The Grammar of mRNA: Language Modelling methods for genomic sequence detection on cardiovascular patient data

Dr Kakia Chatsiou [1] Dr Suha Al Naimi [1] Ben Holmes[1] Dr Santiago Miriuka [2] Dr Carlos Luzzani [2]

[1]*University of Suffolk, UK* [2]MultiplAI

University of Suffolk

Background

Messenger RNA (mRNA) is a crucial component of all living organisms, present in every cell. Acting as a messenger, mRNA interacts with cellular machinery to facilitate **protein synthesis**, carrying specific instructions for the construction of proteins.

Once its role is fulfilled, the mRNA is degraded and does not persist for long periods. The potential of mRNA to instruct the body to produce therapeutic agents has been exemplified by COVID-19 mRNA vaccines, which direct cells to produce specific proteins. **Cardiovascular diseases** represent another potential application area, where the **study of proteins** associated with these conditions could inform treatment strategies. Despite its importance, the **structure of mRNA remains partially understood**, which presents a challenge for scientists.

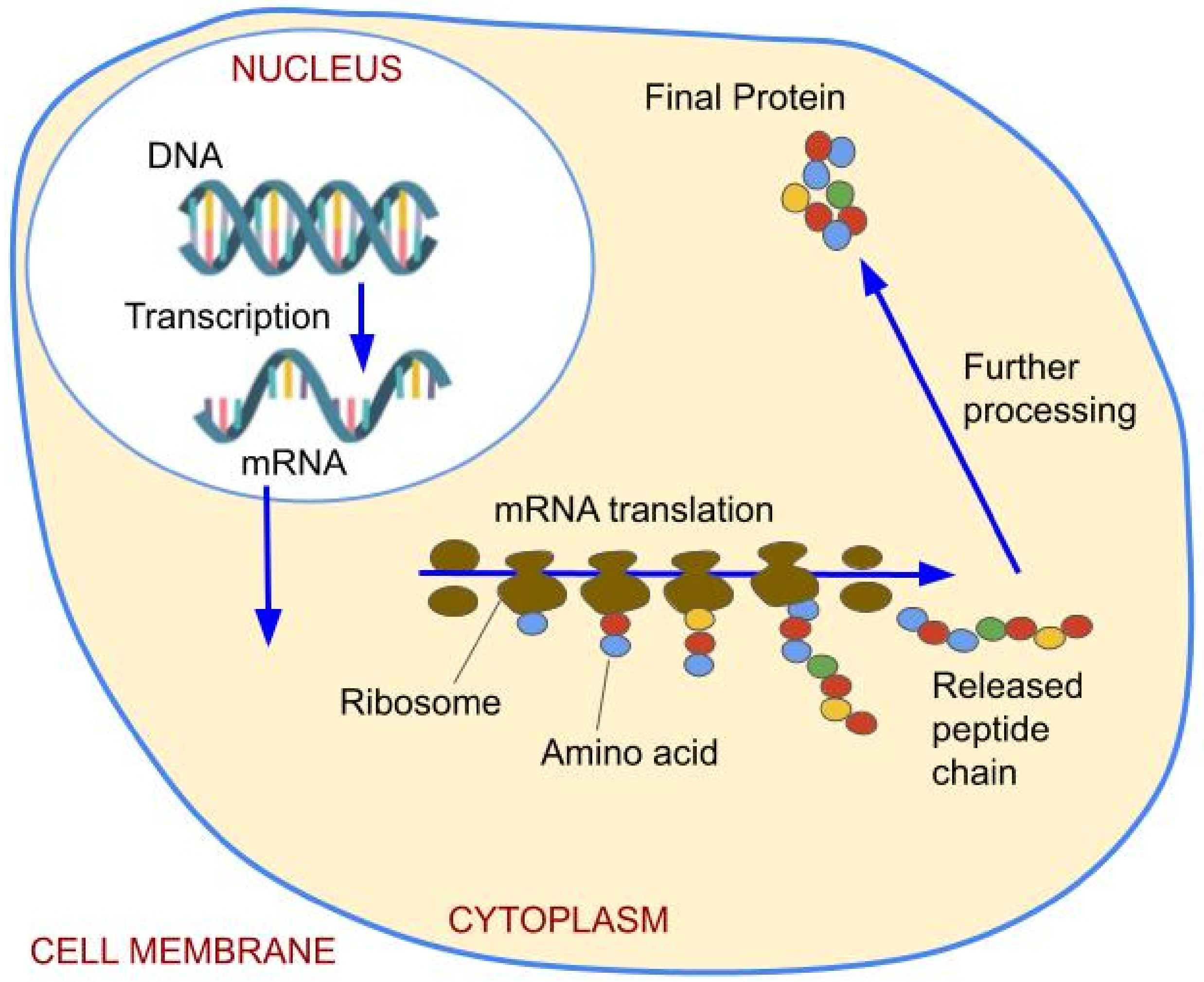

Figure 1. Typical mRNA structure

Project Objectives

This project applied **text mining and machine learning techniques**, typically used in text data analysis, to **study the mRNA structure** and identify its minimal units from existing sequencing data. Its **key objectives** were:

- reviewed the **key literature** to understand the landscape of mRNA sequencing and NLP methods (WP1)
- built a **data pre-processing pipeline for FASTQ** datafiles (WP2)
- explored **FastQ data using machine learning** (clustering) (WP3)

This pilot project was a collaboration between industry partners **MultiplAI Health** and academics from the School of Technology, Business, and Arts at the **University of Suffolk**, funded by an **Innovation Voucher from UKRI**.

Building a Data Pre-Processing Pipeline (WP2)

We experimented with different types of pipelines for mRNA sequencing data that would work with the inhouse data of MultiplAI. Data used was **open mRNA sequencing data** matching MultiplAI's information structure.

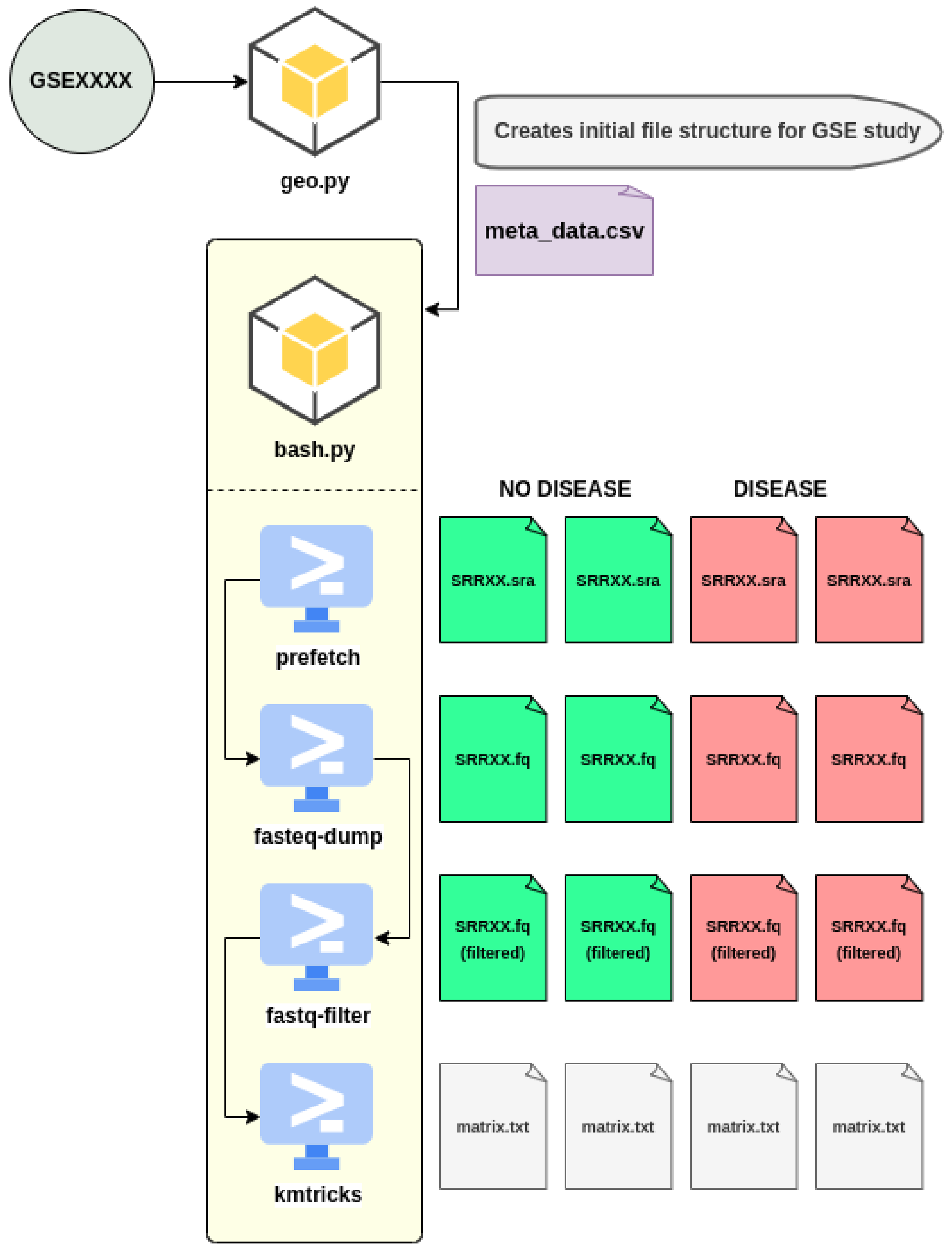

Figure 2. FASTQ pipeline overview

