

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Tatsuhiko Naito

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013 Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed



Ensemble Method for Predicting the Effects of Mutation in MaPSy

Tatsuhiko Naito

July 6, 2018



Department of Neurology
Graduate School of Medicine
The University of Tokyo



Department of Neurology
Tokyo Teishin Hospital

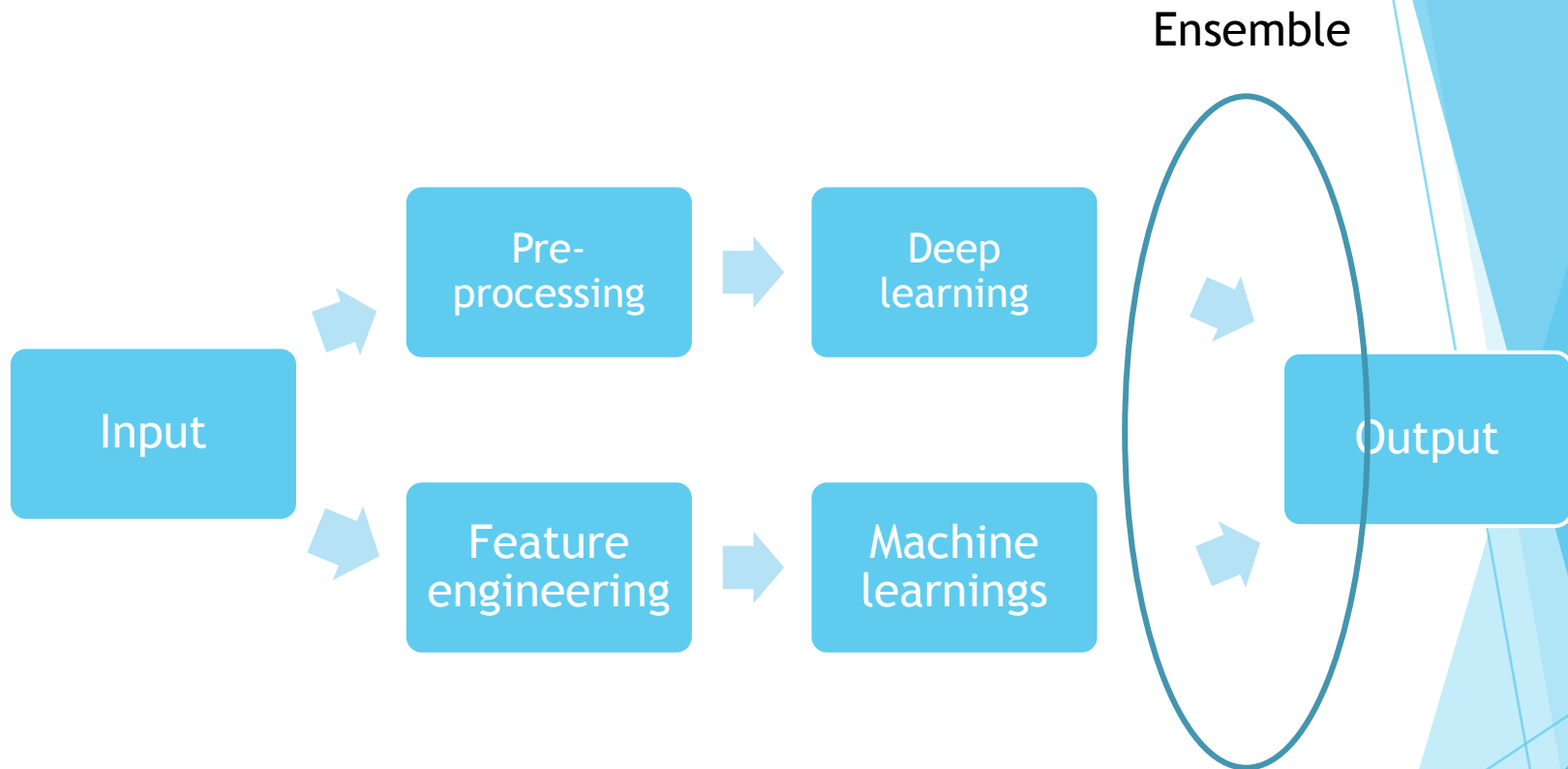
Content

- ▶ Task and data
- ▶ Method
 - ▶ Outline
 - ▶ Deep learning
 - ▶ Feature engineering, machine learning
 - ▶ Ensemble
- ▶ Result
- ▶ Discussion

