Predicting splice sites in DNA sequences with sequence models

2024.05.15

Kuan-Hao Chao

khchao.com

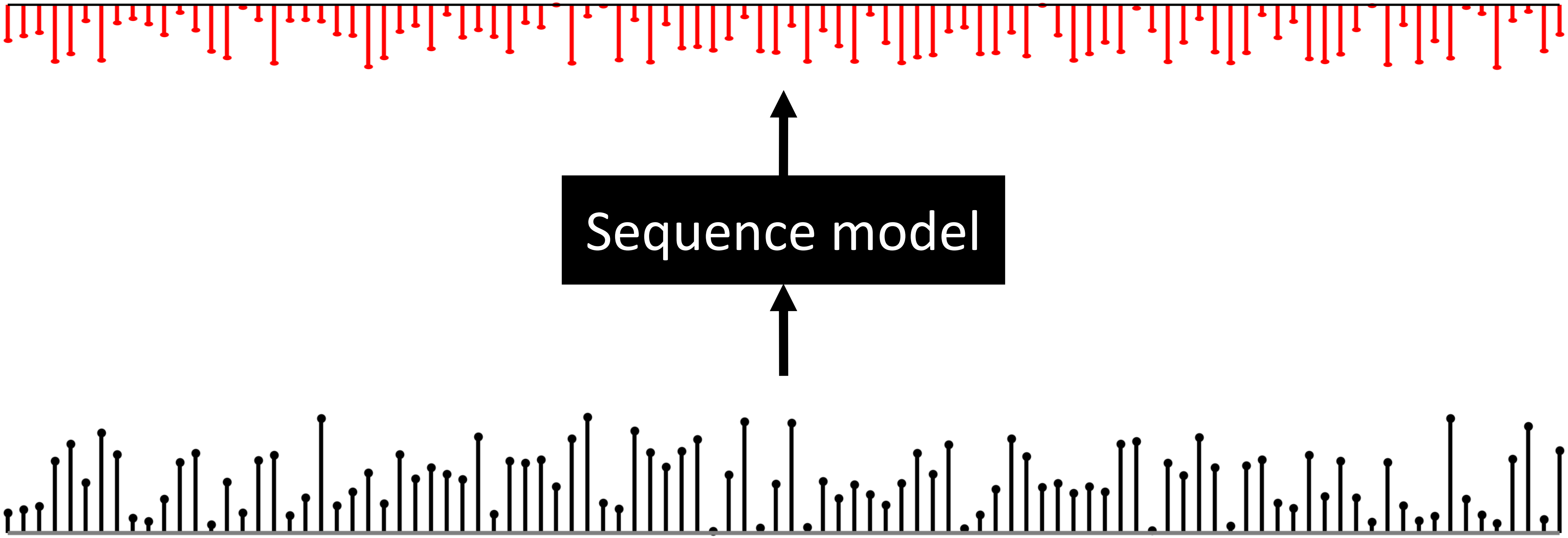
@KuanHaoChao

Kuanhao-Chao

Johns Hopkins University
Center for Computational Biology
CCB

Johns Hopkins
Whiting School of Engineering
Department of Computer Science

Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

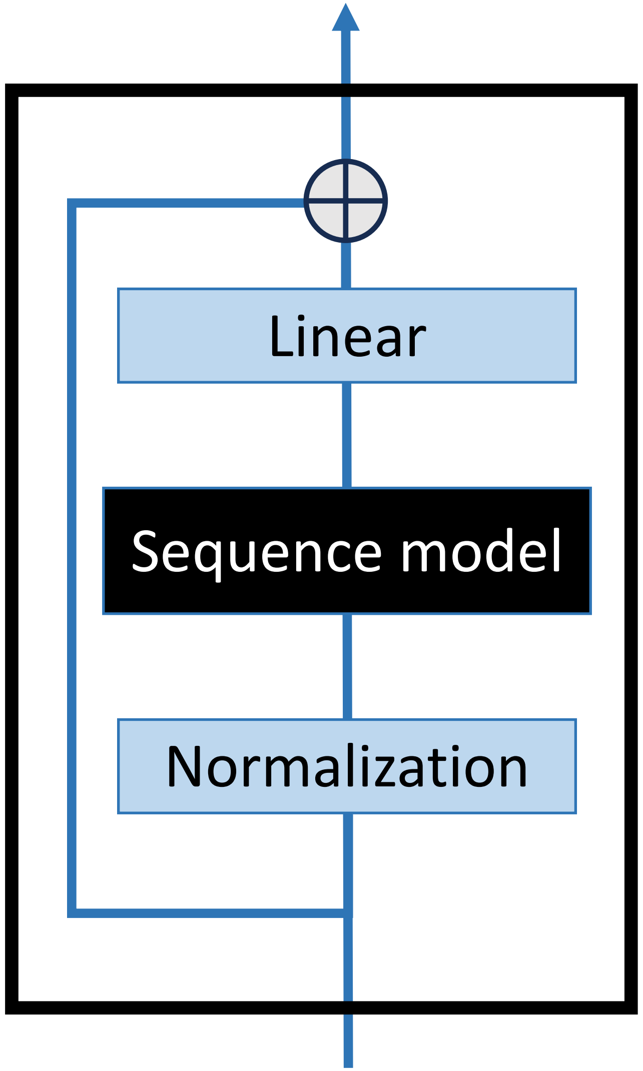


Sequence model

Sequence models map a sequence to a sequence

Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

(batch, length, dim)



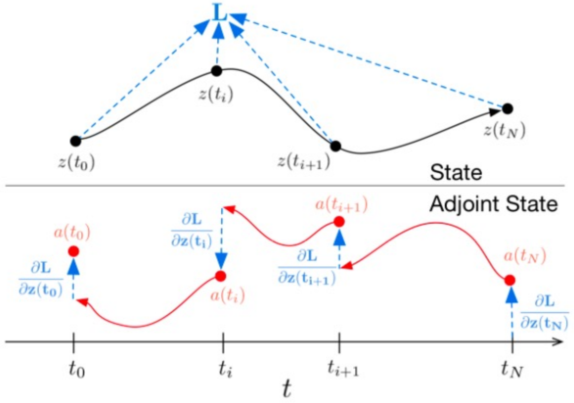
Linear

Sequence model

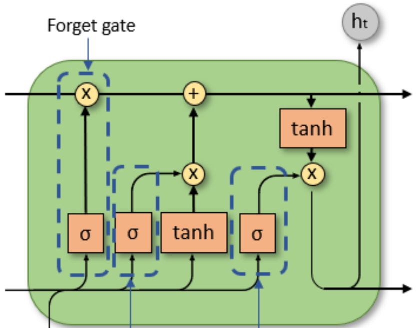
Normalization

(batch, length, dim)

Neural ODEs

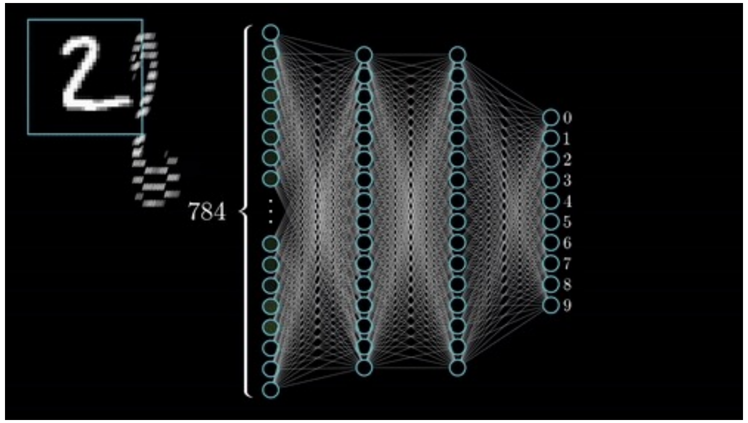

L
$z(t_0)$
$z(t_i)$
$z(t_{i+1})$
$z(t_N)$
State
Adjoint State
$a(t_0)$
$\frac{\partial L}{\partial z(t_0)}$
$\frac{\partial L}{\partial z(t_i)}$
$a(t_i)$
$a(t_{i+1})$
$\frac{\partial L}{\partial z(t_{i+1})}$
$a(t_N)$
$\frac{\partial L}{\partial z(t_N)}$
t_0
t_i
t_{i+1}
t_N
t

RNN



Future work

CNNs



Transformers

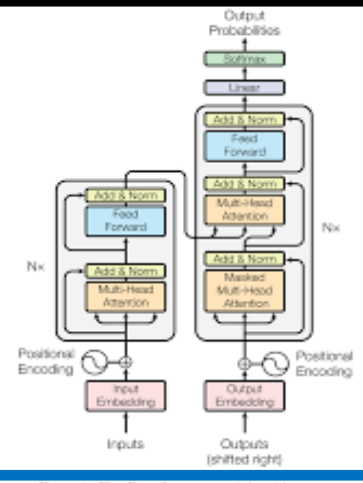
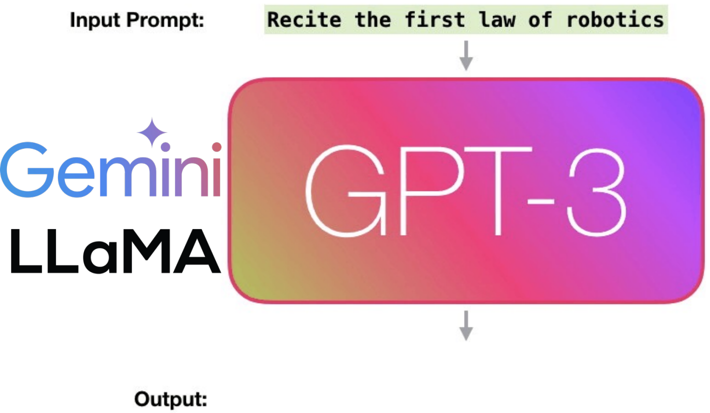
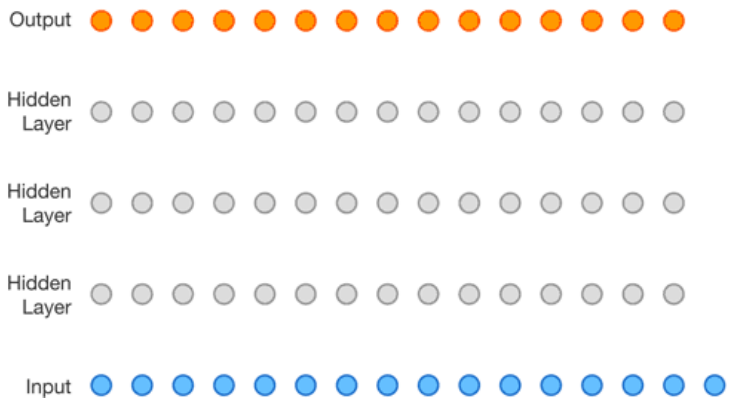


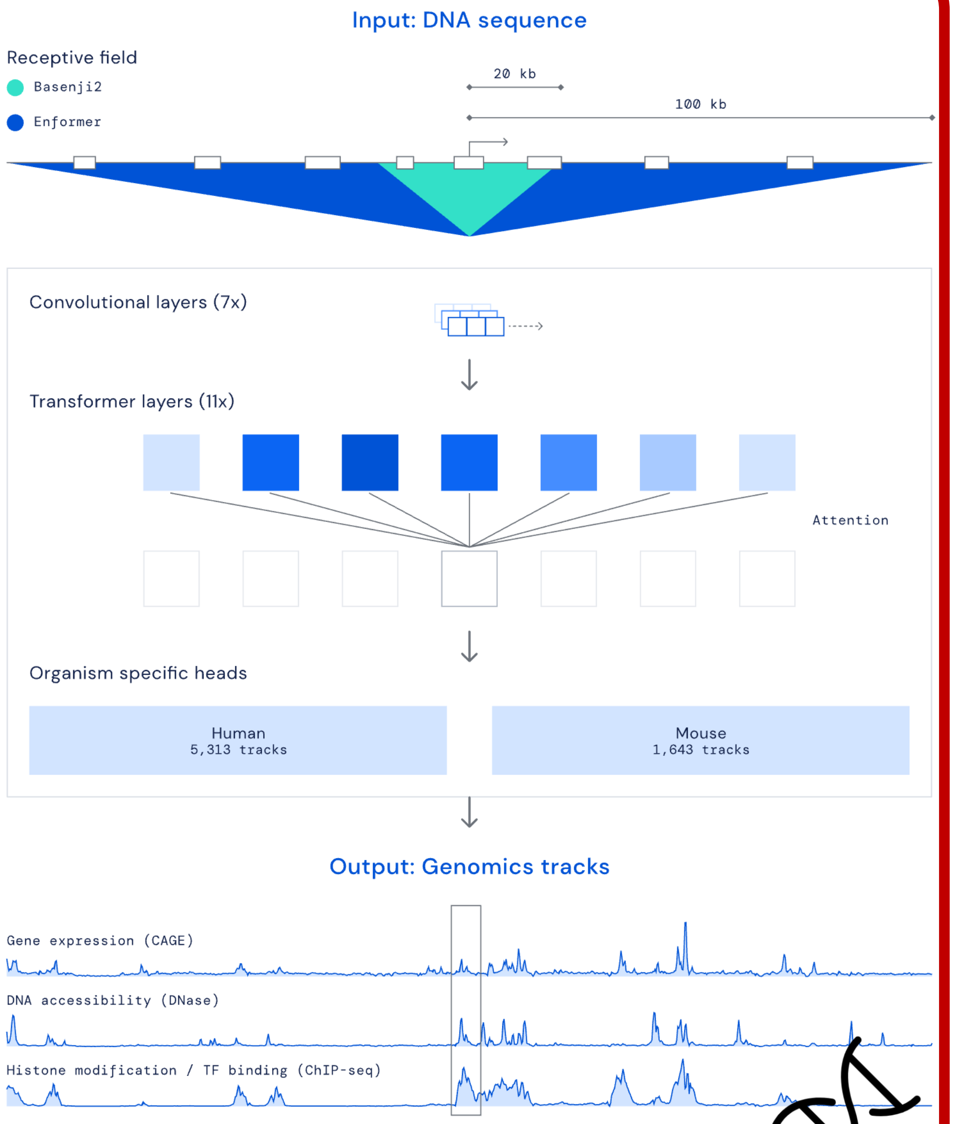
Figure 1: The Transformer - model architecture.

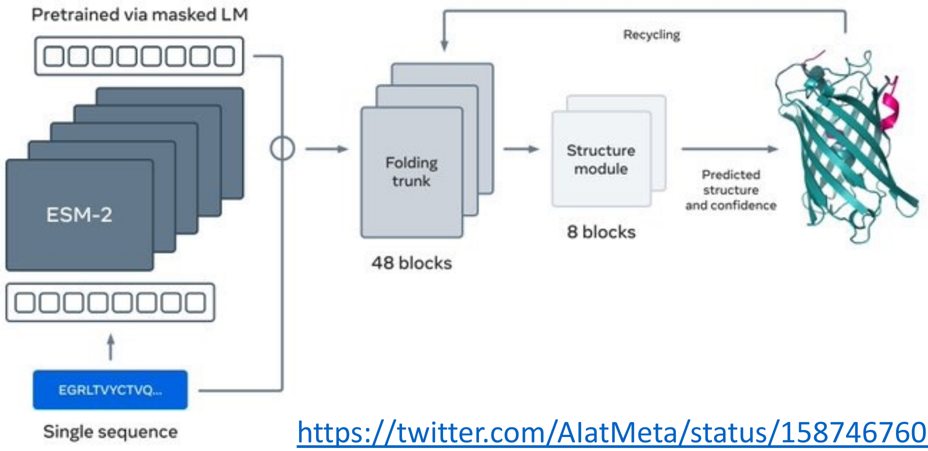




https://jalammar.github.io/how-gpt3-works-visualizations-animations/

https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/





ESMFold

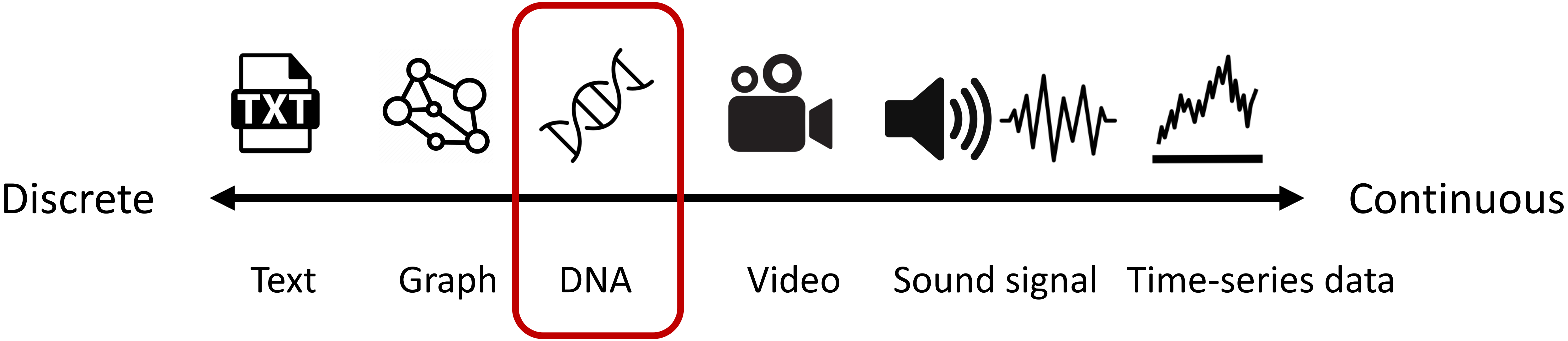
https://twitter.com/AIatMeta/status/1587467600413351937/photo/1

