

CAGI slide pack guidelines

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

____*Jun Cheng*____

- 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



Compositional modeling of core splicing regulation improves prediction of variant effects on splicing

Jun Cheng

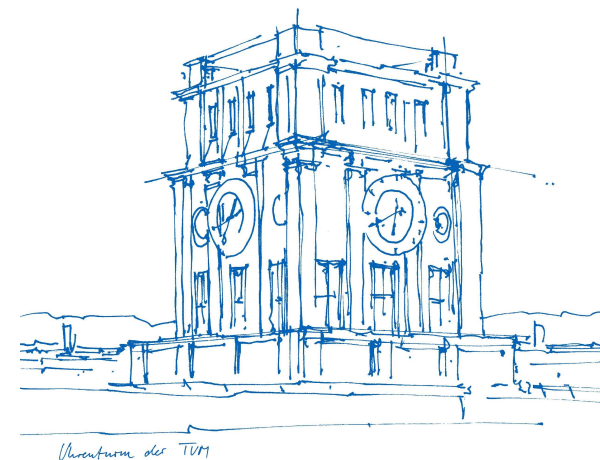
Gagneur lab - Computational biology

Technical University of Munich

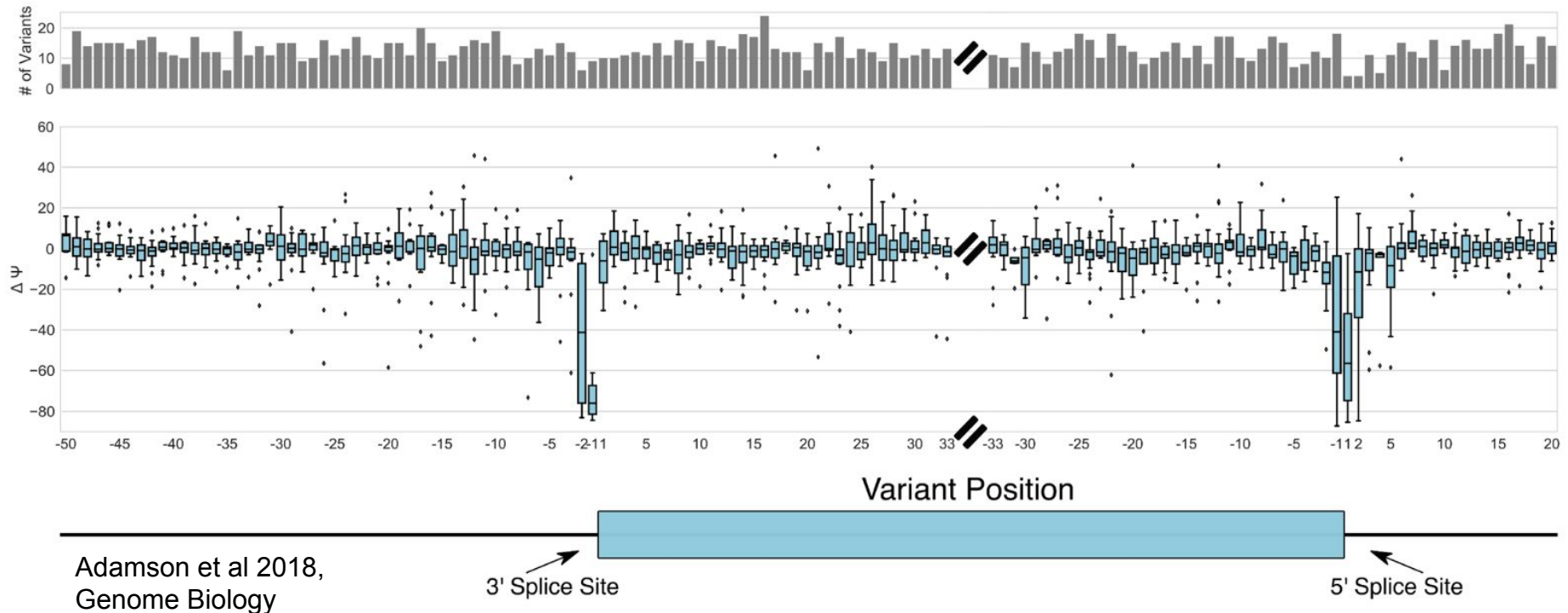


