Using deep learning to illuminate viral dark matter:

algorithms for virus identification and analysis

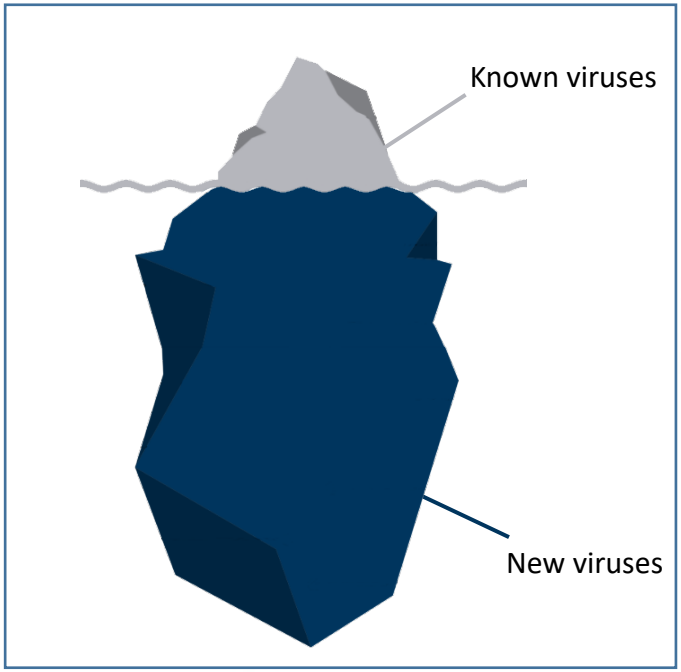
Dr. Yanni Sun
Electrical Engineering, City University of Hong Kong

How to give an interdisciplinary talk: assume zero knowledge but infinite intelligence of the audience. - From Gary Stormo's talk

Viral dark matter

The most diverse and abundant organisms on Earth

Clinical importance (many pathogens);
Important components in natural environments.



Research questions

Who are there?

- Composition analysis in a broad spectrum of samples: human, animal, soil, water etc.
- E.g. detecting all viruses in a patient's throat samples

What do they do?

- Properties and functions of the viruses

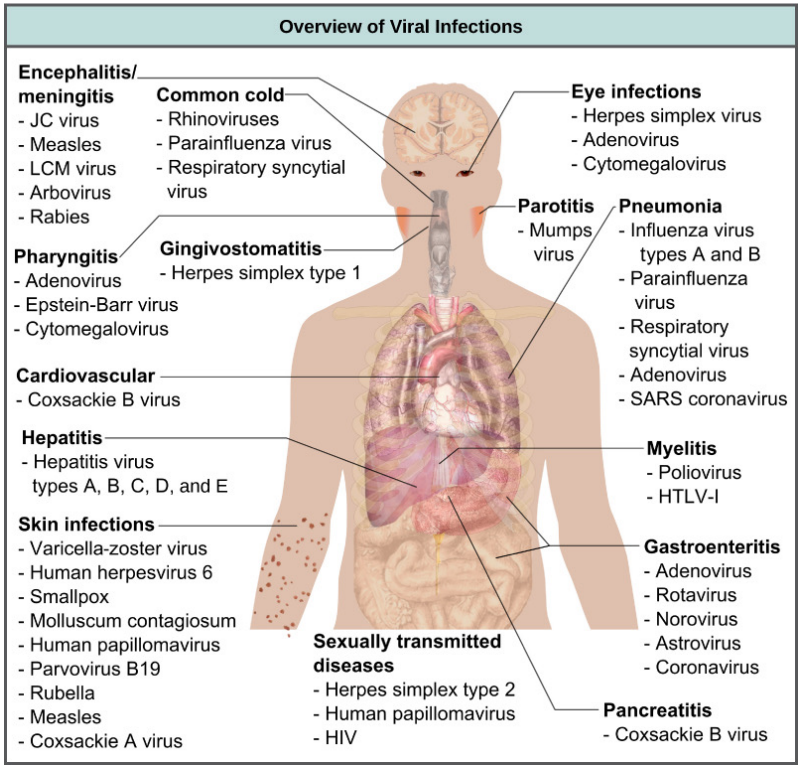
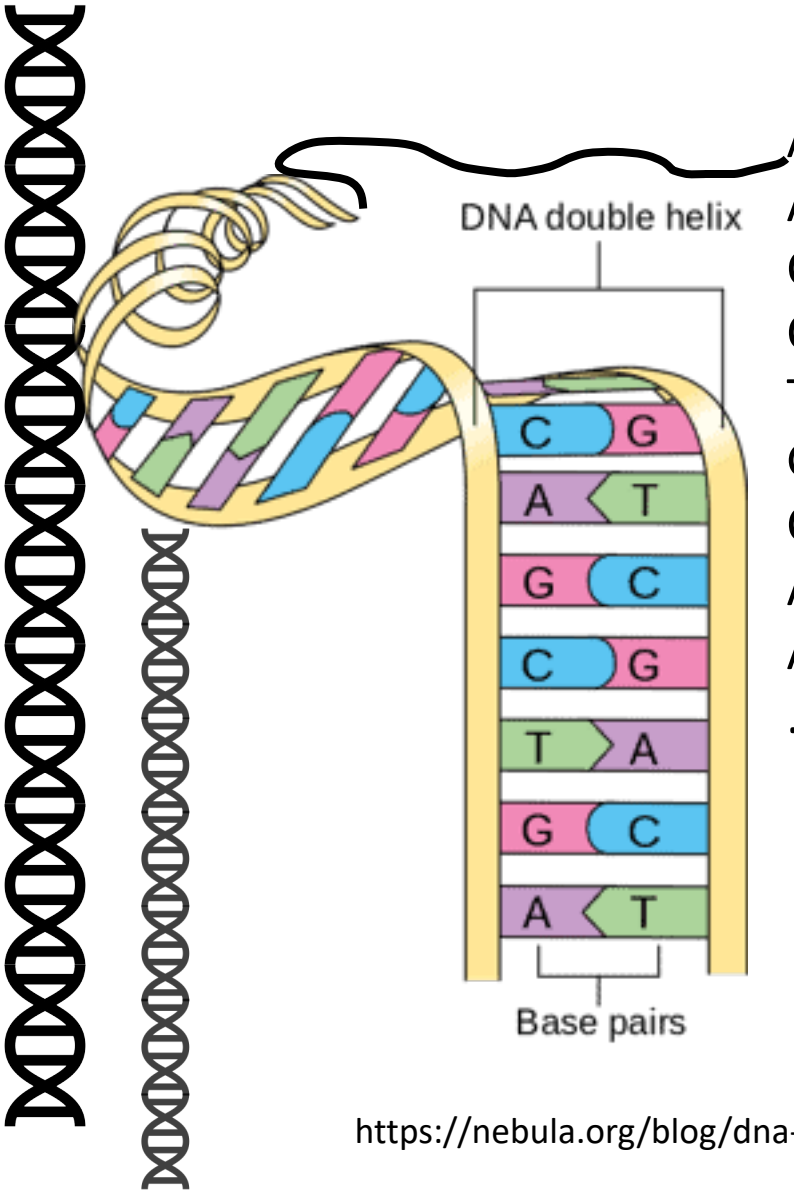


Image from internet



ATGGGCAAGTCAGAAAGTCAGATGG
ATATAACTGATATCAACACTCCAAAG
CCAAAGAAGAAACAGCGATGGACT
CCACTGGAGATCAGCCTCTCGGTCCT
TGTCCTGCTCCTCACCATCATAGCTGT
GACAATGATCGACTCTATGCAACCTA
CGATGATGGTATTTGCAAGTCATCAG
ACTGCATAAAATCAGCTGCTCGACTG
ATCCAAAACATGGATGCCACCACTG
… …

