

CAGI slide pack guidelines

I agree to my slides being used according to the following terms: _____ *Ron Unger* _____

- 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, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



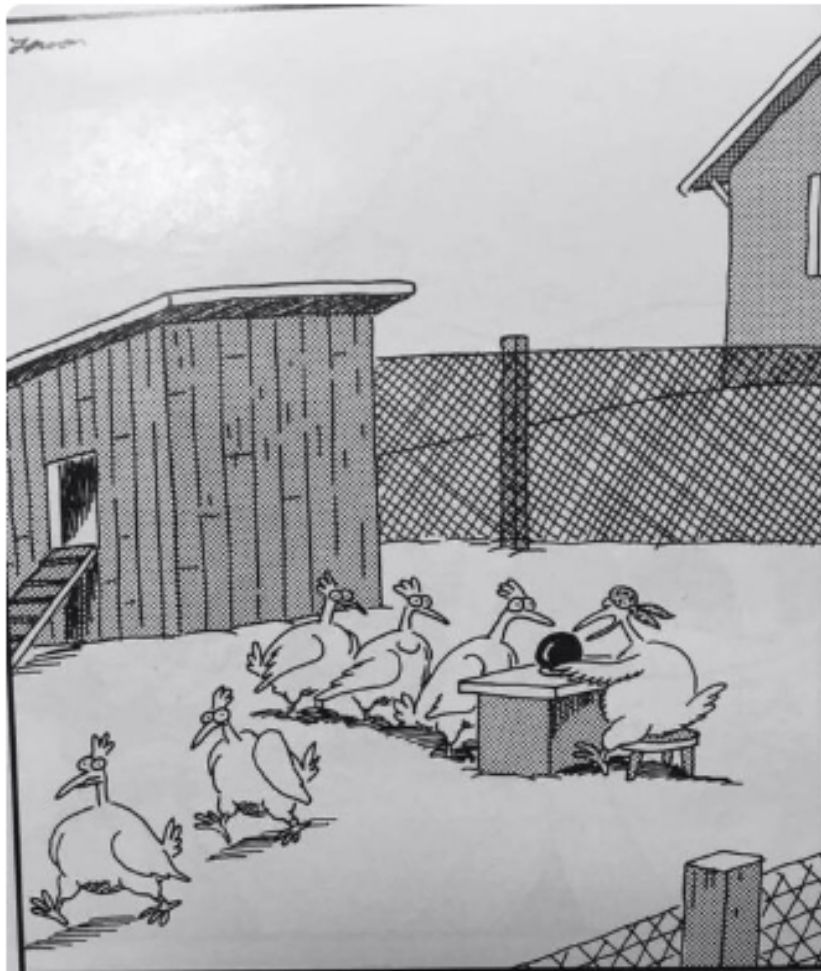
Mapsy and Vex-seq challenges

BIU team: Tzila Fenesh*, Tamar Holzer*, Orit
Adato, Noa Cohen, Liran Carmel and Ron Unger

*Equal Contribution

Goal -

An exercise in using machine learning to identify the impact of genetic variants on exon splicing.



"Whoa! Another bad one! ... I see your severed head lying quietly in the red-stained dirt, a surprised expression still frozen in your lifeless eyes. ... Next."