Enformer

https://deepmind.google/discover/blog/predicting-gene-expression-with-ai/

Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

Spectrum of Sequential Data



Discrete

Continuous

Text

Graph

DNA

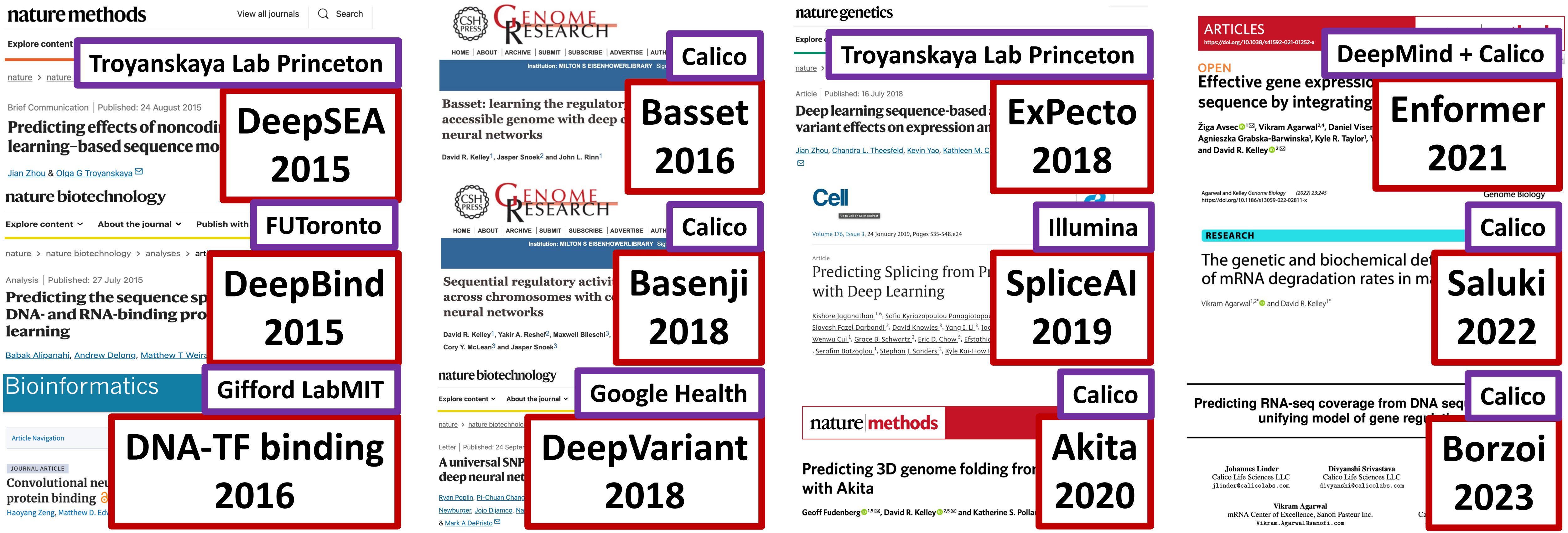
Video

Sound signal

Time-series data

DeepMind Calico illumina Troyanskaya Laboratory

Why Convolutional Neural Network to DNA ?



nature methods | View all journals | Search

Explore content

Troyanskaya Lab Princeton

nature > nature

Brief Communication | Published: 24 August 2015

Predicting effects of noncodi
learning–based sequence mo

Jian Zhou & Olga G Troyanskaya

DeepSEA 2015

nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with

FUToronto

nature > nature biotechnology > analyses > art

Analysis | Published: 27 July 2015

Predicting the sequence sp
DNA- and RNA-binding pro
learning

Babak Alipanahi, Andrew Delong, Matthew T Weira

DeepBind 2015

Bioinformatics

Gifford LabMIT

Article Navigation

JOURNAL ARTICLE

Convolutional neu
protein binding

Haoyang Zeng, Matthew D. Ed

DNA-TF binding 2016

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTH

Institution: MILTON S EISENHOWERLIBRARY Sig

Calico

Basset: learning the regulator
accessible genome with deep
neural networks

David R. Kelley[1], Jasper Snoek[2] and John L. Rinn[1]

Basset 2016

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTH

Institution: MILTON S EISENHOWERLIBRARY Sig

Calico

Sequential regulatory activi
across chromosomes with c
neural networks

David R. Kelley[1], Yakir A. Reshef[2], Maxwell Bileschi[3],
Cory Y. McLean[3] and Jasper Snoek[3]

Basenji 2018

nature biotechnology

Explore content ⌄ About the journal ⌄

Google Health

nature > nature biotechnolo

Letter | Published: 24 Septe

A universal SNP
deep neural net

Ryan Poplin, Pi-Chuan Chang
Newburger, Jojo Dijamco, Na
& Mark A DePristo

DeepVariant 2018

nature genetics

Explore

Troyanskaya Lab Princeton

nature >

Article | Published: 16 July 2018

Deep learning sequence-based
variant effects on expression an

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. C

ExPecto 2018

Cell

Volume 176, Issue 3, 24 January 2019, Pages 535-548.e24

Illumina

Article

Predicting Splicing from Pr
with Deep Learning

Kishore Jaganathan [1] [6], Sofia Kyriazopoulou Panagiotopou
Siavash Fazel Darbandi [2], David Knowles [3], Yang I. Li [3], Ja
Wenwu Cui [1], Grace B. Schwartz [2], Eric D. Chow [5], Efstathi
, Serafim Batzoglou [1], Stephan J. Sanders [2], Kyle Kai-How F

SpliceAI 2019

nature methods

Calico

Predicting 3D genome folding fro
with Akita

Geoff Fudenberg[1,5], David R. Kelley[2,5] and Katherine S. Polla

Akita 2020

ARTICLES

https://doi.org/10.1038/s41592-021-01252-x

DeepMind + Calico

OPEN

Effective gene expressio
sequence by integrating

Žiga Avsec[1], Vikram Agarwal[2,4], Daniel Viser
Agnieszka Grabska-Barwinska[1], Kyle R. Taylor[1],
and David R. Kelley[2]

Enformer 2021

Agarwal and Kelley Genome Biology (2022) 23:245
https://doi.org/10.1186/s13059-022-02811-x

Genome Biology

RESEARCH

Calico

The genetic and biochemical de
of mRNA degradation rates in m

Vikram Agarwal[1,2*] and David R. Kelley[1*]

Saluki 2022

Predicting RNA-seq coverage from DNA seq
unifying model of gene regu

Calico

Johannes Linder
Calico Life Sciences LLC
jlinder@calicolabs.com

Divyanshi Srivastava
Calico Life Sciences LLC
divyanshi@calicolabs.com

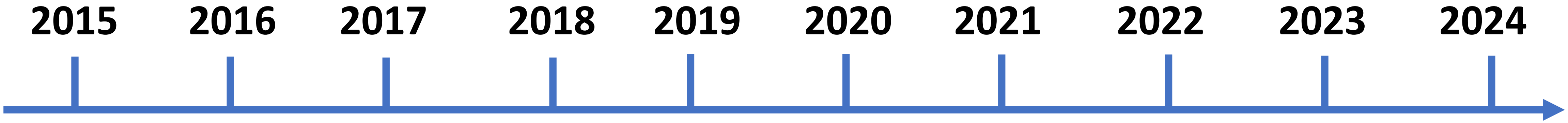
Vikram Agarwal
mRNA Center of Excellence, Sanofi Pasteur Inc.
Vikram.Agarwal@sanofi.com

Ca

Borzoi 2023

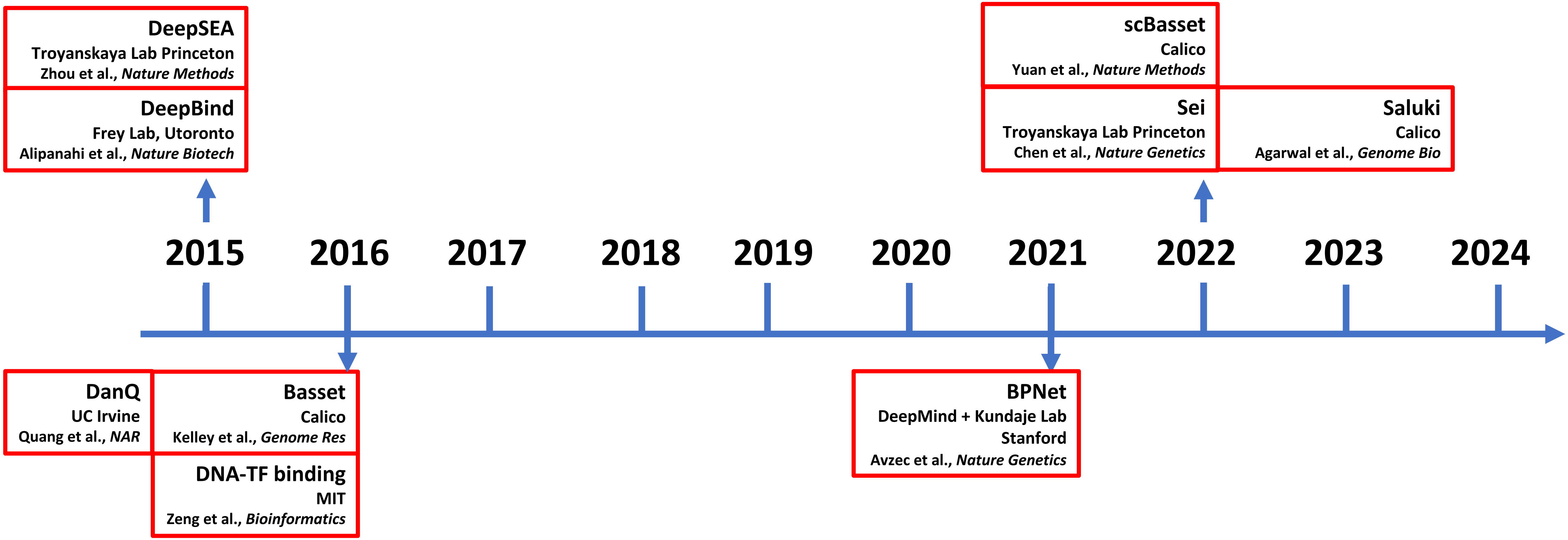
DeepMind Calico illumina® Troyanskaya Laboratory

Why Convolutional Neural Network to DNA ?



2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

Why Convolutional Neural Network to DNA ?



DeepSEA
Troyanskaya Lab Princeton
Zhou et al., *Nature Methods*

DeepBind
Frey Lab, Utoronto
Alipanahi et al., *Nature Biotech*

DanQ
UC Irvine
Quang et al., *NAR*

Basset
Calico
Kelley et al., *Genome Res*

DNA-TF binding
MIT
Zeng et al., *Bioinformatics*

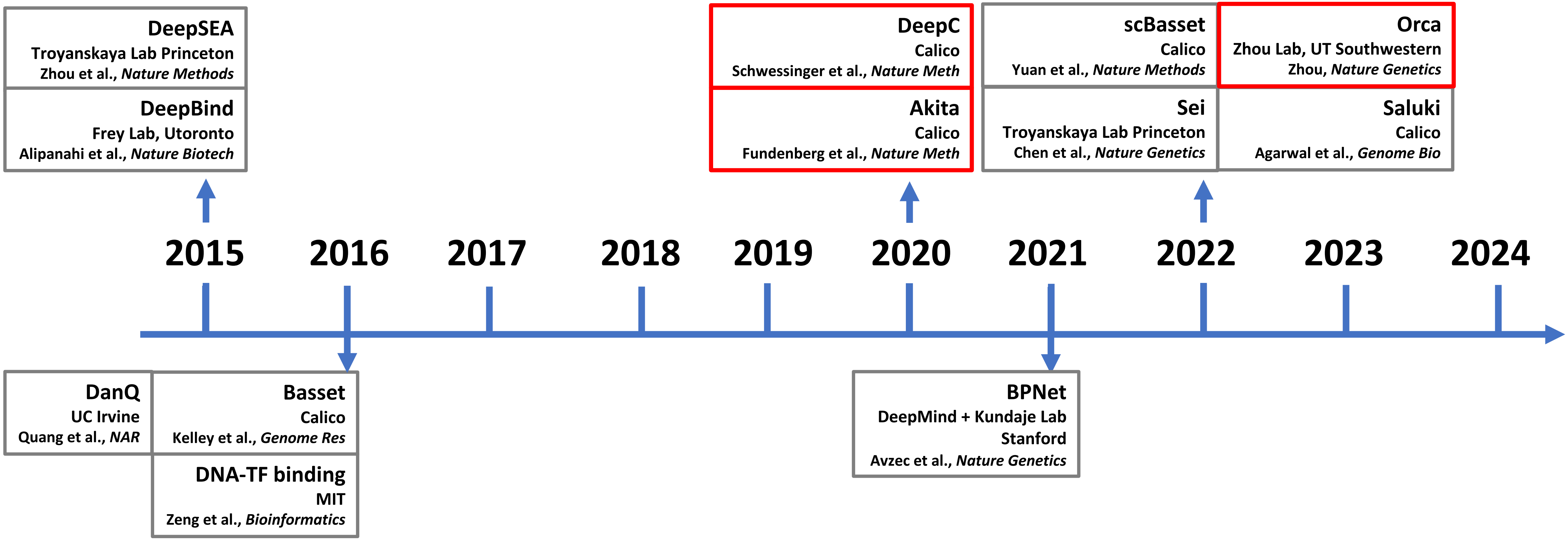
BPNet
DeepMind + Kundaje Lab
Stanford
Avzec et al., *Nature Genetics*

scBasset
Calico
Yuan et al., *Nature Methods*

Sei
Troyanskaya Lab Princeton
Chen et al., *Nature Genetics*

Saluki
Calico
Agarwal et al., *Genome Bio*

Why Convolutional Neural Network to DNA ?



DeepSEA
Troyanskaya Lab Princeton
Zhou et al., *Nature Methods*

DeepBind
Frey Lab, Utoronto
Alipanahi et al., *Nature Biotech*

DeepC
Calico
Schwessinger et al., *Nature Meth*

Akita
Calico
Fundenberg et al., *Nature Meth*

scBasset
Calico
Yuan et al., *Nature Methods*

Sei
Troyanskaya Lab Princeton
Chen et al., *Nature Genetics*

Orca
Zhou Lab, UT Southwestern
Zhou, *Nature Genetics*

Saluki
Calico
Agarwal et al., *Genome Bio*

DanQ
UC Irvine
Quang et al., *NAR*

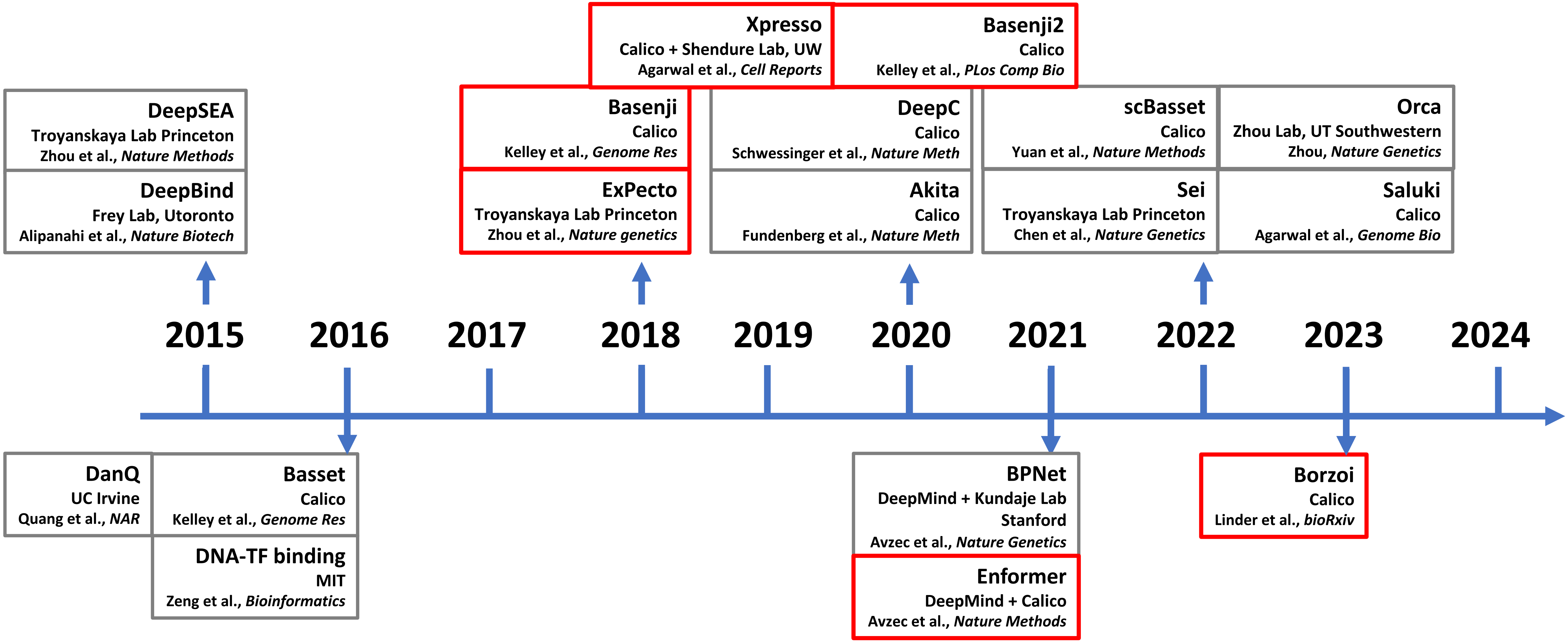
Basset
Calico
Kelley et al., *Genome Res*

DNA-TF binding
MIT
Zeng et al., *Bioinformatics*

BPNet
DeepMind + Kundaje Lab
Stanford
Avzec et al., *Nature Genetics*

DeepMind Calico illumina Troyanskaya Laboratory

Why Convolutional Neural Network to DNA ?