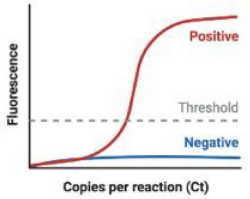
Language of Life

- Contains all functions that make each biological entity unique

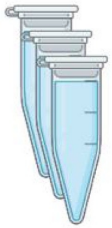
https://nebula.org/blog/dna-structure-model/

Next-generation sequencing technology

Nasopharyngeal swab collected from patients.
Extracted RNA with Direct-zol RNA Miniprep kit



24 positive samples were selected for the downstream experiment



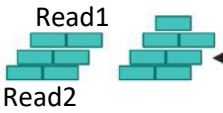
cDNA generated with RT-PCR using the VILO™ Superscript™ III kit

cDNA for each sample mixed with 11 genotyping primers

Multiplex PCR with annealing temperature set at 65 °C

Library for each sample was prepared with ARTIC V3 PCR tiling protocol

8 pmol of the library with 20% PhiX loaded onto the Miseq machine using V3 600 Cycle kit



Sequencing data (sequencing reads)

Bioinformatics analysis

report

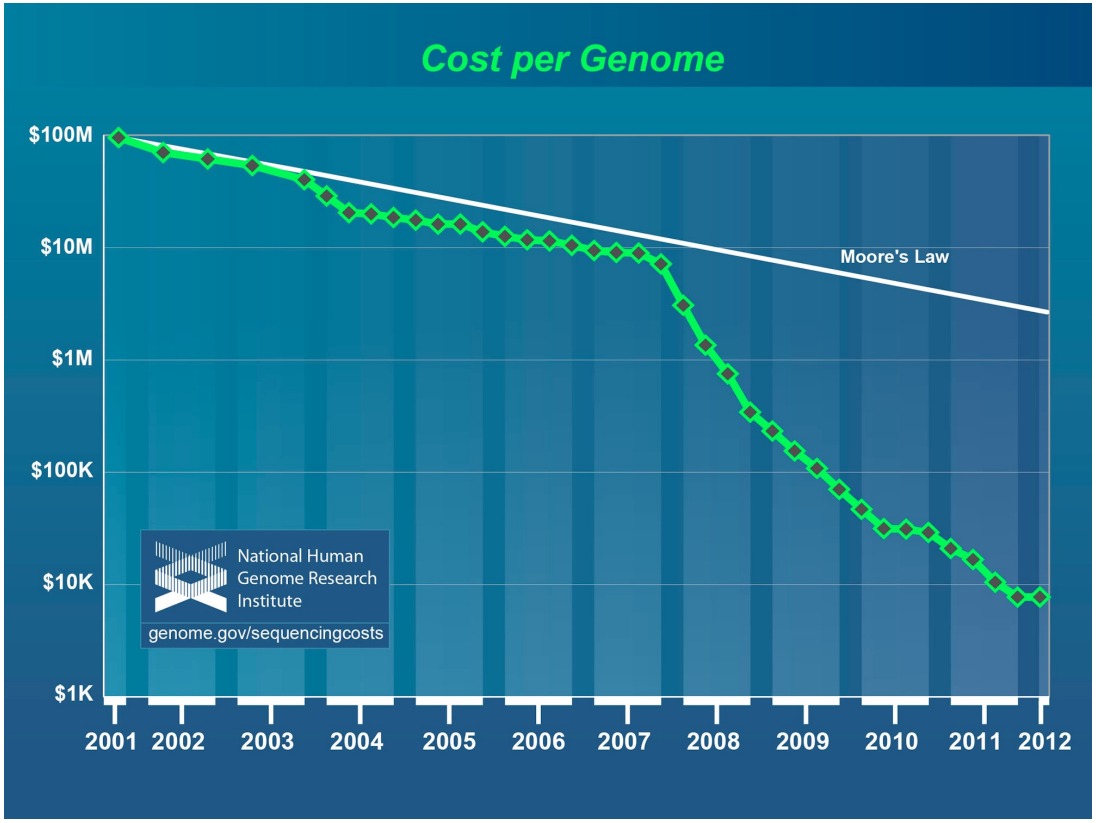
Read1

Read2

```
1) Read Name → @SRR190851.108390742/1
2) Sequence Bases → GAGATTGAGTCTTGCTTTGTCCCCAGGCTGGAGTGCAATGG
3) '+' → +
4) Base Qualities → ;@@@A;>5?B@DABBFA@=EE@E@FEFFHF=BECEFFED>F
1) Read Name → @SRR190851.61391872/1
2) Sequence Bases → CAACATGGTGAAACCCCGTCTCTACTAACATACAAAATTAG
3) '+' → +
4) Base Qualities → CBEBEFIIEIGDJHIJJ?GGHGKFGJEIGGIIIIKKKEIIK
1) Read Name → @SRR190851.22176085/1
2) Sequence Bases → TAGACTGAGGCCTAAGTCTCAGTCTGGGGCCTGGTACATGG
3) '+' → +
4) Base Qualities → @@?CCHECAEBEGDEHFDHEHGFGHB>GFAEHBEE;EGGI>
```

Fastq file (sequencing data/reads)

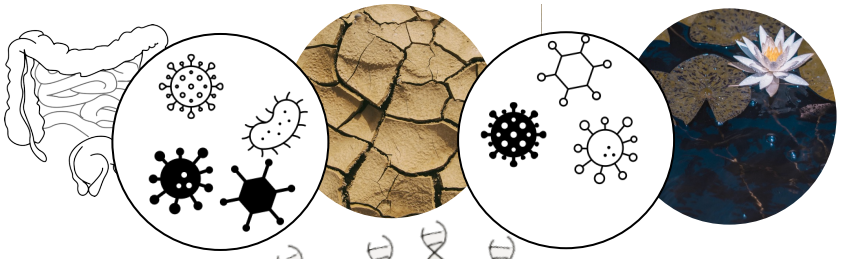
BIG genomic data



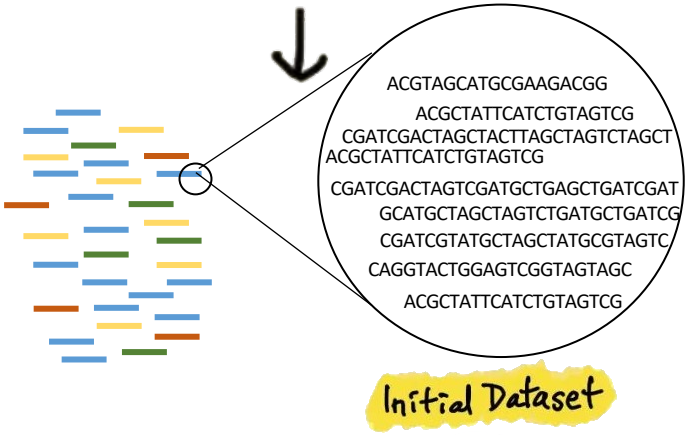
Low cost → fast accumulation of sequencing data

Sequence Read Archive at NIH: 11,141,607,428,443,304 bases (~11,141 terabases)

Research problem formulation: virus identification



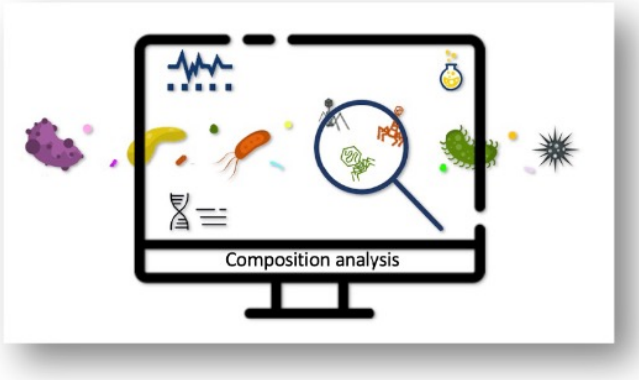
Sequencing machine



Initial Dataset

DNA sequences (**contigs**) obtained using sequencing data analysis

Binary classification



virus identification

Material source: viralzone

Methodology: towards more sensitive and accurate virus identification using

Transformer

Decipher the language of life using AI

Natural language processing

John wanted to go to the coffee shop in Mong Kok.

