JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING
Department of Computer Science

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Computational methods to improve genome annotation, splice site prediction, and gene expression prediction

Kuan-Hao Chao

2024.08.06

khchao.com

@KuanHaoChao

Kuanhao-Chao

Part I

Genome Annotation

- LiftOn: genome annotation lift-over
- Applications: mapping genes between two genomes

Genome annotation

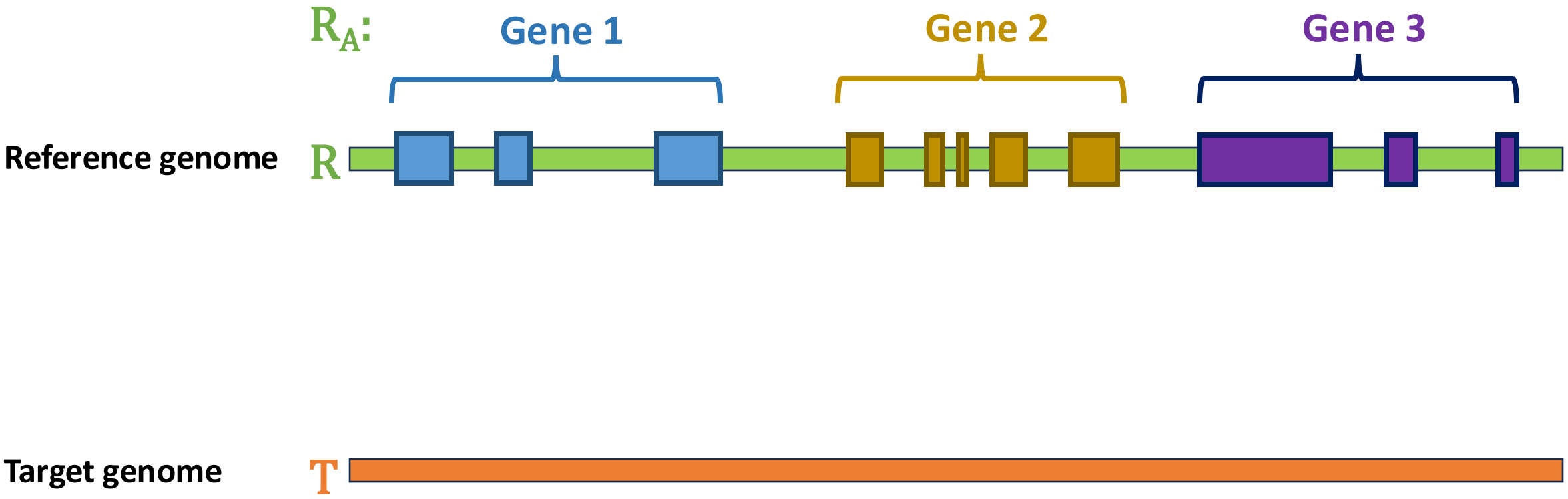
Genome (FASTA)

```
CAGCCCCCGGAGACTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAGTCACATGCGACCGGTgactccctgtccccacccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccagggtgtGCAGCACCACTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCCTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAAACC
ATAGTGCCCAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTCACAGCAGGCATCATCAGTA
```

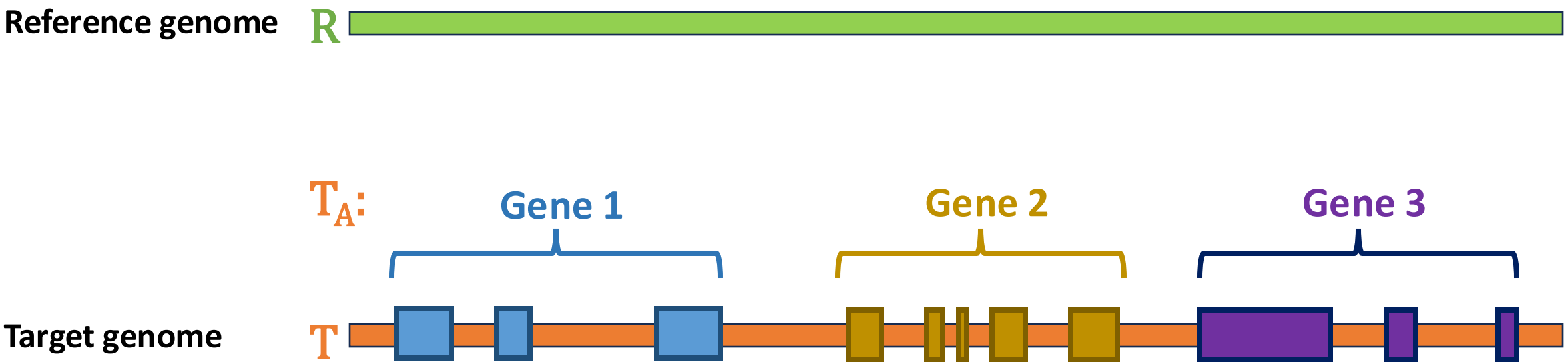
Annotation (GFF / GTF)

```
chr1  BestRefSeq  gene  450740 451678 .  -  .  ID=gene-OR4F29;
chr1  BestRefSeq  mRNA  450740 451678 .  -  .  ID=rna-NM_001005221.2;Parent=gene-OR4F29;
chr1  BestRefSeq  exon  450740 451678 .  -  .  ID=exon-NM_001005221.2-1;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  exon  452658 453675 .  -  .  ID=exon-NM_001005221.2-2;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  exon  454672 459678 .  -  .  ID=exon-NM_001005221.2-3;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  CDS  450740 451678 .  -  0  ID=cds-NP_001005221.2-1;Parent=rna-NM_001005221.2;
chr1  BestRefSeq  CDS  452658 453675 .  -  0  ID=cds-NP_001005221.2-2;Parent=rna-NM_001005221.2;
```

Lift-over Problem Definition:



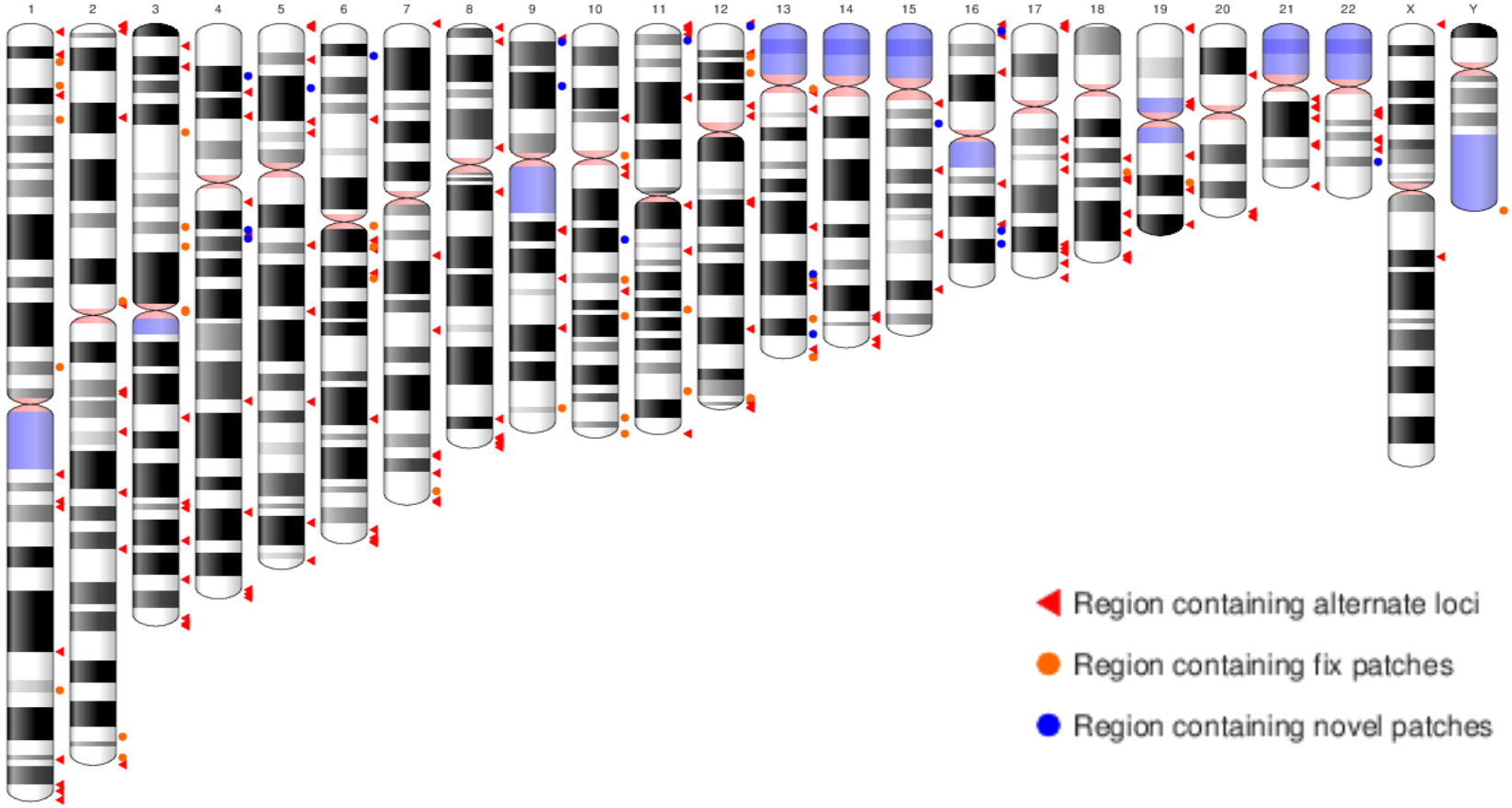
Lift-over Problem Definition:



https://www.sanger.ac.uk/data/genome-reference-consortium/

Application: GRCh38 to T2T-CHM13 lift-over

“**Finished**” the human genome project




nature

Explore content · About the journal · Publish with us

nature › article

Original Article | Published: 01 February 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Science

