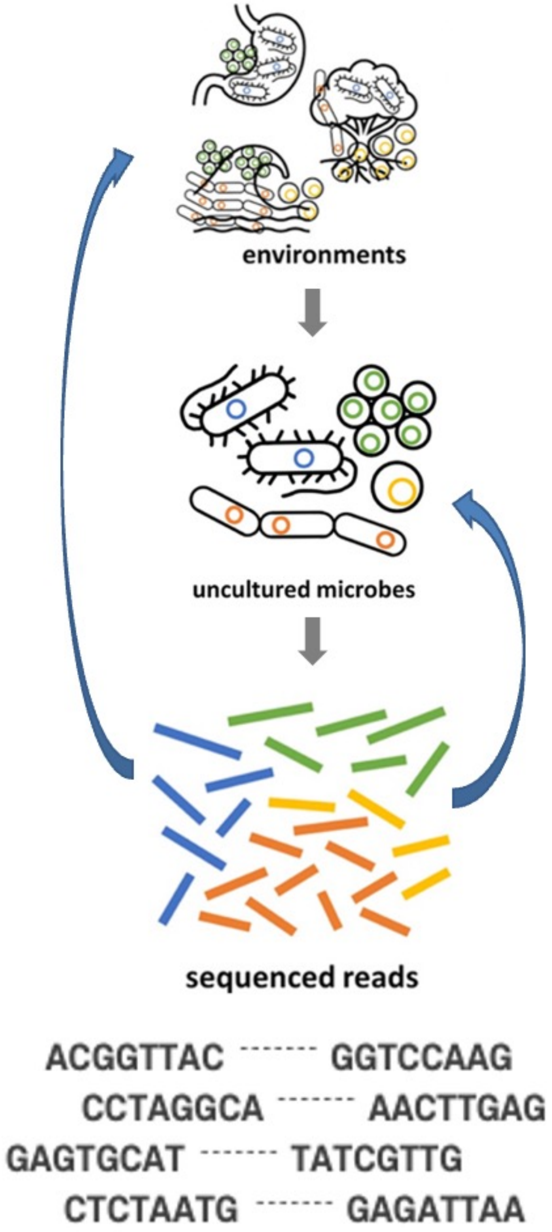
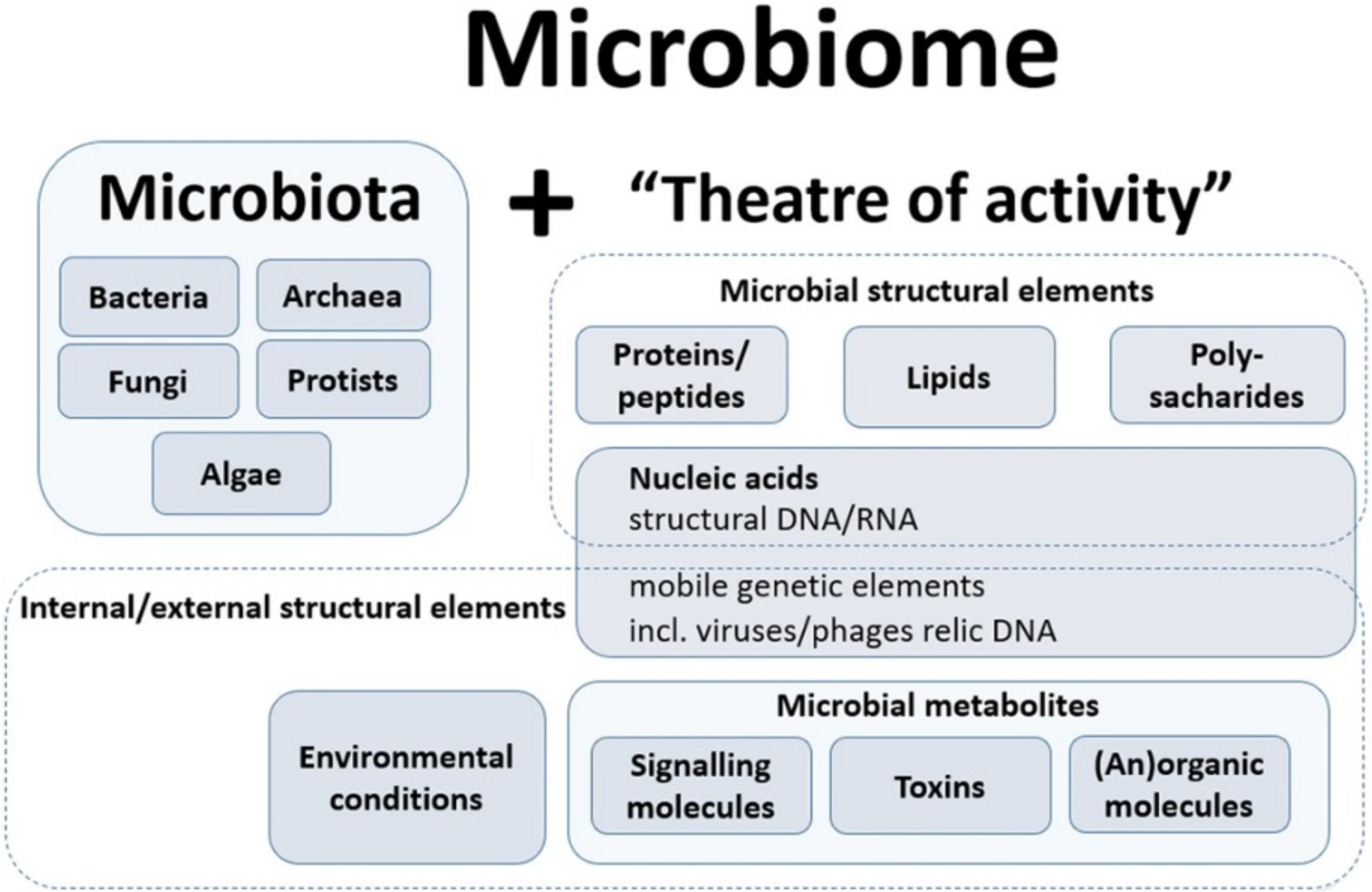
World Ocean Forum 2024

Deep learning approaches for deciphering composition and functional roles of the ocean microbiome

Mina Rho
Hanyang University

Introduction to the microbiome





Microbiome

Microbiota

Bacteria

Archaea

Fungi

Protists

Algae

+ "Theatre of activity"

Microbial structural elements

Proteins/ peptides

Lipids

Poly-sacharides

Nucleic acids

structural DNA/RNA

mobile genetic elements

incl. viruses/phages relic DNA

Internal/external structural elements

Environmental conditions

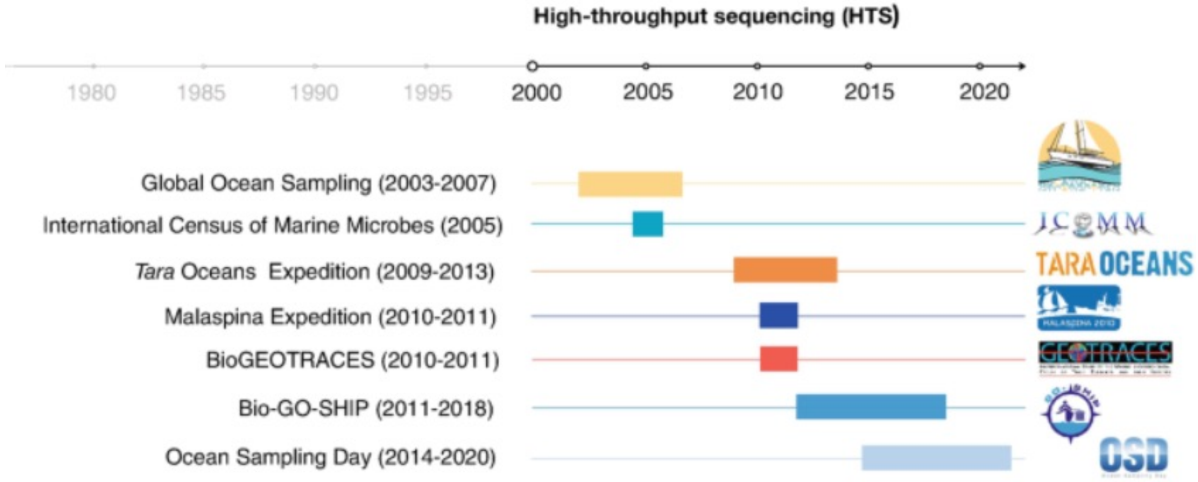
Microbial metabolites

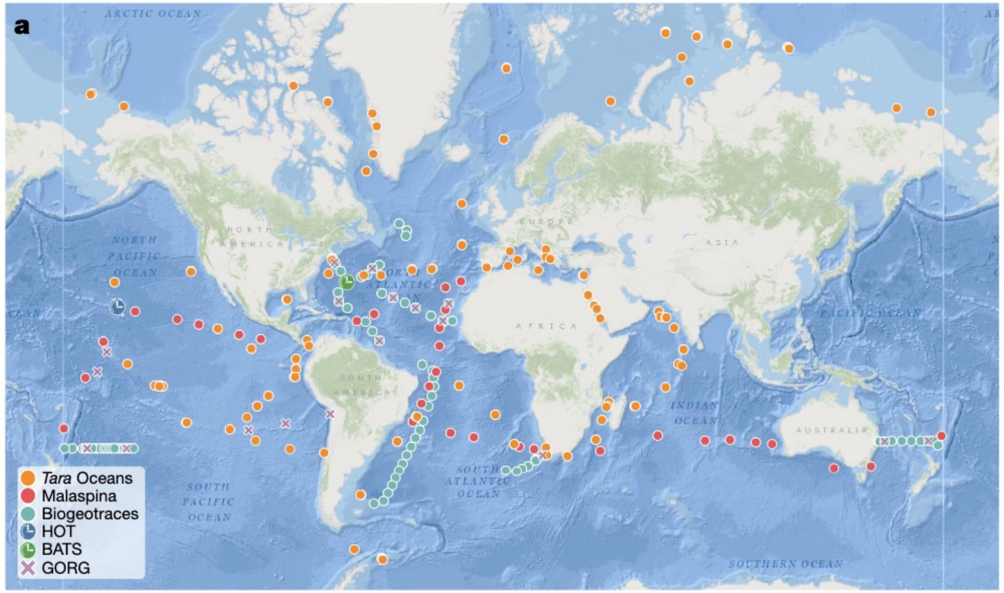
Signalling molecules

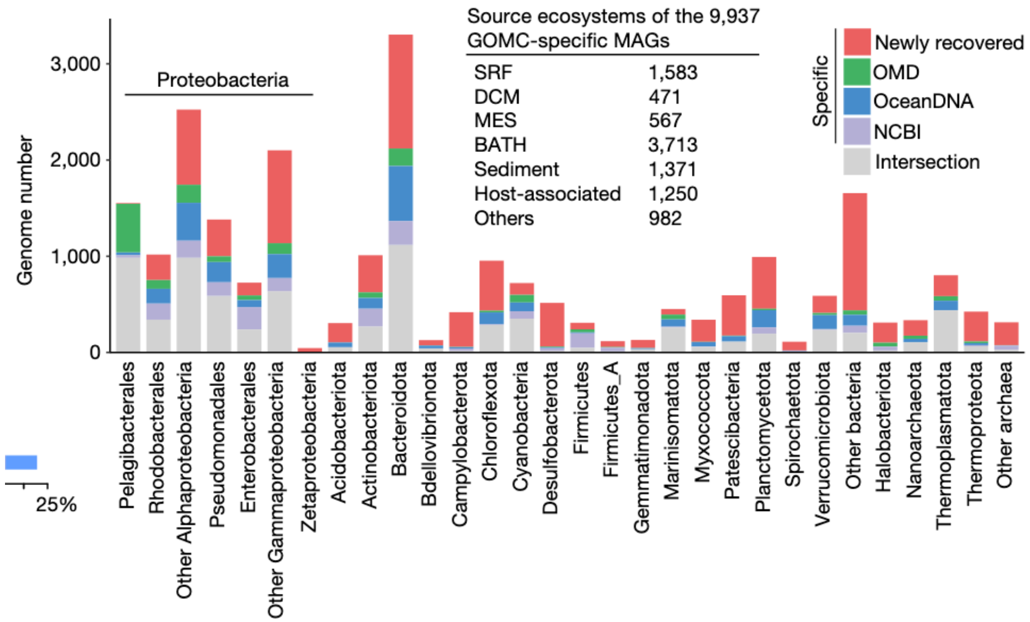
Toxins

(An)organic molecules

Introduction to the marine microbiome

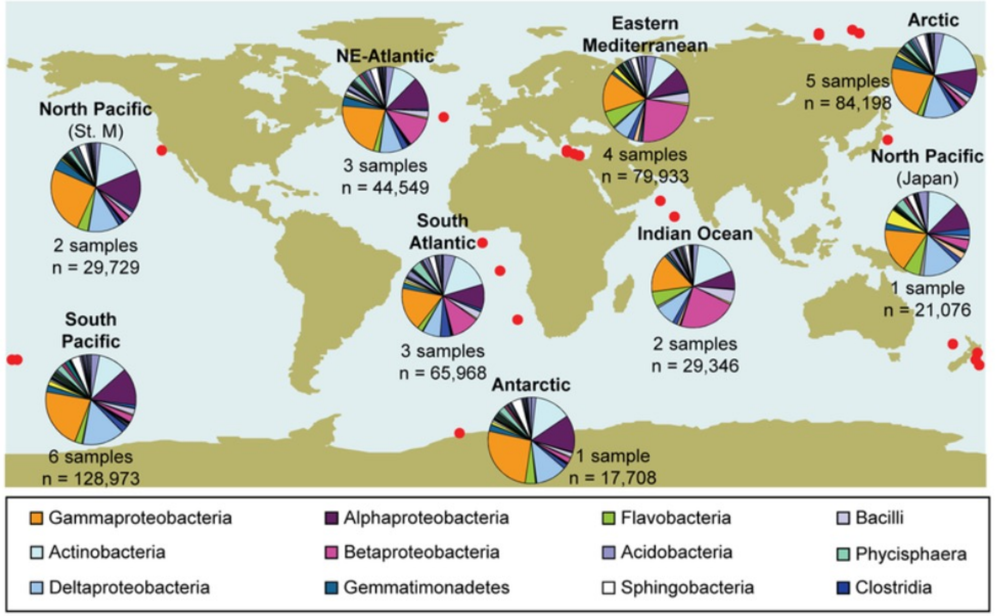






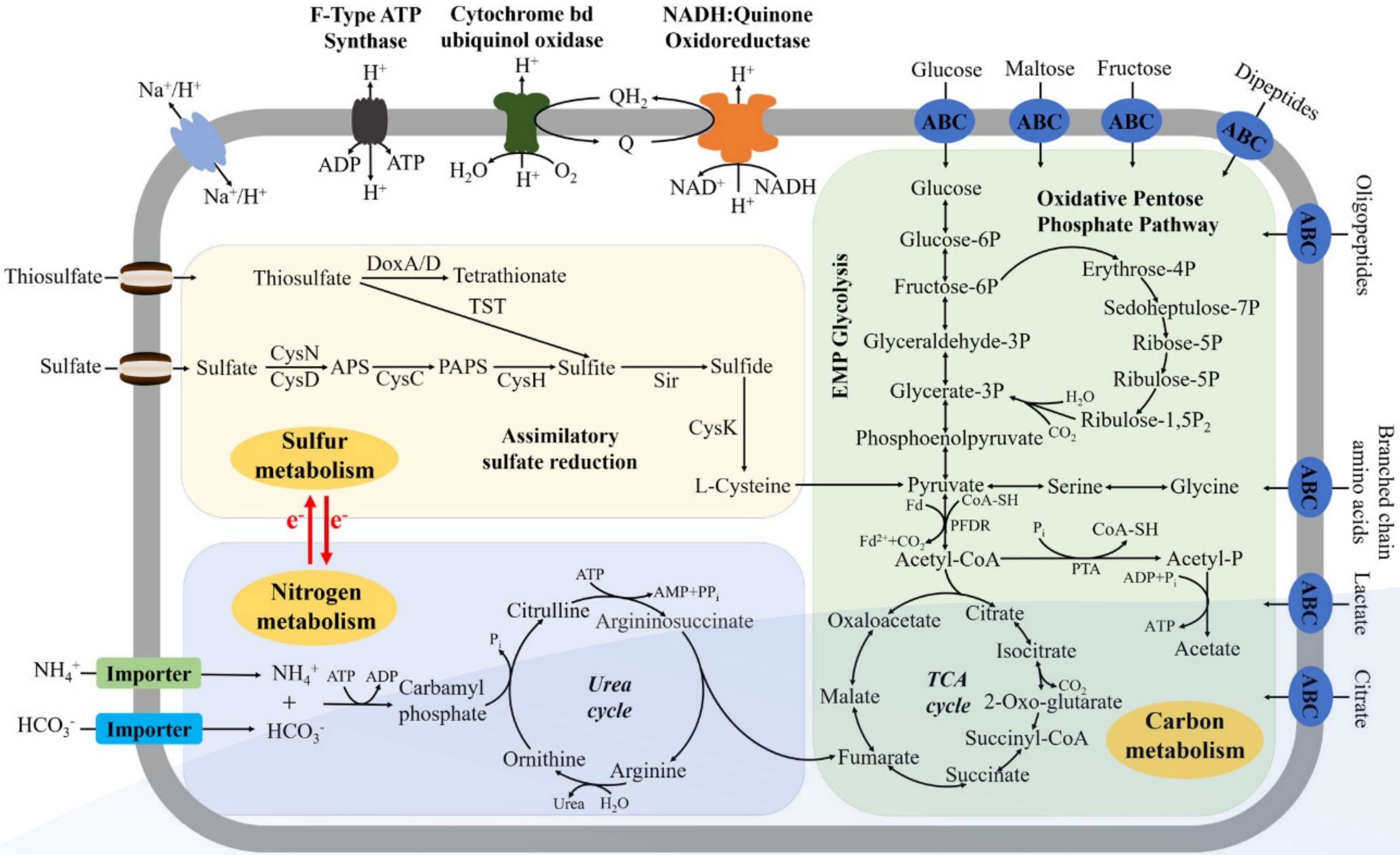
Source ecosystems of the 9,937 GOMC-specific MAGs