Tokenization

[*start*] [J] [o] [h] [n] [*space*] [w] [a] [n] [t] ···

[*start*] [John] [wanted] [to] [go] [to] [the] ···

[*start*] [John] [wanted to] [go to] [the] ···

ChatGPT

Task

Sentiment analysis

Language translation

Part-of-speech tagging

Biological sequence

··· TCGTAGTAGTCGTAGTCGATGTCAGTGTCATGT ···

Tokenization

[*start*] [T] [C] [G] [T] [A] [G] [T] [A] [G] [T] ···

[*start*] [TCG] [TAG] [TAG] [TCG] [TAG] [TCG] ···

[*start*] [TCG] [TAGTA] [GTCG] [TAGTCG] ···

Design deep learning models

AI

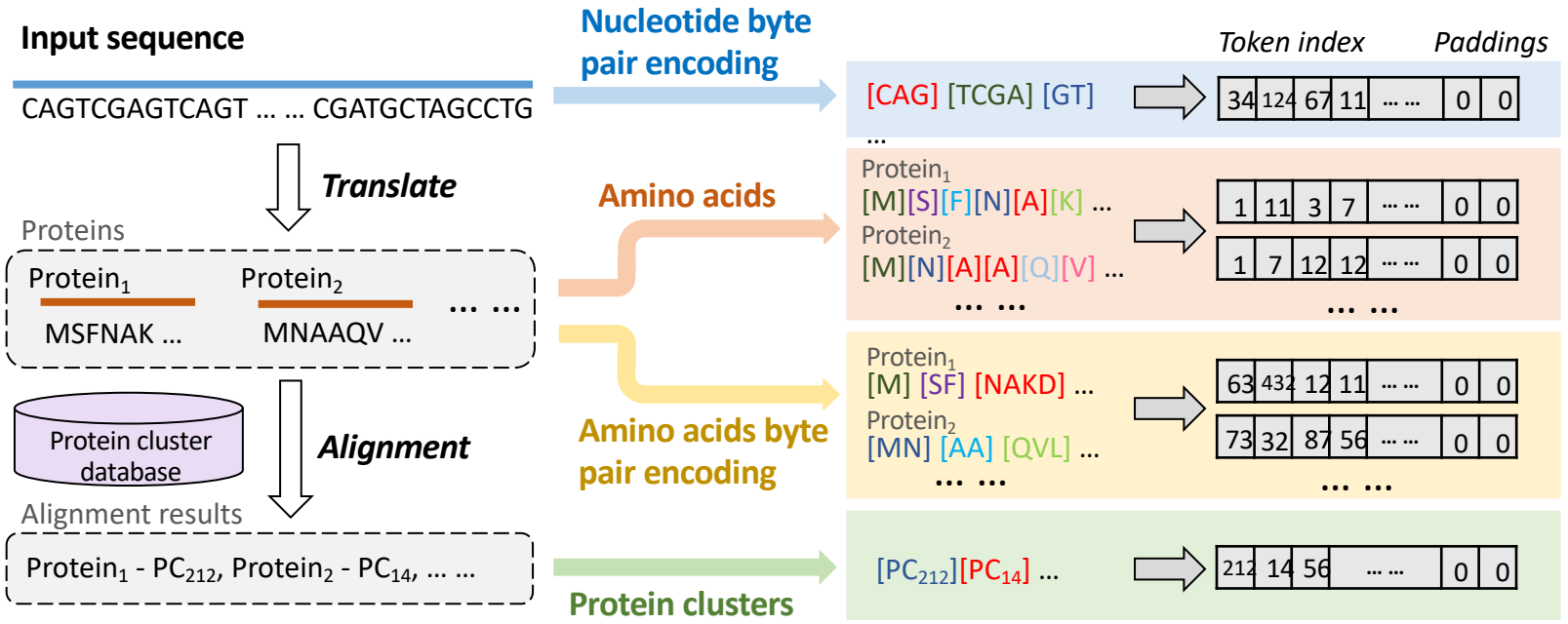
Task

Function prediction

Gene annotation

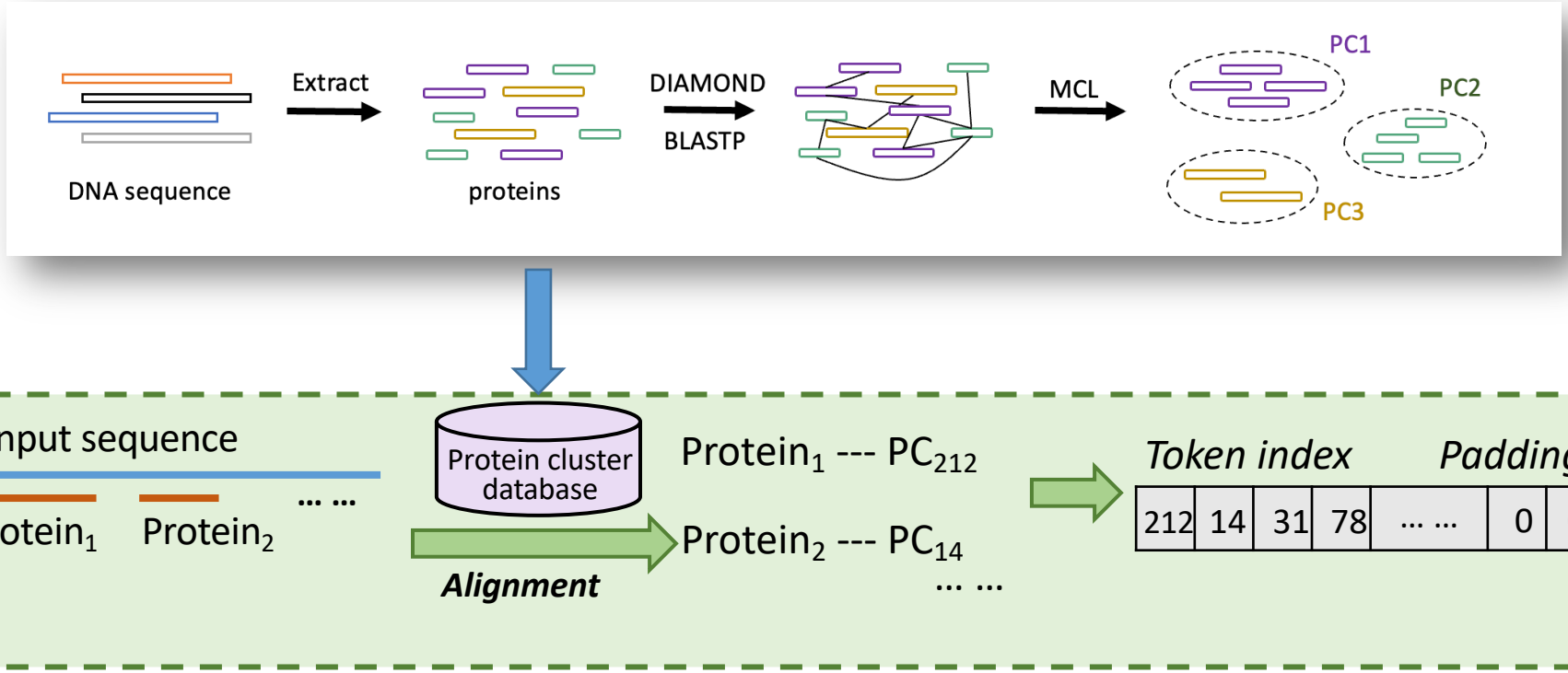
Protein structure prediction

Token (word) construction



