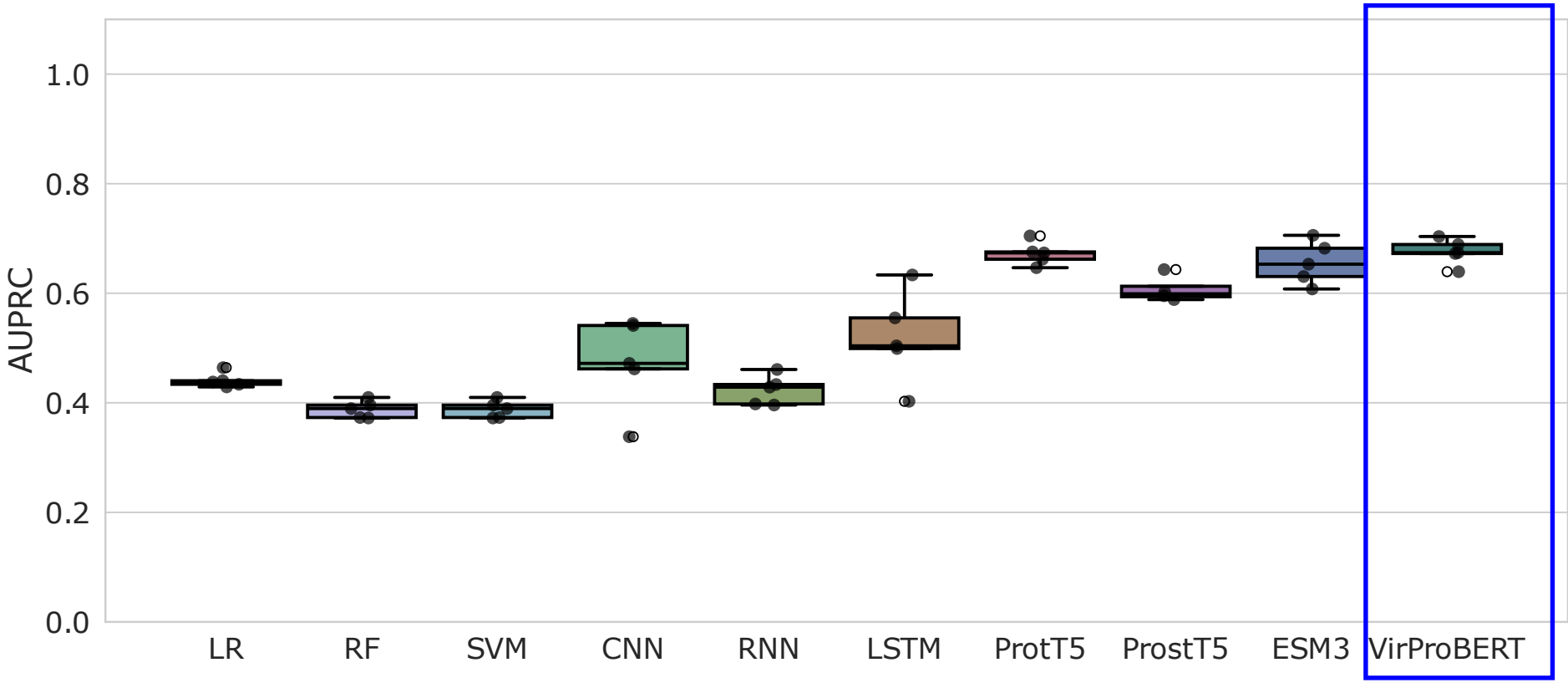
Predicting Common Hosts: Macro AUPRC



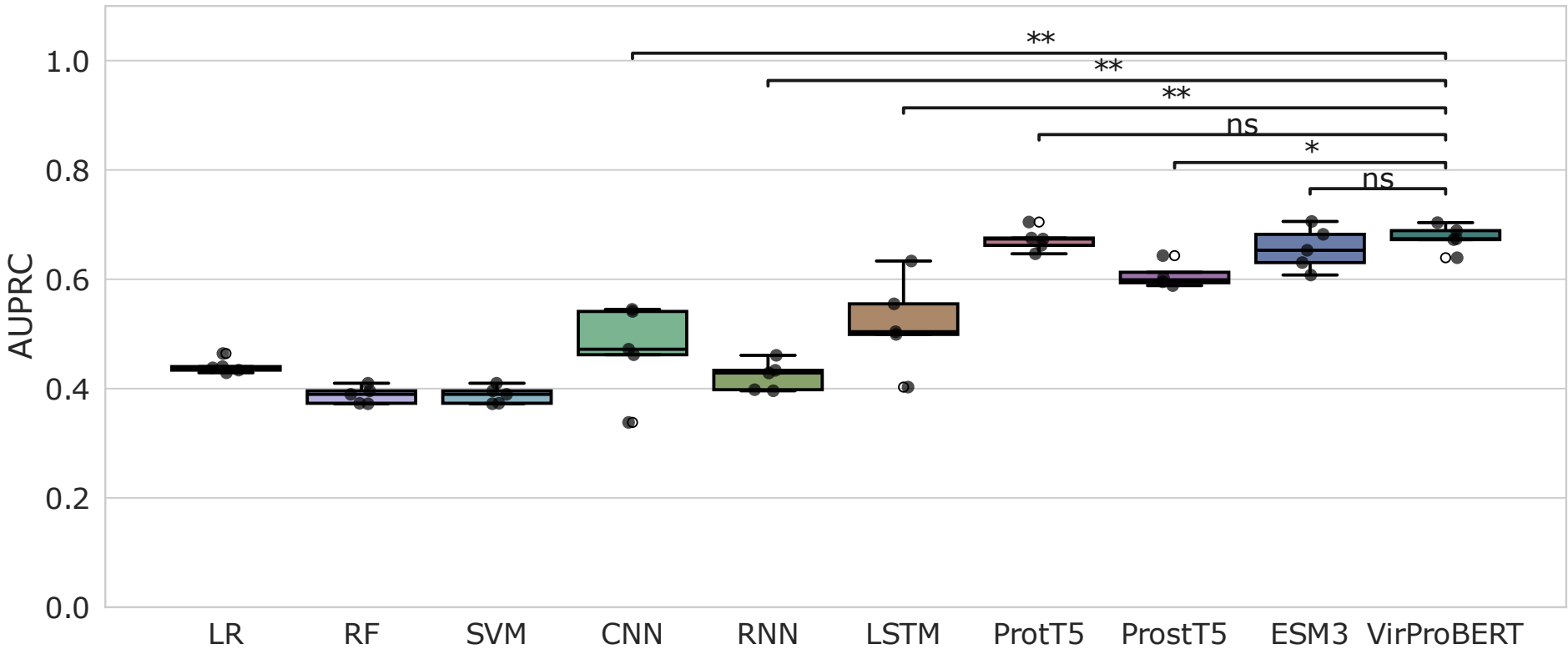
Predicting Common Hosts: Macro AUPRC



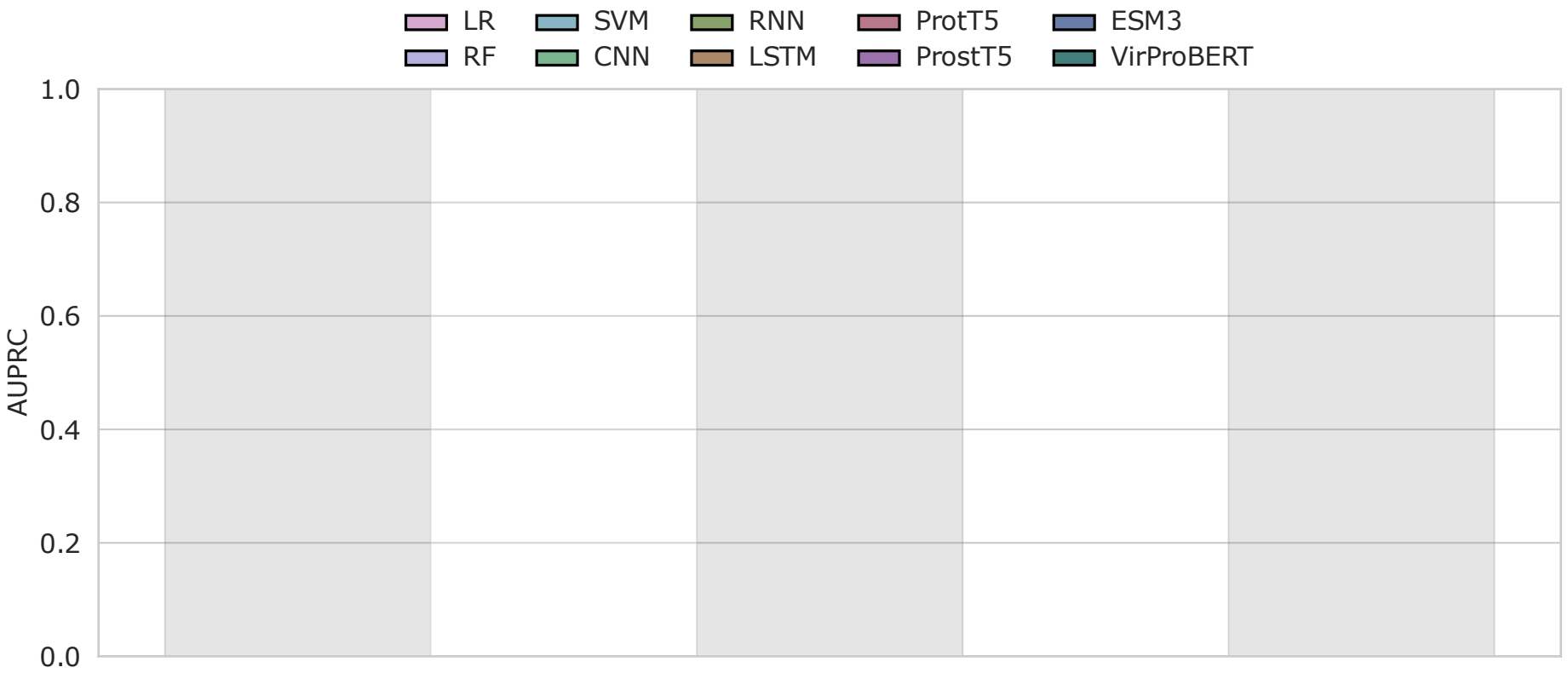
Predicting Common Hosts: Macro AUPRC



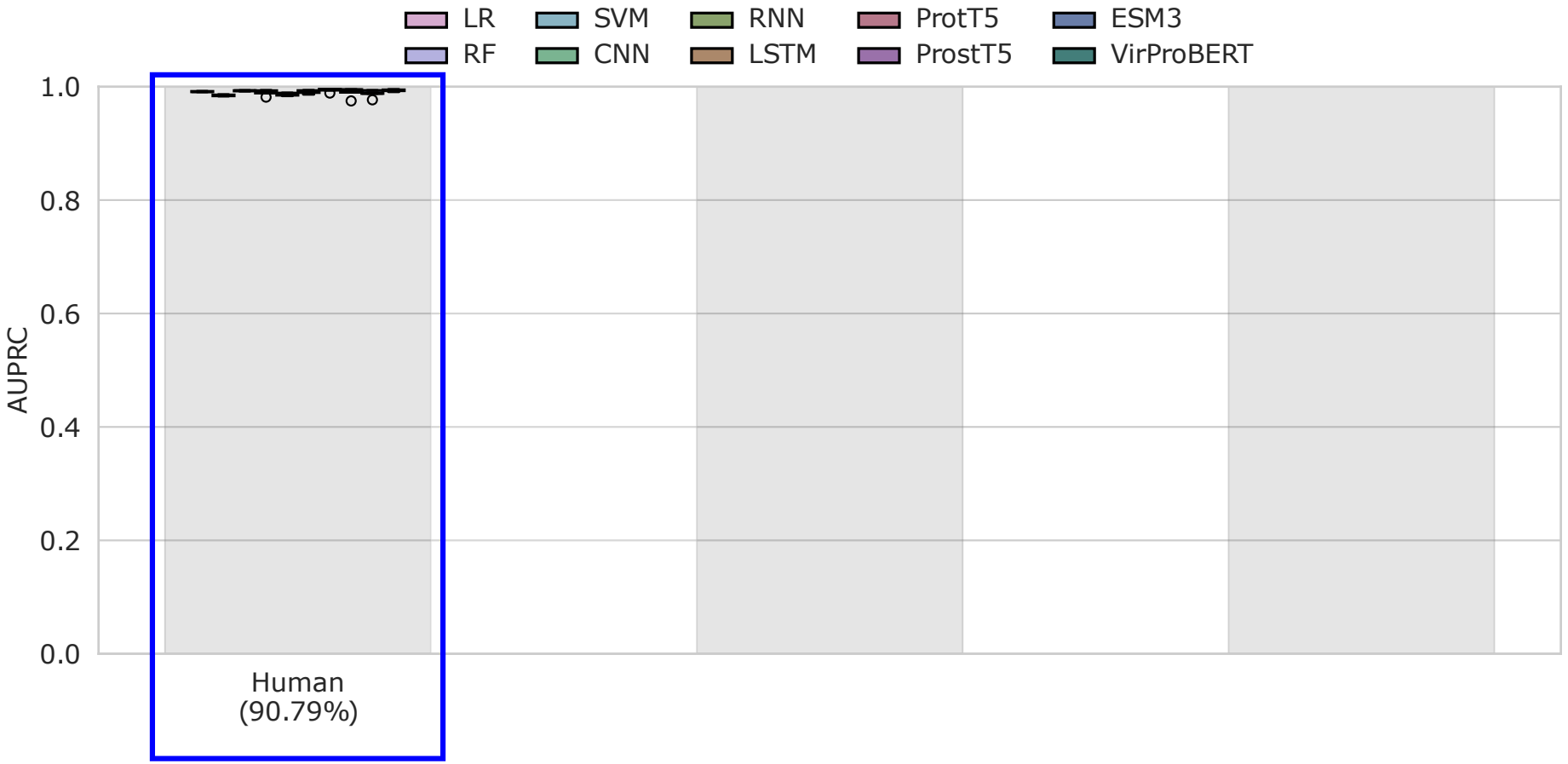
Predicting Common Hosts: Macro AUPRC



Predicting Common Hosts: Class-wise AUPRC

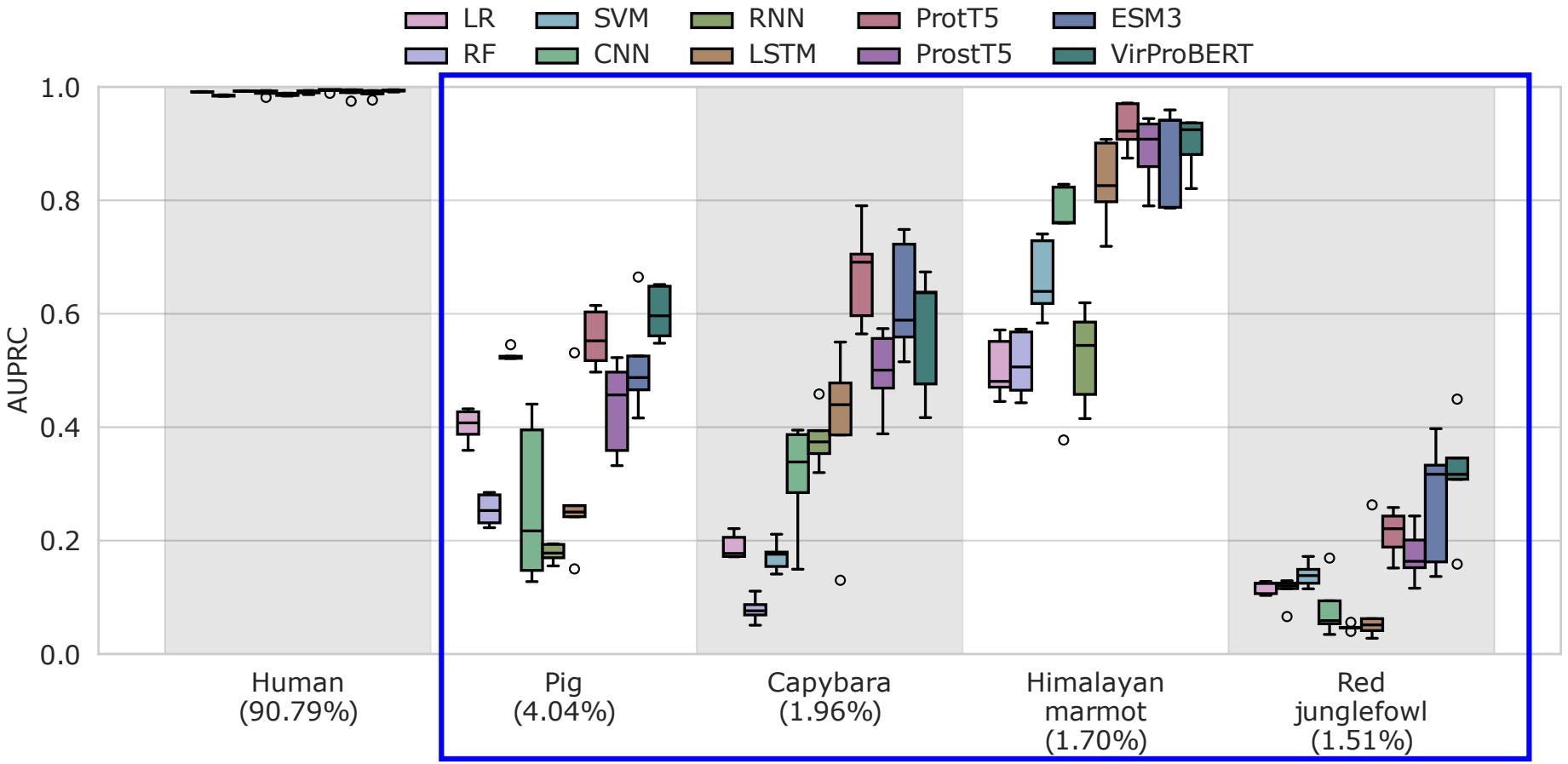


Predicting Common Hosts: Class-wise AUPRC



All models have virtually similar performance for the dominant human class.

Predicting Common Hosts: Class-wise AUPRC



VirProBERT, ProtT5, and ESM3 performed well in predicting the non-human classes despite their relative low prevalence.

Predicting Hosts of *Coronaviridae*

- Spike protein sequences of *Coronaviridae* from UniRef90.

Predicting Hosts of *Coronaviridae*