Task and data

- ▶ From sequence info, make a prediction whether a mutation causes splicing abnormalities in Mapsy
 - ▶ Probability of exonic splicing mutations (ESMs)
 - ▶ ~~Log2~~ vivo/vitro allele ratio (AR)
- ▶ Data
 - ▶ 453 ESM data (positive), 4511 non-ESM data (negative)
 - ▶ Removed misclassified positive data
 - ▶ Imbalanced data -> adjusted with random over-sampling
 - ▶ 10-fold cross-validation

Method: Outline



Method: Deep learning

- ▶ Data pre-processing

- ▶ Removed the leftmost and rightmost 15 nucleotides

- i.e. CTTATGCGAACGTGATAGGGGCCA GGGCTTTGGAAAGGATTGCTGCTC

- ▶ One-hot-encoding on the sequence

- i.e. A -> (1,0,0,0,0), C -> (0,1,0,0,0), **Mutation** -> (0,0,0,0,1)

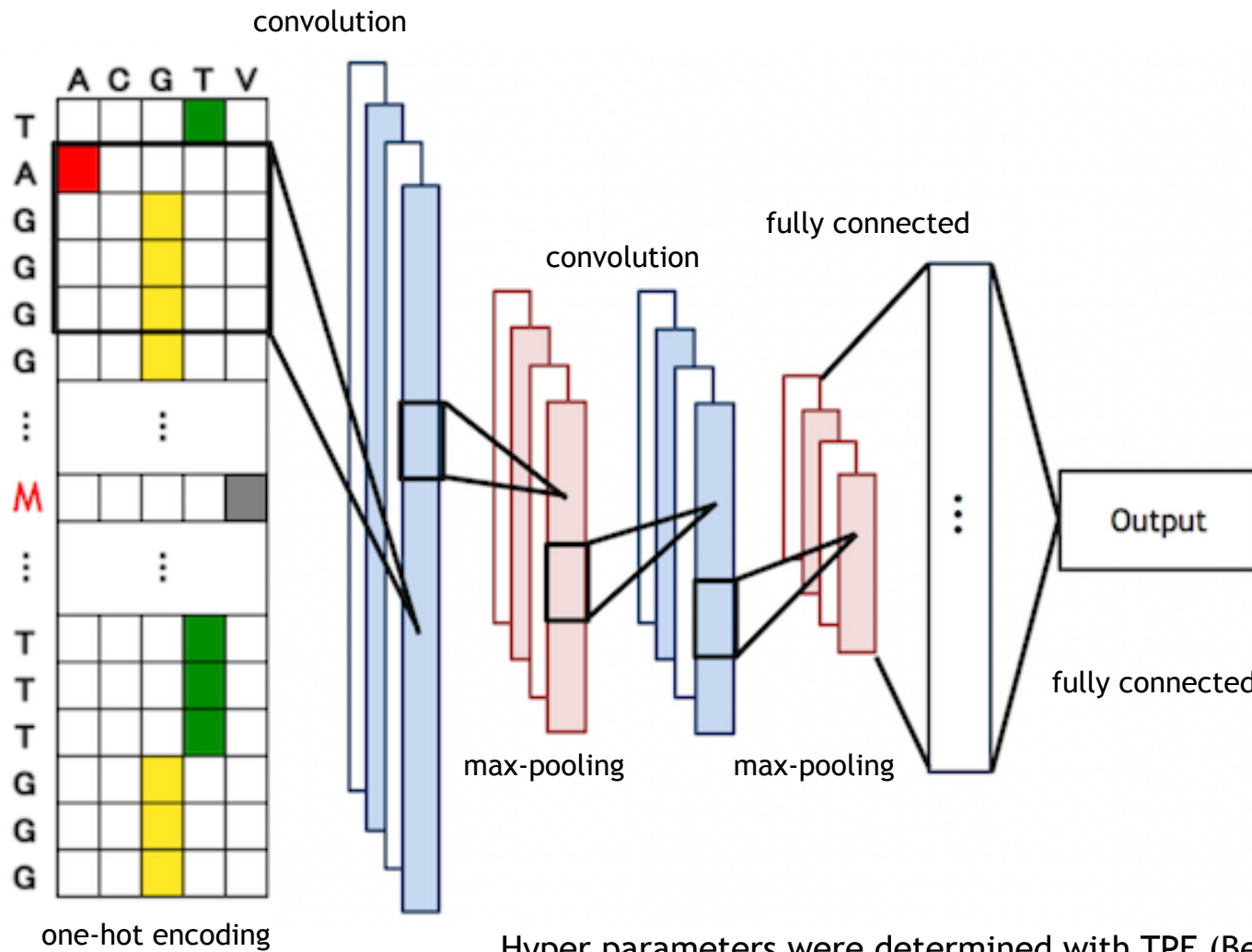
- ▶ Loss function

- ▶ Cross entropy for probability of ESM

- ▶ Mean squared error for AR

- ▶ Optimization: Adam (Kingma, 2014)

► Architecture



Hyper parameters were determined with TPE (Bergstra, 2015)

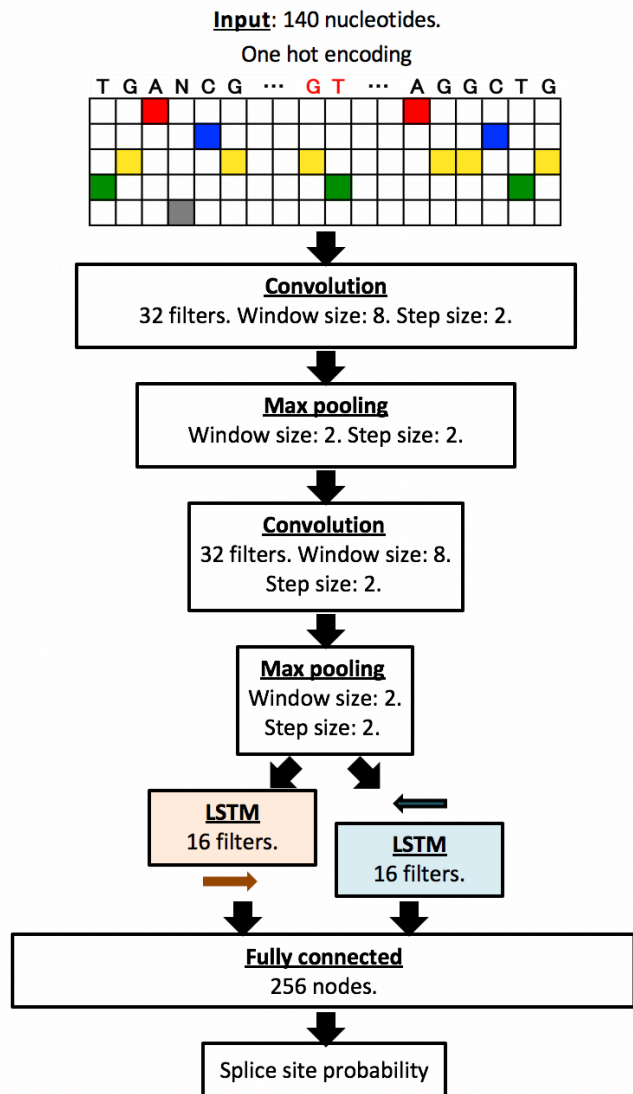
Method: Feature engineering

- ▶ Extracted feature variables manually
- ▶ Applied them to following machine learning methods
 - ▶ Random forests
 - ▶ Logistic regression (Linear regression)
 - ▶ AdaBoost
 - ▶ XGBoost