@gagneurlab

July 2018

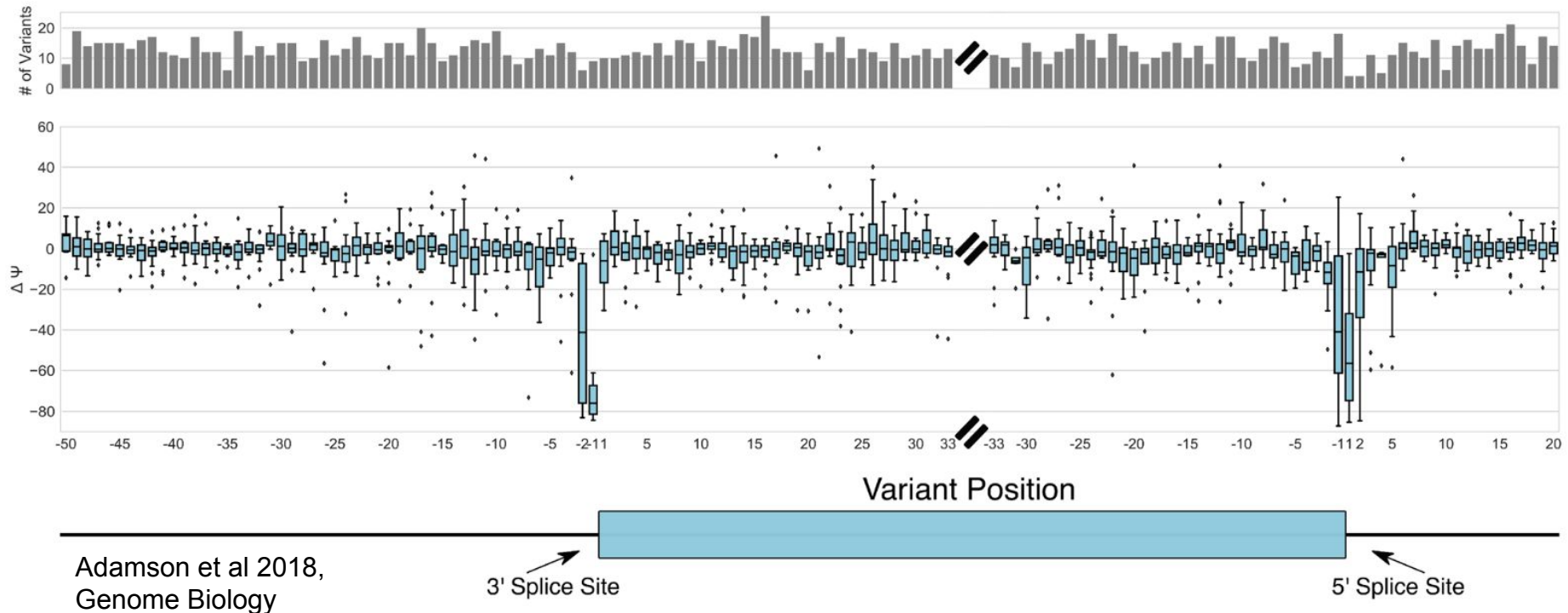


CAGI Vex-seq Challenge: Splicing Variant Effect Prediction



Predict variant effect on splicing with 957 training variants.

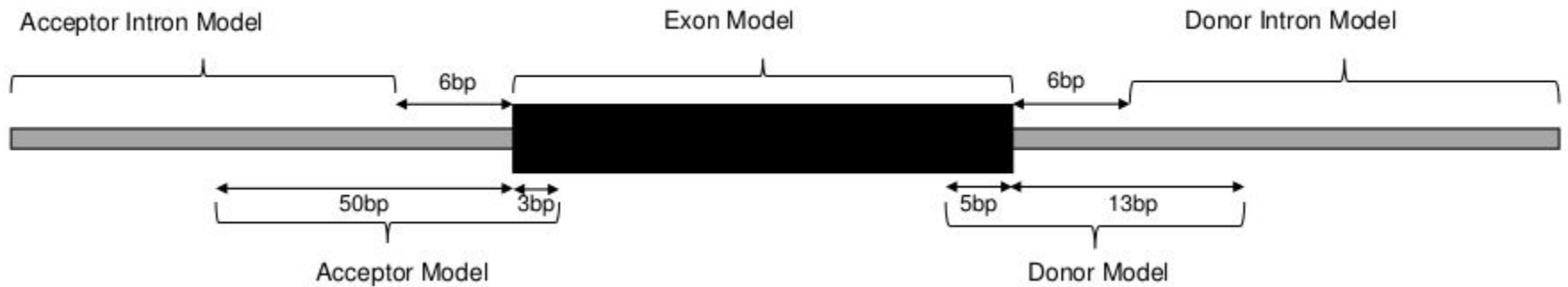
CAGI Vex-seq Challenge: Splicing Variant Effect Prediction



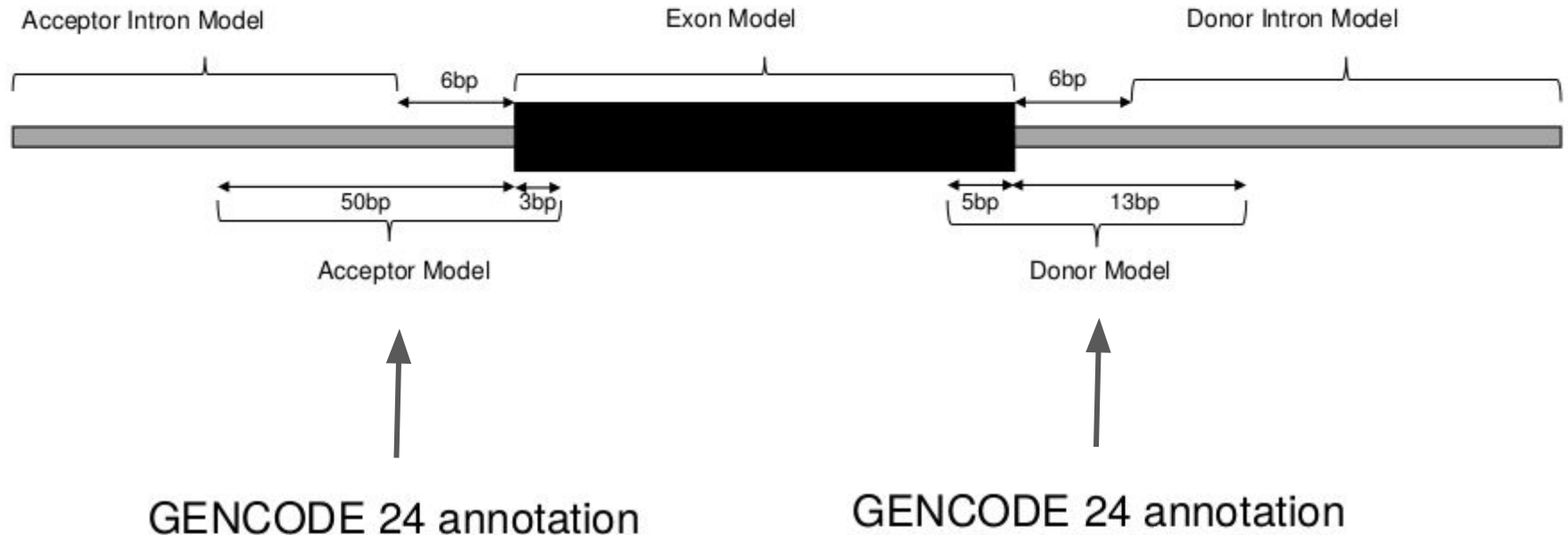
Predict variant effect on splicing with 957 training variants.

Larger dataset needed!