- Spike protein sequences of *Coronaviridae* from UniRef90.

Host	**Prevalence**
Chicken	49.28%
Human	24.64%
Cat	10.14%
Pig	7.25%
Gray wolf	4.35%
Ferret	1.45%
Horshoe bat	1.45%
Chinese rufous horshoe bat	1.45%

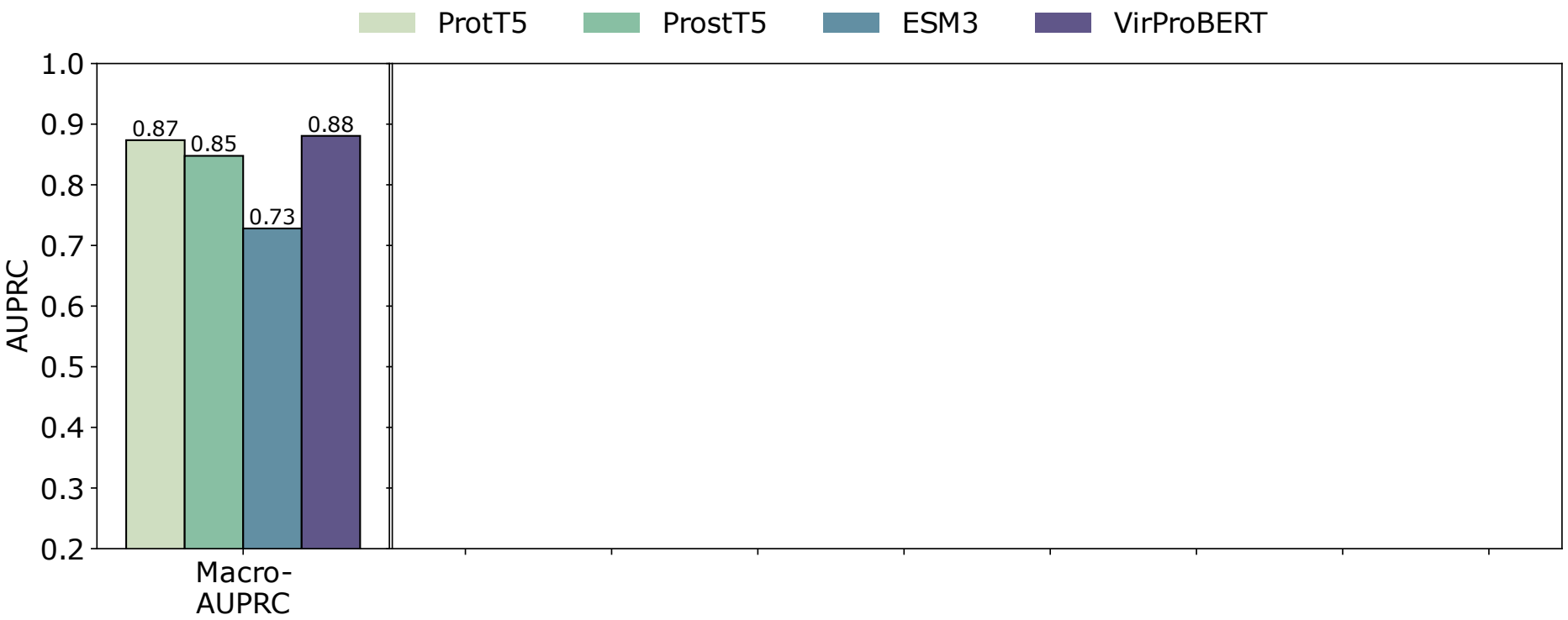
Predicting Hosts of *Coronaviridae*

- Spike protein sequences of *Coronaviridae* from UniRef90.

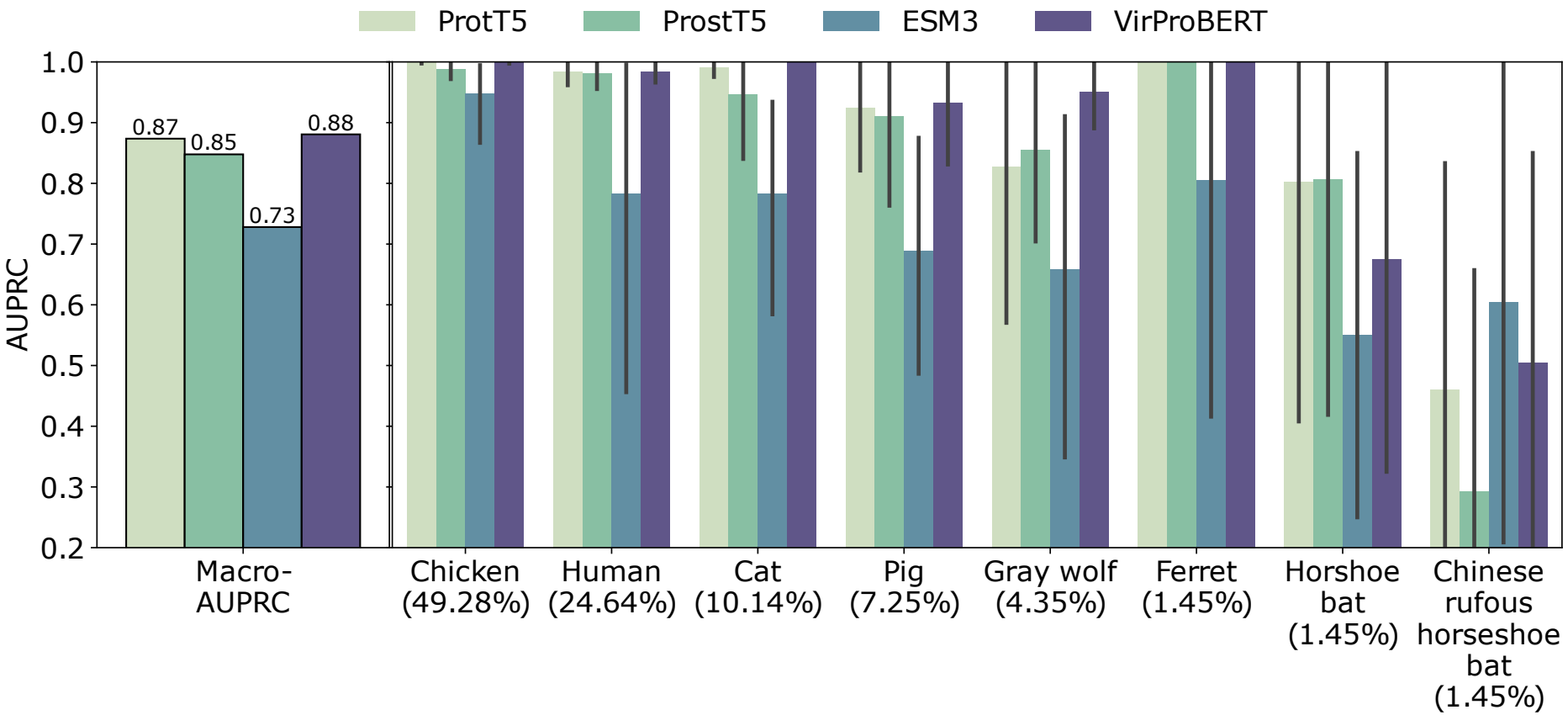
Host	**Prevalence**
Chicken	49.28%
Human	24.64%
Cat	10.14%
Pig	7.25%
Gray wolf	4.35%
Ferret	1.45%
Horshoe bat	1.45%
Chinese rufous horshoe bat	1.45%

- Dataset split: 80% training, 10% validation, 10% testing.

Predicting Hosts of *Coronaviridae*: Macro AUPRC



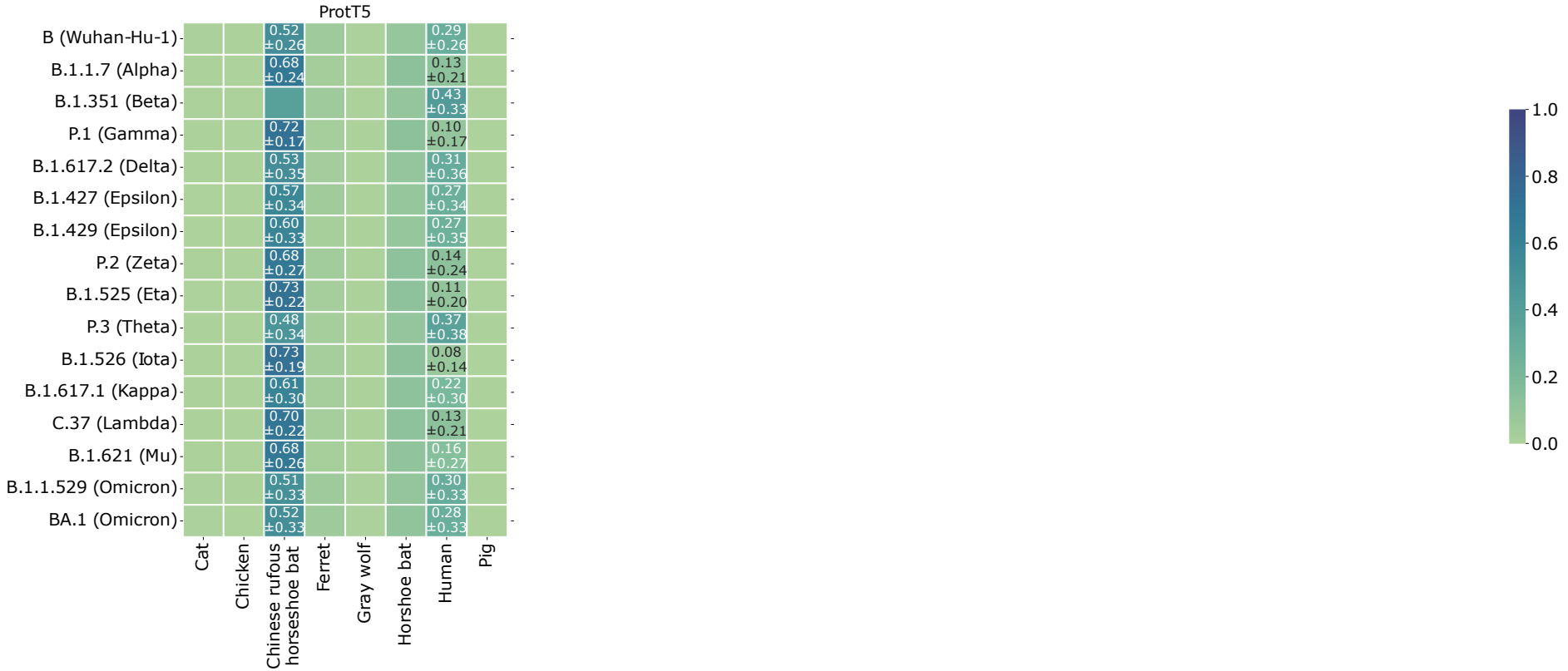
Predicting Hosts of *Coronaviridae*: Macro AUPRC



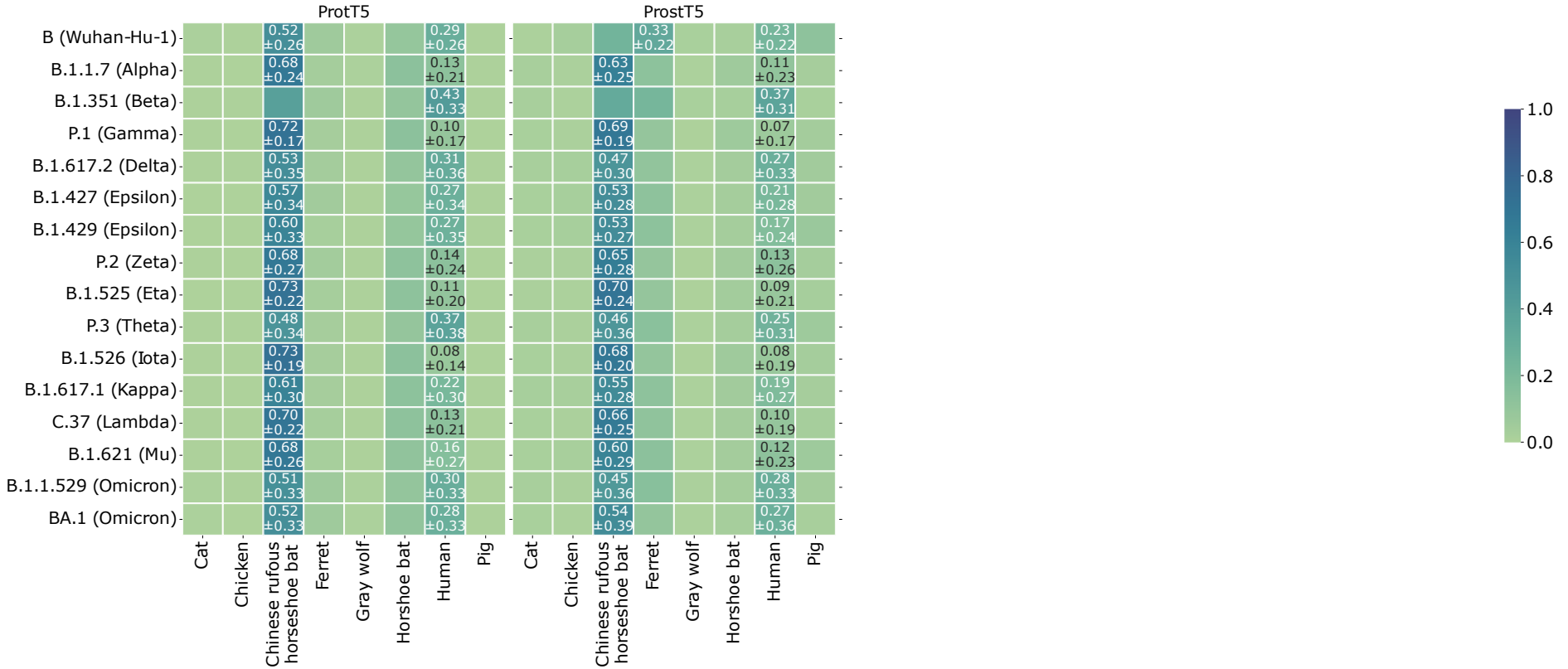
VirProBERT performs at par with SOTA pLMs.