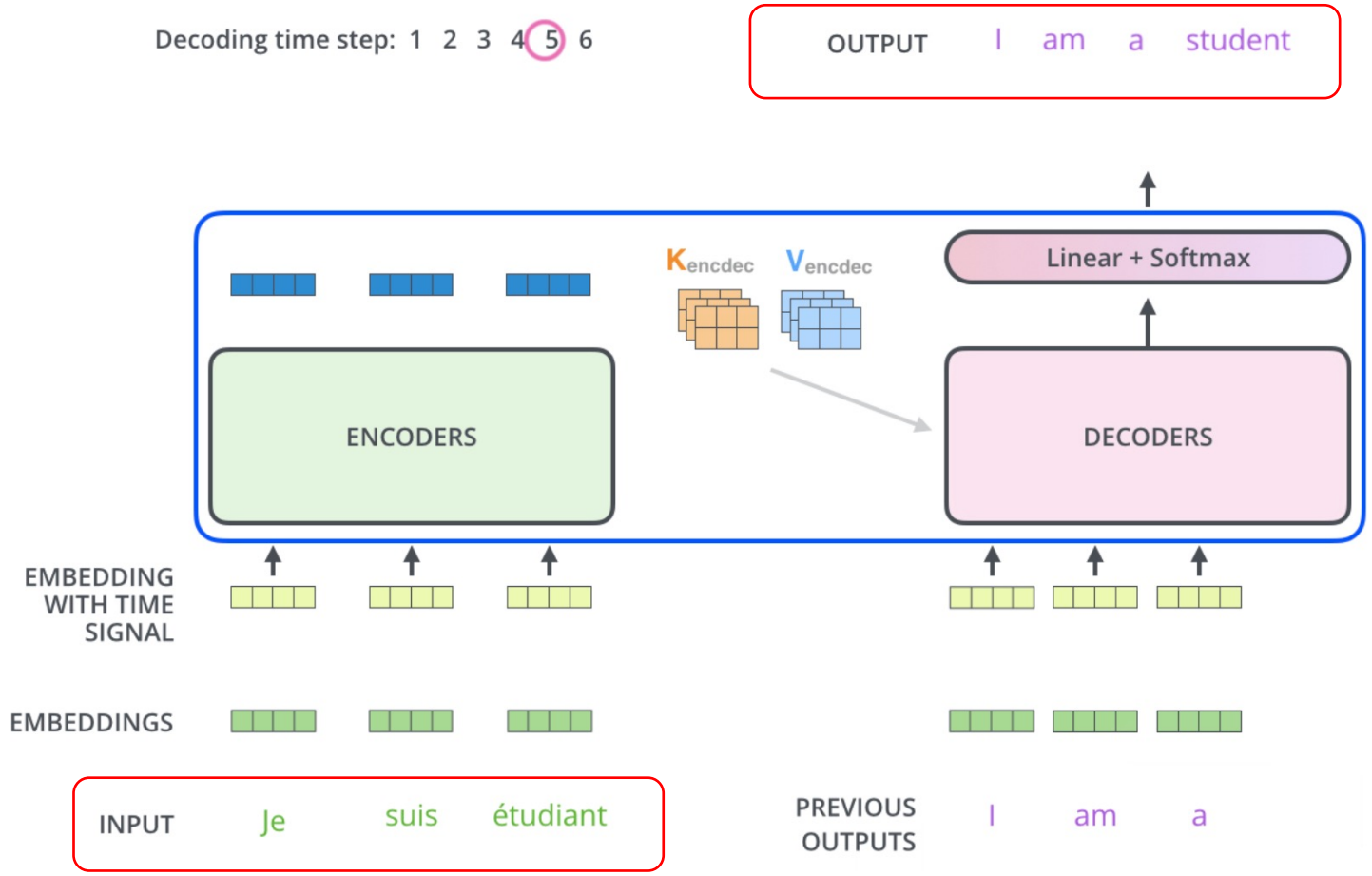
Byte pair encoding (BPE): count the most frequent nucleotide/ amino acid combinations in the corpus.

Protein cluster (PC) tokenizer



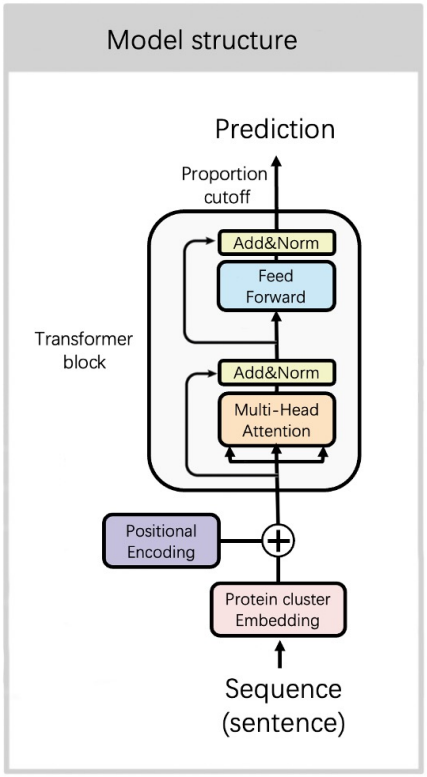
- Learn the importance of proteins
- Capture the correlation between different proteins

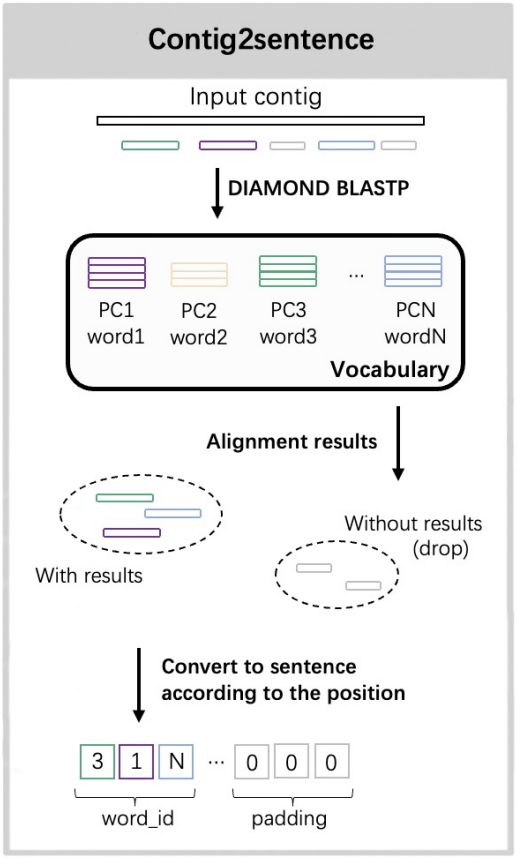
Transformer for translation



https://jalammar.github.io/illustrated-transformer/

Virus Identification with Transformer





Multi-head attention:

- Learn the meaning of the word
- Learn the correlation between word

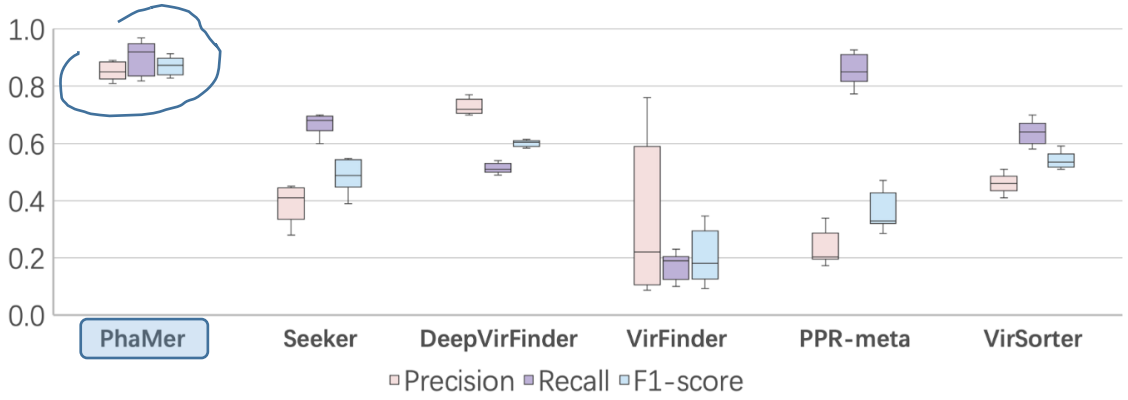
- Learn the importance of the PCs
- Learn the association between proteins