Composite Modeling (Exon, Intron, Splice Site)



Learning the Donor and Acceptor Modules using GENCODE



N = 271,405
True Acceptor

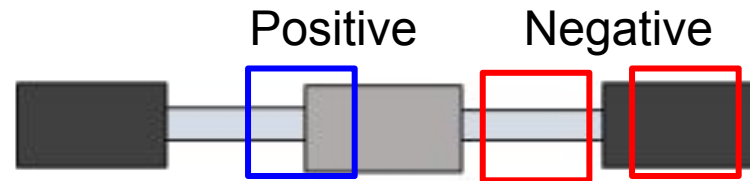
N = 406,841
False Acceptor

N = 273,661
True Donor

N = 410,111
False Donor

Acceptor Scoring Model

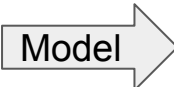
Train a splice site (donor or acceptor) scoring model by training a classifier to classify true/false splice sites



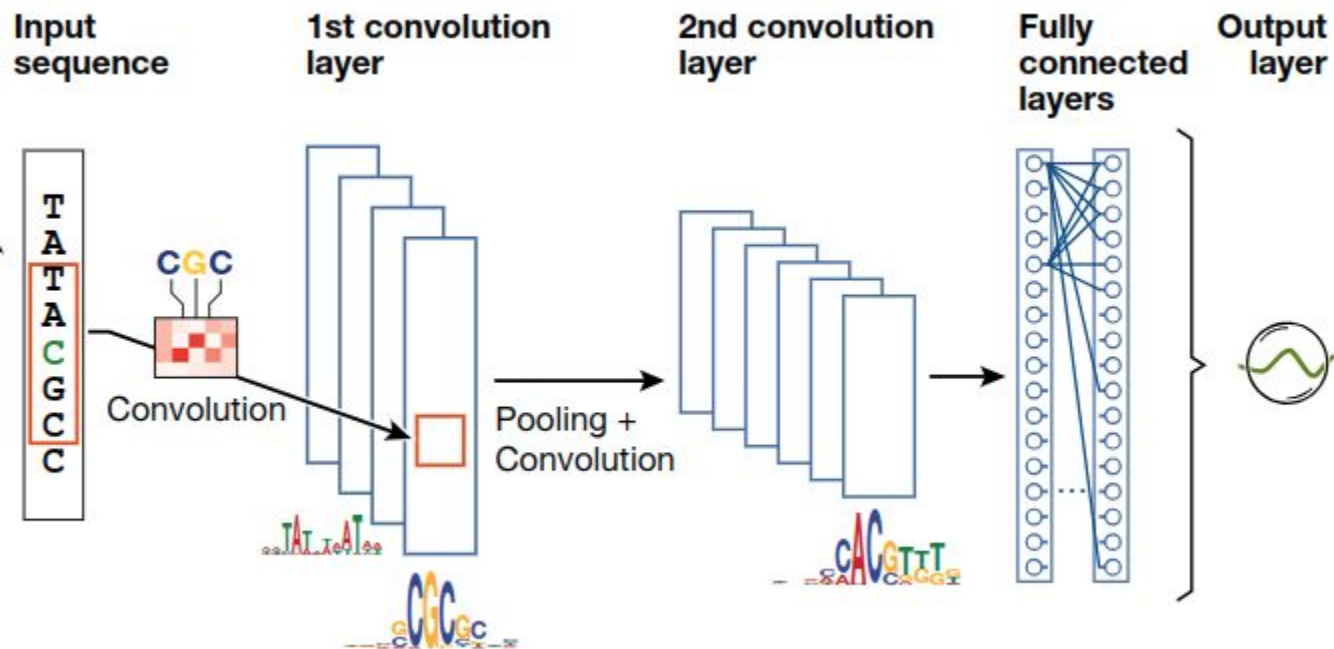
Positive: all annotated acceptors (271,406), 50 bp in intron, 3 bp in exon.

Negative: same length sequence from gene regions that were not annotated as acceptor sites.

- 406841 sequences in total
- 232224 (57%) with AG consensus

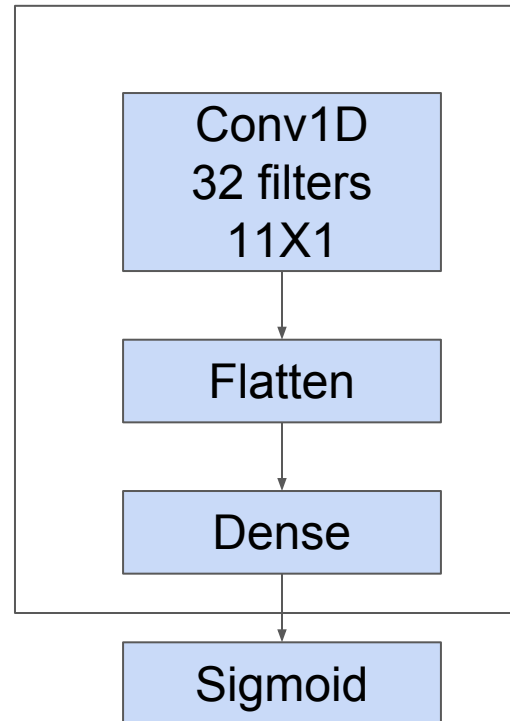
ATGATCG...  Score:
Goodness as an acceptor