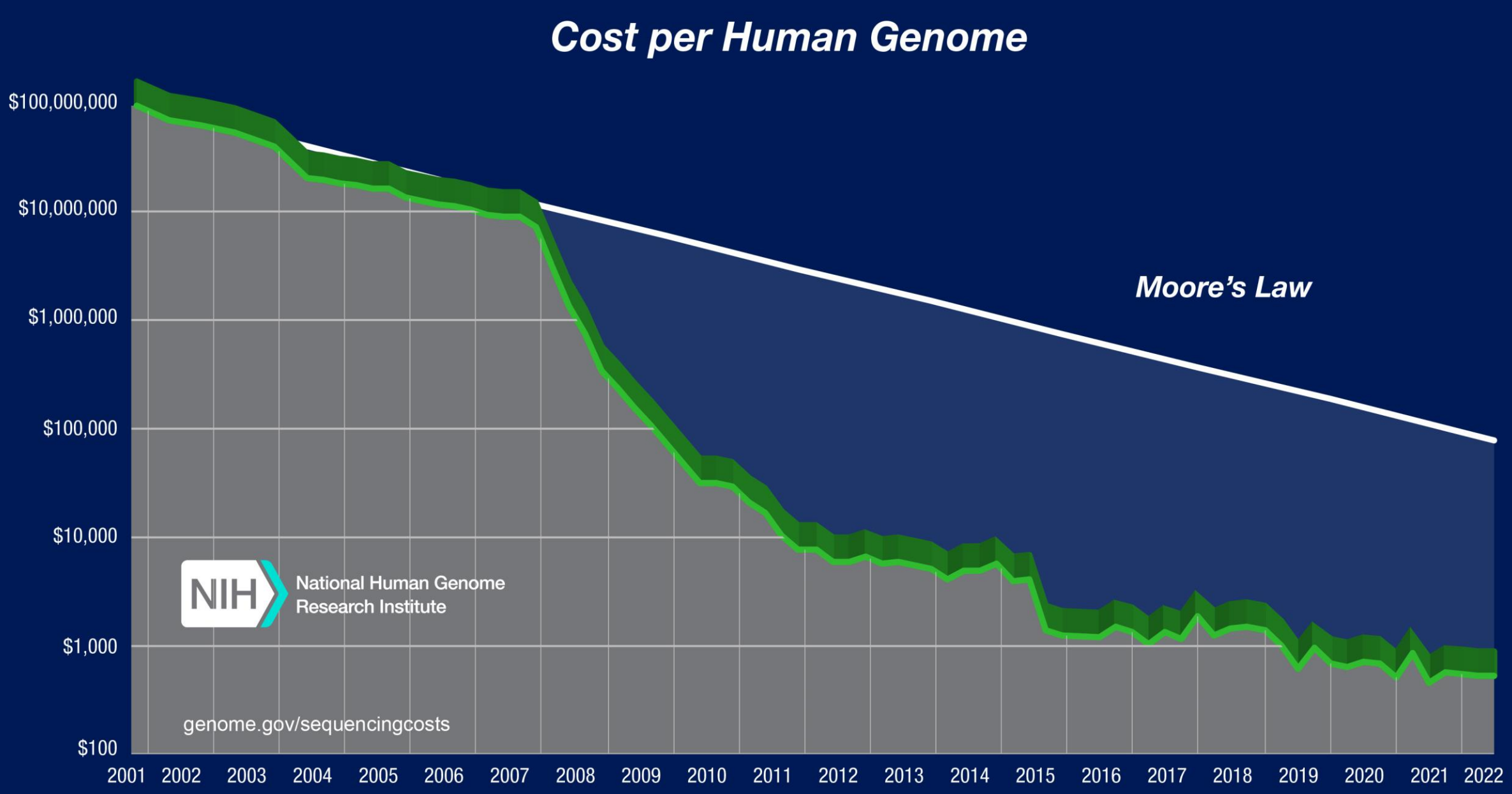
Current Issue · First release papers

HOME › SCIENCE › VOL. 291, NO. 5507 › THE SEQUENCE OF THE HUMAN GENOME

SPECIAL REVIEWS

The Sequence of the Human Genome

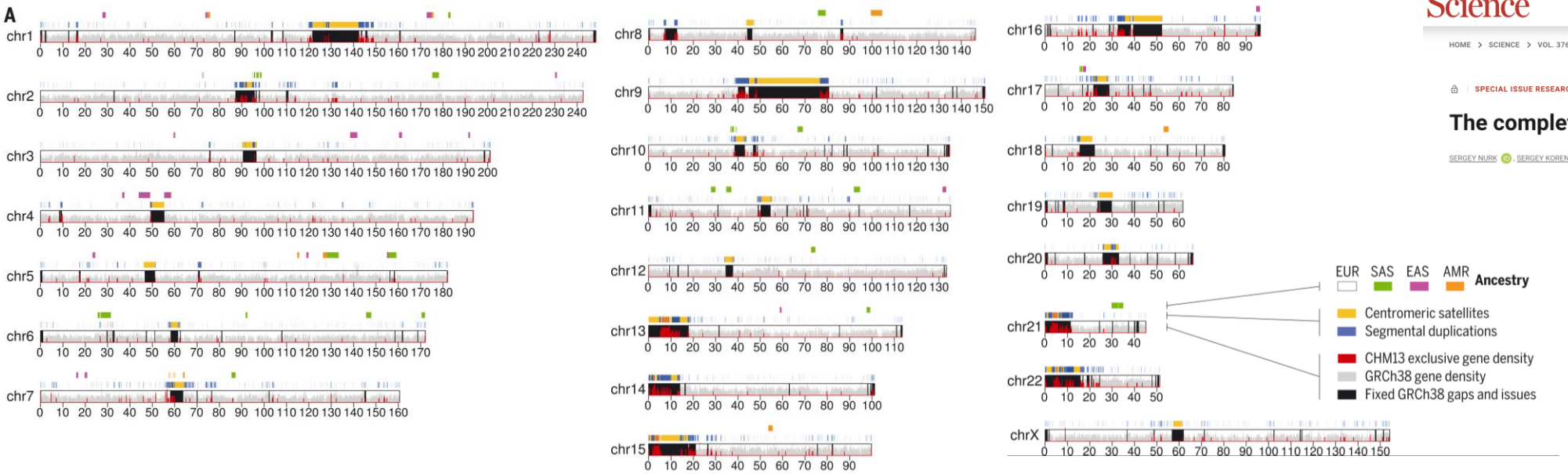
J. CRAIG VENTER, MARK D. ADAMS, EUGENE W. MYERS, PETER W. LI, [...], AND XIAOHONG ZHU +269 authors Authors Info & Affiliations


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

https://youtu.be/MbYvTyIMc84?si=YjijRyLwq31MY5HM

Application: GRCh38 to T2T-CHM13 lift-over

- 238 Mbp added and corrected
- 180 Mbp of centromeric satellites
- 68 Mbp of segmental duplications
- 10 Mbp of rDNAs
- 182 Mbp of entirely novel sequence
- 1956 novel genes including 99 protein-coding



If you were to use a CHM13 annotation …

Which lift-over tool to use?

Telomere-to-Telomere (T2T) consortium slack channel

Giulio Formenti 3:44 PM
if I was to use an annotation for CHM13, which would it be?

(gene annotation)

Arang Rhie 4:11 PM
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.1.gff3.gz or https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.1.bb

Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Article Navigation

JOURNAL ARTICLE

Liftoff: accurate mapping of gene annotations FREE

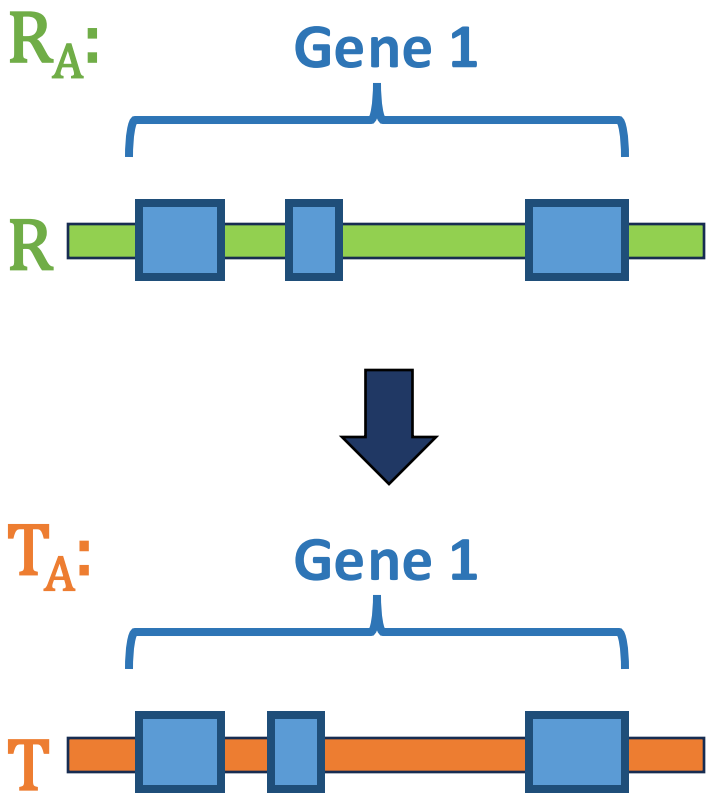
Alaina Shumate ✉, Steven L Salzberg

Bioinformatics, Volume 37, Issue 12, June 2021, Pages 1639–1643,
https://doi.org/10.1093/bioinformatics/btaa1016

Published: 09 May 2021 **Article history** ▾

~ 390 citation

Lift-over problem, what methods are available?