SARS-CoV-2 Variants of Concern

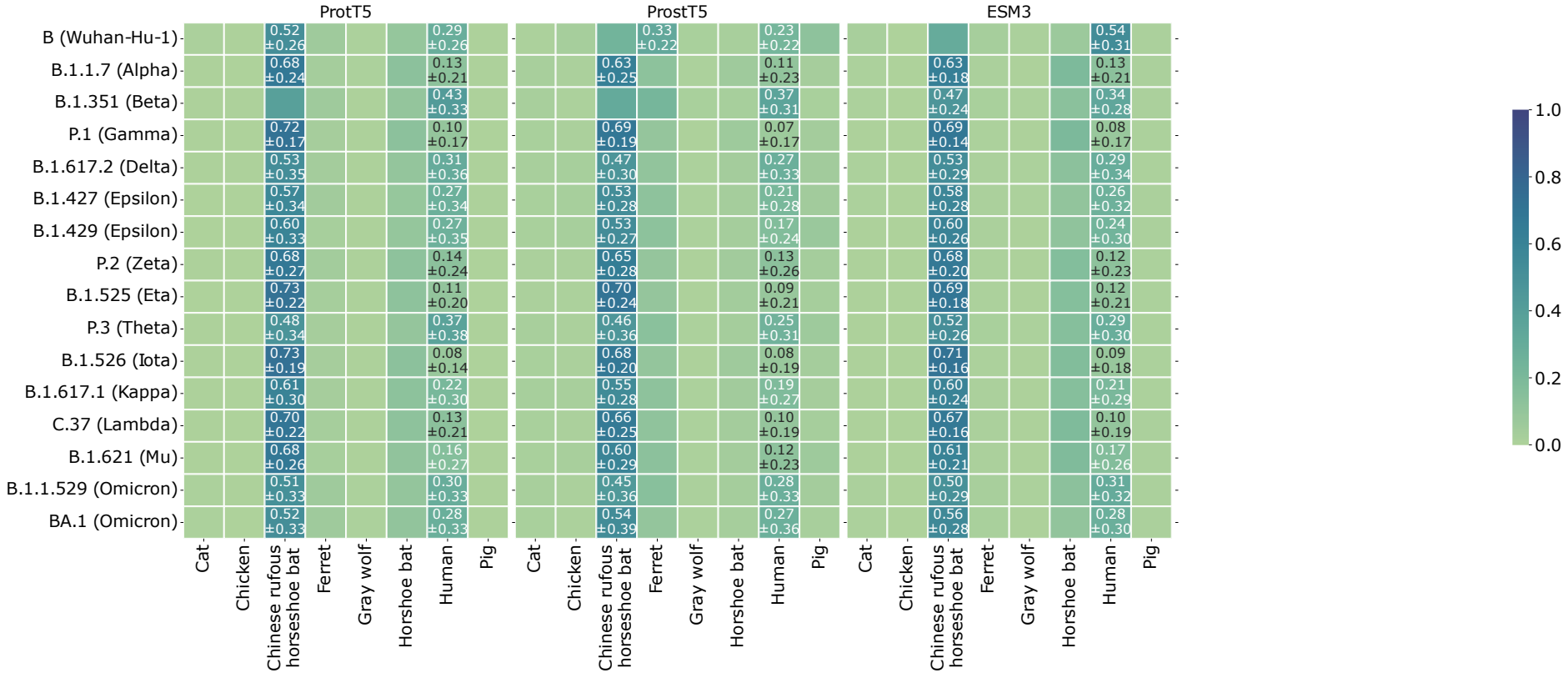
SARS-CoV-2 Variants of Concern



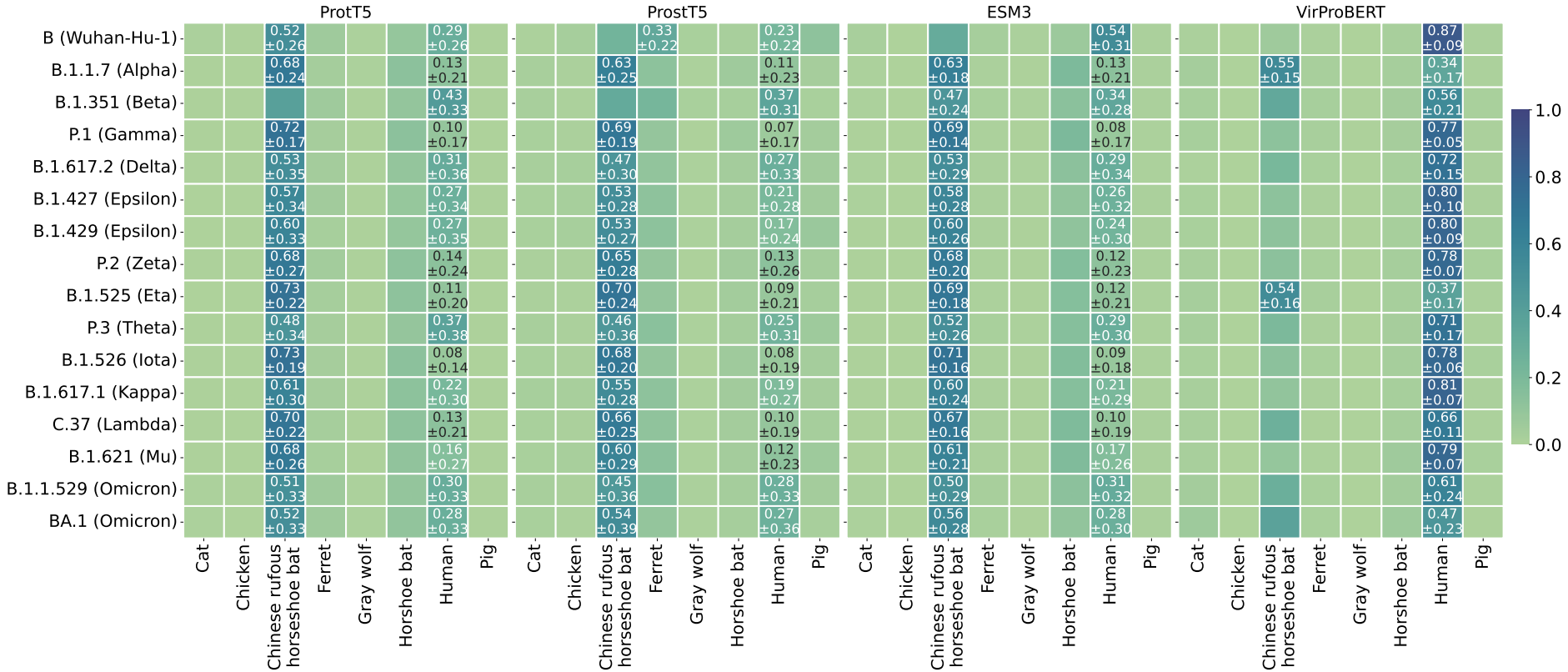
SARS-CoV-2 Variants of Concern



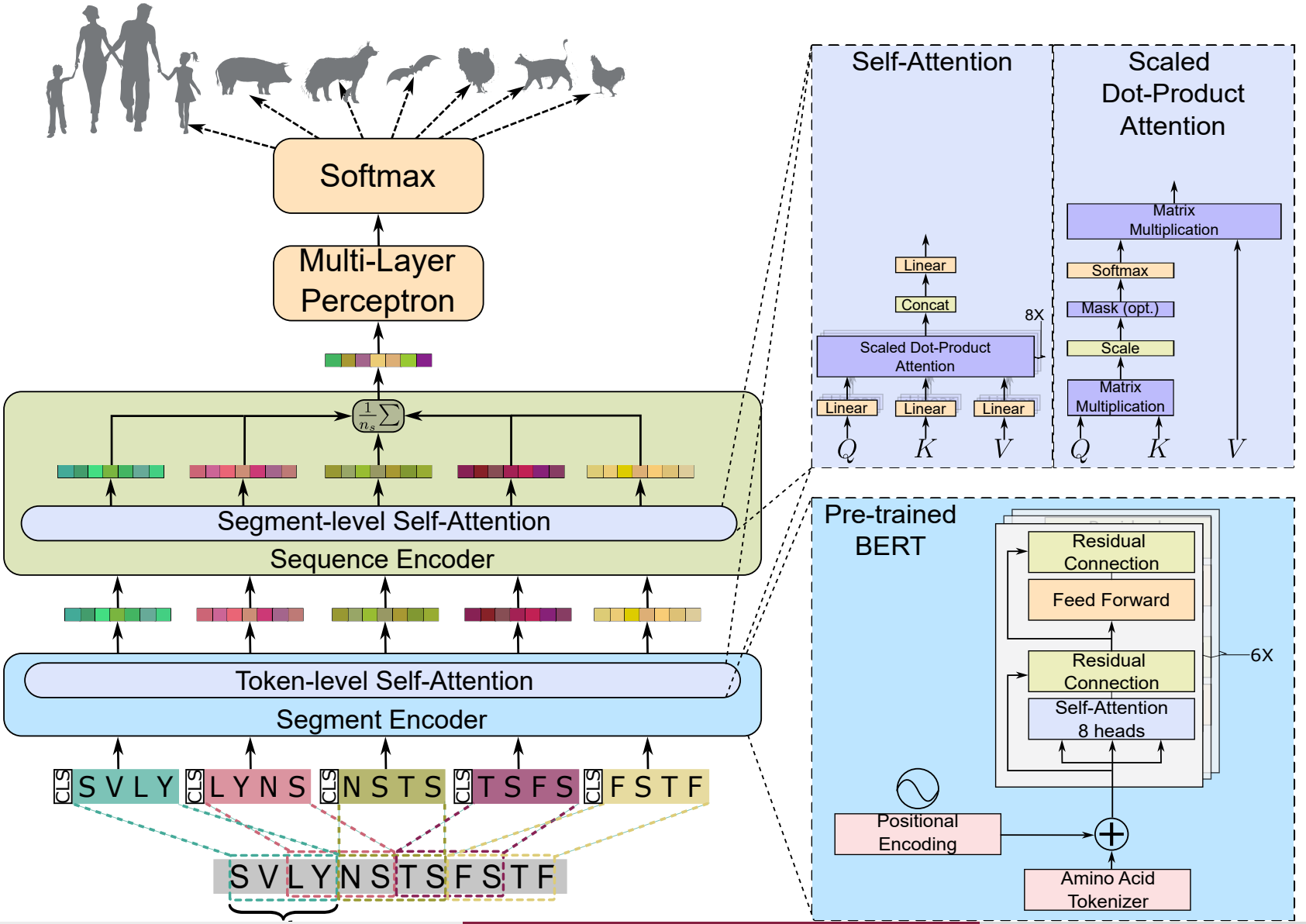
SARS-CoV-2 Variants of Concern



SARS-CoV-2 Variants of Concern



Ablation Study



Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

- without pre-training, segmentation, and hierarchical self-attention (w/o Pre-Tr, w/o Seg, w/o HSA)

Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

- without pre-training, segmentation, and hierarchical self-attention (w/o Pre-Tr, w/o Seg, w/o HSA)
- without pre-training and hierarchical self-attention (w/o Pre-Tr, w/o HSA)

Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

- without pre-training, segmentation, and hierarchical self-attention (w/o Pre-Tr, w/o Seg, w/o HSA)
- without pre-training and hierarchical self-attention (w/o Pre-Tr, w/o HSA)
- without pre-training (w/o Pre-Tr)

Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

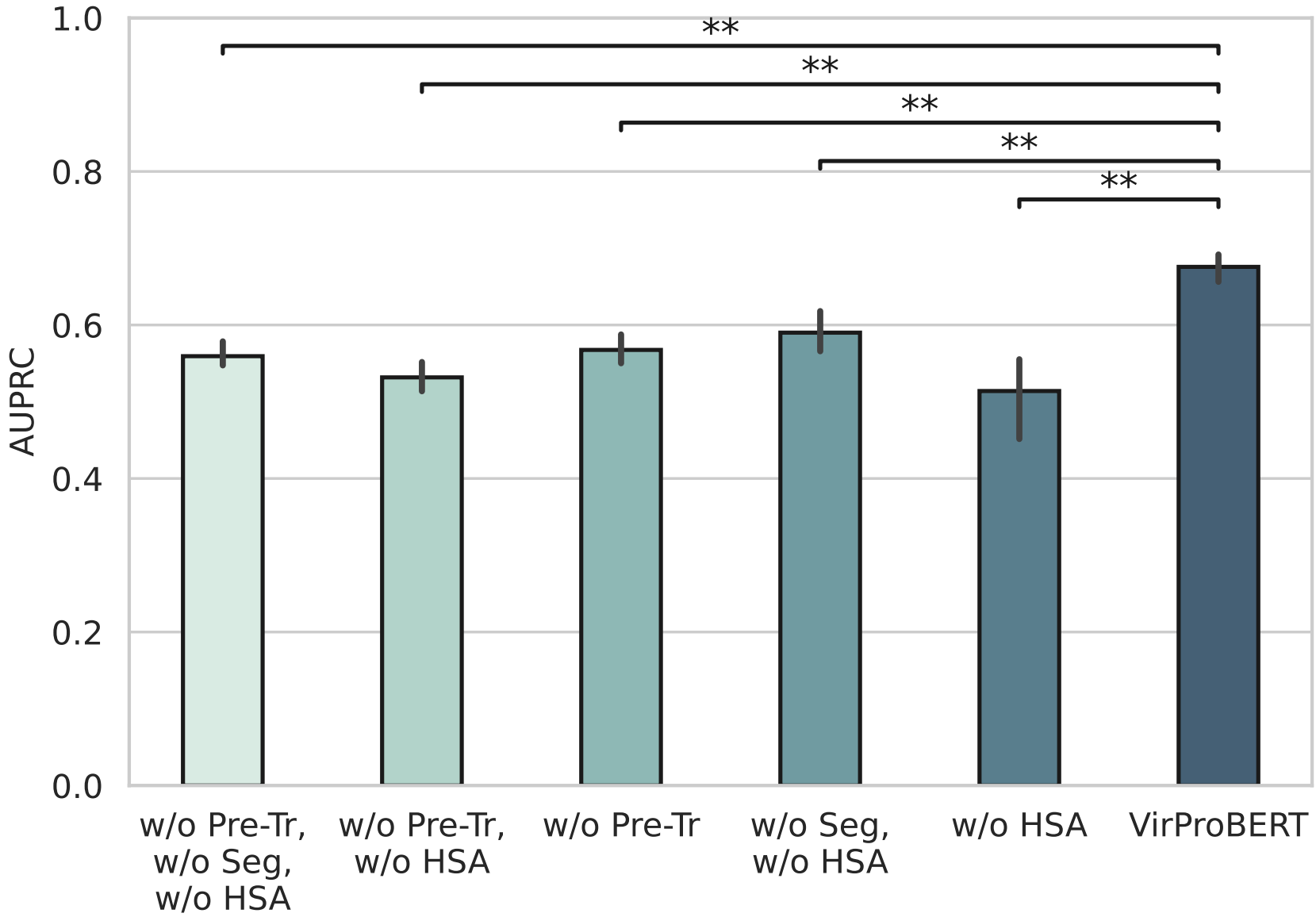
- without pre-training, segmentation, and hierarchical self-attention (w/o Pre-Tr, w/o Seg, w/o HSA)
- without pre-training and hierarchical self-attention (w/o Pre-Tr, w/o HSA)
- without pre-training (w/o Pre-Tr)
- without segmentation and hierarchical self-attention (w/o Seg, w/o HSA)

