nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00791-0

Codon language embeddings provide strong signals for use in protein engineering

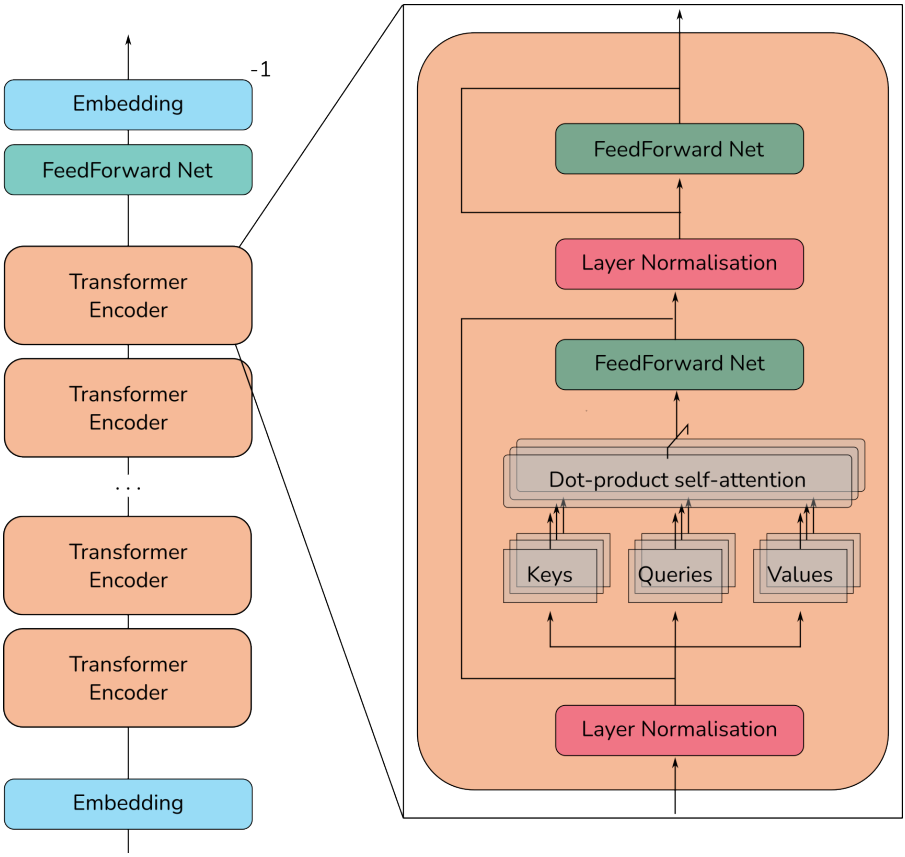
In the format provided by the authors and unedited

Species	Data source	Identifier	Notes
A. thaliana	EMBL	E-GEOD-55866	Whole-organism experiment. Samples taken 16-20 days post-anthesis.
D. melanogaster	EMBL	E-GEOD-18068	Whole-organism experiment. Samples taken from female adults.
E. coli	GEO	GSE205717	Steady state.
H. sapiens	[1]	-	Tissue-level experiment. Samples across 32 tissues were averaged, and entries with dispersion greater than 1 logTPM were removed.
H. volcanii	GEO	GSE204840	Average over untreated batches.
P. pastoris	SRA	SRR10740038	Processed using `kallisto` [2] with default parameters against the GCA_001708105.1 assembly.
S. cerevisiae	EMBL	E-MTAB-8621	

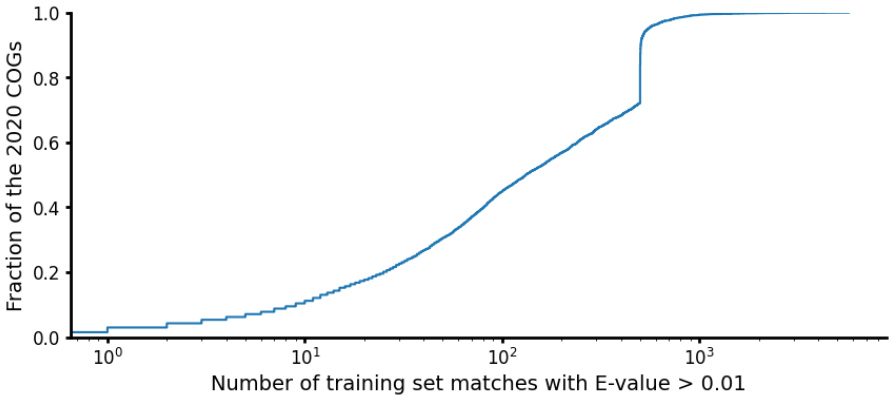
Supplementary Table 1: Transcriptomic dataset sources

