
Biological sequence analysis via deep learning



Dr. Yanni Sun

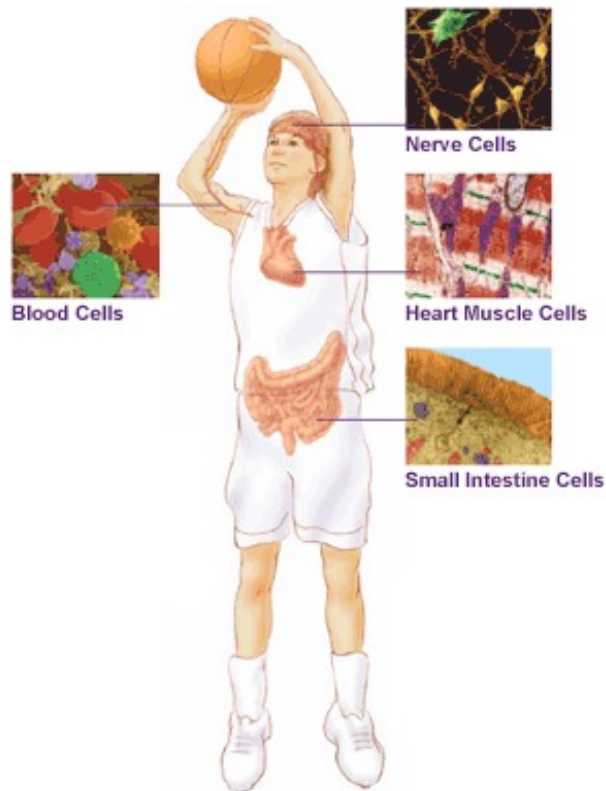
Electrical Engineering, City University of Hong Kong

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

Yanni Sun, PhD in Computer Science & Engineering,
research area: bioinformatics, sequence analysis,
application of deep learning for genomic data analysis

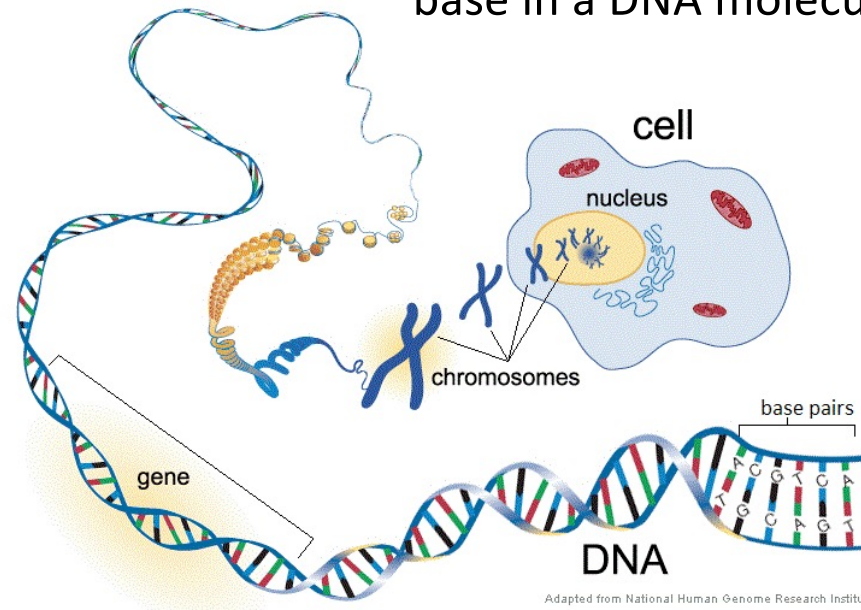


Genomic data: genomes

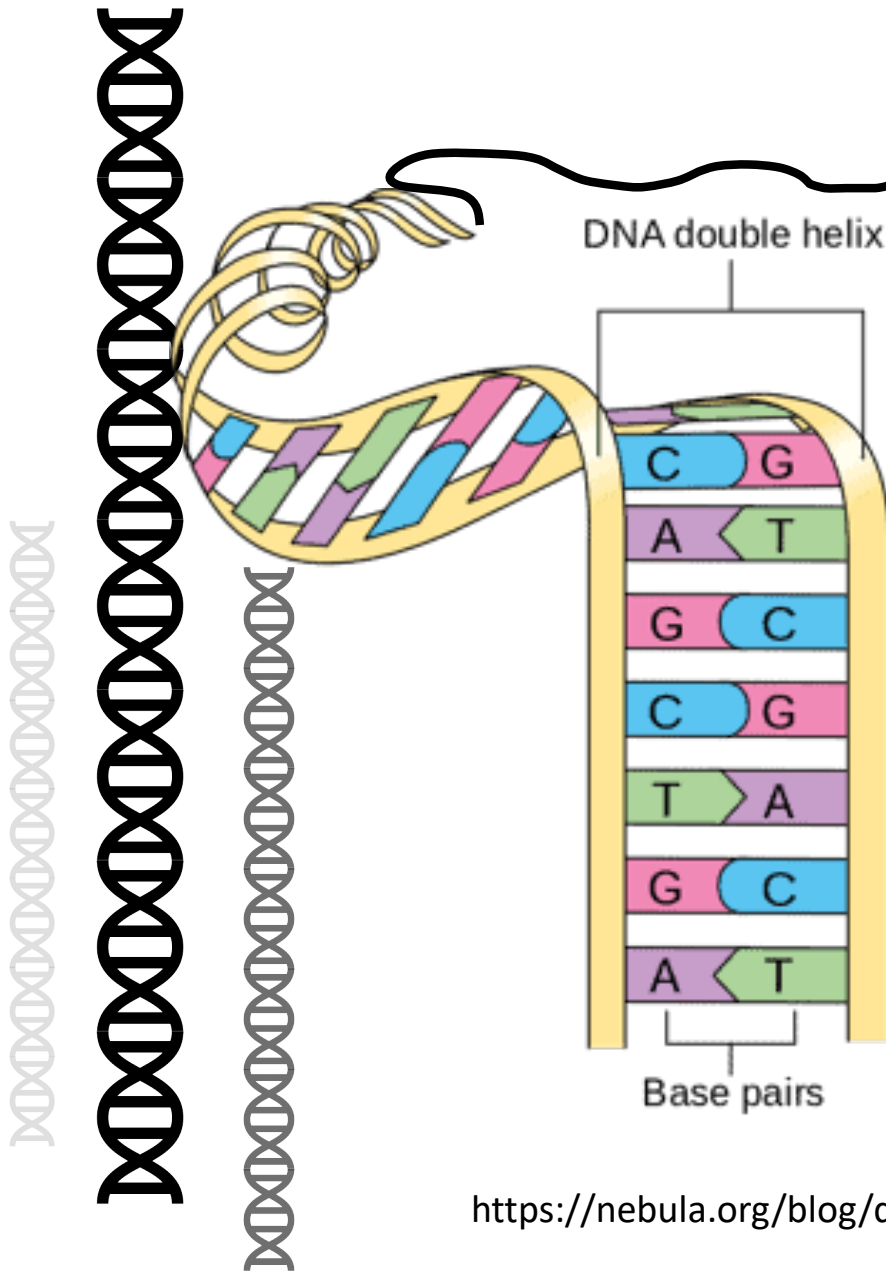


The primary structure of DNA sequence: a string defined on four bases: A, C, G, and T:

Sequencing : determine each base in a DNA molecule



The genome is in every cell except red blood cell and sex cells.



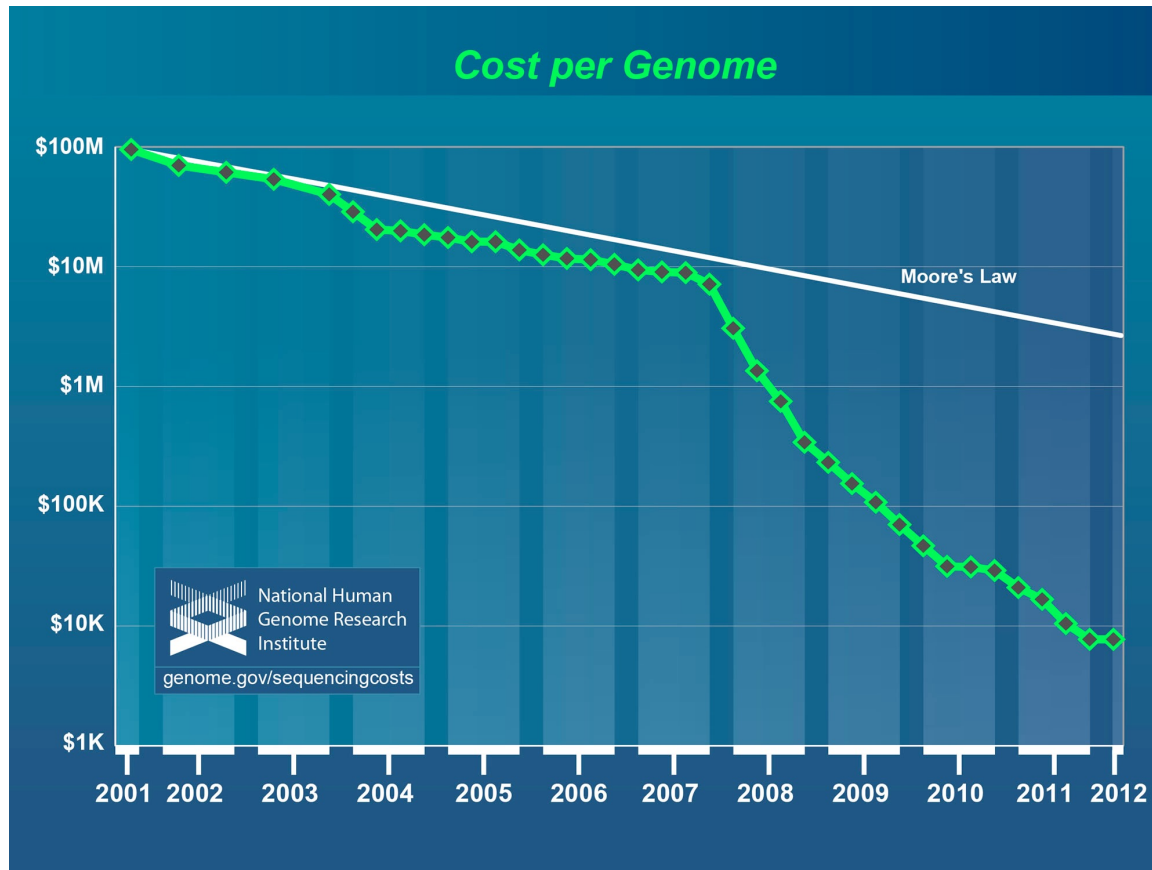
ATGGGCAAGTCAGAAAGTCAGATGG
 ATATAACTGATATCAACACTCCAAAG
 CCAAAGAAGAAACAGCGATGGACT
 CCACTGGAGATCAGCCTCTCGGTCCT
 TGTCTGCTCCTCACCATCATAGCTGT
 GACAATGATCGACTCTATGCAACCTA
 CGATGATGGTATTTGCAAGTCATCAG
 ACTGCATAAAATCAGCTGCTCGACTG
 ATCCAAAACATGGATGCCACCACTG

Language of Life

- Contains all functions that make each of us unique

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






BIG genomic data

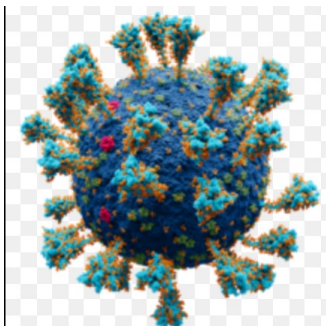


Low cost → fast accumulation of sequencing data

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

Fast accumulation of sequenced genomes

Species	<i>T2 phage</i>	<i>Escherichia coli</i>	<i>Drosophila melanogaster</i>	<i>Homo sapiens</i>	<i>Paris japonica</i>
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