SRF	1,583
DCM	471
MES	567
BATH	3,713
Sediment	1,371
Host-associated	1,250
Others	982



Nature 2024

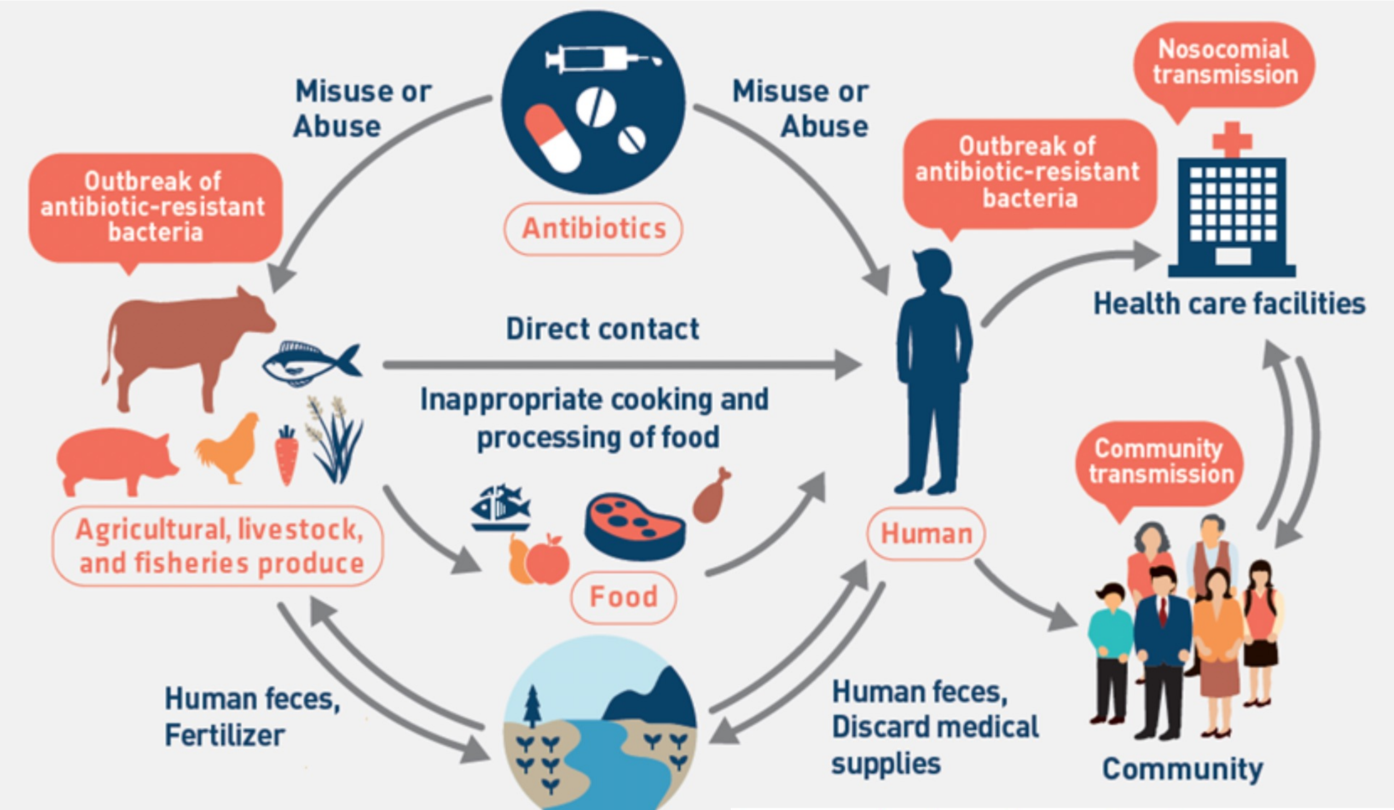
Nutrient cycling in marine microbes



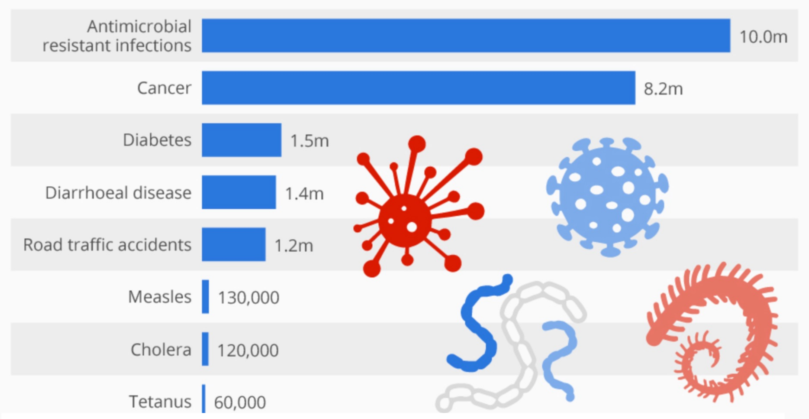
An example of energy production in *Phototrophicus methaneseepsis* ZRK3

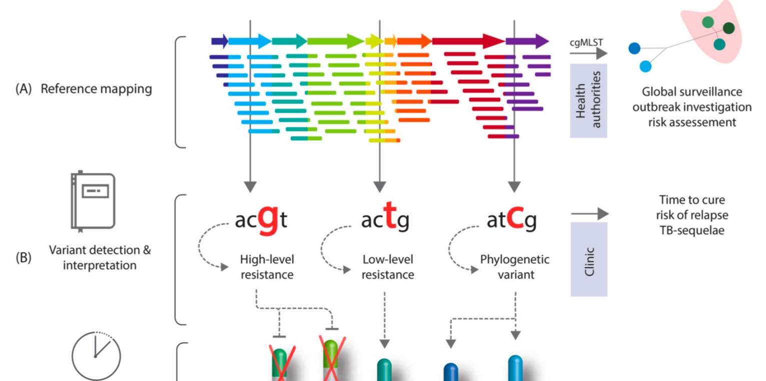
mBio 2024

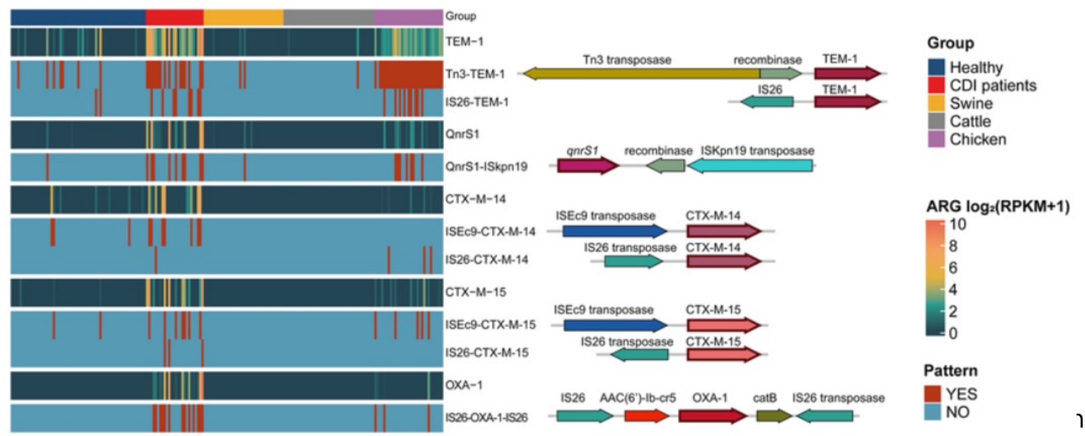
Close relationship with human as one health framework


Misuse or Abuse
Antibiotics
Misuse or Abuse
Nosocomial transmission
Outbreak of antibiotic-resistant bacteria
Outbreak of antibiotic-resistant bacteria
Health care facilities
Direct contact
Inappropriate cooking and processing of food
Agricultural, livestock, and fisheries produce
Food
Human
Community transmission
Human feces, Fertilizer
Human feces, Discard medical supplies
Community
Environment

https://nih.go.kr/

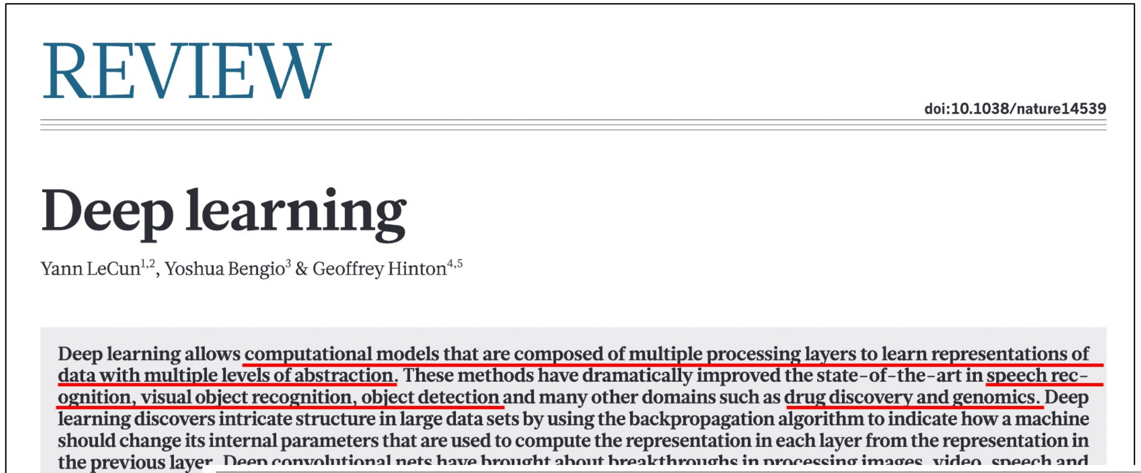






ıt. Microbiol.(2016) Front Cell Infect Microbiol 2023

AIs for biomedical problems



REVIEW

doi:10.1038/nature14539

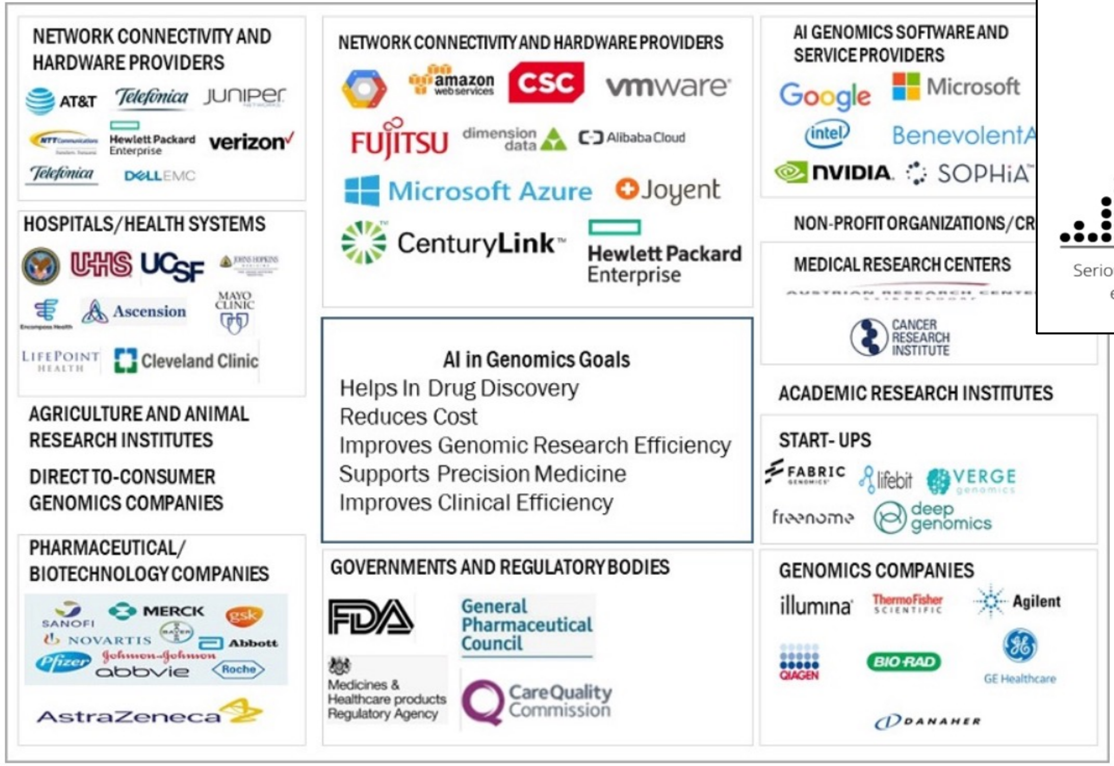
Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas r

[1]Facebook AI Research, 770 Br
Research Université de Montré
94043, USA. [5]Department of C

436 | NATURE | VO



NETWORK CONNECTIVITY AND HARDWARE PROVIDERS

NETWORK CONNECTIVITY AND HARDWARE PROVIDERS

AI GENOMICS SOFTWARE AND SERVICE PROVIDERS

HOSPITALS/HEALTH SYSTEMS

NON-PROFIT ORGANIZATIONS/CR

MEDICAL RESEARCH CENTERS

AI in Genomics Goals

- Helps In Drug Discovery
- Reduces Cost
- Improves Genomic Research Efficiency
- Supports Precision Medicine
- Improves Clinical Efficiency