Chain file

UCSU liftOver

Kuhn et al.

CrossMap

Zhao et al.

DNA-based

Liftoff

github.com/agshumate/Liftoff

Shumate et al.

minimap2

github.com/lh3/minimap2

Li

Protein-based

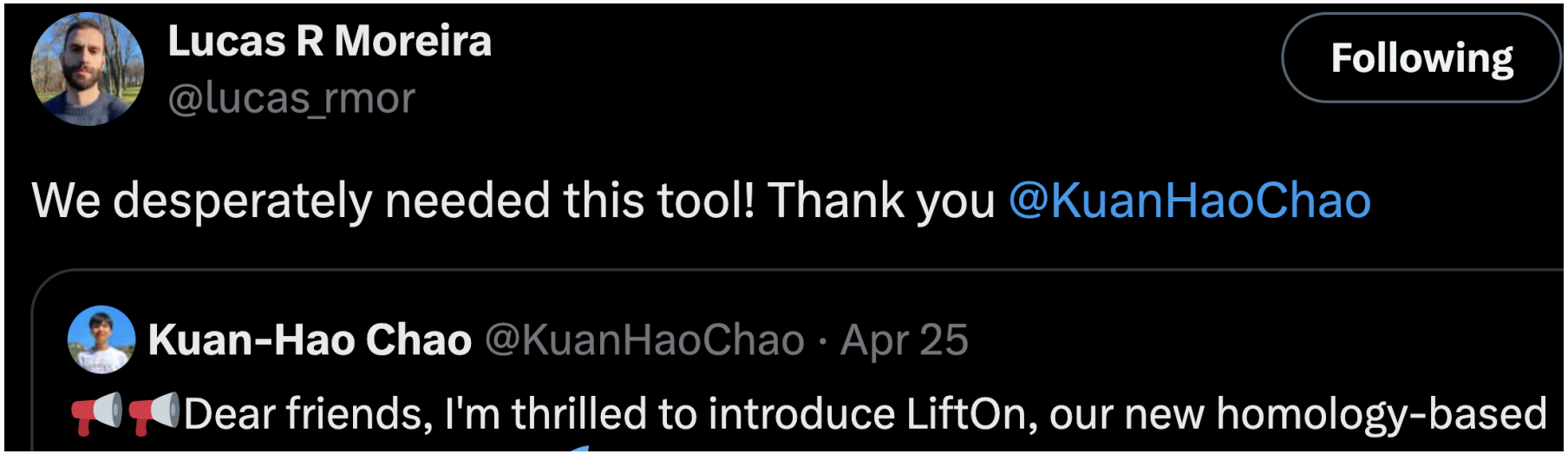
miniprot

github.com/lh3/miniprot

Li

Methods – LiftOn

- How can we do better?
  - Combining DNA- and protein-based alignments!

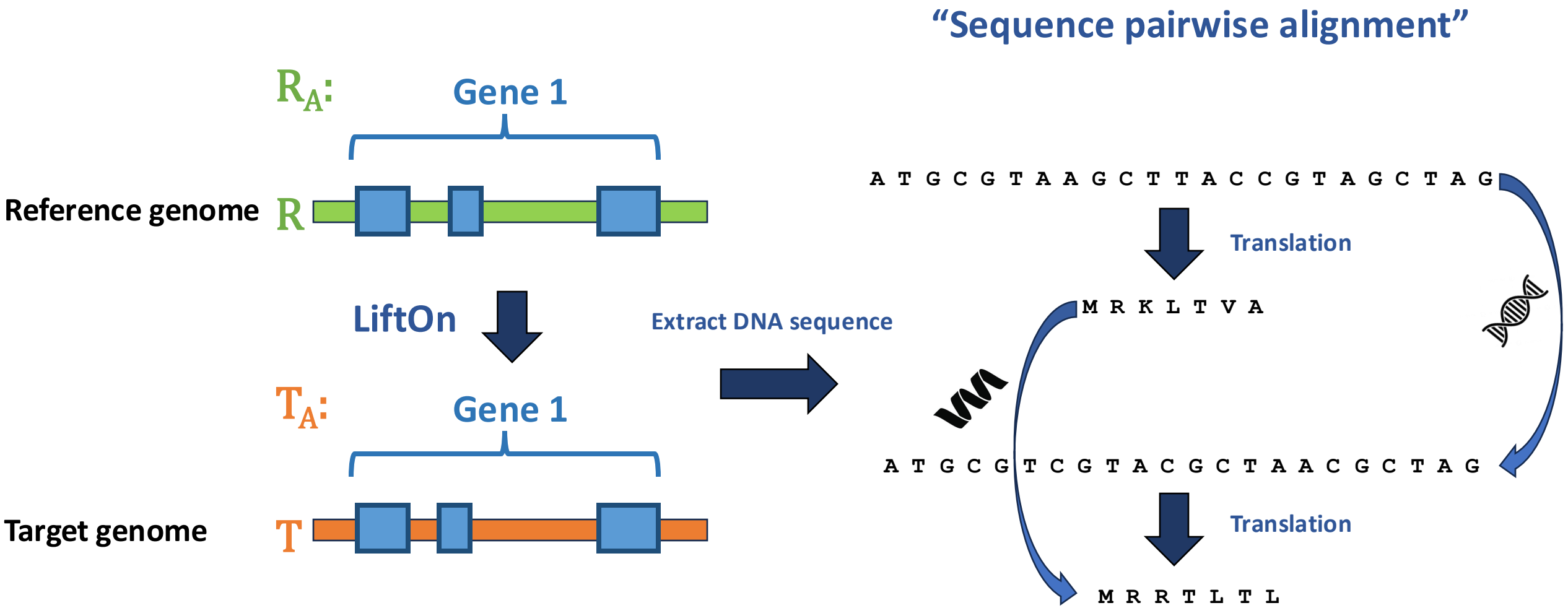


LiftOn

Results

- outperforms state-of-the-art DNA- and protein-based liftover methods
- improves the annotation of protein-coding genes in T2T-CHM13 genome
- Improves annotation lift-over between distant species, such as mouse and rat

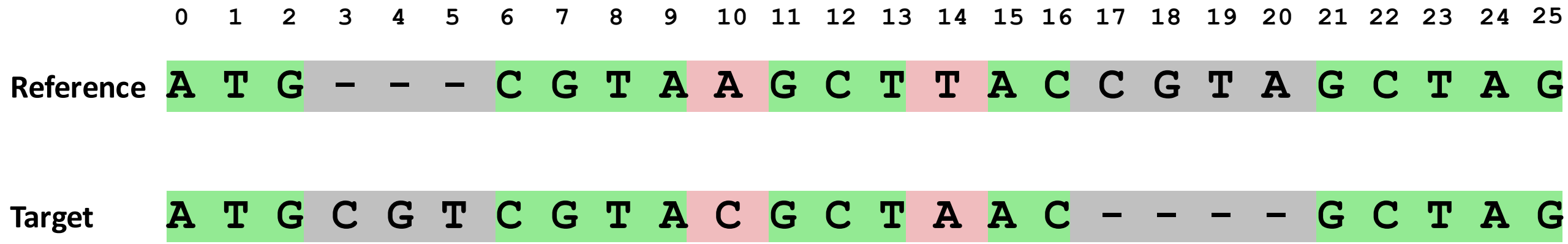
Evaluation Metrics



https://lh3.github.io/2018/11/25/on-the-definition-of-sequence-identity

Evaluation Metrics

DNA sequence identity



$$\frac{\#Matched\_nucleotide}{\#alignment\ column} = \frac{17}{26} = 65.4\%$$

https://lh3.github.io/2018/11/25/on-the-definition-of-sequence-identity

Evaluation Metrics

Protein sequence identity

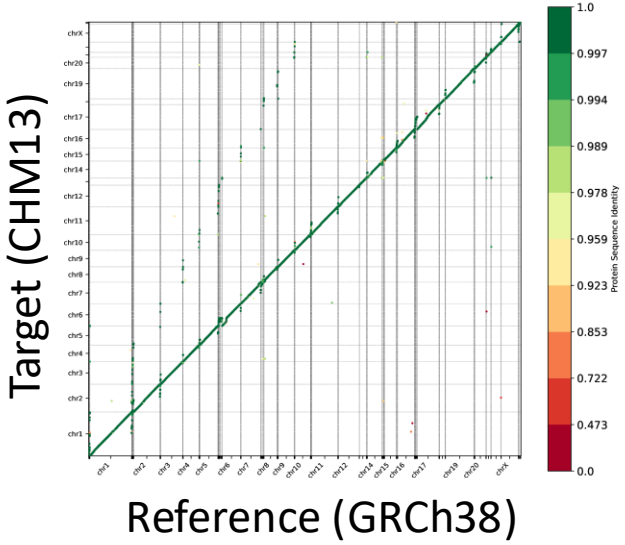
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Reference	M	G	L	V	-	-	-	-	R	W	S	Y	K	K	N	P	T	A	F	E	H	I	I	C	D	*
Target	M	G	L	V	R	W	S	S	R	W	S	Y	Q	K	N	P	T	A	-	-	H	I	-	C	D	*

$$\frac{\#Matched\_AA}{\#alignment\ column\ -\ \#gaps\ in\ reference\ protein} = \frac{18}{26-4} = 81.8\%$$