DeepSEA
Troyanskaya Lab Princeton
Zhou et al., *Nature Methods*

DeepBind
Frey Lab, Utoronto
Alipanahi et al., *Nature Biotech*

Xpresso
Calico + Shendure Lab, UW
Agarwal et al., *Cell Reports*

Basenji2
Calico
Kelley et al., *PLos Comp Bio*

Basenji
Calico
Kelley et al., *Genome Res*

ExPecto
Troyanskaya Lab Princeton
Zhou et al., *Nature genetics*

DeepC
Calico
Schwessinger et al., *Nature Meth*

Akita
Calico
Fundenberg et al., *Nature Meth*

scBasset
Calico
Yuan et al., *Nature Methods*

Sei
Troyanskaya Lab Princeton
Chen et al., *Nature Genetics*

Orca
Zhou Lab, UT Southwestern
Zhou, *Nature Genetics*

Saluki
Calico
Agarwal et al., *Genome Bio*

DanQ
UC Irvine
Quang et al., *NAR*

Basset
Calico
Kelley et al., *Genome Res*

DNA-TF binding
MIT
Zeng et al., *Bioinformatics*

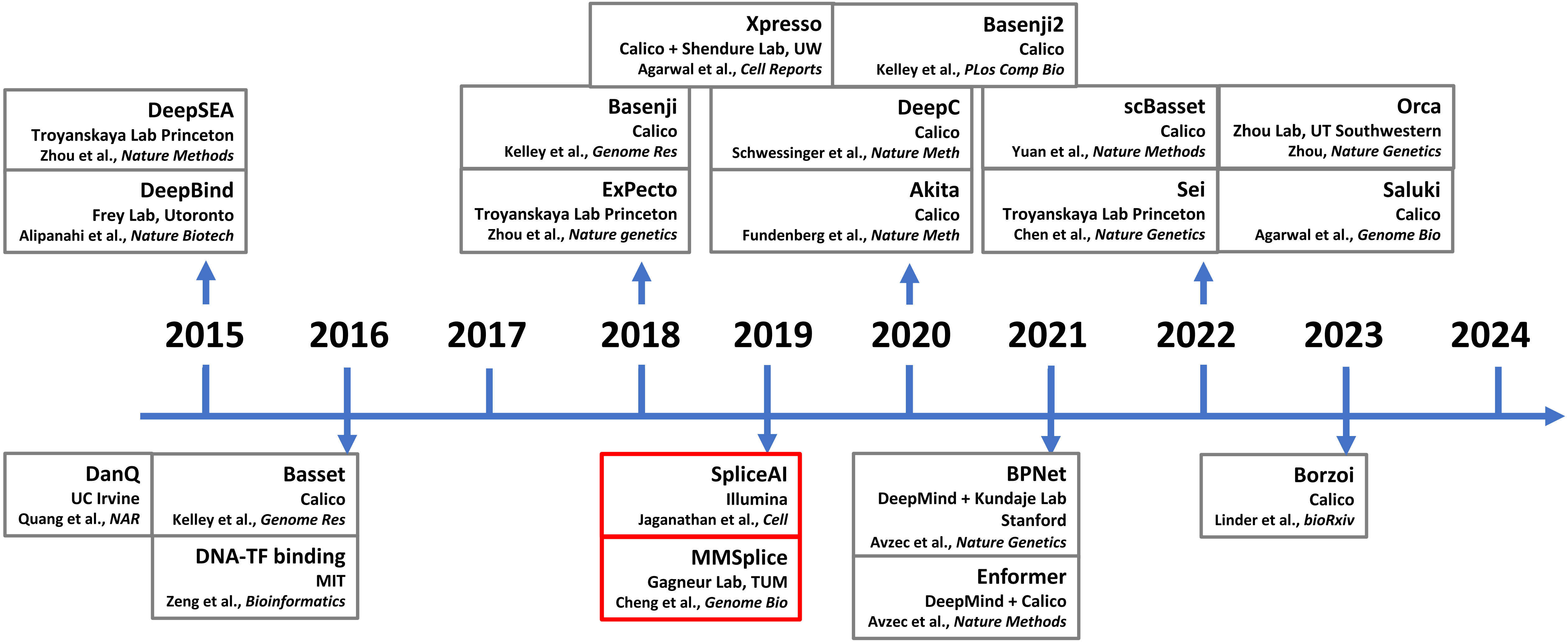
BPNet
DeepMind + Kundaje Lab
Stanford
Avzec et al., *Nature Genetics*

Enformer
DeepMind + Calico
Avzec et al., *Nature Methods*

Borzoi
Calico
Linder et al., *bioRxiv*

DeepMind Calico illumina® Troyanskaya Laboratory

Why Convolutional Neural Network to DNA ?



DeepSEA
Troyanskaya Lab Princeton
Zhou et al., *Nature Methods*

DeepBind
Frey Lab, Utoronto
Alipanahi et al., *Nature Biotech*

DanQ
UC Irvine
Quang et al., *NAR*

Basset
Calico
Kelley et al., *Genome Res*

DNA-TF binding
MIT
Zeng et al., *Bioinformatics*

Xpresso
Calico + Shendure Lab, UW
Agarwal et al., *Cell Reports*

Basenji
Calico
Kelley et al., *Genome Res*

ExPecto
Troyanskaya Lab Princeton
Zhou et al., *Nature genetics*

SpliceAI
Illumina
Jaganathan et al., *Cell*

MMSplice
Gagneur Lab, TUM
Cheng et al., *Genome Bio*

Basenji2
Calico
Kelley et al., *PLos Comp Bio*

DeepC
Calico
Schwessinger et al., *Nature Meth*

Akita
Calico
Fundenberg et al., *Nature Meth*

BPNet
DeepMind + Kundaje Lab
Stanford
Avzec et al., *Nature Genetics*

Enformer
DeepMind + Calico
Avzec et al., *Nature Methods*

scBasset
Calico
Yuan et al., *Nature Methods*

Sei
Troyanskaya Lab Princeton
Chen et al., *Nature Genetics*

Orca
Zhou Lab, UT Southwestern
Zhou, *Nature Genetics*

Saluki
Calico
Agarwal et al., *Genome Bio*

Borzoi
Calico
Linder et al., *bioRxiv*

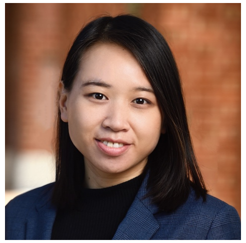
“All models are wrong, but some are useful”

- George Box, 1978

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015).
Time series analysis: forecasting and control. John Wiley & Sons.

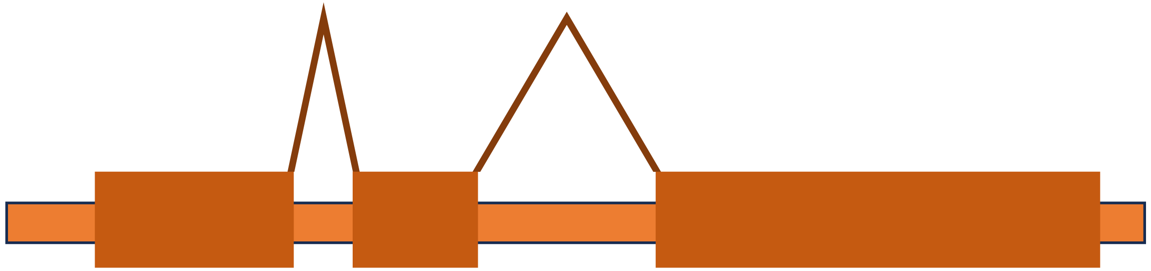
Steven Salzberg

Mihaela Pertea


Anqi Liu

SpliceAI-toolkit: splice site predictor

Pre-mRNA

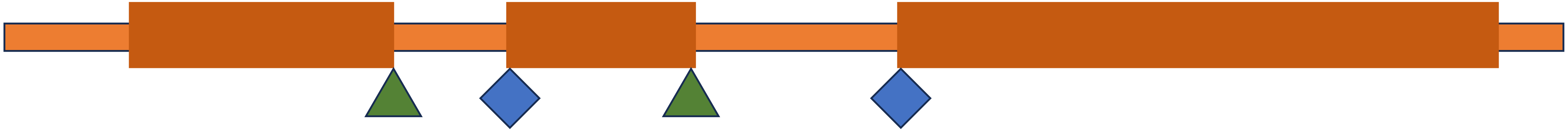


<span>**Chao, K. H.**</span>, Mao, A., Liu, Anqi, Salzberg, S. L., & Pertea, M. (**2024**). SpliceAI-toolkit. Manuscript in preparation. 📕 **https://ccb.jhu.edu/spliceai-toolkit/**

Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (**2023**). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** 📕 **https://ccb.jhu.edu/splam/**

Can we predict splice sites using only DNA?

Yes!



X AGACTCAGCCCCCGGAGACTTAGTTAGAGGAAGAAAAAGGTAGGACAGAAGAAAAAGGCAGGACATACAAGGTGCTGGCCCAGGGCGG

Y 0000000000000000000002000000010000000020000000000001000000000000000000000000000000000000

▲ Donor: 2

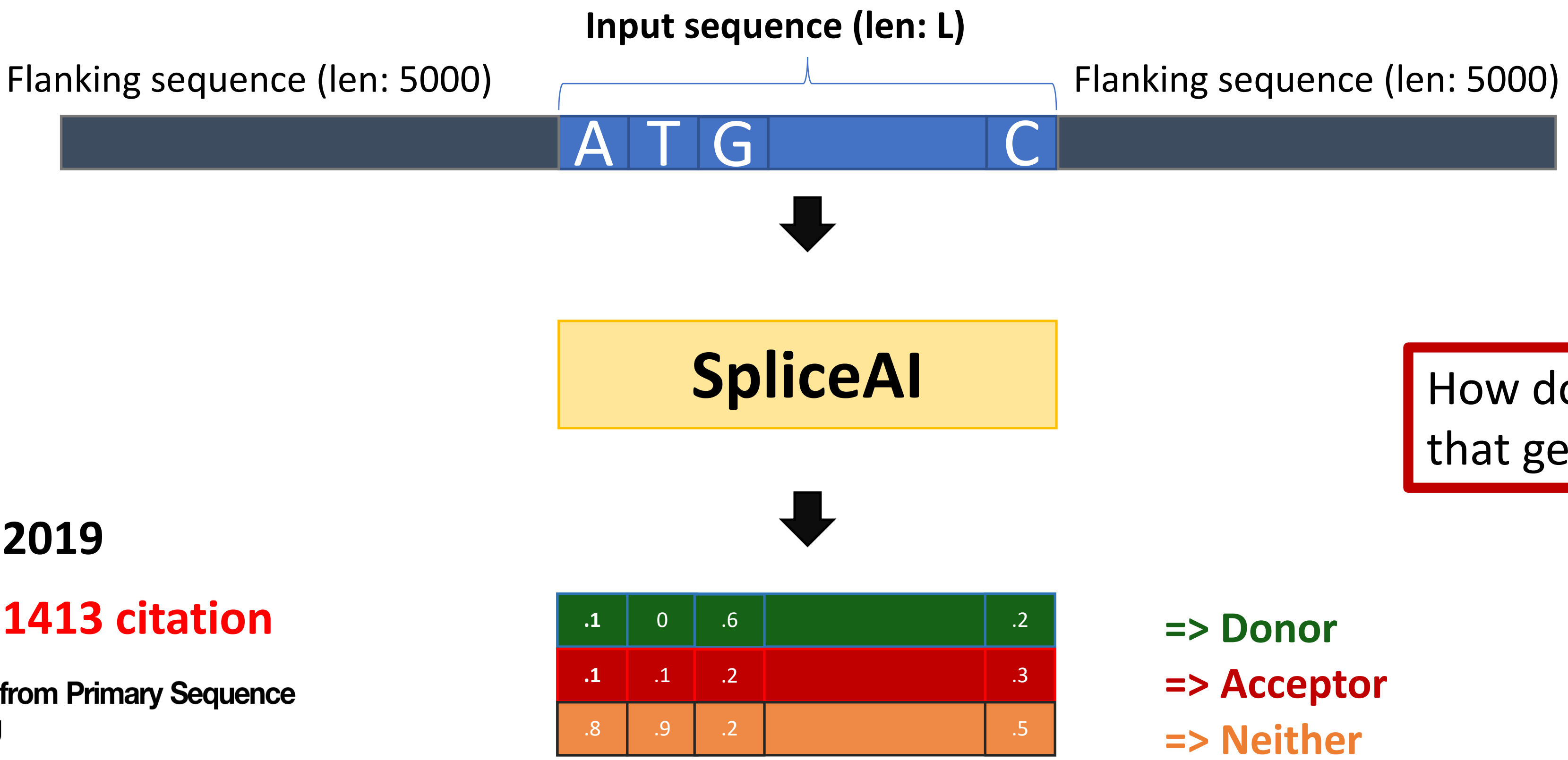
◆ Acceptor: 1

Neither: 0

SpliceAI: splite site predictor

$$\begin{bmatrix} A\\ C\\ G\\ T\\ N\end{bmatrix} = \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix}$$

$$\begin{bmatrix} Acceptor\\ Donor\\ Neither\\ Padding\end{bmatrix} = \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\\0&0&0\end{bmatrix}$$



.1	0	.6		.2
.1	.1	.2		.3
.8	.9	.2		.5

=> Donor
=> Acceptor
=> Neither

Dimension:

$\mathbf{X}$: $(L+10000) * 4$

How do we train the model given that gene sequences vary?

$\mathbf{Y}$: $L * 3$

2019

Cell

1413 citation

Predicting Splicing from Primary Sequence with Deep Learning

Kishore Jaganathan,[1,6] Sofia Kyriazopoulou Panagiotopoulou,[1,6] Jeremy F. McRae,[1,6] Siavash Fazel Darbandi,[2] David Knowles,[3] Yang I. Li,[3] Jack A. Kosmicki,[1,4] Juan Arbelaez,[2] Wenwu Cui,[1] Grace B. Schwartz,[2] Eric D. Chow,[5] Efstathios Kanterakis,[1] Hong Gao,[1] Amirali Kia,[1] Serafim Batzoglou,[1] Stephan J. Sanders,[2] and Kyle Kai-How Farh[1,7,*]

SpliceAI: data preprocessing

$\frac{F}{2}$ W $\frac{F}{2}$

X NN … NN**ATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCA**NN … NN

Y **00000002000001000000002000001000002000100000200000001000000**

X= [NN … NN**ATGTCGTGTCGAGTTG**]

Y= [**0000000200**]

SpliceAI: data preprocessing

$\frac{F}{2}$ W $\frac{F}{2}$

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 0000000200000100000002000001000002000100000200000001000000

X= [

NN … NNATGTCGTGTCGAGTTG

TCGTGTCGAGTTGTCGTGTTCAG

]

Y= [

0000000200

0001000000

]

SpliceAI: data preprocessing

$\frac{F}{2}$ W $\frac{F}{2}$

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 00000002000001000000020000010000020001000002000000001000000

X= [NN … NNATGTCGTGTCGAGTTG
TCGTGTCGAGTTGTCGTGTTCAG
TTGTCGTGTTCAGGTCAGTCAGG]

Y= [0000000200
0001000000
2000001000]

SpliceAI: data preprocessing

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 00000002000001000000002000001000002000100000200000001000000