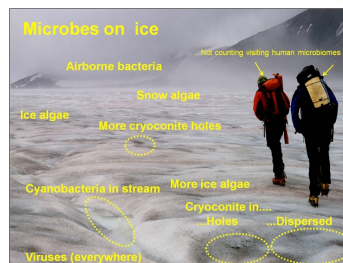


Length: ~30,000 bp

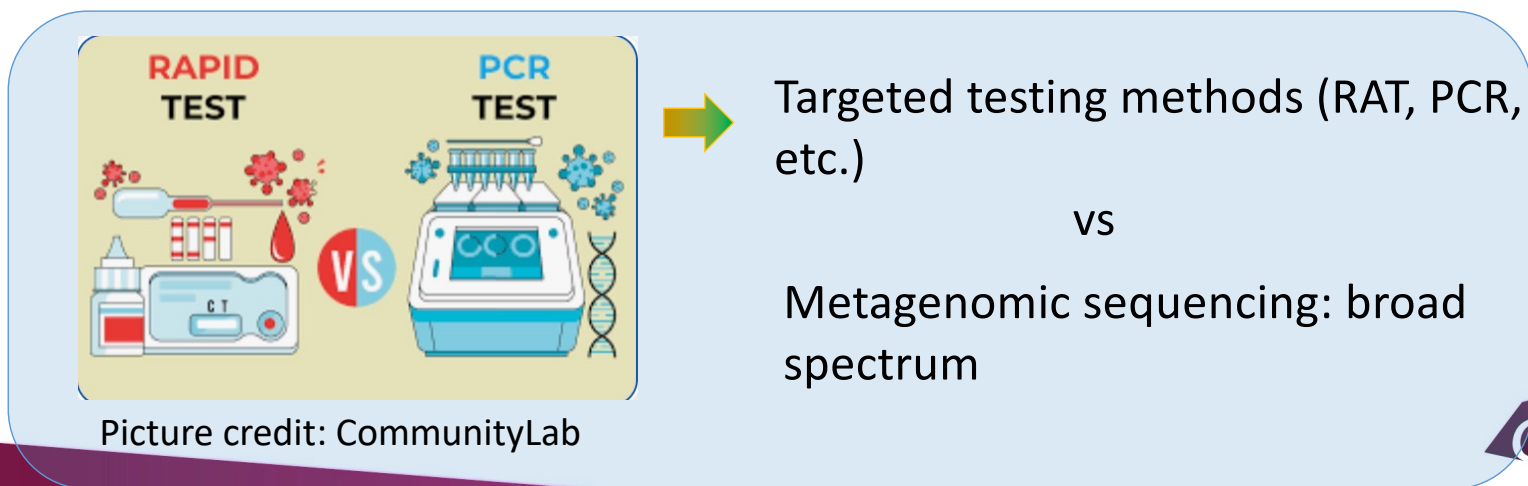
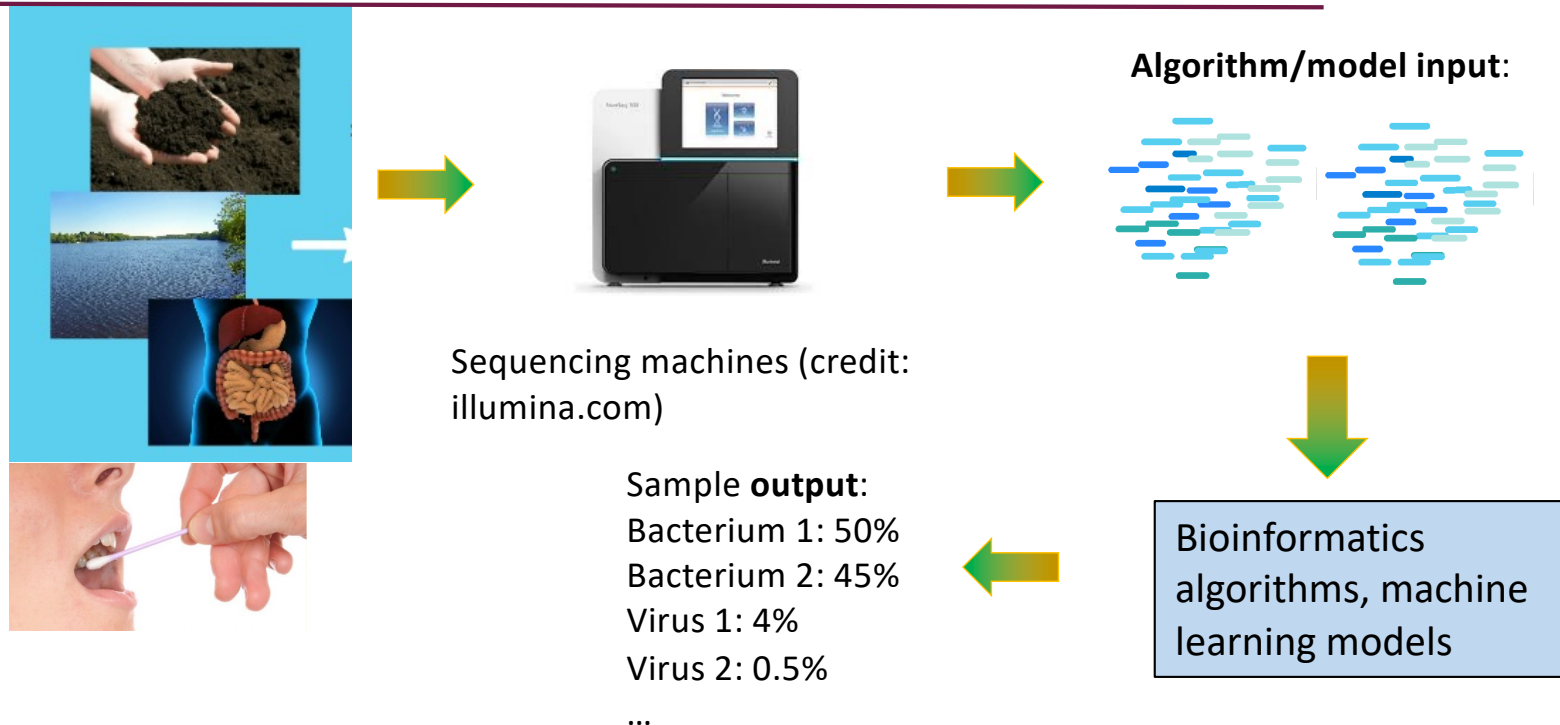
Pictures downloaded from Wiki

Analyze microbial communities using next-generation sequencing data

- Microbial communities: groups of microbes (such as bacteria, viruses) that share a common living space

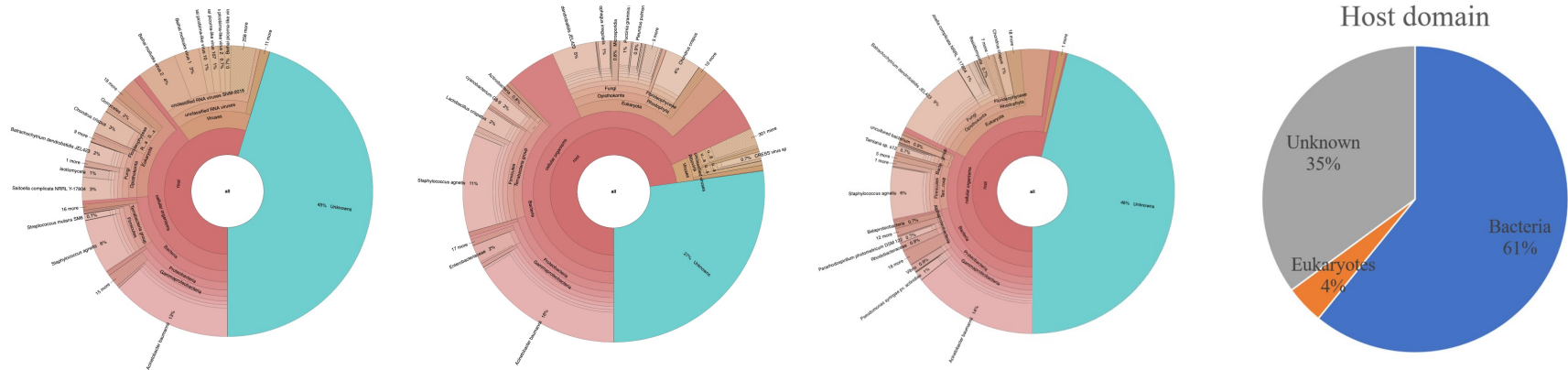


Powerful method to study microbiome: metagenomic sequencing

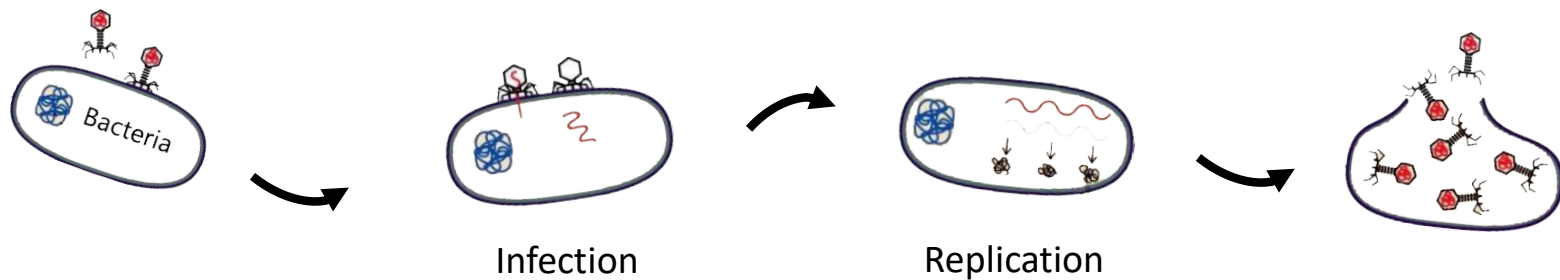


Dark matter in microbial communities: prokaryotic viruses

➤ The most common and diverse entities



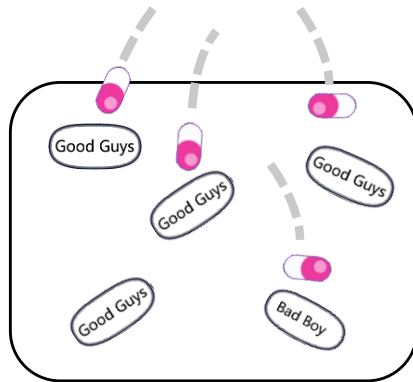
➤ Life Cycle (lytic Cycle)



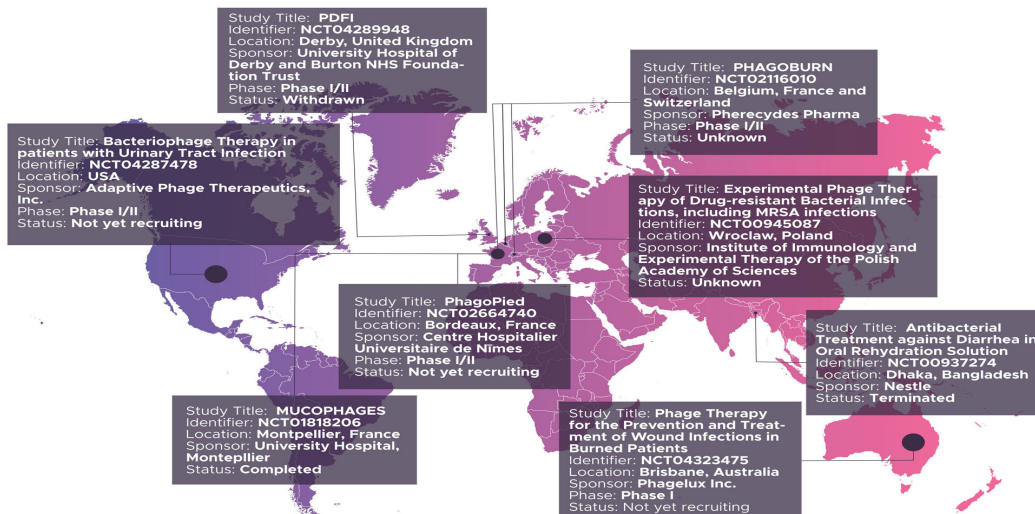
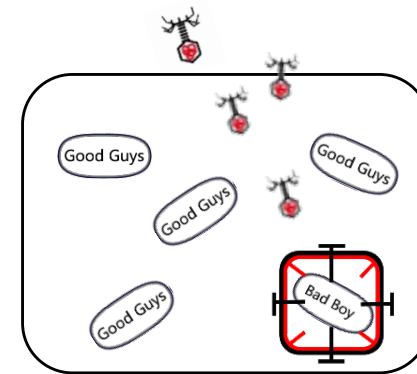
Phage therapy

Used as antibiotics

Antibiotic (Area of effect)



Phages (targeted)



Adesanya et al. An exegesis of bacteriophage therapy: An emerging player in the fight against anti-microbial resistance. 2020

