RSDexport

Yichen Han

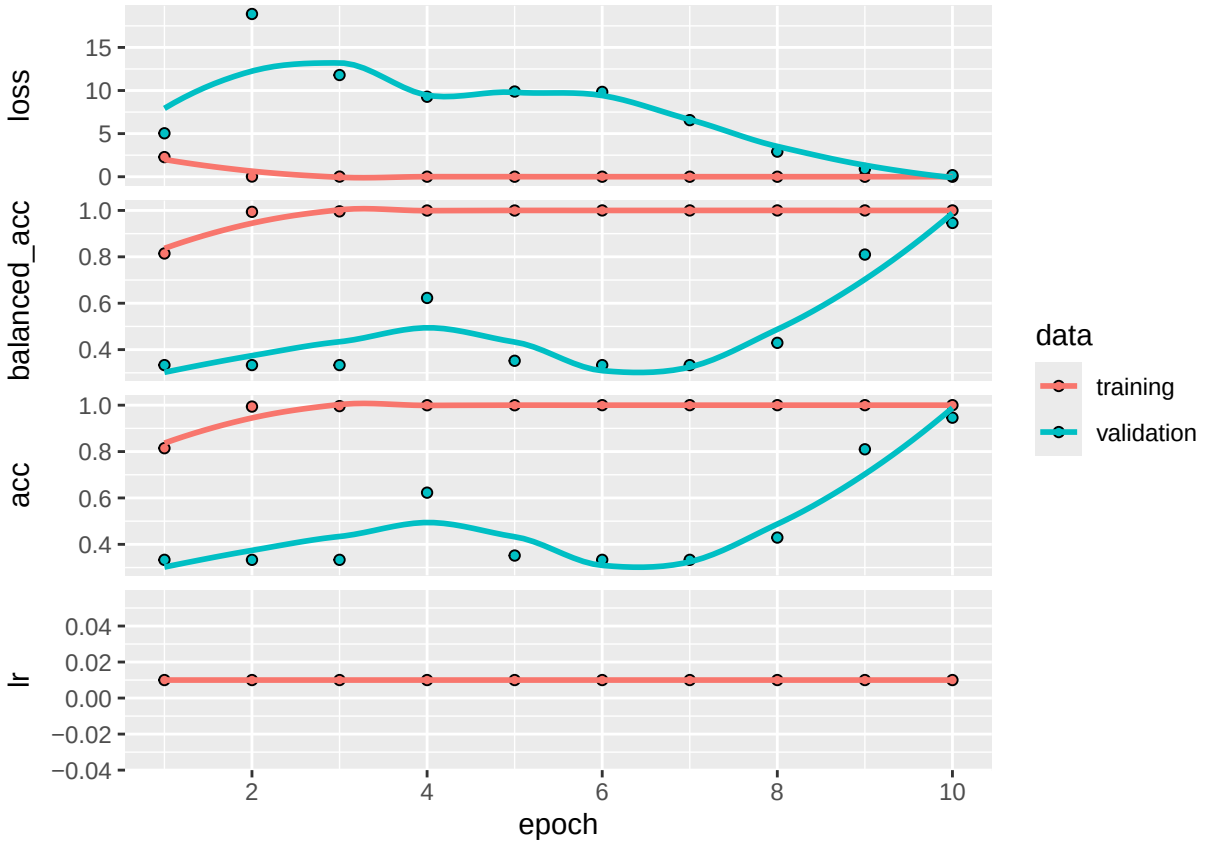
2024-08-11

Experiment 39

Spec.region excluded for "abnormal". Functional change not guaranteed.

Call: `GenePermutation(triplets, keyed, num.perm=15000, min.subs=10, max.subs=30, spec.region=30:60)`

Model: `maxlen = 477, batch_size = 64, steps_per_epoch = 45, epochs = 10, step = c(1, 1, 1)`




```
## Using checkpoint checkpoints/rsd-permutation_39/Ep.010-val_loss0.18-val_acc0.946.hdf5
```

Model Evaluation

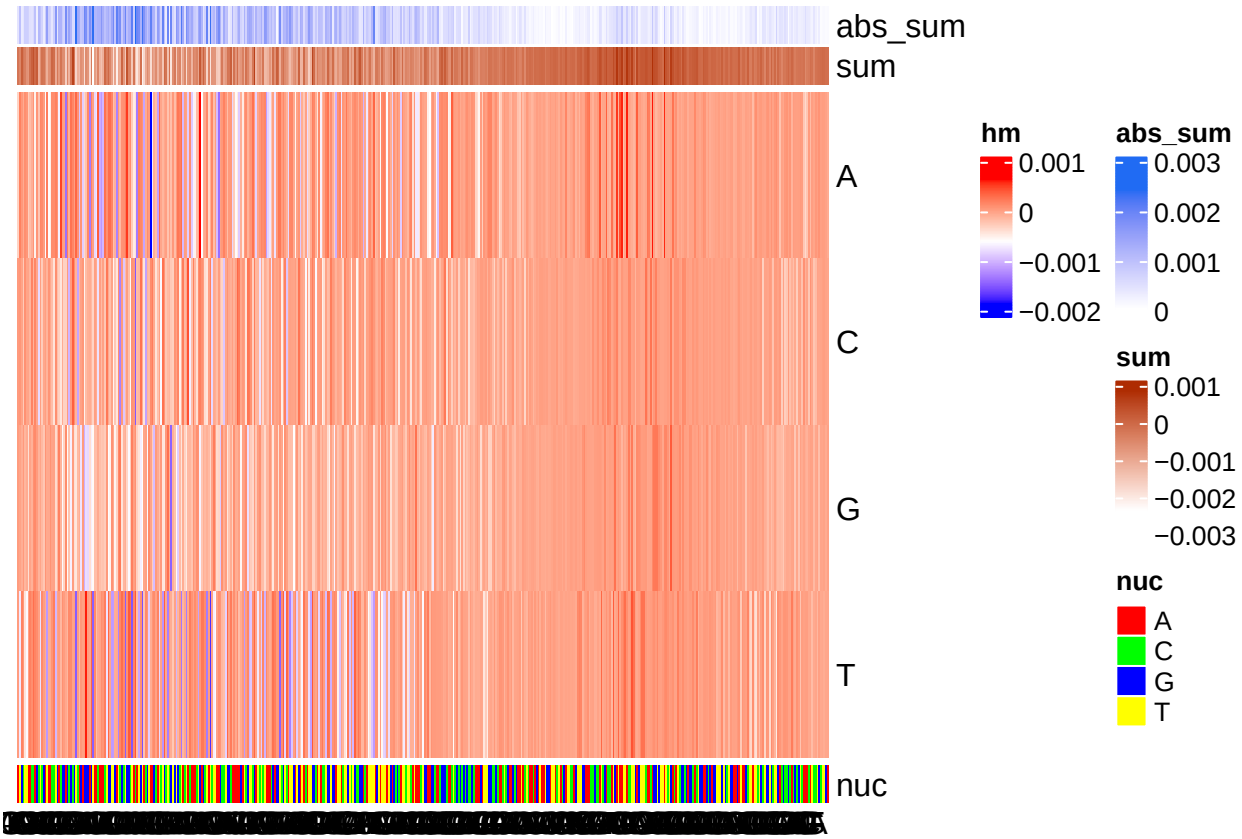
```
## [[1]]
## [[1]]$confusion_matrix
```

```
##           Truth
## Prediction normal abnormal special
##   normal      254       82       1
##   abnormal      0      163       0
##   