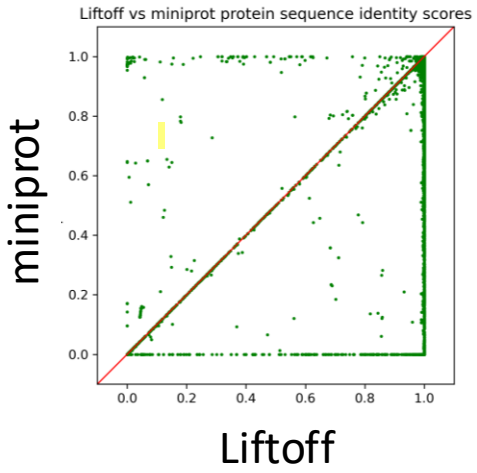
“Do not penalize longer proteins”

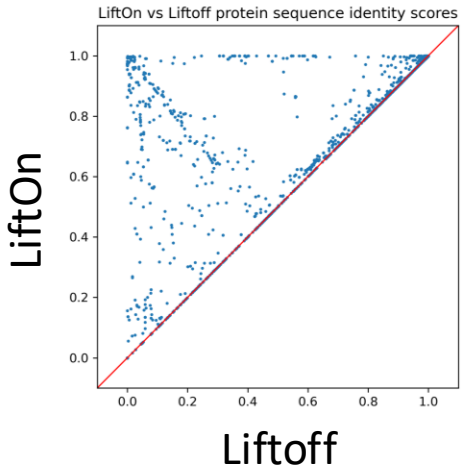
Result 1: improves DNA & protein-based lift-over

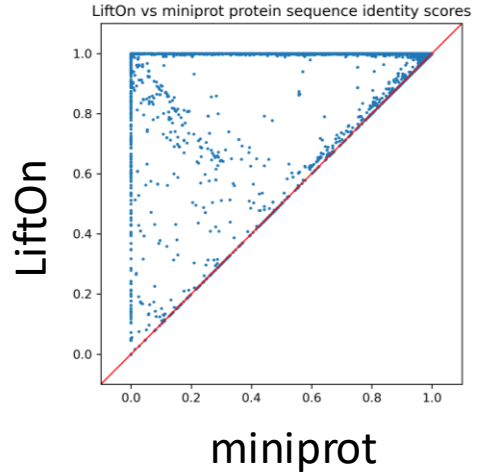
Map RefSeq v220 from GRCh38 -> CHM13V2.0

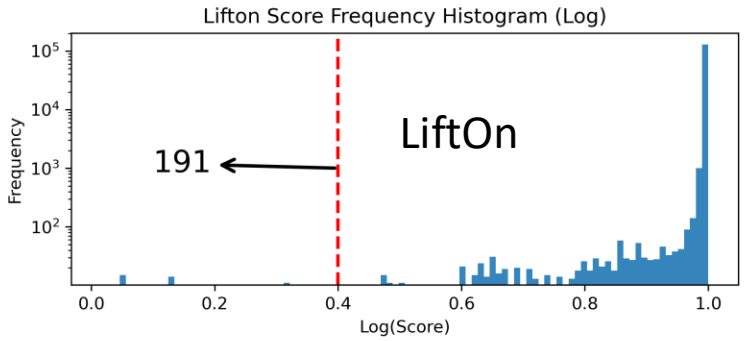


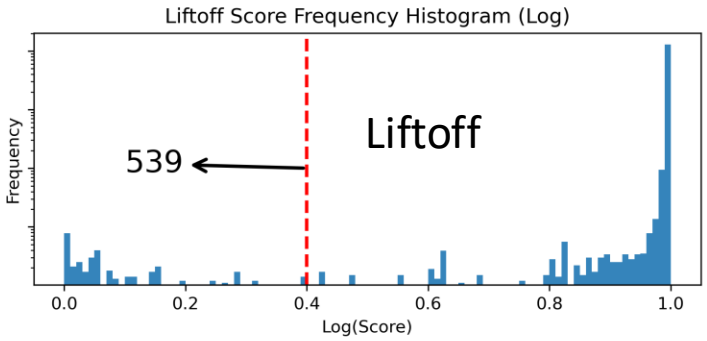
Compressed-gap protein sequence identity

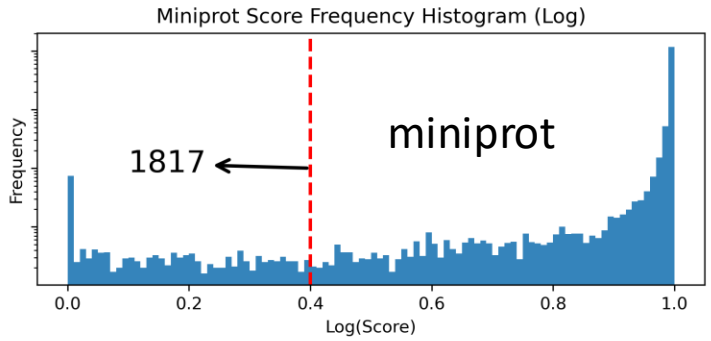








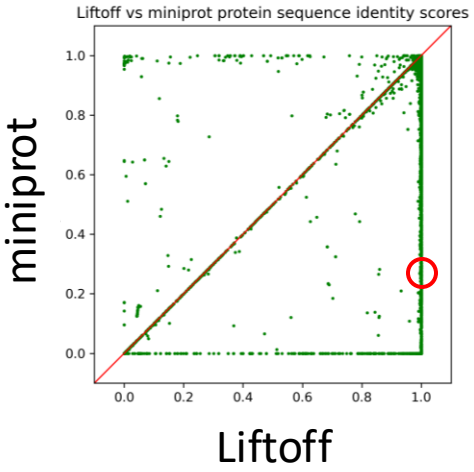


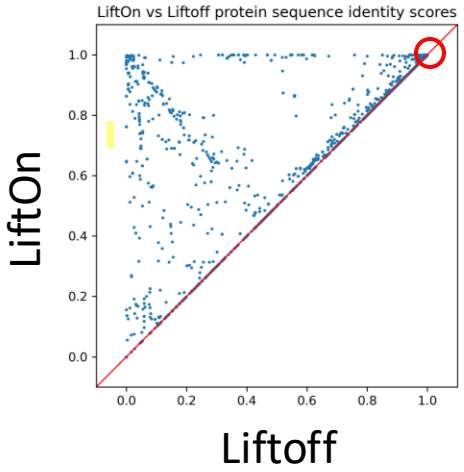


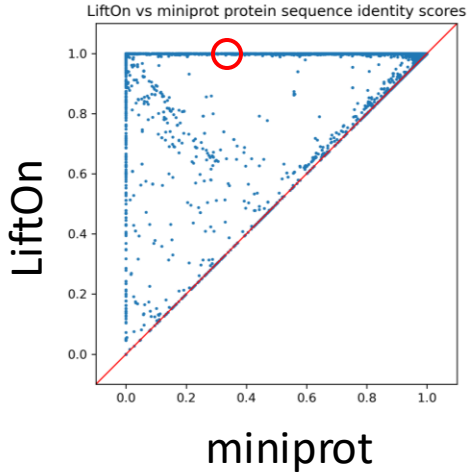
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

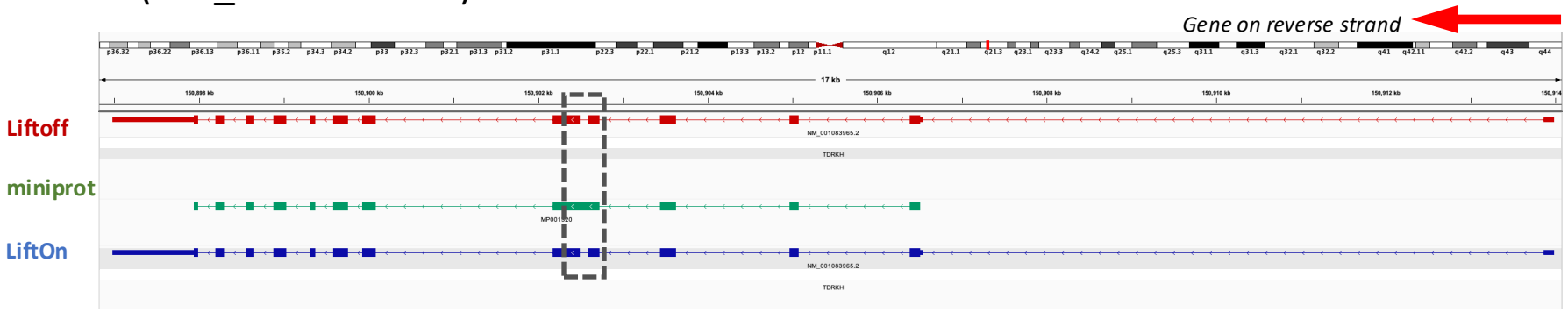
Compressed-gap protein sequence identity