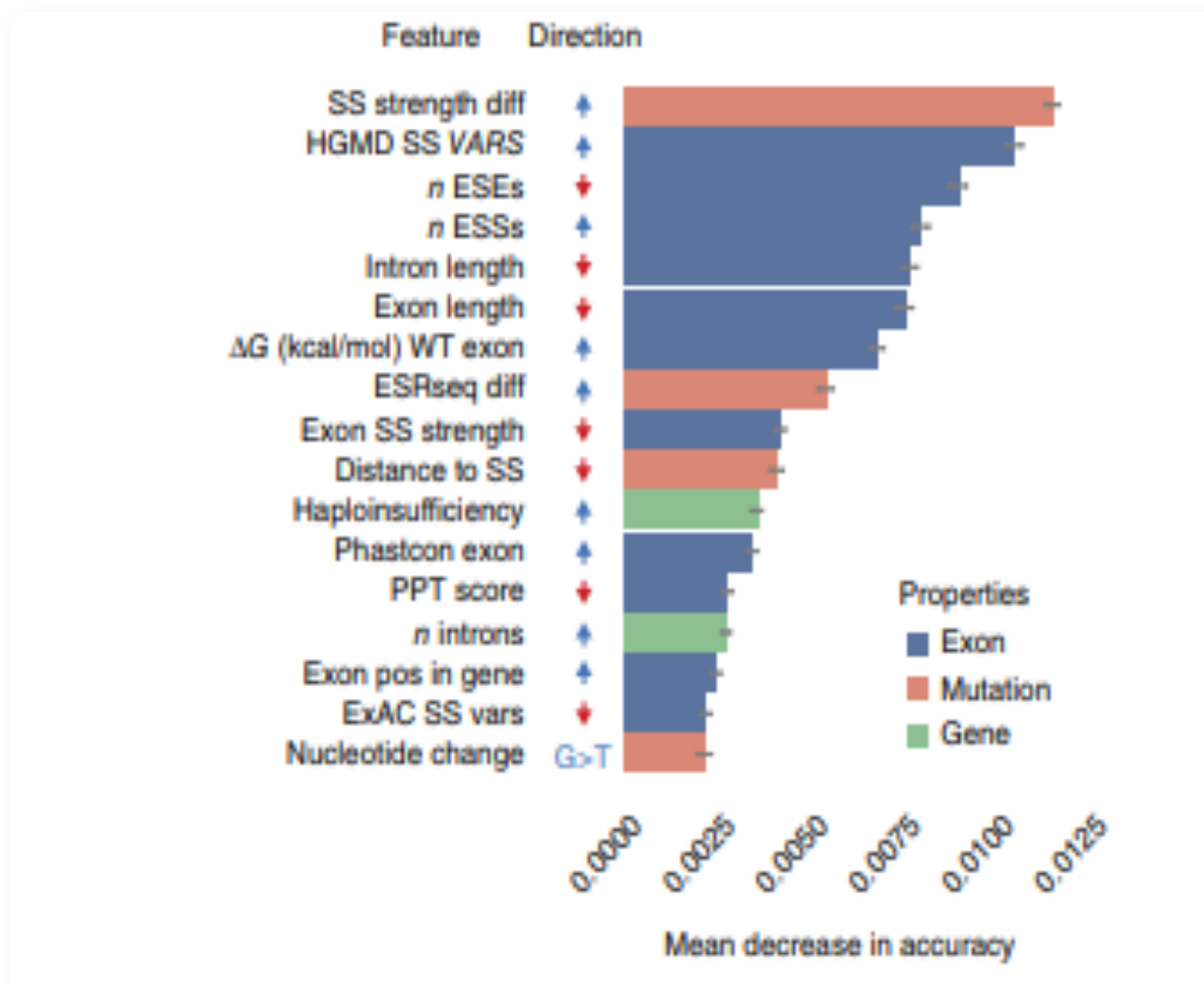
We based our Predication on

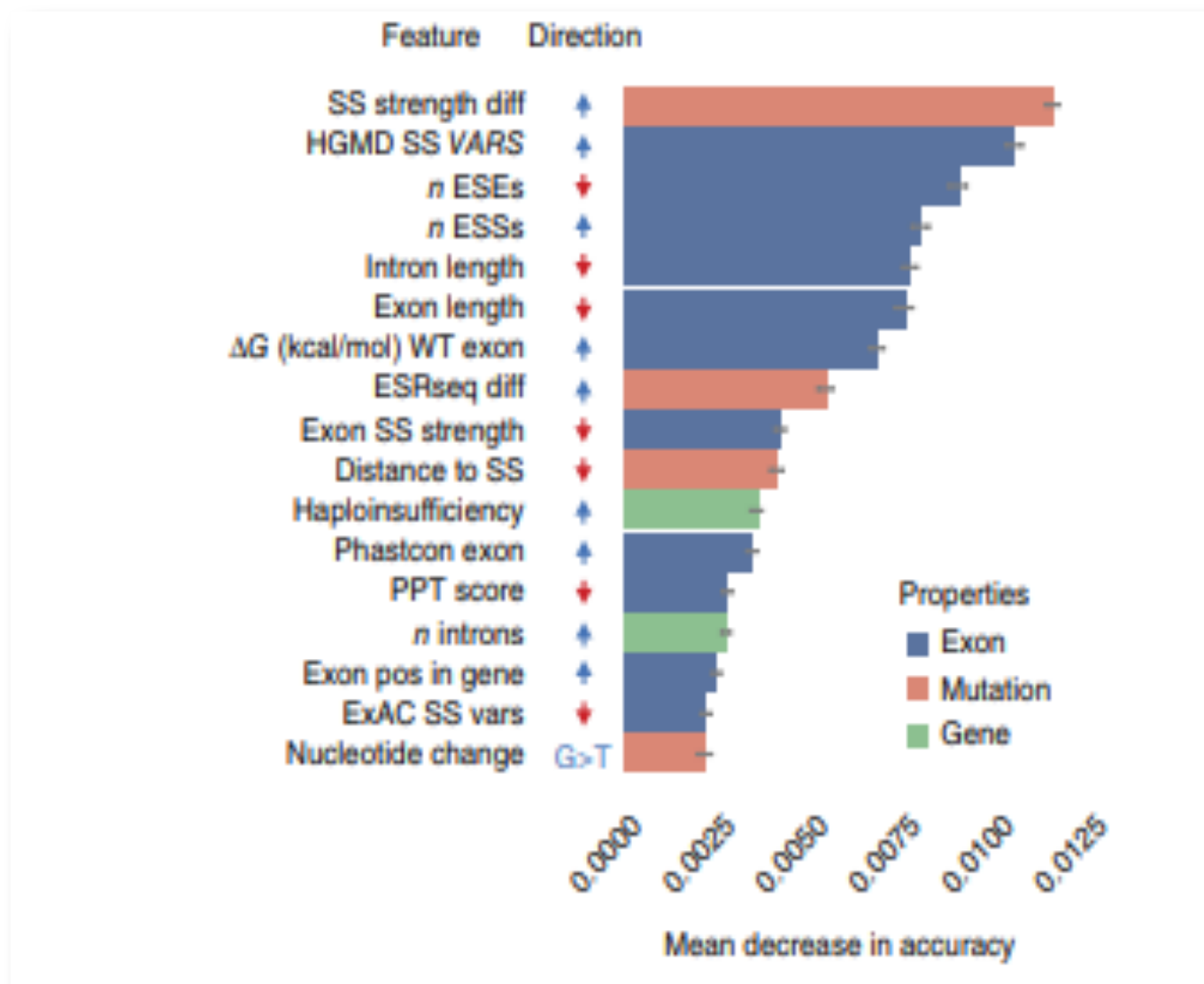
ARTICLES

nature
genetics

Pathogenic variants that alter protein code often disrupt splicing