► Feature variables

- Exon len: length of exon
- Exon GC rate: GC rate of exon
- Exon num: number of exons composing the gene
- Exon pos: exon position number divided by the exon num
- Mt dist: the distance from the mutation point to the nearer splicing site
- DS/AS MaxEnt: splice site score of DS/AS calculated with MaxEntScan (Yeo, 2004)
- ESRseq: Hexamer splicing scores of WT (Shengdong, 2011)
- ESRseq diff: WT/MT difference of ESRseq
- ESRseq 5'/3' intron: ESRseq of 5'/3'intron
- HI score/proba: haploinsufficiency score and probability (Huang et al, 2010)
- MFE: the maximum free energy computed with ViennaRNA packages (Package et al, 2011)
- MFE diff: WT/MT difference of MFE
- DS/AS proba: probability of DS/AS of a MT calculated with DSSP
- DS/AS proba diff: WT/MT difference of DS/AS proba

► Splice site probability calculated with DSSP

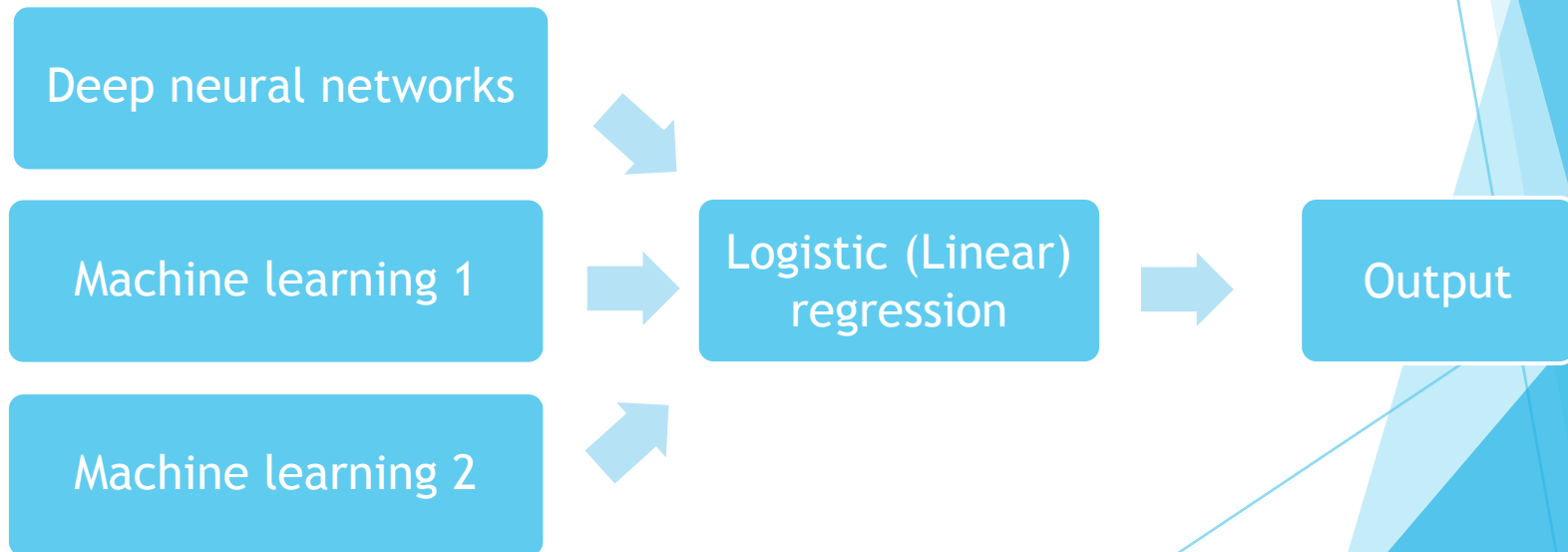


(Naito, 2018)