ACADEMIC RESEARCH INSTITUTES

AGRICULTURE AND ANIMAL RESEARCH INSTITUTES

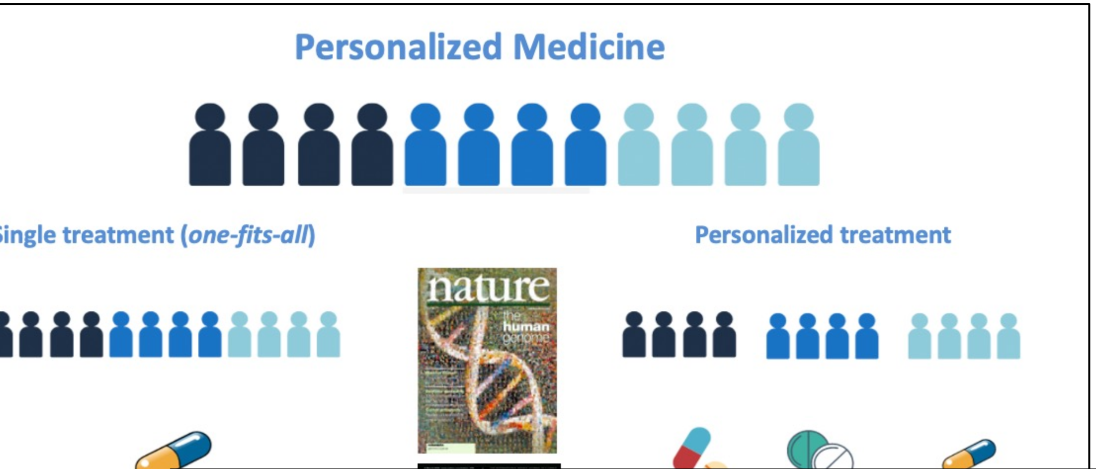
DIRECT TO-CONSUMER GENOMICS COMPANIES

START- UPS

PHARMACEUTICAL/ BIOTECHNOLOGY COMPANIES

GOVERNMENTS AND REGULATORY BODIES

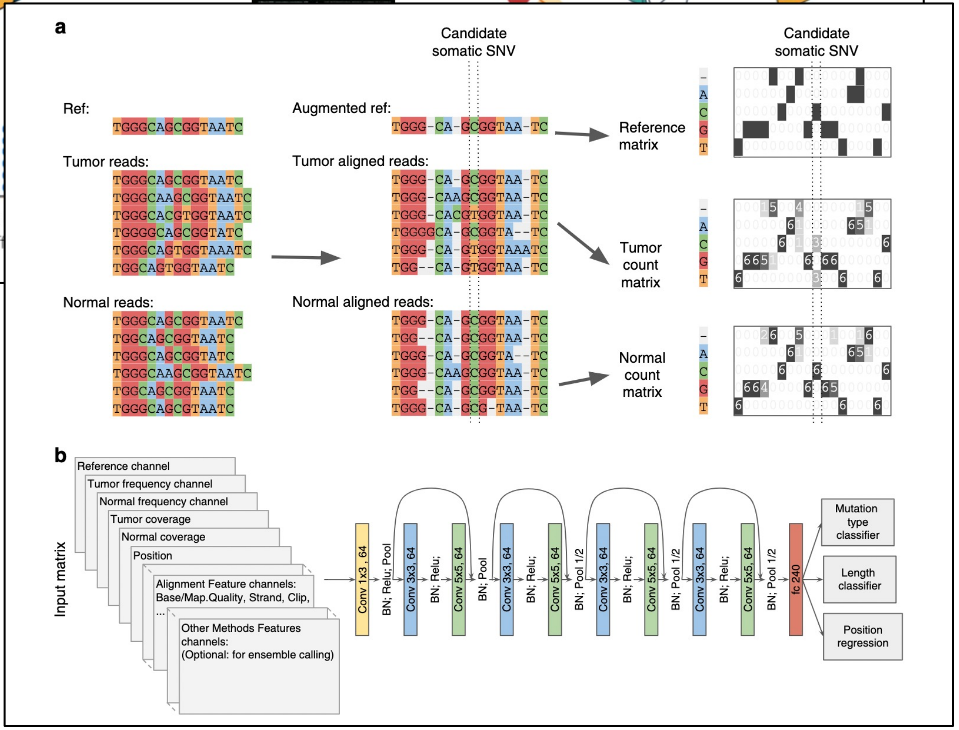
GENOMICS COMPANIES



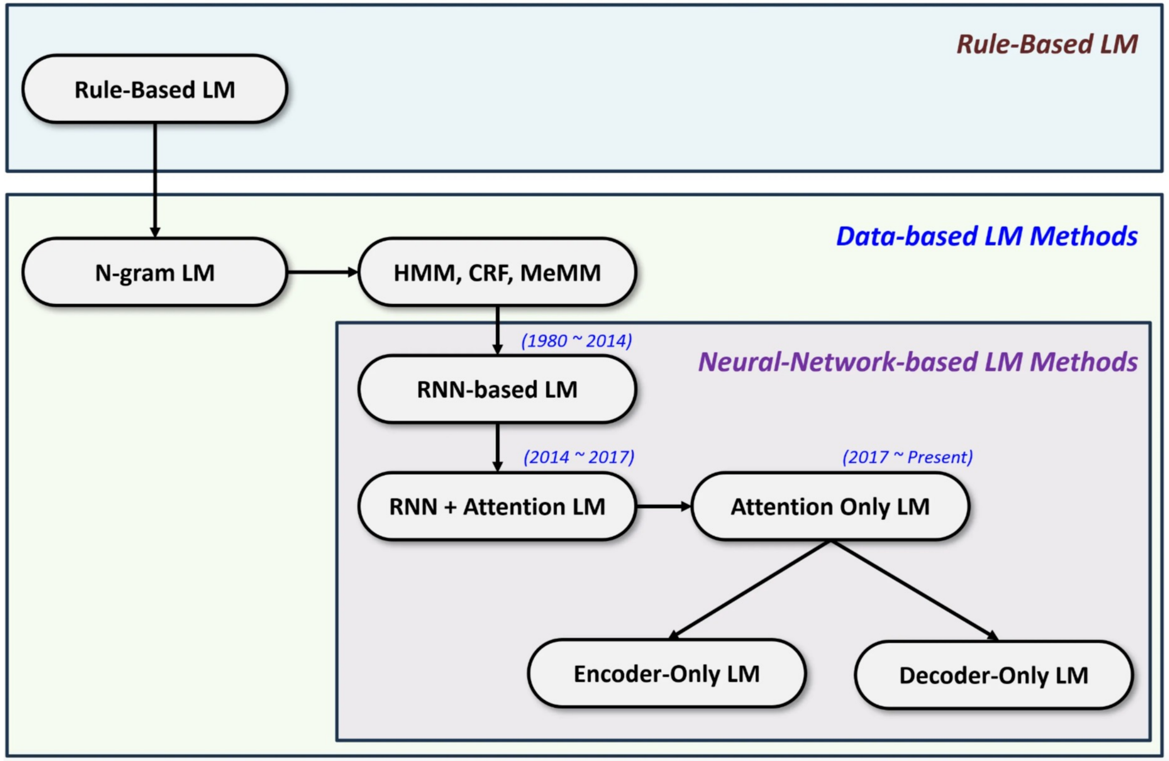
Personalized Medicine

Single treatment (*one-fits-all*)

Personalized treatment



Large language models



Rule-Based LM

Rule-Based LM

Data-based LM Methods

N-gram LM

HMM, CRF, MeMM

Neural-Network-based LM Methods

(1980 ~ 2014)

RNN-based LM

(2014 ~ 2017)

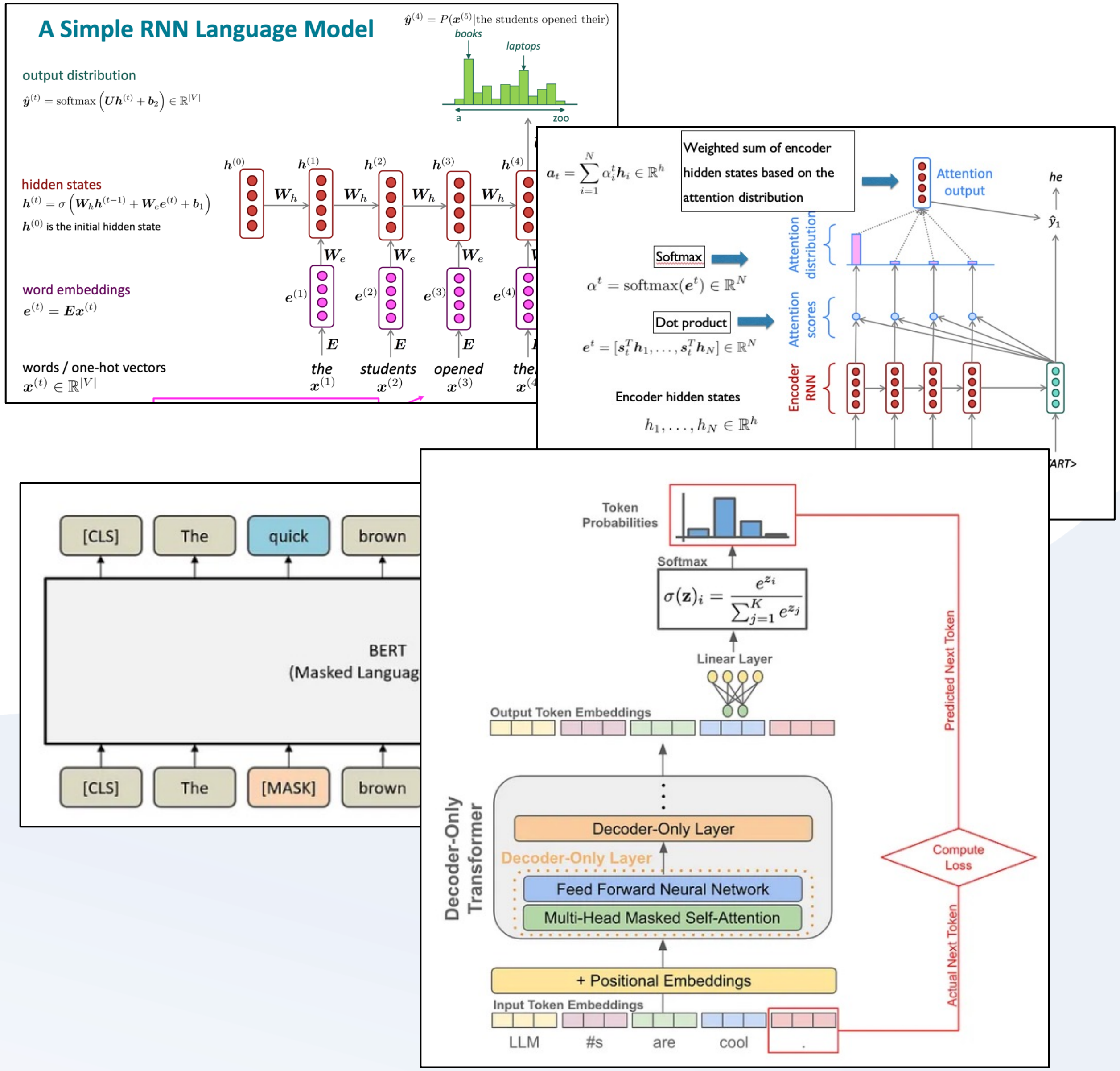
RNN + Attention LM

(2017 ~ Present)

Attention Only LM

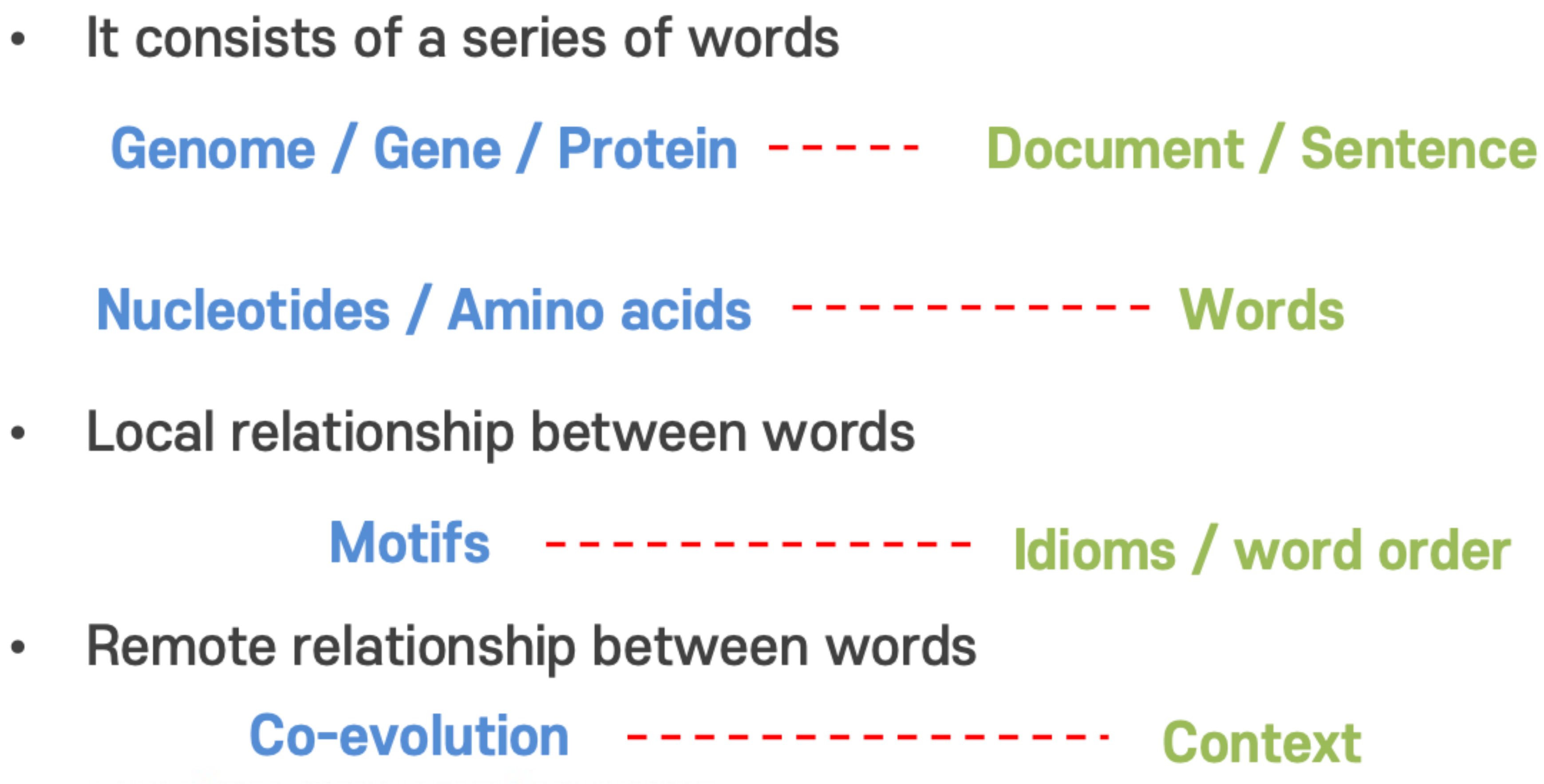
Encoder-Only LM

Decoder-Only LM



A Simple RNN Language Model
$\hat{y}^{(4)} = P(x^{(5)}|\text{the students opened their})$
books
laptops
a
zoo
output distribution
$\hat{y}^{(t)} = \text{softmax}\left(Uh^{(t)} + b_2\right) \in \mathbb{R}^{|V|}$
hidden states
$h^{(t)} = \sigma\left(W_h h^{(t-1)} + W_e e^{(t)} + b_1\right)$
$h^{(0)}$ is the initial hidden state
word embeddings
$e^{(t)} = Ex^{(t)}$
words / one-hot vectors
$x^{(t)} \in \mathbb{R}^{|V|}$
$h^{(0)}$ $h^{(1)}$ $h^{(2)}$ $h^{(3)}$ $h^{(4)}$
W_h W_e E
$e^{(1)}$ $e^{(2)}$ $e^{(3)}$ $e^{(4)}$
the
students
opened
the
$x^{(1)}$ $x^{(2)}$ $x^{(3)}$ $x^{(4)}$
$a_t = \sum_{i=1}^{N} \alpha_i^t h_i \in \mathbb{R}^h$
Weighted sum of encoder hidden states based on the attention distribution
Attention output
Softmax
$\alpha^t = \text{softmax}(e^t) \in \mathbb{R}^N$
Dot product
$e^t = [s_t^T h_1, \ldots, s_t^T h_N] \in \mathbb{R}^N$
Encoder hidden states
$h_1, \ldots, h_N \in \mathbb{R}^h$
Attention distribution
Attention scores
Encoder RNN
he
$\hat{y}_1$
ART>
[CLS]
The
quick
brown
BERT
(Masked Languag
[CLS]
The
[MASK]
brown
Token Probabilities
Softmax
$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
Linear Layer
Output Token Embeddings
Decoder-Only Transformer
Decoder-Only Layer
Decoder-Only Layer
Feed Forward Neural Network
Multi-Head Masked Self-Attention
+ Positional Embeddings
Input Token Embeddings
LLM
#s
are
cool
.
Predicted Next Token
Compute Loss
Actual Next Token