Two research problems

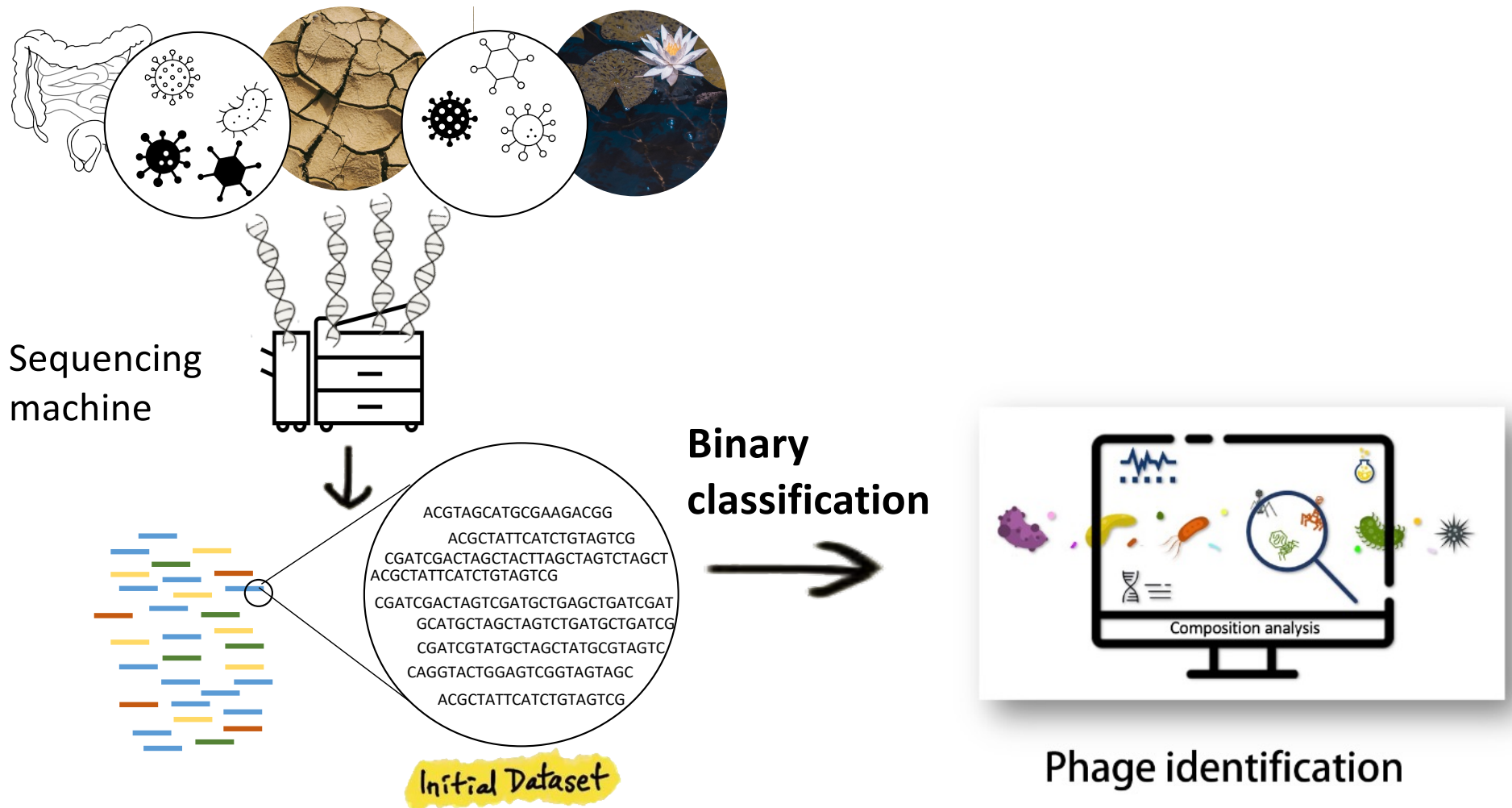
1. Identify phages from metagenomic data
2. Find the phage-bacteria interactions

Jiayu Shang, Xubo Tang, Ruocheng Guo, **Yanni Sun***, Accurate identification of bacteriophages from metagenomic data using Transformer, *Briefings in Bioinformatics*, Volume 23, Issue 4, July 2022, bbac258,

Jiayu Shang, Jinzhe Jiang, and **Yanni Sun***, "Bacteriophage classification for assembled contigs using Graph Convolutional Network", the 29th Annual International Conference on Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology (*ISMB/ECCB 2021*) July 25, 2021 **acceptance rate: 18%**

Jiayu Shang and **Yanni Sun***, "Detecting the hosts of bacteriophages using GCN-based semi-supervised learning", *BMC Biology*, 2021

Phage identification: problem formulation



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

**Methodology: towards more sensitive
and accurate phage identification using**

Transformer

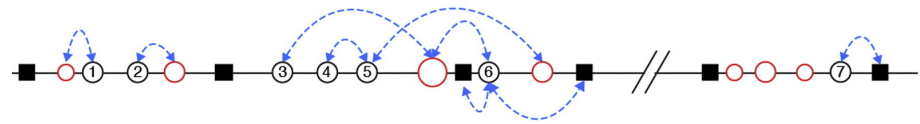
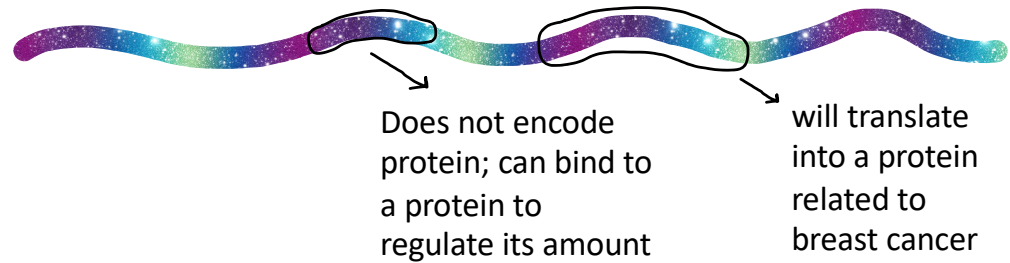
Natural language vs. language of life

- **Alphabet**

- English: {A,B,C,D, ..., Z}
- DNA: {A,C,G,T}

- **Words**

- {happy, sad, student...}: delimited by spaces
- What are the words for DNA sequences?

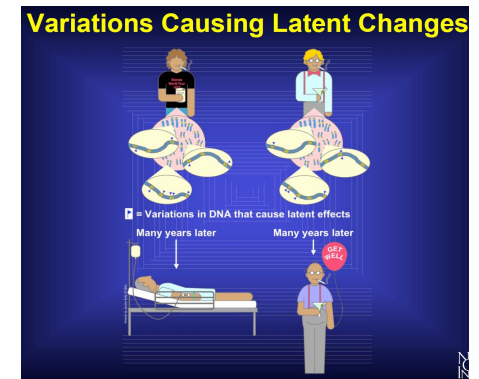
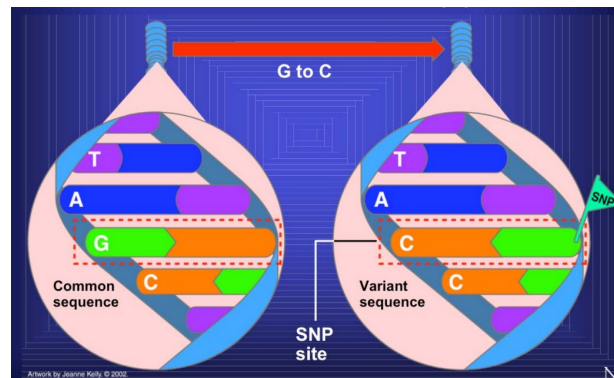


- **Interactions/associations**

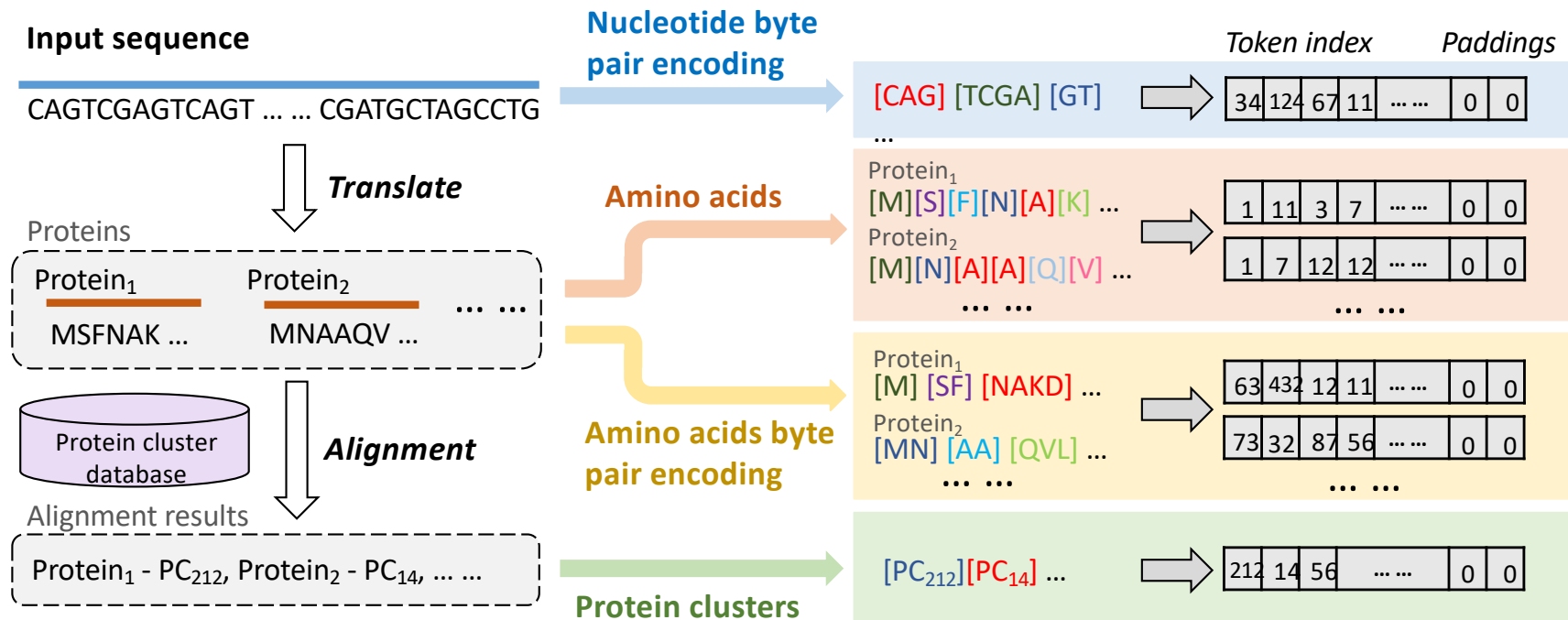
- She quickly corrected her mistakes. She corrected her mistakes quickly.

- **Errors in spelling**

- I like my bok.
- Variations in DNA → diseases



Token (word) construction



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

Nucleotide byte pair encoding tokenizer

Input sequence

CAGTCGAGT CGATGCG



Token index				Paddings		
34	124	67	11	...	0	0

Build the BPE vocabulary

Iteration	Sequence	BPE Vocabulary
0	A C A C G A C G T	{A, C, G, T}
1	AC AC G AC G T	{A, C, G, T, AC}
2	AC ACG ACG T	{A, C, G, T, AC, ACG}
⋮	⋮	⋮

Start with single nucleotides (A, T, C, G)

Expand the vocabulary by adding DNA substrings with high occurrence frequency

Iterate and repeat to synthesize longer tokens

Count the most frequent nucleotide combinations.

Easy to implement, but the encoded vector of the input sequence is usually too long.

Amino acids tokenizer

Protein

MSFNAKDVKQKCQVAA

[M] [S] [F] [N] [A] ...

Token index				Paddings		
1	11	3	7	...	0	0

20 standard amino acids (*A, R, N, D, ...*), undefined amino acid (*X*), and other amino acids (*OTHER*).