The data pre-processing pipeline used **Python**, **Conda**, the **SRA toolkit** library, **Pandas**, **Parsel**, **Pysradb**, **Fastq-filter**, and **Kmtricks**. The script fetches metadata, retrieves SRA data files, generates FASTQ files, filters these files based on Phred quality scores, and finally generates a k-mer count matrix.

Machine Learning on FastQ data (WP3)

Exploring **vectorised FastQ data** using machine learning methods to identify patterns and structures was next. **Clustering** was performed using various combinations of **k-mer clustering units** and *k-means* and *hierarchical clustering* to identify gene or protein sequences relevant to cardiovascular diseases. While preliminary, this exploration confirmed the potential of these methods for sequencing data analysis.

Next Steps

1. **RNA-seq to vector encoding**: Nucleic acid sequences can be encoded using one-hot encoding, word2vec, embedding layers, or k-mer counts.

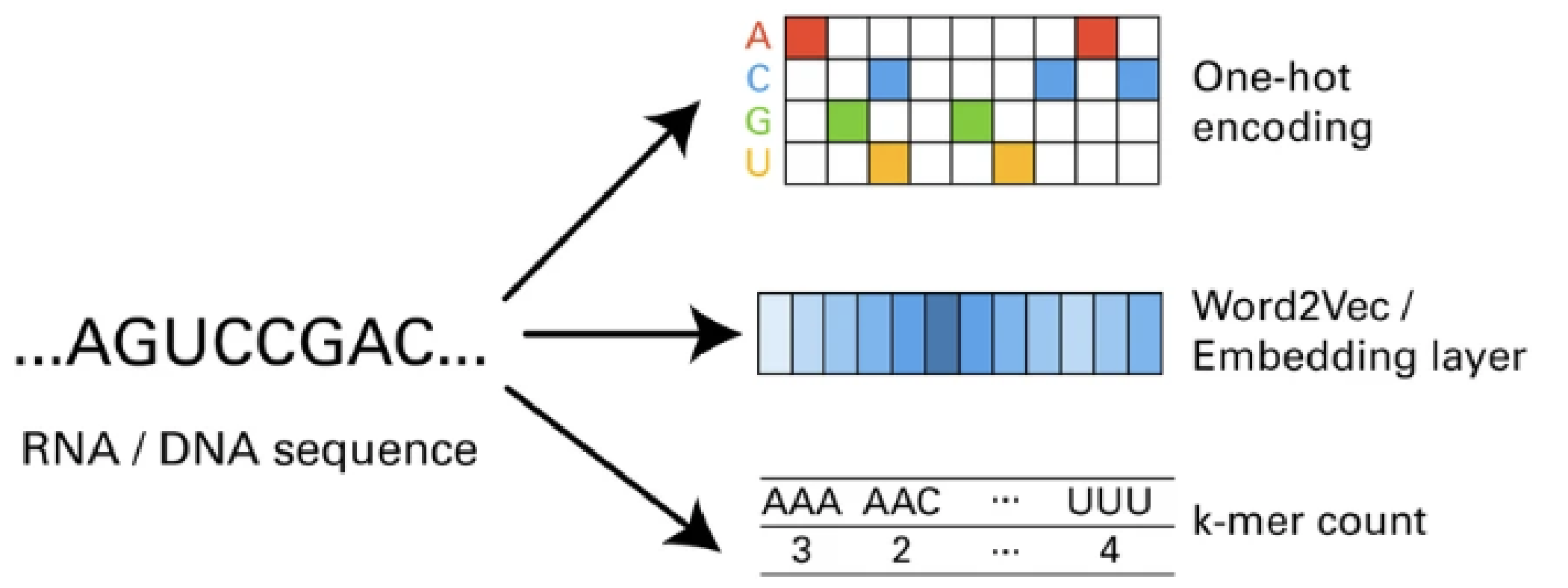

Figure 3. Processing RB data into input features of DL models - Hwang et al (2024).