Virus Identification – Experimental Results

Dataset

- Using viruses as positive sample and their host as negative samples
  - May share common regions
- Testing on several independent datasets:
  - Low-similarity data
  - Mock dataset
  - IMG/VR database

Results

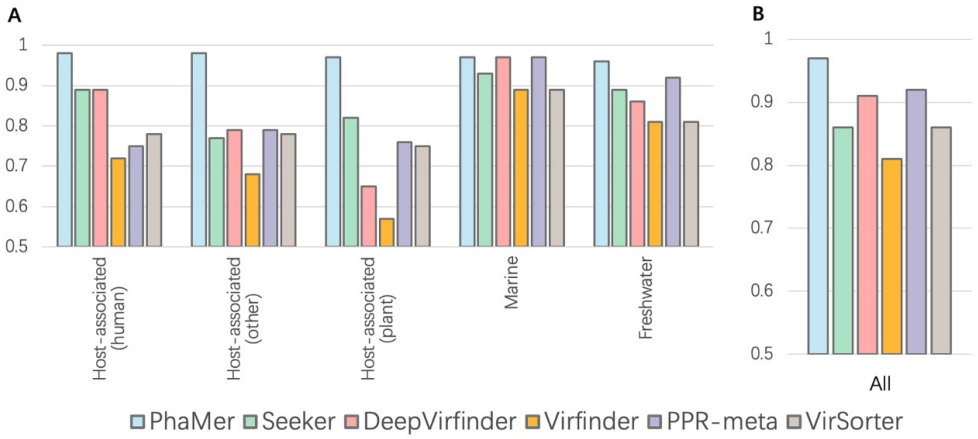


Mock dataset (European Nucleotide Archive PRJEB19901, ~30 species/strains)

Virus Identification – Experimental Results

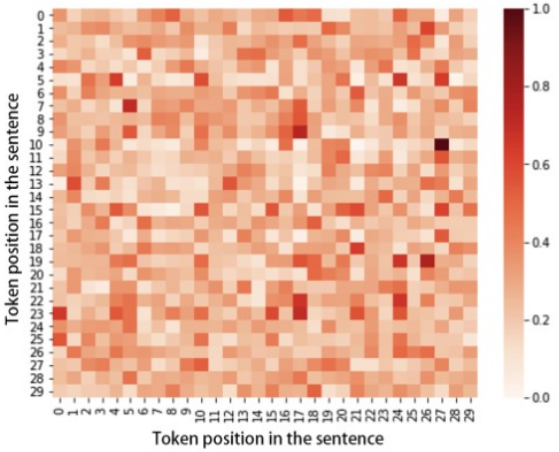
Results

Y-axis: recall



IMG/VR database

Visualization

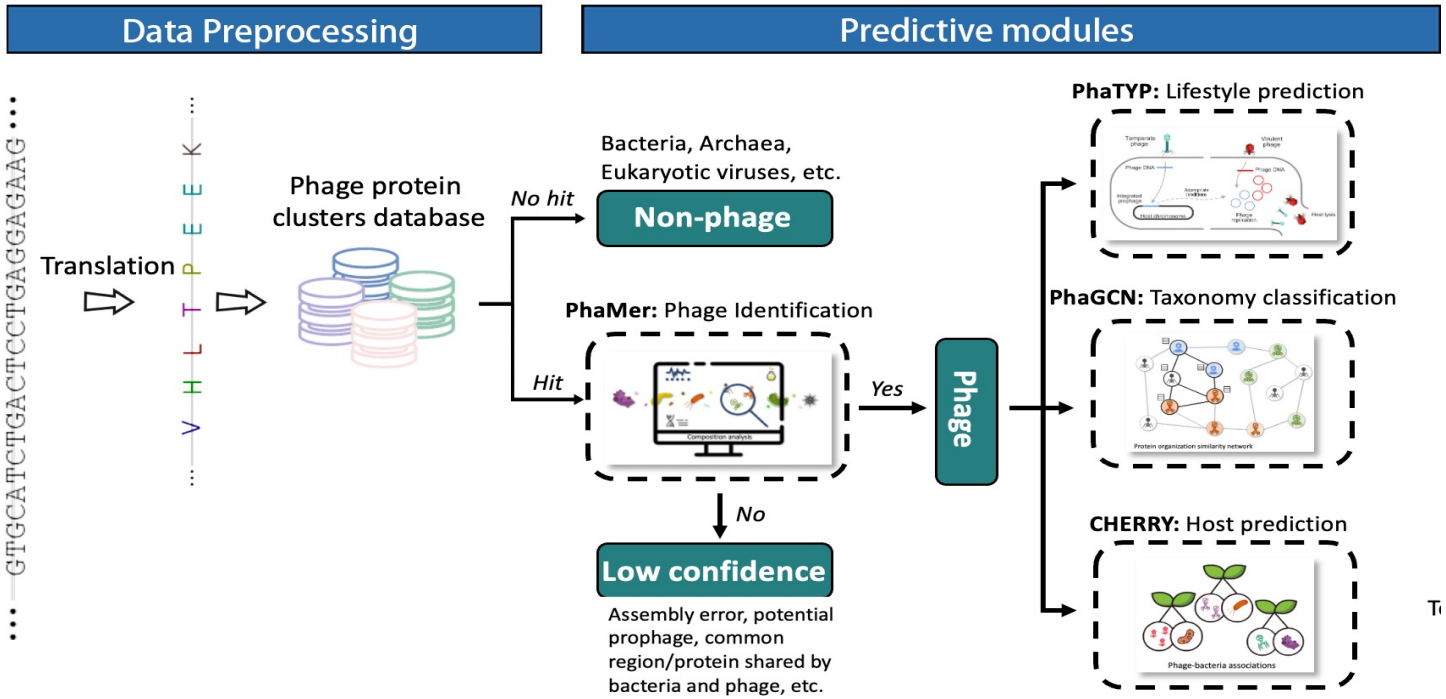


Attention score matrix in Transformer

- Attention score matrix in Transformer

  The high score are the PCs contains viral structural proteins (tail fiber/baseplate/…)

PhaBOX

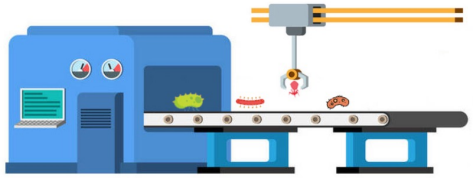


https://phage.ee.cityu.edu.hk/

Our web server

Our tools for phage sequence analysis

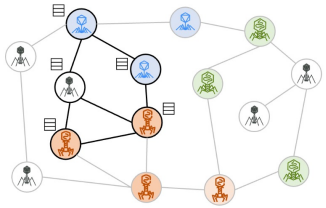
PhaMer



PhaMer is a python library for identifying bacteriophages from metagenomic data. PhaMer is based on a Transorfer model and rely on protein-based vocabulary to convert DNA sequences into sentences.

https://github.com/KennthShang/PhaMer

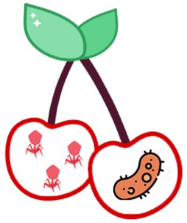
PhaGCN



PhaGCN is a GCN based model, which can learn the species masking feature via deep learning classifier, for new Phage taxonomy classification. To use PhaGCN, you only need to input your contigs to the program.

https://github.com/KennthShang/PhaGCN

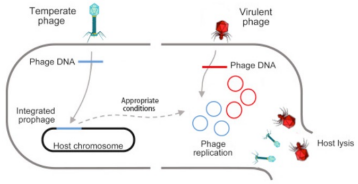
CHERRY



CHERRY is a python library for predicting the interactions between viral and prokaryotic genomes. CHERRY is based on a deep learning model, which consists of a graph convolutional encoder and a link prediction decoder.

https://github.com/KennthShang/CHERRY

Phage TYPE



PhaTYP is a python library for bacteriophages' lifestyle prediction. PhaTYP is a BERT-based model and rely on protein-based vocabulary to convert DNA sequences into sentences for prediction.

https://github.com/KennthShang/PhaTYP

17

Acknowledgement

Funding: HKIDS, GRF, ITF, and City University of Hong Kong

Hong Kong Science Museum

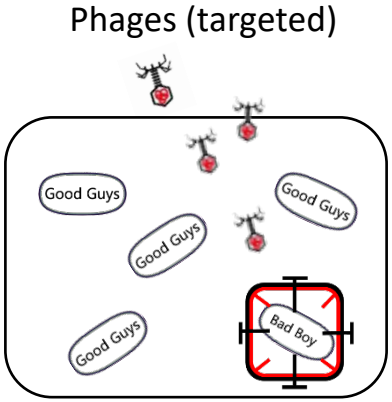
Questions?