TDRKH (NM_001083965.2) chr1:150896981 - 150913985

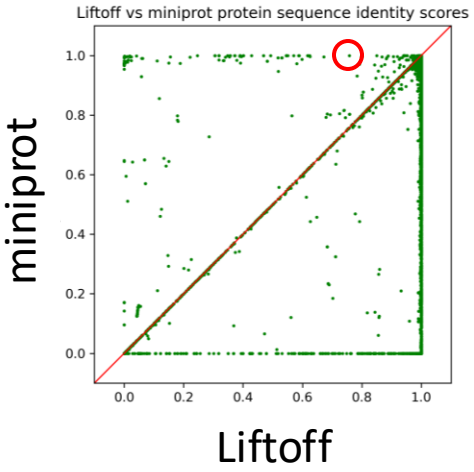


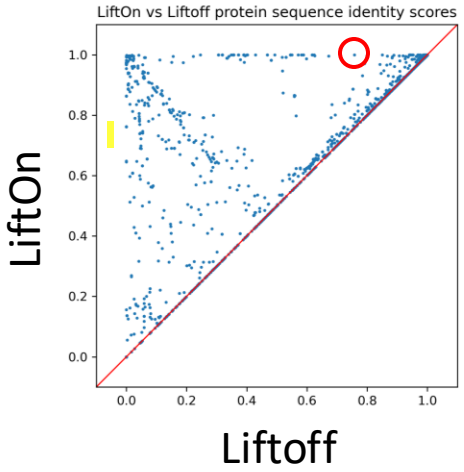
	Protein identity
Liftoff	100%
miniprot	31.48%
LiftOn	100%

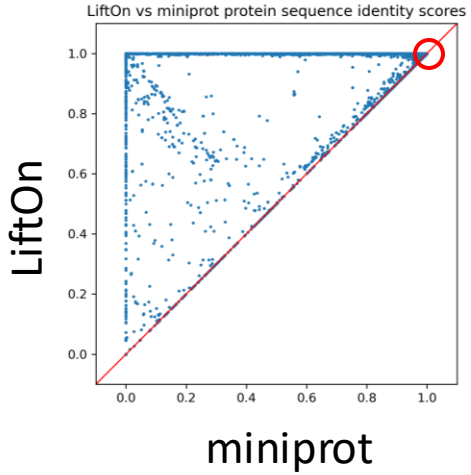
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

Compressed-gap protein sequence identity







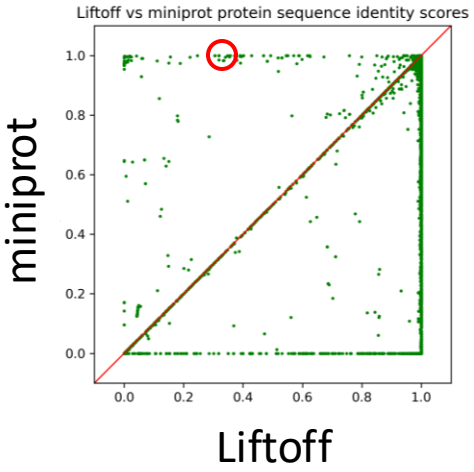
SLC22A31 (NM_001384763.1) chr16:95276205 - 95280662

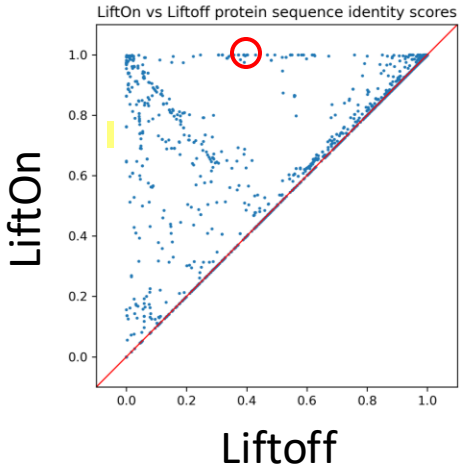
	Protein identity
Liftoff	76.70%
miniprot	99.55%
LiftOn	99.55%

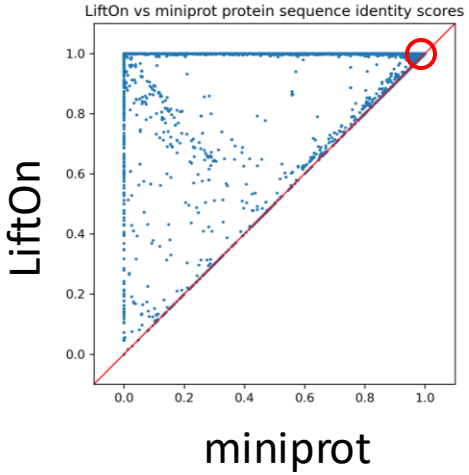
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

Compressed-gap protein sequence identity

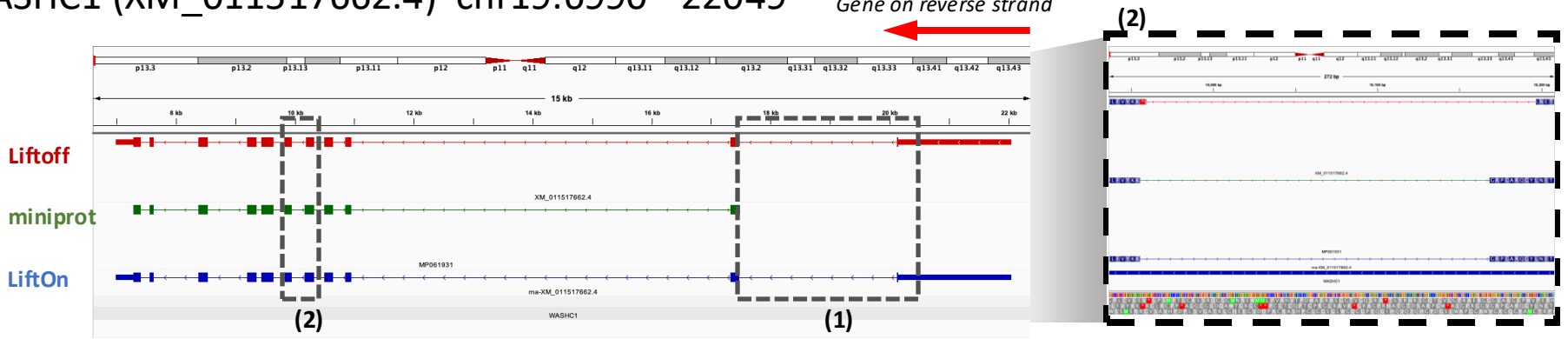






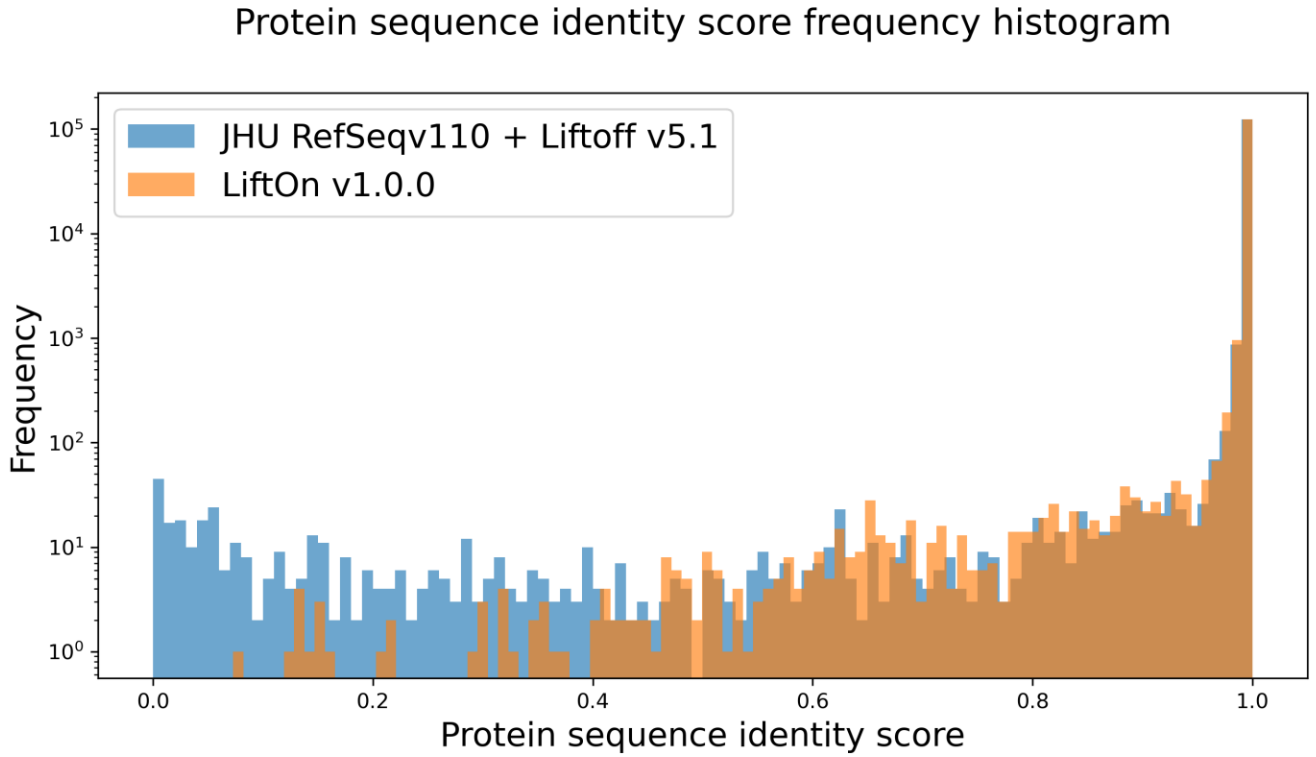
WASHC1 (XM_011517662.4) chr19:6990 - 22049