2. **Dimensionality reduction and Choice of Deep learning model:**

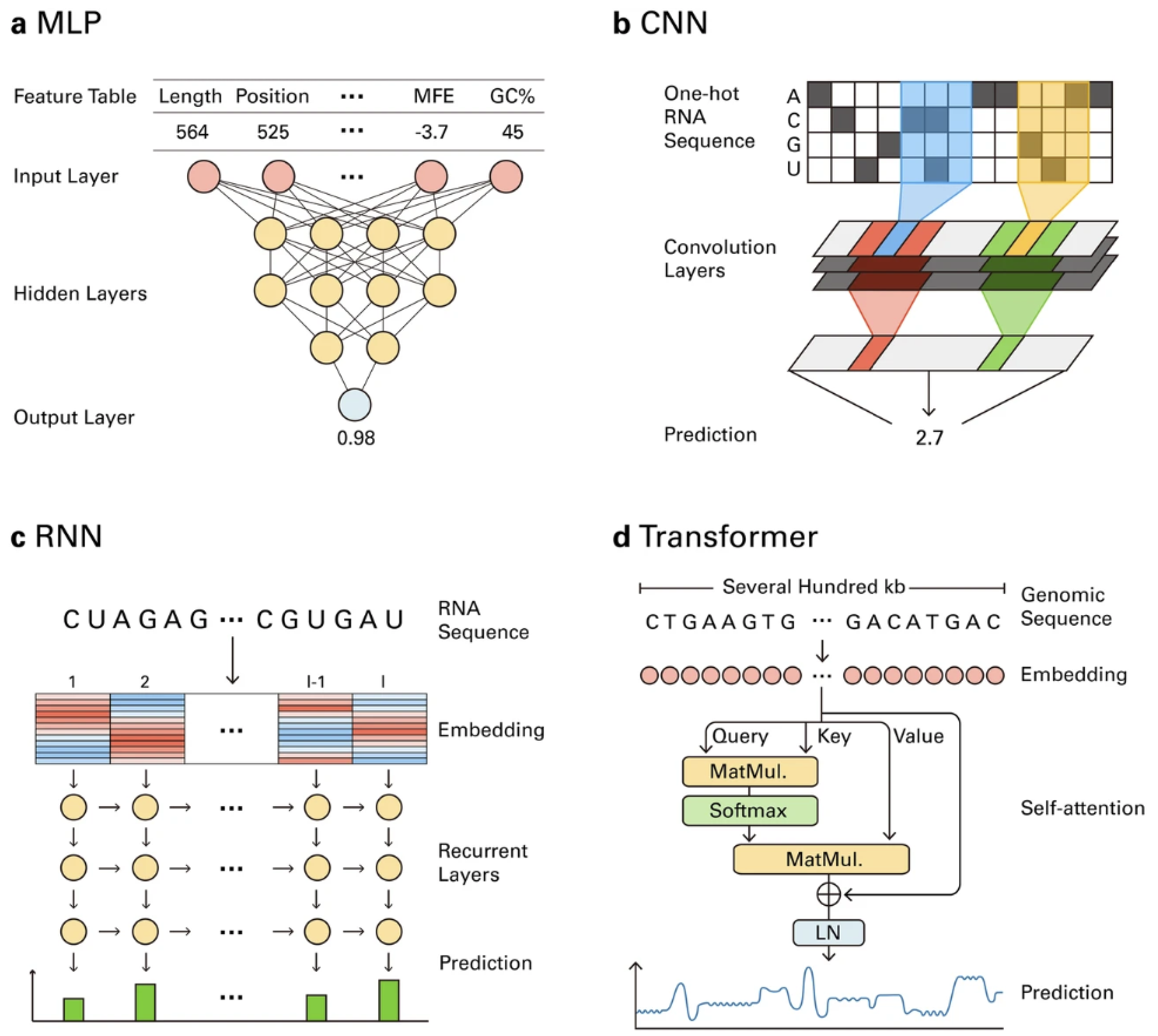

Figure 4. DL architectures for RB models - Hwang et al (2024)

a MLPs can make probabilistic predictions from biological feature tables. **b** CNNs can predict biological features from one-hot-encoded biological sequences by capturing local patterns. **c** RNNs can process embeddings of RNA sequences to provide basewise predictions of biological features. **d** GNNs can operate on gene–gene interaction networks derived from gene correlation matrices to predict genewise biological features.

Hwang, H., Jeon, H., Yeo, N. et al. (2024) Big data and deep learning for RNA biology. Exp Mol Med 56, 1293–1321. https://doi.org/10.1038/s12276-024-01243-w

https://github.com/MultiplAI-UoSuffolk School of Allied Health Sciences Research Symposium 2024, University of Suffolk k.chatsiou@uos.ac.uk