Revisiting K-mer Profile for Effective and Scalable Genome Representation Learning

Abdulkadir Çelikkanat, Andres R. Masegosa, Thomas D. Nielsen

Introduction

Metagenomic binning

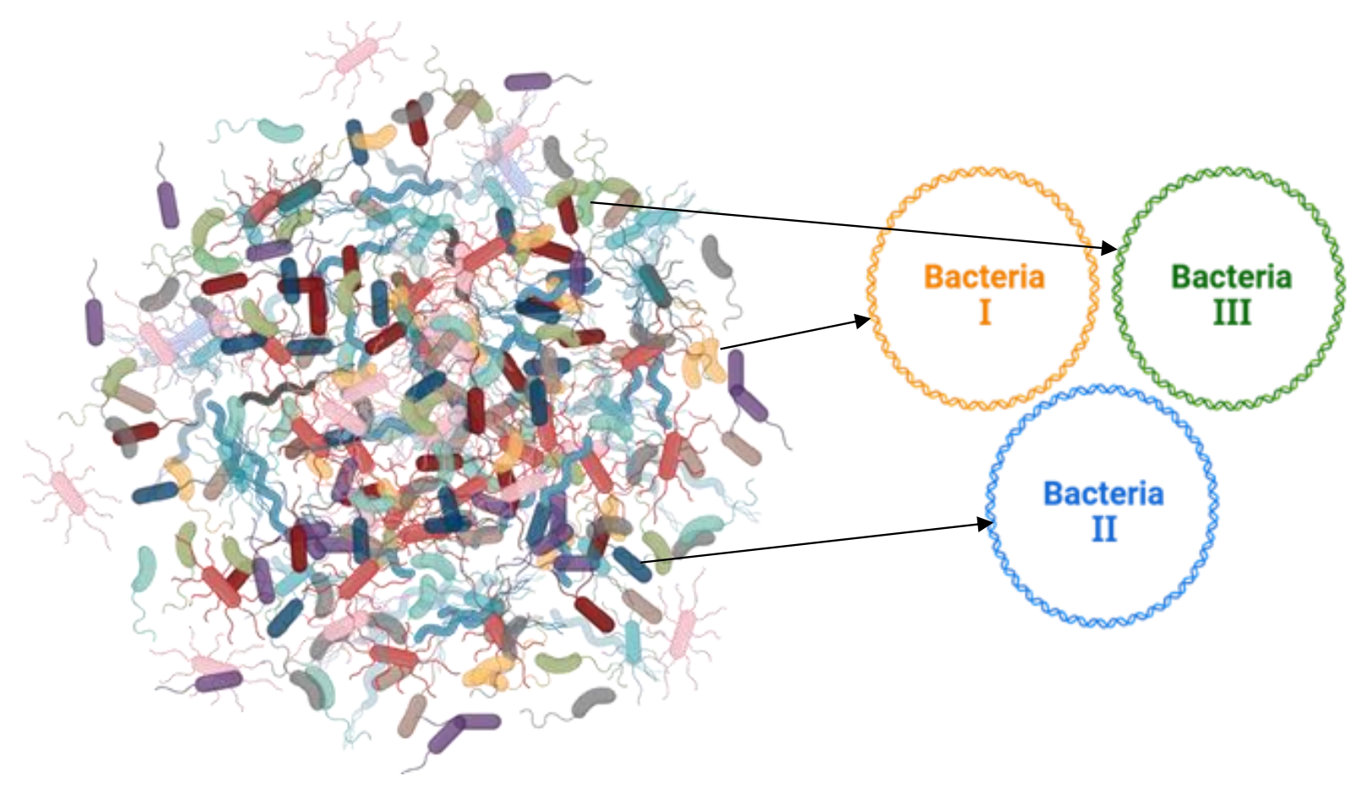
- Sequencing a complex microbial sample using current DNA sequencing technologies rarely produces full DNA sequences, but rather a mixture of DNA fragments (called **reads**) of the microbes present in the sample.
- In order to recover the full microbial genomes, a subsequent binning/clustering step is performed, where individual DNA fragments are clustered together according to their genomic origins.