Large language models to biological sequence data



- It consists of a series of words

Genome / Gene / Protein ------ Document / Sentence

Nucleotides / Amino acids ------------- Words

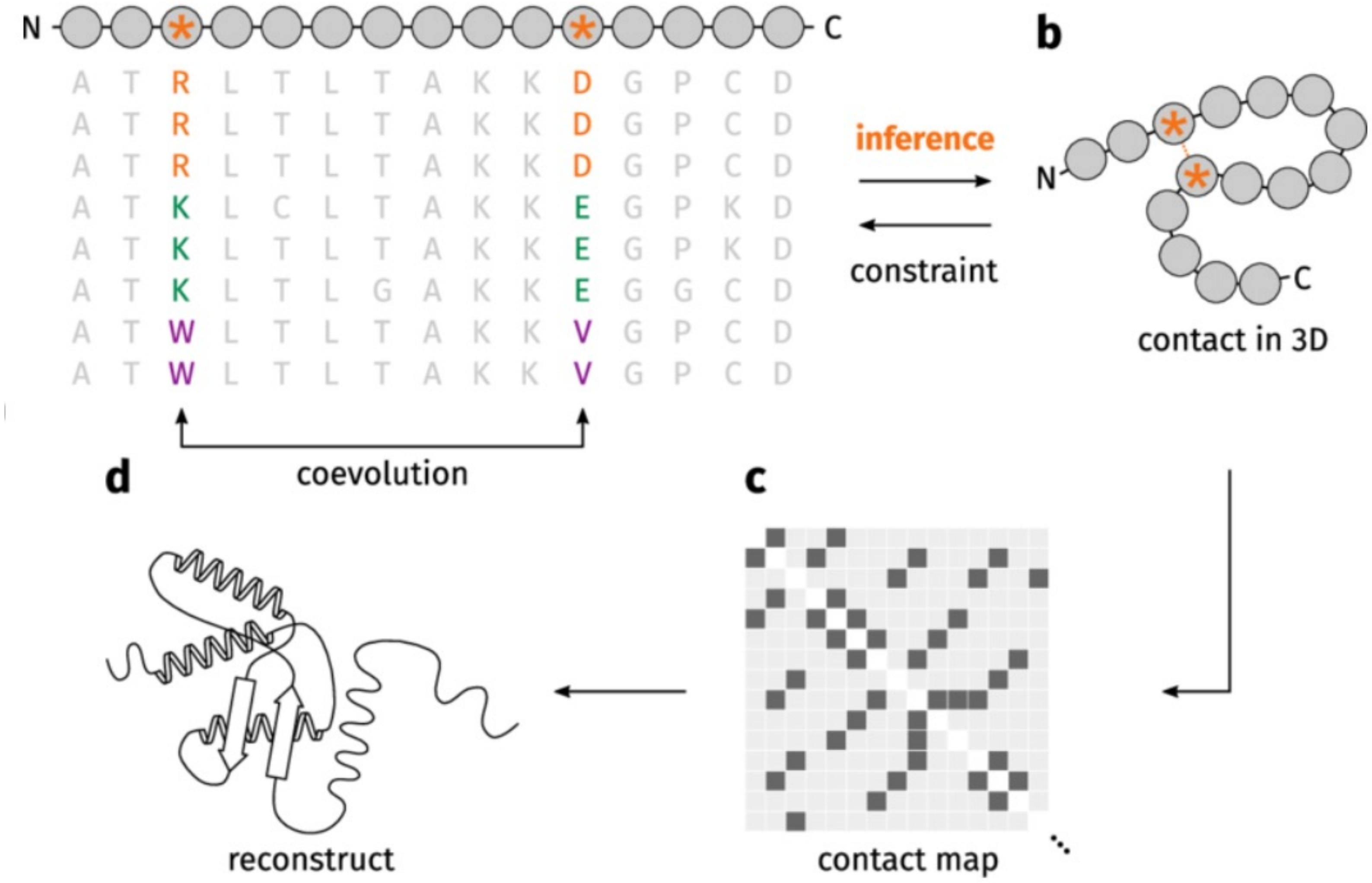
- Local relationship between words

Motifs -------------- Idioms / word order

- Remote relationship between words

Co-evolution ---------------- Context

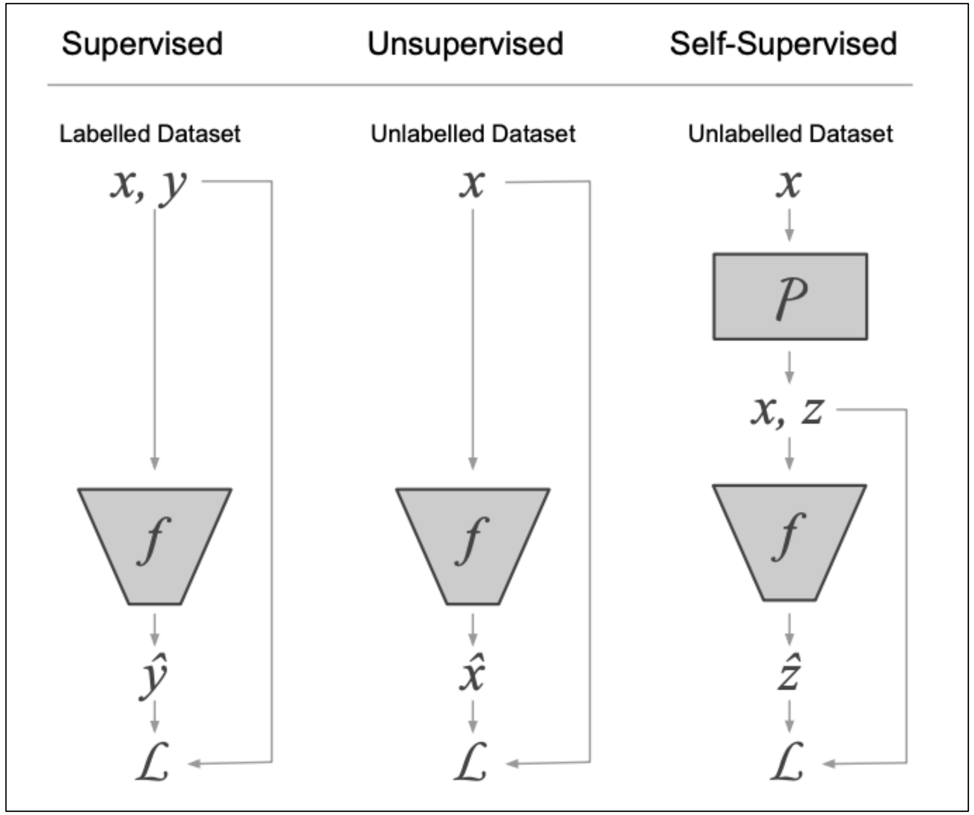
Large language models to biological sequence data



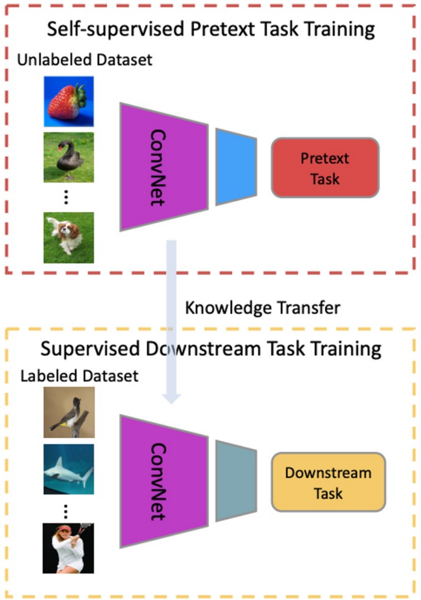
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

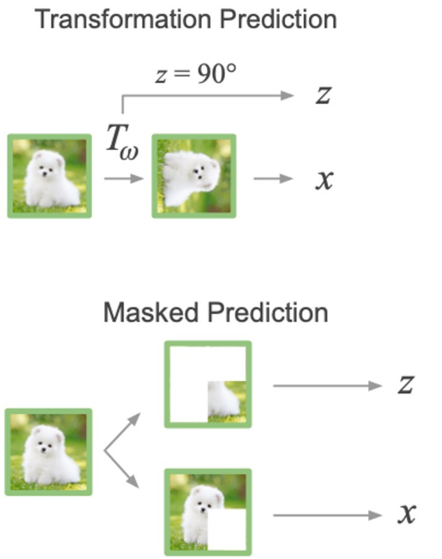
Nature 2021

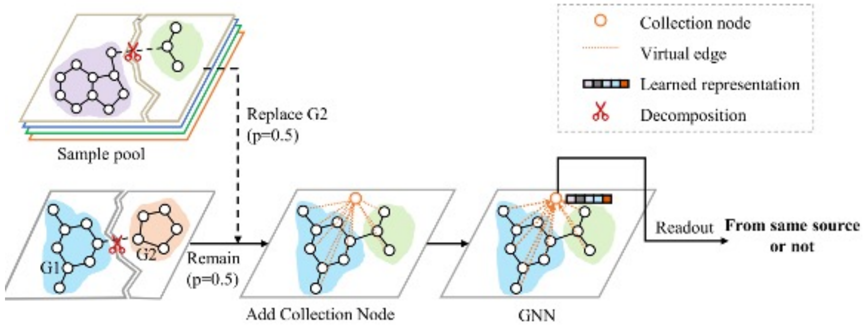
Self-supervised learning to biological sequencing data

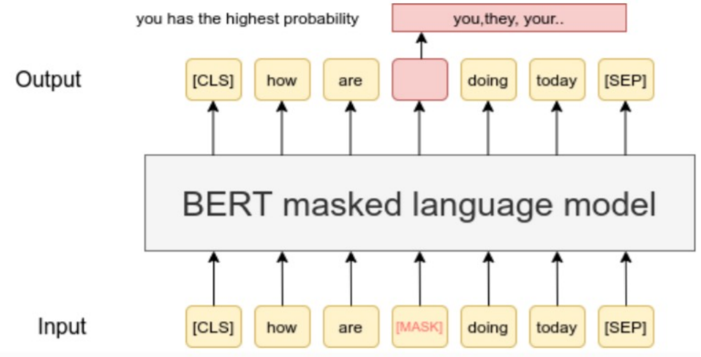


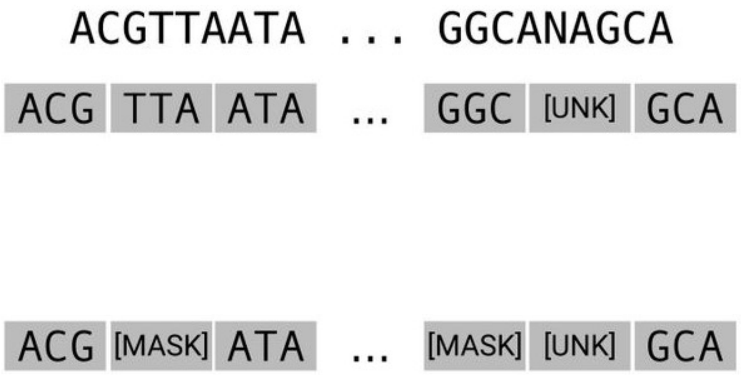
Pretext task



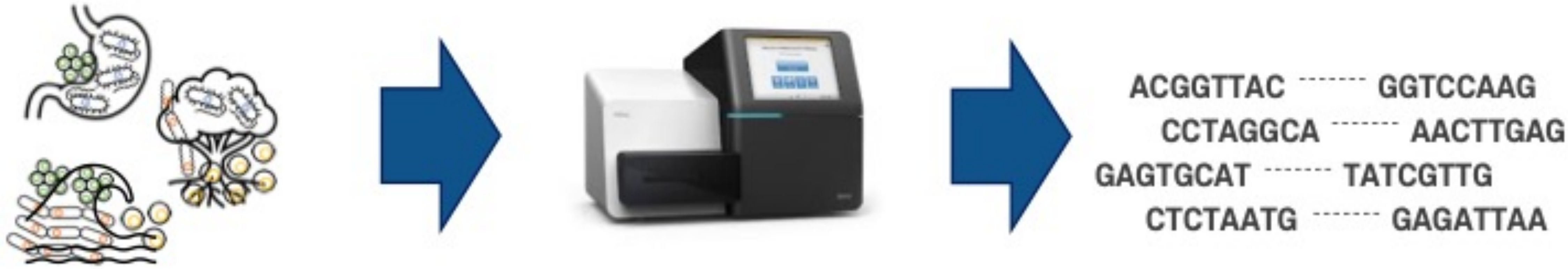






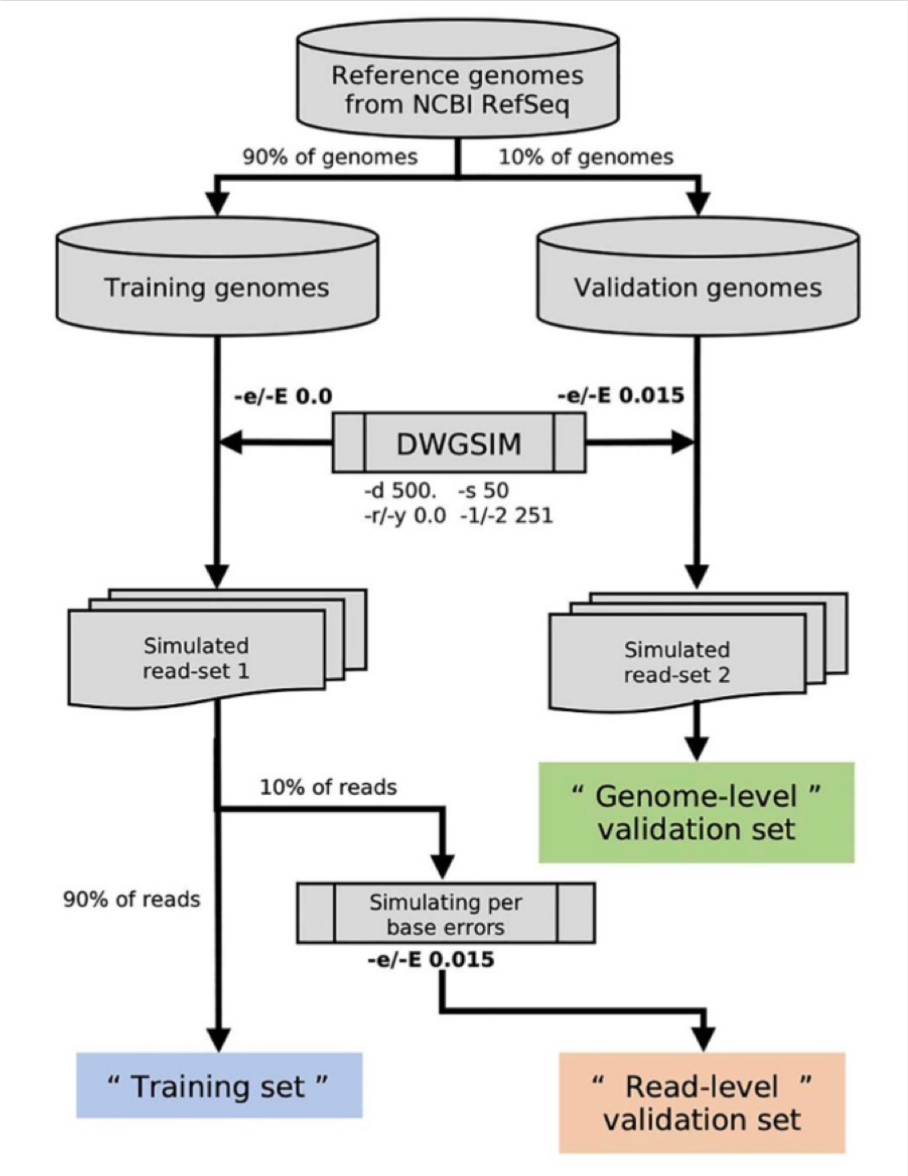


1. Taxonomy classification from sequencing read sequence


ACGGTTAC ------- GGTCCAAG
CCTAGGCA ------- AACTTGAG
GAGTGCAT ------- TATCGTTG
CTCTAATG ------- GAGATTAA

- Using sequence fragments of 150 -250 bps, the sequence should be assigned (classified) to one of the taxa at a certain taxonomy level