Amino acids byte pair encoding tokenizer

Protein

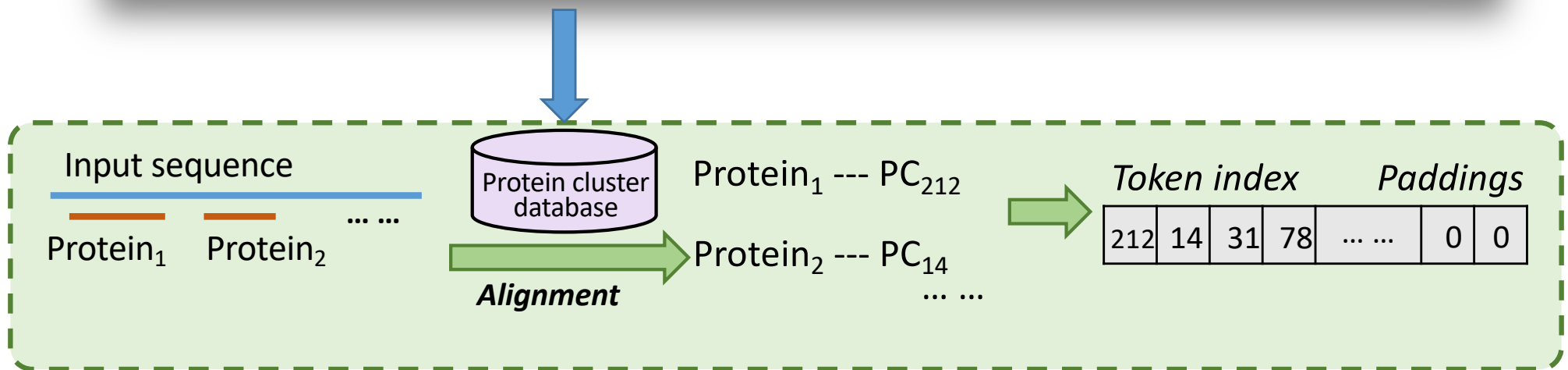
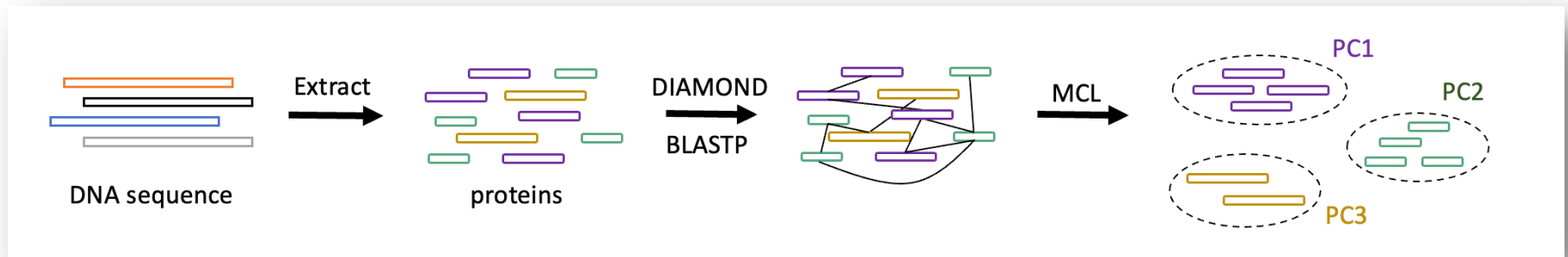
MSFNAKDVKQKCQVAA

[M] [SF] [NAKD] ...

Token index				Paddings		
63	432	12	11	...	0	0

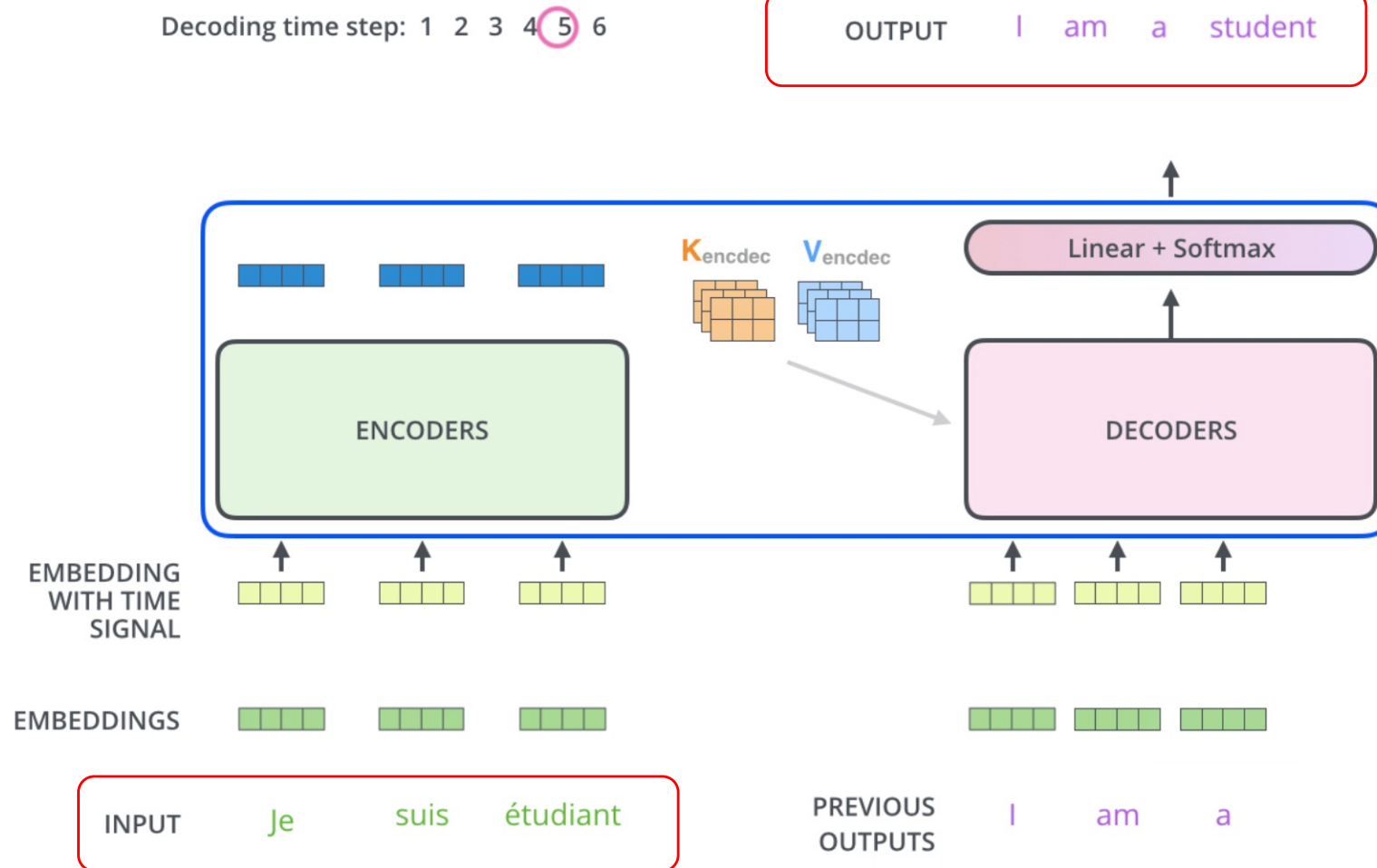
Amino acids BPE vocabulary: *count the most frequent amino acid combinations.*

Protein cluster (PC) tokenizer



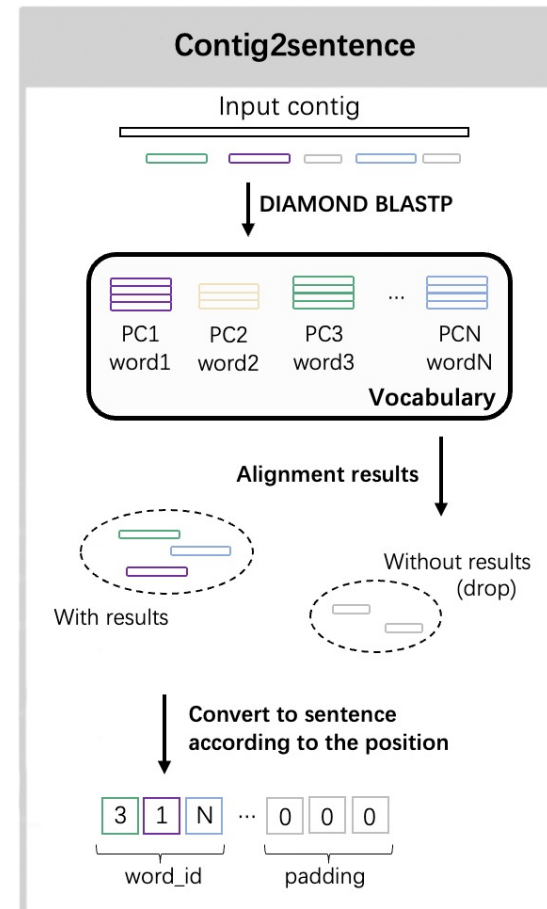
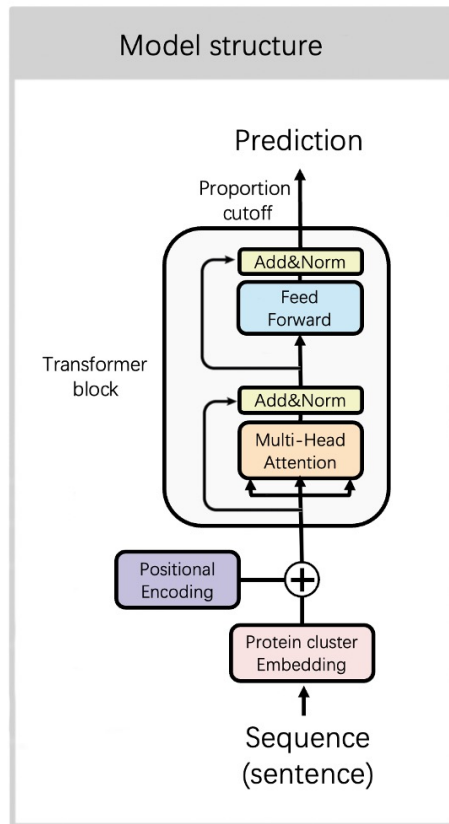
- Learn the importance of proteins
- Capture the correlation between different proteins

Transformer for translation



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

Phage Identification with Transformer



Multi-head attention:

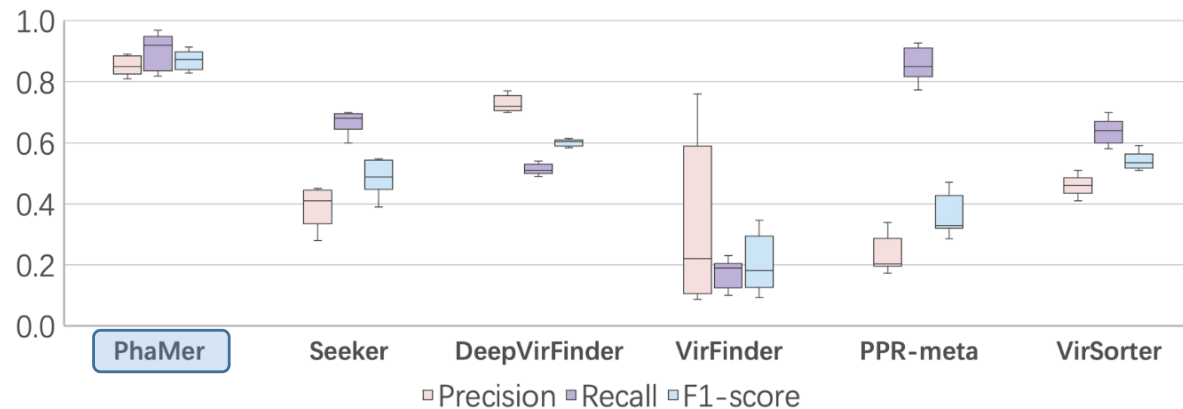
- Learn the meaning of the word
- Learn the correlation between word
- Learn the importance of the PCs
- Learn the association between proteins