Ablation Study

Three main components of VirProBERT: pre-training, segmentation, and hierarchical self-attention.
Five combinations of components:

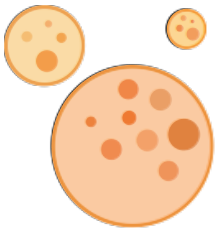
- without pre-training, segmentation, and hierarchical self-attention (w/o Pre-Tr, w/o Seg, w/o HSA)
- without pre-training and hierarchical self-attention (w/o Pre-Tr, w/o HSA)
- without pre-training (w/o Pre-Tr)
- without segmentation and hierarchical self-attention (w/o Seg, w/o HSA)
- without hierarchical self-attention (w/o HSA)

Ablation Study

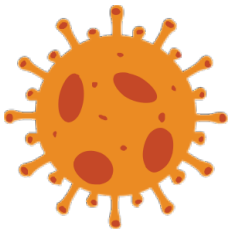


Outline

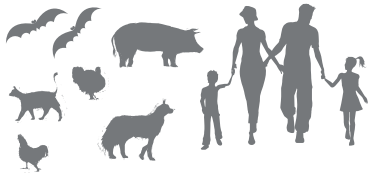
Generalizing to Rare & Unseen Hosts and Unseen Viruses



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Pre-training Dataset

1.2 million sequences

267,865 sequences

3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

Hosts with atleast 1% prevalence

Hosts with prevalence < 1%

Rare Hosts

16,074 sequences
143 virus hosts

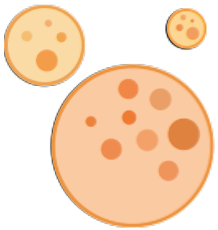
Immunodeficiency virus (IDV)

220,068 sequences
7 viruses
40 virus hosts

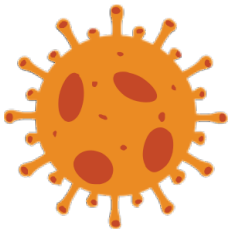
Non-Immunodeficiency virus (Non-IDV)

47,792 sequences
3,772 viruses
1,304 virus hosts

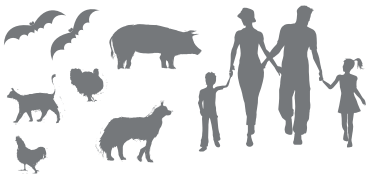
Generalizing to Rare & Unseen Hosts and Unseen Viruses



Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Pre-training Dataset

1.2 million sequences

267,865 sequences

3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

Hosts with atleast 1% prevalence

Hosts with prevalence < 1%

Rare Hosts

16,074 sequences
143 virus hosts

Unseen Hosts

Unseen Viruses

Immunodeficiency virus (IDV)

220,068 sequences
7 viruses
40 virus hosts

Non-Immunodeficiency virus (Non-IDV)

47,792 sequences
3,772 viruses
1,304 virus hosts

Few-Shot Learning (FSL)

Train ML models to predict from a small number of examples.
Terminology:

- N: Number of classes.
- K: Number of **support** examples or **shots**.
- N-way, K-shot classification.

FSL Data Splits

An **episode** is composed of N classes, $K \times N$ **support** samples, and $Q \times N$ **query** samples.

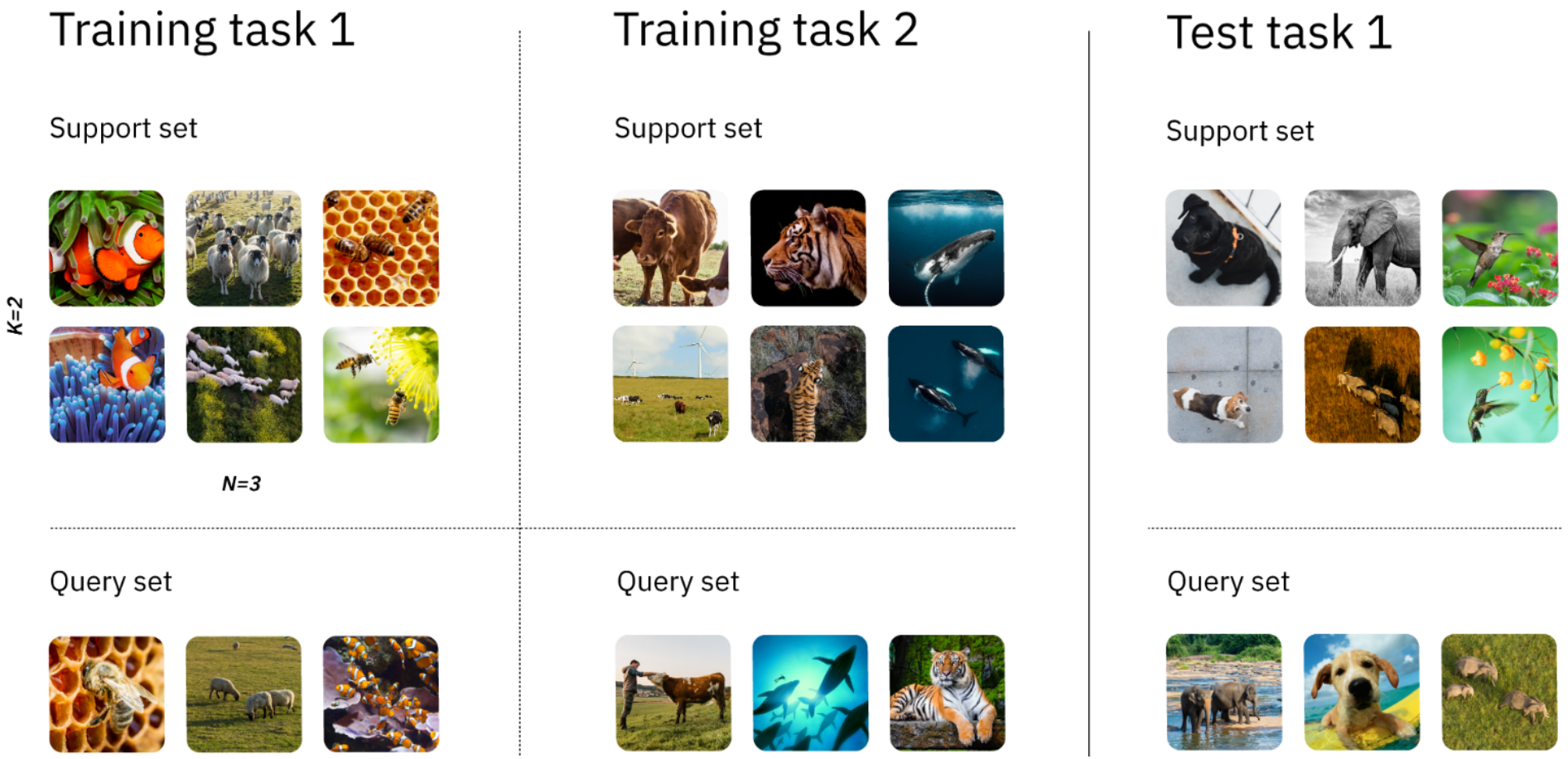
https://www.ibm.com/think/topics/few-shot-learning

FSL Data Splits

An **episode** is composed of N classes, $K \times N$ **support** samples, and $Q \times N$ **query** samples.

https://www.ibm.com/think/topics/few-shot-learning

FSL Data Splits

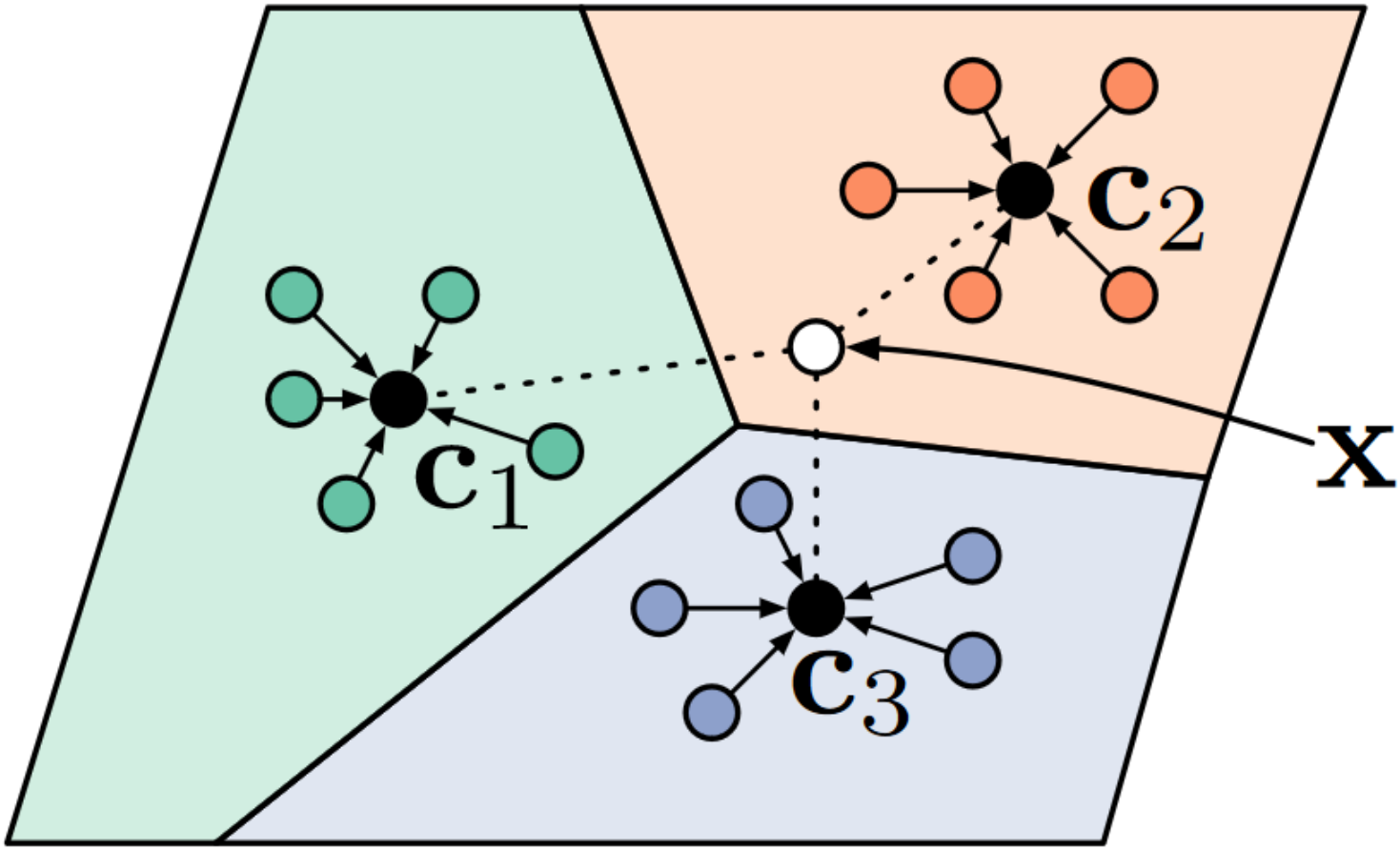


Dataset is split into training and testing sets based on **class labels**.

https://www.ibm.com/think/topics/few-shot-learning

Prototypical Networks

Probability for each class is **inversely proportional** to the euclidean distance from the class prototype.



J. Snell, et. al., "Prototypical Networks for Few-shot Learning," *Neural Information Processing Systems.*, 2017.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for Q query sequences: probabilities for each of the N classes.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for Q query sequences: probabilities for each of the N classes.
  - Compute cross entropy loss on query sequence predictions.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for Q query sequences: probabilities for each of the N classes.
  - Compute cross entropy loss on query sequence predictions.
  - Back propagate loss to update weights of VirProBERT.
- Testing:
  - Compute N prototypes using $N \times K$ support sequences.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for Q query sequences: probabilities for each of the N classes.
  - Compute cross entropy loss on query sequence predictions.
  - Back propagate loss to update weights of VirProBERT.
- Testing:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for **all remaining** sequences: probabilities for each of the N classes.

FSL in a Nutshell

- Dataset: Rare classes in Non-IV dataset.
- Each episode (batch): N classes, K support sequences for each class, Q query sequences for each class.
- Training, Validation, and Testing dataset split based on **labels**.
- Training:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for Q query sequences: probabilities for each of the N classes.
  - Compute cross entropy loss on query sequence predictions.
  - Back propagate loss to update weights of VirProBERT.
- Testing:
  - Compute N prototypes using $N \times K$ support sequences.
  - Prediction for **all remaining** sequences: probabilities for each of the N classes.
  - AUPRC for each class in the episode.

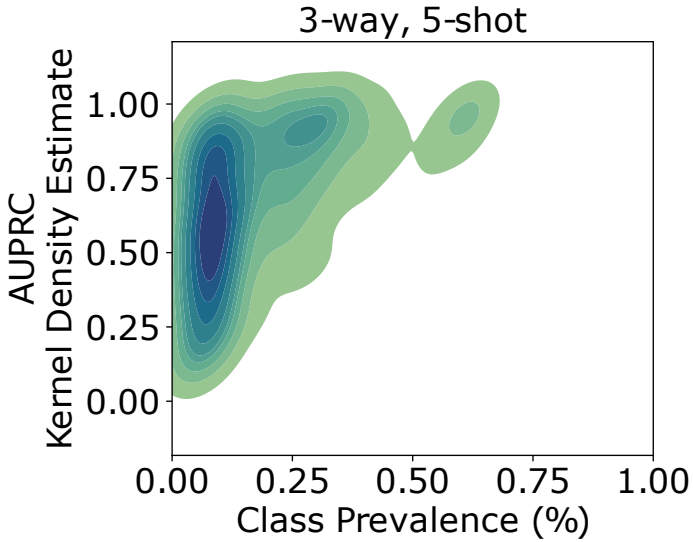
Selecting Best Configuration for FSL

Dataset: Rare classes in Non-IV dataset.