Taxonomy classification from sequencing read sequence



10,119 viral genomes

2,293 eukaryotic DNA viruses

2,733 eukaryotic RNA viruses

5,529 bacterial genomes

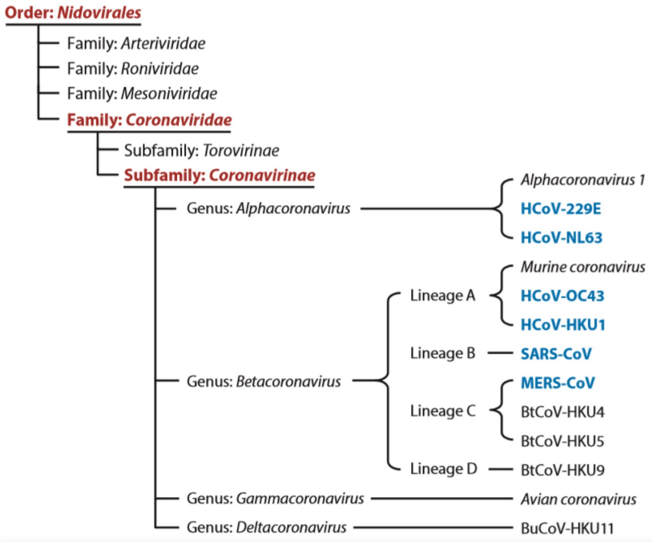
(one genome for each species)

Eukaryote-related viruses were retained

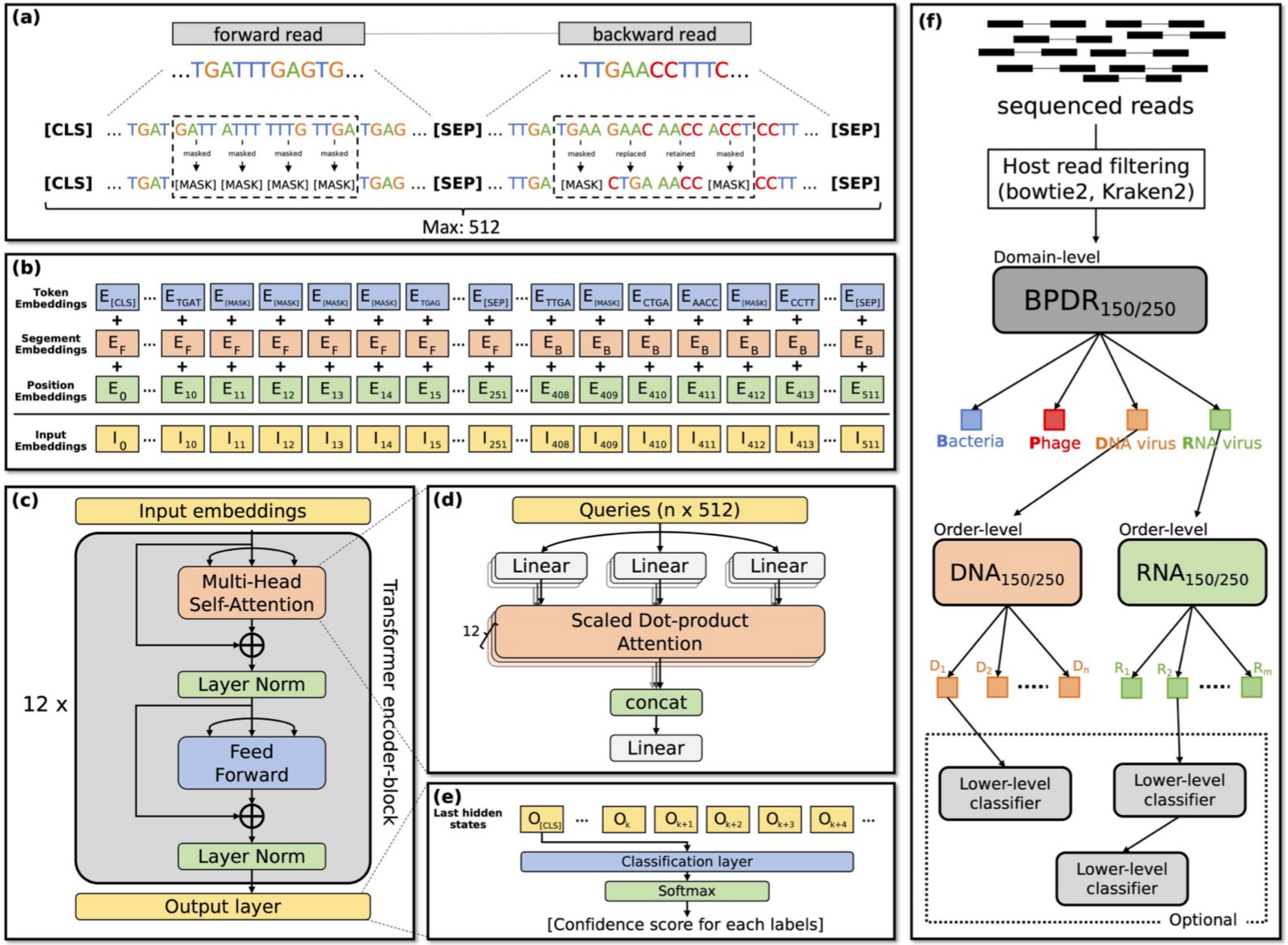
(Based on ICTV annotation)

The length of reads was set to 151 and 251

Insert size from 300 to 800



VIBE: Taxonomy classification from sequencing read sequence

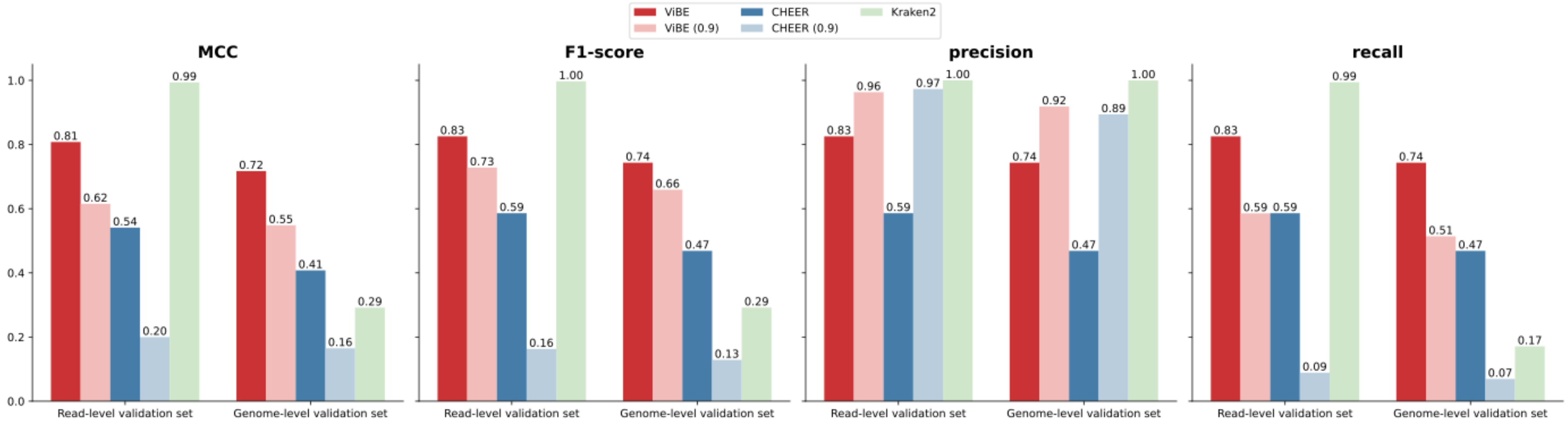


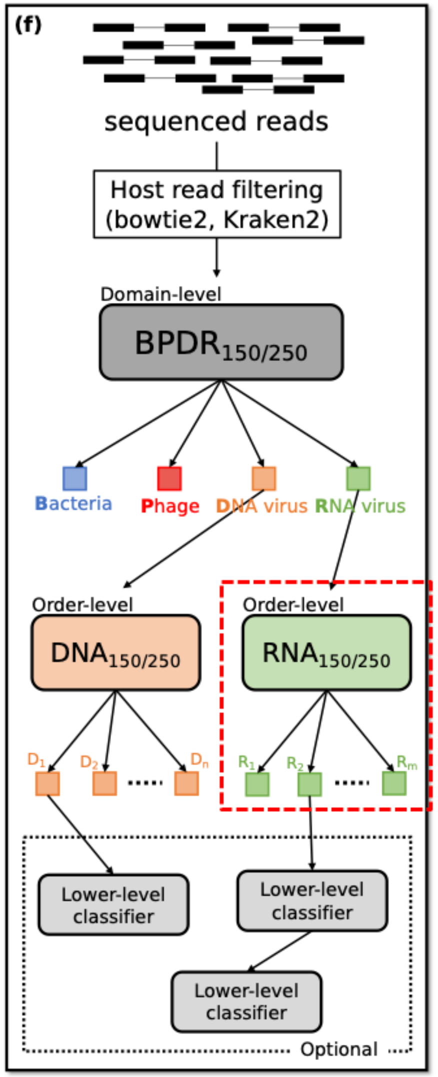
Brief Bioinfo 2022

VIBE: Taxonomy classification from sequencing read sequence

Order-level classifier for RNA viruses

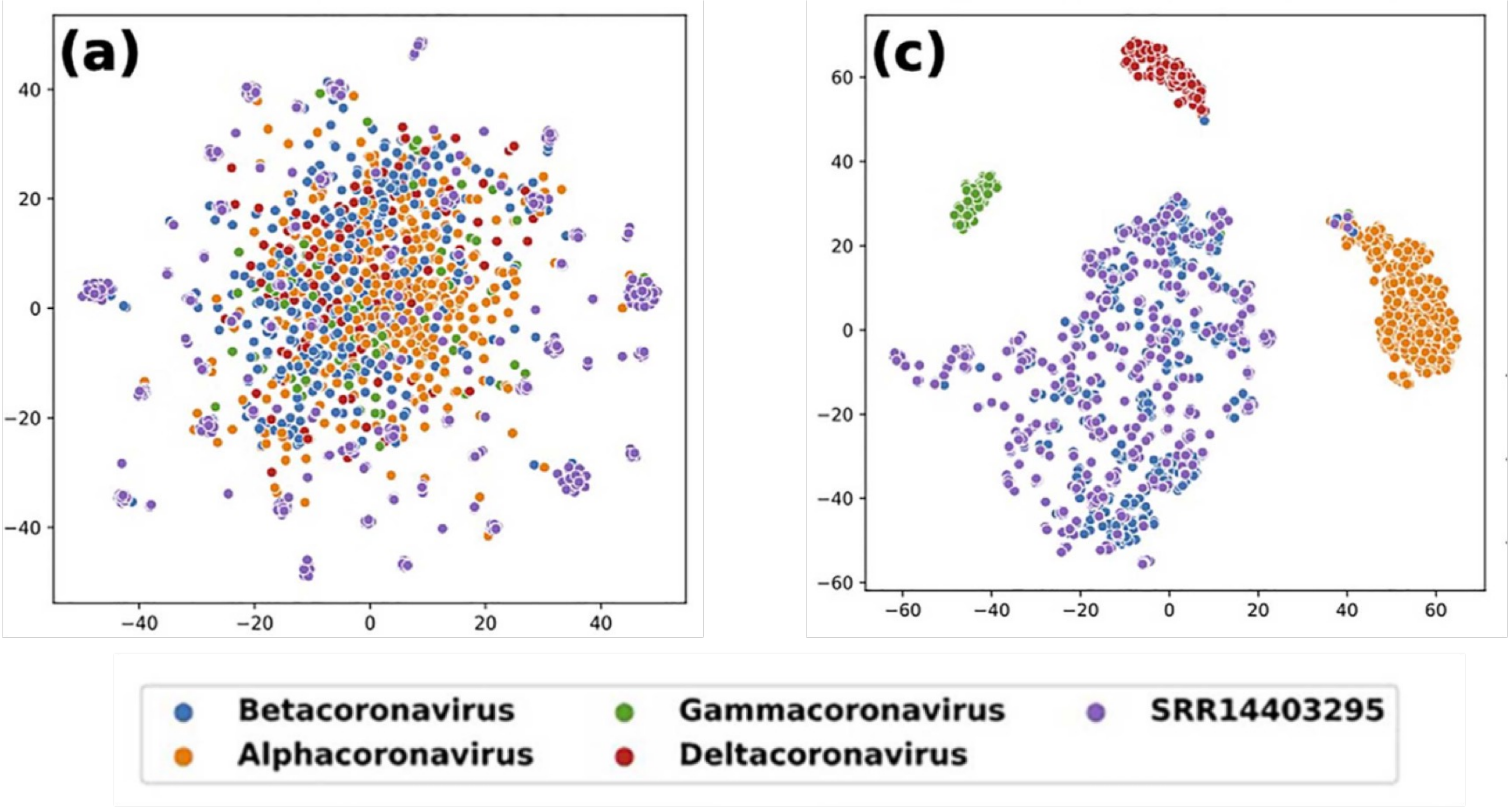
- SOTA methods:
  - CHEER / Skip-Gram + parallel CNN model
  - Kraken2 / *k*-mer homology method
- Test data:
  - Read-level validation set
  - Genome-level validation set





Brief Bioinfo 2022

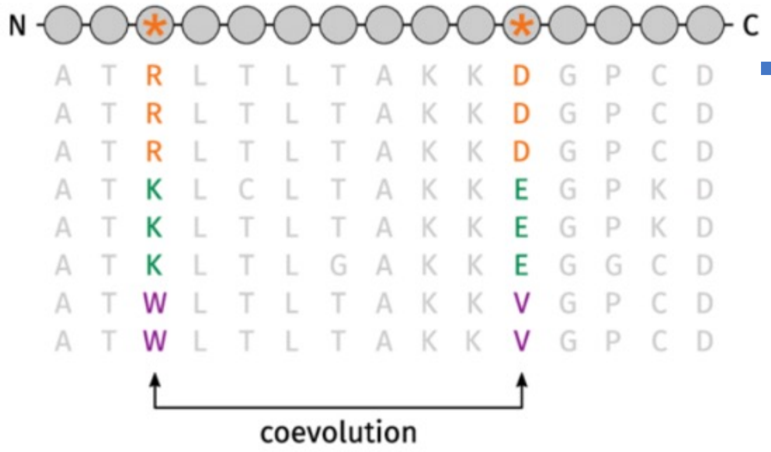
VIBE: Taxonomy classification from sequencing read sequence



- The ViBE was re-trained without the SARS-CoV-2 reference genome
- COVID-19 samples were tested with the model
- (A) 4-mer frequency, (C) embedded vector by fine-tuned model