Shape: $\left\lceil \frac{L}{W} \right\rceil * (\frac{F}{2} + W + \frac{F}{2})$

X =

NN … NNATGTCGTGTCGAGTTG

TCGTGTCGAGTTGTCGTGTTCAG

TTGTCGTGTTCAGGTCAGTCAGG

⋮

AAGTAGAGCTCANNNNN … NN

Shape: $\left\lceil \frac{L}{W} \right\rceil * (W)$

Y =

0000000200

0001000000

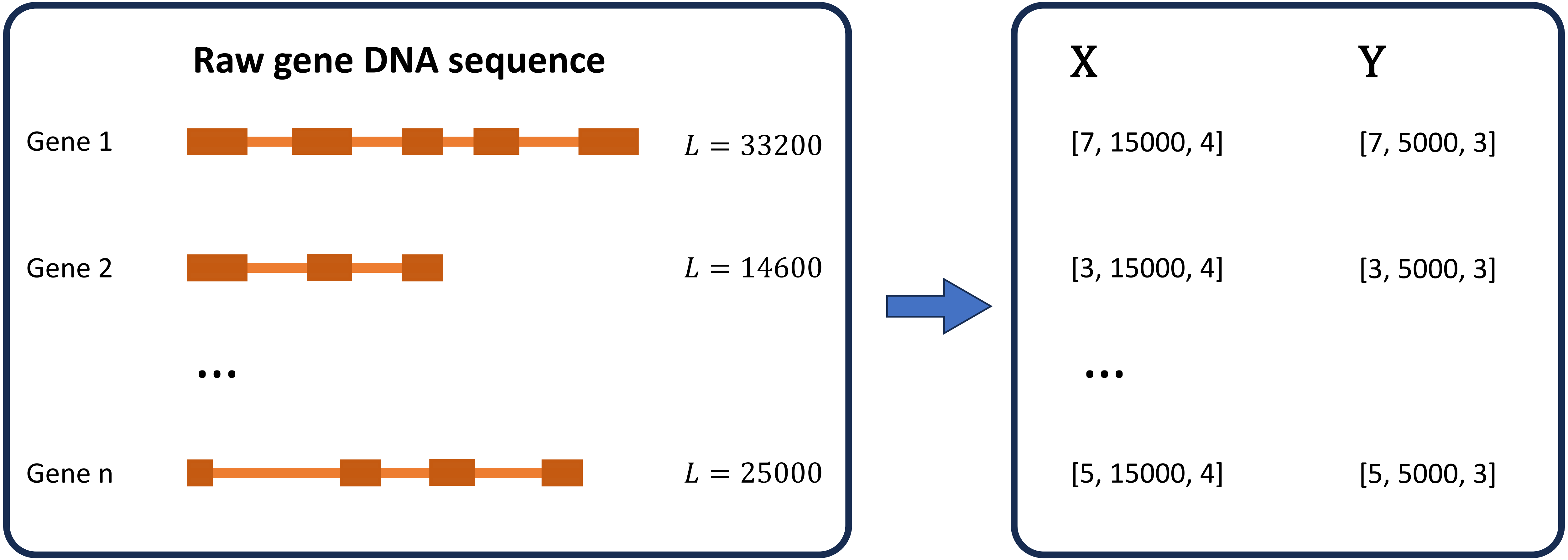
2000001000

⋮

0000000000

SpliceAI: data preprocessing

$W = 5000$ $F = 10{,}000$



Raw gene DNA sequence

Gene 1 $L = 33200$

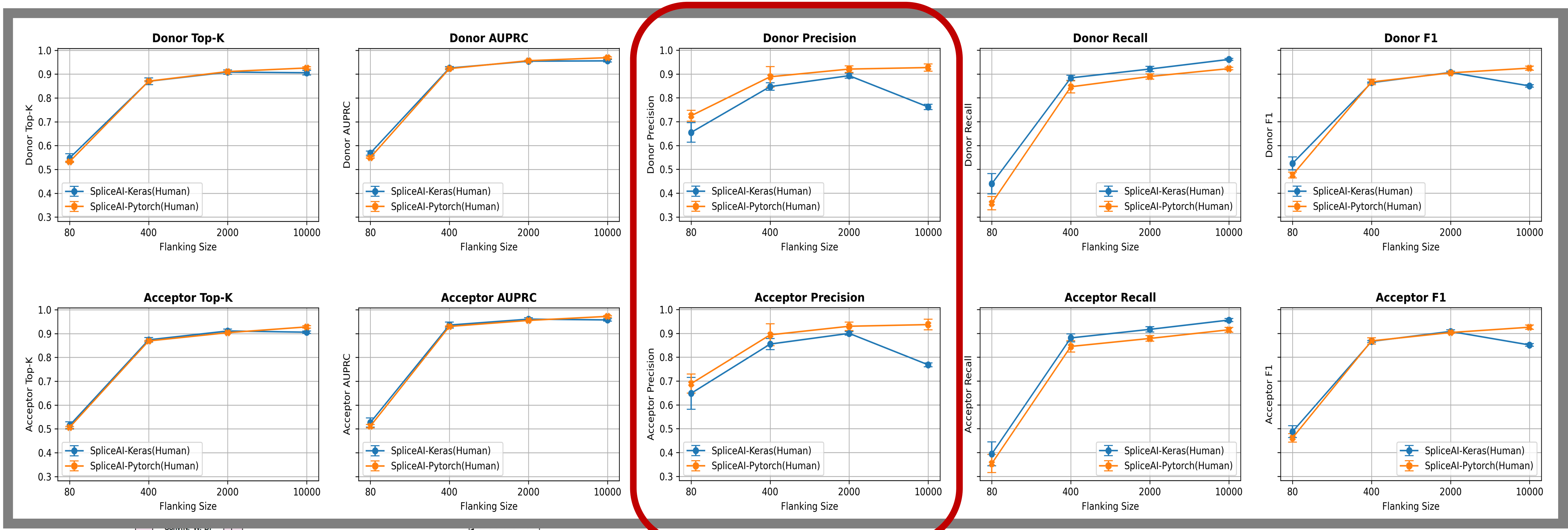
Gene 2 $L = 14600$

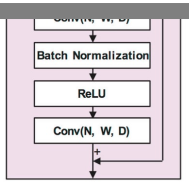
…

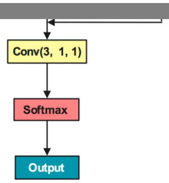
Gene n $L = 25000$

X	Y
[7, 15000, 4]	[7, 5000, 3]
[3, 15000, 4]	[3, 5000, 3]
…	
[5, 15000, 4]	[5, 5000, 3]

SpliceAI-toolkit : better than SpliceAI!



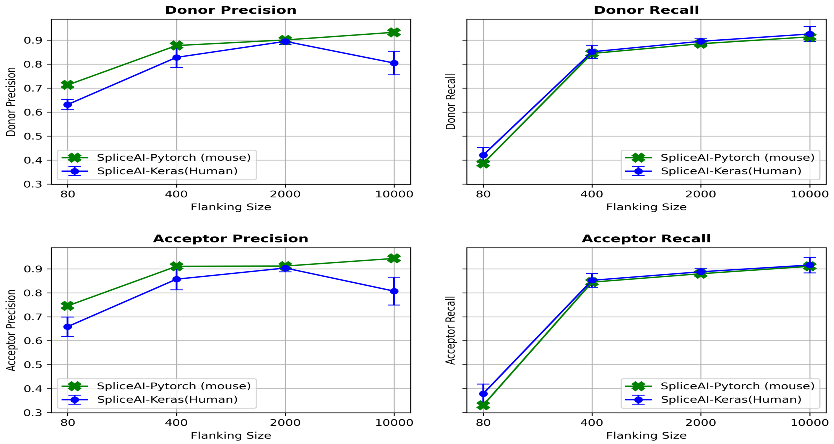




SpliceAI-toolkit : retrain on different species

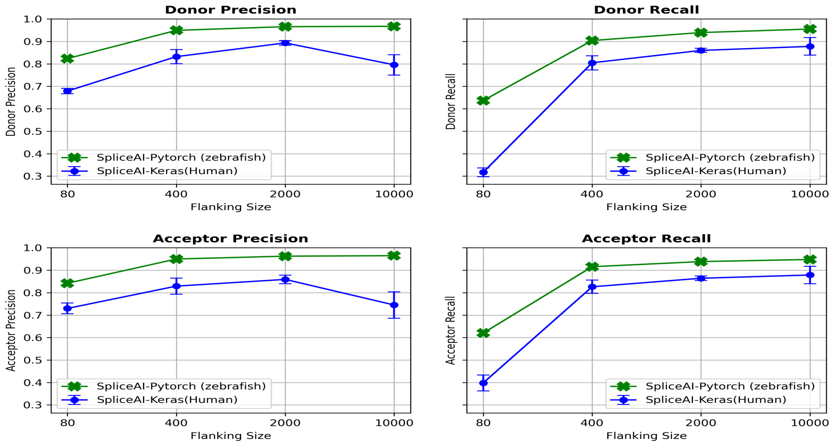
A

Splice site prediction metrics for mouse



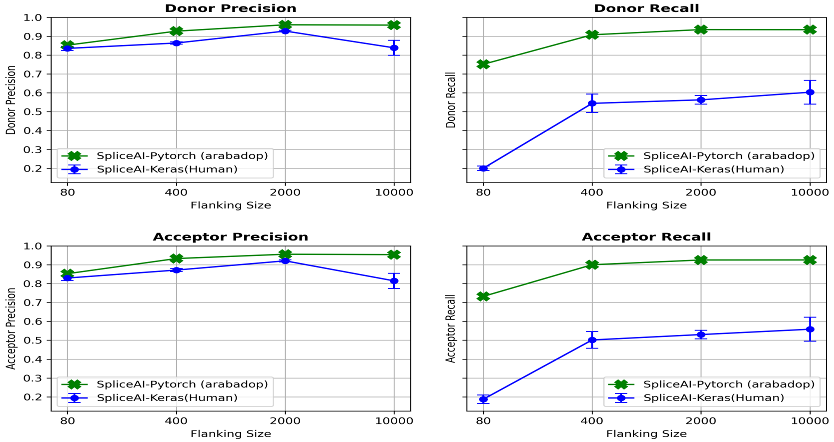
B

Splice site prediction metrics for zebrafish



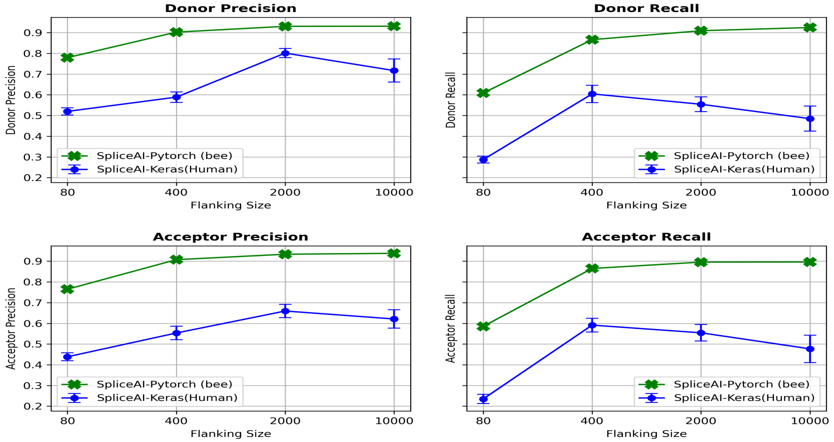
C

Splice site prediction metrics for arabadop

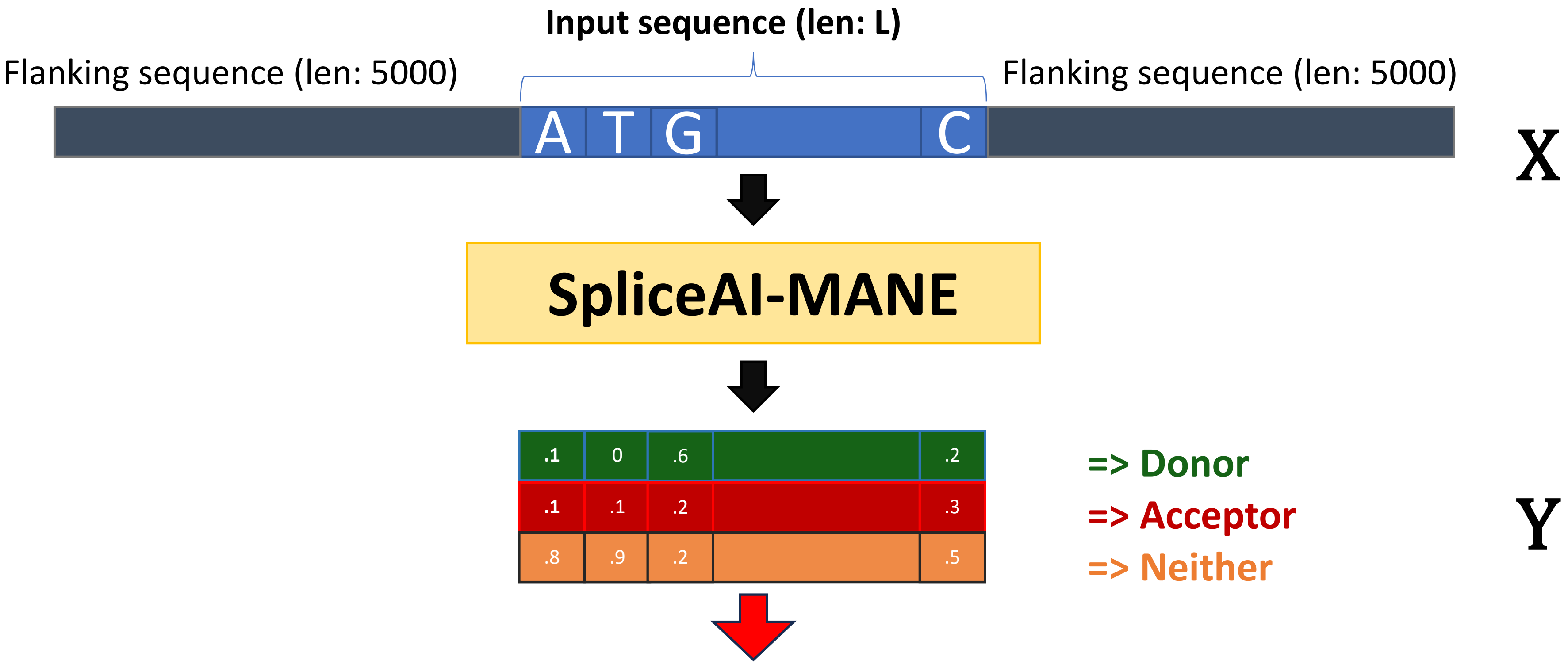


D

Splice site prediction metrics for bee



SpliceAI-toolkit : new concept – Calibration

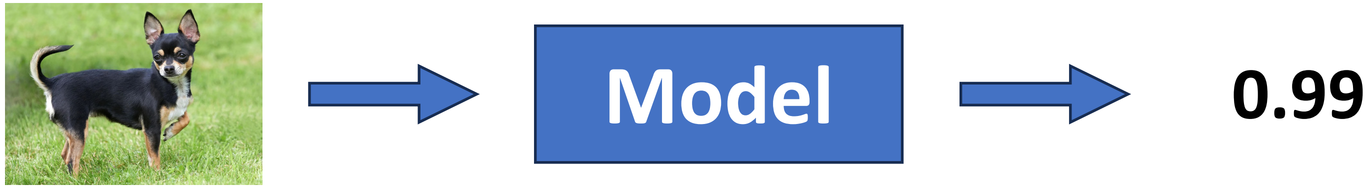


.1	0	.6		.2
.1	.1	.2		.3
.8	.9	.2		.5

- What do SpliceAI-MANE scores signify?
- Do the model's predicted probabilities accurately represent the true likelihood of an event occurring?"

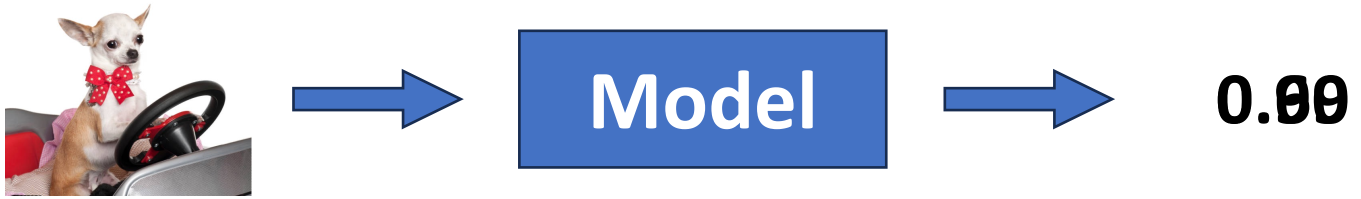
Binary classification example

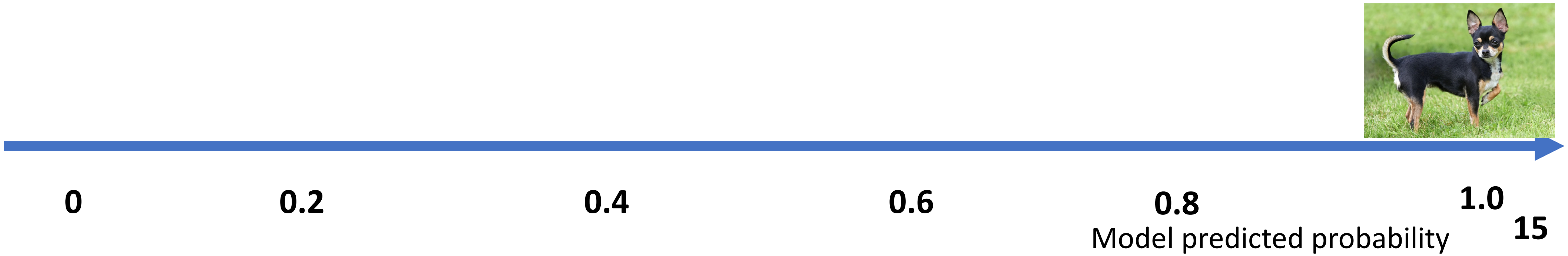
SpliceAI-toolkit : new concept – Calibration