Convolutional Neural Network in Genomics



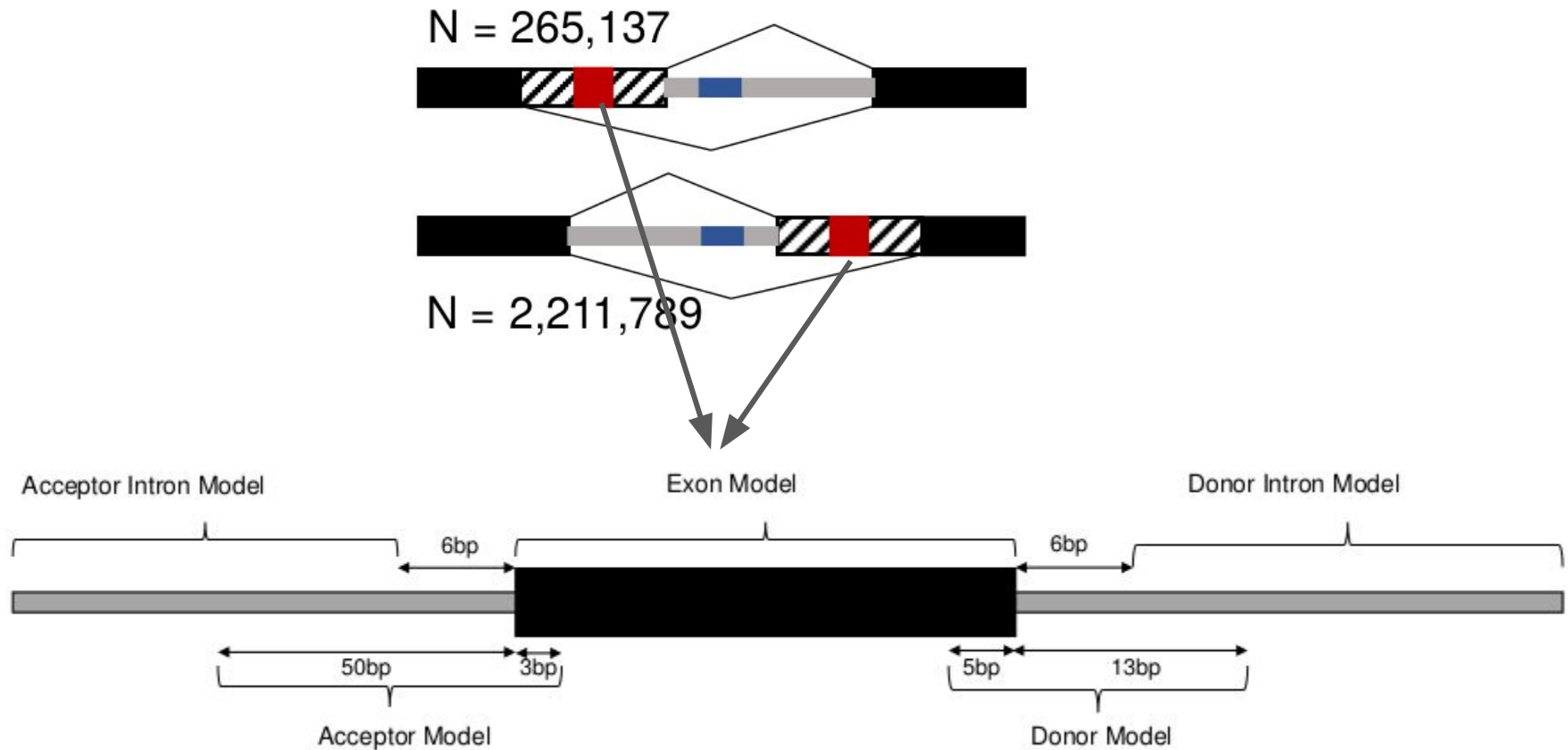
Angermueller et al, 2016, MSB

Acceptor Scoring Model



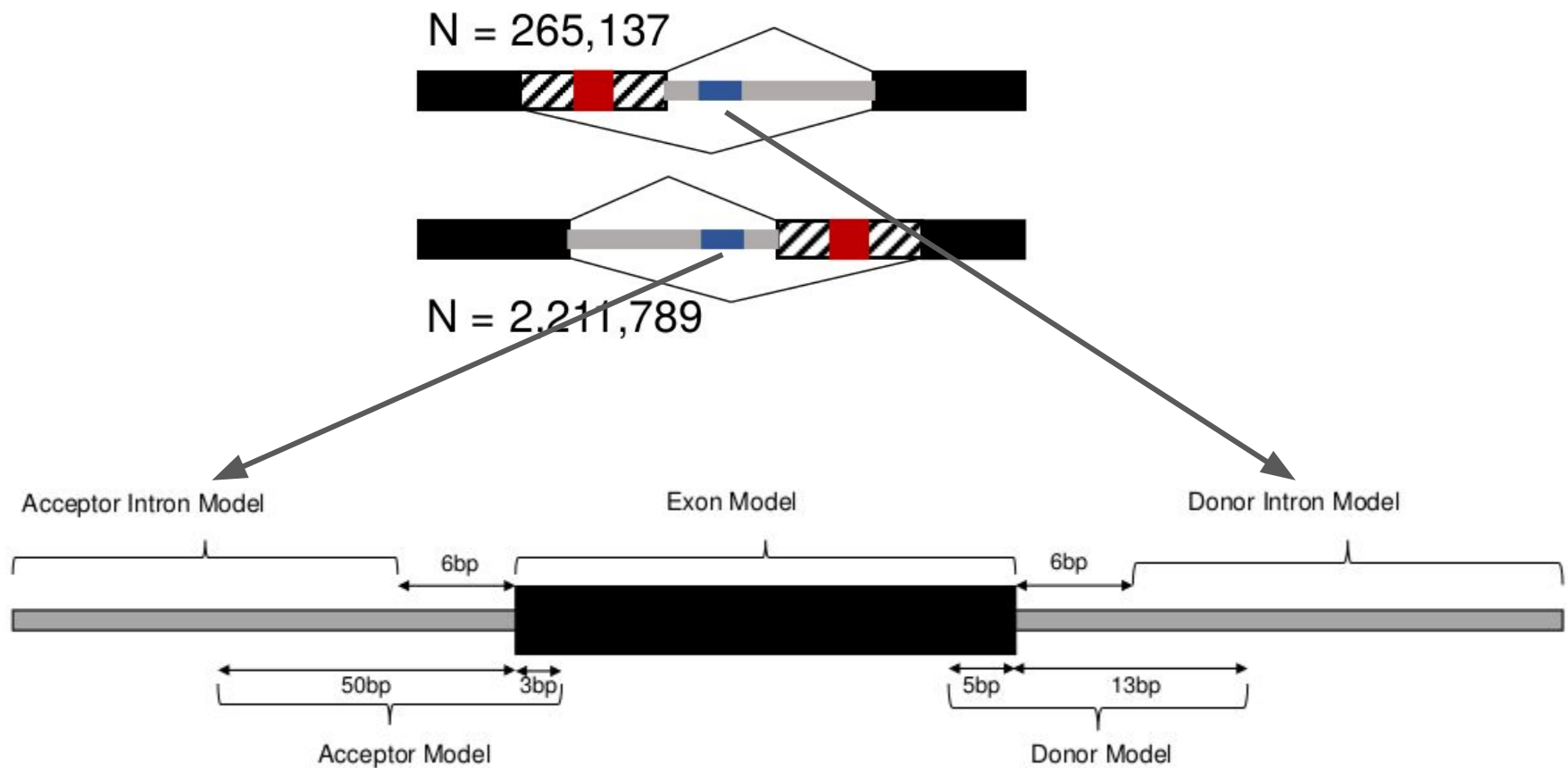
Learning the Exon Module Based on a MPRA Dataset