special       2       11     255
##
## [[1]]$accuracy
## [1] 0.875
##
## [[1]]$categorical_crossentropy_loss
## [1] 0.6876215
##
## [[1]]$AUC
## NULL
##
## [[1]]$AUPRC
## NULL
```

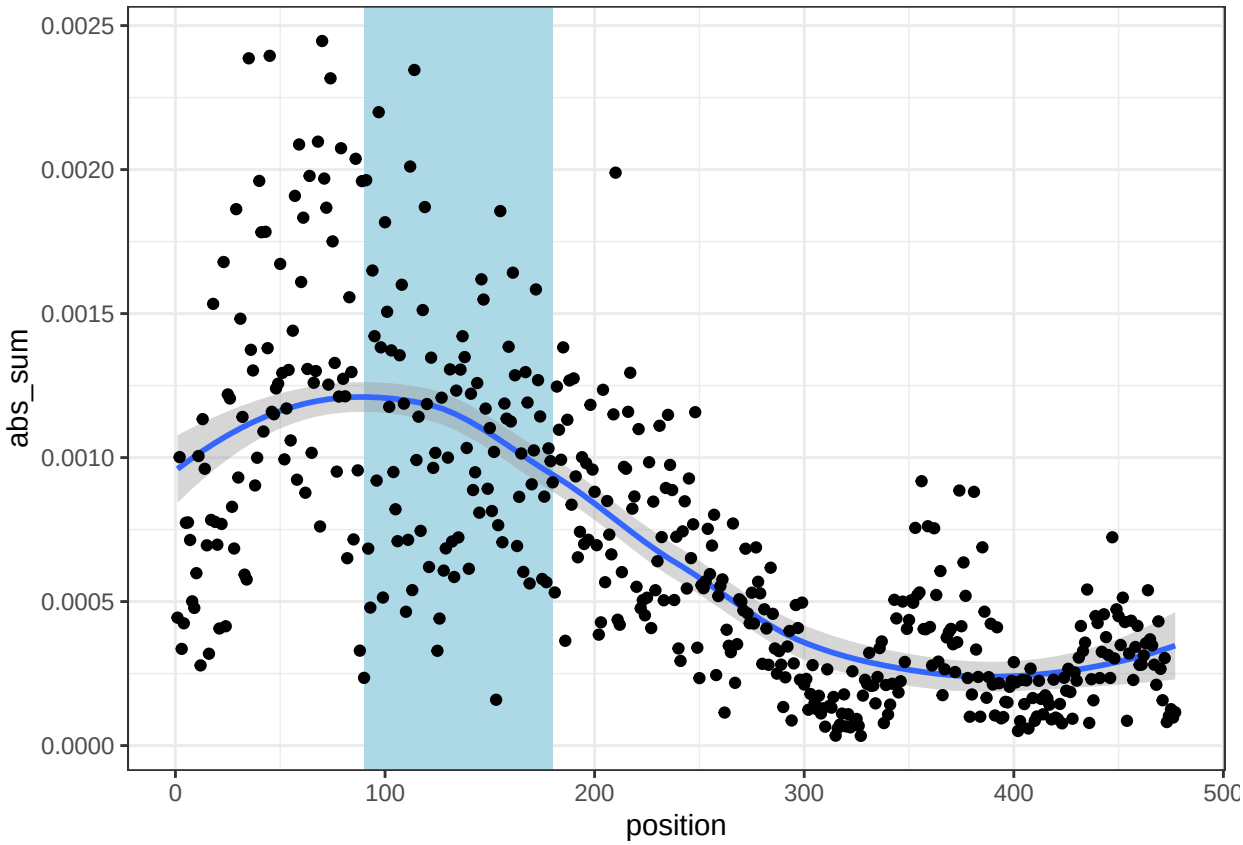
```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    0    0    1
## [3,]    0    0    1    0
## [4,]    0    1    0    0
## [5,]    0    0    0    1
## [6,]    0    0    0    1
```

```
##              [,1]         [,2] [,3]
## [1,] 2.989441e-09 3.559504e-18    1
```

```
## [[1]]
```




```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



We notice high importance assigned to the starting region of the sequence, despite most substitutions happening between AA 30 to AA 60.