Ideally

Extract a genome for each bacteria in the sample

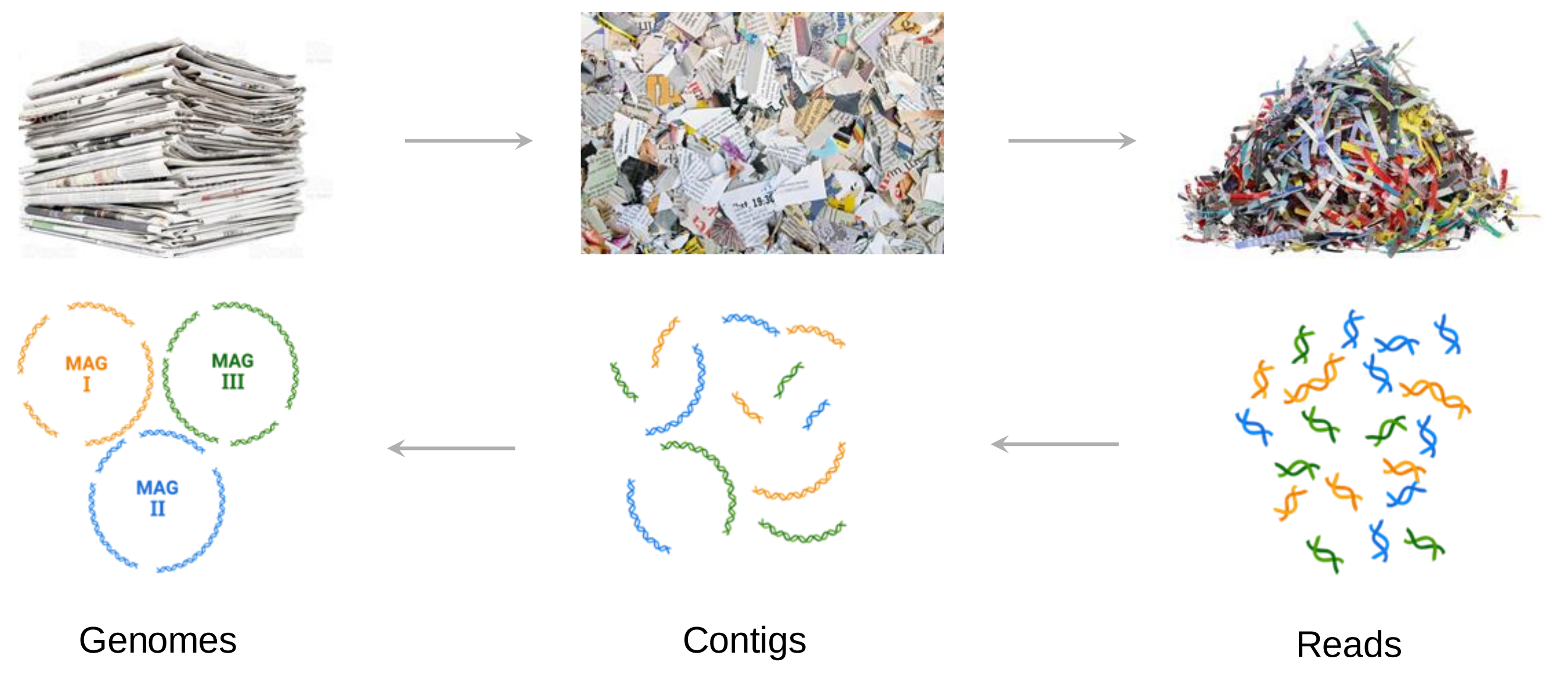
Reality

Recreate genomes of very similar organisms using available information

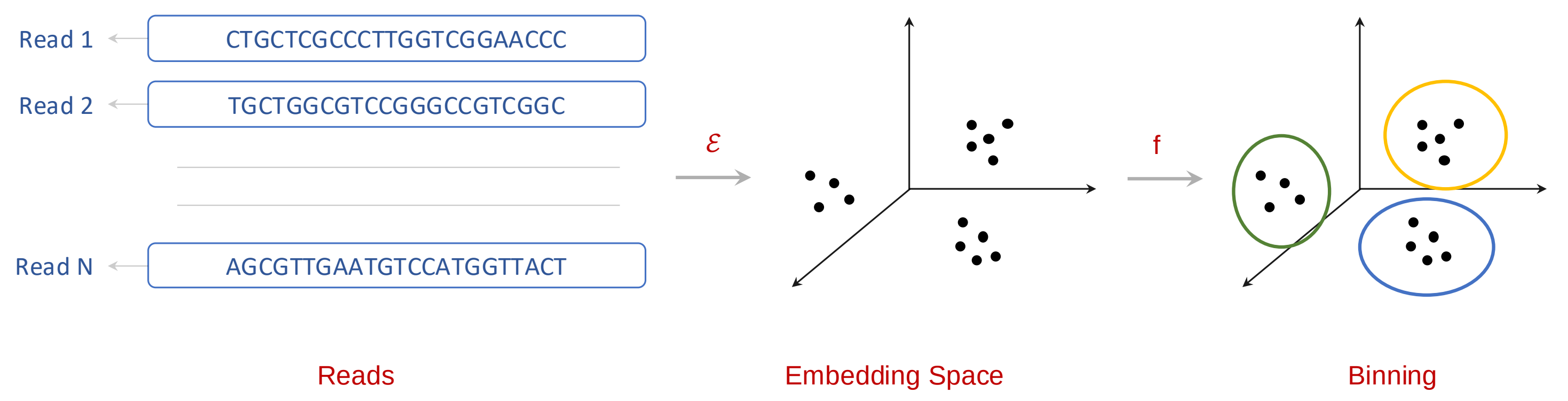


- ~ 50,000 prokaryotic species in the genome databases
- Estimates of millions to billions or even trillions of species.

Metagenomics Binning Problem



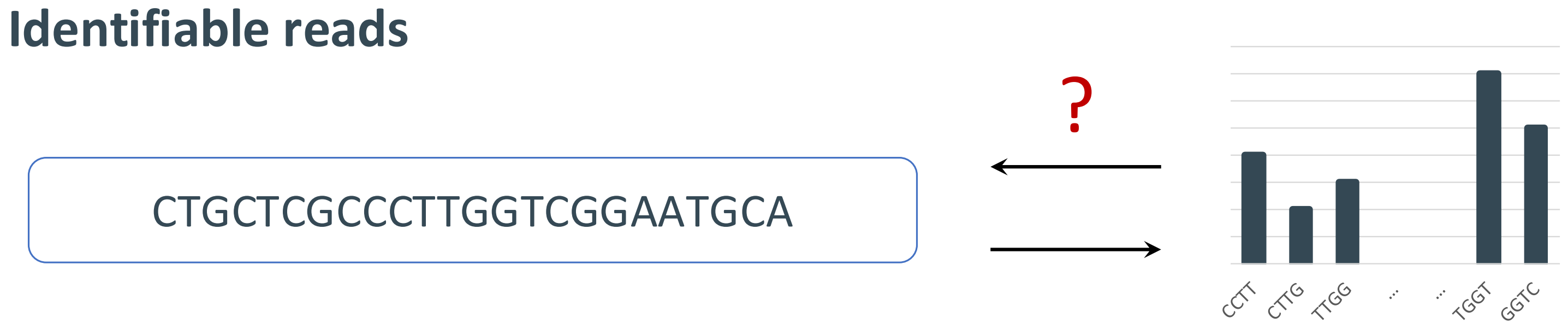
Learning latent representations of reads



Problem Definition. Let $\mathcal{R} \subset \Sigma^+$ be a finite set of reads with a genome mapping function ℓ where $\Sigma = \{A, C, T, G\}$. For a given threshold value $\gamma \in \mathbb{R}^+$, the objective is to learn an embedding function $\mathcal{E} : \mathcal{R} \to \mathbb{R}^d$ that embeds reads into a low-dimensional metric space $(\mathrm{X}, d_{\mathrm{X}})$, usually a Euclidean space, such that $d_{\mathrm{X}}(\mathcal{E}(\mathbf{r}), \mathcal{E}(\mathbf{q})) \leq \gamma$ if and only if $\ell(\mathbf{r}) = \ell(\mathbf{q})$ for all reads $\mathbf{r}, \mathbf{q} \in \mathcal{R}$ where $d \ll |\mathcal{R}|$.