Massively parallel splicing reporter assay from Rosenberg et al 2015

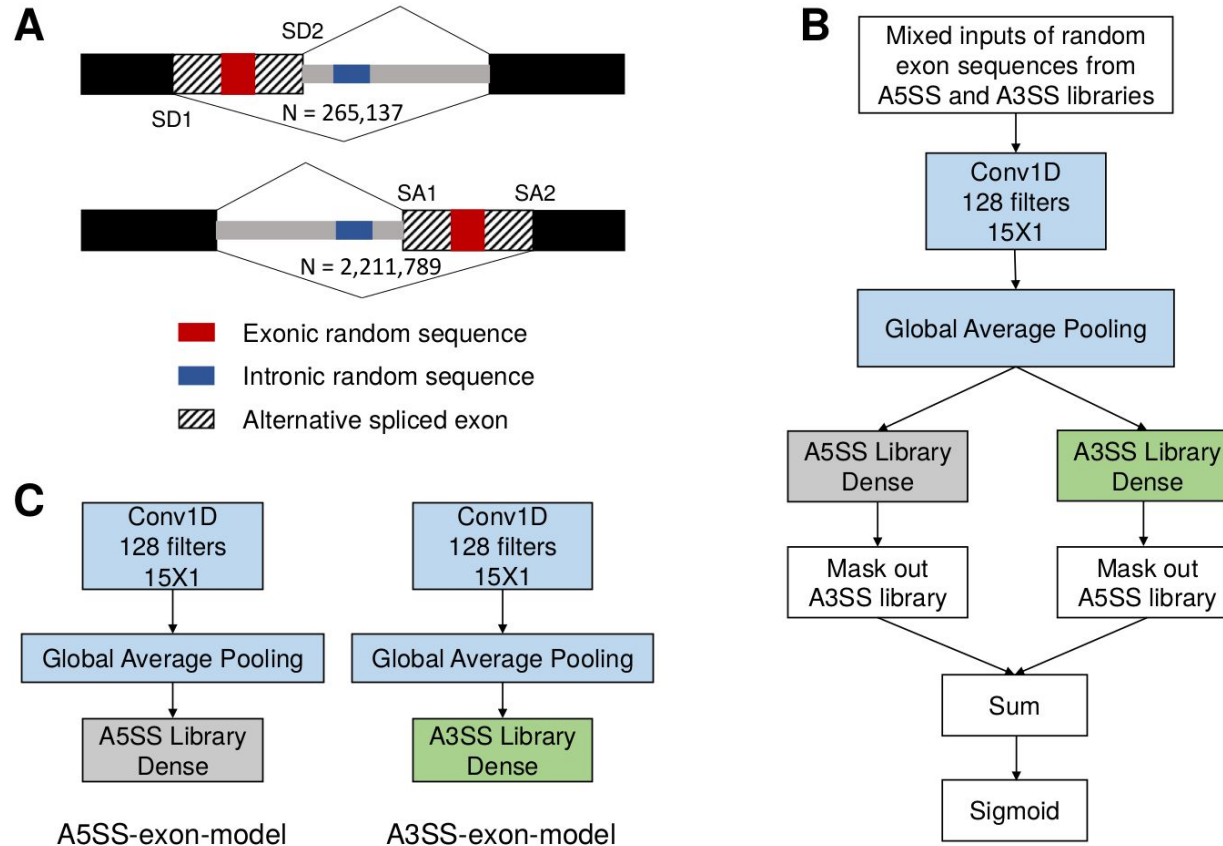


Learning the Intron Modules Based on a MPRA Dataset

Massively parallel splicing reporter assay from Rosenberg et al 2015

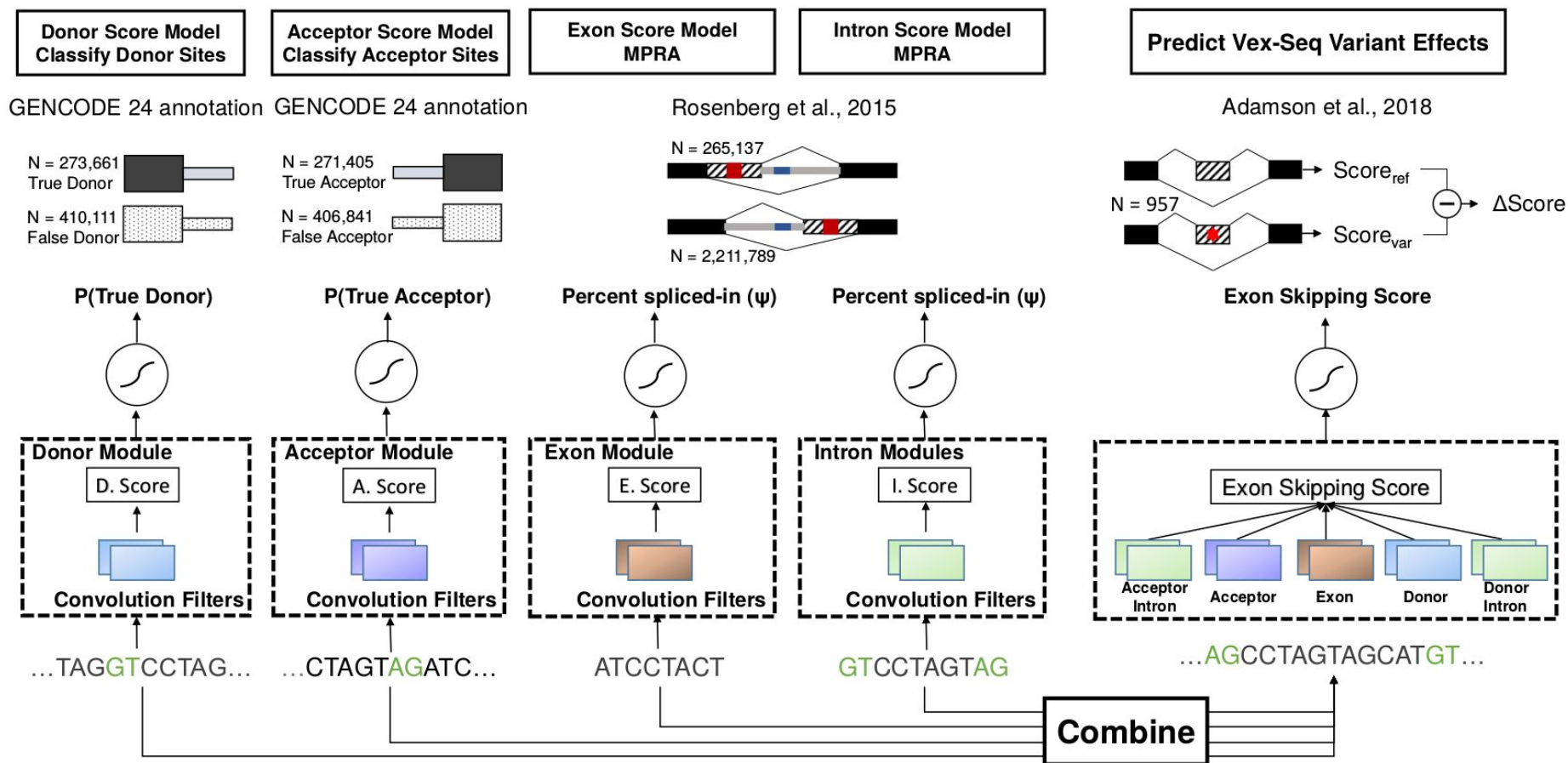


Train a Joint Exon Regulator Model



Same principle applied to intron models with MPRA data

Composite Modeling (Exon, Intron, Splice Site)



Variant Effect Prediction

Effect of mutation on PSI depends on basal PSI:

Assume a mutation change PSI 5% $\rightarrow$ 50% in one gene, however, in another gene with PSI = 99%, it can hardly make any change.