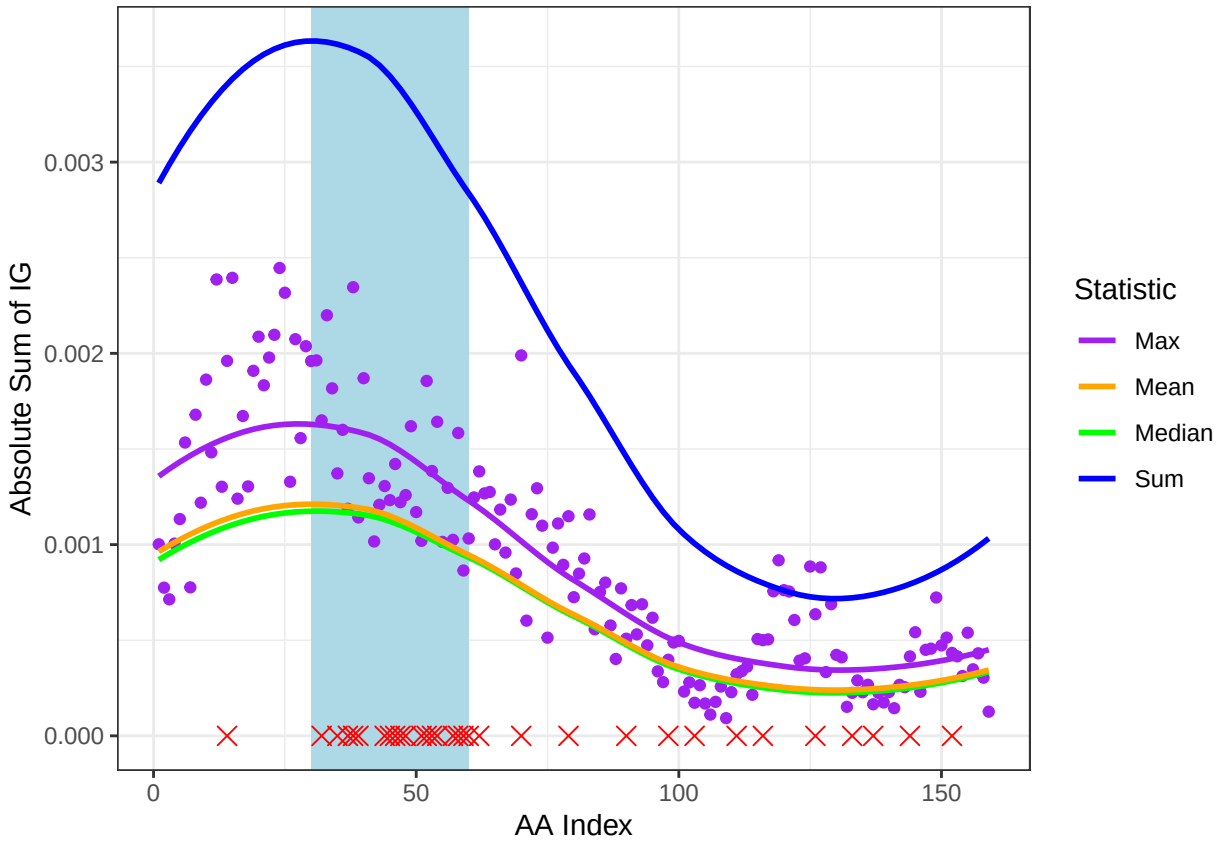
We speculate that it flows in due to the shuffled baseline – random shuffling almost certainly causes non-synonymous substitution across the sequence. It is enough for the model to look at the beginning to know if it is "abnormal" or not. Ideally, it should be compared with a normal sequence and only the blue marked area should have high importance.

This plot is enough proof for the need of improving the baseline.

Alternative Display

We combine every 3 loci and use methods similar to pooling to aggregate the information. This captures the codon structure. For importance measured by absolute locus gradients sum, all pooling methods have a similar effect.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