For more information, please visit the lab website:
https://yannisun.github.io/

Host Prediction using GCN

Problem formulation

- Given a phage sequence, identify its bacterial hosts
  - Hosts' strains, species, genus, family etc.



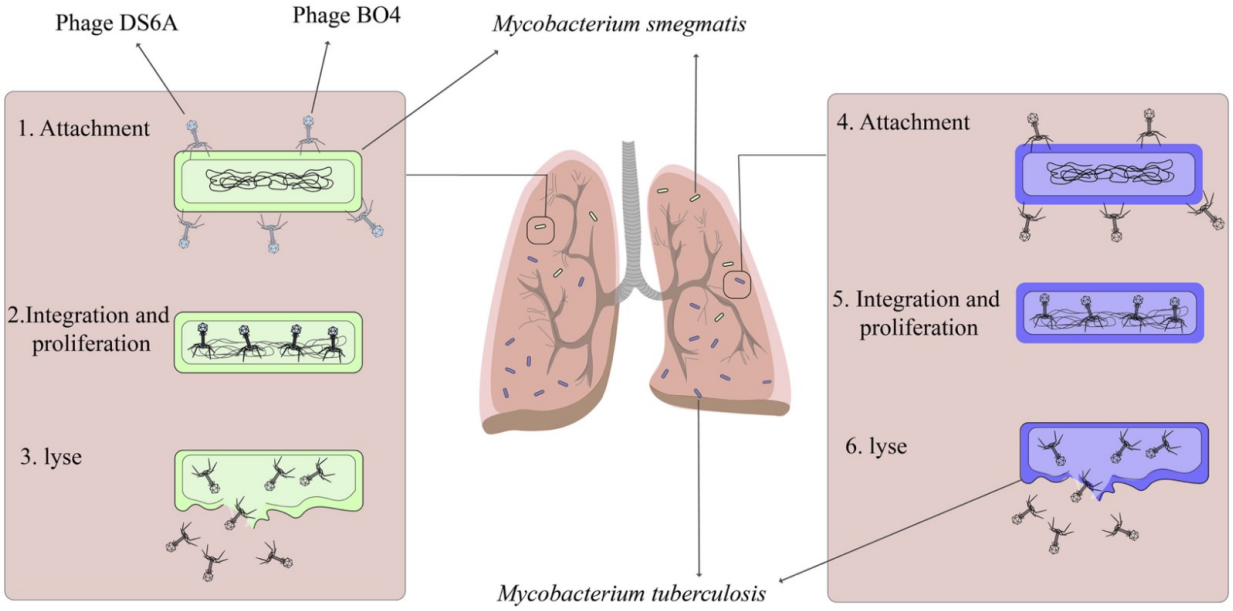
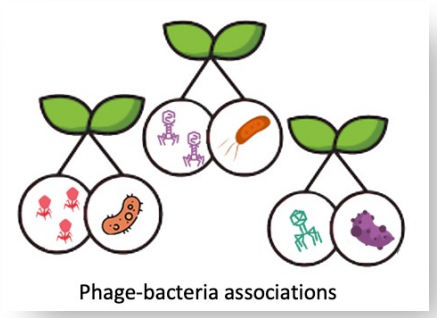


Figure 1 Steps involved in phage mediated *Mycobacterium tuberculosis* lysis using *Mycobacterium smegmatis*.

Azimi, T., Mosadegh, M., Nasiri, M. J., Sabour, S., Karimaei, S., & Nasser, A. (2019). Phage therapy as a renewed therapeutic approach to mycobacterial infections: a comprehensive review. *Infection and drug resistance*, *12*, 2943.

Host prediction: challenges

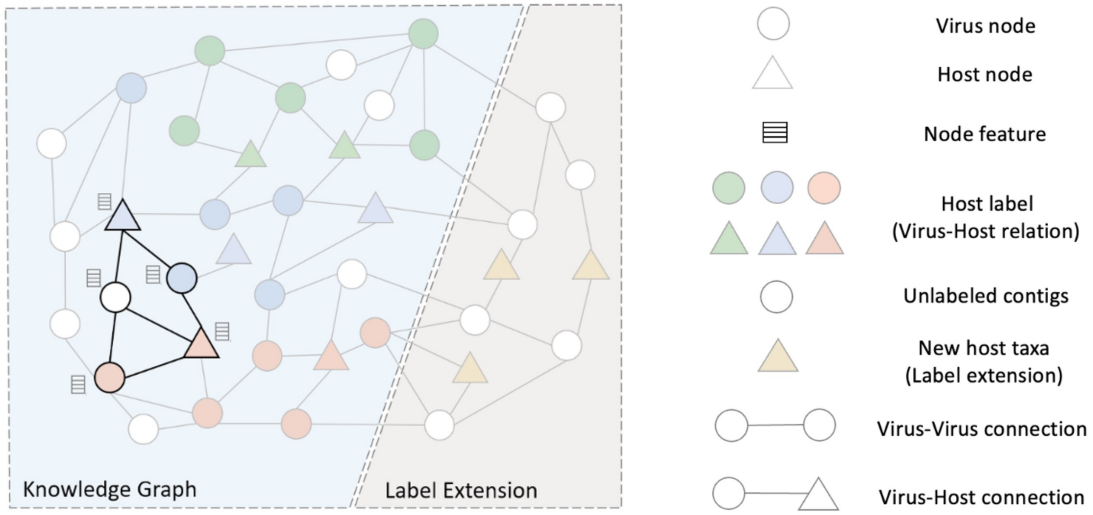


- Lack known virus-host interactions
  - The number of known interactions dated up to 2020 only accounted for ~40% (1,940) of the phages at the NCBI RefSeq
  - Among the 60,105 prokaryotic genomes at the NCBI RefSeq, only 223 kinds of species have annotated interactions
- Not all phages share common regions with their host genomes
  - ~24% phages do not have significant alignments (e-value < 1e-5) with their hosts
  - Thus, alignment-based methods have limited host prediction ability

Semi-supervised learning

- Limited known phages and large sequencing data → semi-supervised learning
  - Training on both labeled (known phages) and unlabeled (test) sequences

- Graph convolutional network (**GCN**)
  - Modeling the topological relationship between samples