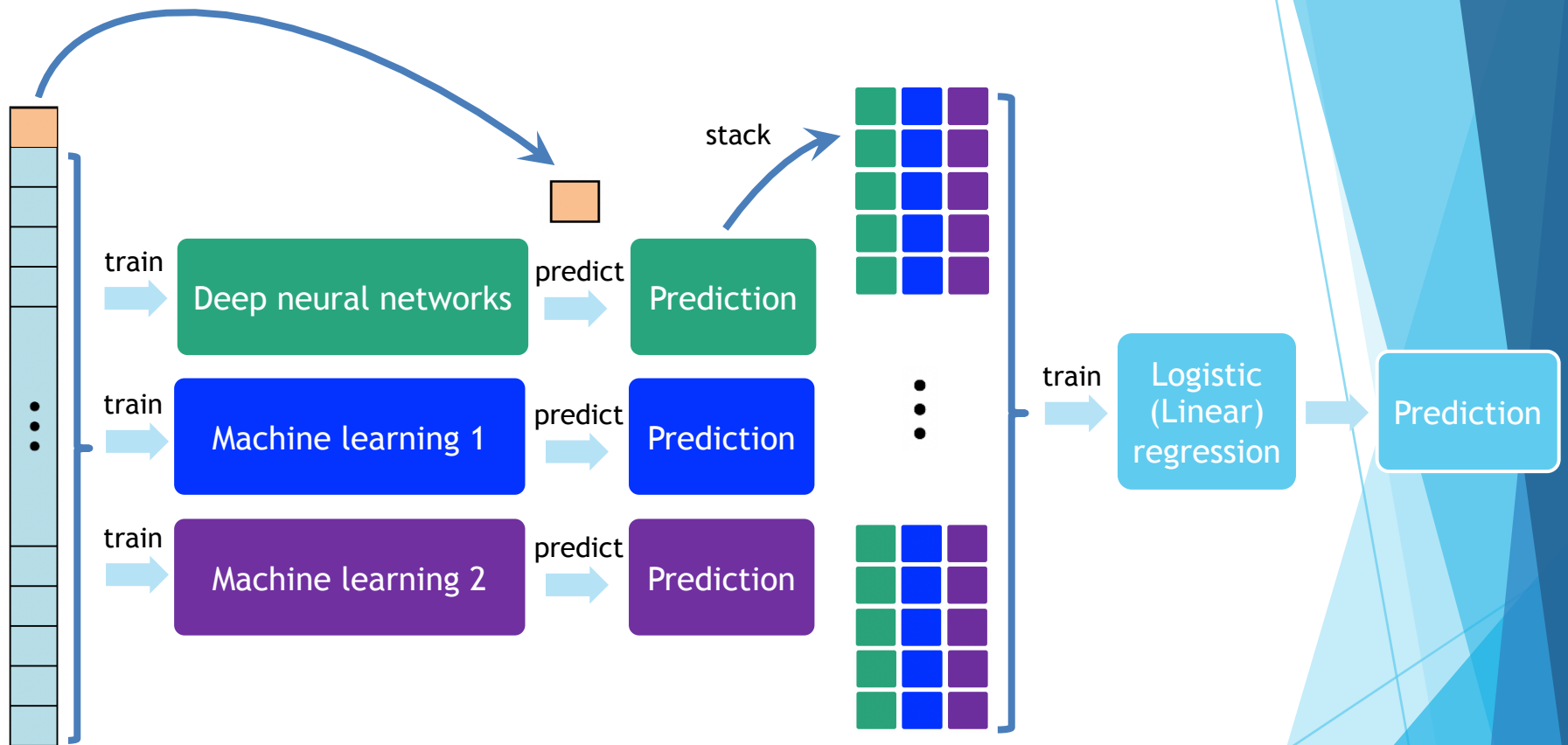
- Calculate splice site probability with deep learning from 140 nucleotides with the consensus sequence in the middle
- Evaluate probabilities for DS and AS respectively
- DS proba (WT)
- AS proba (WT)
- DS proba-diff (between WT and MT)
- AS proba-diff (between WT and MT)

Method: Ensemble

- ▶ Stacking: ensemble method for reducing generalization error
- ▶ Used regression models for ensemble