Variant Effect Prediction Requires the Reference PSI Level

$\Delta\hat{\psi}$ is predicted with:

$$\hat{\psi}_{var} = \sigma(\Delta_{\hat{s}} + \text{logit}(\psi_{ref})) \quad (1)$$

$$\Delta_{\hat{\psi}} = \hat{\psi}_{var} - \psi_{ref} \quad (2)$$

$$= \sigma(\Delta_{\hat{s}} + \text{logit}(\psi_{ref})) - \psi_{ref} \quad (3)$$

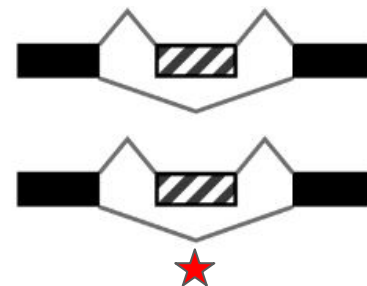
where:

$\Delta_{\hat{s}}$: predicted score differences between WT and mutations.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

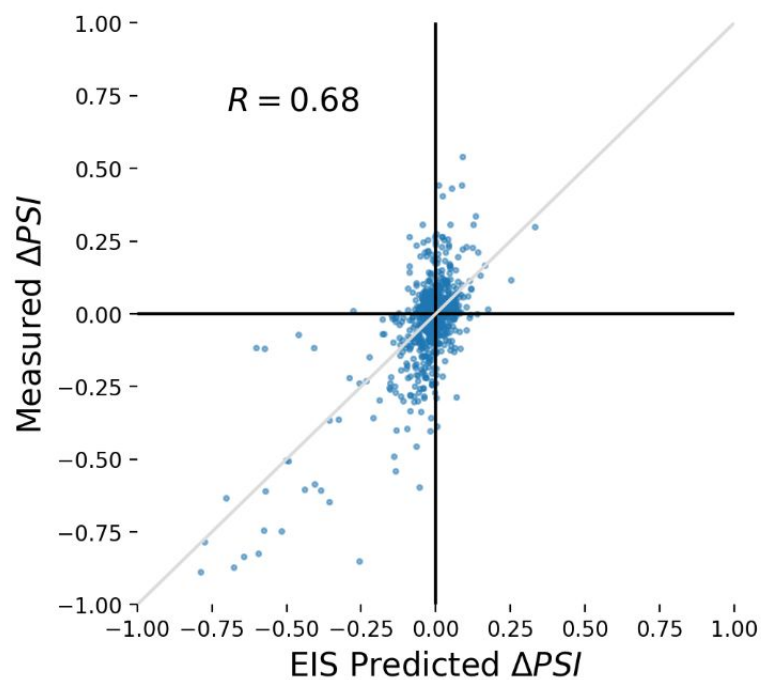
$$\text{logit}(x) = \log \frac{x}{1 - x}$$