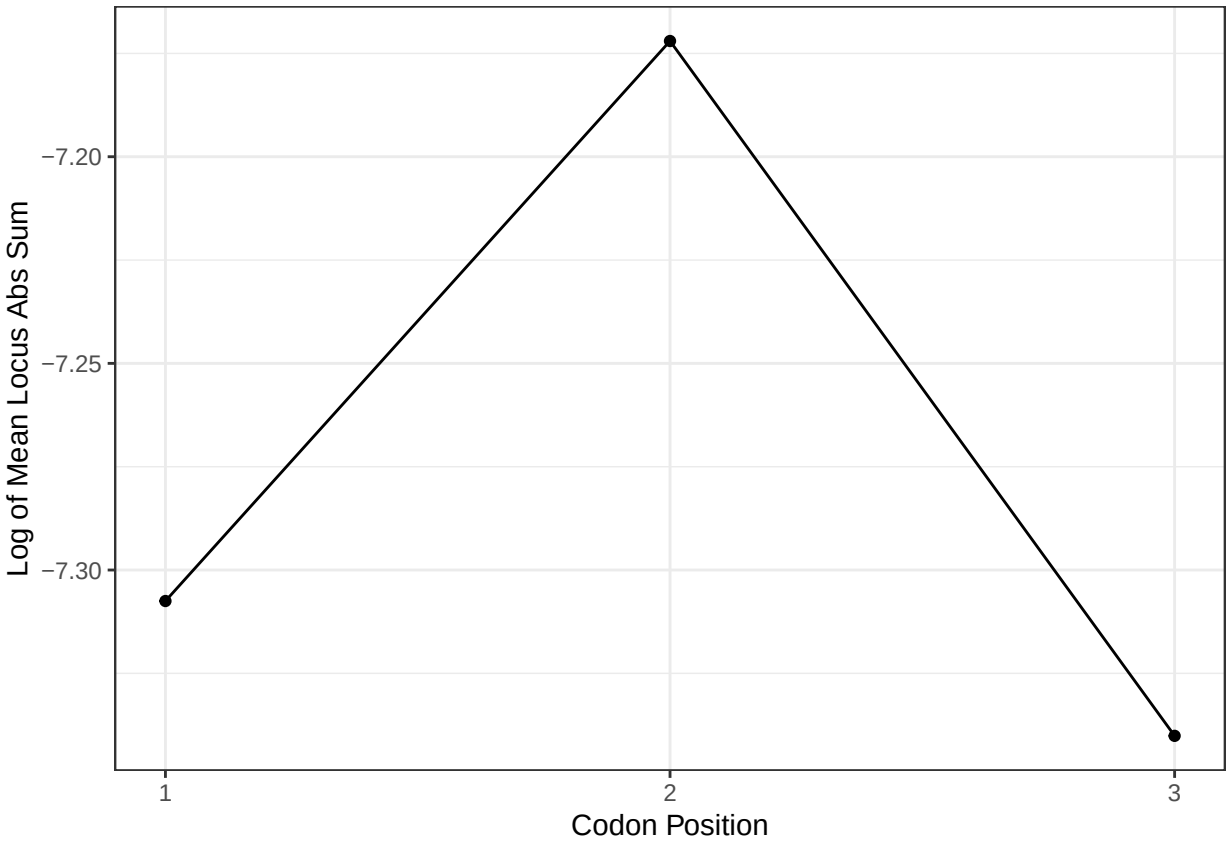


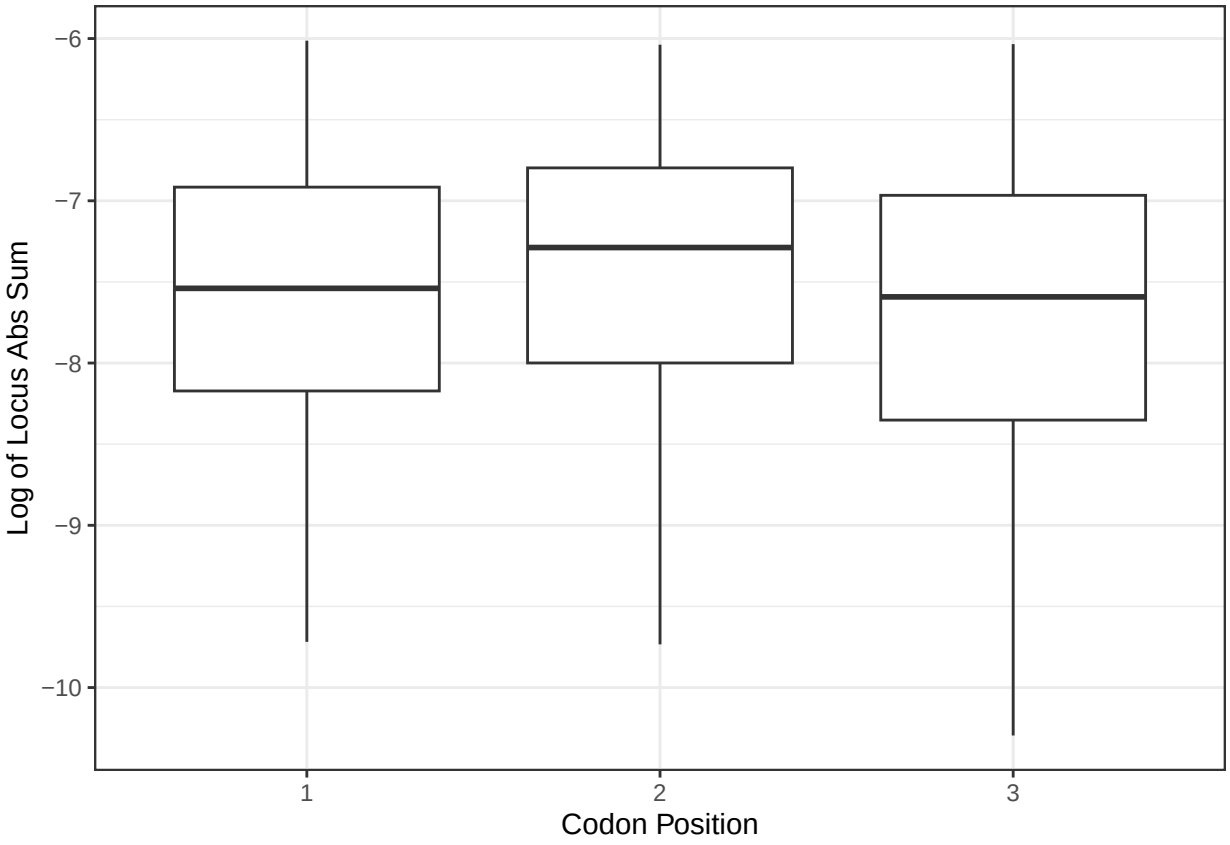
Ideas based on Durrant & Bhatt (2021)

The paper documented average importance of the 3 loci within a codon and found out the highest importance is likely to be assigned to the second locus, while the third (wobble position) receives the least importance.

We were able to replicate this output here, simply based on our synthetic data. However, the paper did not test if the difference is statistically significant. According to our test result, this effect could still be random.