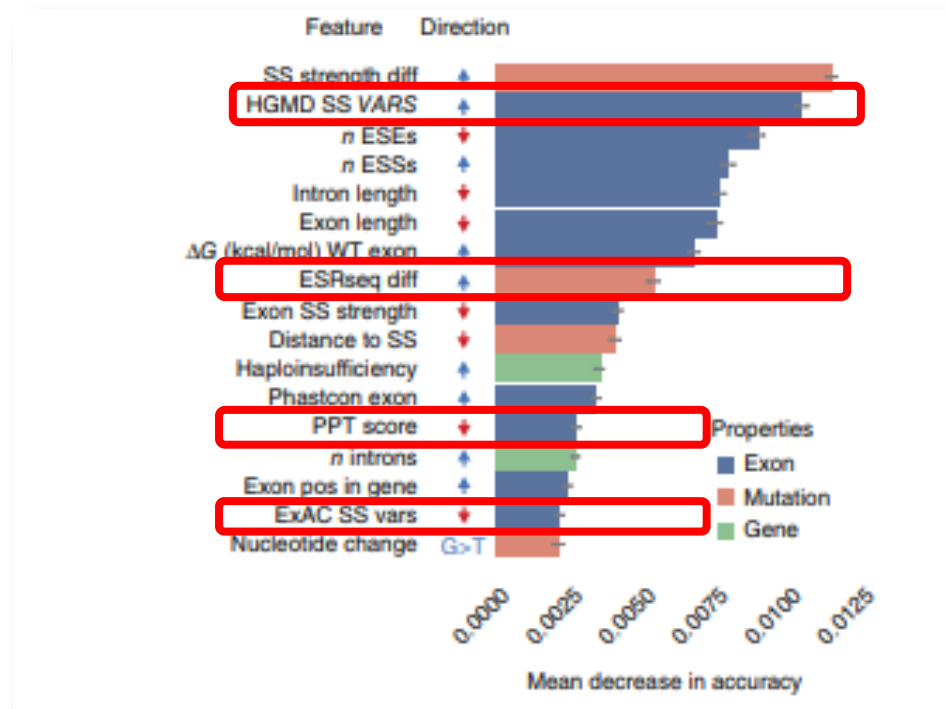
Rachel Soemedi^{1,2,7}, Kamil J Cygan^{1,2,7} , Christy L Rhine², Jing Wang², Charlston Bulacan³, John Yang⁴,
Pinar Bayrak-Toydemir⁵, Jamie McDonald⁵ & William G Fairbrother^{1,2,6}