$$\psi_{ref} \in (0, 1)$$

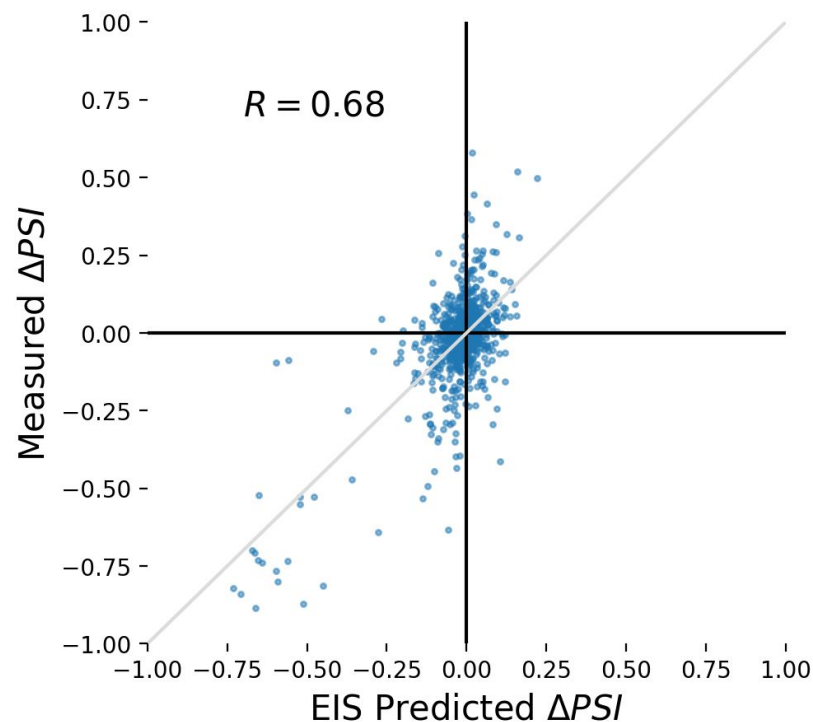


Model Validation on Vex-Seq Data

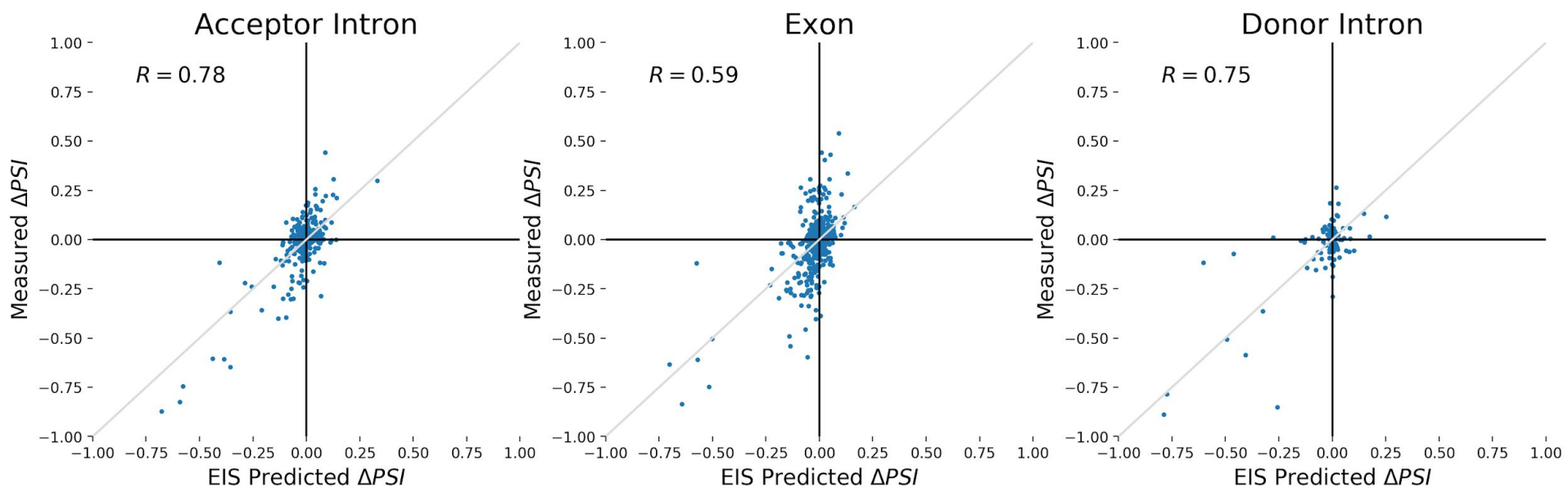
Training



Testing

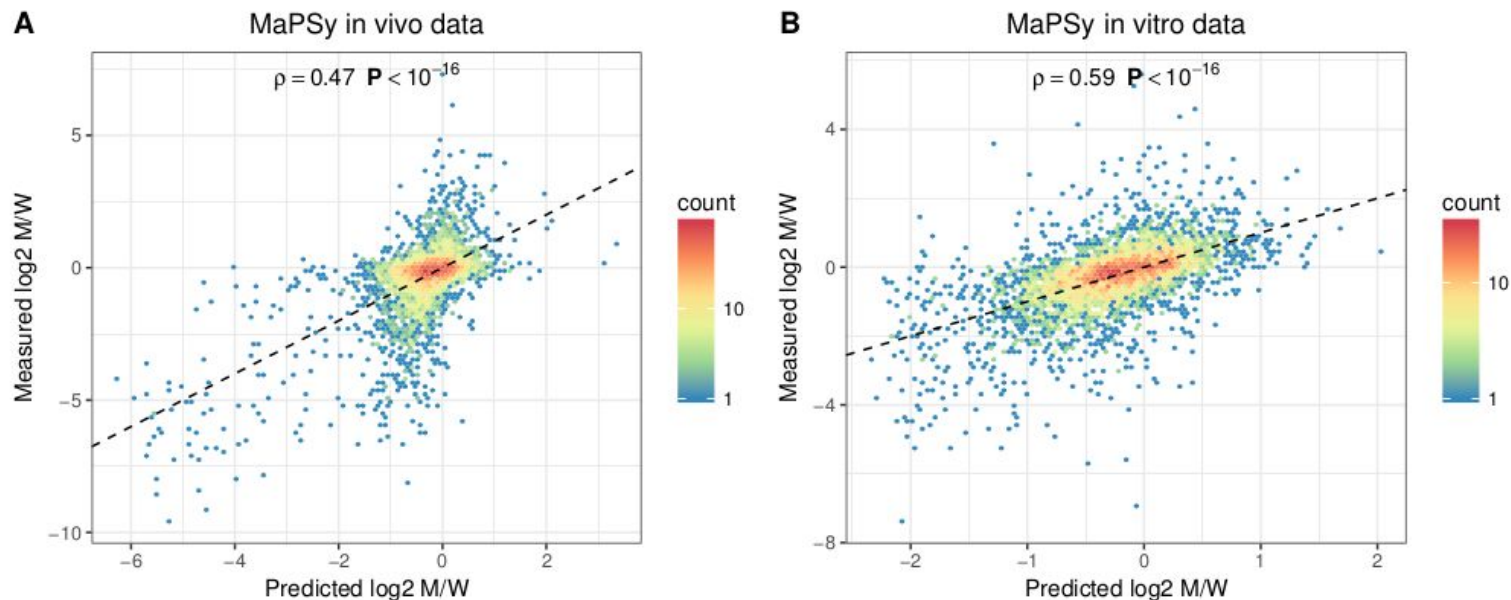


Model Validation on Vex-Seq Data

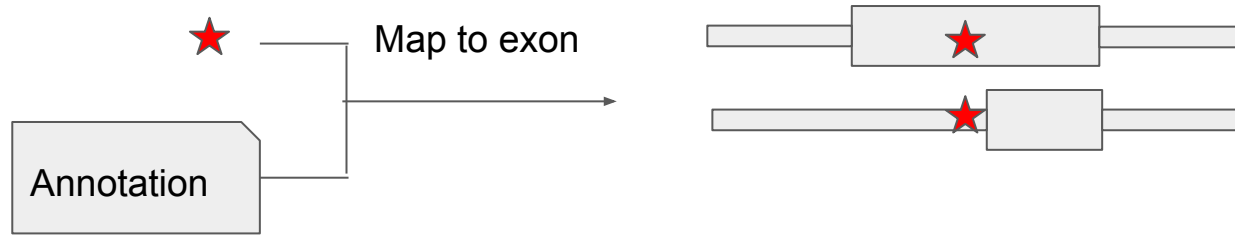


Model Validation on MaPSy Data

Analyzed **4,964** published **disease-causing exonic** mutations using a massively parallel splicing assay (MaPSy). Approximately **10%** of exonic mutations **altered splicing**.