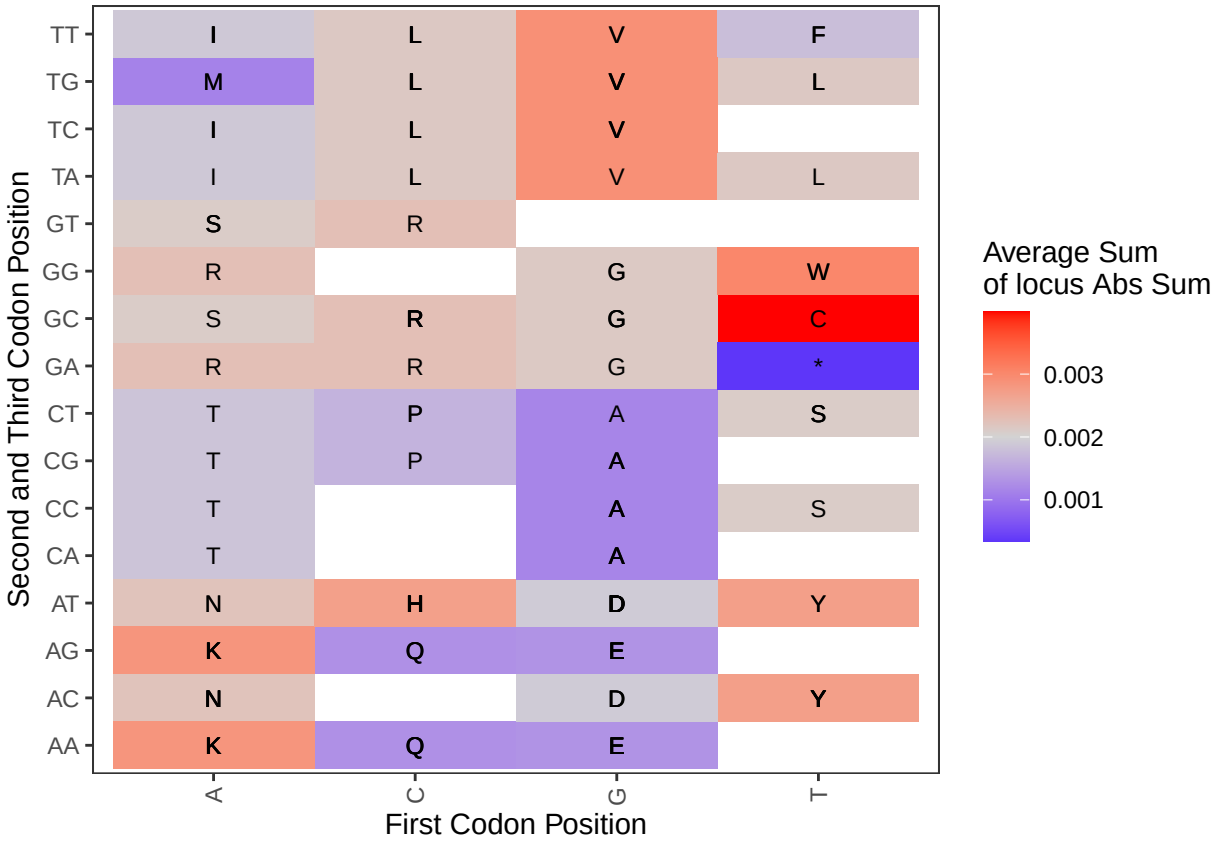
The possible capturing of wobleness, or special emphasis on the second locus, however, indicates some inner workings of our genomic model.


Log of Mean Locus Abs Sum
−7.20
−7.25
−7.30
1
2
3
Codon Position

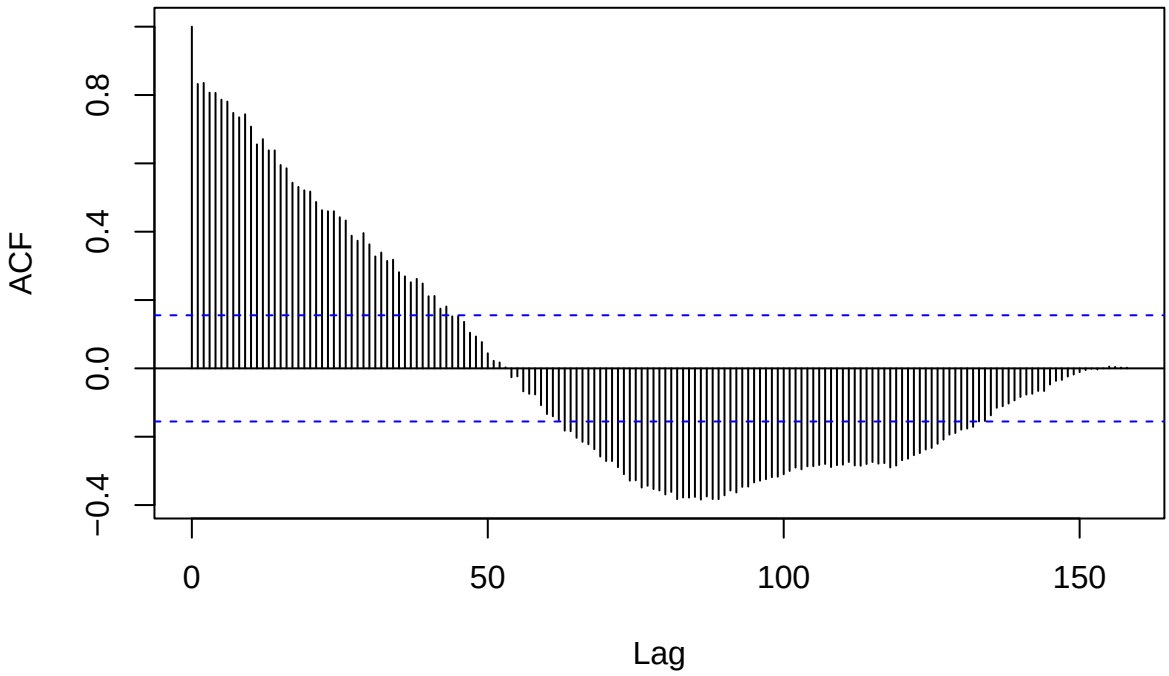

Log of Locus Abs Sum
−6
−7
−8
−9
−10
1
2
3
Codon Position

```
##
## Call:
## lm(formula = log(abs_sum) ~ factor(position), data = codon_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62161 -0.55317  0.09533  0.70736  1.63964
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -7.63479    0.06965 -109.624   <2e-16 ***
## factor(position)2  0.15296    0.09849    1.553    0.121
## factor(position)3 -0.03918    0.09849   -0.398    0.691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8782 on 474 degrees of freedom
## Multiple R-squared:  0.008887,   Adjusted R-squared:  0.004705
## F-statistic: 2.125 on 2 and 474 DF,  p-value: 0.1205
```

The paper also created a heatmap for the average importance for each amino acid across the whole sequence. This method is here replicated. Due to the nature of our synthetic data, it cannot be of use in this report, but might be helpful later.