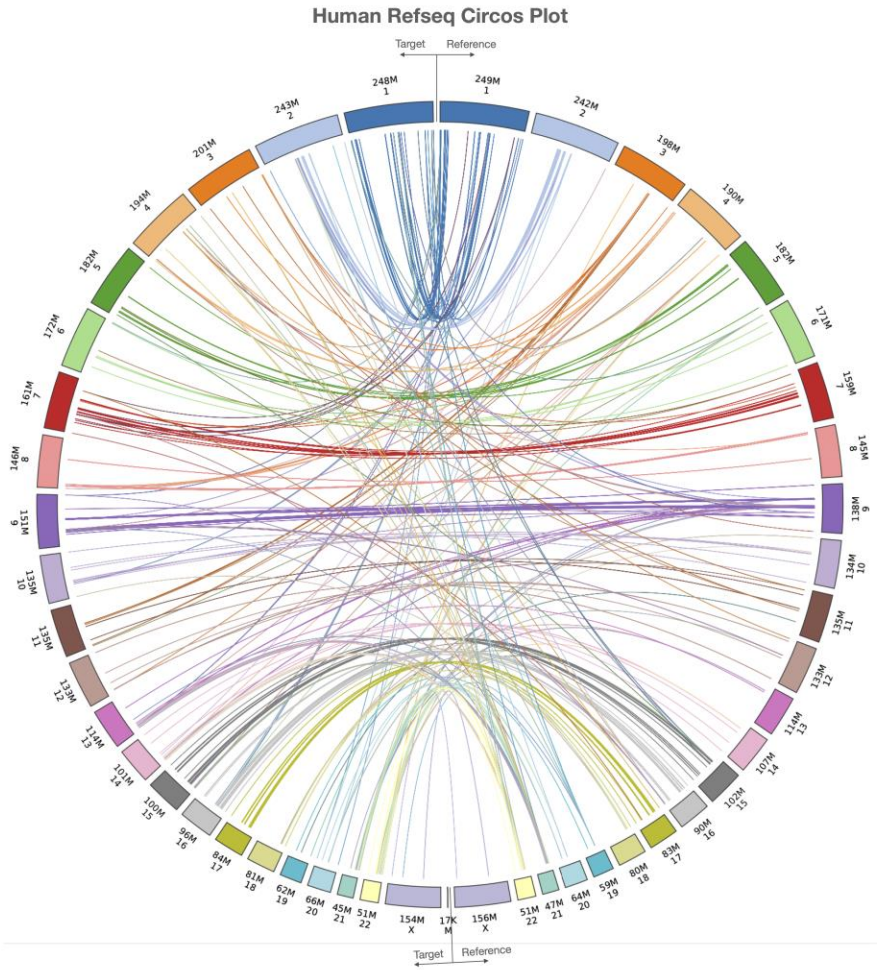
Gene on reverse strand



	Protein identity
Liftoff	38.92%
miniprot	99.14%
LiftOn	99.35%

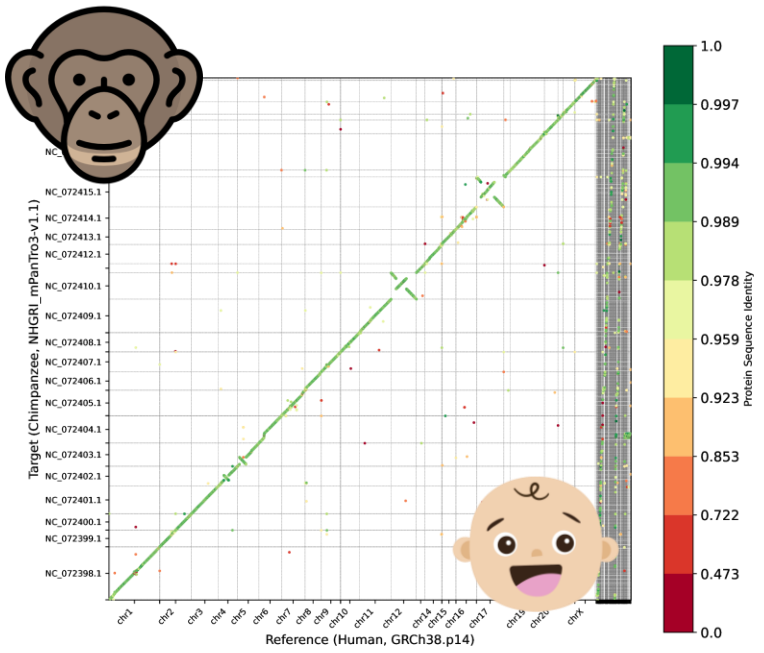
Result 2: improve CHM13 protein annotations





Result 3: improve distant species lift-over

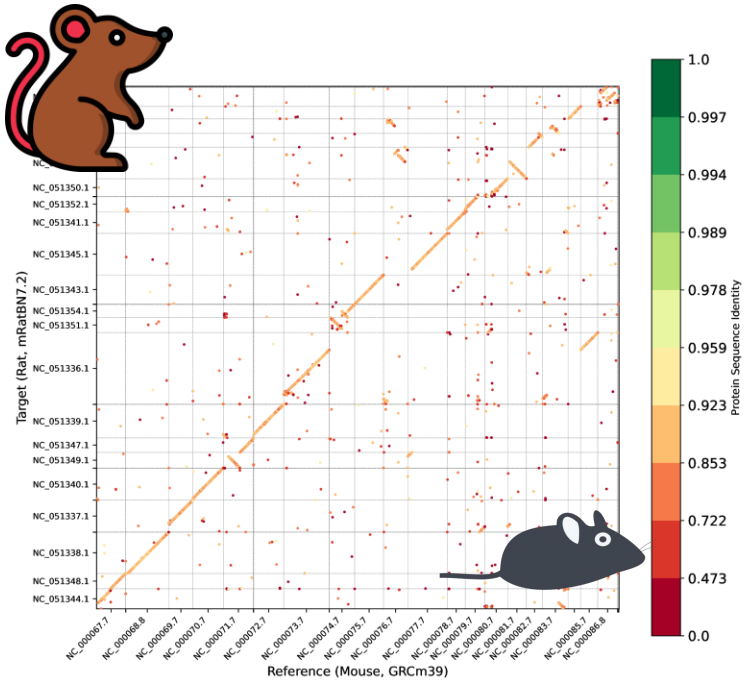
human to chimp



Mash : 0.013

Dashing2 : 0.47

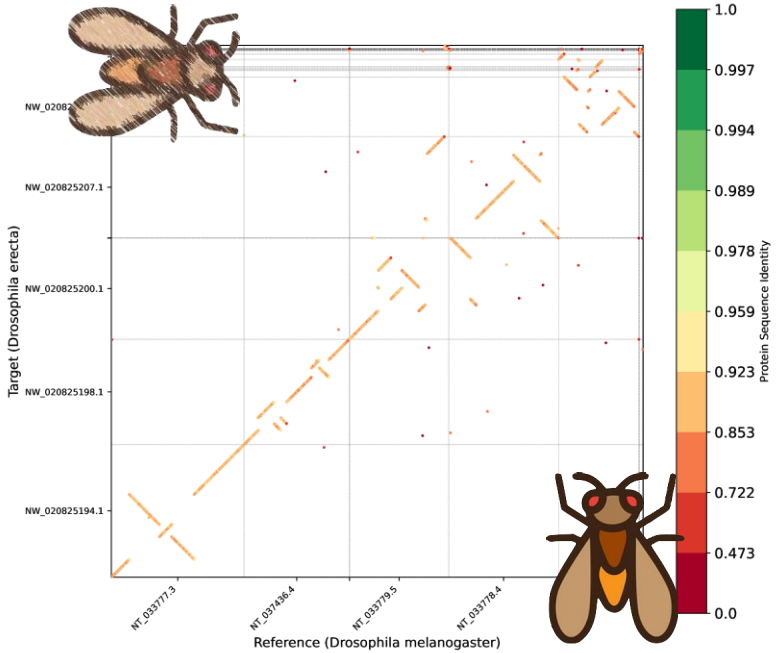
mouse to rat



Mash : 0.120

Dashing2 : 0.01

Drosophila m. to *Drosophila e.*



Mash : 0.077

Dashing2 : 0.07

Method

Genomic sketching with multiplicities and locality-sensitive hashing using Dashing 2