Variant Effect Prediction from VCF File



Delta logit(Psi)

Var1	Exon1	-0.5
Var1	Exon2	-0.9
Var2	Exon2	2.1

One variant can affect the splicing of multiple exons, one exons can be affected by many variants.

Variants can be SNPs or indels, can be exonic or intronic.

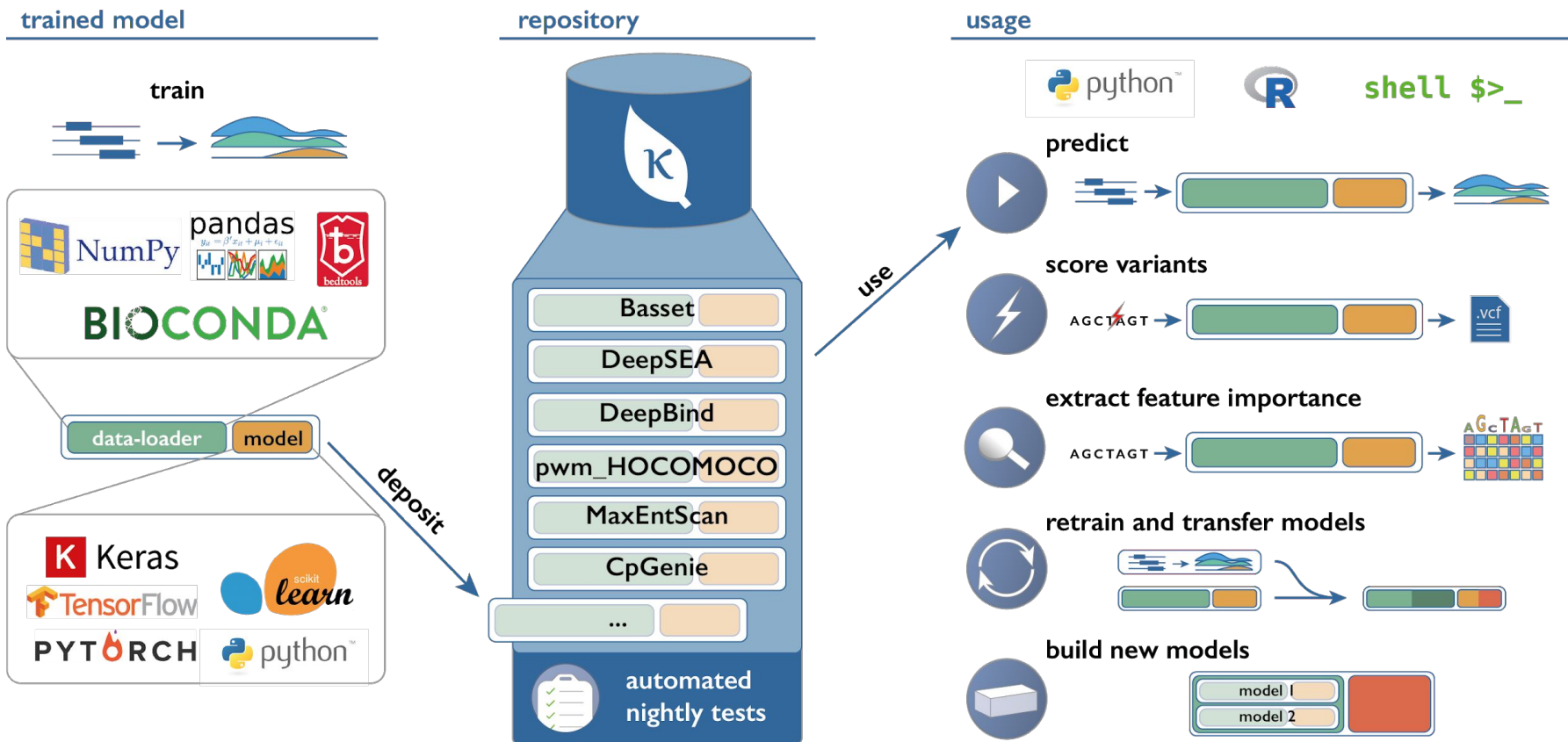
Conclusion

Modular architecture.

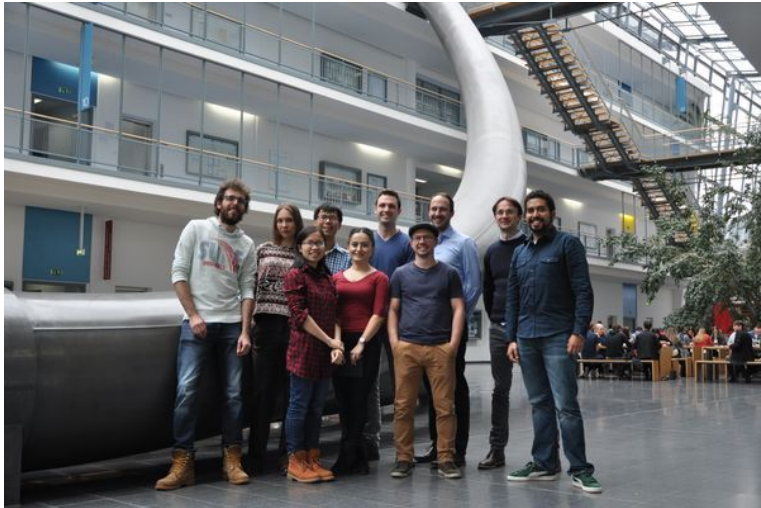
Modules trained on complementary rich data.

Runnable on VCF, including indels.

Kipoi: accelerating the community exchange and reuse of predictive models for genomics



Acknowledgements



Whole Gagneur Lab, especially:

- Julien Gagneur
- Ziga Avsec
- Nguyen Thi Yen Duong

Technical University of Munich

Graduate School
Quantitative Biosciences Munich
KONWHIR

CAGI Team
Gaia Andreoletti
Scott Adamson