We noticed strong autocorrelation among the averange absolute sums. We speculate this to be a model artifact.



The paper proposed a measurement to quantify the similarity of assigned importance by the IML method to triplets coding the same amino acid.

Codon Synonym Similarity

CSS score by Durrant & Bhatt:

$$CSS := \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}_i$$

Null distribution approximated by random AA substitution.

We argue that this method has many flaws. First of all, the importance measure is unbounded and scaled specific to data and model. The sample variance is scale-sensitive, so that the quantity is not comparable between scenarios. Secondly, CSS score calculates the mean grouped by amino acid across the whole sequence, regardless of their position. In most cases, position on the DNA strand matters more than its corresponding amino acid. Thus, this measure does not capture the positional importance of the amino acid, and delivers less valuable information.

CSS score using coefficient of variance: to correct the scaling problem, we use coefficient of variation as the metric.

$$CSS_{CV} := \frac{1}{k}\sum_{i=1}^{k}\frac{\hat{\sigma}_i}{\hat{\mu}_i}$$

To further improve the method, we propose the following.

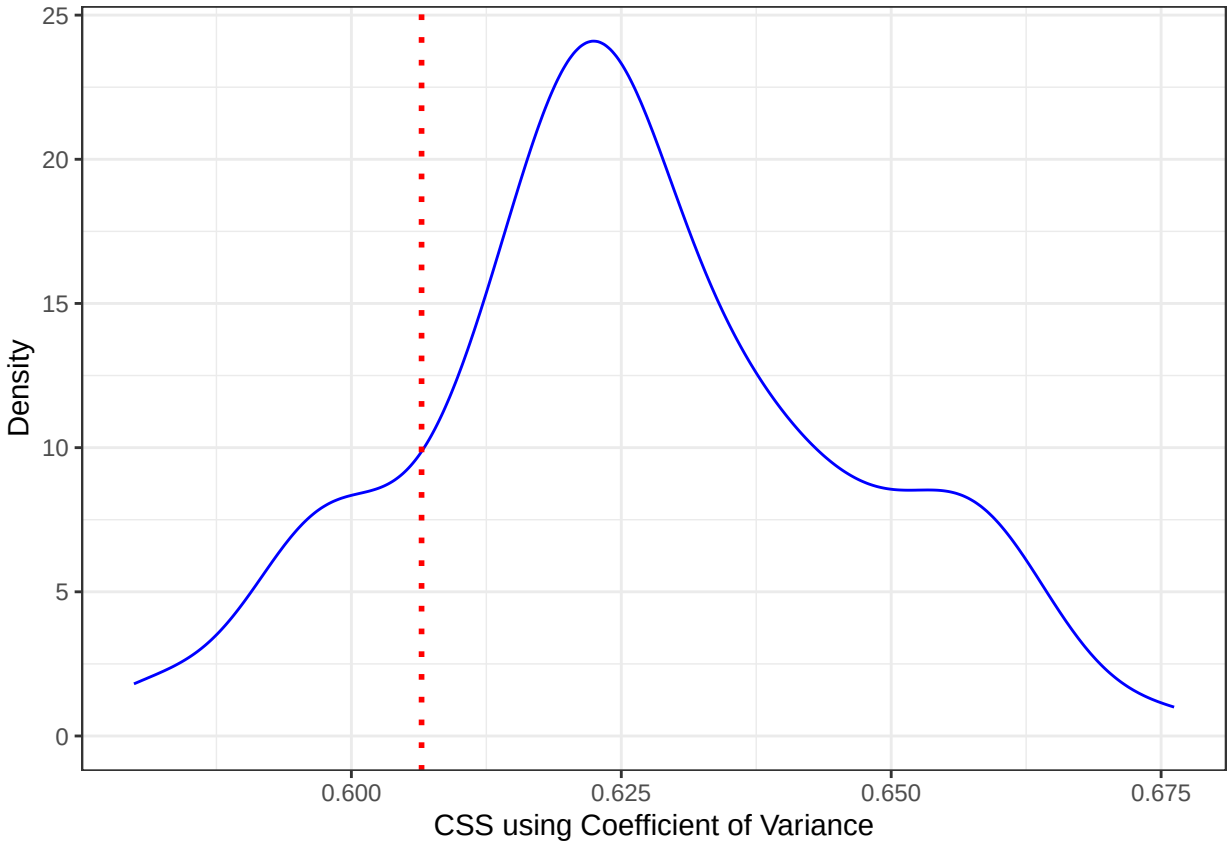
The metric should demonstrate how reliably the IML method / the model treats synonymous codons, given that they are on the same position. It should be bounded between 0 and 1, making the quantity easy to interpret and comparable across models and interpretation methods.