Proposed Models

Identifiable reads



Theorem. Let $\mathbf{r}$ be a read of length ℓ. There exists no other distinct read having the same k-mer profile if and only if it does not satisfy any of the following conditions:

1. $r_1 \cdots r_{k-1} = r_{\ell-k-2} \cdots r_\ell$ and $r_i \neq r_1$ for some $1 < i < \ell - k - 2$.
2. $r_i \cdots r_{i+k-2} = r_j \cdots r_{j+k-2}$ and $r_g \cdots r_{g+k-2} = r_h \cdots r_{h+k-2}$ for some indices $1 \leq i < g < j < h \leq \ell - k + 2$ where $r_{i+k-1} \cdots r_{g-1} \neq r_{j+k-1} \cdots r_{h-1}$.
3. $r_i \cdots r_{i+k-2} = r_j \cdots r_{j+k-2} = r_h \cdots r_{h+k-2}$ for some indices $1 \leq i < j < h \leq \ell - k + 2$ where $r_{i+k-1} \cdots r_{j-1} \neq r_{j+k-1} \cdots r_{h-1}$.

- *Identifiable reads* can be uniquely reconstructed from their k-mer profile.

Lipschitz equivalent spaces.

Proposition. Let $M_1 = (\aleph_\ell, d_{\mathcal{H}})$ and $M_2 = (\mathbb{N}^{|\Sigma^k|}, \|\cdot\|_1)$ be the metric spaces denoting the set of identifiable reads and their corresponding k-mer profiles equipped with edit and ℓ_1 distances, respectively. The k-mer profile function, $c : M_1 \to M_2$, mapping given any read, $\mathbf{r}$, to its corresponding k-mer profile, $c_{\mathbf{r}} := c(\mathbf{r})$, is a Lipschitz equivalence, i.e. it satisfies

$$\forall \mathbf{r}, \mathbf{q} \in \Sigma^\ell \quad \alpha_l d_{\mathcal{H}}(\mathbf{r}, \mathbf{q}) \leq \|c_{\mathbf{r}} - c_{\mathbf{q}}\|_1 \leq \alpha_u d_{\mathcal{H}}(\mathbf{r}, \mathbf{q}) \tag{1}$$

for $\alpha_l = 1/\ell$ and $\alpha_u = k|\Sigma|^k$, so M_1 and M_2 are Lipschitz equivalent.

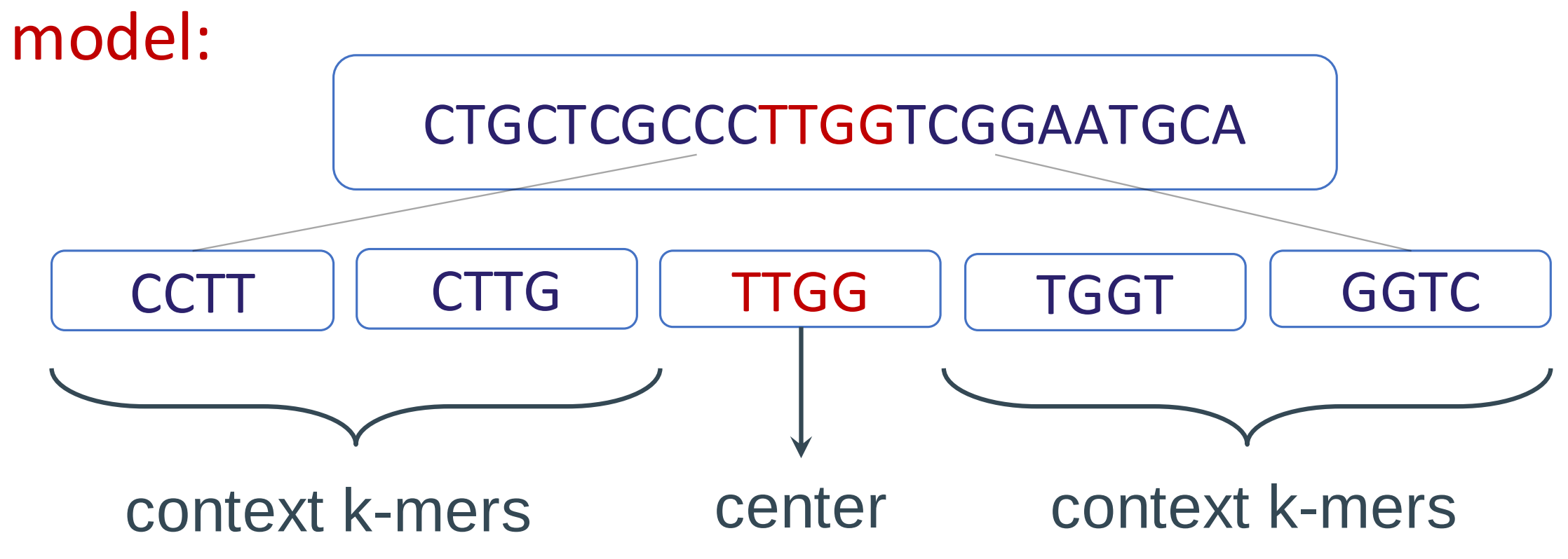
k-mer profile: First, consider the definition of k-mer profiles:

$$\mathcal{E}_{\mathrm{kmer}}(\mathrm{r}) \coloneqq \sum_{\mathrm{x} \in \Sigma^{\mathrm{k}}} \mathrm{c_r(x) z_x}$$

where $\mathrm{z_x}$ represents the canonical basis vector.

k-mers are not independent!

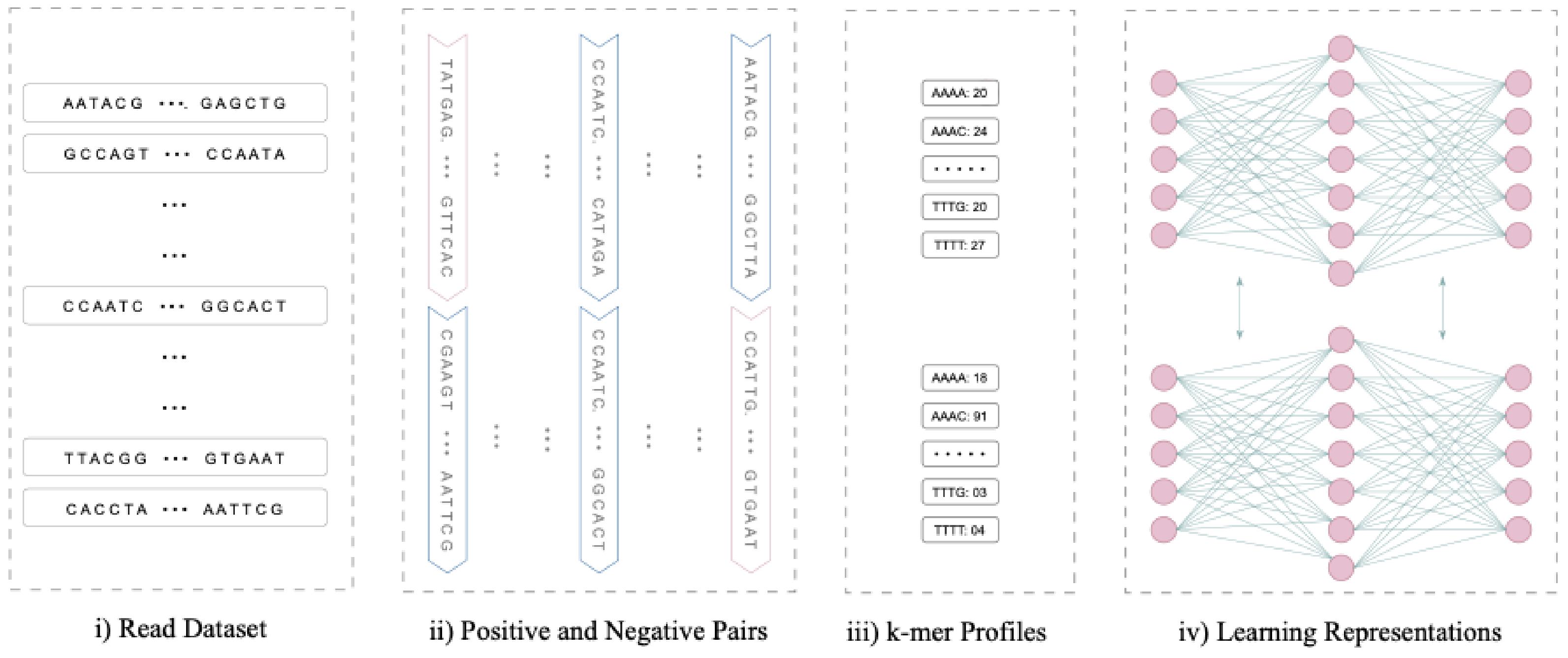
Poisson model:



- $o_{\mathbf{x},\mathbf{y}}$ indicates the number of average co-appearances of k-mers i $\mathbf{x}$ and $\boldsymbol{y}$ per read within a window size ω

$$o_{\mathbf{x},\mathbf{y}} \sim Pois(\lambda_{\mathbf{x},\mathbf{y}}) \qquad \lambda_{\mathbf{x},\mathbf{y}} \coloneqq \exp(-\|\mathbf{z_x} - \mathbf{z}_y\|^2)$$

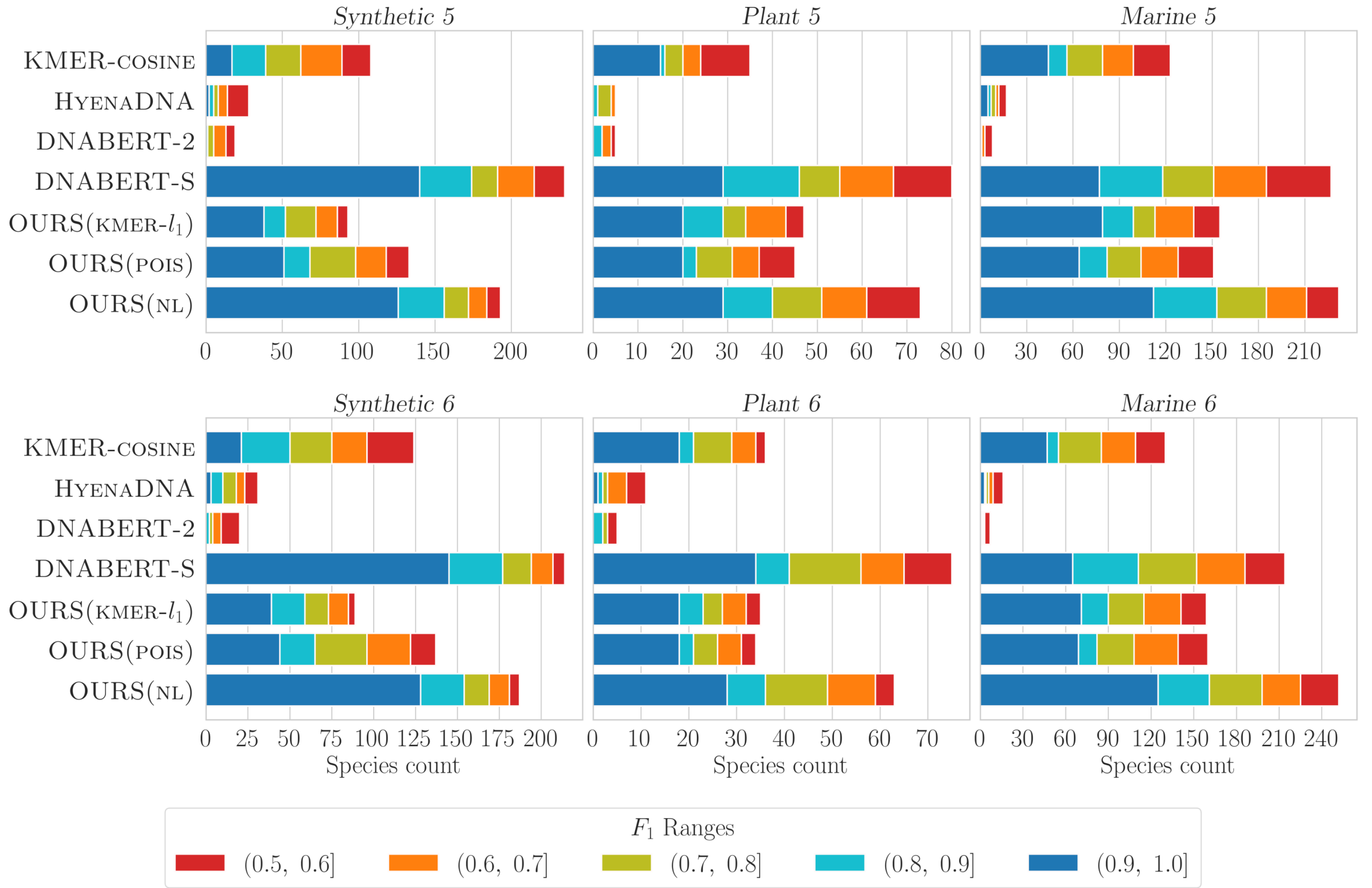
Non-linear read embeddings



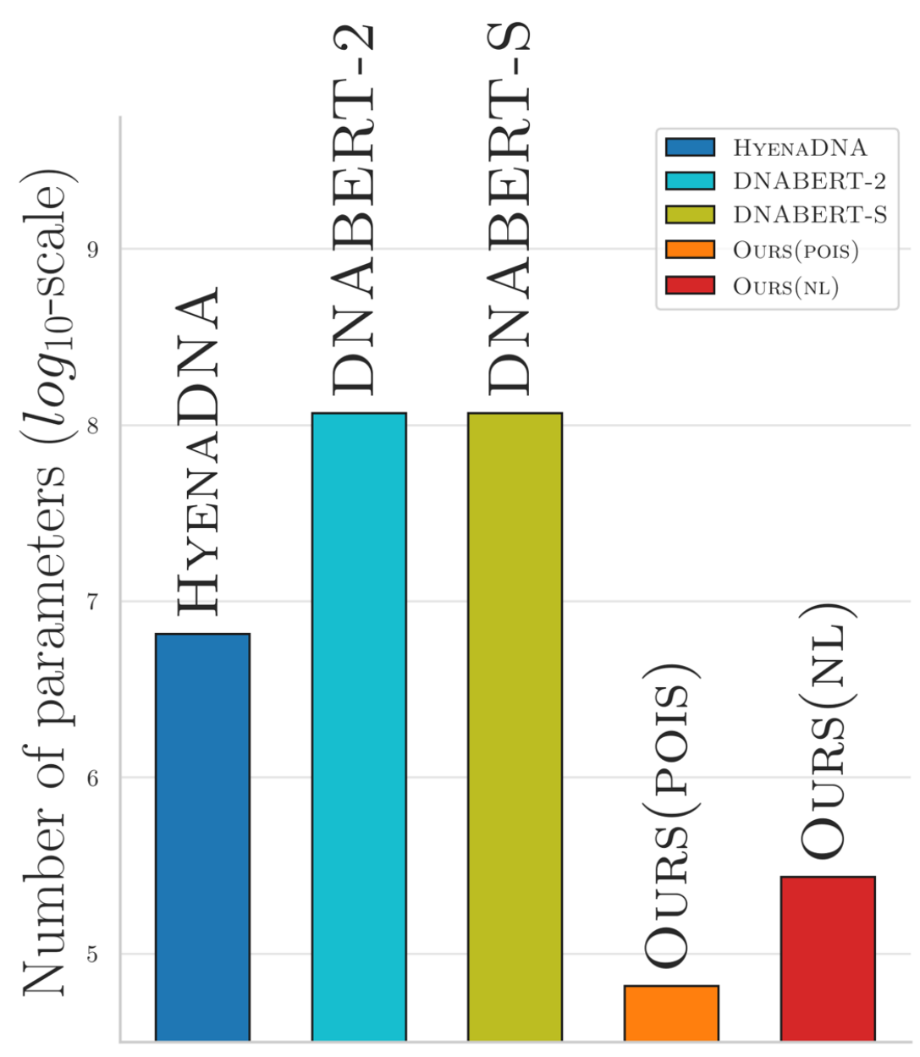
i) Read Dataset ii) Positive and Negative Pairs iii) k-mer Profiles iv) Learning Representations