- 143 hosts with prevalence bewteen 0.05% and 1%.
- Hosts with at least *six* samples.

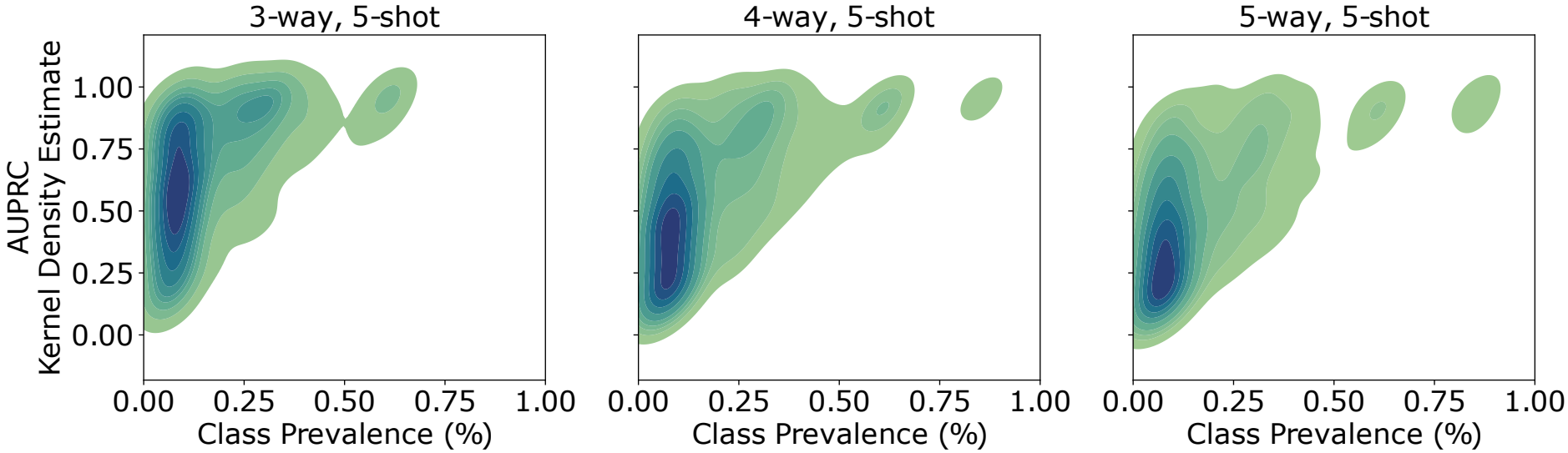
Selecting Best Configuration for FSL

Predicting **unseen and rare** hosts in Non-IV dataset.



Selecting Best Configuration for FSL

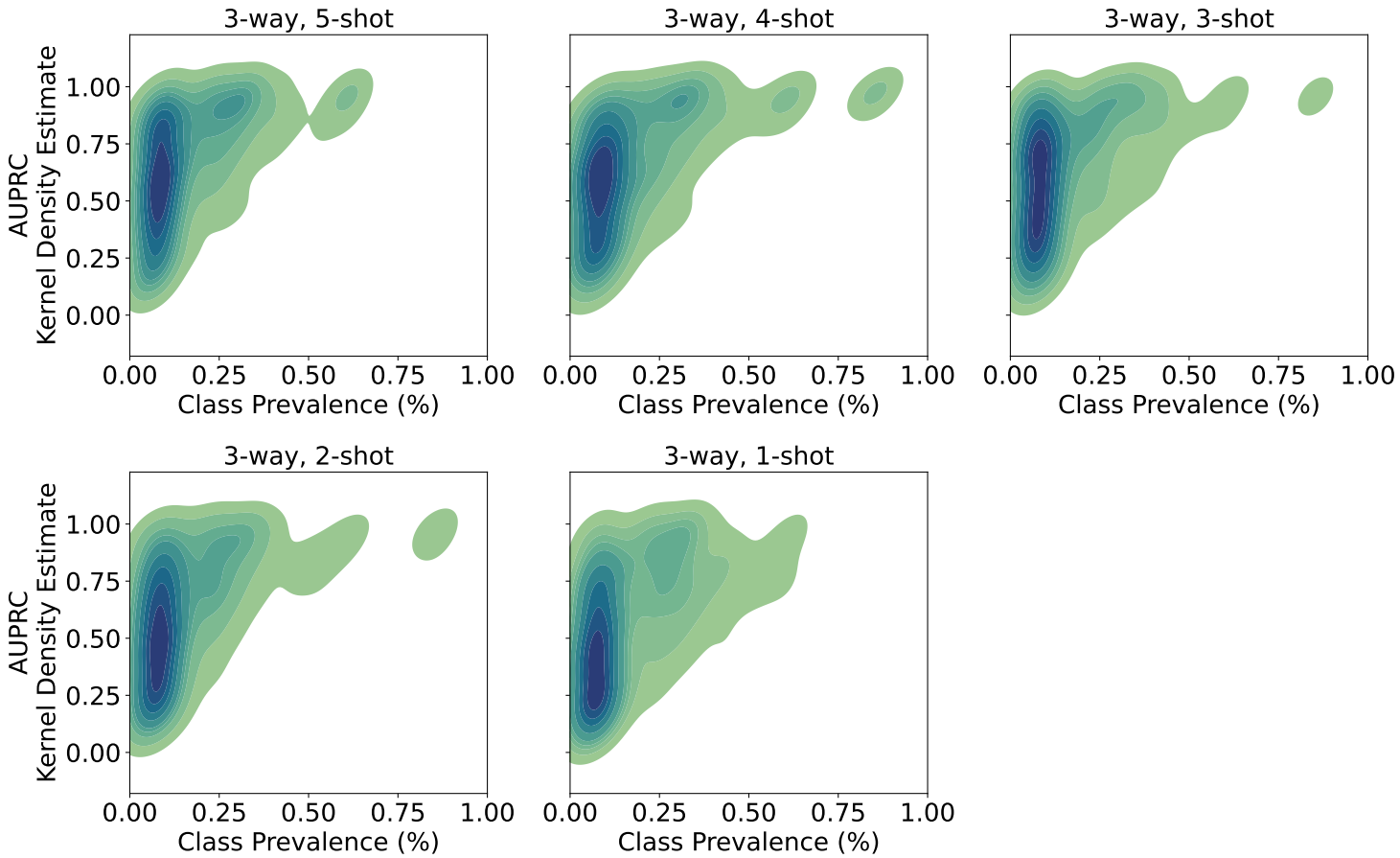
Predicting **unseen and rare** hosts in Non-IV dataset.



N-**way,** 5-**shot**: Performance decreases as the number of classes increases.

Selecting Best Configuration for FSL

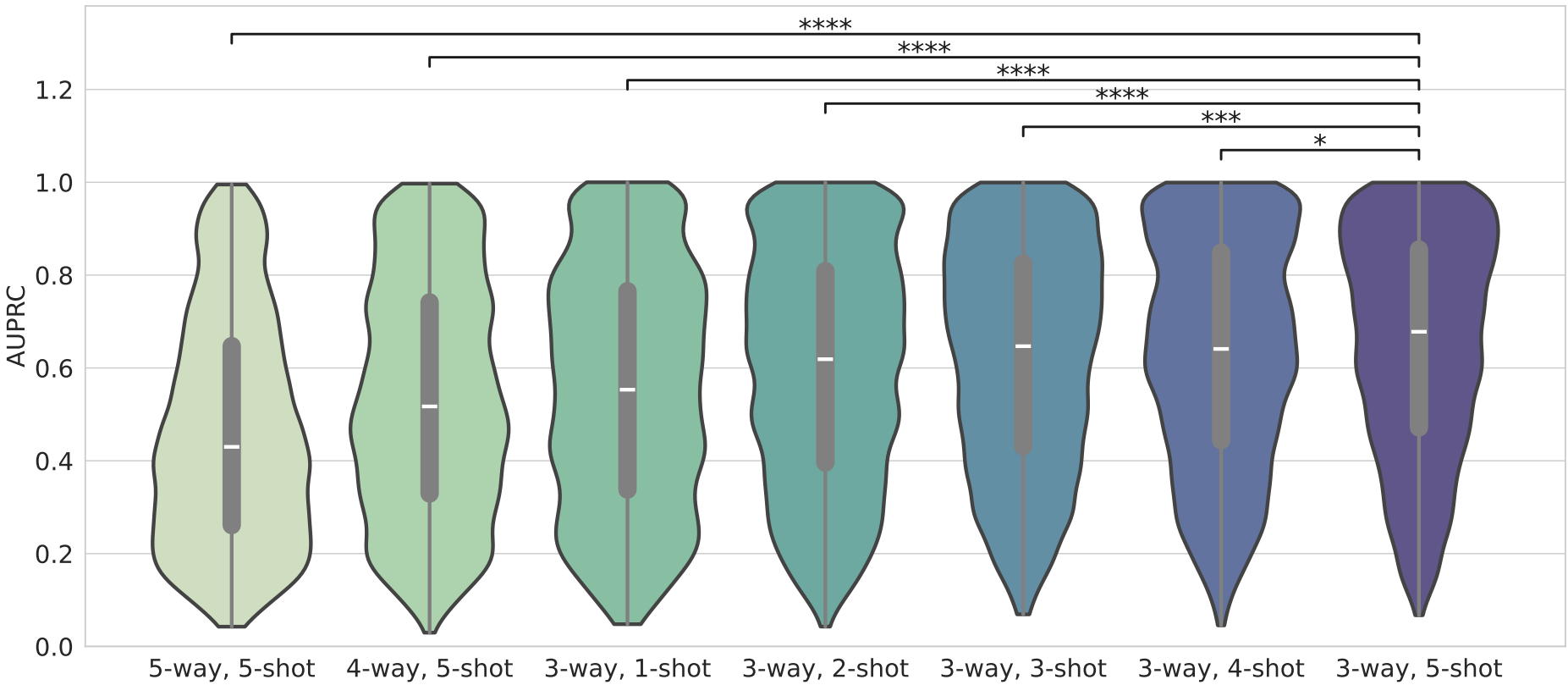
Predicting **unseen and rare** hosts in Non-IV dataset.



3-**way,** K-**shot**: Performance decreases as the number of shots decreases.

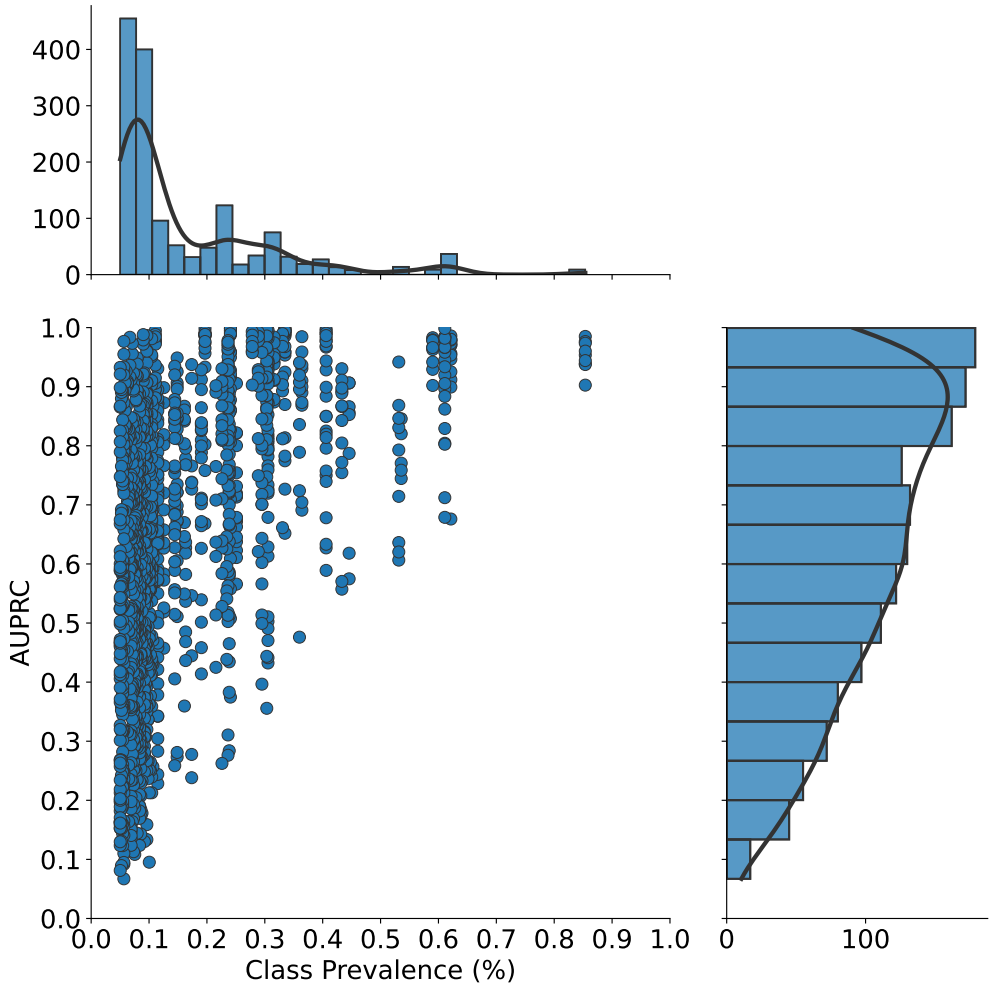
Selecting Best Configuration for FSL

Predicting **unseen and rare** hosts in Non-IV dataset.



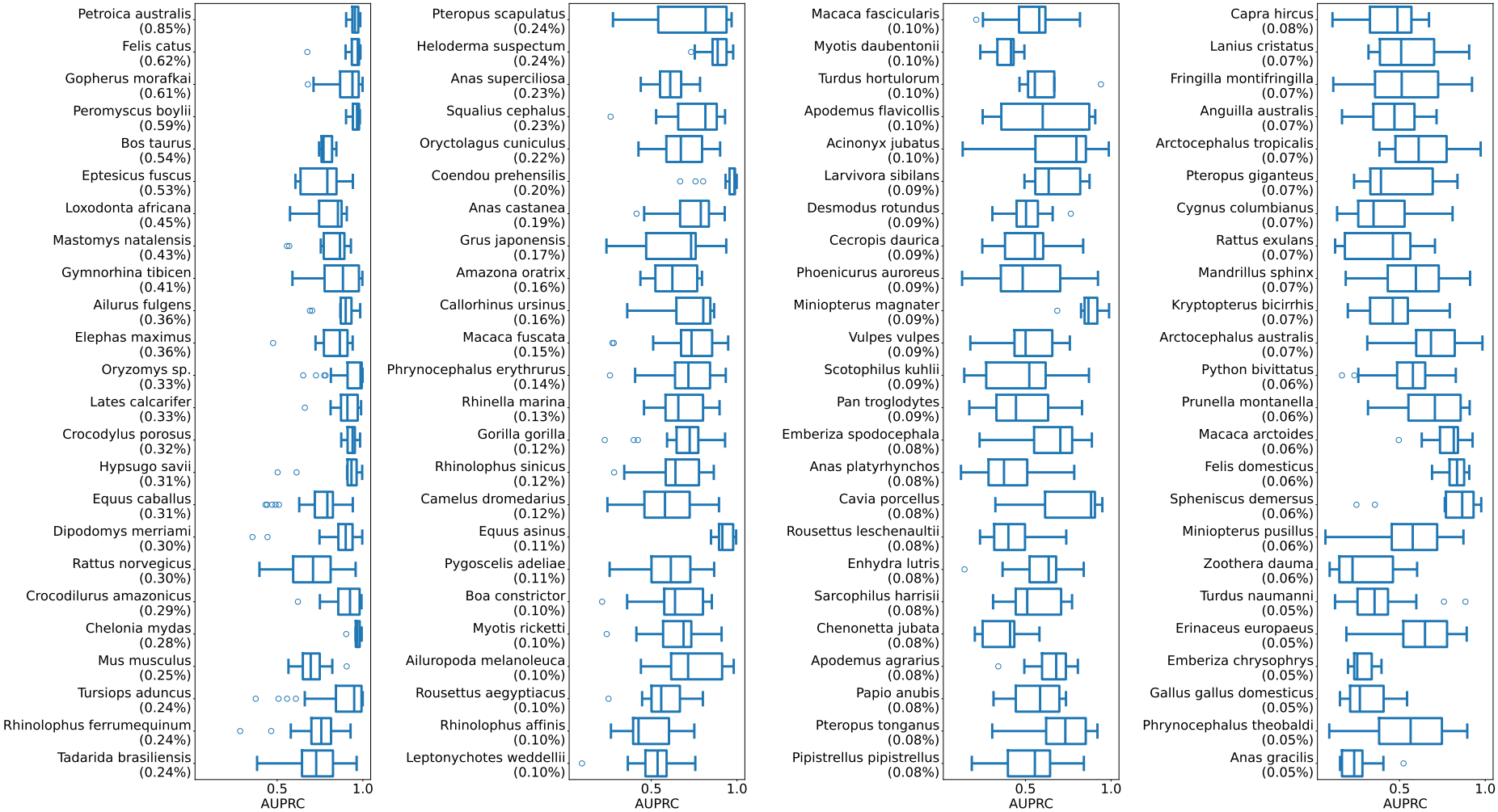
3-way, 5-shot classifier performs significantly better.