Positional Codon Synonym Consistency score: perform ok-mutate based on instance. Group aggregated IG scores (here using sum of absolute gradients sum) by their AA position, and calculate the standard

deviation of the mean IG score for each triplet, corrected by its mean and scaled using inverse transformation. => No distribution needed, if near 1, then low variability / high consistency.

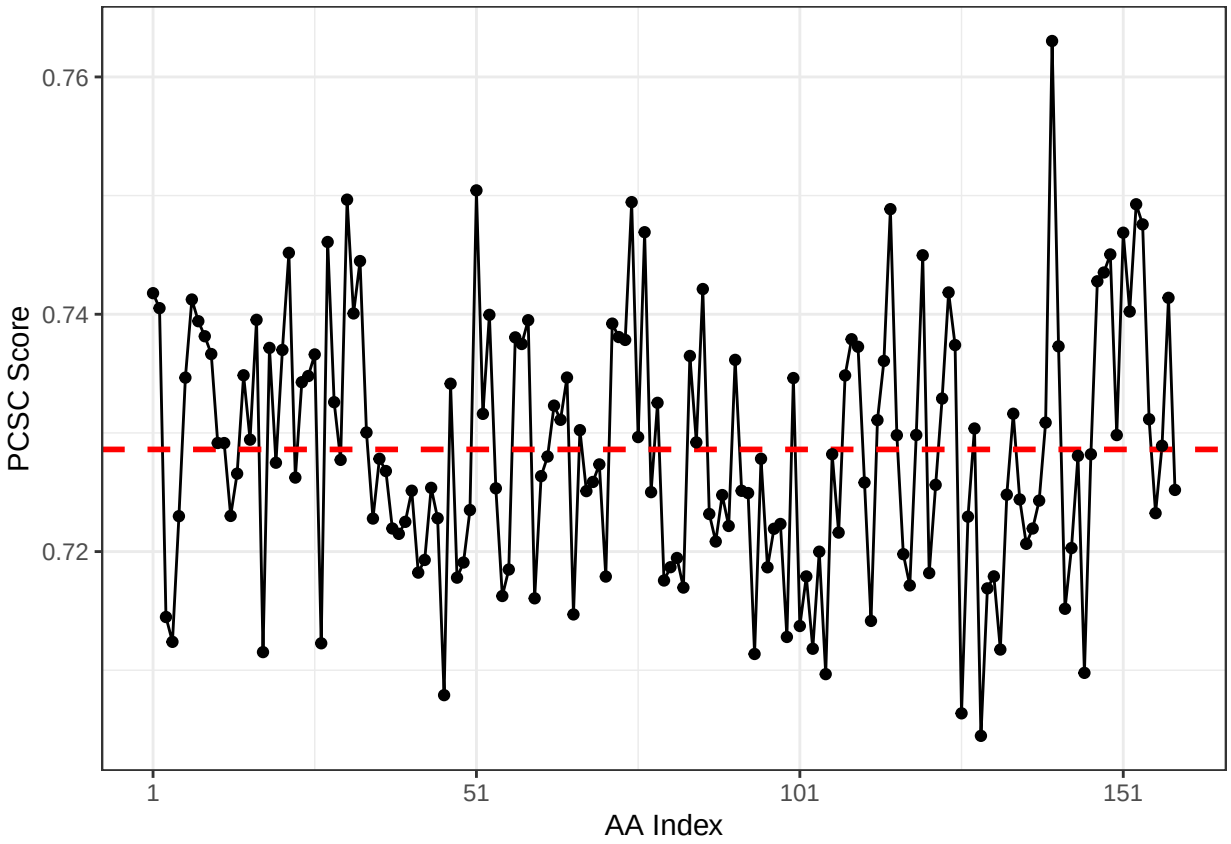
$$\text{PCSC} := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha\cdot\hat{c}_v}, \quad \hat{c}_v = \frac{\hat{\sigma}_i}{\hat{\mu}_i}, \; N = \frac{n}{3}$$

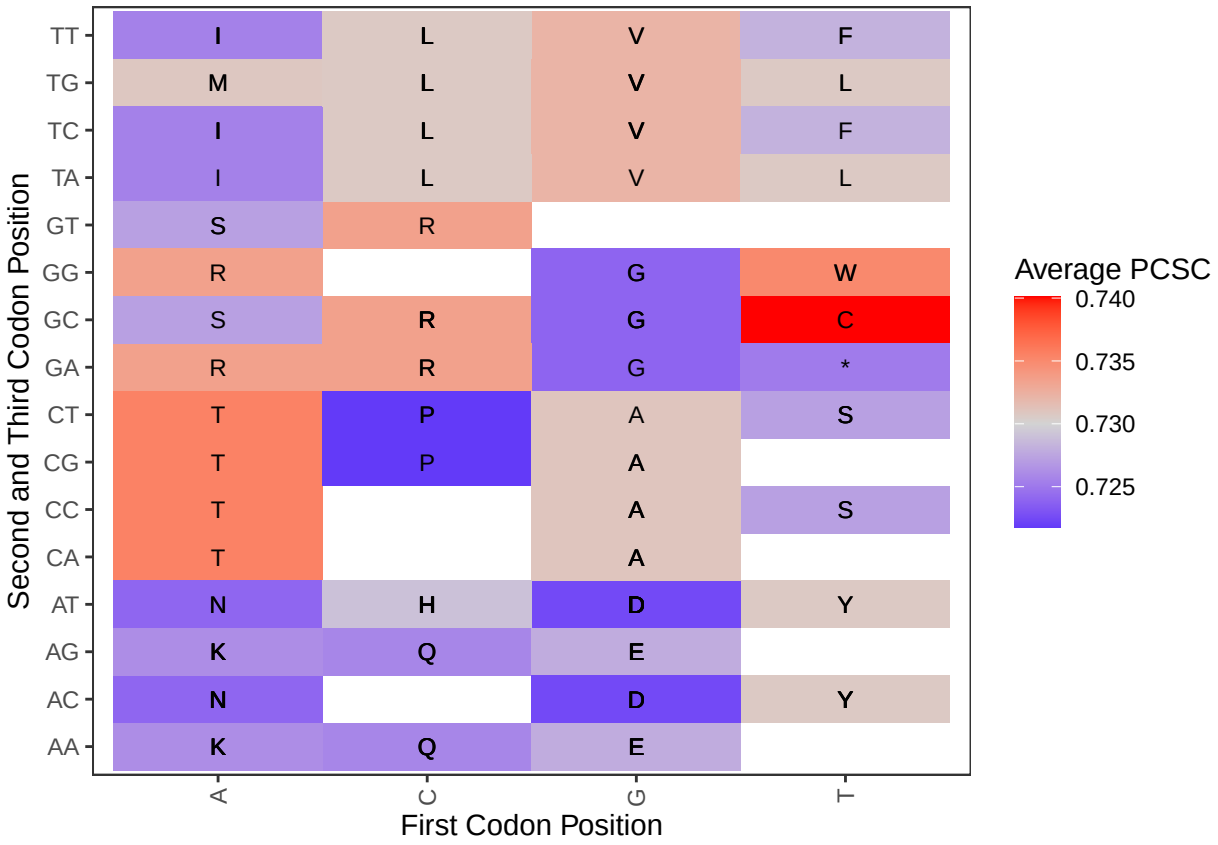
We also implemented the paper's metric here.