Phage Identification – Experimental results

Dataset

- Using phages 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

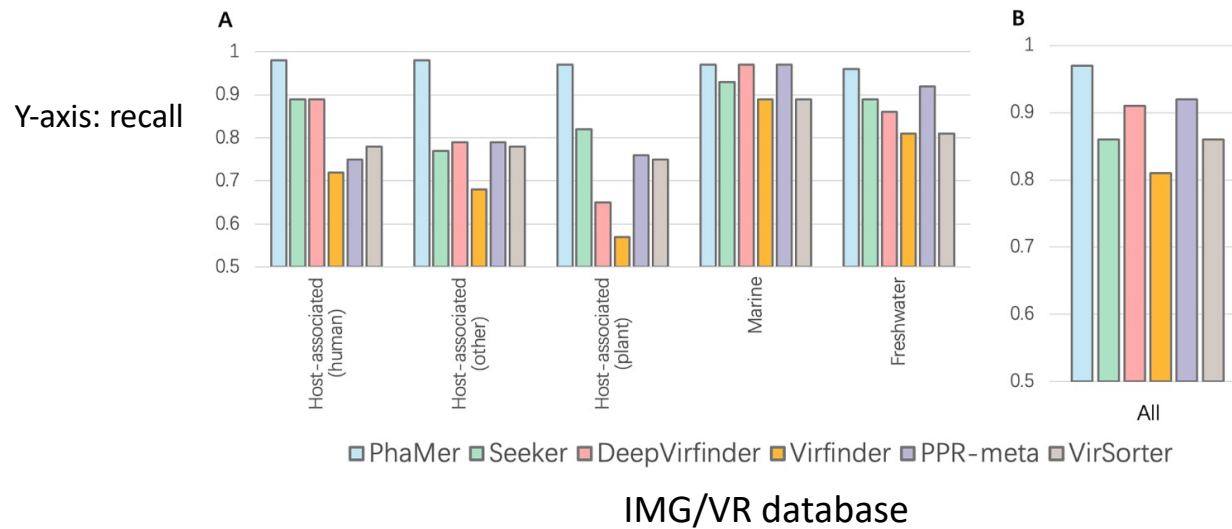
Results



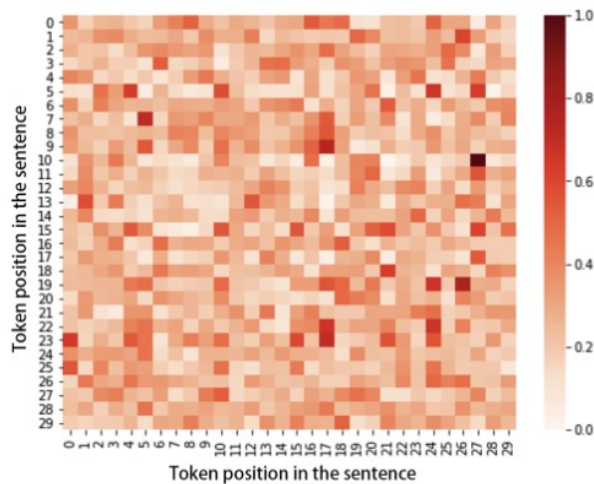
Mock dataset (European Nucleotide Archive PRJEB19901, ~30 species/strains)

Phage Identification – Experimental results

Results



Visualization



Attention score matrix in Transformer

- Attention score matrix in Transformer

The high score are the PCs contains structural protein (tail fiber/baseplate/...)

Host Prediction using GCN

Problem formulation

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

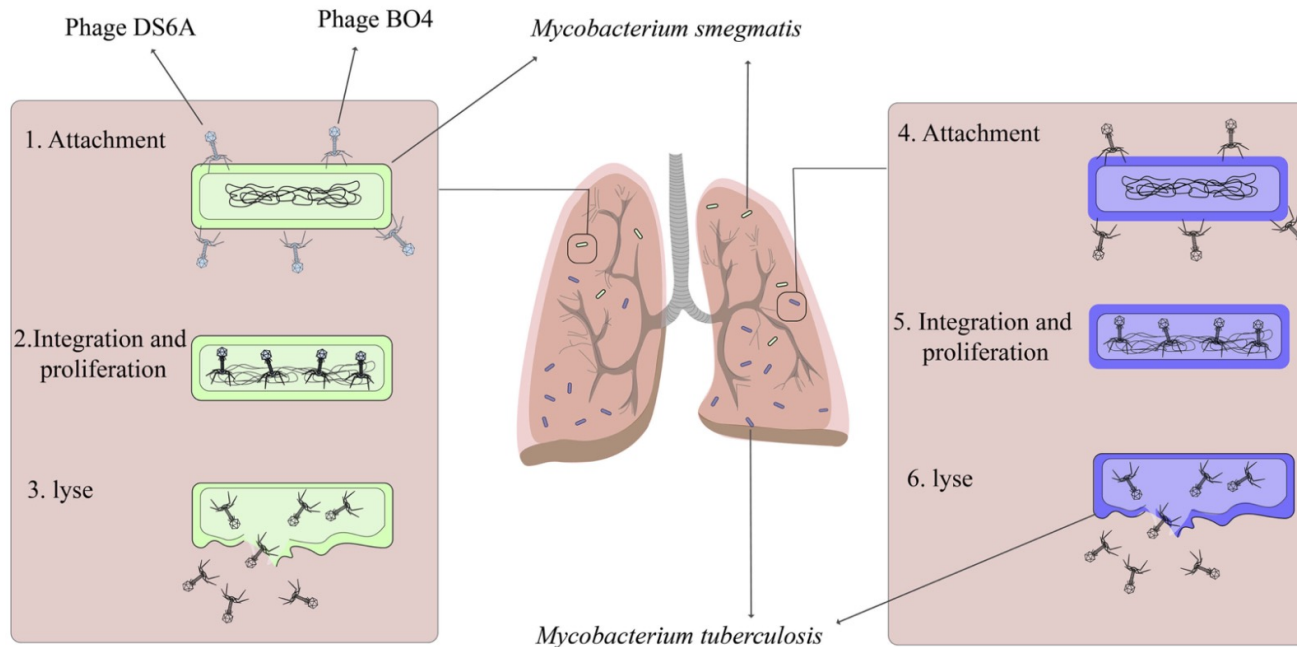
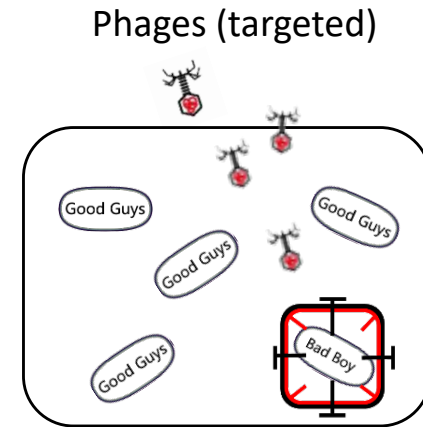
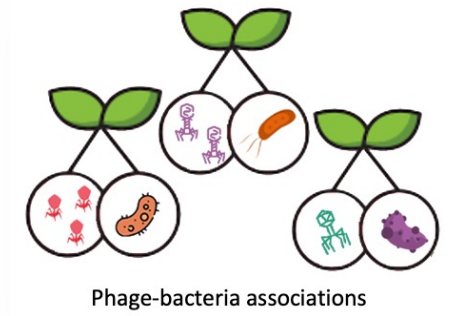


Figure 1 Steps involved in phage mediated *Mycobacterium tuberculosis* lysis using *Mycobacterium smegmatis*.

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 ($e\text{-value} < 1e\text{-5}$) with their hosts
 - Thus, alignment-based methods have limited host prediction ability

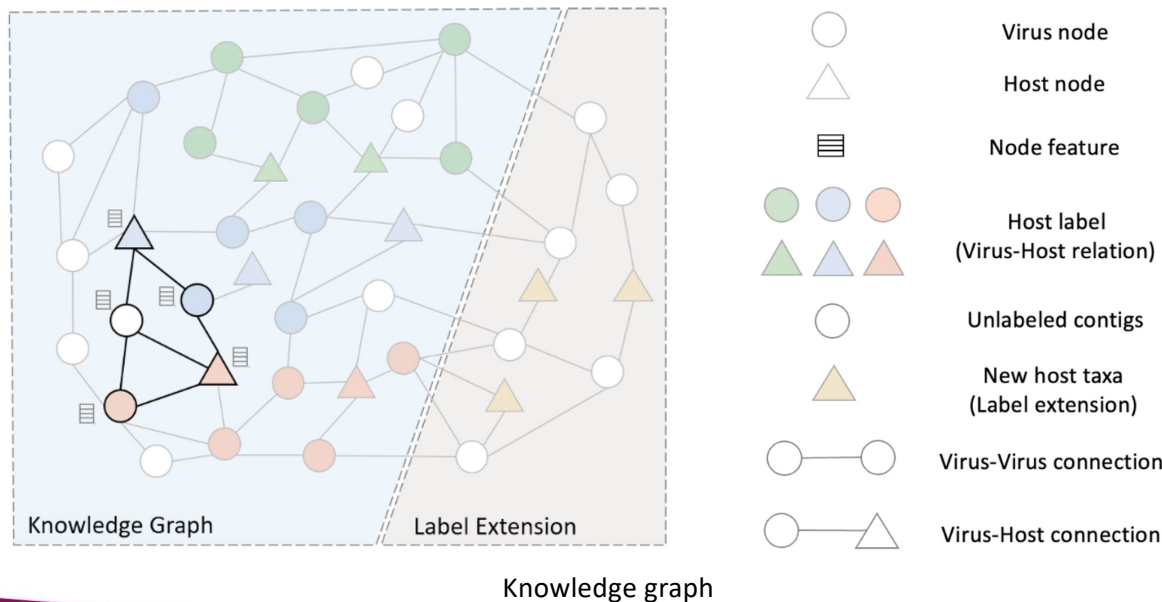
Semi-supervised learning

➤ Limited known phages and large sequencing data → semi-supervised learning

- Training on both labeled (known phages) and unlabeled (test) sequences

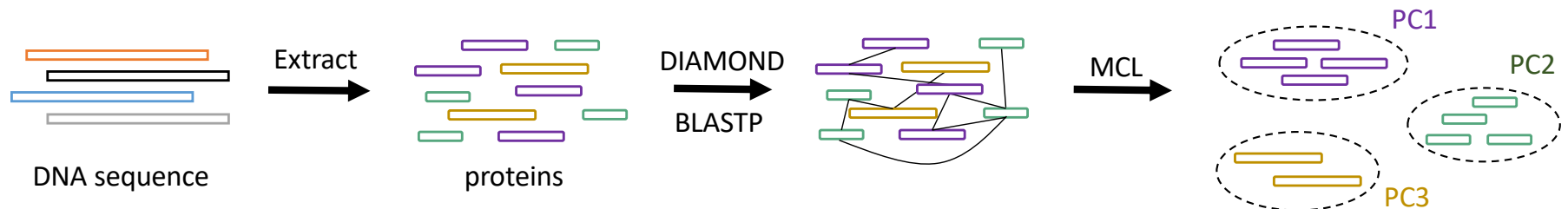
➤ Graph convolutional network (**GCN**)

- Modeling the topological relationship between samples



Edge construction I: virus-vs-virus

- **Similar protein organizations** -> might infect the same host
- **Protein cluster** construction using Markov Chain clustering (MCL)



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Phage A	Green	White	Green	White	Green	Green	White	White
Phage B	Green	White	White	White	Green	Green	White	Green
Phage C	White	Green	White	Green	White	White	Green	White
Phage D	White	White	Green	White	White	Green	White	White

Example (A and B):

$$c = 3, n = 8, a = b = 4$$

Share at least c protein clusters

$$P(X \geq c) = \sum_{i=c}^{\min(a,b)} \frac{C_a^i C_{n-a}^{b-i}}{C_n^b}$$



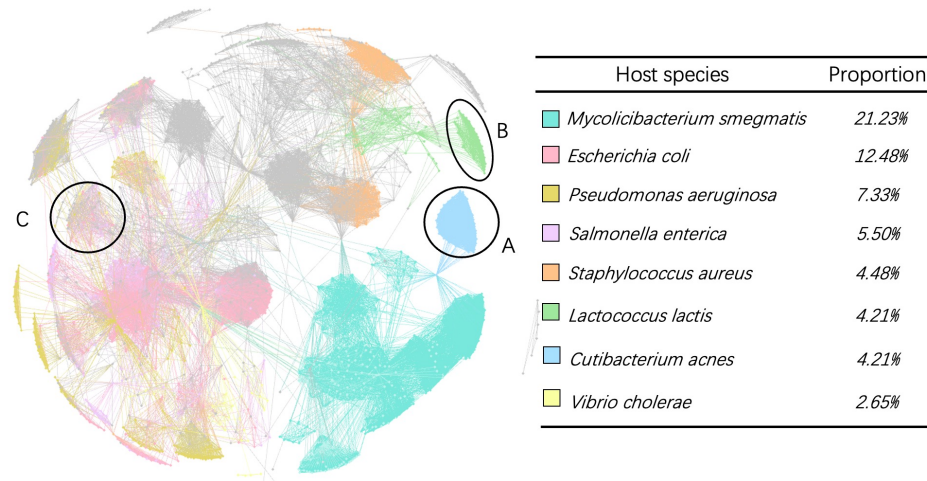
[1] Bolduc, Benjamin, et al. "vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria." PeerJ 5 (2017)

Edge construction II

➤ Virus vs bacteria

$$virus-prokaryote = \begin{cases} 1 & \text{if } \exists \text{ CRISPR alignment} \\ & \text{or BLASTN } E_{value} < \tau_2 \\ & \text{or } \exists \text{ interaction in dataset} \\ 0 & \text{otherwise} \end{cases}$$

