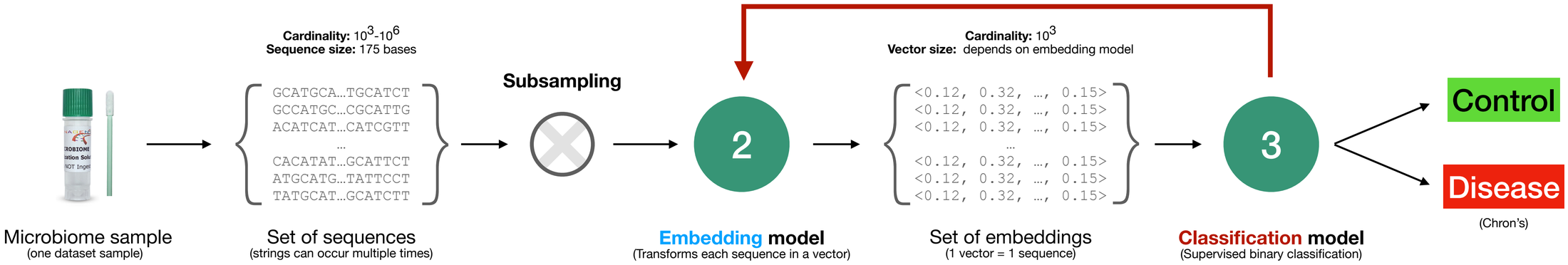
The impact of embedding models on disease detection tasks from microbiome sequencing data

Mattia Strocchi* [1], **Supervisors:** Gabriele Corso[2], Pietro Liò[3], **Responsible professor:** Jasmijn Baaijens[1]

(*m.strocchi@student.tudelft.com)

1 TU Delft
2 MIT
3 University of Cambridge



1. Overview

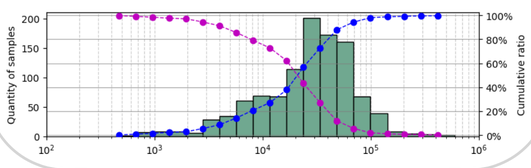
★ **Background:** given a set of microbiome (16S-V4 rRNA) sequences, predict if the host has Crohn's disease.

Goal: take various **embedding models** (2) and measure their efficacy via the **classifier** (3) performance.

★ **Research Question 1:** which *embedding model* yields the best *disease detection* performance?

★ **Research Question 2:** does the *set classification formulation* improve the overall disease detection performance?

Dataset overview: each sample (patient) has a variable number of sequences (microbes). Below the distribution of the cardinality of the samples.



2. Embedding models

1) k-mer frequency: use k-mer distributions as vectors (MicroPheno) [3]

2) Learnable embeddings: sequences are embedded in a vector space that preserves edit-distance (NeuroSEED) [4]

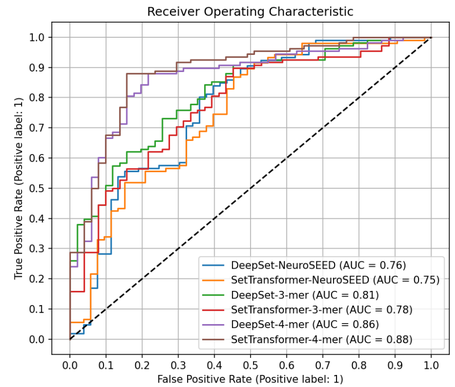
3. Classification models

★ **Literature:** a common approach is combining the set of vectors into a single one by using a permutation-invariant function (e.g. the mean) and then applying a vector classification model.

Issue: combining the vectors this way (referred to as *baseline*) does not allow the classifier to learn the complex interactions among the sequences.

★ **Proposed approach:** Instead of combining the vectors, we use models working over sets, namely: **Deep Set** [1], and **Set Transformer** [2].

4. Results



Setup: k-mer frequency vectors were tested for k = 3 and k = 4. NeuroSEED embeddings are produced by a CNN on euclidian space.

★ **Results:** The graph shows the ROC curves for each combination of embedding model and classifier. 4-mer embeddings perform 8.6% better compared to the best model on 3-mers, that are 6.1% better than NeuroSEED's.

5. Conclusions

		AUROC		
	Model	NeuroSEED	3-mers	4-mers
Baselines	RF	0.721	0.755	0.826
	KNN	0.688	0.757	0.792
	MLP	0.718	0.678	0.743
	RBF SVM	0.646	0.769	0.822
	Average	0.693	0.740	0.796
Set	Set Transformer	0.751	0.779	0.884
	Deep Set	0.765	0.814	0.864
	Average	0.758	0.796	0.874
	Overall avg.	0.725	0.768	0.835
	Overall std.	0.043	0.045	0.051

★ The best model achieves an AUC of 0.884 on 4-mers, which is an excellent result considering the signal-to-noise ratio of the problem **(RQ1)**. Additionally, the *set formulation* of the task **(RQ2)** is shown to be more effective, scoring a better AUC on all the embeddings.

6. Limitations

- Tests run on a *single disease* dataset.
- Going above k = 4 was not possible.
- Missing alignment-based embeddings.
- Influence of the *subsampling step* should be further investigated.

7. References

[1] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., & Smola, A. (2017). Deep Sets. arXiv. https://doi.org/10.48550/ARXIV.1703.06114

[2] Lee, J., Lee, Y., Kim, J., Kosiorek, A. R., Choi, S., & Teh, Y. W. (2018). Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. arXiv. https://doi.org/10.48550/ARXIV.1810.00825

[3] Asgari, E., Garakani, K., McHardy, A. C., & Mofrad, M. R. K. (2018). MicroPheno: Predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. BioRxiv. https://doi.org/10.1101/255018

[4] Corso, G., Ying, R., Pándy, M., Veličković, P., Leskovec, J., & Liò, P. (2021). Neural Distance Embeddings for Biological Sequences. arXiv. https://doi.org/10.48550/ARXIV.2109.09740