Here begins our metric.

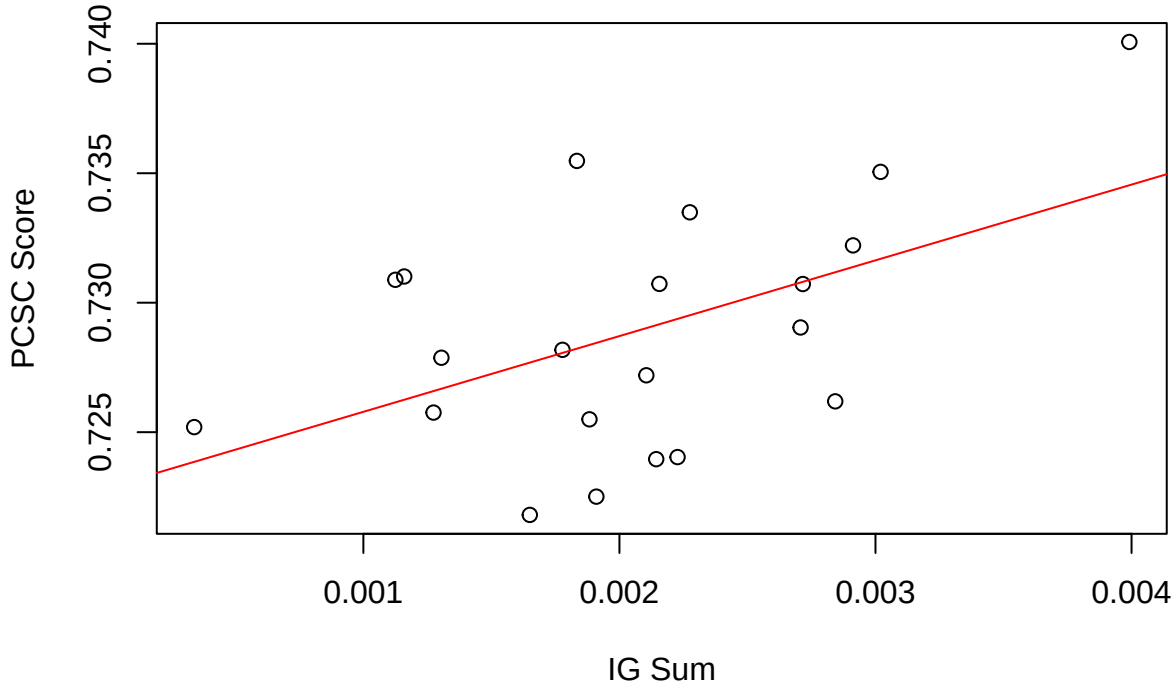
```
## Loaded result_df from existing CSV file.
```


PCSC Score
0.76
0.74
0.72
1
51
101
151
AA Index


Second and Third Codon Position
TT I L V F
TG M L V L
TC I L V F
TA I L V L
GT S R
GG R G W
GC S R G C
GA R R G *
CT T P A S
CG T P A
CC T A S
CA T A
AT N H D Y
AG K Q E
AC N D Y
AA K Q E
A C G T
First Codon Position
Average PCSC
0.740
0.735
0.730
0.725

We noticed some correlation between AA feature importance and AA importance consistency. We speculate this to be a random effect and more of a model artifact.

```
##
##  Pearson's product-moment correlation
##
## data:  data_cor$value.y and data_cor$value.x
## t = 2.554, df = 19, p-value = 0.01939
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09447585 0.76933420
## sample estimates:
##       cor
## 0.5055445
```




```
##
## Call:
## lm(formula = data_cor$value.y ~ data_cor$value.x)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0059364 -0.0028727  0.0001192  0.0033585  0.0072505
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.722861   0.002532 285.526   <2e-16 ***
## data_cor$value.x 2.925000   1.145249   2.554   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004148 on 19 degrees of freedom
```

```
## Multiple R-squared:  0.2556, Adjusted R-squared:  0.2164
## F-statistic: 6.523 on 1 and 19 DF,  p-value: 0.01939
```