Predicting Rare & Unseen Hosts using 3-Way, 5-Shot FSL



The median class prevalence was 0.09% (45 samples), yet the median AUPRC was as high as 0.68

Predicting Rare & Unseen Hosts using 3-Way, 5-Shot FSL

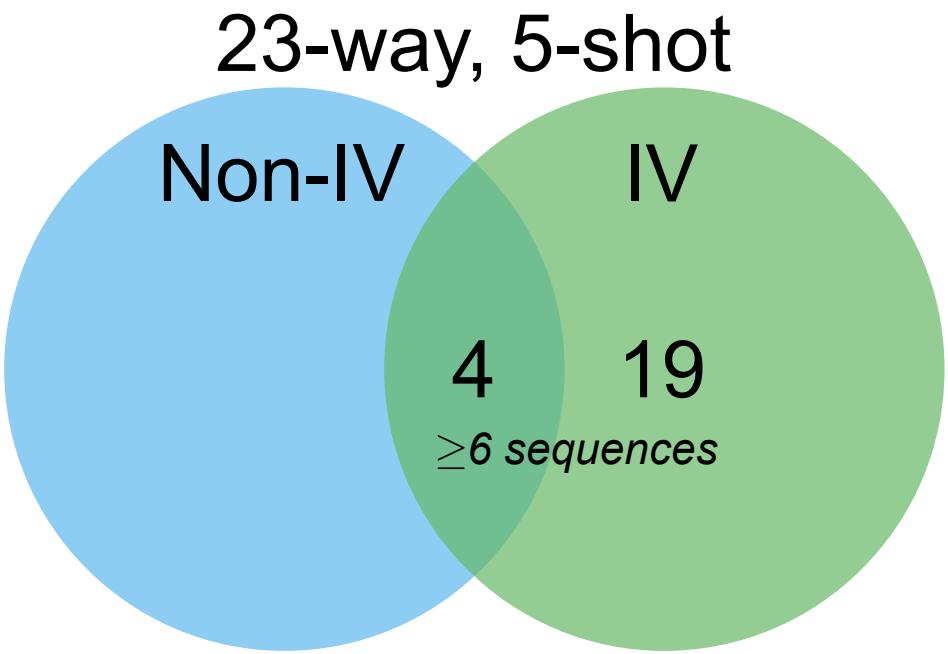


Mean AUPRC for each class ranged from 0.96 to 0.26 with a decreasing trend as the prevalence of rare classes changed from 1% to 0.05%.

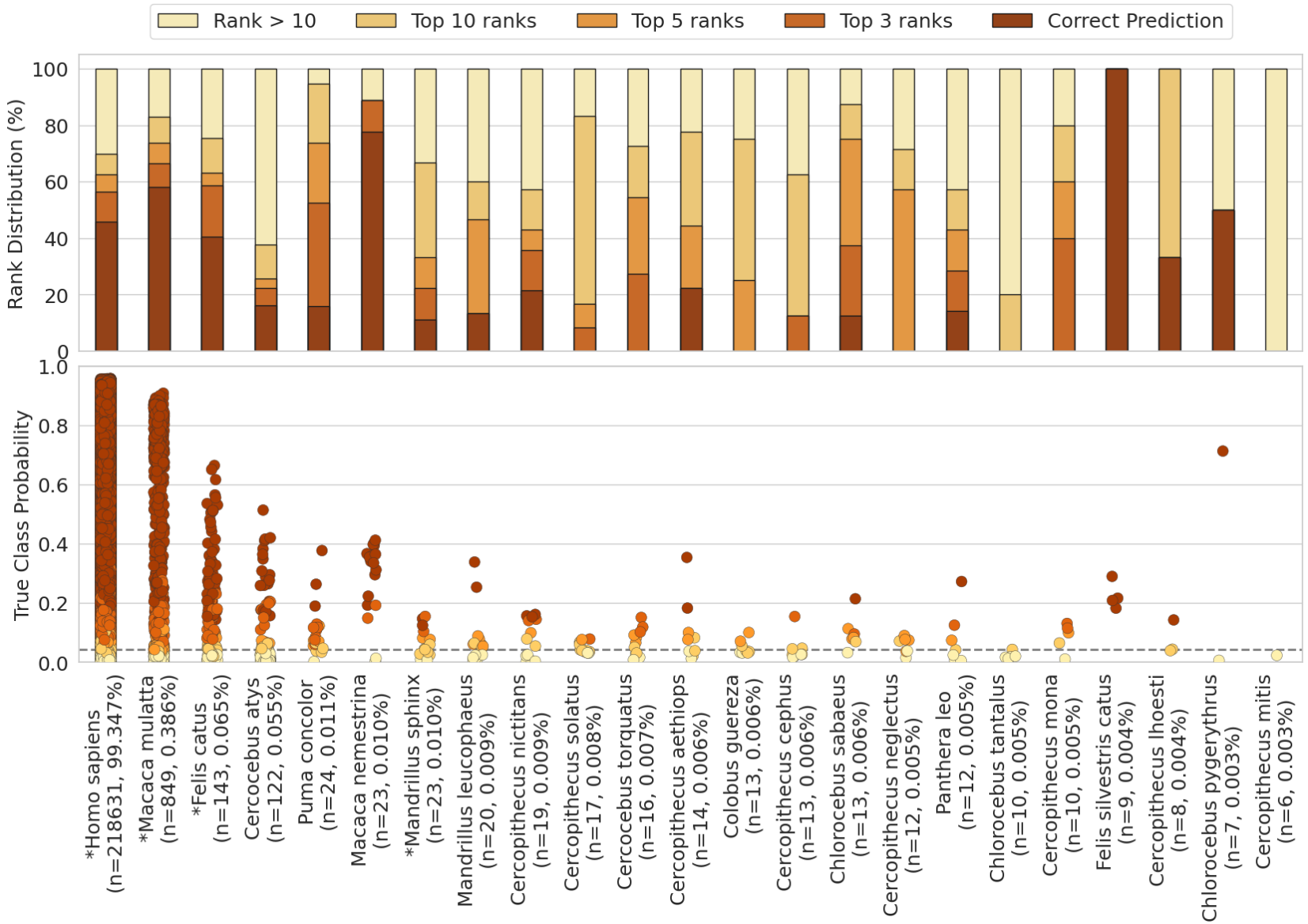
Predicting Hosts of Unseen Viruses

Dataset: Hosts in IV dataset.

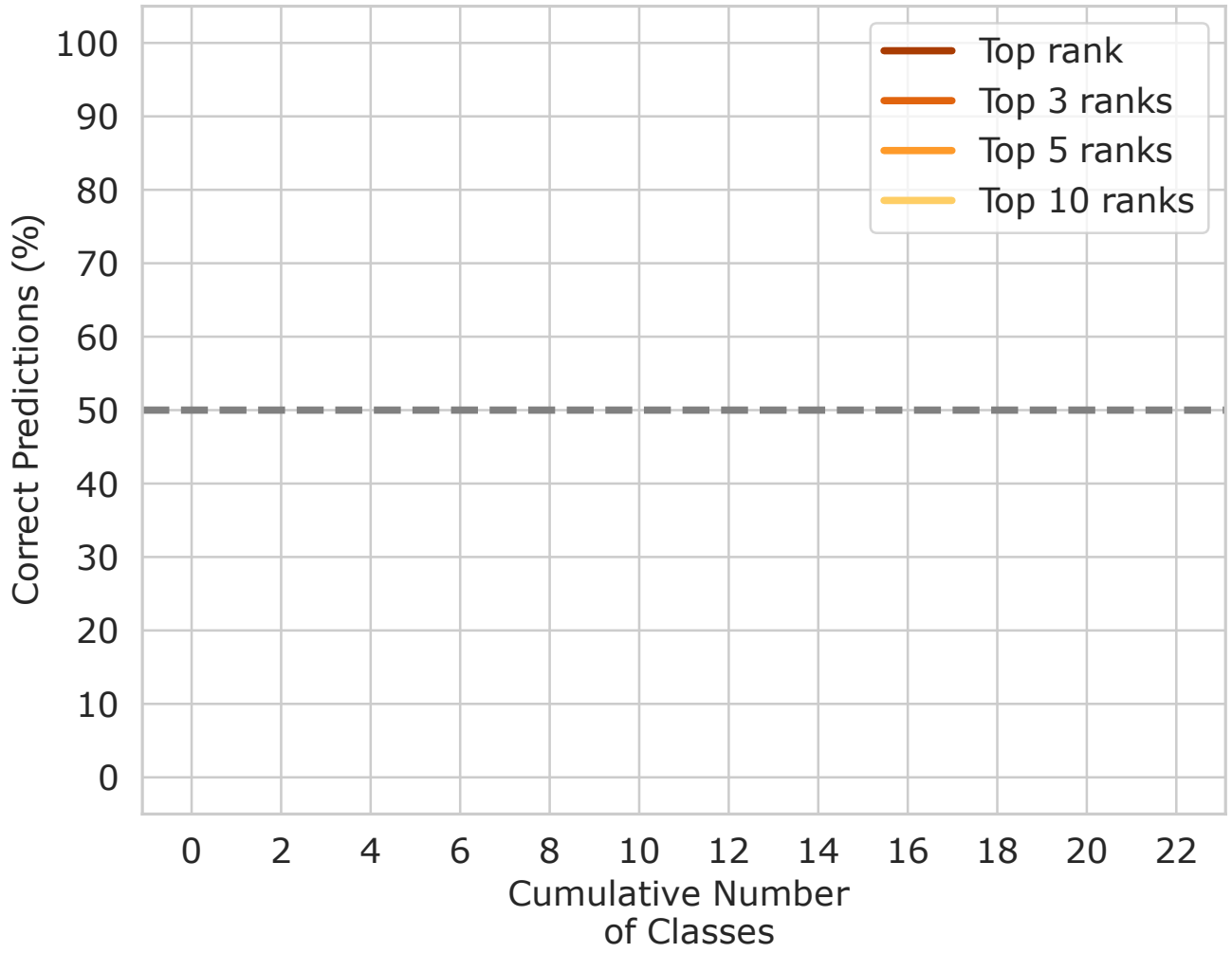
- 23-way, 5-shot few-shot classifier prediction in a **purely evaluative** setting.
- Hosts with **atleast six** samples.
- Four hosts were seen in Non-IV dataset.