Knowledge graph

Graph Convolutional Neural Network

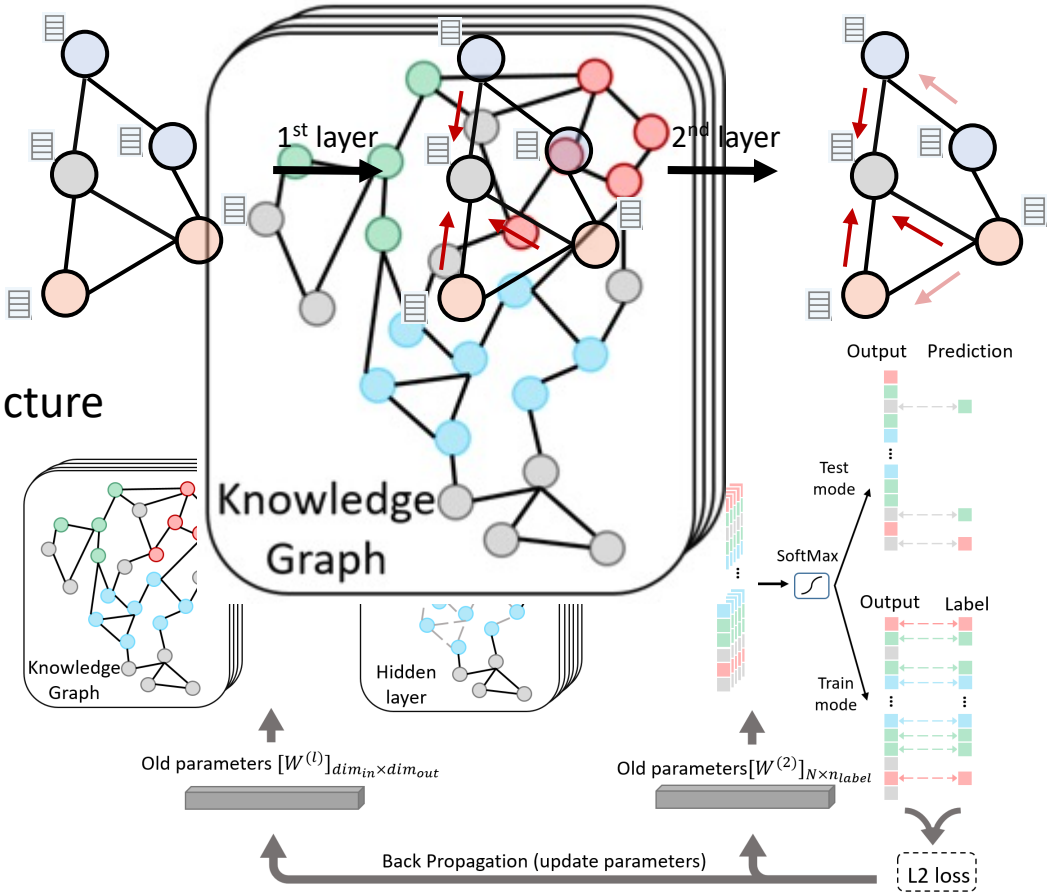
Key insight

$$H^{(l+1)} = ReLU(\tilde{D}^{-\frac{1}{2}}\tilde{G}\tilde{D}^{\frac{1}{2}}H^{(l)}W^{(l)})$$

$$Out = SoftMax(H^{(2)}W^{(2)})$$

- Using features from neighborhood

Structure

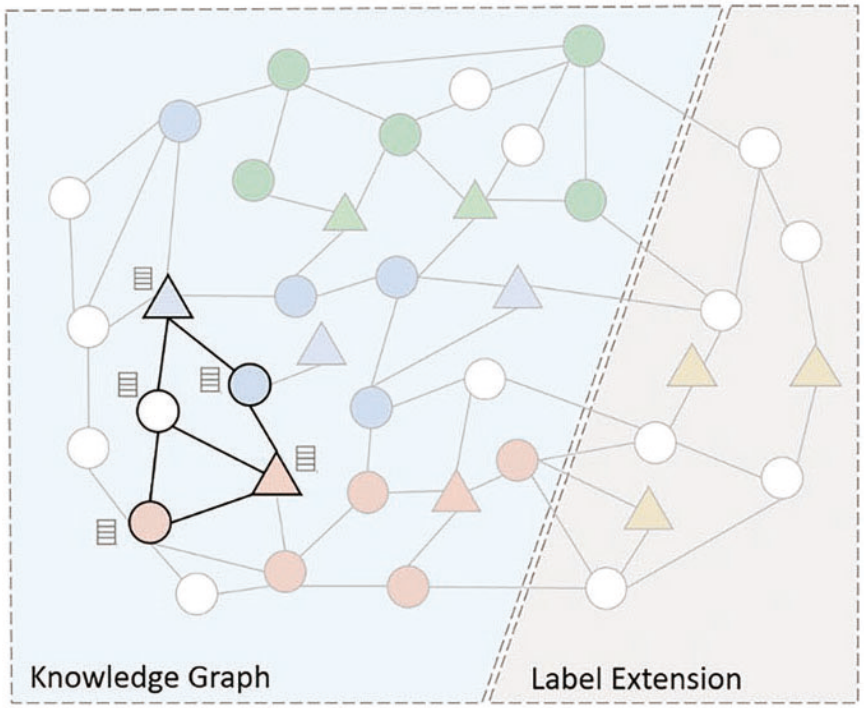


Graph convolutional neural network structure

Adapted Improvement 1

Graph extension

- Connecting the new bacteria by adding nodes and edges
- Training with the bacteria nodes (new labels)