$$\mathcal{L}_{\mathrm{NL}}\Big(\{y_{ij}\}_{(i,j)\in\mathcal{I}}|\Omega\Big) := -\frac{1}{|\mathcal{I}|}\sum_{(i,j)\in\mathcal{I}} y_{ij}\log p_{ij} + (1-y_{ij})\log(1-p_{ij}) \qquad p_{ij} = \exp\left(-\|\mathcal{E}_{\mathrm{NL}}(\mathbf{r}_i) - \mathcal{E}_{\mathrm{NL}}(\mathbf{r}_j)\|^2\right)$$

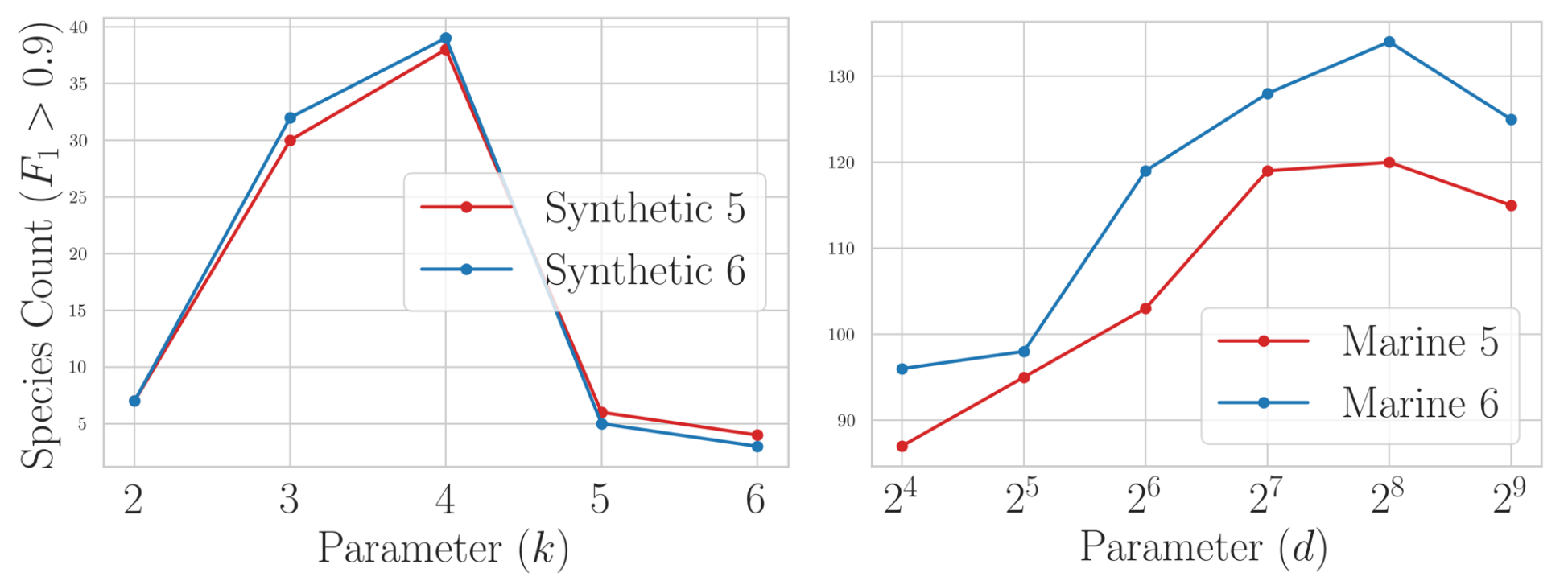
Experiments



Number of parameters



Ablation study



This work was supported by a research grant (VIL50093) from VILLUM FONDEN

SCAN ME