Note that very few features are related to the specific mutations

We did not use all the features



We added few features

Weighted the values of ESS/ESE

We consider the spacing between ESS/ESE

We added Ex-Skip results (<http://ex-skip.img.cas.cz/>)

We try to add the results of prediction model of Noa Cohen and Liran Carmal (Hebrew University, work in progress), did not work on the test by itself, but helped a bit as an additional feature

Methods -

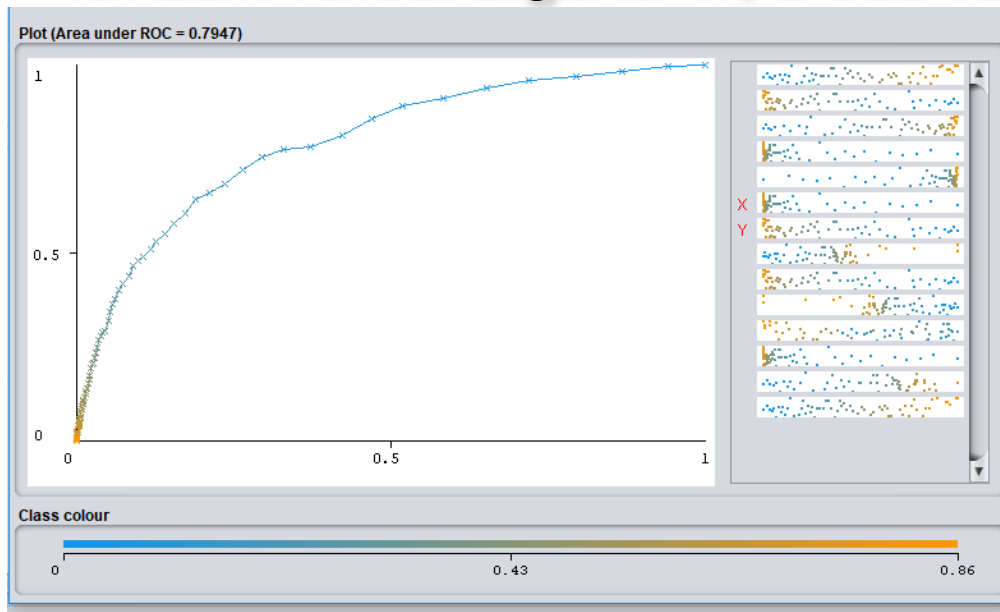
- Rna-fold – for the delta G feature.
- Ucsc table – for various features (conservation, number of exons, etc.).
- MaxEntScan – for the splice site strength feature.
- Ex-skip.
- Article about ESS/ESE motives.
- Haploinsufficiency table.
- Etc.

Algorithms-

- Decision tree.
- Random forest.
- Neural network.
- Logistic regression.

Results –

MaPsy -Roc curve (on the training set using random forest algorithm, cross validation)