Binary classification example

SpliceAI-toolkit : new concept – Calibration



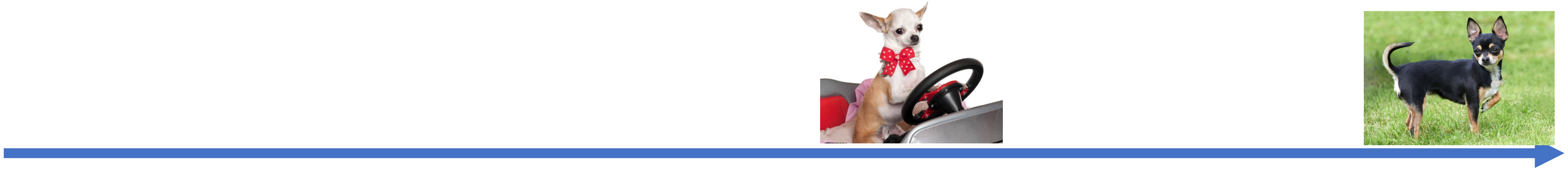


Binary classification example

SpliceAI-toolkit : new concept – Calibration

→ **Model** →

0.10
0.69



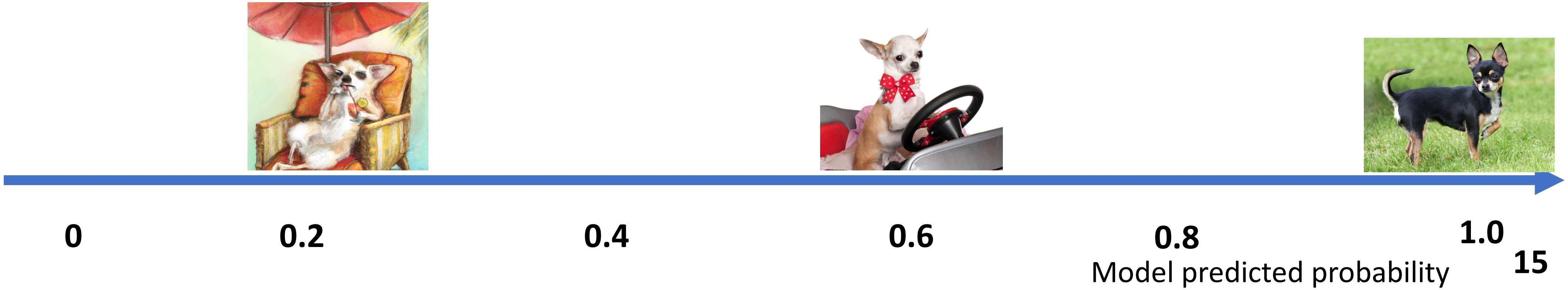
Binary classification example

SpliceAI-toolkit : new concept – Calibration

Model

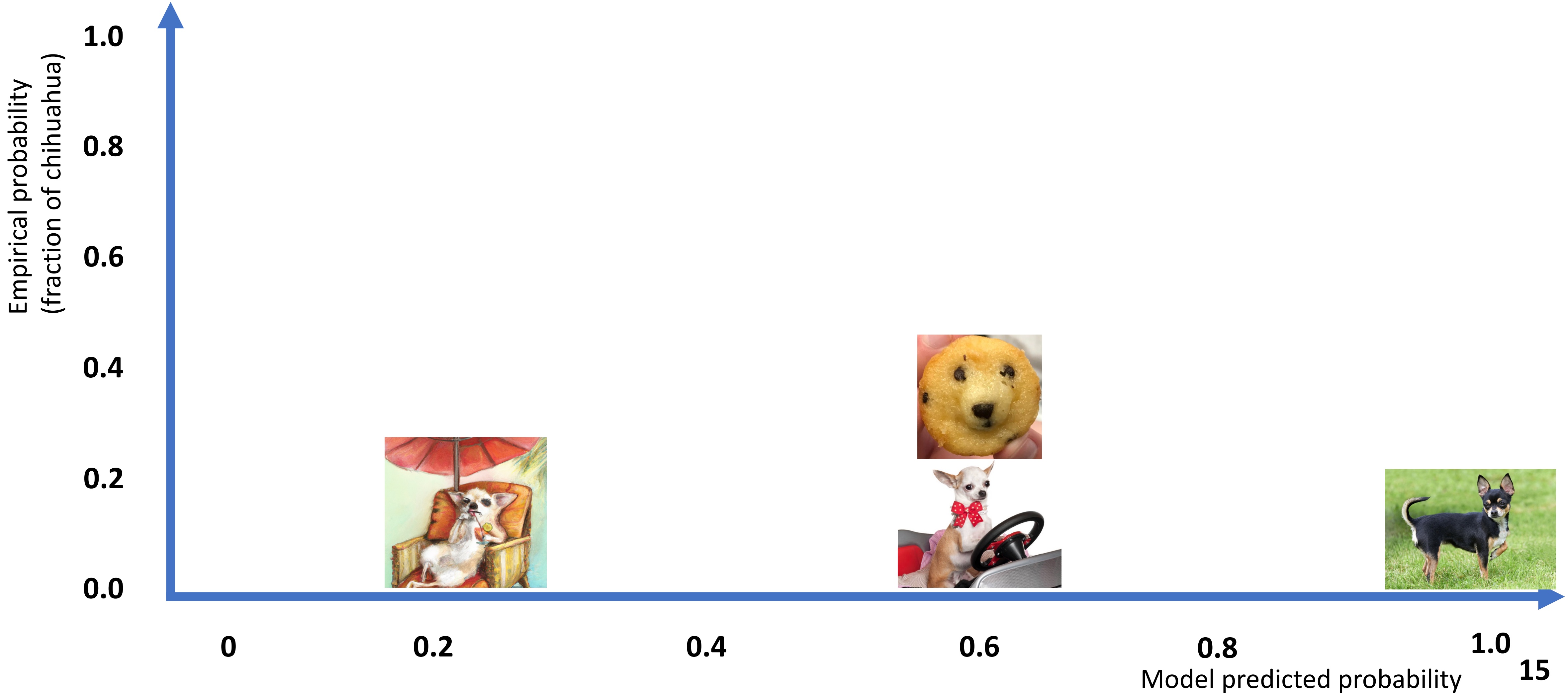
0.62

0.19



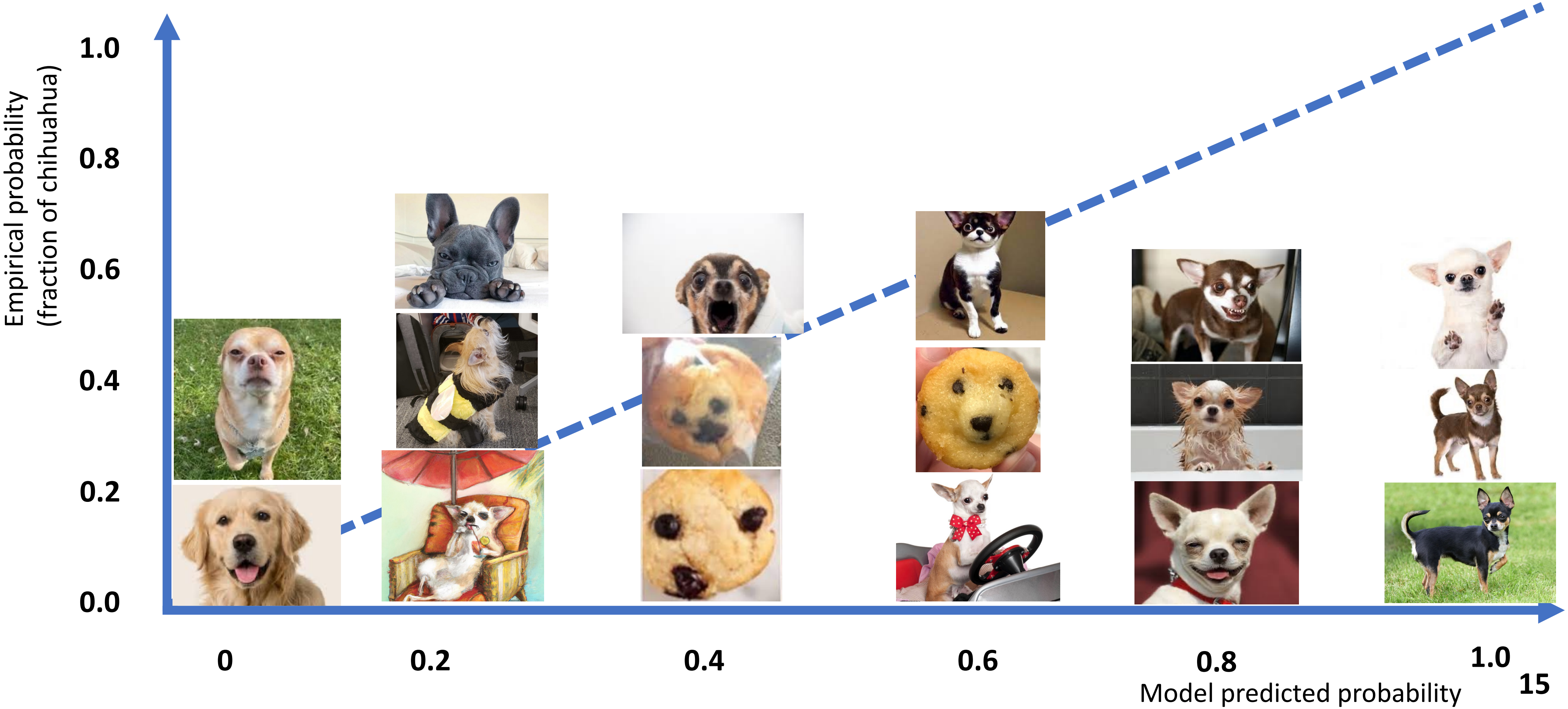
Binary classification example

SpliceAI-toolkit : new concept – Calibration



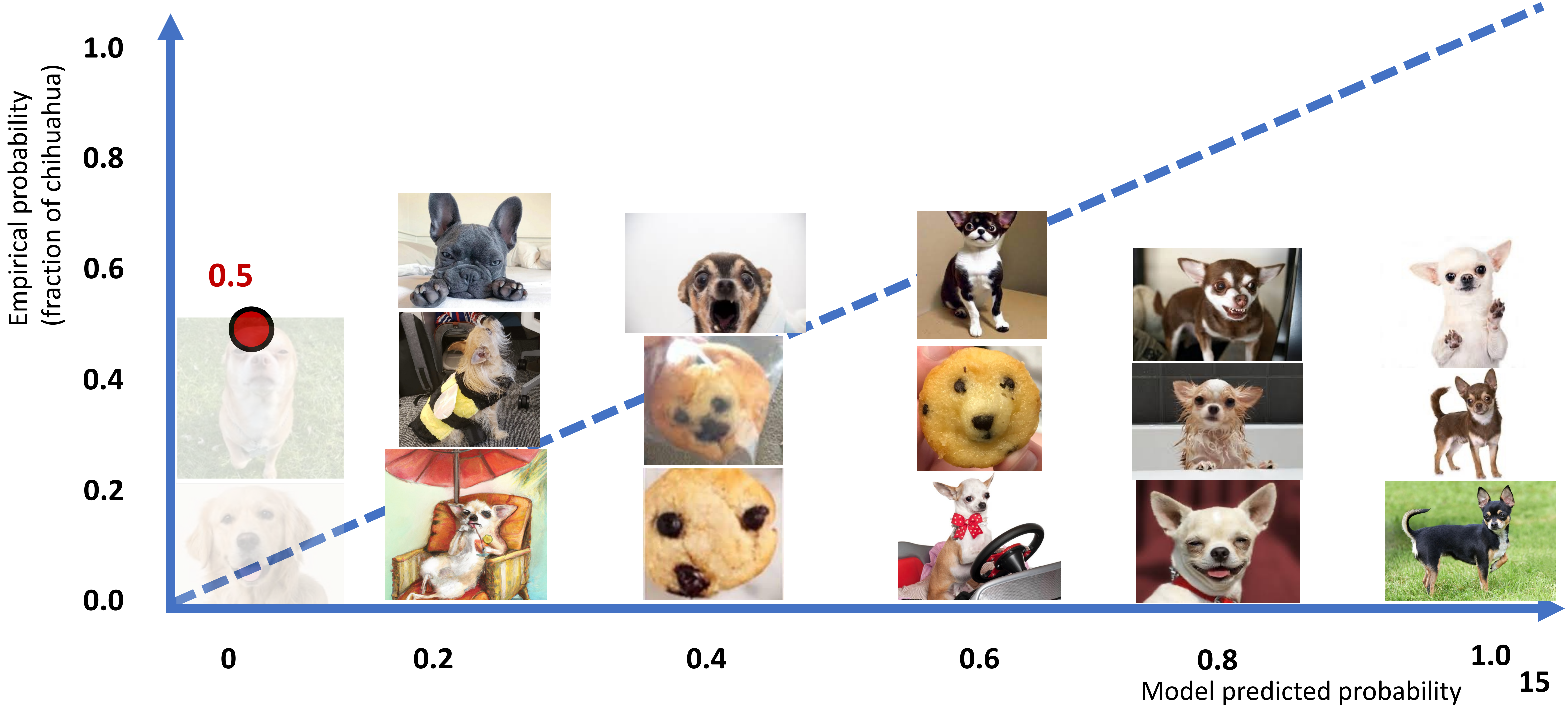
Binary classification example

SpliceAI-toolkit : new concept – Calibration



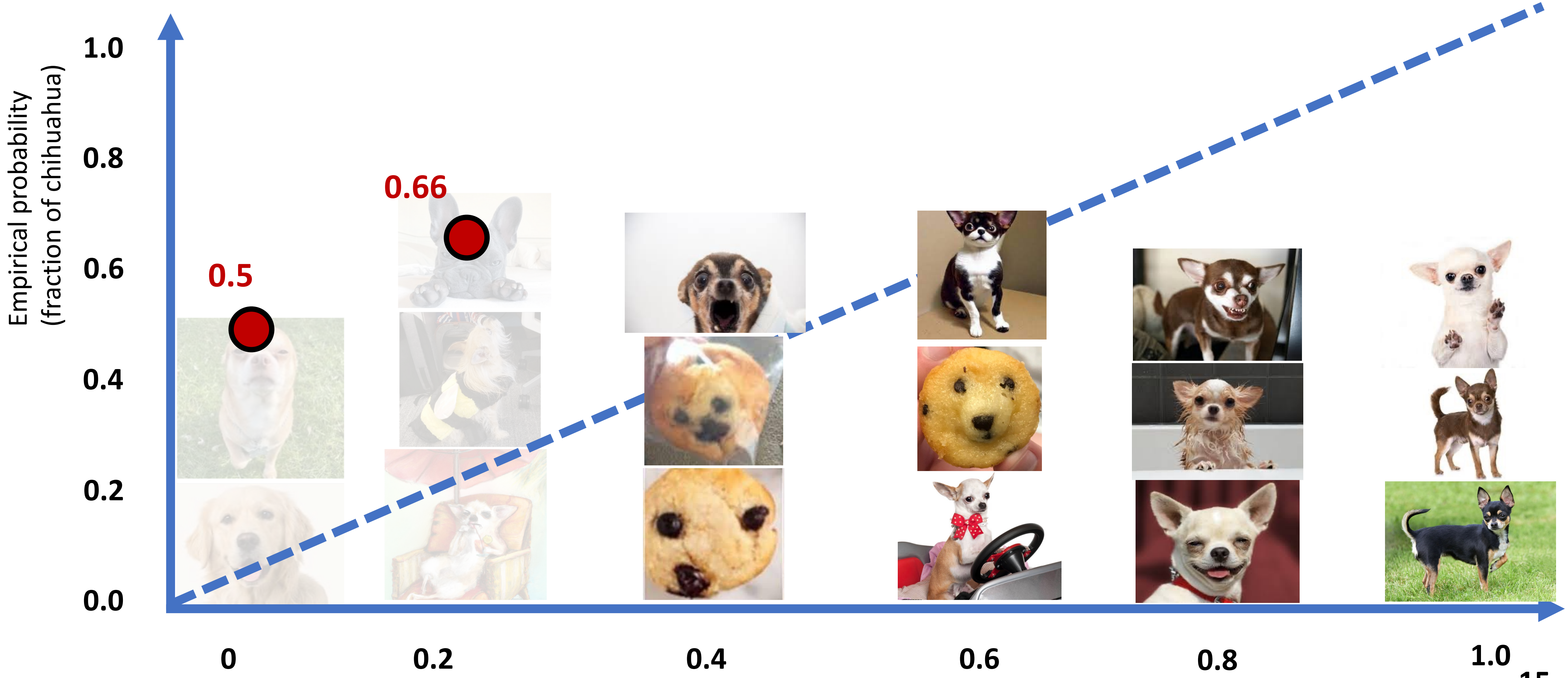
Binary classification example

SpliceAI-toolkit : new concept – Calibration



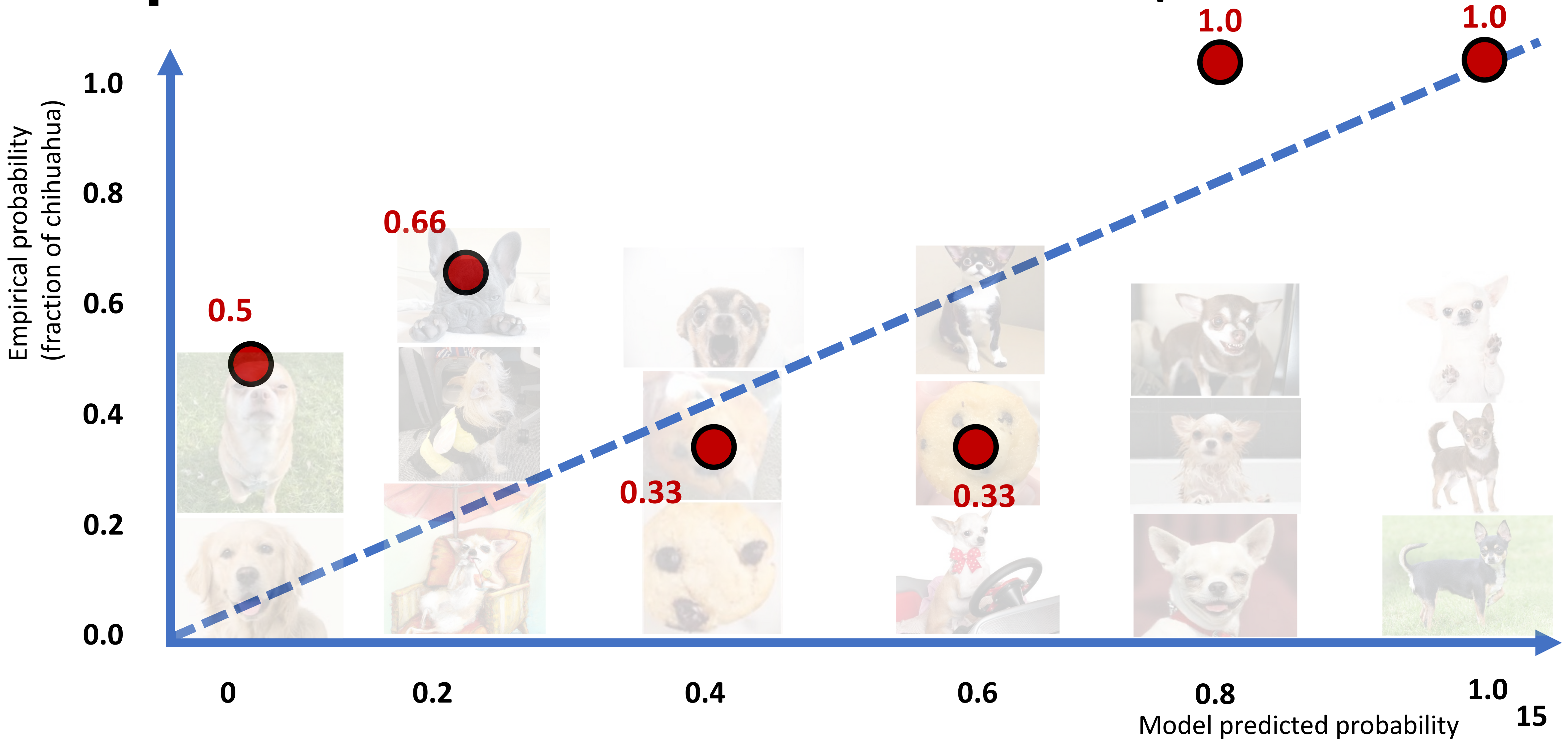
Binary classification example

SpliceAI-toolkit : new concept – Calibration



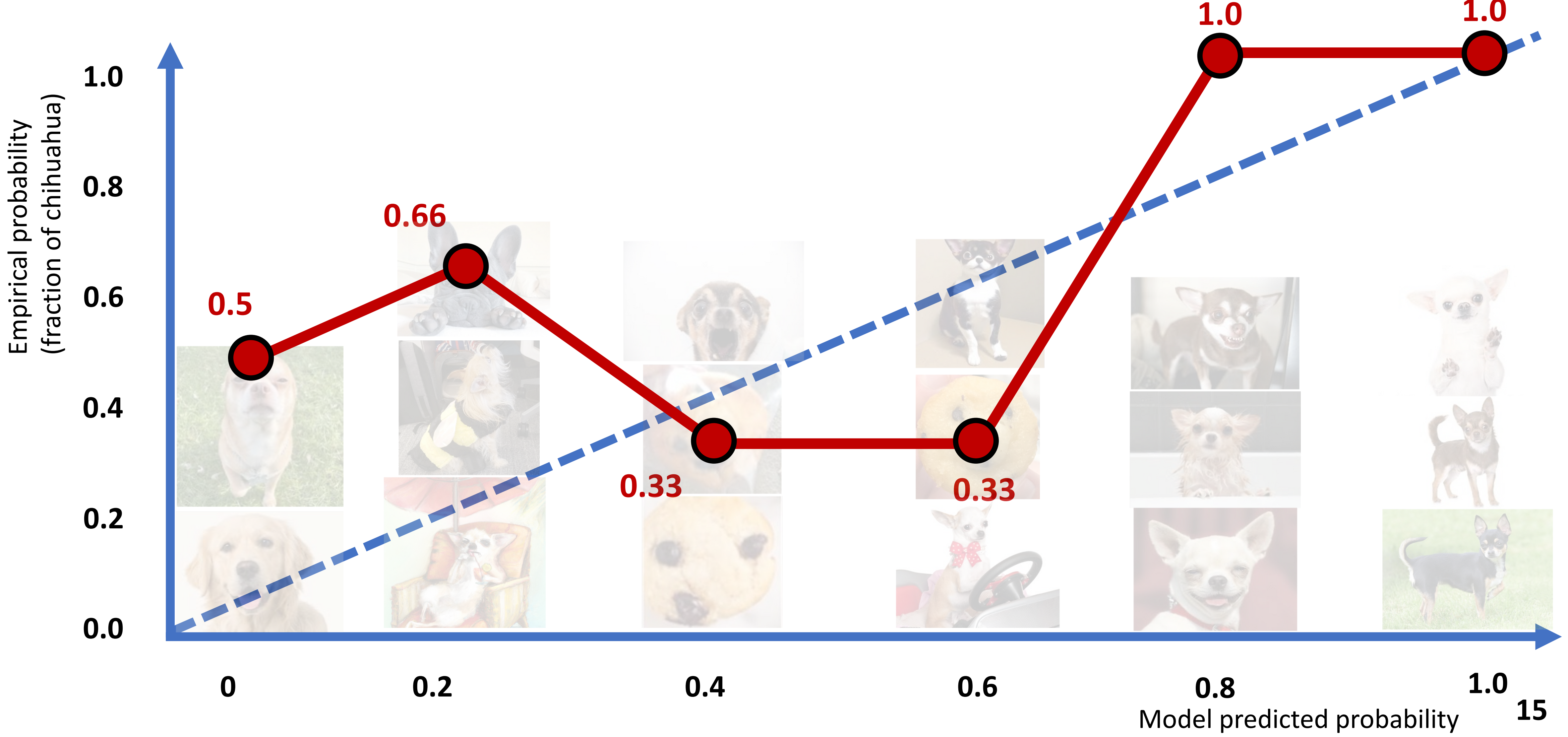
Binary classification example

SpliceAI-toolkit : new concept – Calibration



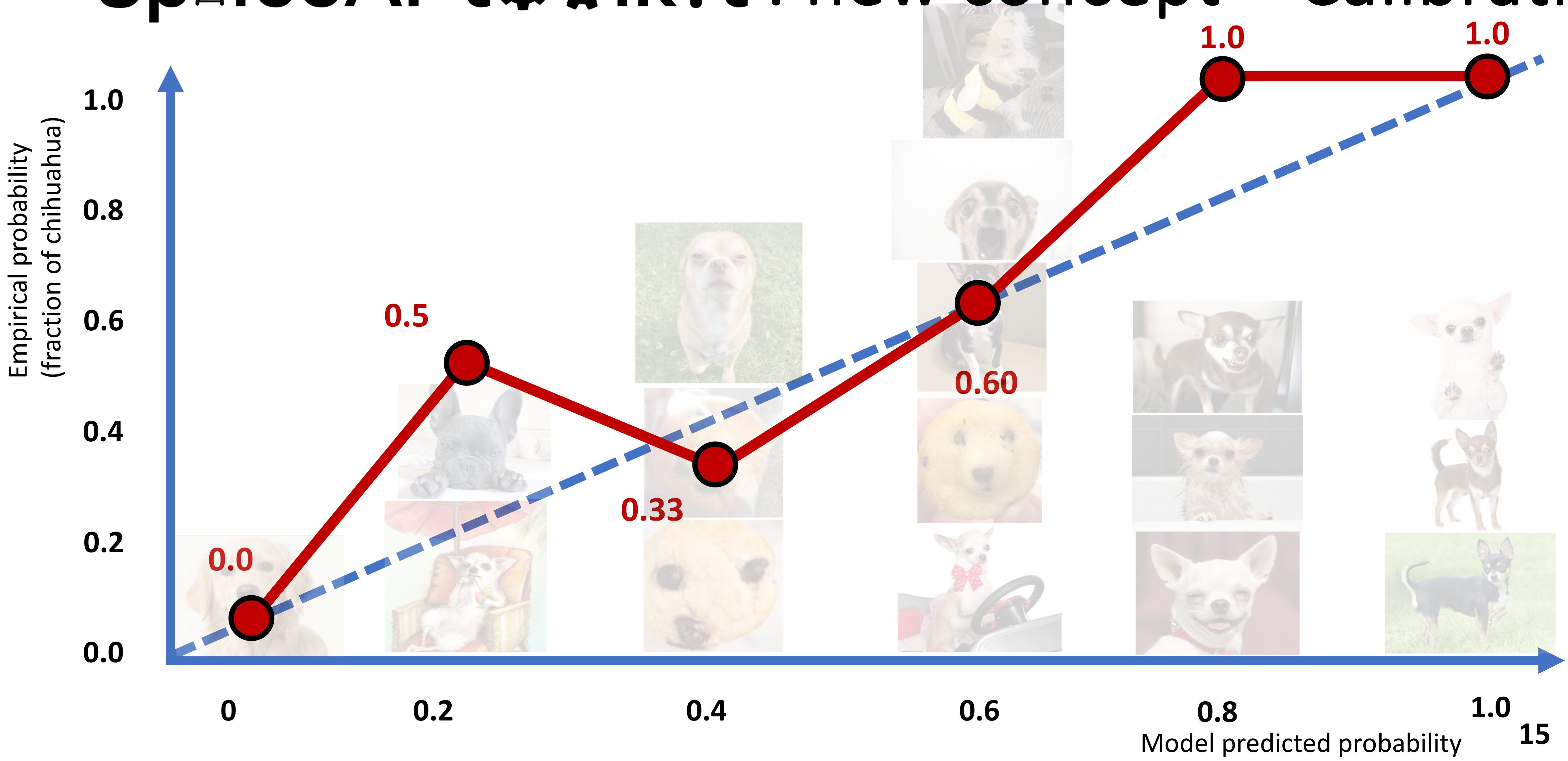
Binary classification example

SpliceAI-toolkit : new concept – Calibration



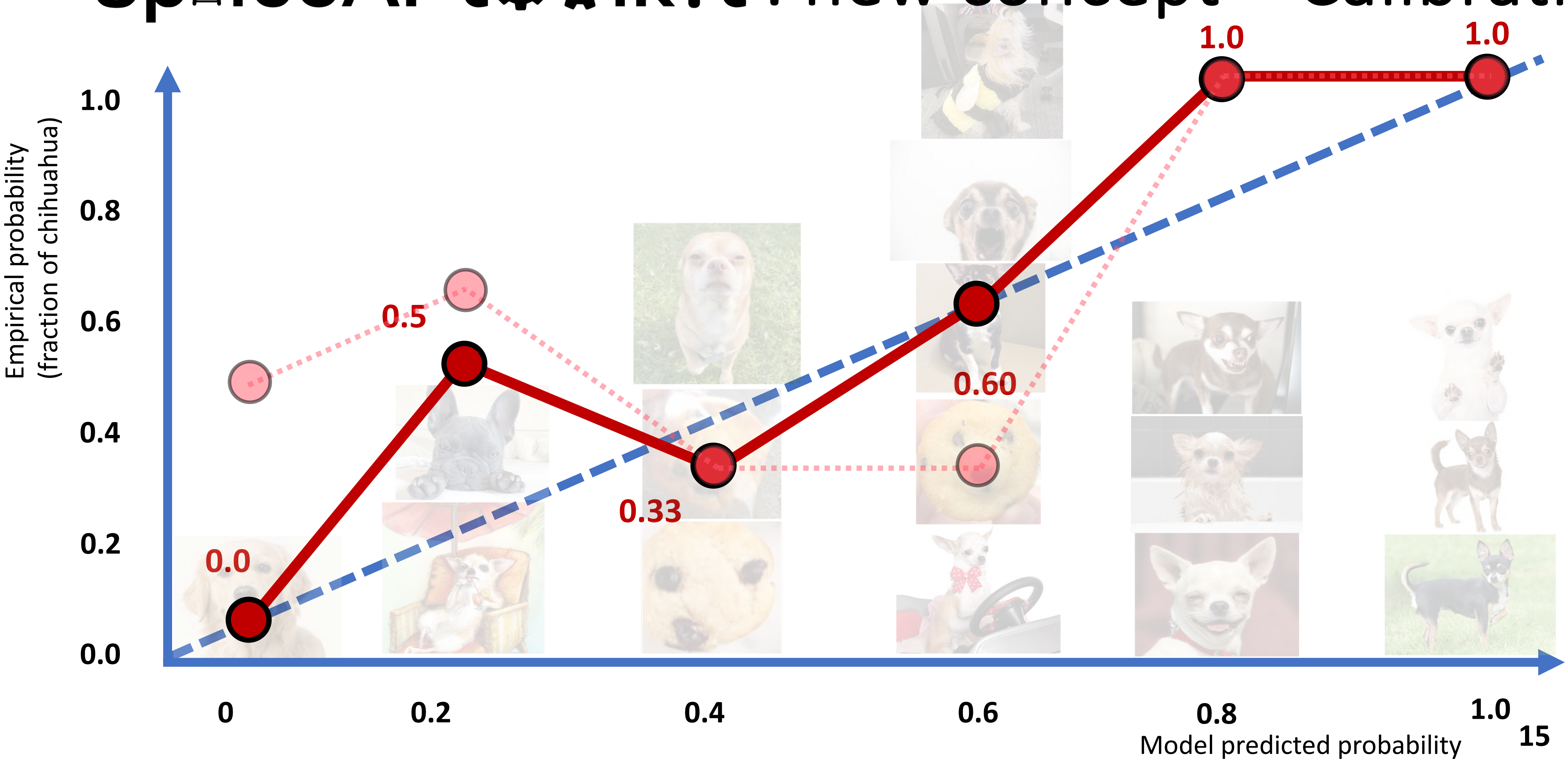