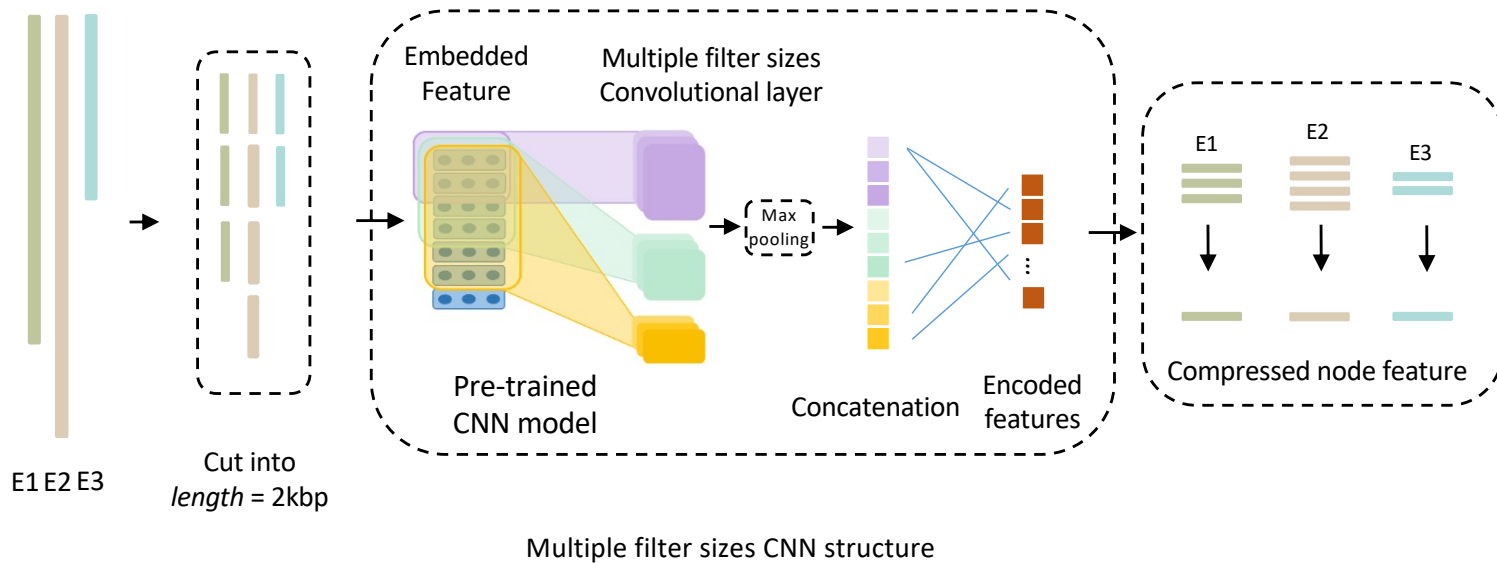
➤ Overview of the knowledge graph



Node feature encoding

➤ Capture **motif-related** patterns from the DNA sequences

- Different **filter sizes** -> different **length** of motifs



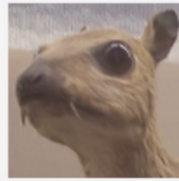
Motifs and convolution filters in CNN

- Conserved sequence patterns:
 - important features for genomic sequence classification
 - Can be represented by convolution filters

One hot encoding:

$$\text{AACG} \rightarrow \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$


Input image



Convolution Kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Feature map



Edge detector kernel
<https://developer.nvidia.com/discover/convolution>

Motif finding problem parameters

$L=35$

```

tacatAGAAGAAAGGggcgtagcacacggttagcgccg
tttgagcagatttagtcctggaaaCAATAAAACGa
tgggatgacttAAAATAATGGtcggatcattcga
ggatgCAAAAAAAGGtccacgcaaaggcaaggaga
ggtaaggctggttagctagATAATAAAGGctatag
    
```

$$f = \frac{1}{5} \begin{bmatrix} 3 & 4 & 5 & 4 & 2 & 5 & 5 & 4 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4 & 5 \\ 0 & 1 & 0 & 1 & 2 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix}$$

$W=10$ Position weight matrix (PW for this motif (width $W=10$))

Zia, Amin & Moses, Alan. (2012).

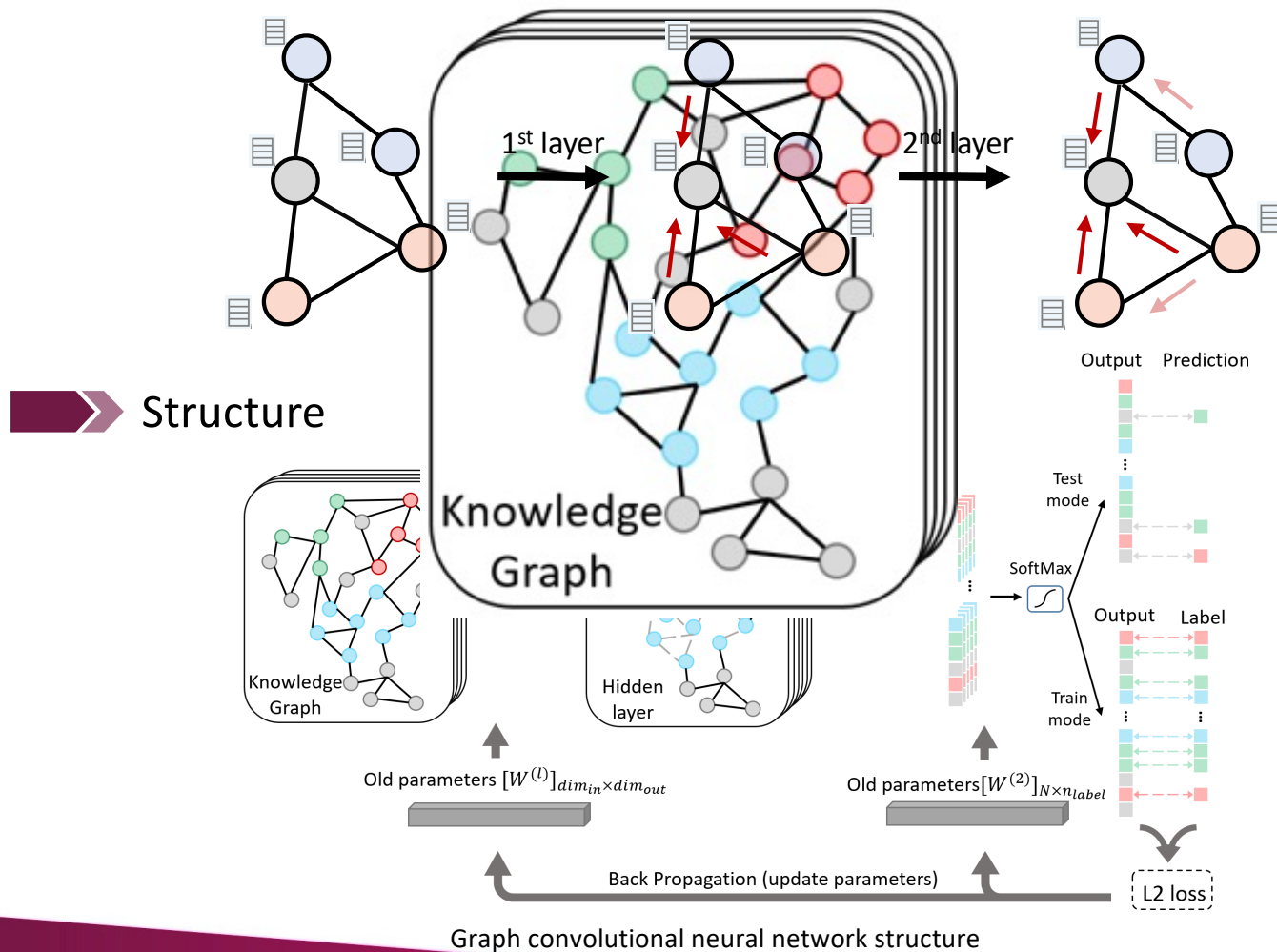
Graph Convolutional Neural Network

Key insight

- Using features from neighborhood

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

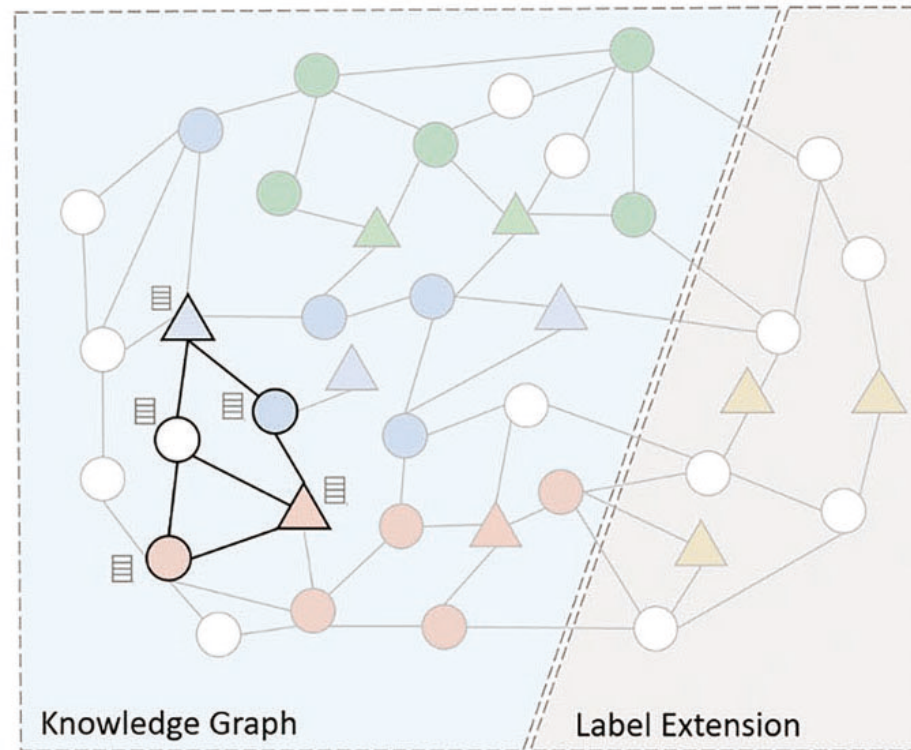
$$\text{Out} = \text{SoftMax}(H^{(2)} W^{(2)})$$