Area under ROC = 0.79

However-The confusion matrix is not that good

```
=== Confusion Matrix ===
```

a	b	<-- classified as
4464	47	a = 0
398	42	b = 1

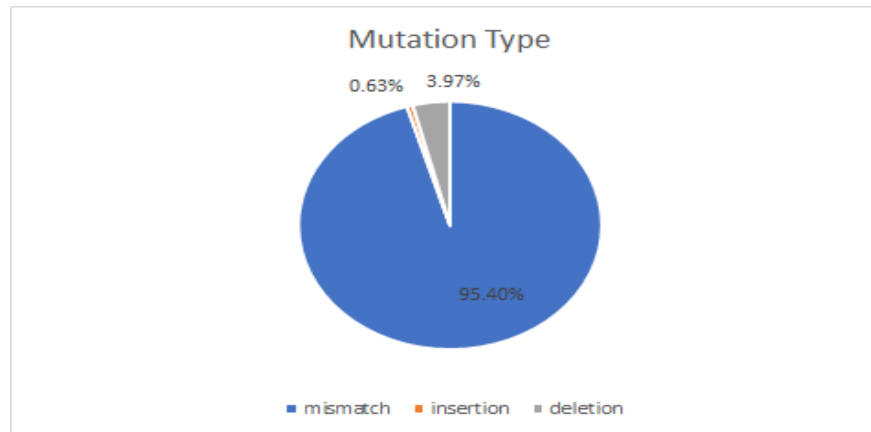
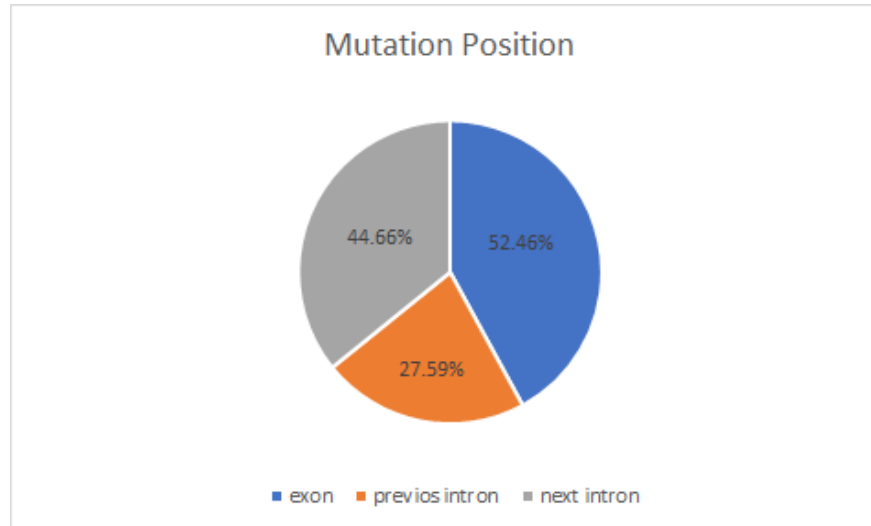
Results – VexSeq

Comparison of Vex-Seq and MaPSY-

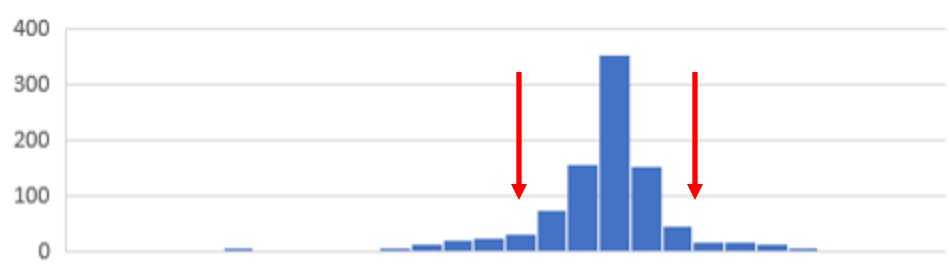
	Vex-seq	MaPSY
Mutation type	Insertion / deletion / mismatch	mismatch
Mutation position	Intron / exon	exon