Daniel N. Baker and Ben Langmead

Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218-2683, USA

Ondov *et al. Genome Biology* (2016) 17:132
DOI 10.1186/s13059-016-0997-x

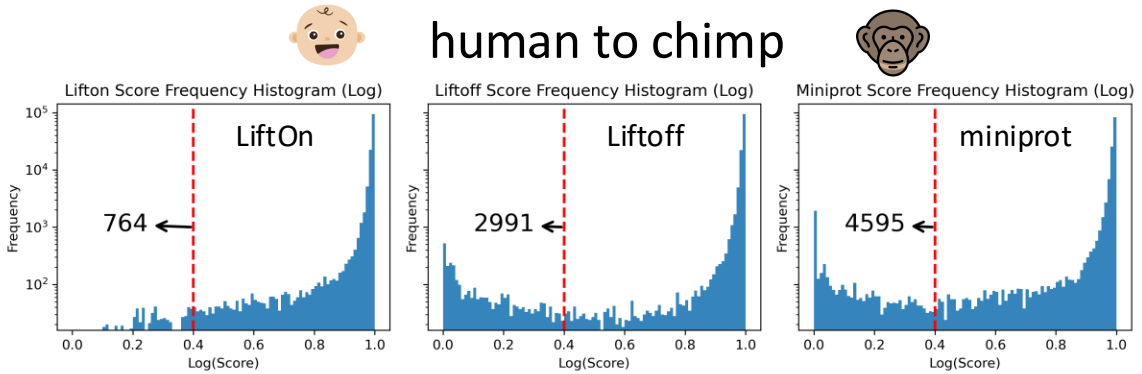
Genome Biology

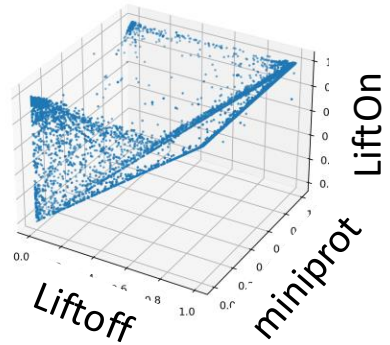
SOFTWARE Open Access

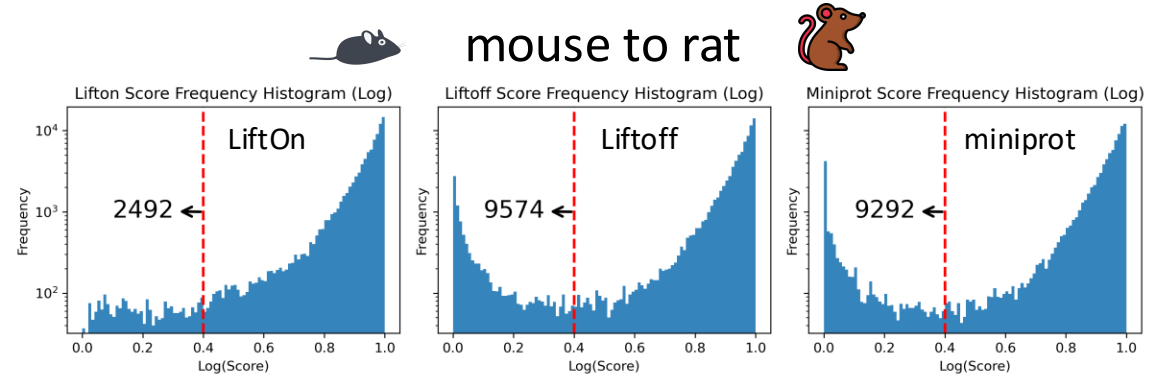
Mash: fast genome and metagenome distance estimation using MinHash

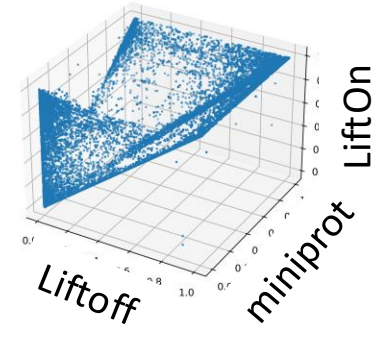
Brian D. Ondov[1], Todd J. Treangen[1], Páll Melsted[2], Adam B. Mallonee[1], Nicholas H. Bergman[1], Sergey Koren[3] and Adam M. Phillippy[3*]

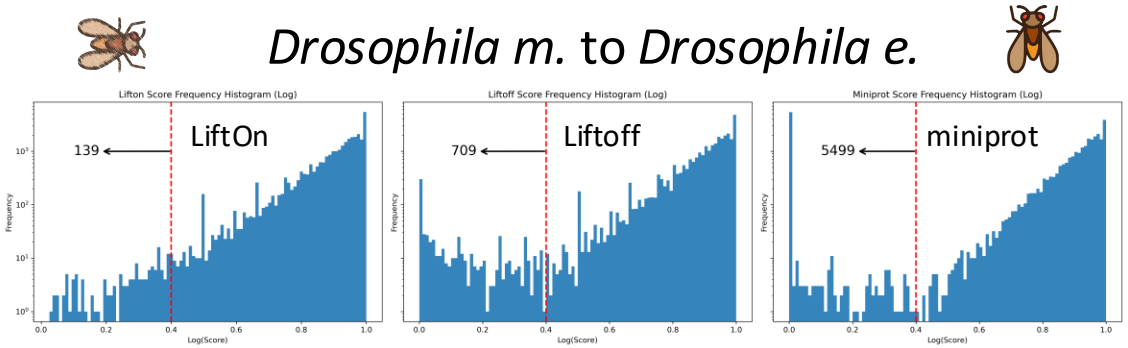
Result 3: improve distant species lift-over

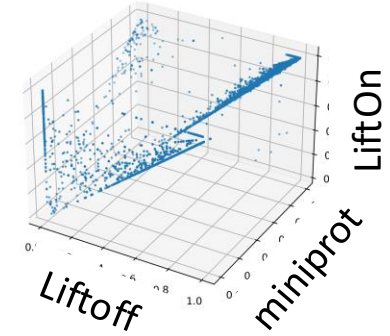












Result 3: improve distant species lift-over

House mouse
(*Mus musculus*)

Yeast
(*Saccharomyces cerevisiae*)

Honey bee
(*Apis mellifera*)

Thale cress
(Arabidopsis thaliana)

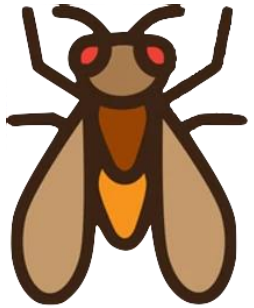
New Results

Follow this preprint

Combining DNA and protein alignments to improve genome annotation with LiftOn

Kuan-Hao Chao, Jakob M. Heinz, Celine Hoh, Alan Mao, Alaina Shumate, Mihaela Pertea, Steven L Salzberg

doi: https://doi.org/10.1101/2024.05.16.593026



fruit fly
(*Drosophila melanogaster*)

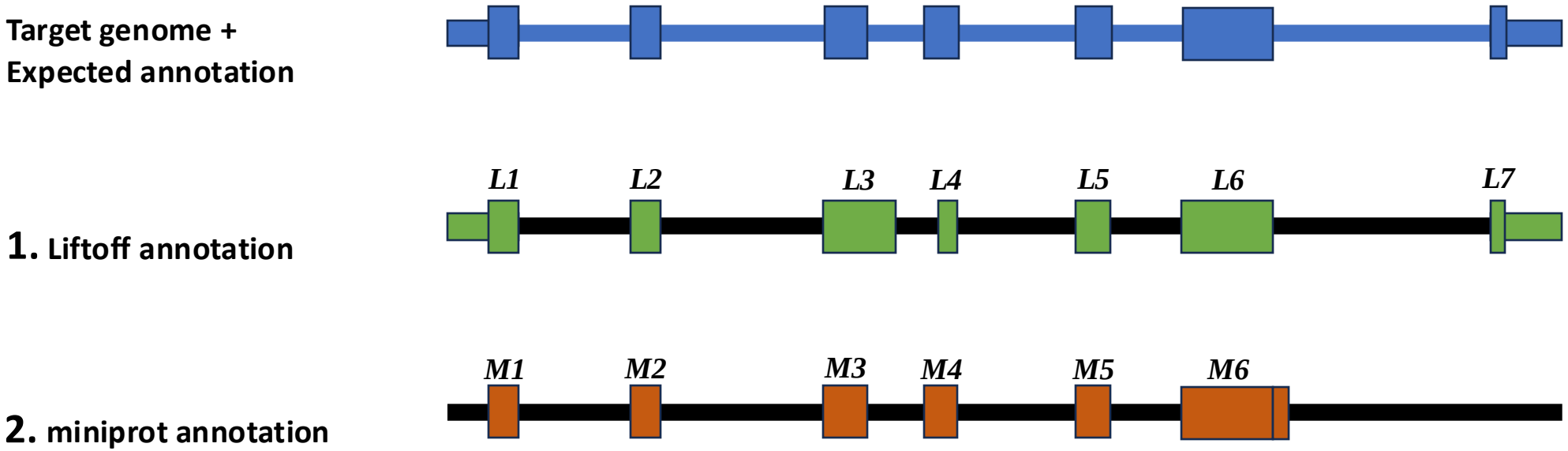
Rice
(Oryza sativa)