Adapted Improvement 1

➤ Graph extension

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

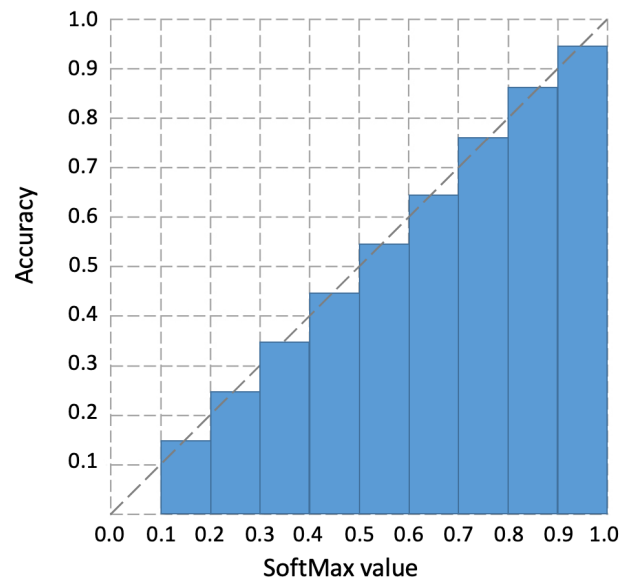


Label extension

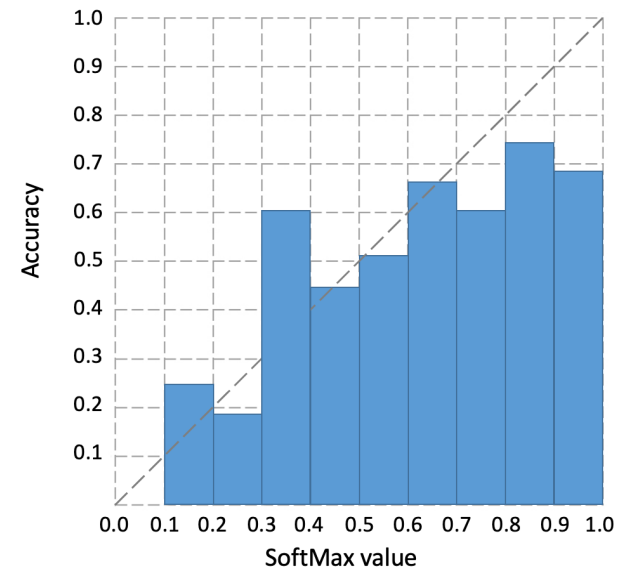
Adapted Improvement 2

➤ How to represent confidence of the predictions

What we want



The real case



High SoftMax value \neq High confidence

Adapted Improvement 2

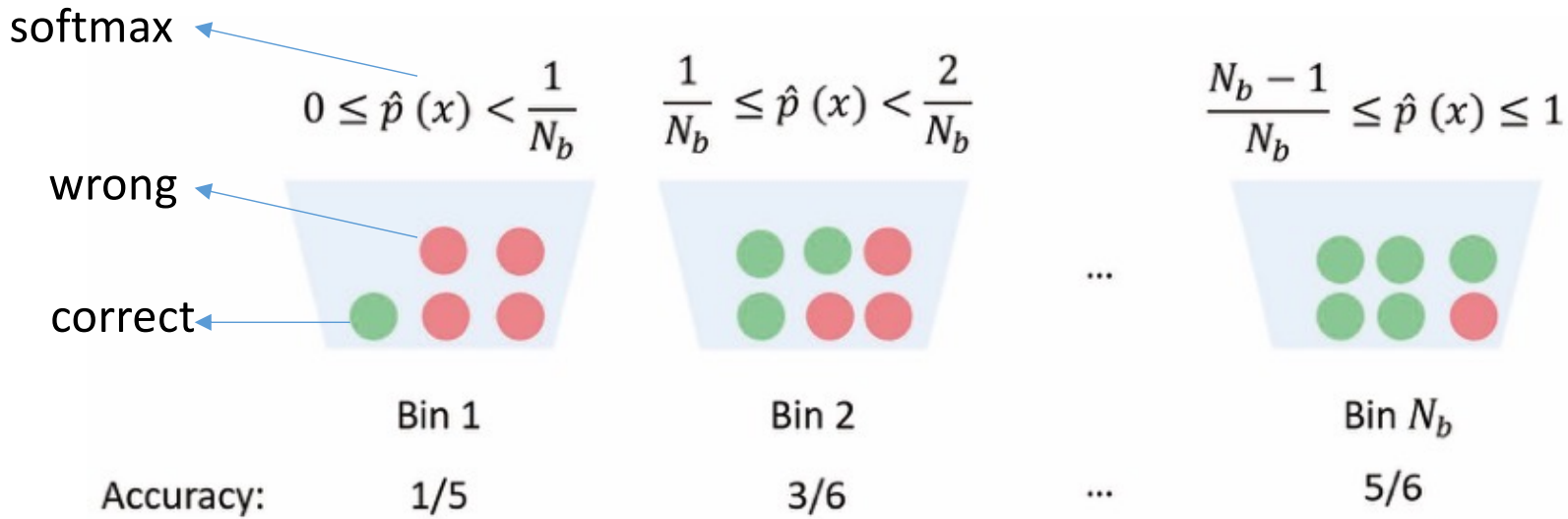
Expected calibrated error (ECE)

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

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

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

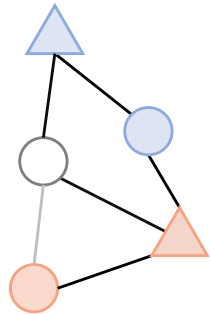
$$conf_i = \frac{\sum_j^{T_i} \hat{p}(x_{ij})}{S_i}$$



Theory of expected calibrated error (ECE)

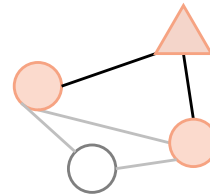
Hard cases for alignment-based methods

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.

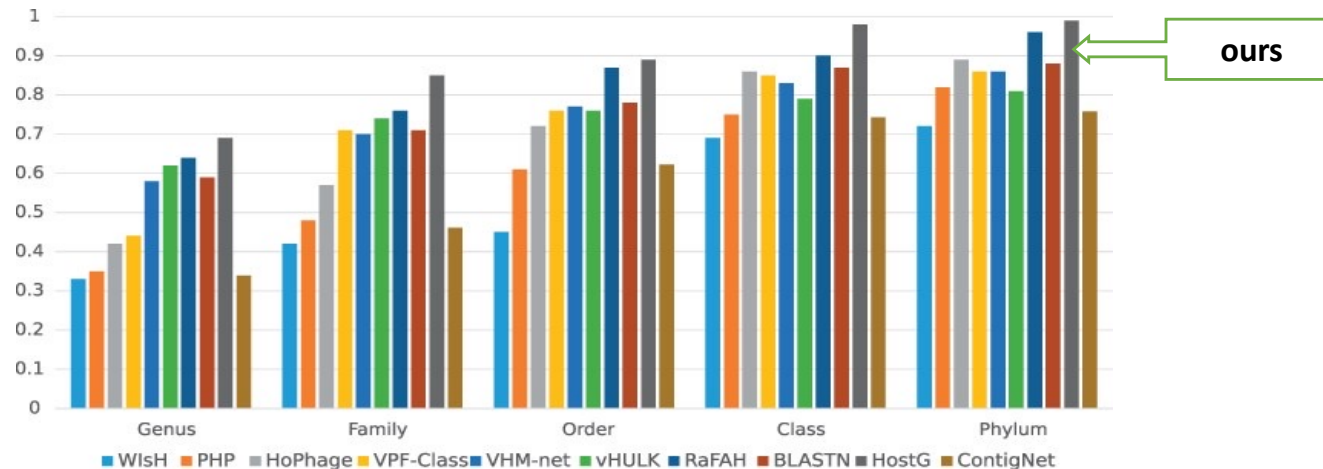
B) Has no alignment result with bacteria



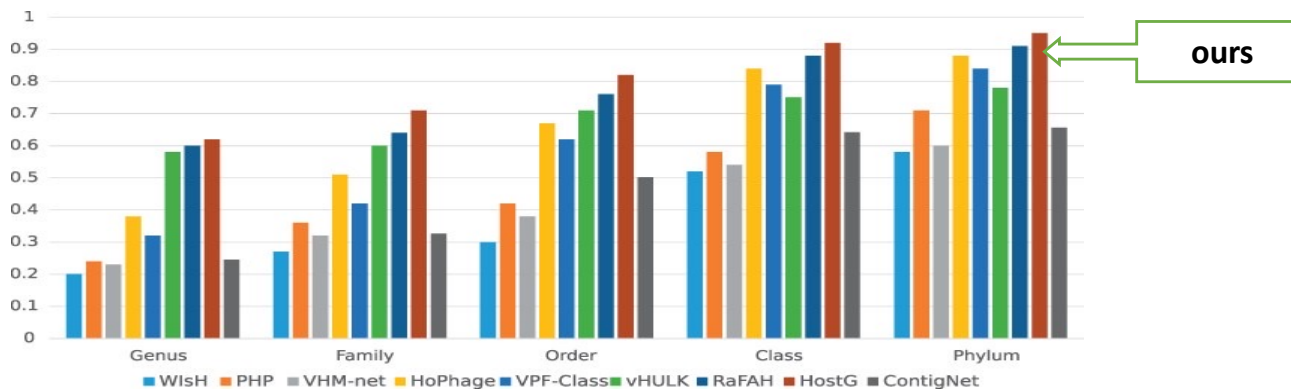
770/1426 phages have no alignment

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



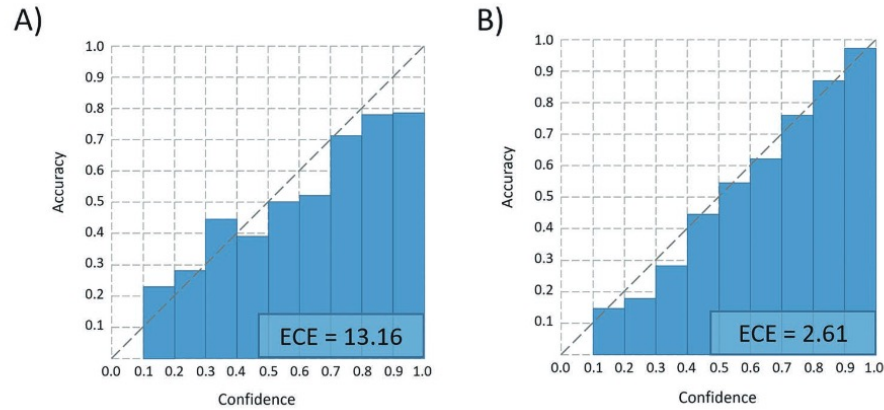
Bioinformatics, Volume 38, Issue Supplement_1, July 2022, Pages i45–i52, <https://doi.org/10.1093/bioinformatics/btac239>

The content of this slide may be subject to copyright: please see the slide notes for details.

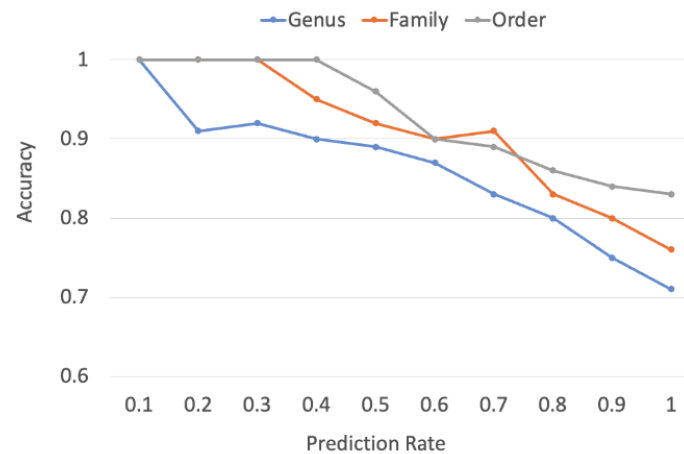
Experimental results

Improvement with ECE

- Accuracy vs. confidence (SoftMax value)