Species	Data source	Identifier	Notes
A. thaliana	GenBank	GCA_000001735.1	
D. melanogaster	Ensembl	BDGP6.32	
E. coli	GenBank	GCA_000259695.1	
H. sapiens	Ensembl	GRCh38.107	
H. volcanii	GenBank	GCA_000025685.1	
P. pastoris	GenBank	GCA_001708105.1	
S. cerevisiae	GenBank	GCA_000146045.2	

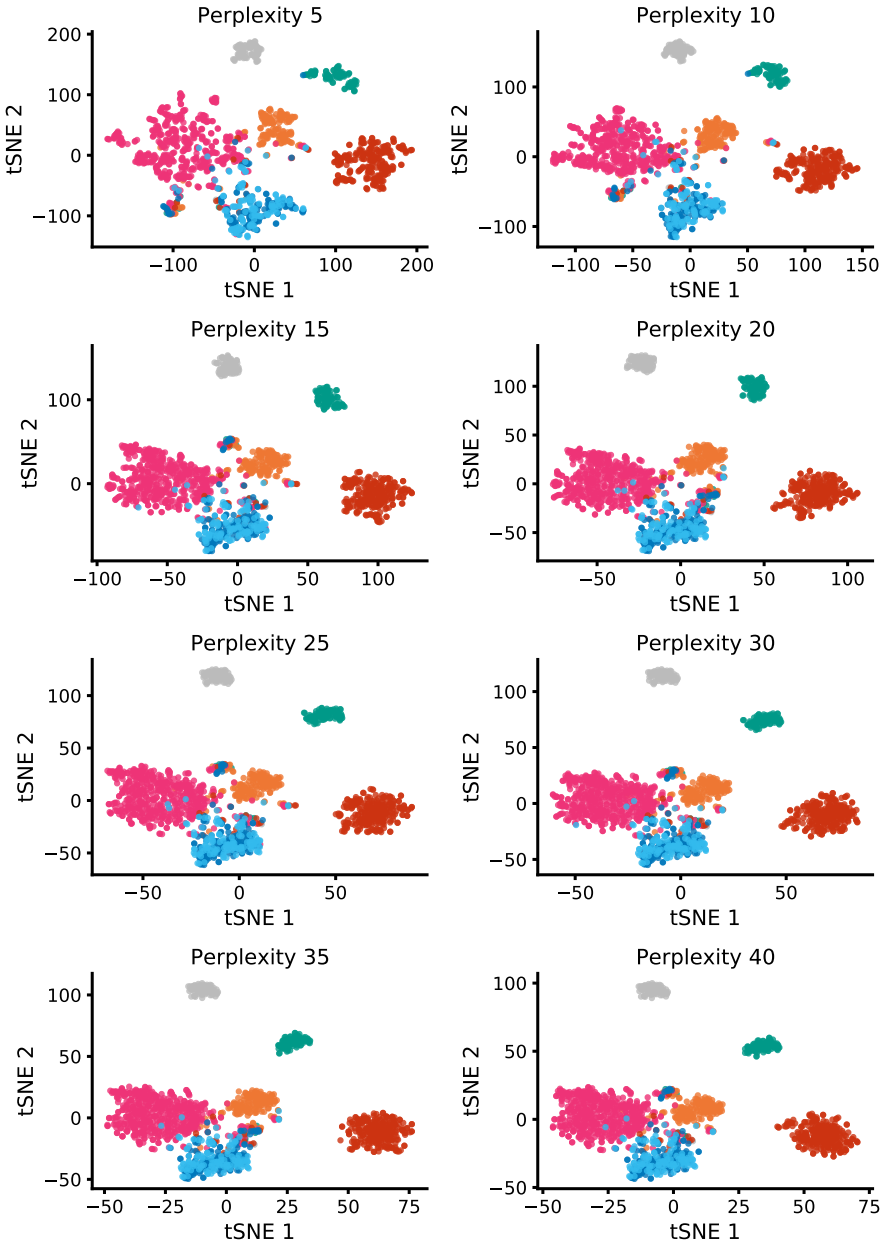
Supplementary Table 2: Assemblies



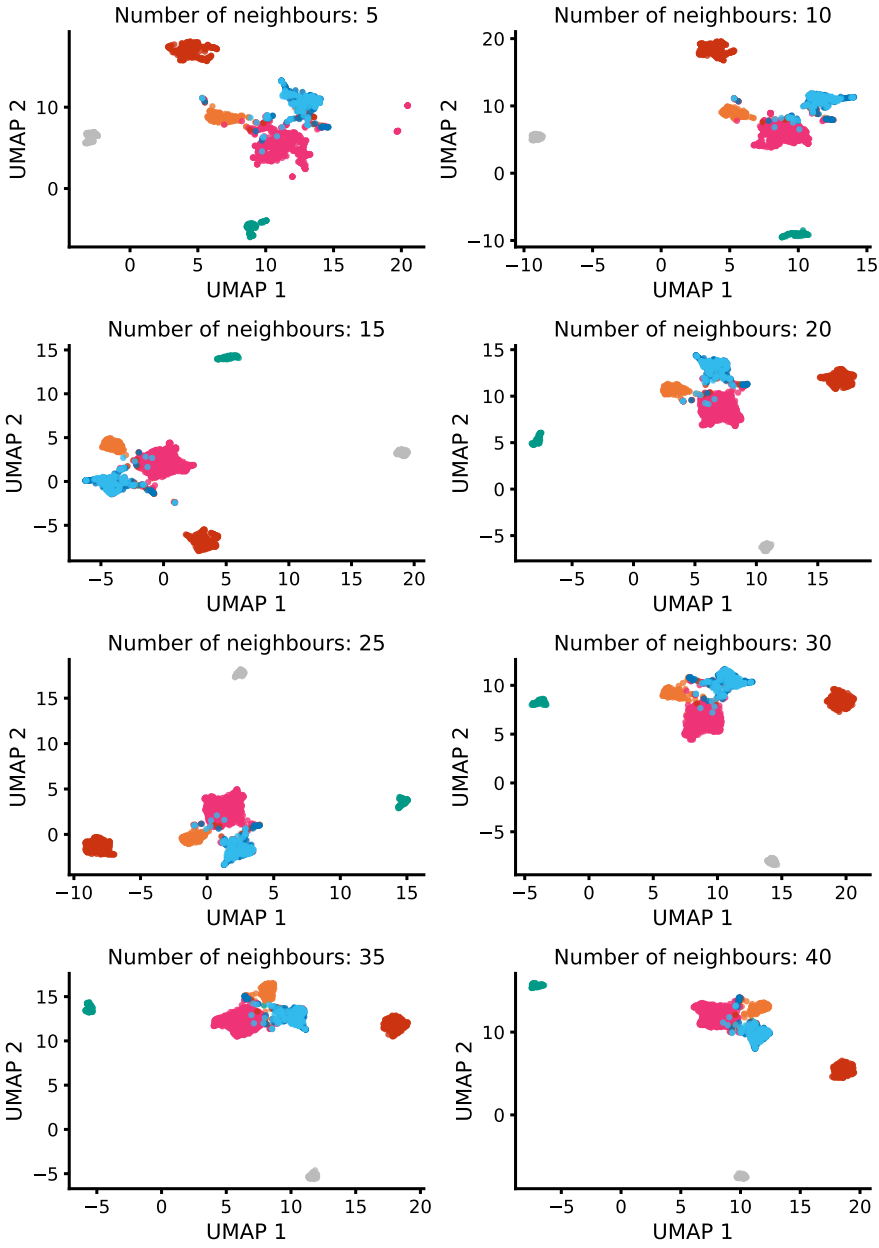
Supplementary Figure 1: Architectural diagram of the deep learning architecture used in CaLM.