We added this information as features

Characteristics of Vex-Seq data



Converting continuous data to 3 classes



The “no-change” range was $-12.2 < \text{value} < 3.08$

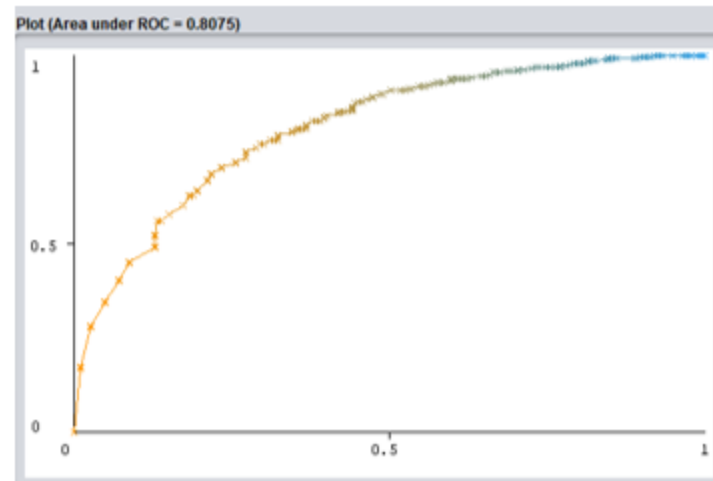
More splicing > 3.08

Less splicing < -12.2

Best results on the training set used Random Forest

=== Confusion Matrix ===

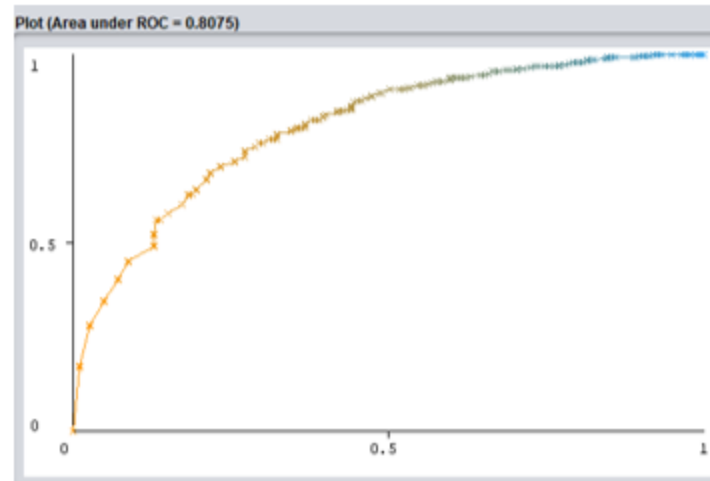
a	b	c	<-- classified as
739	31	8	a = 0
85	25	7	b = 1
35	6	21	c = 2



Best results on the training set used Random Forest

=== Confusion Matrix ===

a	b	c	<-- classified as
739	31	8	a = 0
85	25	7	b = 1
35	6	21	c = 2



On the challenge data Random Forest predicted all mutations as no change so we used the results of decision tree algorithm

Converting back the data to continuous range

We invested quite an effort in finding the best way to convert the discrete, 3 classes data, to continuous range.

It turns out that these procedures were not very successful, and our better submission was used on direct prediction of continues values.

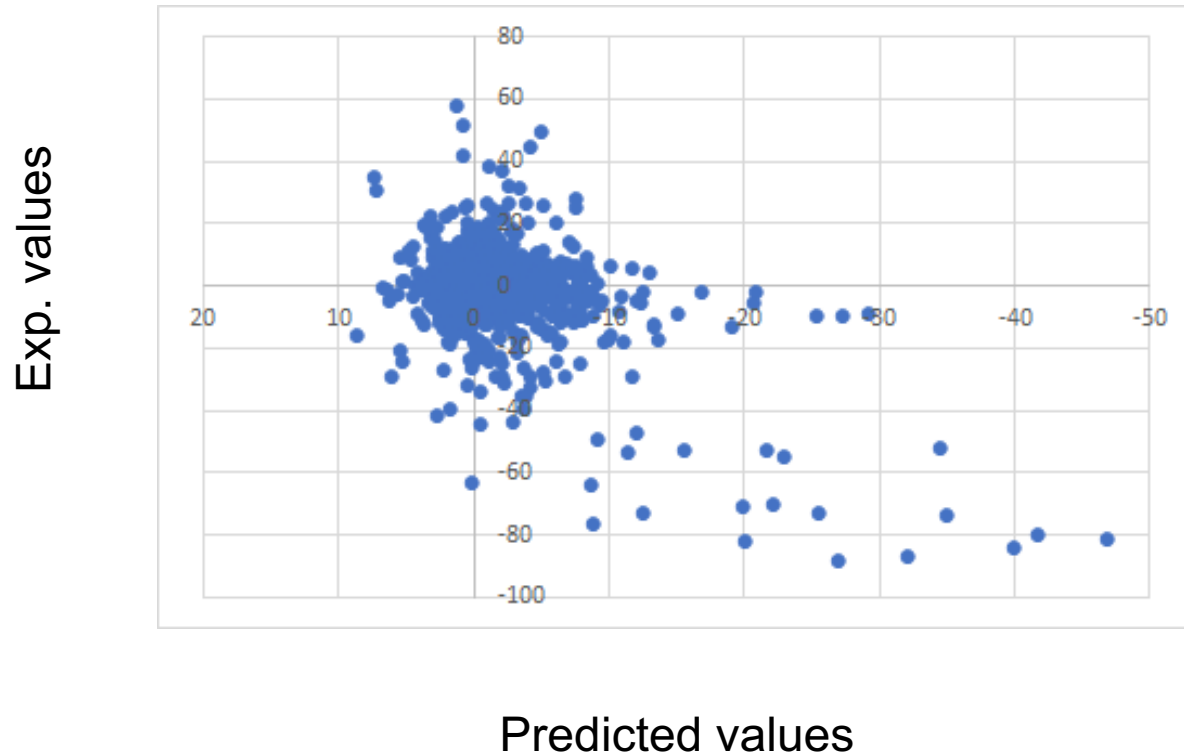
Converting back the data to continuous range