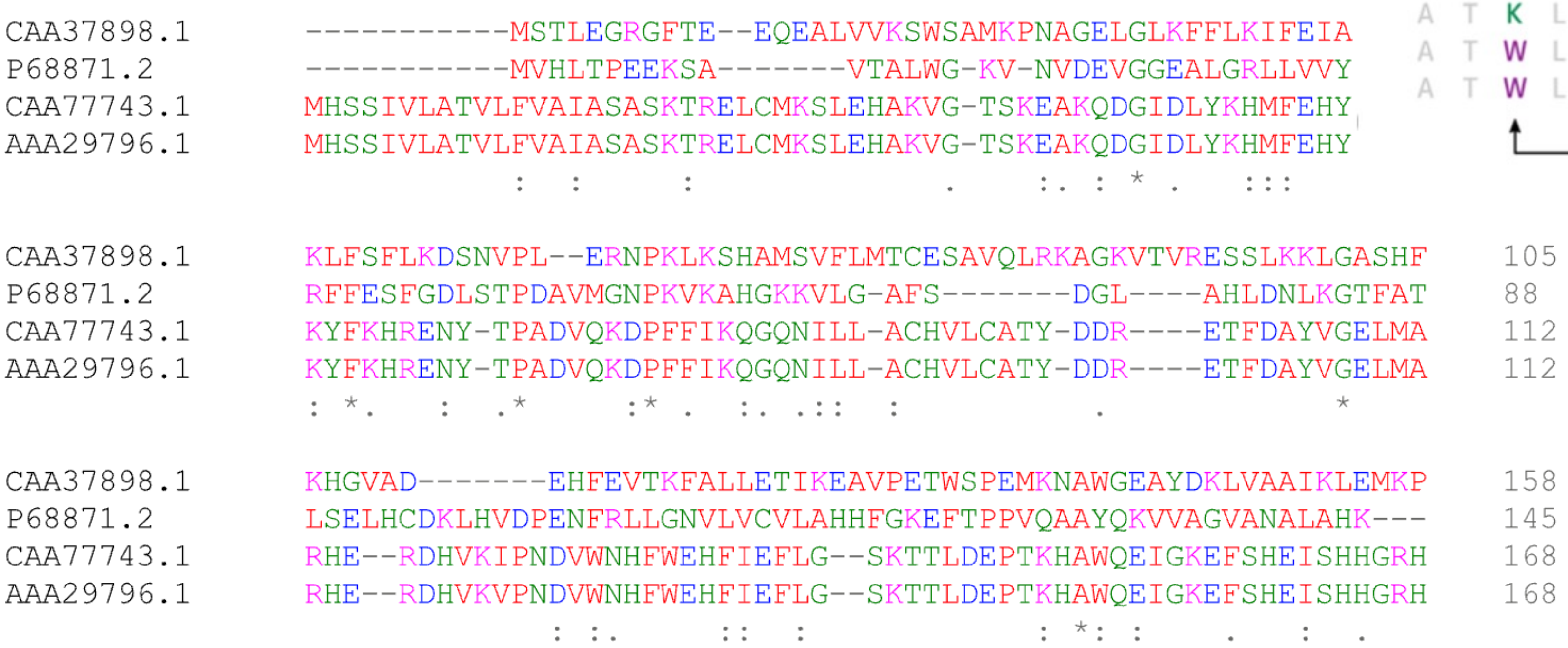
Brief Bioinfo 2022

2. Function prediction from protein sequences



```
N                             C
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Four different hemoglobin protein sequences



```
CAA37898.1      -----------MSTLEGRGFTE--EQEALVVKSWSAMKPNAGELGLKFFLKIFEIA
P68871.2        -----------MVHLTPEEKSA-------VTALWG-KV-NVDEVGGEALGRLLVVY
CAA77743.1      MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVG-TSKEAKQDGIDLYKHMFEHY
AAA29796.1      MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVG-TSKEAKQDGIDLYKHMFEHY
                           :  :     :             .    :. : * .   :::

CAA37898.1      KLFSFLKDSNVPL--ERNPKLKSHAMSVFLMTCESAVQLRKAGKVTVRESSLKKLGASHF    105
P68871.2        RFFESFGDLSTPDAVMGNPKVKAHGKKVLG-AFS-------DGL----AHLDNLKGTFAT    88
CAA77743.1      KYFKHRENY-TPADVQKDPFFIKQGQNILL-ACHVLCATY-DDR----ETFDAYVGELMA    112
AAA29796.1      KYFKHRENY-TPADVQKDPFFIKQGQNILL-ACHVLCATY-DDR----ETFDAYVGELMA    112
                : *.   :  .*     :* .  :. .::  :          .            *