Binary classification example

SpliceAI-toolkit : new concept – Calibration



Binary classification example

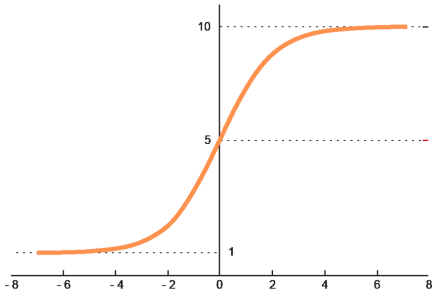
SpliceAI-toolkit : new concept – Calibration

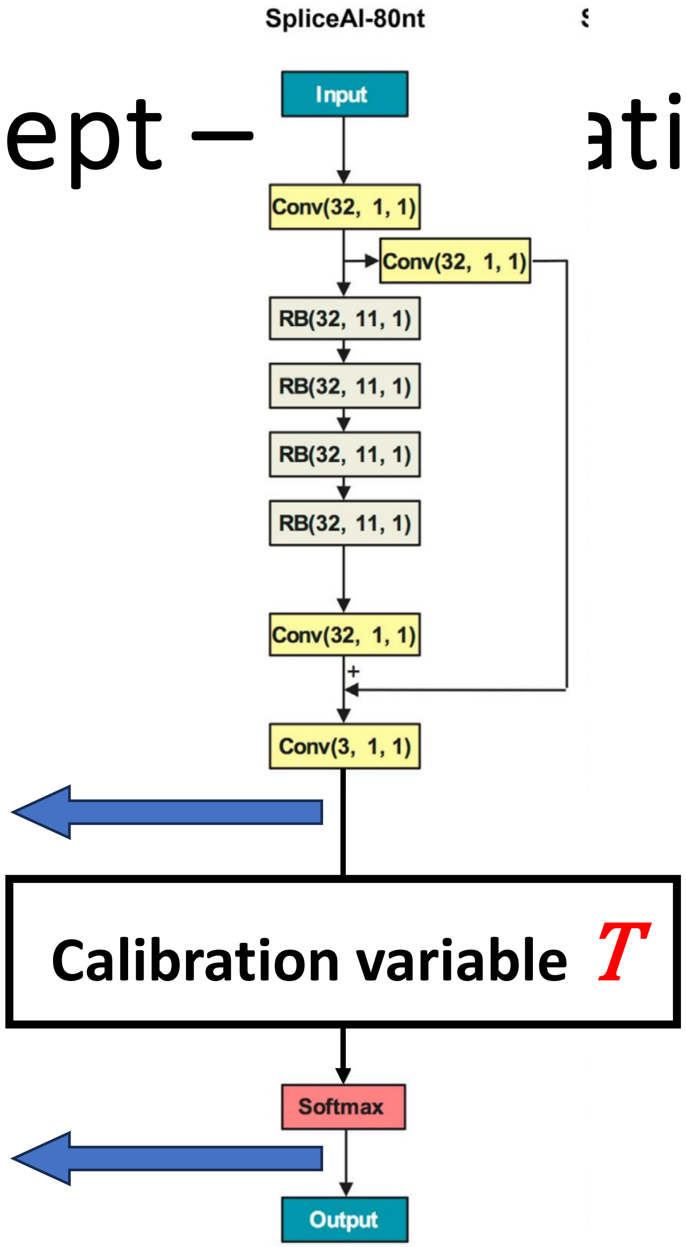


SpliceAI-toolkit : new concept – ation

- Platt scaling (Temperature scaling)

$$P_i = \frac{e^{\frac{y_i}{T}}}{\sum_{k=1}^{n} e^{\frac{y_k}{T}}}$$





Logits(y_i)

Probability (P_i)

$i \in \{Neither, Acceptor, Donor\}$

Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." Advances in large margin classifiers 10.3 (1999): 61-74.

SpliceAI-toolkit : new concept – Calibration

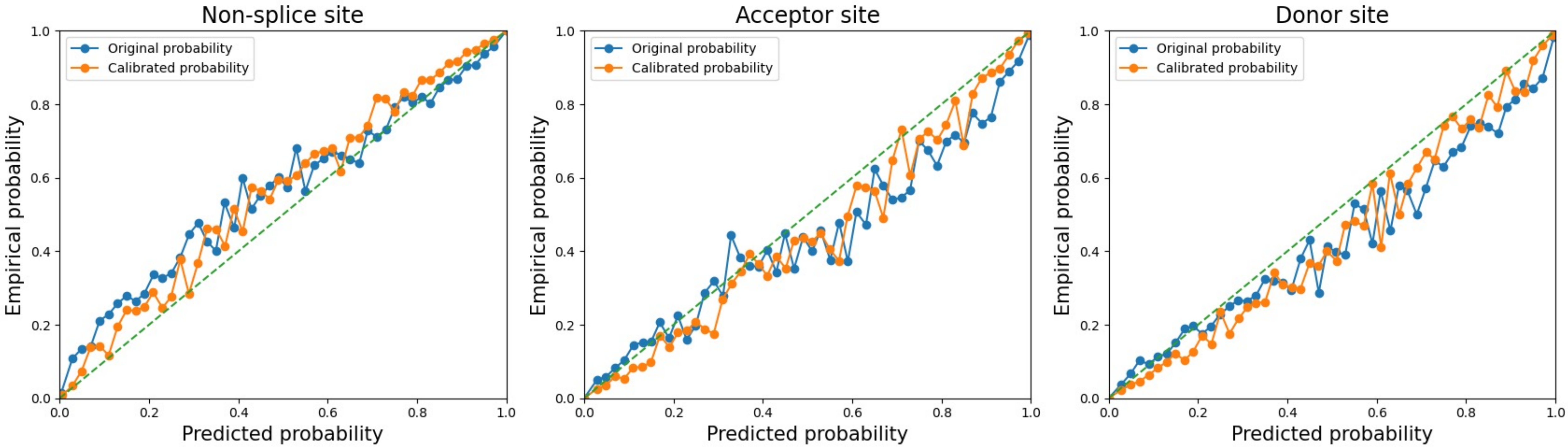
$$conf(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i \qquad acc(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} 1(\hat{y}_i = y_i)$$

Before calibration - **NLL** : 0.13310300, **ECE**: 0.00001282

Optimal temperature: 1.28049

After calibration - **NLL**: 0.11896934, **ECE** : 0.00000079

$$\mathcal{L}_{NLL} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i|x_i)) \qquad \mathcal{L}_{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$



Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

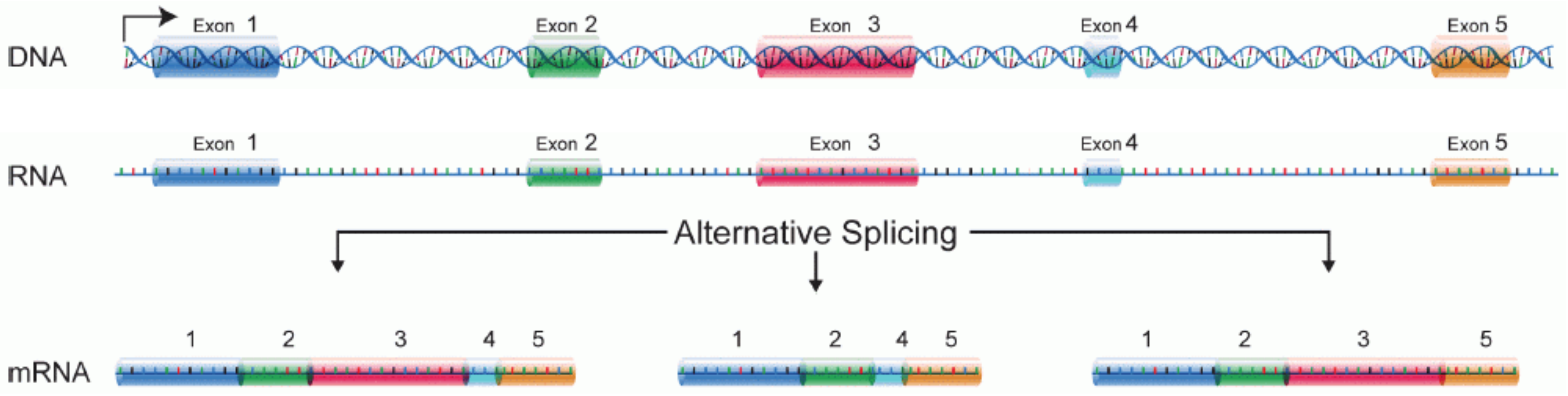
SpliceAI-toolkit : Summary

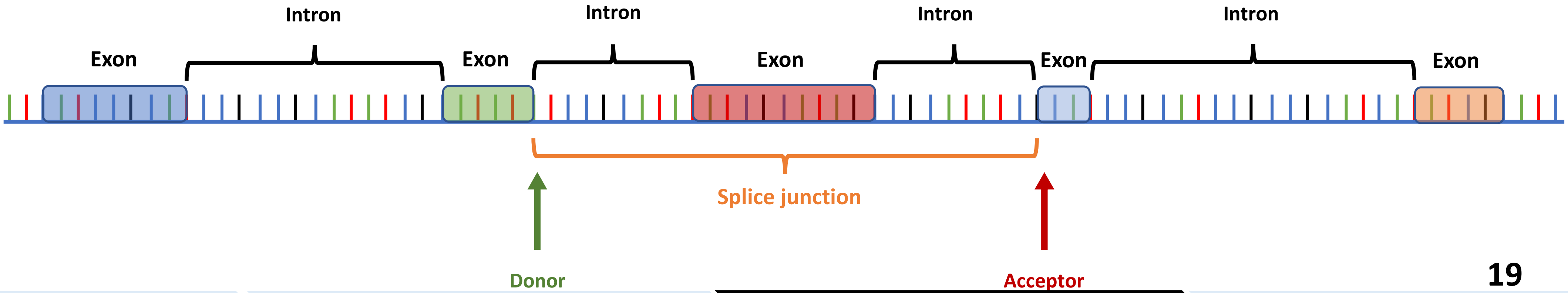
1. Data preprocessing: sliding window chunking
2. Easy-to-run framework to train your own SpliceAI
3. Pretrained SpliceAI-MANE
4. Pretrained SpliceAI on different species
5. Predict genetic variants' effect on splice sites
6. Model calibration: temperature scaling

Problem solved?

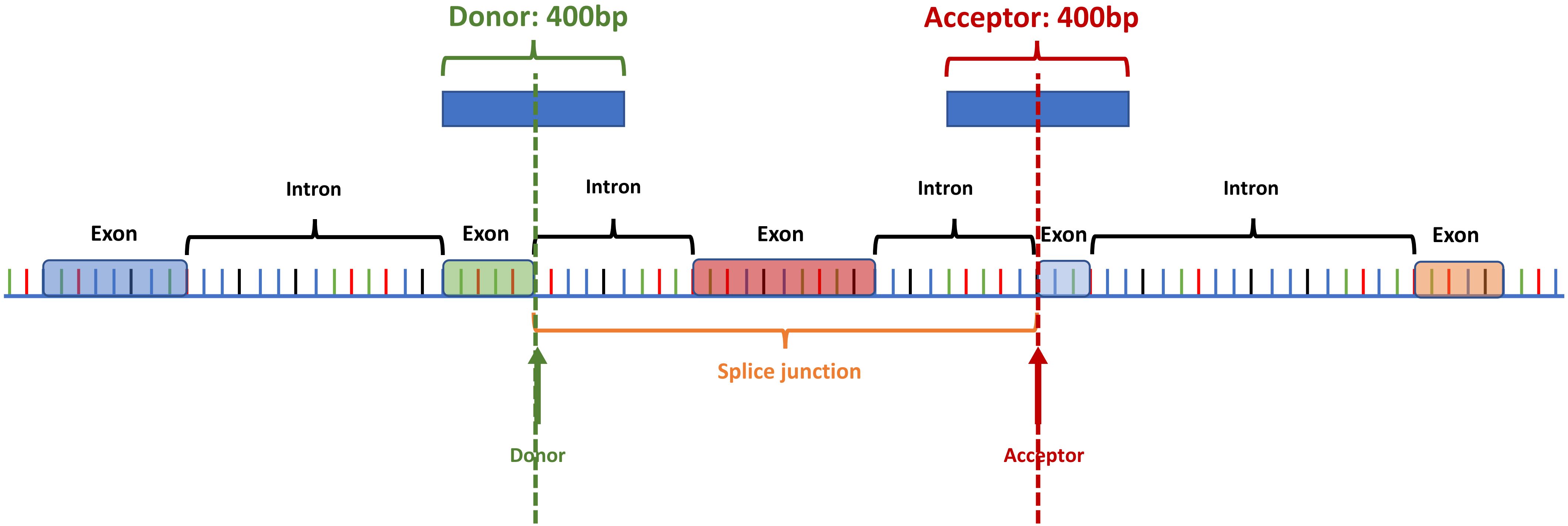
Chao, K. H., Mao, A., Liu, Anqi, Salzberg, S. L., & Pertea, M. (**2024**). SpliceAI-toolkit. Manuscript in preparation. 📕https://ccb.jhu.edu/spliceai-toolkit/ **(in preparation)**

🚨 Is canonical labelling approach correct?



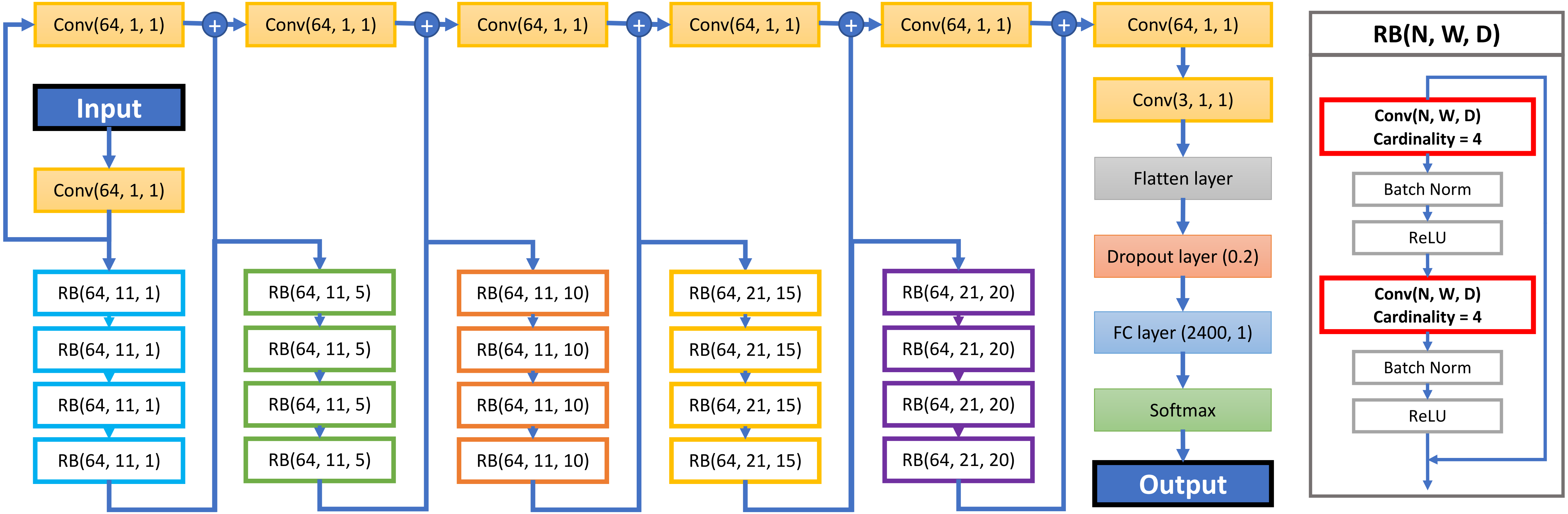


SPLAM : Data Processing



Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** 📕**https://ccb.jhu.edu/splam/**

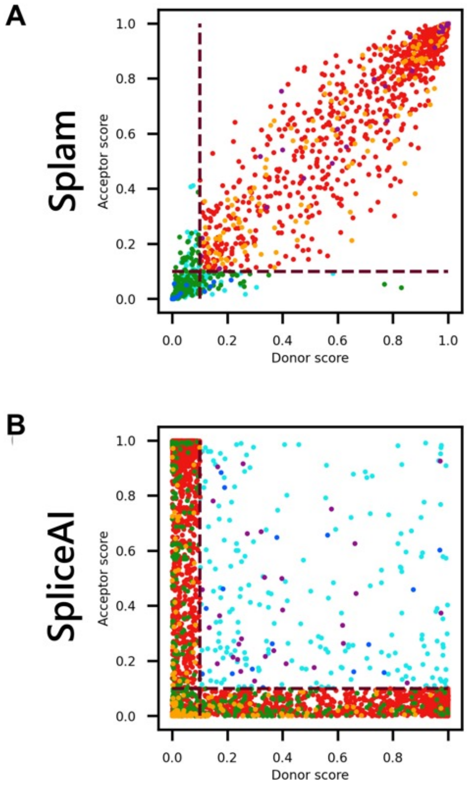
SPLAM : Splam Model Architecture



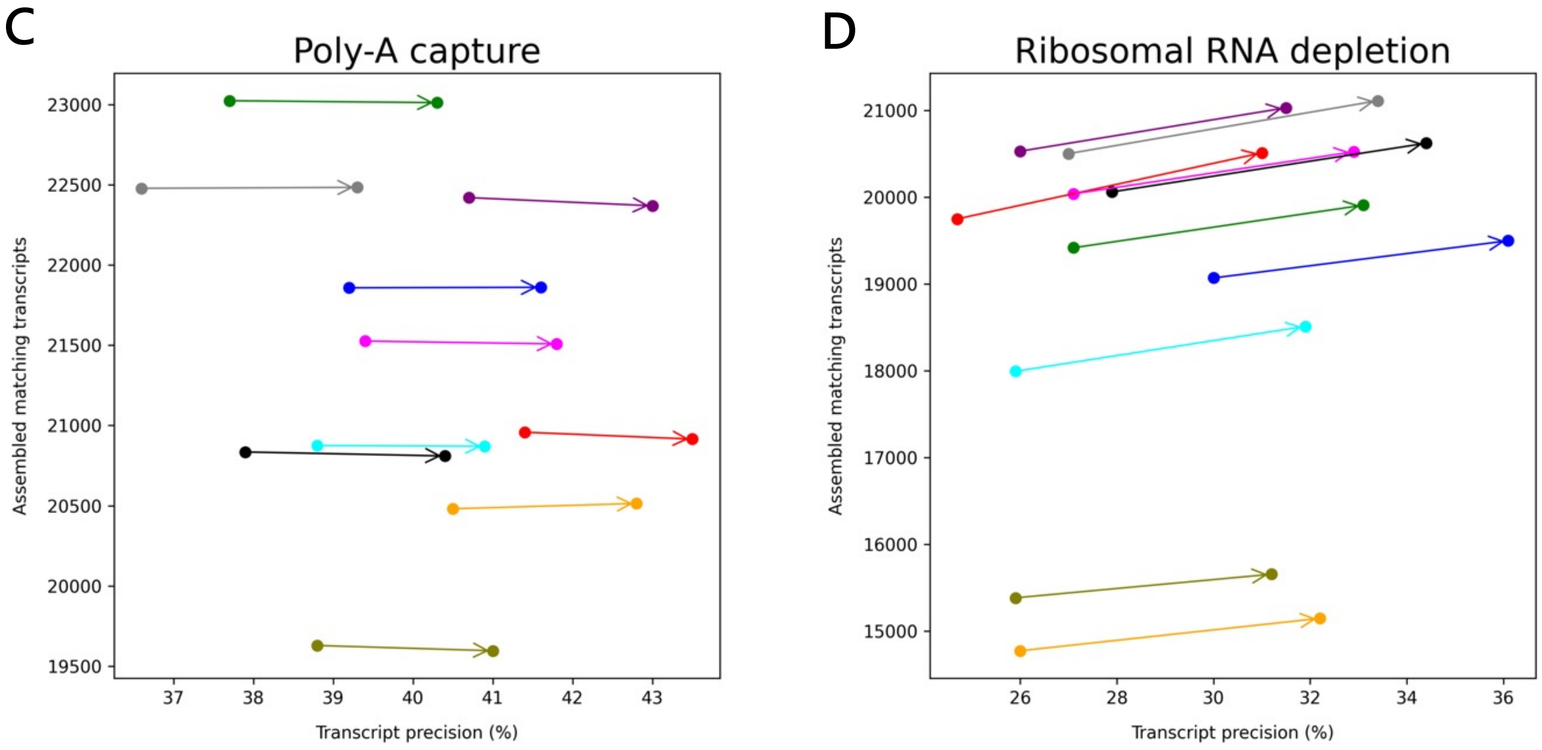
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** 📕**https://ccb.jhu.edu/splam/**

SPLAM : deep-learning splice site predictor

Score stability



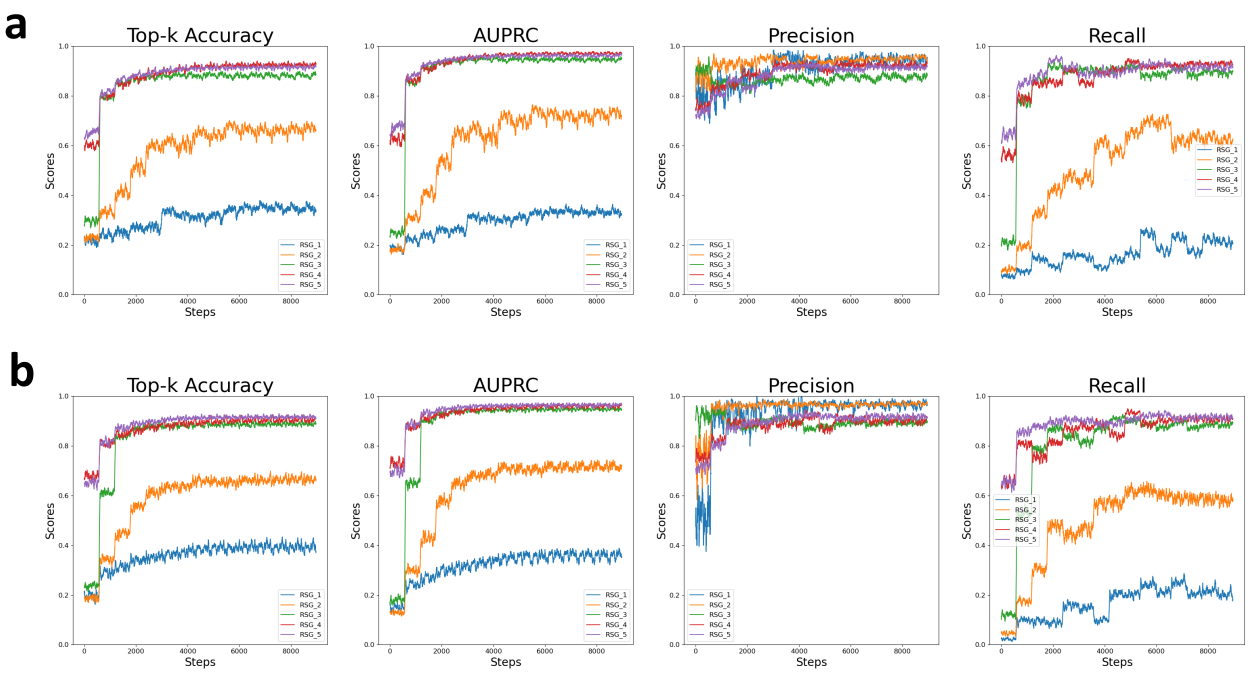
Transcriptome assembly improvement



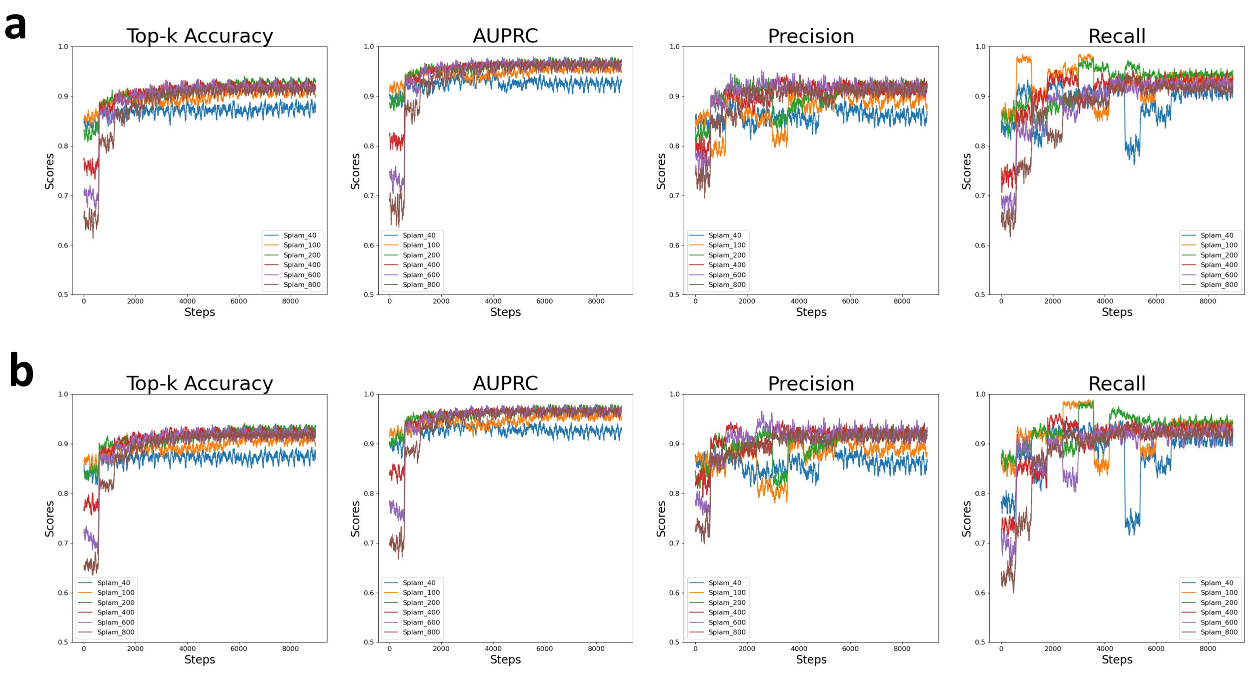
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** 📕**https://ccb.jhu.edu/splam/**