- Prediction with confidence

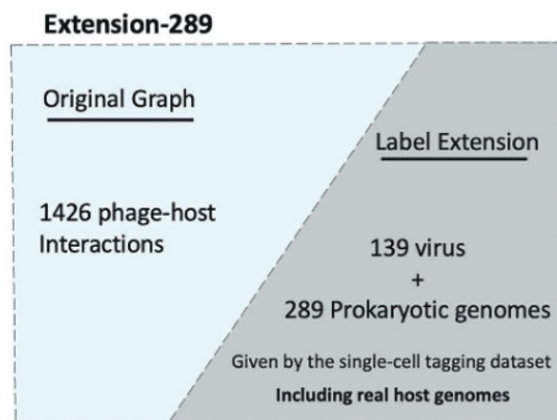


Experimental results

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

➤ Improvement with label extension

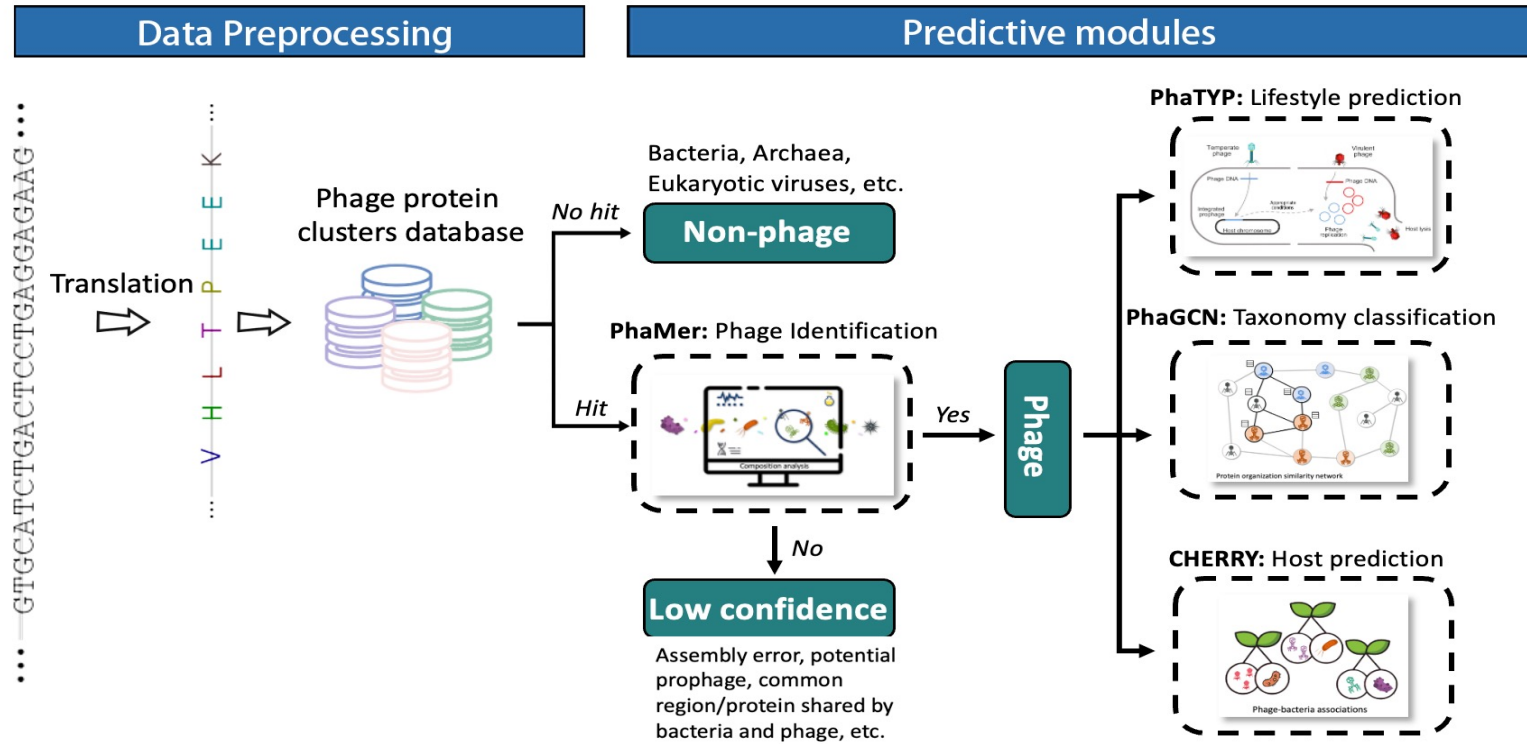
- Extension of the knowledge graph



- Performance of the label extension



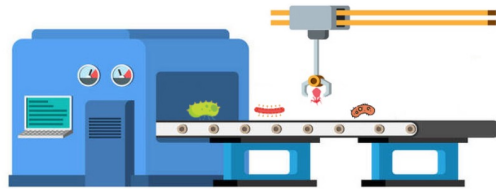
PhaBOX



<https://phage.ee.cityu.edu.hk/>
Our web server

Our tools for phage sequence analysis

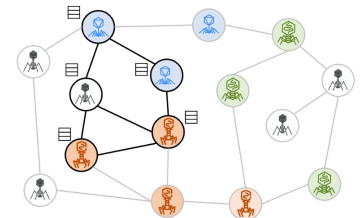
PhaMer



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

<https://github.com/KennthShang/PhaMer>

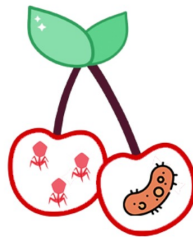
PhaGCN



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>

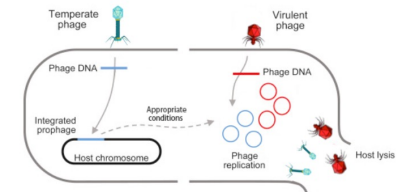
CHERRY



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>



Acknowledgement

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

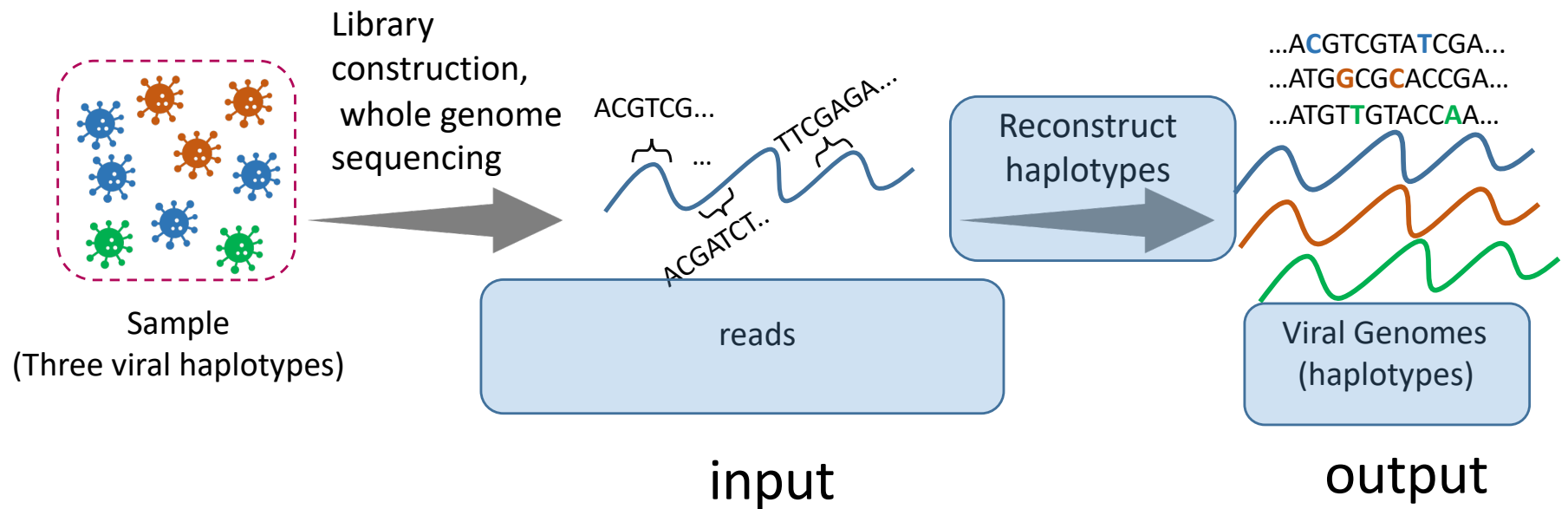


Hong Kong Science Museum

Questions?

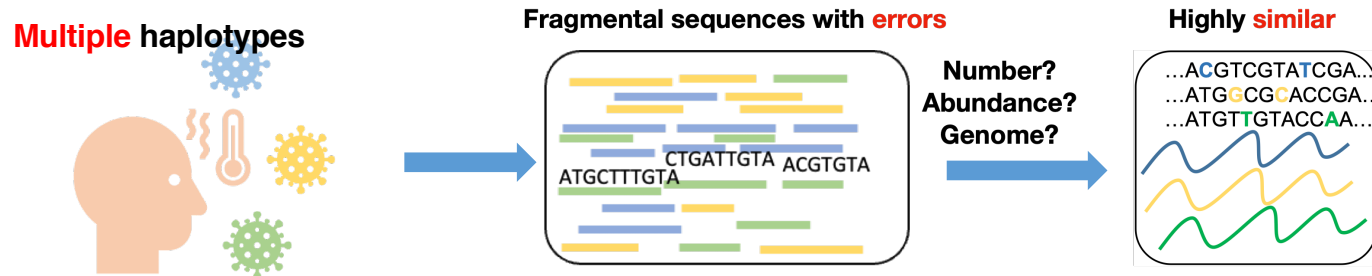
For more information, please visit the lab website:
<https://yannisun.github.io/>

Viral haplotype characterization using TGS

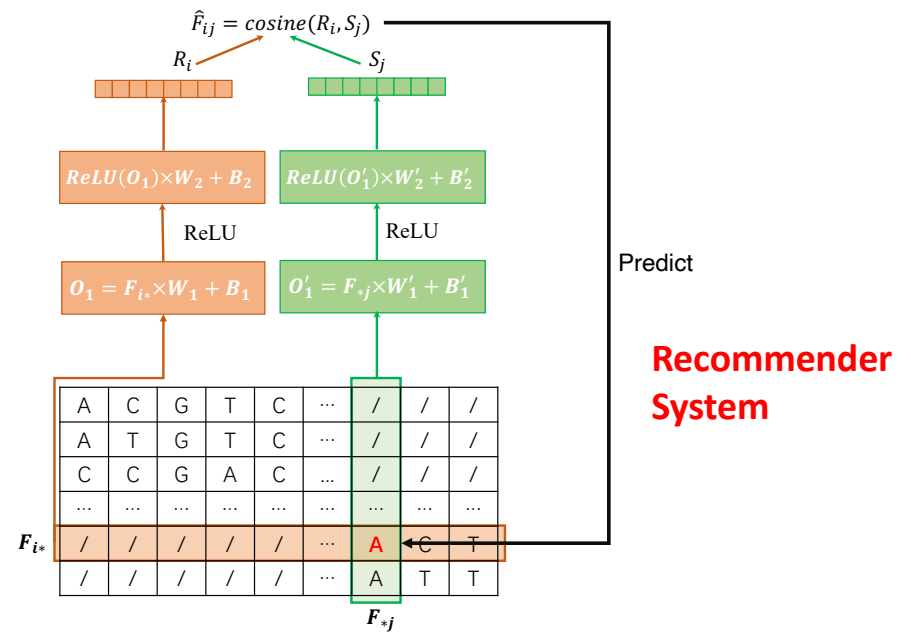
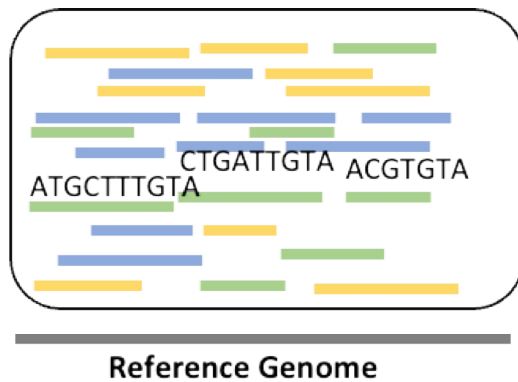


Generic assembly tools cannot produce all haplotypes

Viral haplotype reconstruction from TGS sequencing data



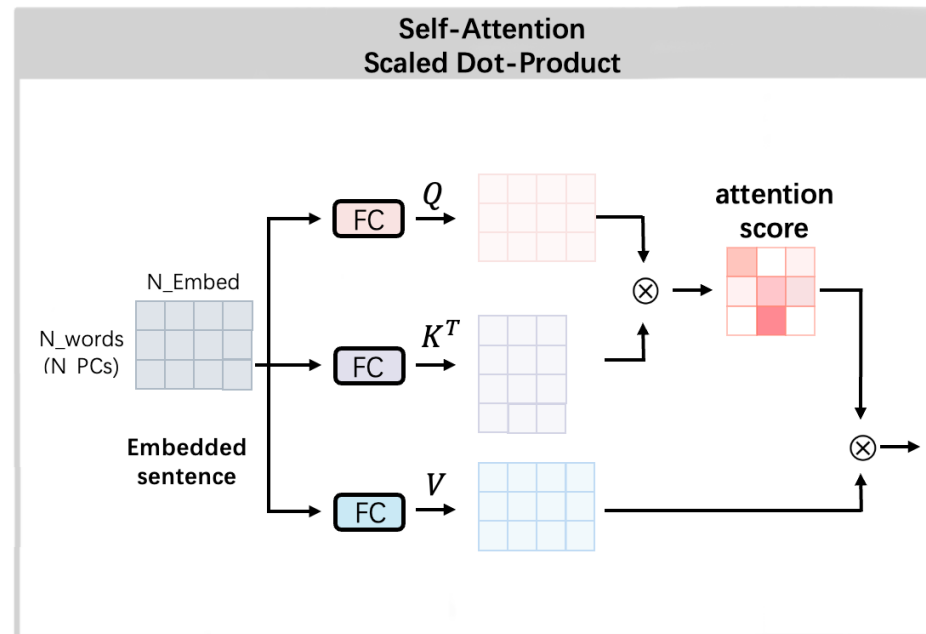
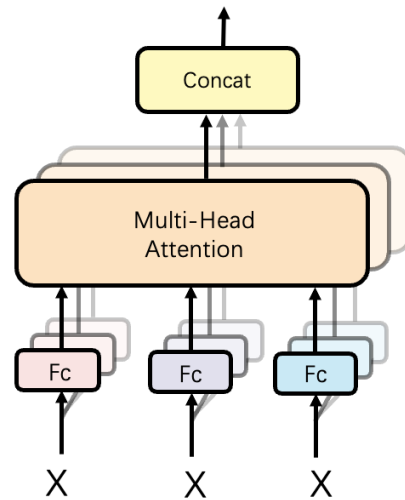
Learning latent features for distinguishing reads from different strains



Dehan Cai, Jiayu Shang, Sun Yanni. HaploDMF: viral Haplotype reconstruction from long reads via Deep Matrix Factorization. *Bioinformatics*. 2022; 29:btac708.

Dehan Cai & Yanni Sun. Reconstructing viral haplotypes using long reads. *Bioinformatics*. 2022; btac089.

Phage Identification with Transformer



- The embedding of the protein sentence will be fed into multi-head attention module
- Value in the attention score matrix represents the strength of associations between two proteins