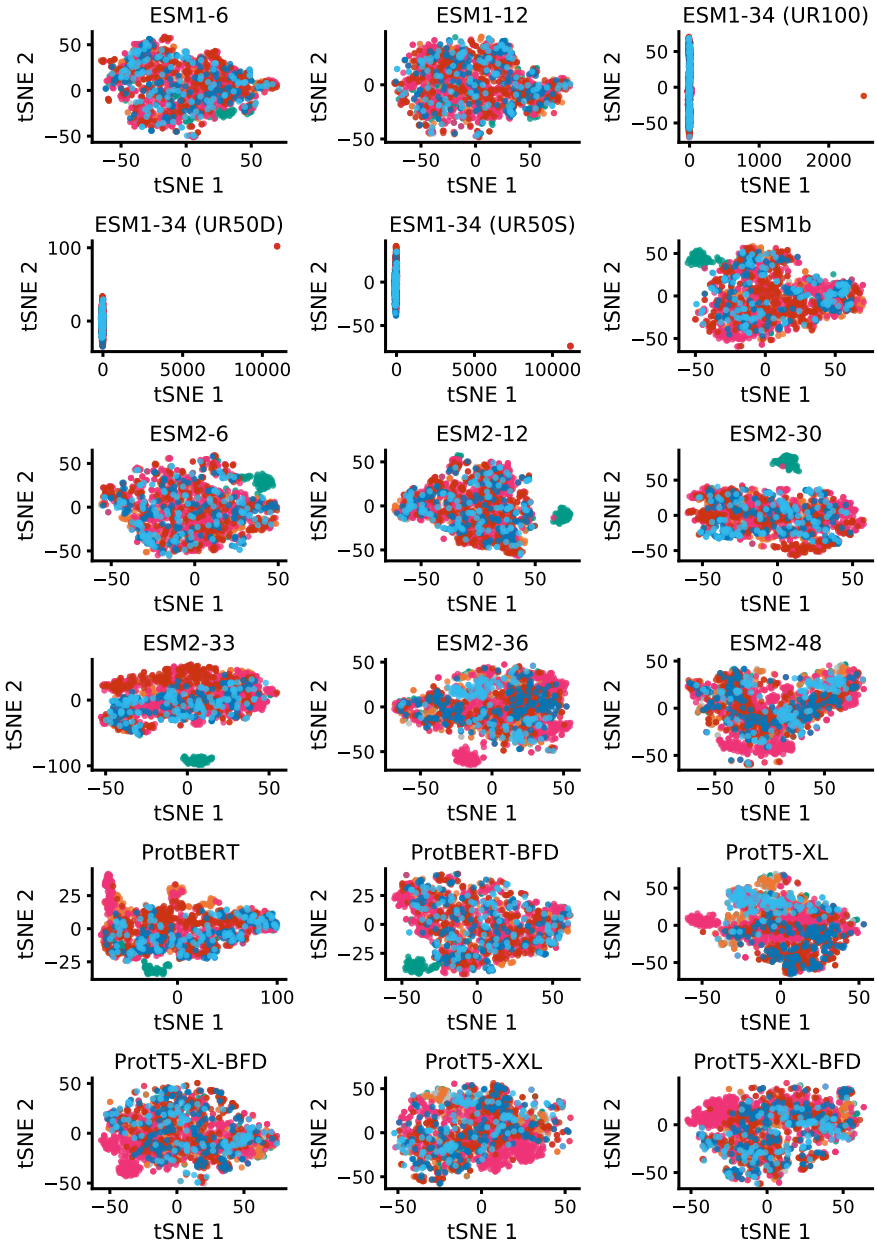
Supplementary Figure 2: **Verification that the training set contains proteins with high presence across the tree of life.** We observe matches with E-value < 0.01 for 98.5% of the COGs (lower significance matches are found for all but two of the COGs); in particular, for two thirds of the COGs we observe at least 50 matches with E-value < 0.01.