CAA37898.1      KHGVAD-------EHFEVTKFALLETIKEAVPETWSPEMKNAWGEAYDKLVAAIKLEMKP    158
P68871.2        LSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHK---    145
CAA77743.1      RHE--RDHVKIPNDVWNHFWEHFIEFLG--SKTTLDEPTKHAWQEIGKEFSHEISHHGRH    168
AAA29796.1      RHE--RDHVKVPNDVWNHFWEHFIEFLG--SKTTLDEPTKHAWQEIGKEFSHEISHHGRH    168
                             : :.     ::  :            : *: :    .   :  .
```

- There is no hard sequence similarity threshold for “safe” function prediction
- Sequences that are more than 30-40% identical are considered as having the same or a very similar function

His-Me finger endonucleases

His-Me finger nuclease

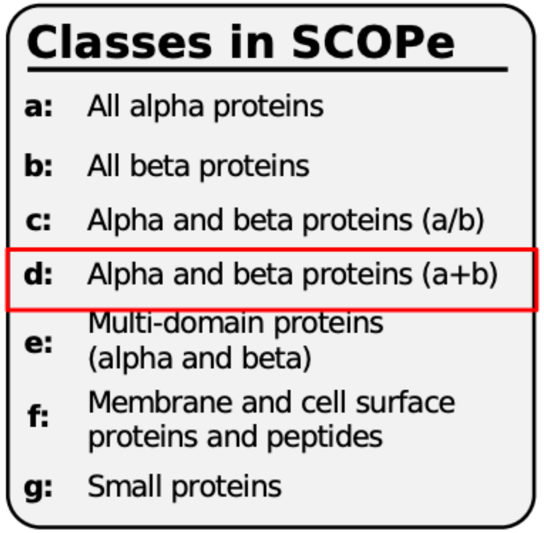
- Conserved Histidine residue (His)
- Catalytic metal ion (Me)
- Finger-like structure (finger)

Example

- Cas9 enzyme
  in the CRISPR-Cas9 genome editing technology

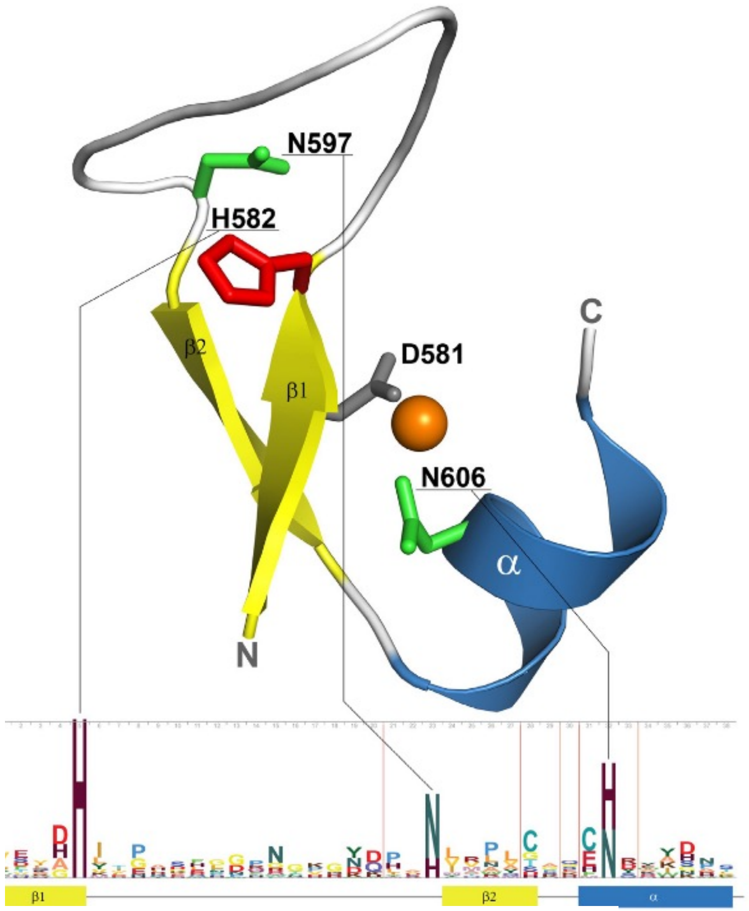
Capturing conserved residues and secondary structure

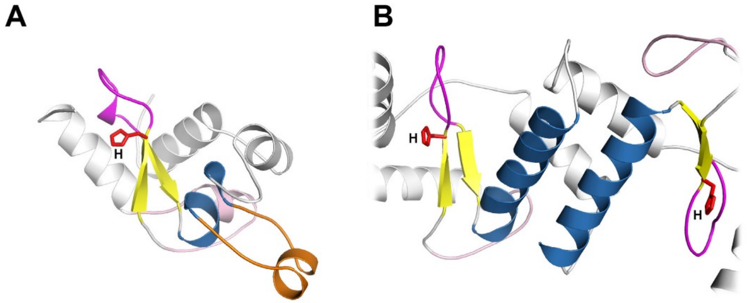
is essential to classify His-Me finger nuclease



Classes in SCOPe

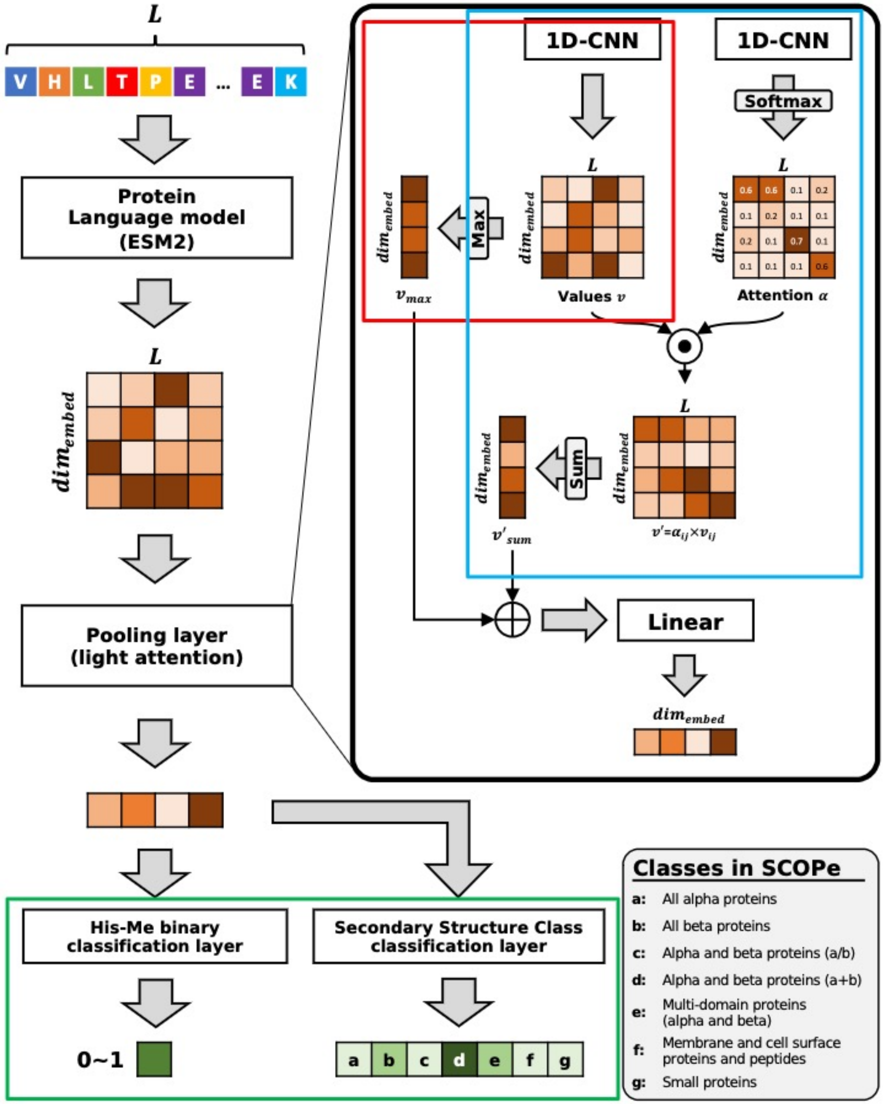
a:	All alpha proteins
b:	All beta proteins
c:	Alpha and beta proteins (a/b)
d:	Alpha and beta proteins (a+b)
e:	Multi-domain proteins (alpha and beta)
f:	Membrane and cell surface proteins and peptides
g:	Small proteins





NAR 2017

FuncPred: Function prediction from protein sequences



Classes in SCOPe

a: All alpha proteins
b: All beta proteins
c: Alpha and beta proteins (a/b)
d: Alpha and beta proteins (a+b)
e: Multi-domain proteins (alpha and beta)
f: Membrane and cell surface proteins and peptides
g: Small proteins

Capturing conserved residues and secondary structure is essential to classify His-Me finger nuclease

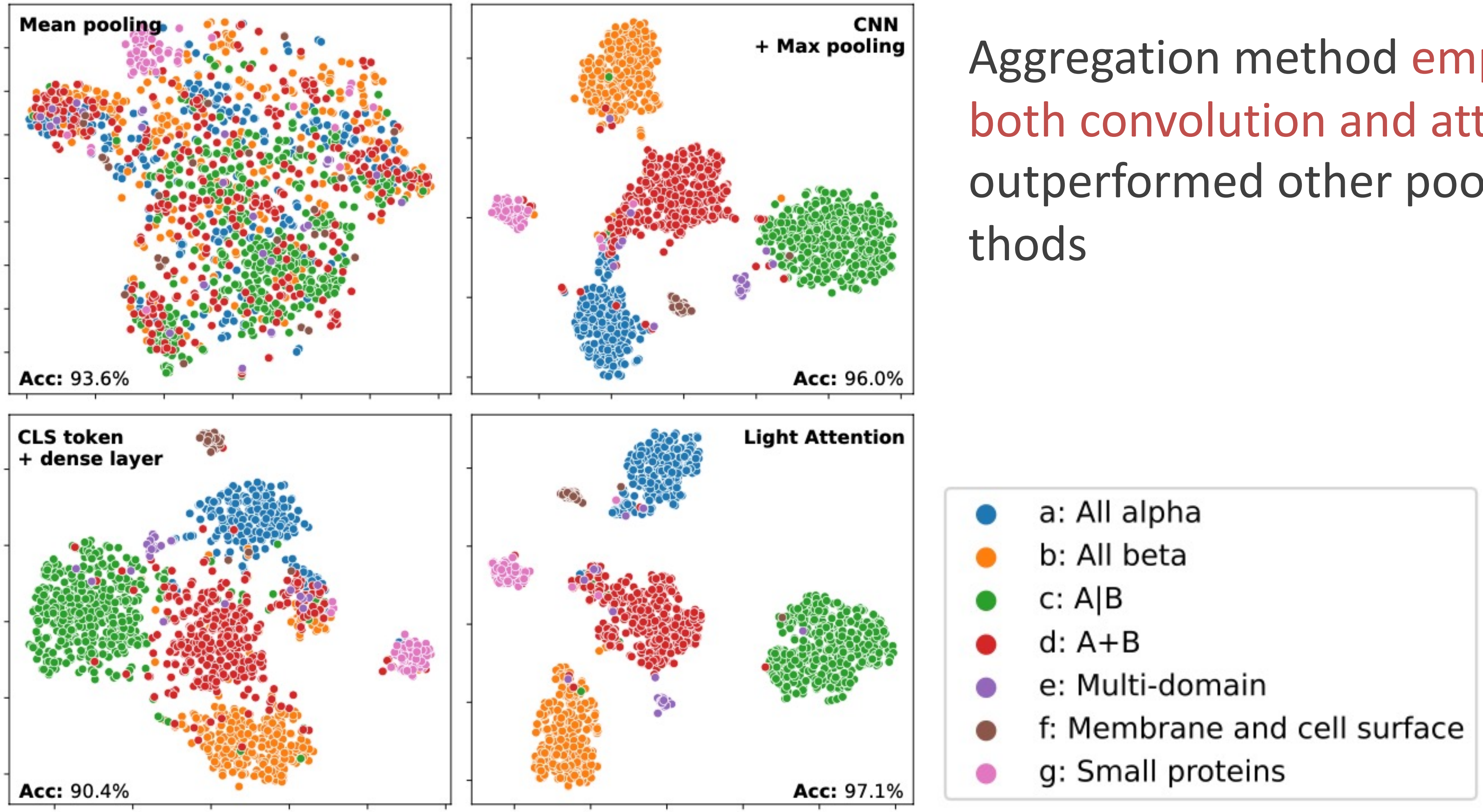
Pooling: light attention

- **1D-CNN + max pooling**: capturing secondary structure
- **Convolution x attention**: capturing conserved residues

Using sequence-level representation, binary classification for detecting His-Me finger nuclease and secondary structure class classification were performed simultaneously.

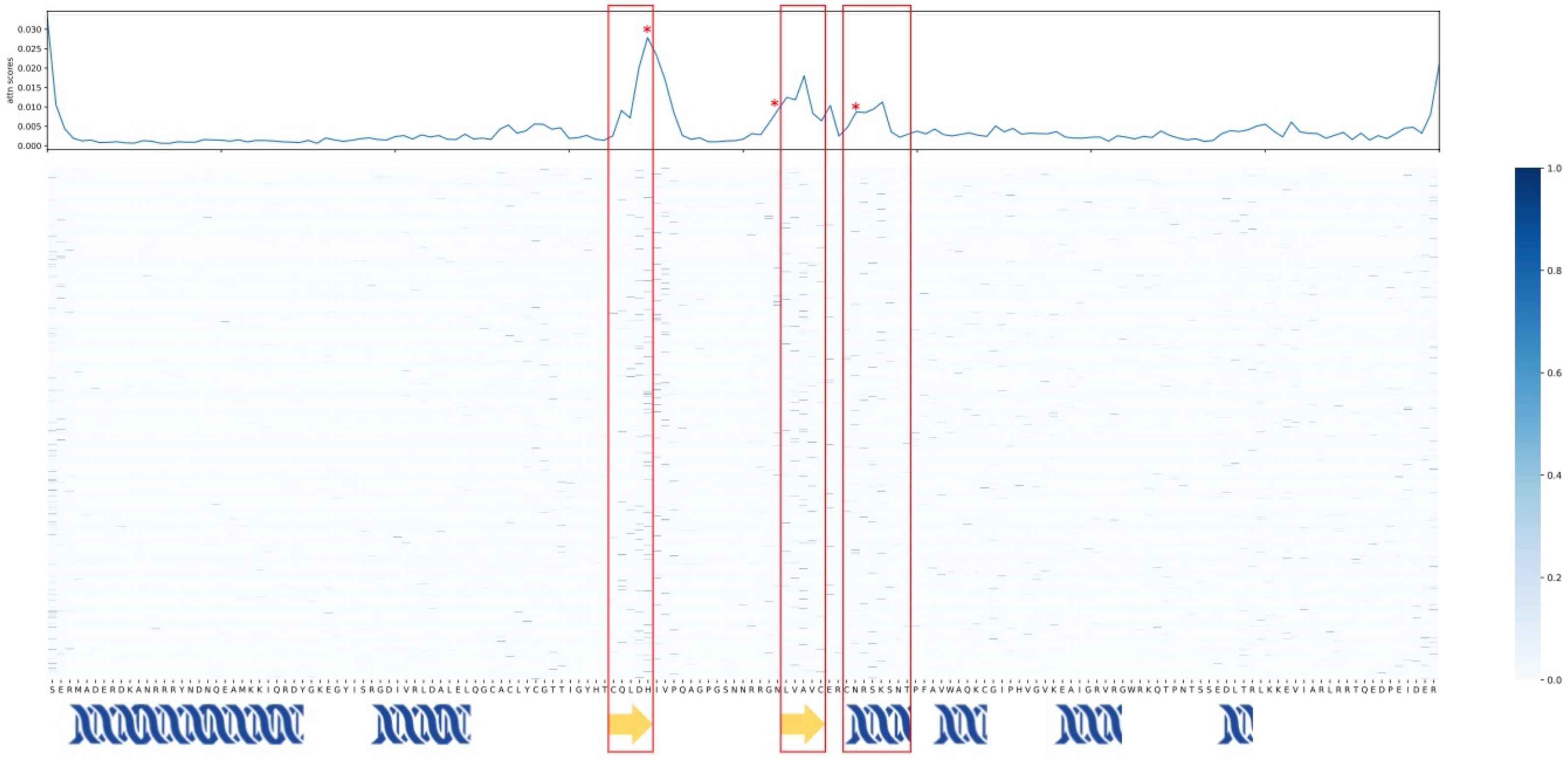
FuncPred: Function prediction from protein sequences

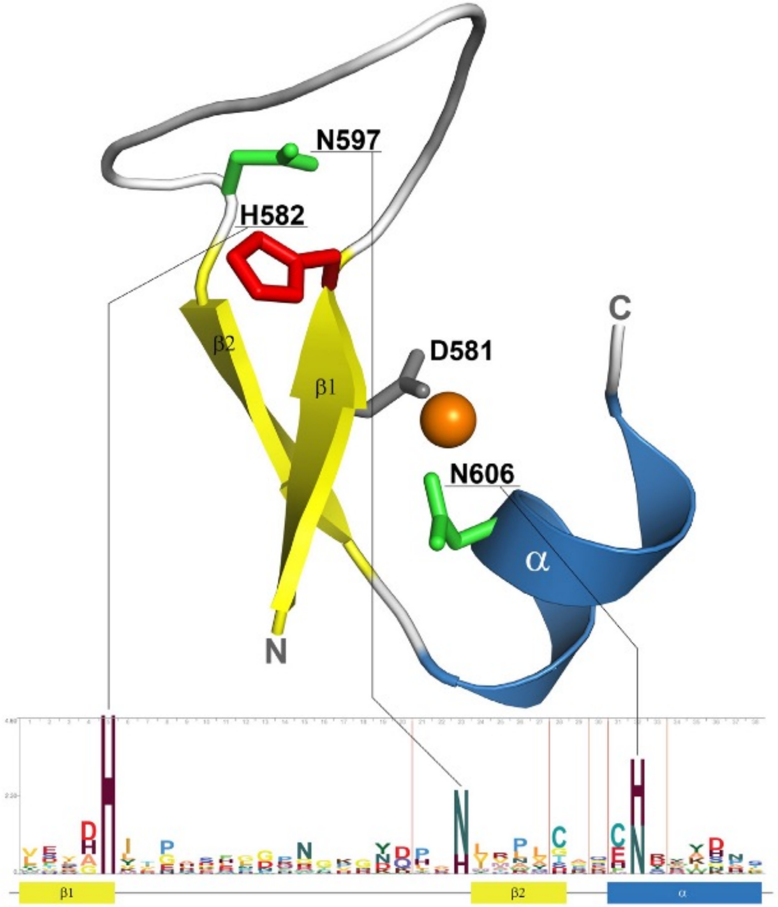
Aggregation method employing both convolution and attention outperformed other pooling methods



FuncPred: Function prediction from protein sequences

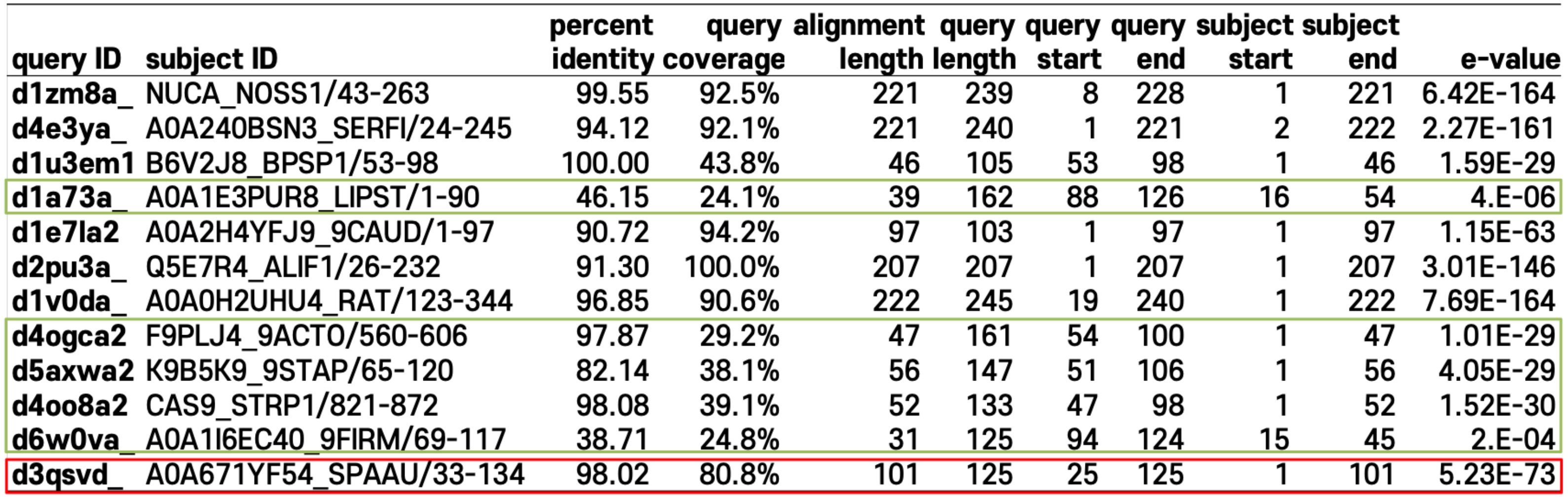
Aggregation method successfully focused on both conserved residues and conserved structures





FuncPred: Function prediction from protein sequences

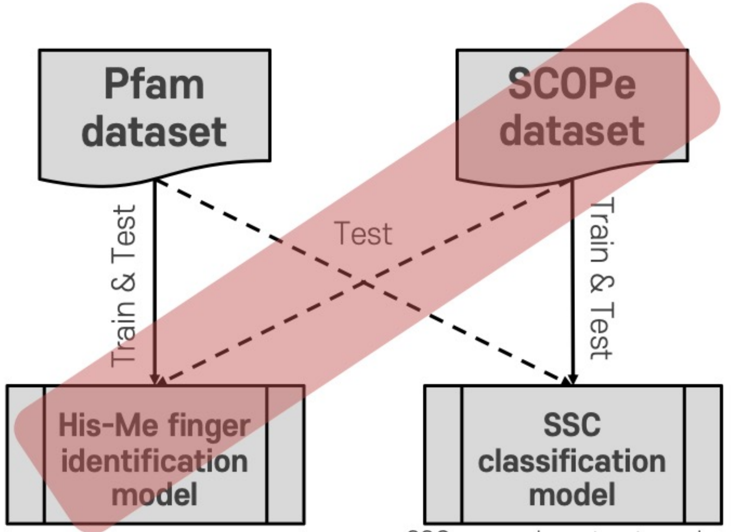
His-Me finger proteins in the SCOPe database are remote homologous against the Pfam database



query ID	subject ID	percent identity	query coverage	alignment length	query length	query start	query end	subject start	subject end	e-value