► Stacking



Result

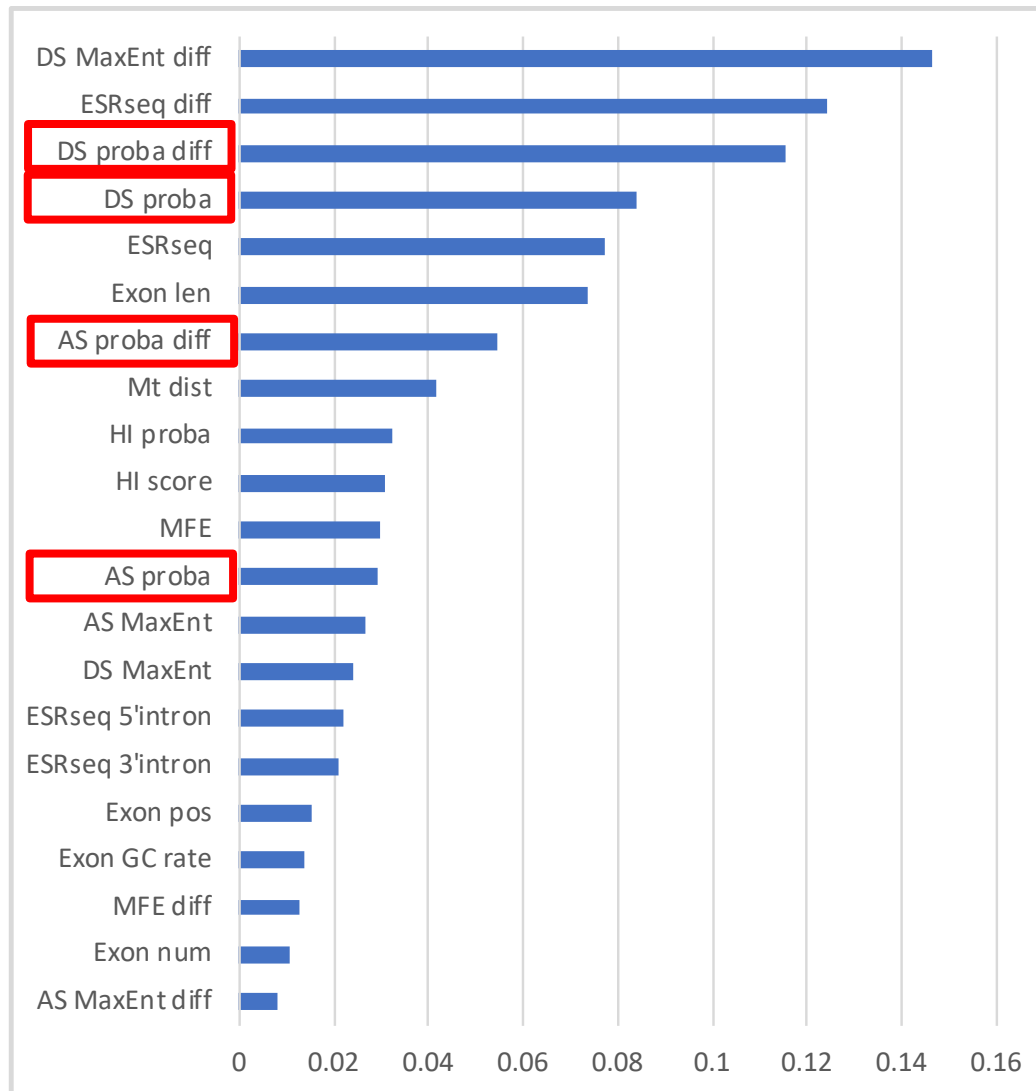
- ▶ Tried combining a lot of models, but eventually following models were used.
- ▶ Probability of ESM

	Deep learning	Logistic regression	Random forests	Ensemble
ROC-AUC	0.753	0.670	0.749	0.760

- ▶ Vivo/vitro AR

	Deep learning	Linear regression	Random forests	Ensemble
vivo MSE	1.15	-	1.11	1.11
vitro MSE	0.43	0.42	-	0.41

► Feature ranking



Discussion

- ▶ Combination of feature learning and feature engineering lead to better performance
- ▶ Original features may have enhanced the performance

Acknowledgement

- ▶ CAGI and data providers for the challenge