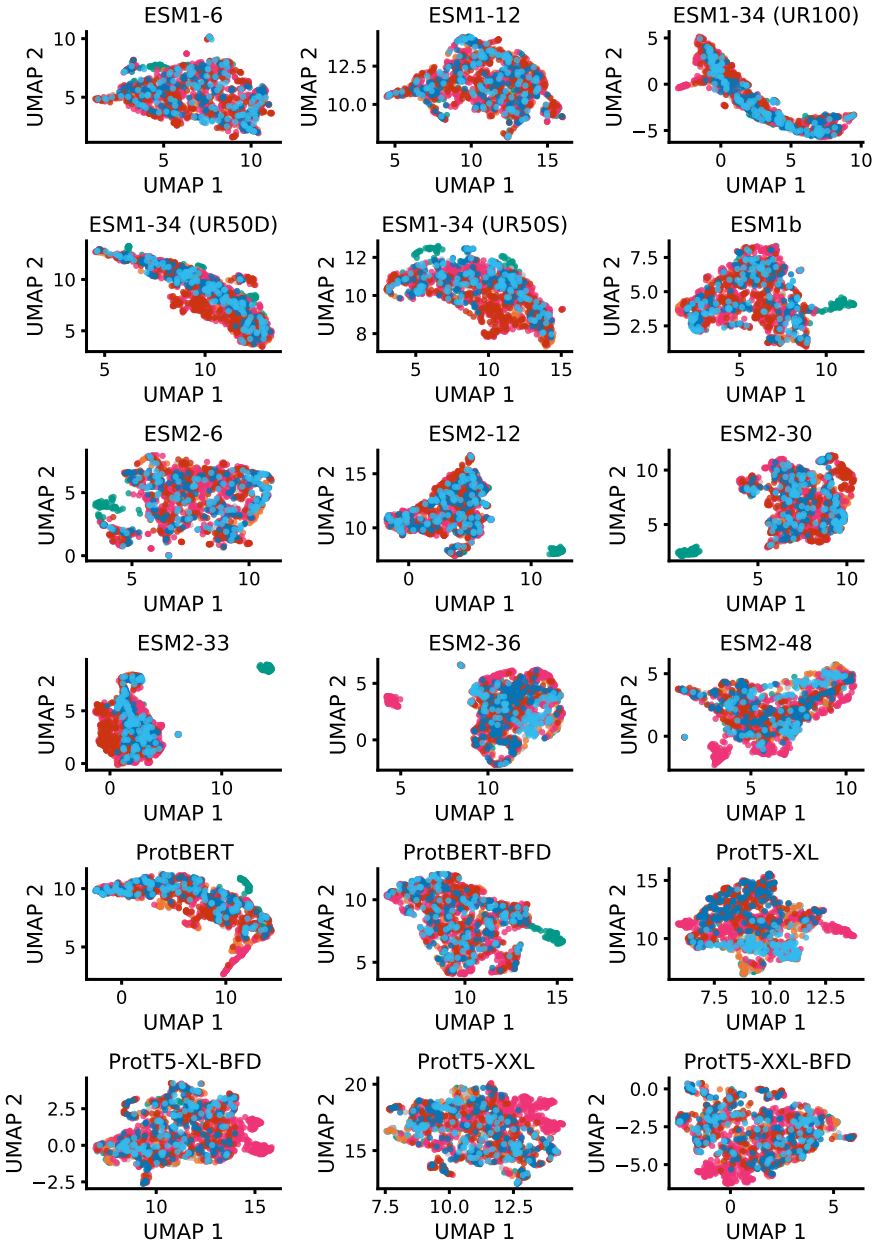
Supplementary Figure 3: **Comparison of the tSNE embedding presented in Figure 2c with different perplexity values.**



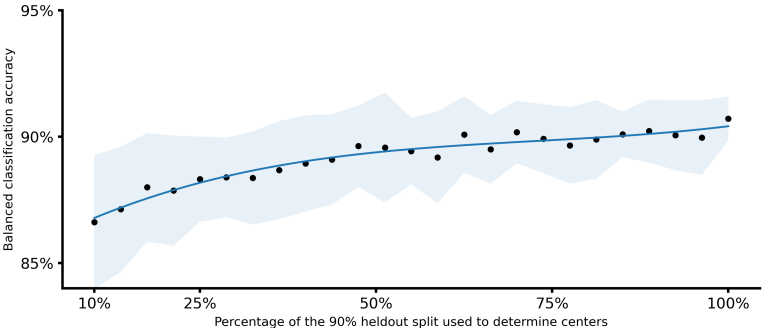
Supplementary Figure 4: Comparison of the tSNE embedding presented in Figure 2c using UMAP and different numbers of neighbours.