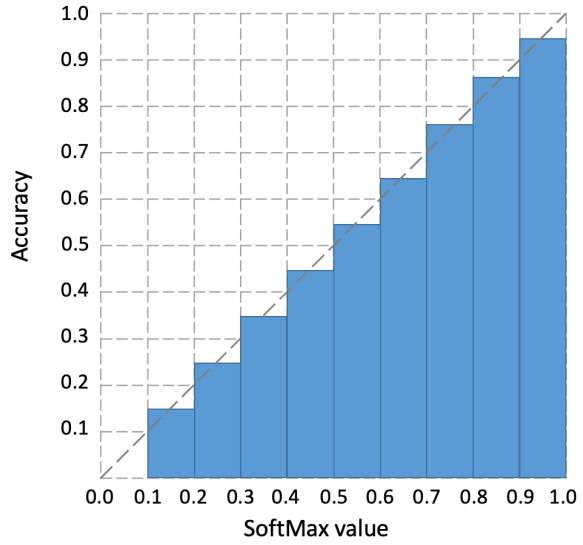


Label extension

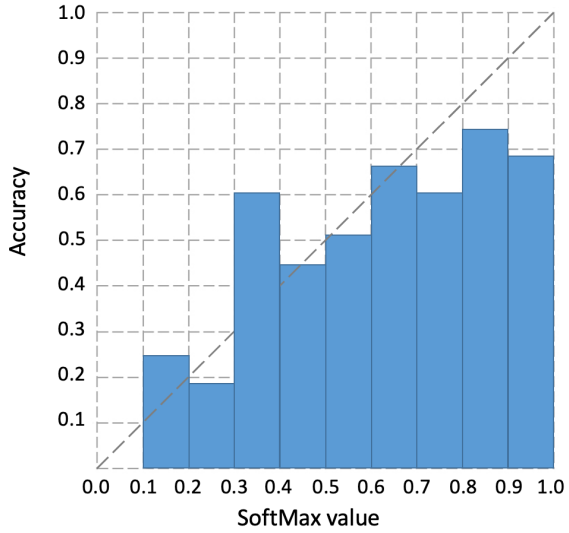
Adapted Improvement 2

How to represent confidence of the predictions

What we want



The real case



High SoftMax value ≠ High confidence

Adapted Improvement 2

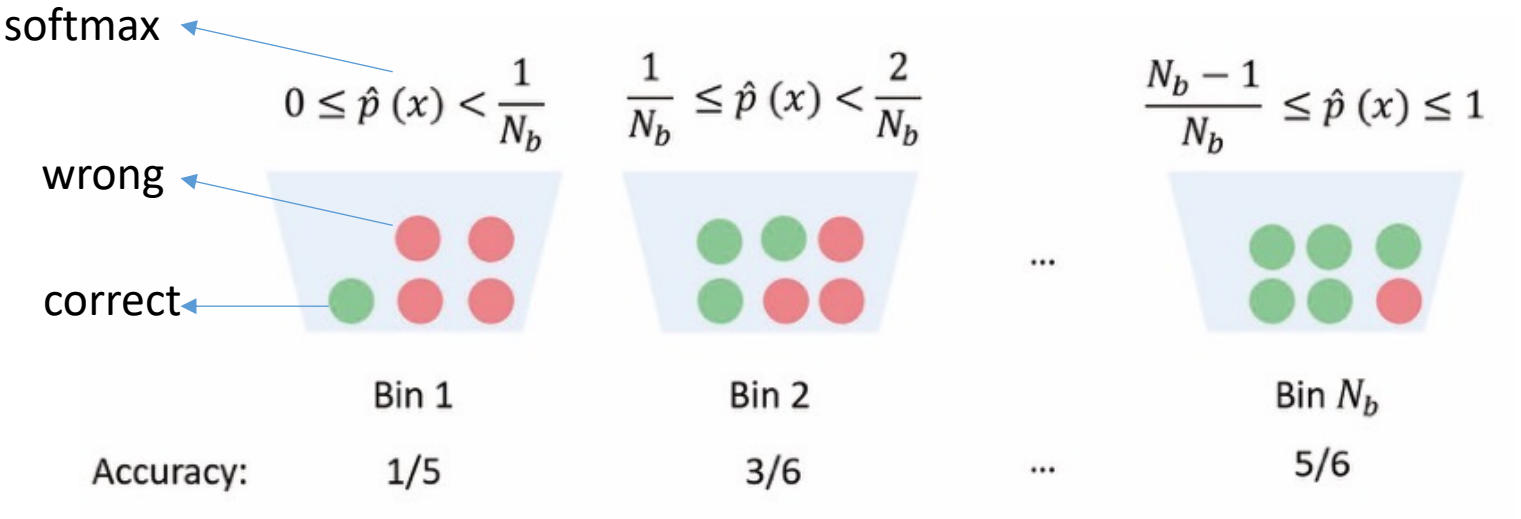
- Expected calibrated error (ECE)

$$\mathcal{L} = ECE + L2$$

$$ECE = \sum_i^{N_b} \frac{T_i}{T} |Acc_i - conf_i|$$

- Divided the confidence score into several bins
- Minimize $\|accuracy - confidence\|$ in each bin

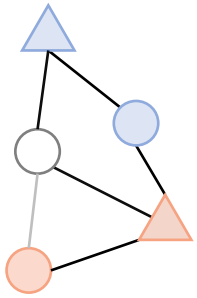
$$conf_i = \frac{\sum_j^{T_i} \hat{p}(x_{ij})}{S_i}$$



Theory of expected calibrated error (ECE)

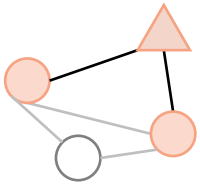
Hard cases for alignment-based methods

A) Has alignment results with bacteria in different taxa



269/656 phages have alignments with bacteria in different taxa at order level; 566/656 at family level; 656/656 at genus level.

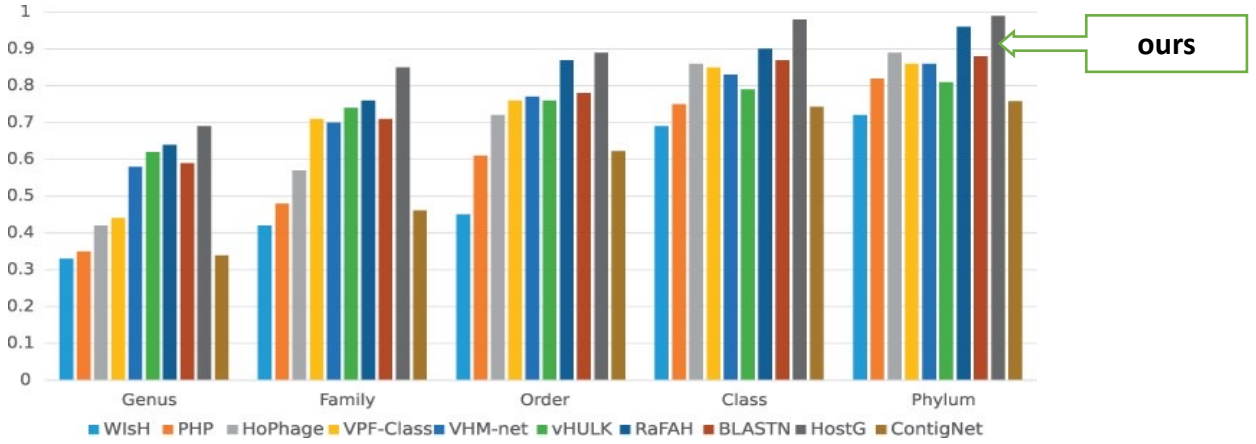
B) Has no alignment result with bacteria



770/1426 phages have no alignment

Third-party evaluation

Host prediction accuracy for whole genomes from genus to phylum



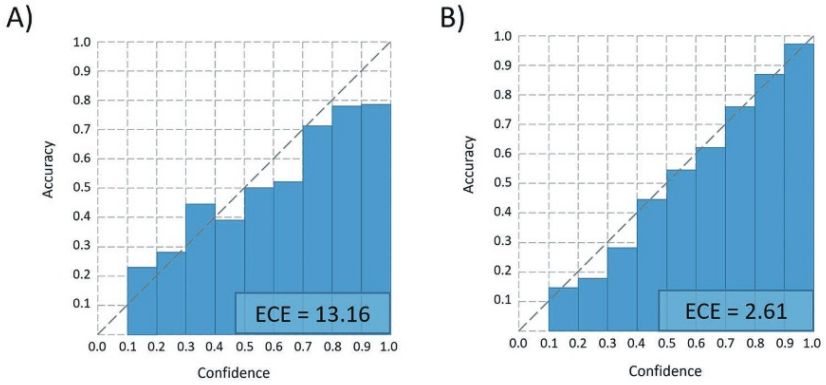
Host prediction accuracy for whole genomes without alignment results from genus to phylum

Bioinformatics, Volume 38, Issue Supplement_1, July 2022, Pages i45–i52, https://doi.org/10.1093/bioinformatics/btac239

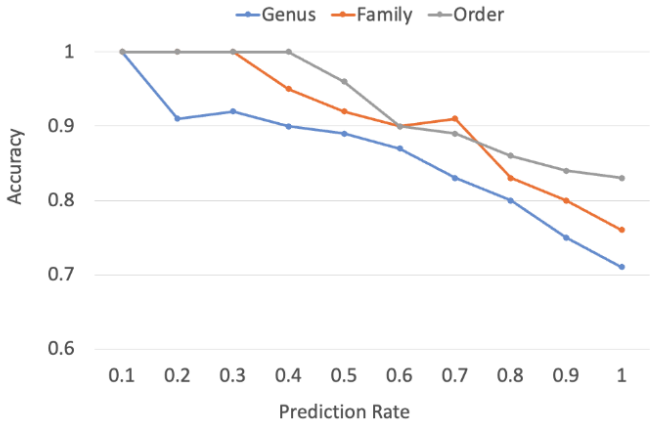
Experimental results

Improvement with ECE

- Accuracy vs. confidence (SoftMax value)



- Prediction with confidence

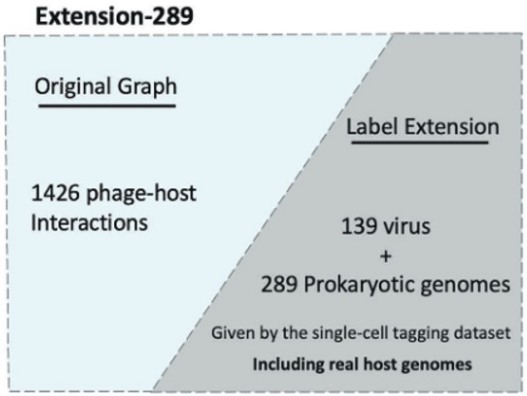


Experimental results

—— Single-cell viral tagging using a human stool sample

Improvement with label extension

- Extension of the knowledge graph



- Performance of the label extension