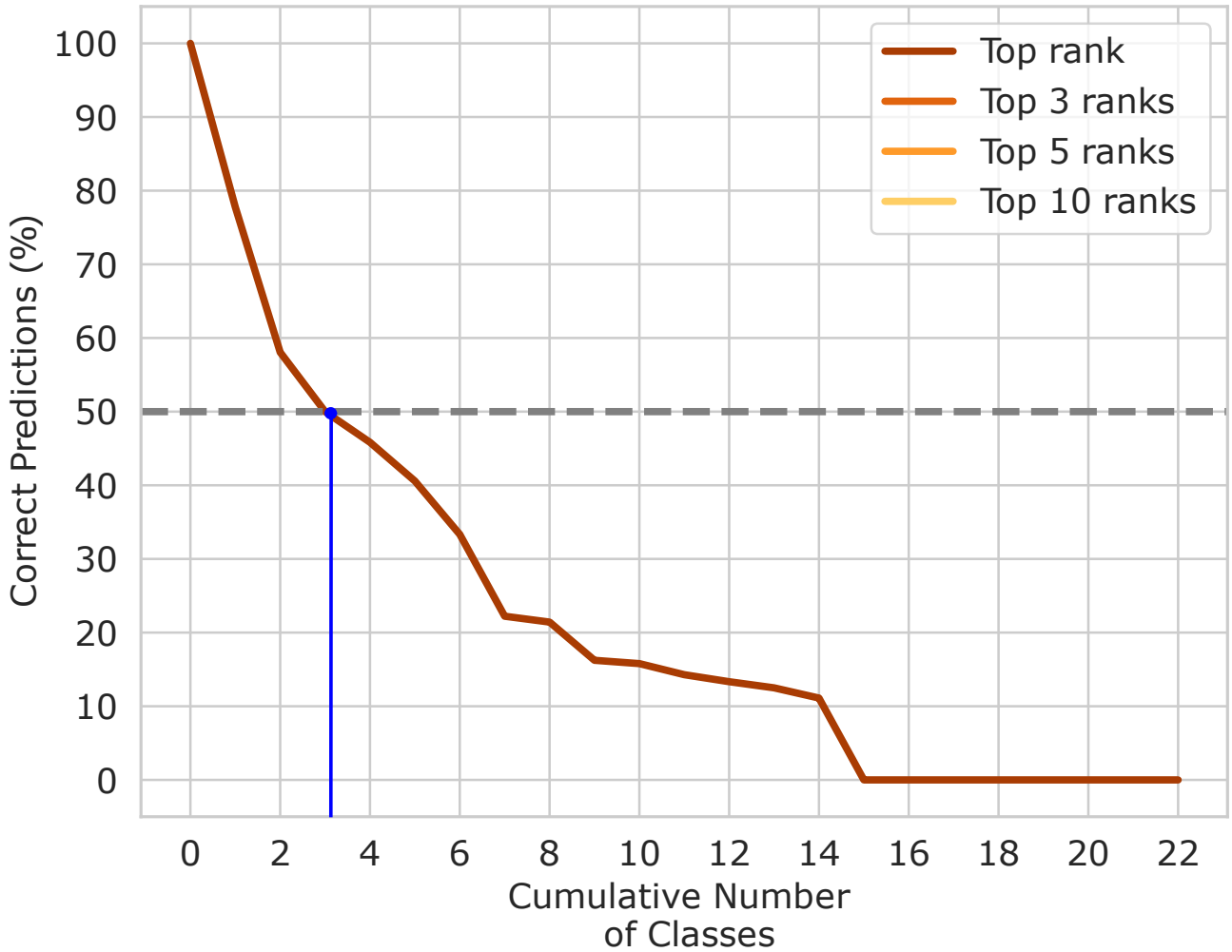
Predicting Hosts of Unseen Viruses



Predicting Hosts of Unseen Viruses

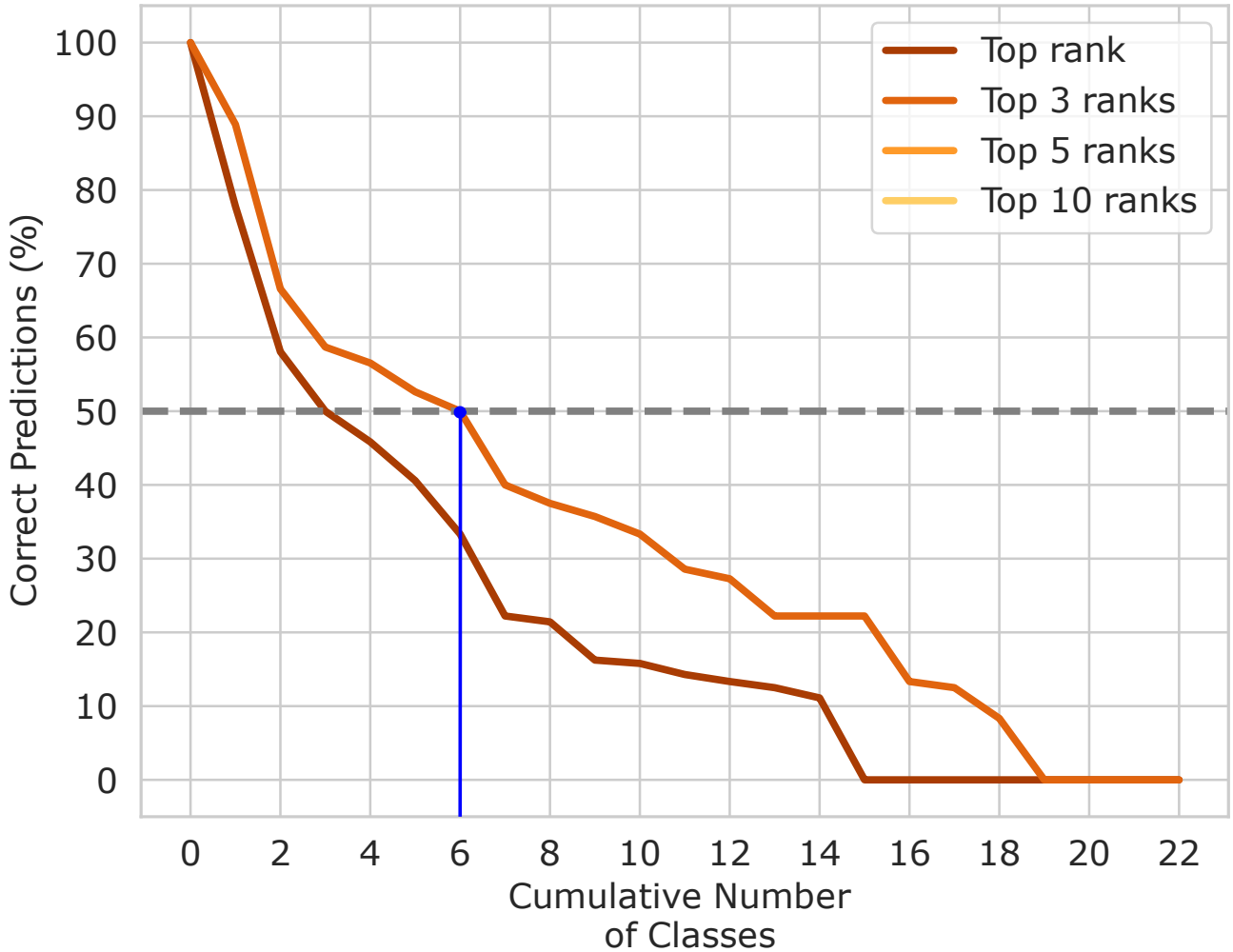


Predicting Hosts of Unseen Viruses



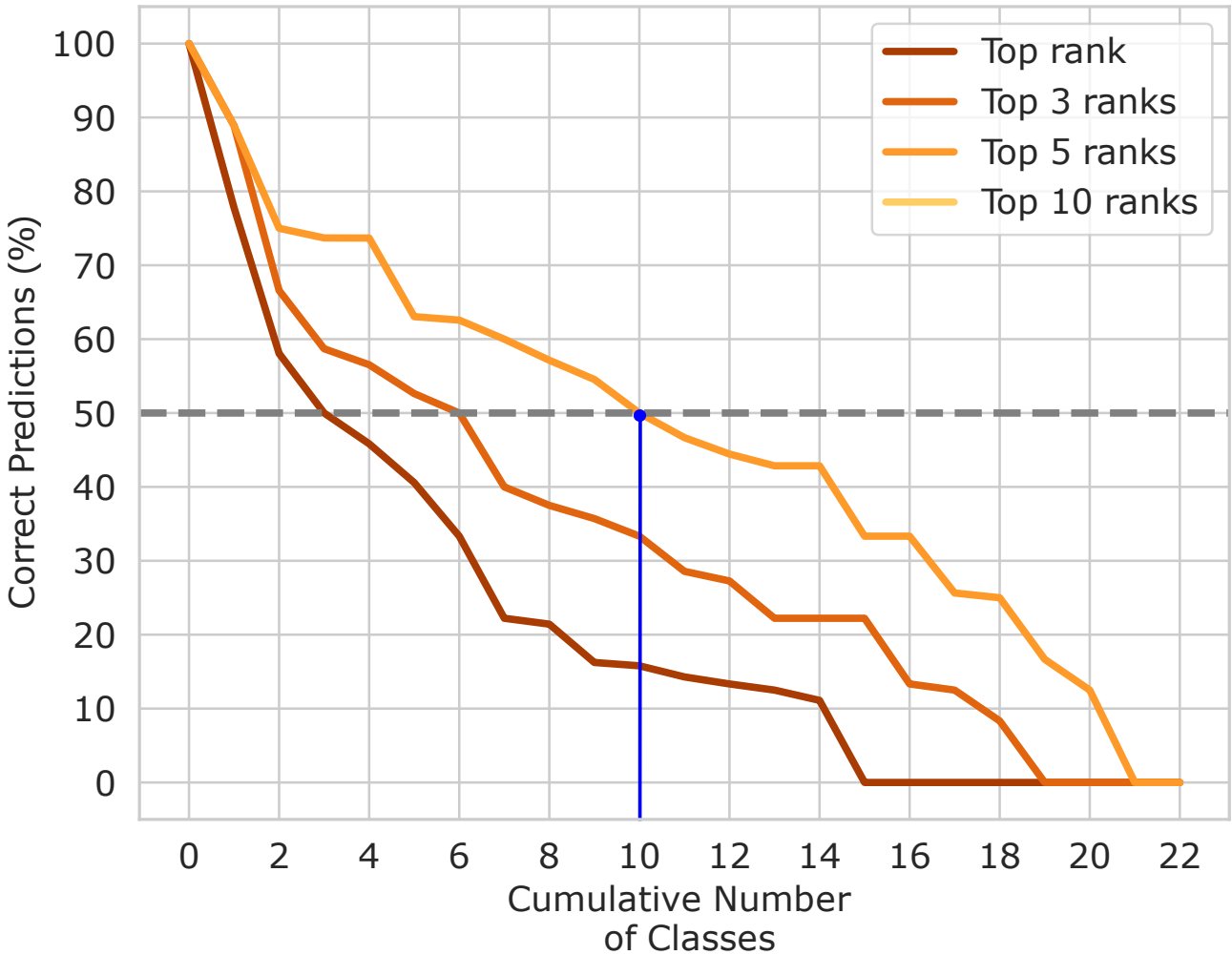
3 classes had atleast 50% of the sequences with correct predictions.

Predicting Hosts of Unseen Viruses



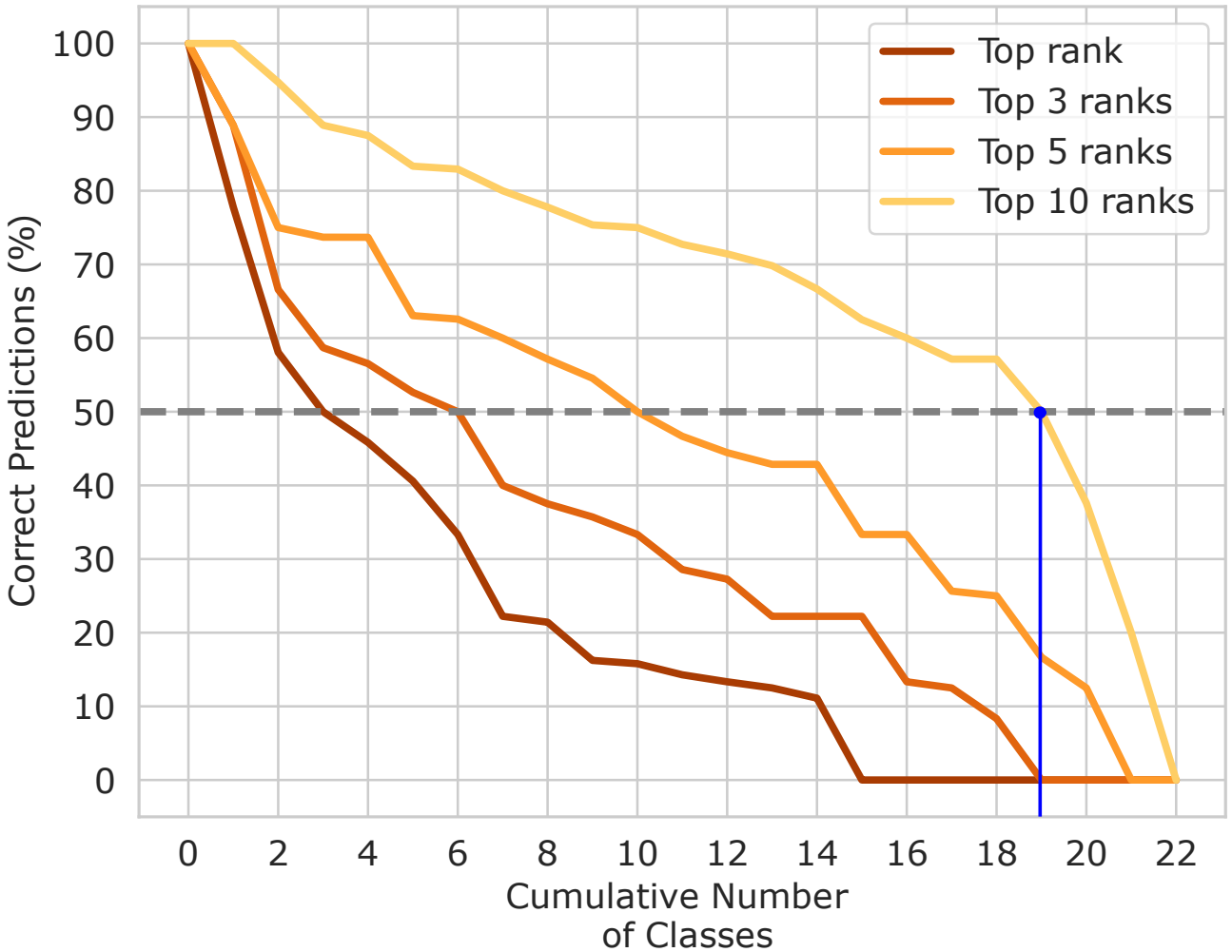
6 classes had true predictions for atleast 50% of the sequences within top 3 ranks.

Predicting Hosts of Unseen Viruses



10 classes had true predictions for atleast 50% of the sequences within top 5 ranks.

Predicting Hosts of Unseen Viruses

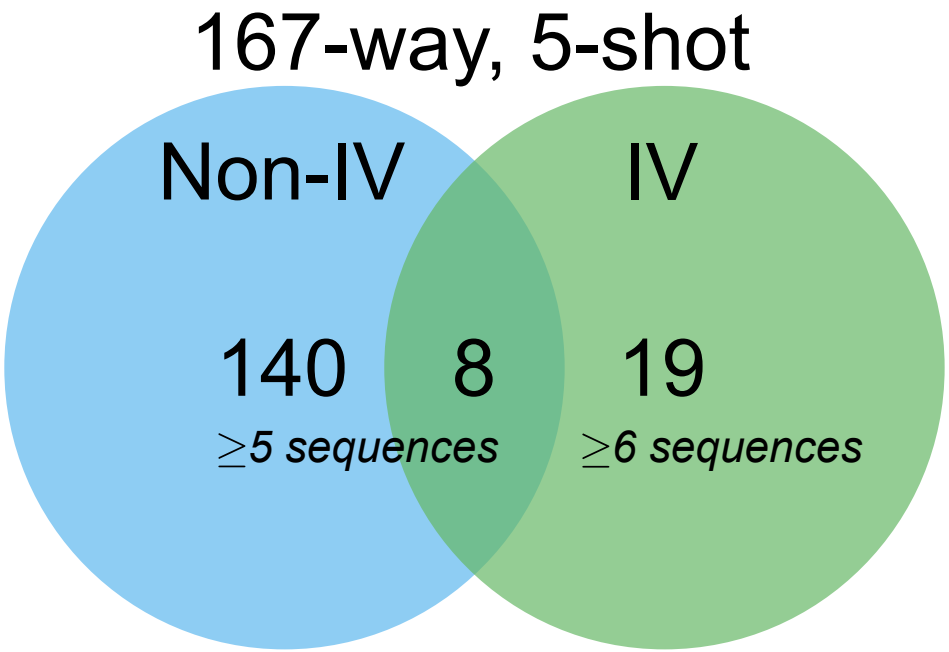


19 classes had true predictions for atleast 50% of the sequences within top 10 ranks.