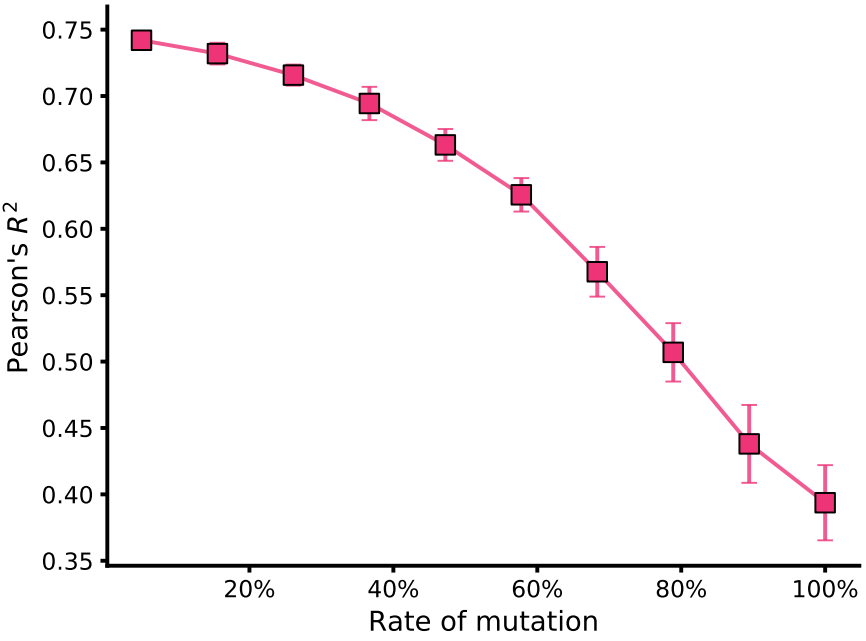
Supplementary Figure 5: Replicates of the tSNE embedding presented in Figure 2c using different amino acid language models.



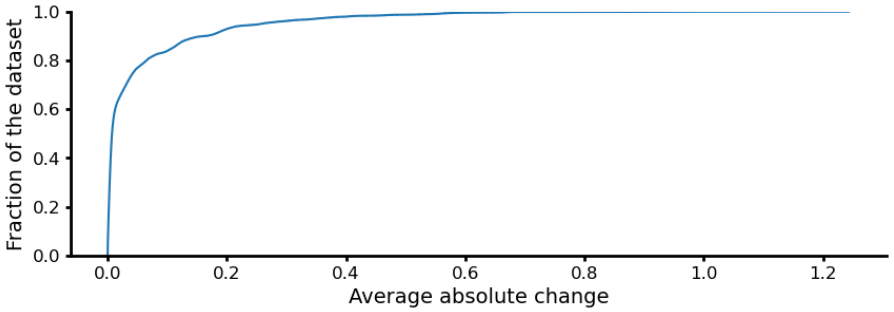
Supplementary Figure 6: Replicates of the tSNE embedding presented in Figure 2c using UMAP and different amino acid language models.



Supplementary Figure 7: **Validation of the number of examples used to train the k-nearest centers species classifier described in Section 2.3.** We split the heldout dataset into a 90% set to determine the centers, and a 10% set to validate prediction accuracy, stratifying by species. We used increasing percentages of the 90% set, between 10% and 90%, and evaluated the balanced classification accuracy; we repeated this process 25 times to determine confidence intervals. We observe that the results monotonically increase accuracy, although a representative result is obtained with 33% of the data. The blue line represents a cubic polynomial fit to the data.



Supplementary Figure 8: **CaLM's performance at predicting melting point with increasing rates of synonymous codon mutations.** The correlation between predictions and ground truth values drops by nearly half as the rate of mutations approaches 100%, suggesting that codon usage information is fundamental for CaLM's performance. Data are presented as mean values with error bars representing the standard deviation calculated from 5-fold cross-validation



Supplementary Figure 9: **Average absolute change in the prediction when individual codons are mutated to their synonymous alternatives.** Over 90% of the dataset experiences a change of less than 0.2 units (for reference, the standard deviation of the dataset is 2.06 units). This result supports the hypothesis that the model is learning global features of the sequences..

References

[1] Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., *et al.*: Tissue-based map of the human proteome. Science **347**(6220), 1260419 (2015)

[2] Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic RNA-seq quantification. Nature biotechnology **34**(5), 525–527 (2016)