SPLAM : deep-learning splice site predictor

Interpretability: ablation study



Interpretability: input sequence



Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** **https://ccb.jhu.edu/splam/**

SPLAM : Summary

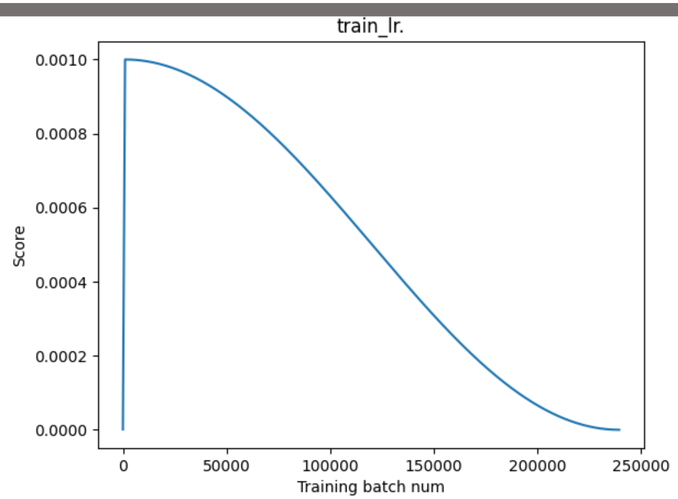
- Better than Splic
- Generalize to no

$$Loss_{CLE} = - \sum_{class \in \{donor,acceptor,neither\}} I_{class} \times \log(P_{class}) \quad (2)$$

$$Loss_{FL} = - \sum_{class \in \{donor,acceptor,neither\}} I_{class} \times (1 - P_{class})^{\gamma} \times \log(P_{class}) \quad (3)$$

Technical takeaways

- Focal loss improves cross entropy loss
- Learing rate warm up + sinusoidal decay
- Residual connection is powerful
- Grouped convolution helps (cardinality)



Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. **bioRxiv.** **https://ccb.jhu.edu/splam/**

Future sequence models in genomics?

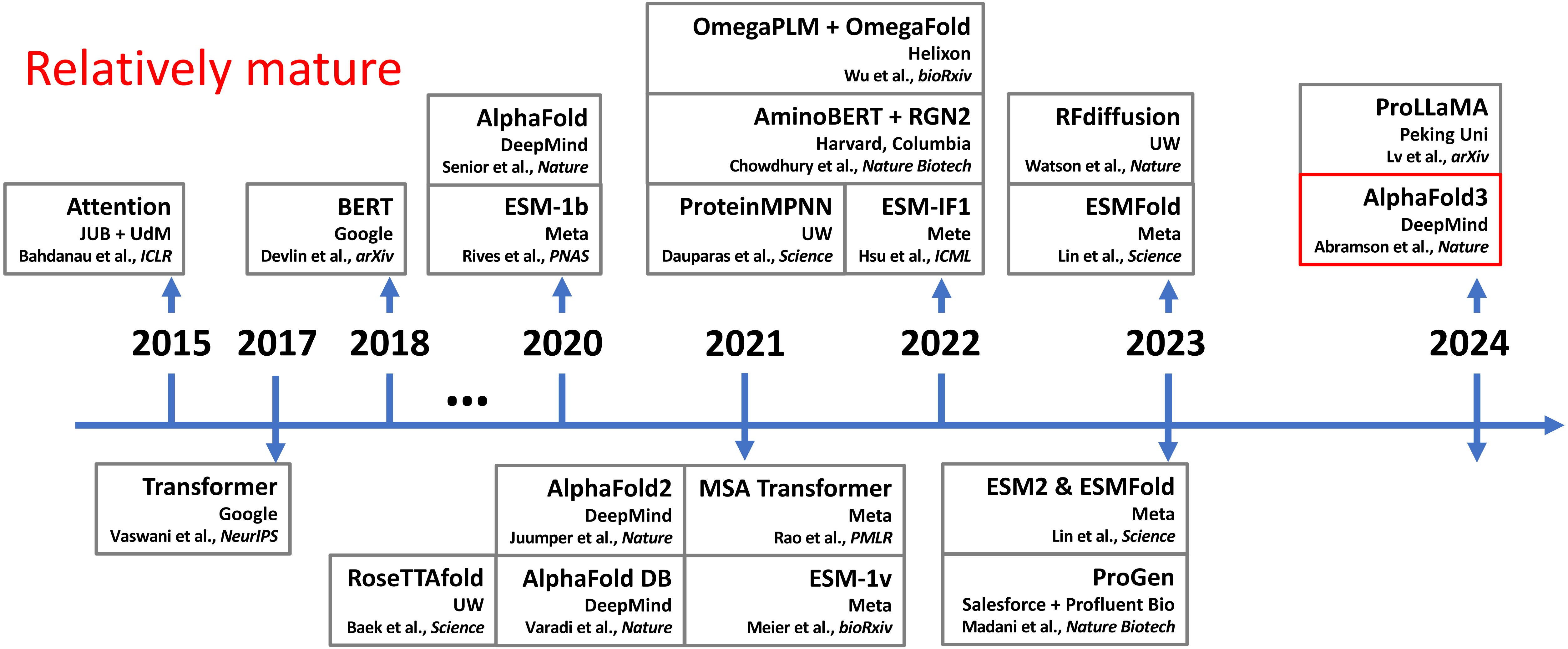
CNN or ?

Future sequence models in genomics?

CNN or **/and Transformer**?

Future? – Protein transformer-based models

Relatively mature



Attention
JUB + UdM
Bahdanau et al., *ICLR*

Transformer
Google
Vaswani et al., *NeurIPS*

BERT
Google
Devlin et al., *arXiv*

RoseTTAfold
UW
Baek et al., *Science*

AlphaFold
DeepMind
Senior et al., *Nature*

ESM-1b
Meta
Rives et al., *PNAS*

AlphaFold2
DeepMind
Juumper et al., *Nature*

AlphaFold DB
DeepMind
Varadi et al., *Nature*

MSA Transformer
Meta
Rao et al., *PMLR*

ESM-1v
Meta
Meier et al., *bioRxiv*

OmegaPLM + OmegaFold
Helixon
Wu et al., *bioRxiv*

AminoBERT + RGN2
Harvard, Columbia
Chowdhury et al., *Nature Biotech*

ProteinMPNN
UW
Dauparas et al., *Science*

ESM-IF1
Mete
Hsu et al., *ICML*

RFdiffusion
UW
Watson et al., *Nature*

ESMFold
Meta
Lin et al., *Science*

ESM2 & ESMFold
Meta
Lin et al., *Science*

ProGen
Salesforce + Profluent Bio
Madani et al., *Nature Biotech*

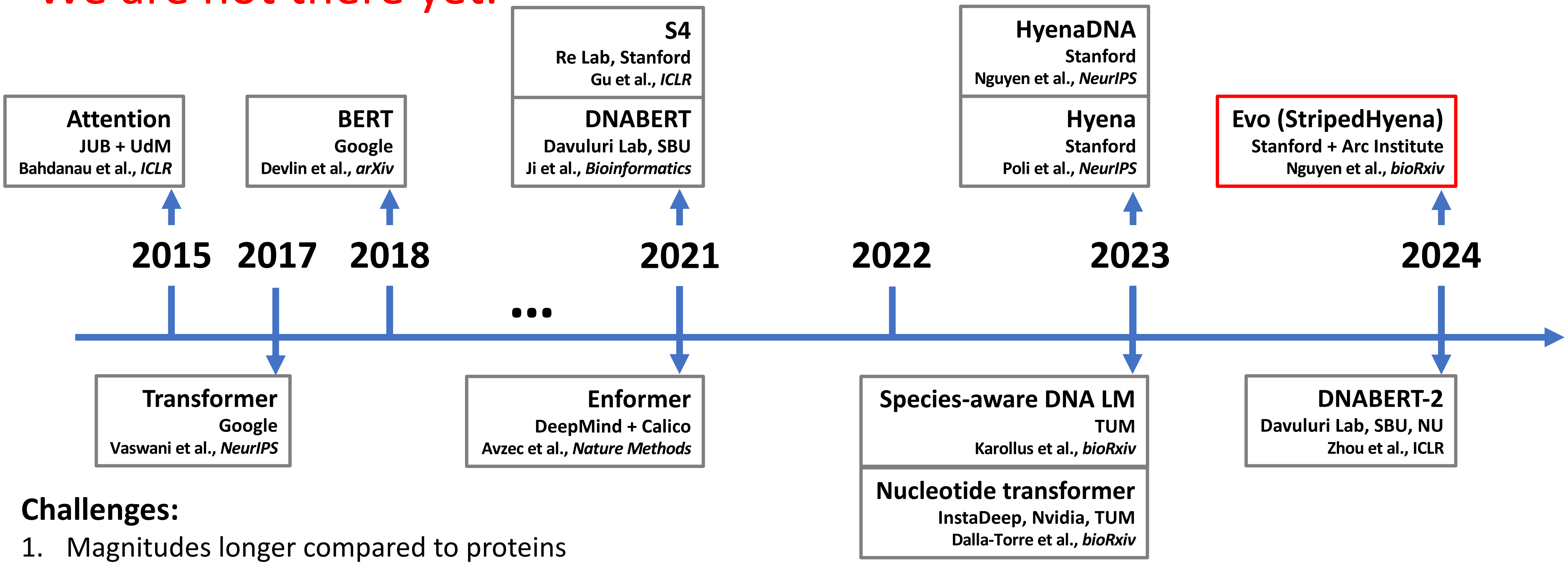
ProLLaMA
Peking Uni
Lv et al., *arXiv*

AlphaFold3
DeepMind
Abramson et al., *Nature*

DeepMind Calico Arc Institute InstaDeep™

Future? – DNA transformer-based models

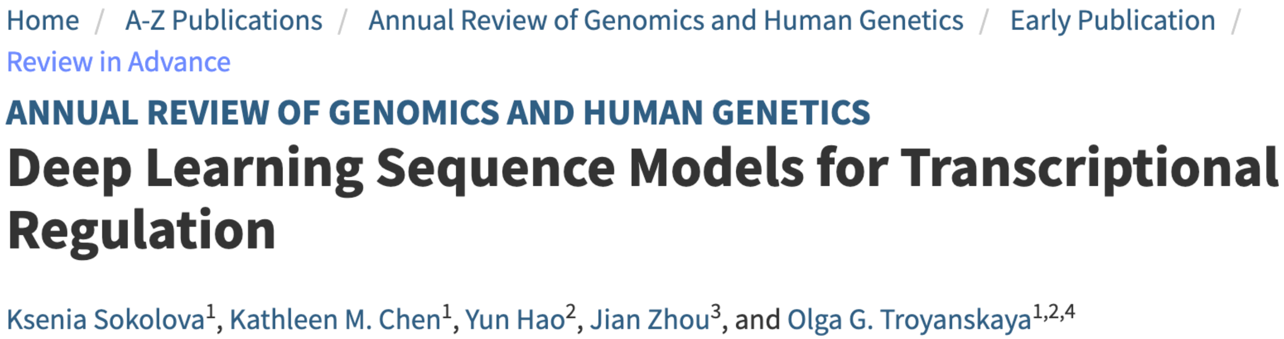
We are not there yet.



Challenges:

1. Magnitudes longer compared to proteins
2. Long-range dependencies & interactions spanning over 100k+ nt.

Application? – Genome annotation

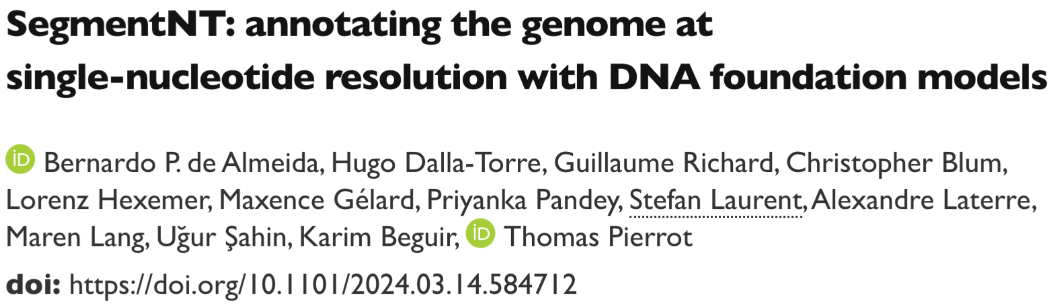


Home / A-Z Publications / Annual Review of Genomics and Human Genetics / Early Publication / Review in Advance

ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS

Deep Learning Sequence Models for Transcriptional Regulation

Ksenia Sokolova[1], Kathleen M. Chen[1], Yun Hao[2], Jian Zhou[3], and Olga G. Troyanskaya[1,2,4]



SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models

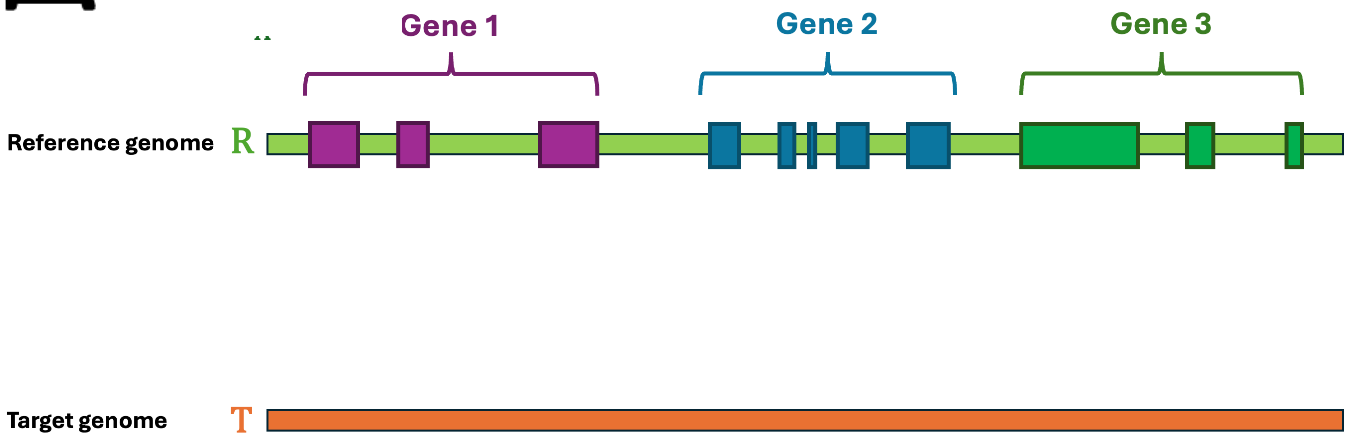
Bernardo P. de Almeida, Hugo Dalla-Torre, Guillaume Richard, Christopher Blum, Lorenz Hexemer, Maxence Gélard, Priyanka Pandey, Stefan Laurent, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, Thomas Pierrot

doi: https://doi.org/10.1101/2024.03.14.584712

LiftOn

ISMB 2024 · JULY · 12-16 ISCB MONTREAL CANADA

RECOMB-SEQ



github.com/Kuanhao-Chao/LiftOn

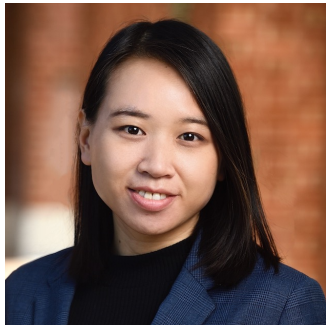
ccb.jhu.edu/lifton Preprint coming soon!

Acknowledge

khchao.com @KuanHaoChao Kuanhao-Chao

Steven Salzberg

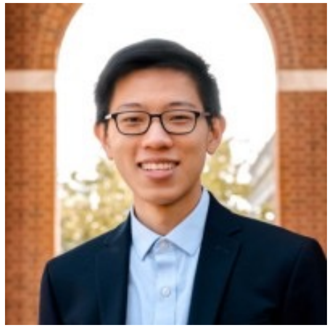
Mihaela Pertea


Anqi Liu

Alaina Shumate

Jakob Heinz

Celine Hoh


Alan Mao

- All members in Salzberg lab, Pertea lab
- All friends at Malone & CCB
- All friends at JHU Computational Biology Group

> *"If you think of mathematics as the perfect description language for physics, then AI might be the perfect one for biology."*

Demis Hassabis, CEO of DeepMind, 2022