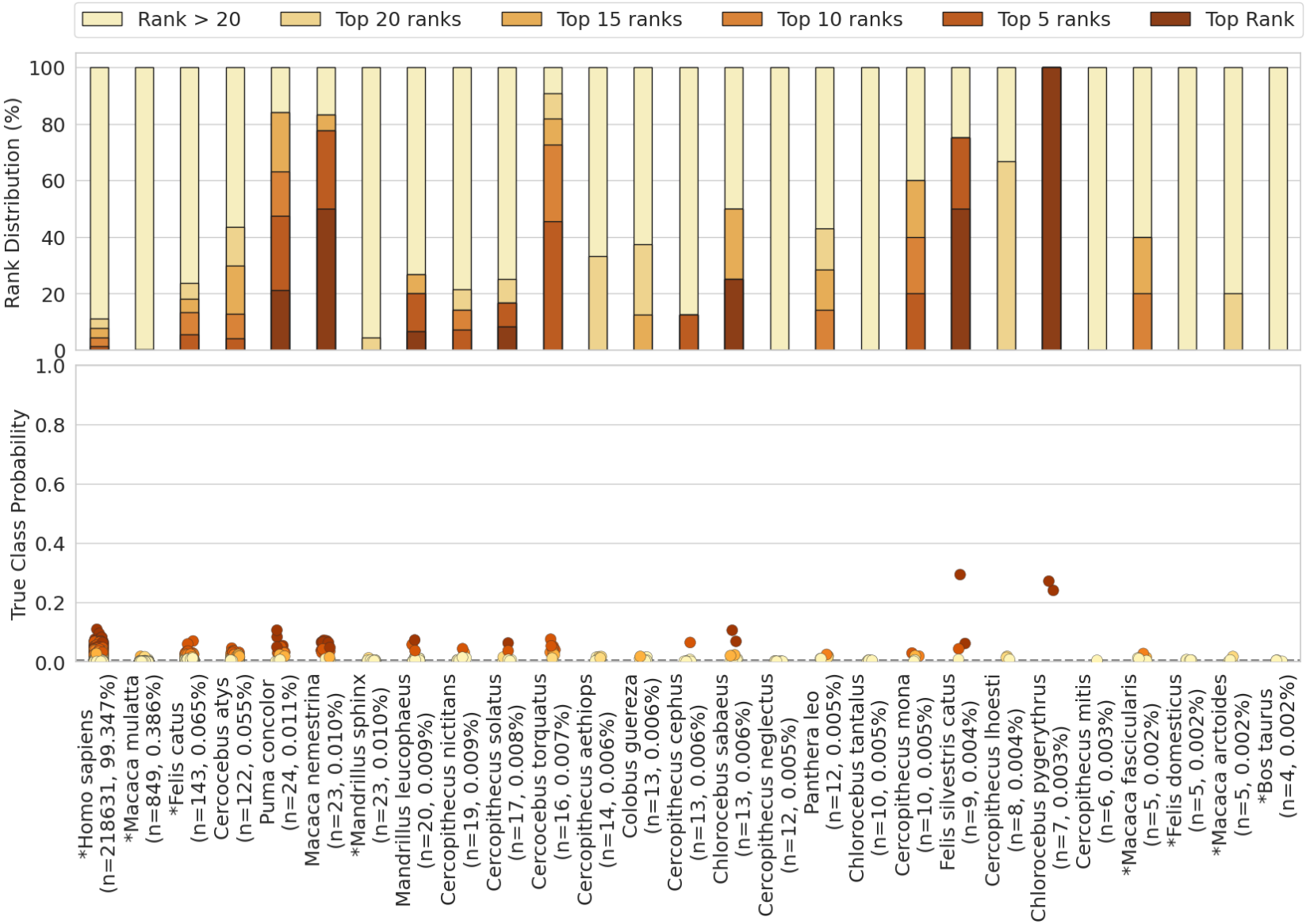
Predicting Hosts of Unseen Viruses

Dataset: Hosts in IV dataset.

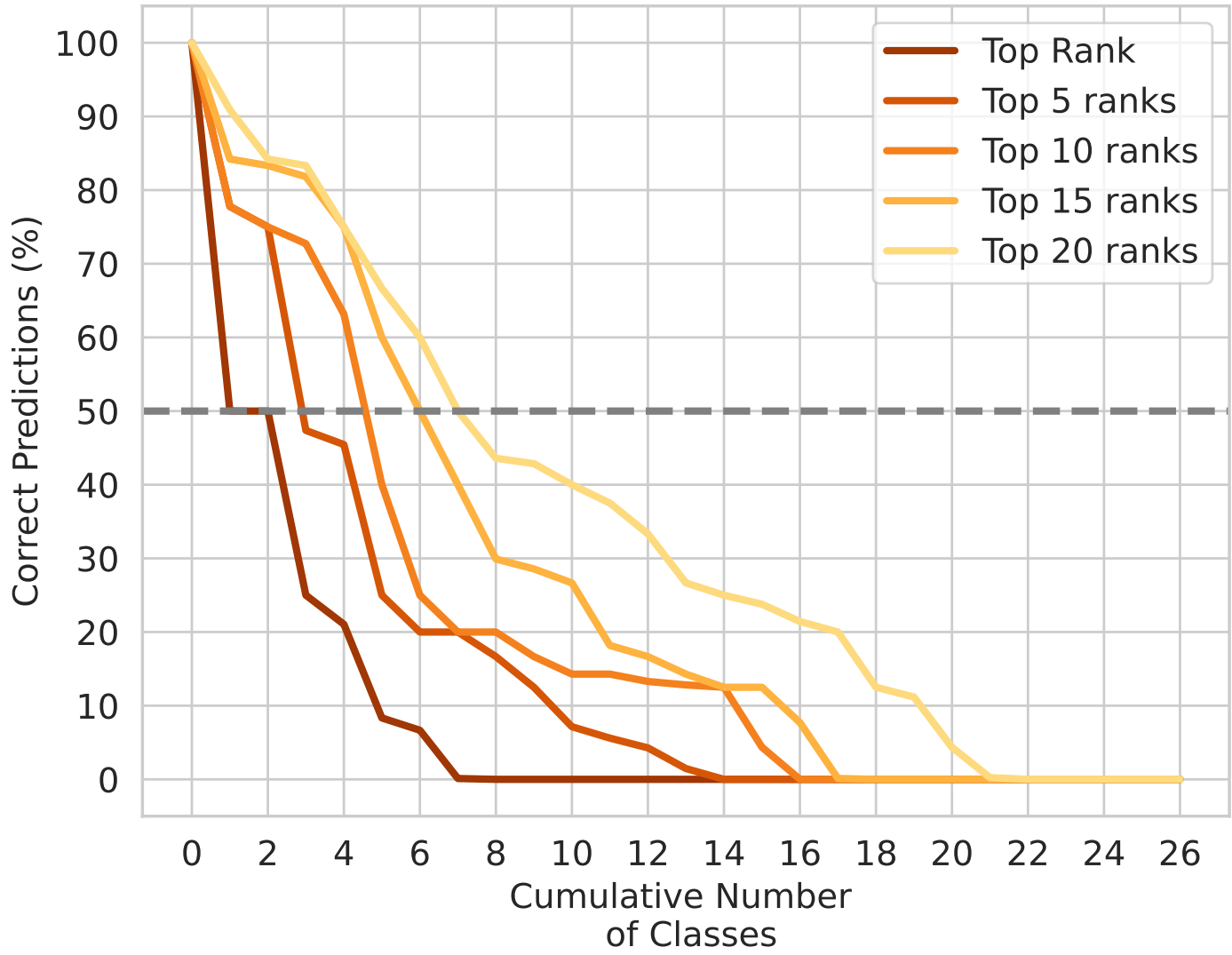
- 167-way, 5-shot few-shot classifier prediction in a **purely evaluative** setting.
- Expanded the set of 23 hosts by adding 148 hosts from the non-IV dataset that had **atleast five** sequences.
- Eight hosts were seen in Non-IV dataset.



Predicting Hosts of Unseen Viruses



Predicting Hosts of Unseen Viruses



Outline

Summary

- Model architecture comprising segmenting the input sequence and hierarchical self-attention facilitated VirProBERT to learn superior embeddings for protein sequences of any length.

Summary

- Model architecture comprising segmenting the input sequence and hierarchical self-attention facilitated VirProBERT to learn superior embeddings for protein sequences of any length.
- Comprehensive pre-training dataset of viral protein sequences was highly effective.

Summary

- Model architecture comprising segmenting the input sequence and hierarchical self-attention facilitated VirProBERT to learn superior embeddings for protein sequences of any length.
- Comprehensive pre-training dataset of viral protein sequences was highly effective.
- VirProBERT achieved prediction performance for common classes that was on par with SOTA pLMs despite its minimal model size.

Summary

- Model architecture comprising segmenting the input sequence and hierarchical self-attention facilitated VirProBERT to learn superior embeddings for protein sequences of any length.
- Comprehensive pre-training dataset of viral protein sequences was highly effective.
- VirProBERT achieved prediction performance for common classes that was on par with SOTA pLMs despite its minimal model size.
- VirProBERT can generalize to rare, unseen hosts and unseen viruses.

Summary

- Model architecture comprising segmenting the input sequence and hierarchical self-attention facilitated VirProBERT to learn superior embeddings for protein sequences of any length.
- Comprehensive pre-training dataset of viral protein sequences was highly effective.
- VirProBERT achieved prediction performance for common classes that was on par with SOTA pLMs despite its minimal model size.
- VirProBERT can generalize to rare, unseen hosts and unseen viruses.
- Generality of FSL framework permits the addition of a new host (with few labeled sequences) to the model without any explicit fine-tuning.

References

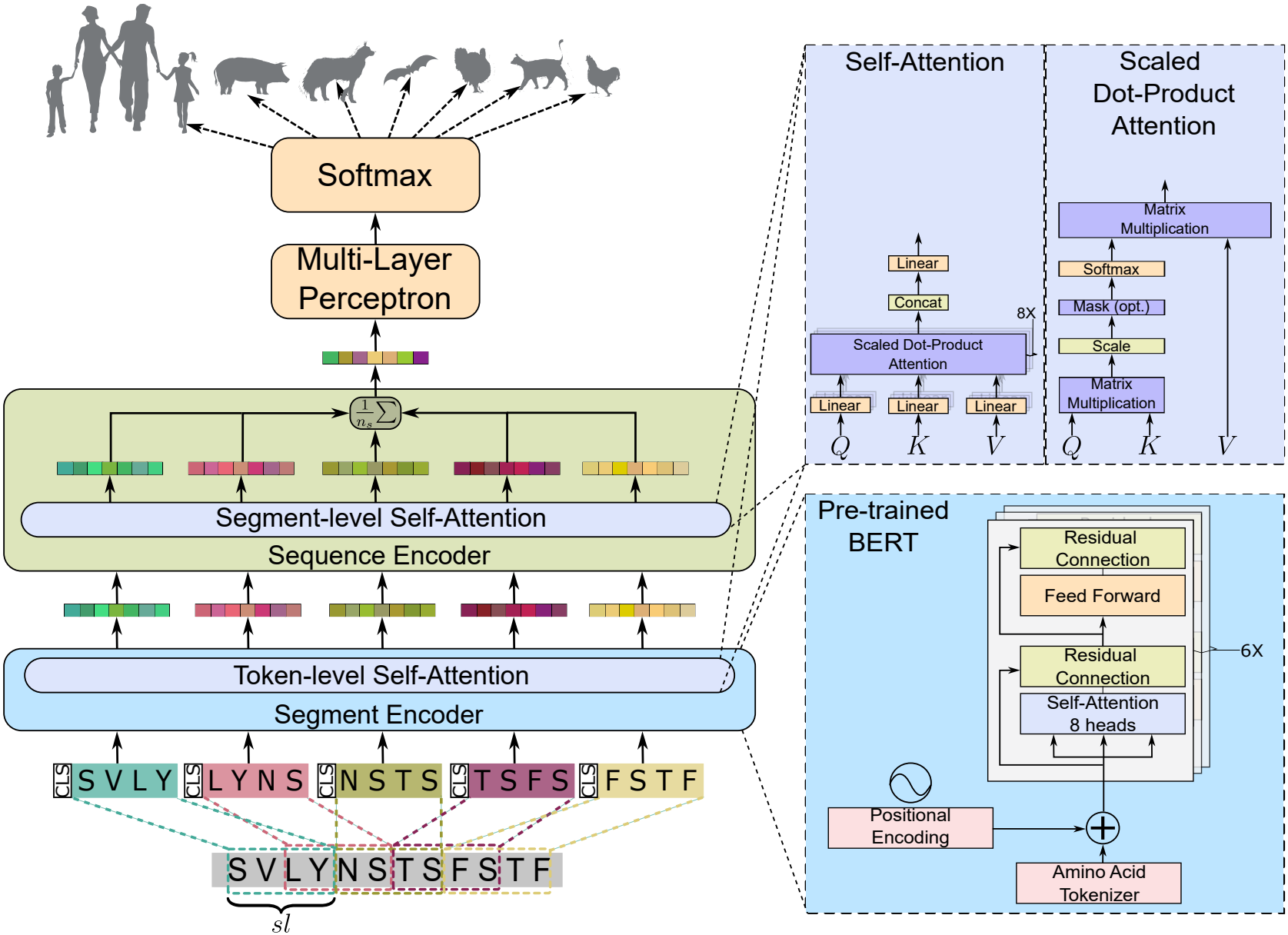
Wang Liu-Wei, Şenay Kafkas, Jun Chen, Nicholas J. Dimonaco, Jesper Tegnér, and Robert Hoehndorf.
DeepViral: Prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes.
Bioinformatics, 37(17):2722–2729, September 2021.

Katie K. Tseng, Heather Koehler, Daniel J. Becker, Rory Gibb, Colin J. Carlson, Maria del Pilar Fernandez, and Stephanie N. Seifert.
Viral genomic features predict orthopoxvirus reservoir hosts, October 2023.

Outline

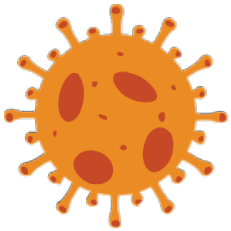
Ideas for Course Projects

1. Input: Split data to avoid leakage:
   - All proteins from a genome sequence, virus, or family are either in the training set or in the test set.
   - Will require tracing the genome a protein comes from, organizing proteins by genome, and implementing a data split.
2. Pre-training: use Transformer decoder or T5 models.
3. Baseline models:
   - Currently, the pipeline converts each sequence to a vector of 3-mer frequencies as input to LR, SVM, and RF.
   - Replace these vectors with pre-trained embeddings in VirProBERT.
4. Output: Implement a hierarchical classifier
   - Predict the complete taxonomy of a host, e.g, *Homo sapiens*.
   - How will we measure accuracy?
5. Output: Predict an "unknown" label.
   - For each host in our collection, train a one-class or binary classifier. Prediction of "No" for every host $\equiv$ overall prediction is "unknown".
   - Evaluate on existing multi-class dataset and separately on plant/animal/microbial proteins.

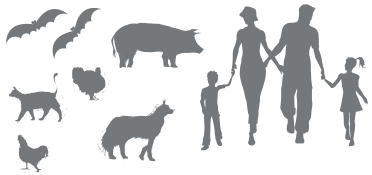

Softmax
Multi-Layer Perceptron
Segment-level Self-Attention
Sequence Encoder
Token-level Self-Attention
Segment Encoder
CLS S V L Y
CLS L Y N S
CLS N S T S
CLS T S F S
CLS F S T F
S V L Y N S T S F S T F
sl
Self-Attention
Linear
Concat
8X
Scaled Dot-Product Attention
Linear
Linear
Linear
Q
K
V
Scaled Dot-Product Attention
Matrix Multiplication
Softmax
Mask (opt.)
Scale
Matrix Multiplication
Q
K
V
Pre-trained BERT
Residual Connection
Feed Forward
Residual Connection
Self-Attention 8 heads
6X
Positional Encoding
Amino Acid Tokenizer

UniRef90 → Viruses → Vertebrates → Final Datasets

Protein sequences with atleast 90% sequence similarity



Retain sequences from clusters of viral proteins



Retain clusters where the host is a vertebrate

Common Hosts	Prevalence
Human	60.23%
Pig	2.69%
Capybara	1.31%
Himalayan marmot	1.13%
Red junglefowl	1.01%

Pre-training Dataset

1.2 million sequences

267,865 sequences

3,779 viruses (~82% Immuno-deficiency virus)

1,314 virus hosts (~92% Humans)

Hosts with atleast 1% prevalence

Hosts with prevalence < 1%

Rare Hosts

16,074 sequences
143 virus hosts

Immunodeficiency virus (IDV)

220,068 sequences
7 viruses
40 virus hosts

Non-Immunodeficiency virus (Non-IDV)

47,792 sequences
3,772 viruses
1,304 virus hosts