We invested quite an effort in finding the best way to convert the discrete, 3 classes data, to continuous range.

It turns out that these procedures were not very successful, and our better submission was used on direct prediction of continues values with a Random Forest. .

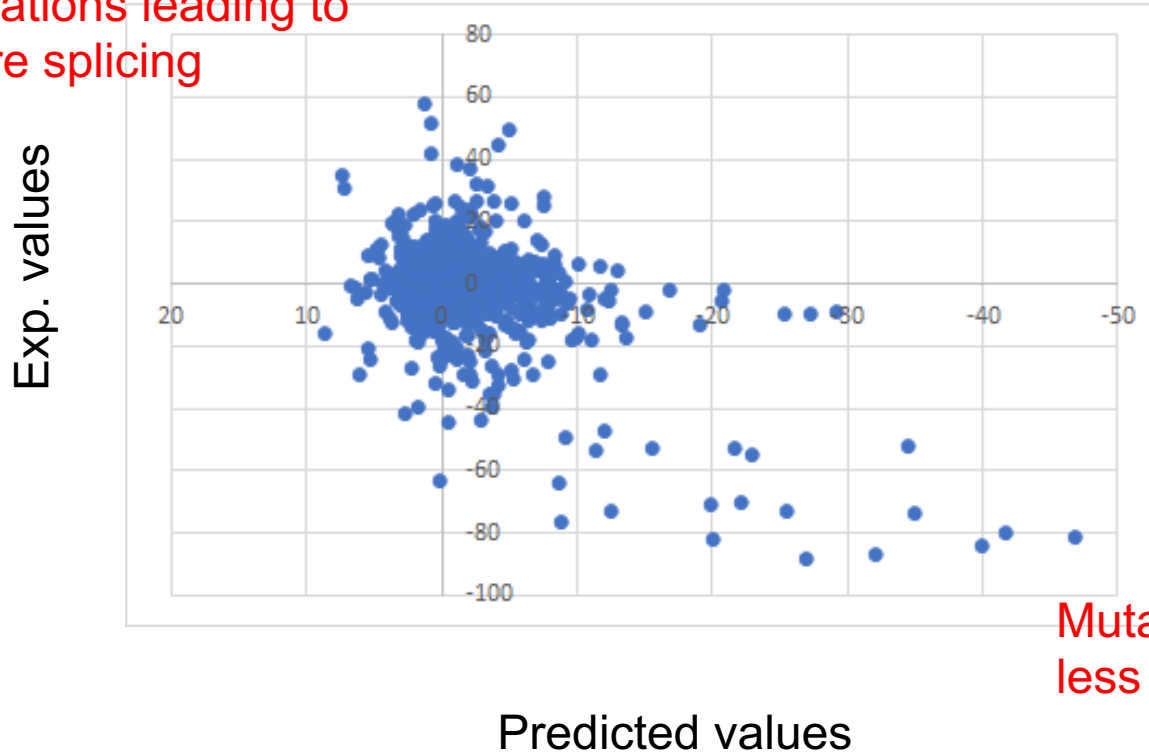
<i>results</i>	<i>sub5</i>	<i>sub4</i>	<i>sub3</i>	<i>sub2</i>	<i>sub1</i>	
					1	sub1
				1	0.162718	sub2
			1	0.196721	0.547668	sub3
		1	0.392351	0.215142	0.777046	sub4
	1	0.304025	0.377344	0.551217	0.3265	sub5
1	0.490003	0.072077	0.096844	0.326902	0.070846	results

Comparing the predicted and experimental results



Comparing the predicted and experimental results

Mutations leading to
more splicing



Mutations leading to
less splicing

We did reasonably well with the extreme values