d1zm8a_	NUCA_NOSS1/43-263	99.55	92.5%	221	239	8	228	1	221	6.42E-164
d4e3ya_	A0A240BSN3_SERFI/24-245	94.12	92.1%	221	240	1	221	2	222	2.27E-161
d1u3em1	B6V2J8_BPSP1/53-98	100.00	43.8%	46	105	53	98	1	46	1.59E-29
d1a73a_	A0A1E3PUR8_LIPST/1-90	46.15	24.1%	39	162	88	126	16	54	4.E-06
d1e7la2	A0A2H4YFJ9_9CAUD/1-97	90.72	94.2%	97	103	1	97	1	97	1.15E-63
d2pu3a_	Q5E7R4_ALIF1/26-232	91.30	100.0%	207	207	1	207	1	207	3.01E-146
d1v0da_	A0A0H2UHU4_RAT/123-344	96.85	90.6%	222	245	19	240	1	222	7.69E-164
d4ogca2	F9PLJ4_9ACTO/560-606	97.87	29.2%	47	161	54	100	1	47	1.01E-29
d5axwa2	K9B5K9_9STAP/65-120	82.14	38.1%	56	147	51	106	1	56	4.05E-29
d4oo8a2	CAS9_STRP1/821-872	98.08	39.1%	52	133	47	98	1	52	1.52E-30
d6w0va_	A0A1I6EC40_9FIRM/69-117	38.71	24.8%	31	125	94	124	15	45	2.E-04
d3qsvd_	A0A671YF54_SPAAU/33-134	98.02	80.8%	101	125	25	125	1	101	5.23E-73

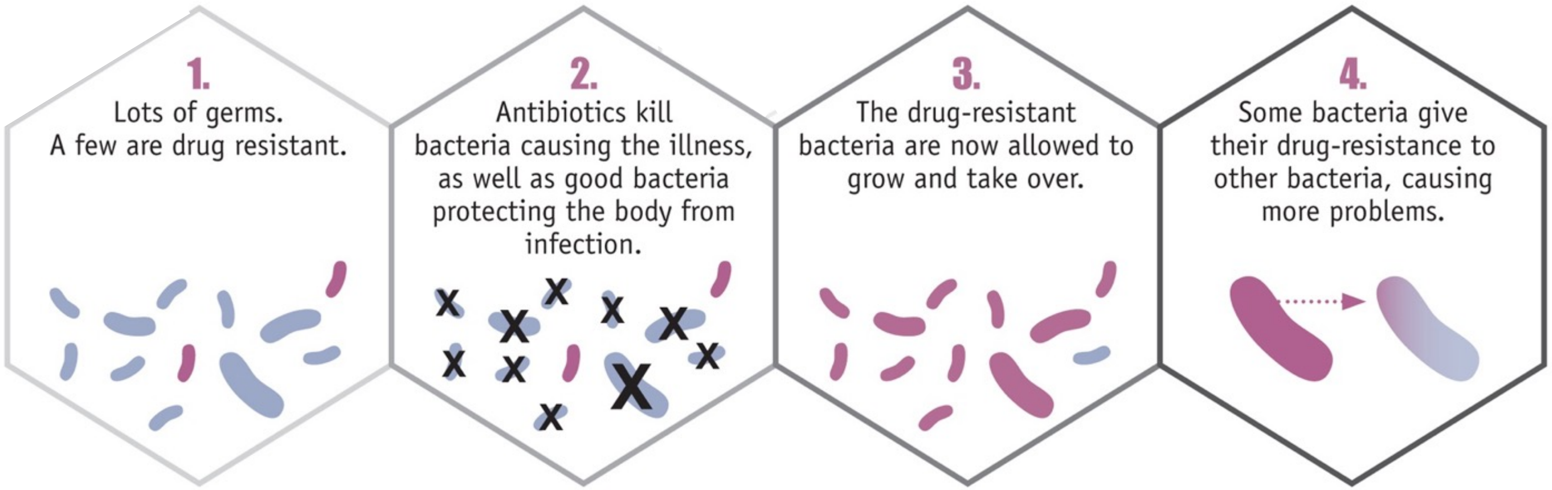
🟩 Remote homologous proteins

🟥 Highly homologous proteins, which was not reported as His-Me finger nuclease



SSC: secondary structure class

3. Genomics exchanges: antibiotic resistance



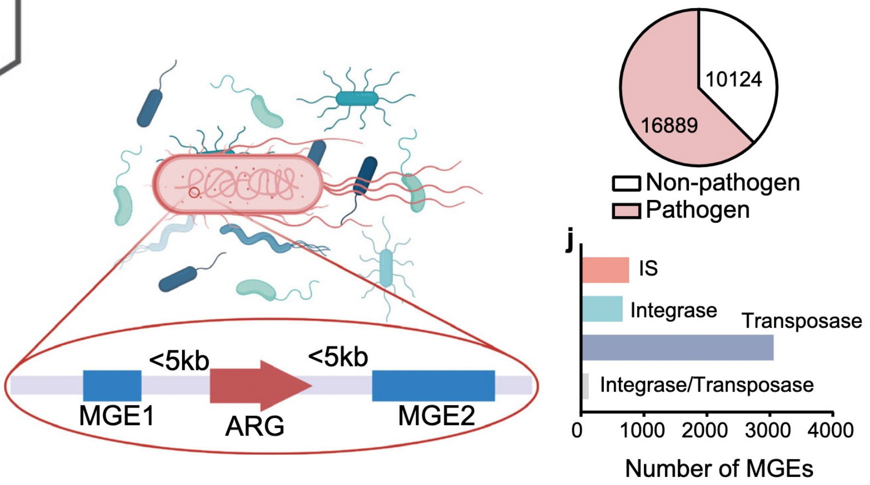
1. Lots of germs. A few are drug resistant.

2. Antibiotics kill bacteria causing the illness, as well as good bacteria protecting the body from infection.

3. The drug-resistant bacteria are now allowed to grow and take over.

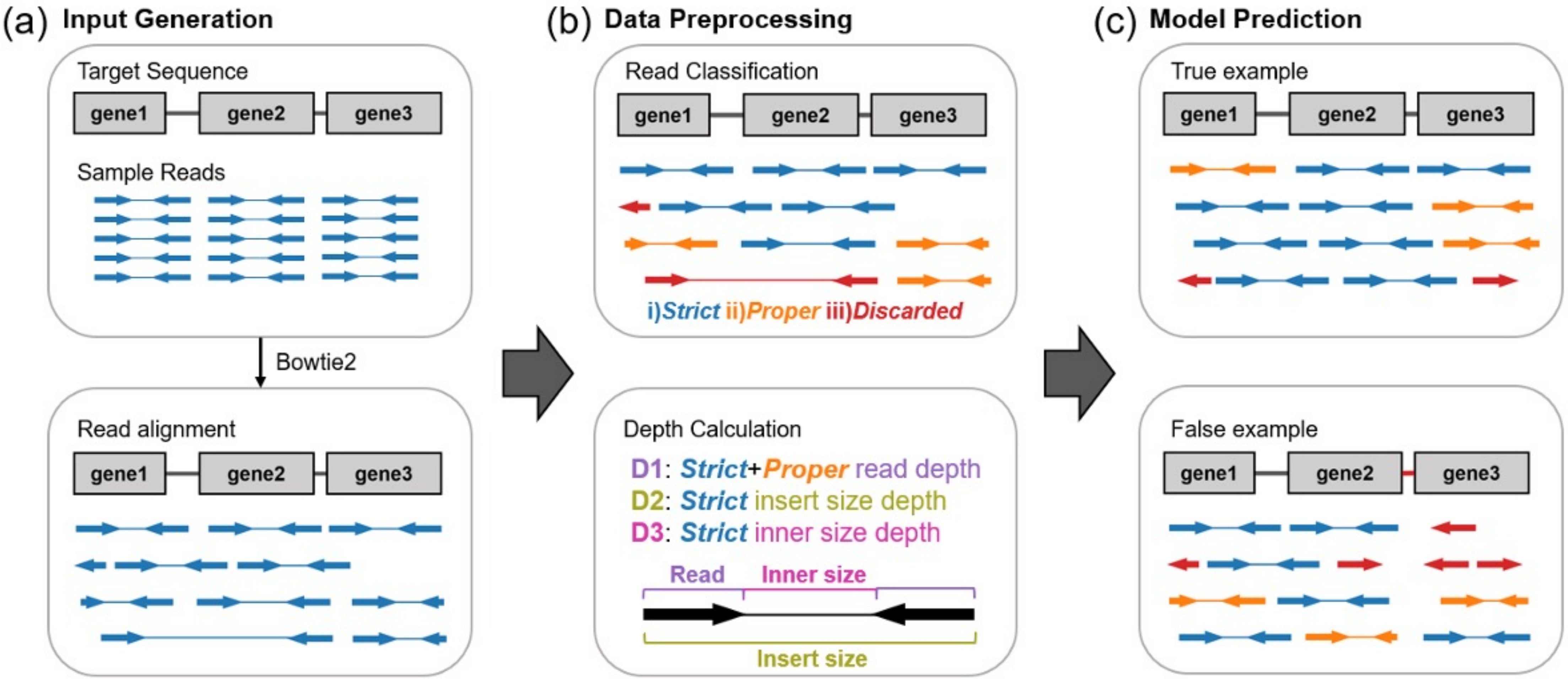
4. Some bacteria give their drug-resistance to other bacteria, causing more problems.

- Antibiotic resistance genes are often located on plasmids or transposons and can be transferred from cell to cell by conjugation, transformation, or transduction
- Resistome is a collection of all the antibiotic resistance genes and related elements in bacteria.

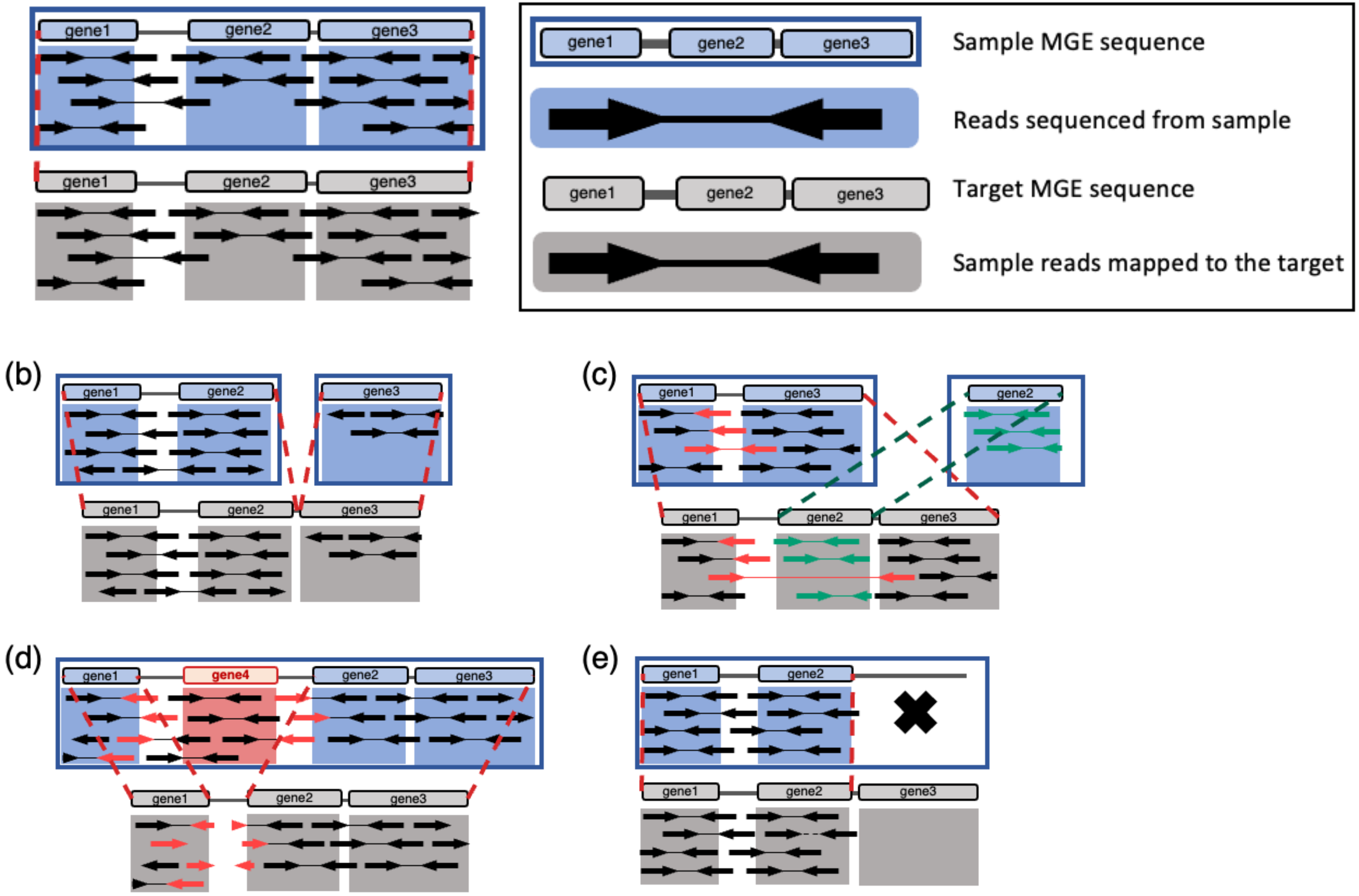


Nature Comm (2022)

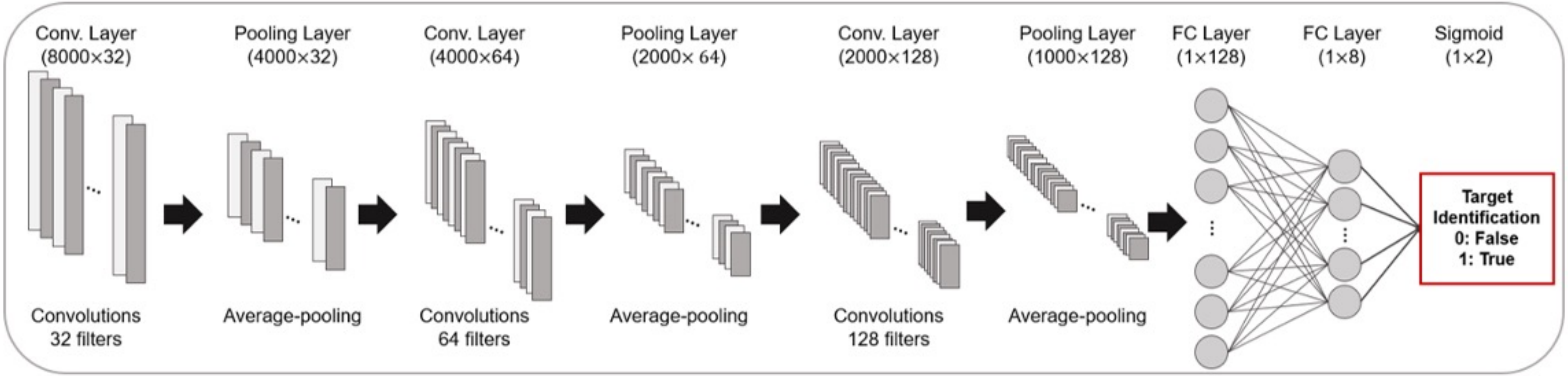
DeepMobilome: identifying ARGs from microbiome sequencing data



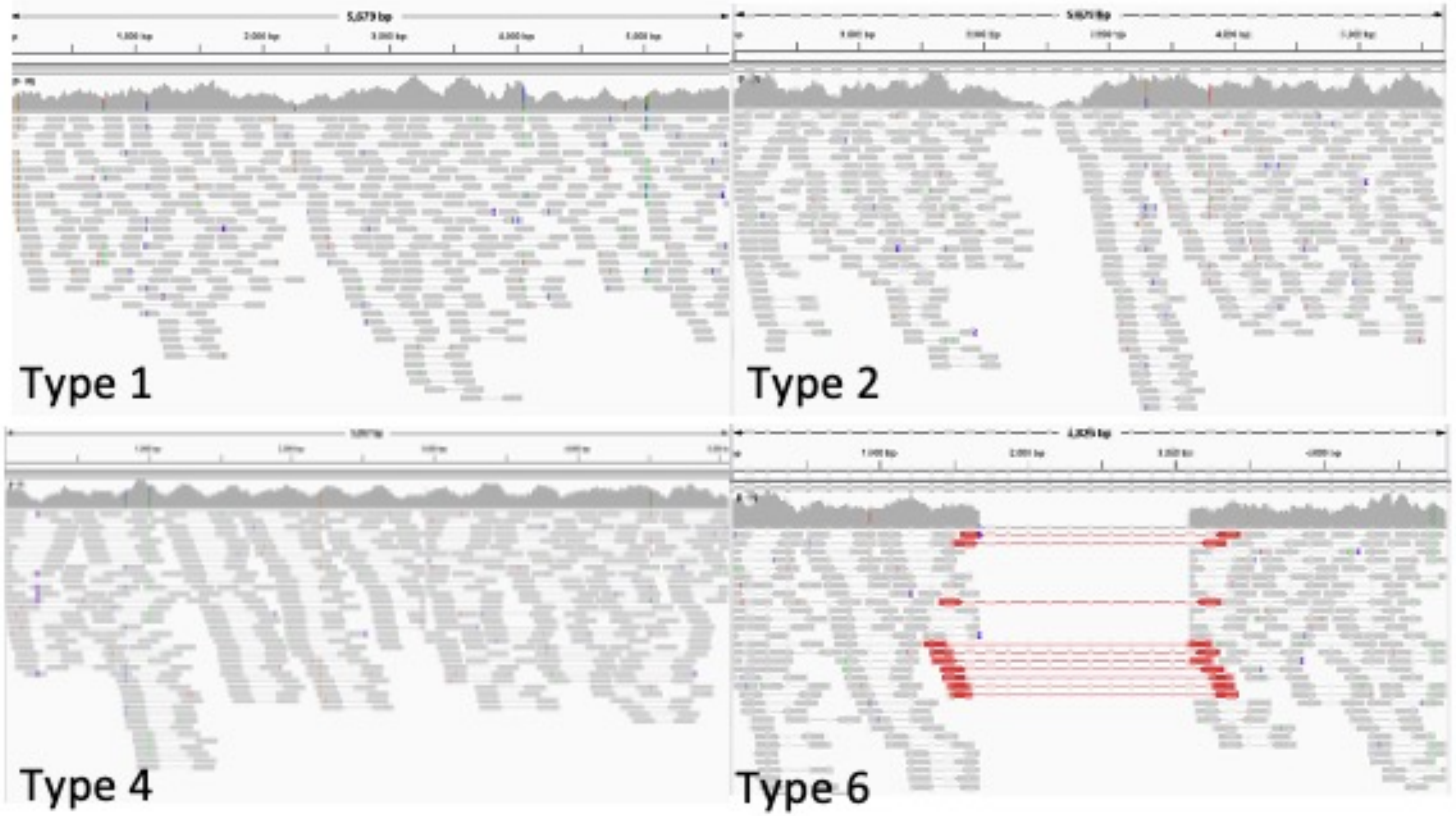
MGE patterns vs read alignments: MGEs

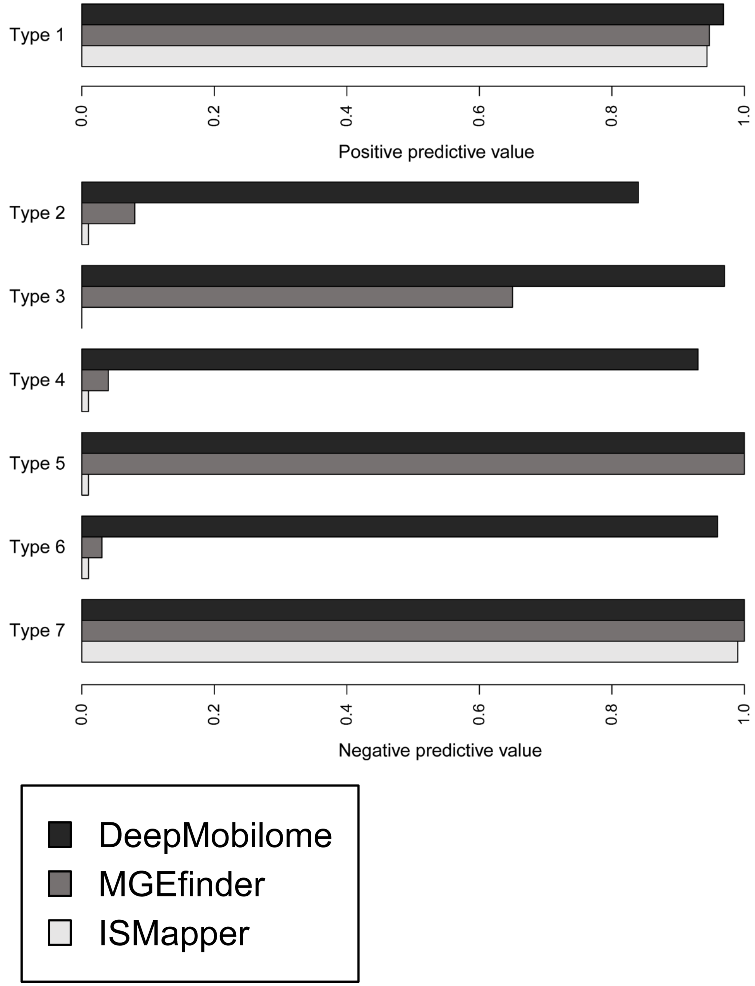


DeepMobilome: identifying ARGs from microbiome sequencing data



MGE patterns vs read alignments: pseudo MGEs





Take-home message

- Deep learning discovers multiple representations with different levels of abstractions for genomes and chemical compounds

- Generalization and specialization can be achieved by self-supervised learning in the pre-training process and task-specific learning in the fine-tuning process

- Large language model can be used to understand genomic characteristics and interactions of various taxa and functions

Acknowledgements

Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries, Korea(20220517)

Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University))

Q
A