Methods in Details

- Protein-maximization algorithm
  - Step 1: chaining CDSs
  - Step 2: ORF search

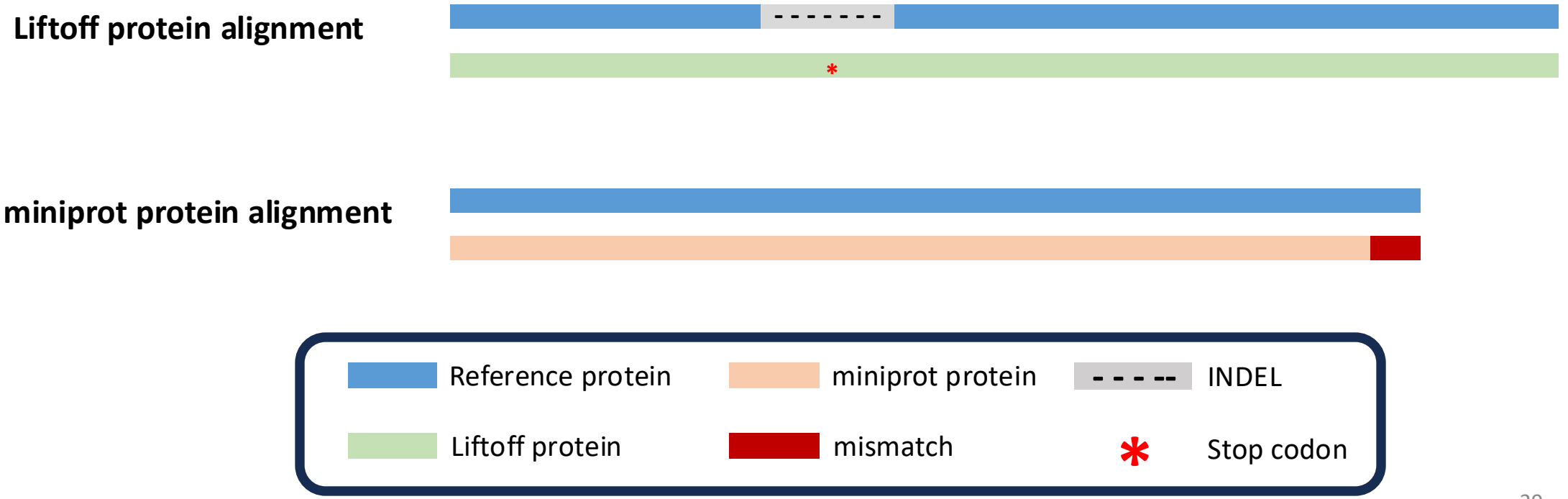
LiftOn: Protein-maximization algorithm

A



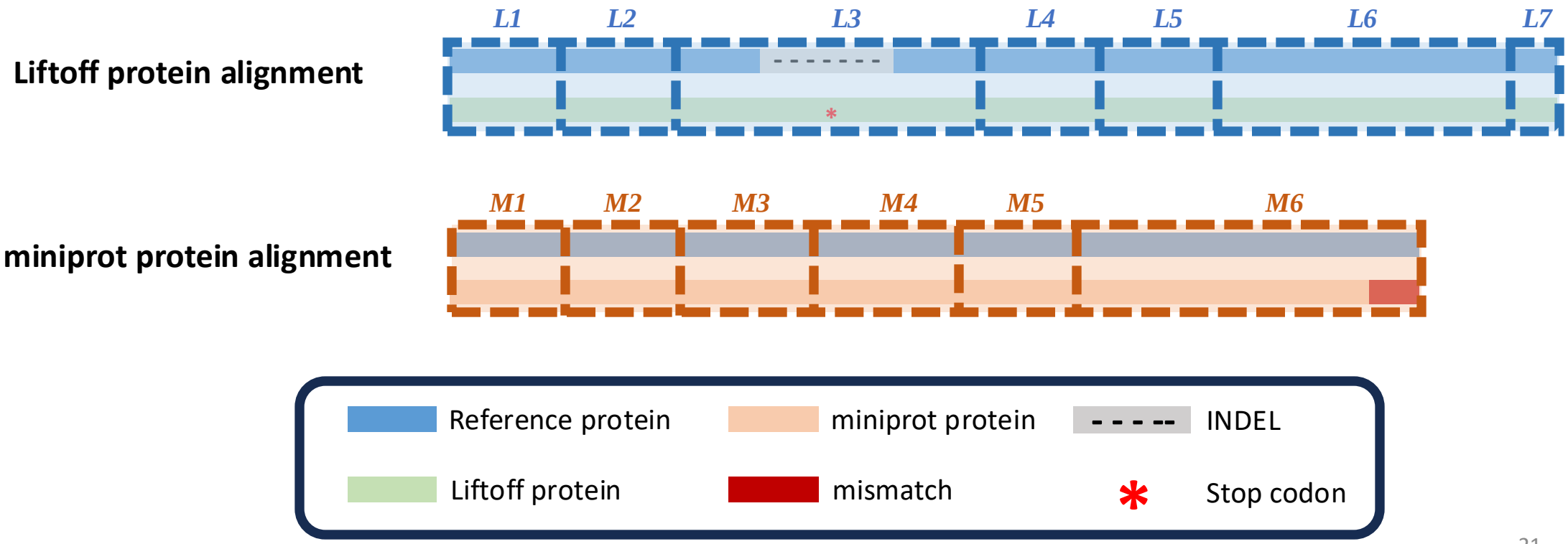
LiftOn: Protein-maximization algorithm

B **Step 1: Align Liftoff & miniprot proteins to reference protein**



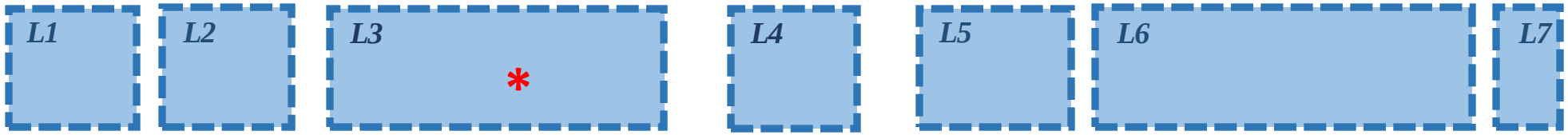
LiftOn: Protein-maximization algorithm

C **Step 2: Mapped CDS boundaries onto Liftoff & miniprot protein alignments**



LiftOn : Protein-maximization algorithm

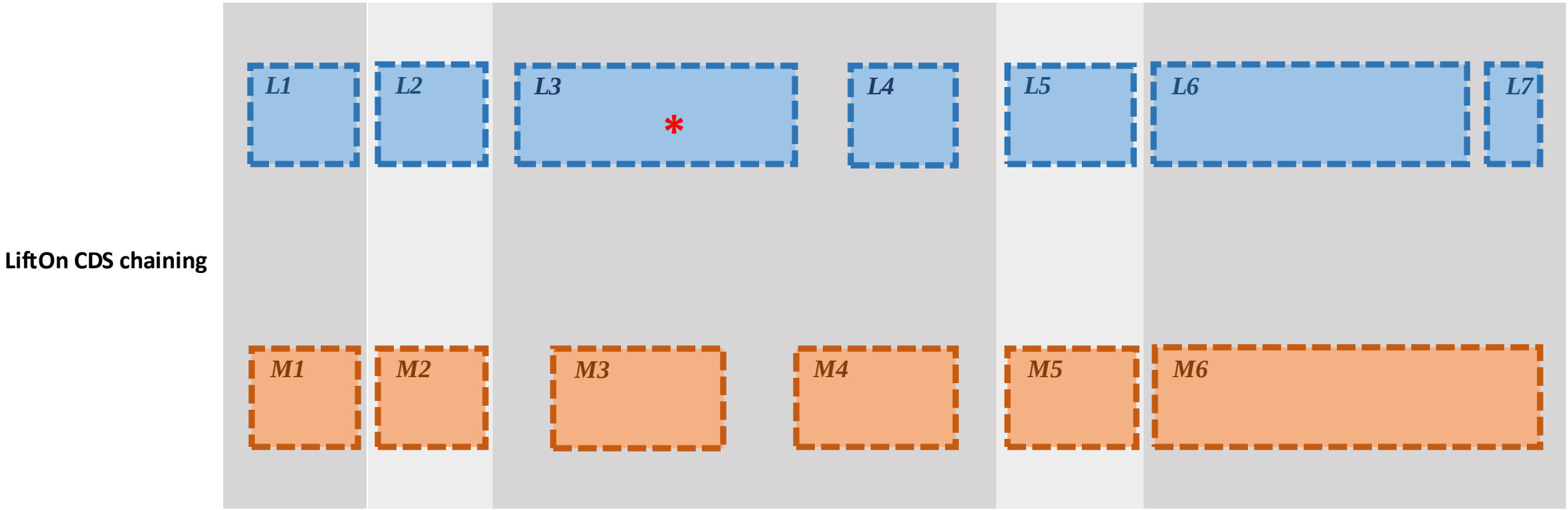
D Step 3: group CDSs by “accumulated AA in the reference protein”



LiftOn CDS chaining

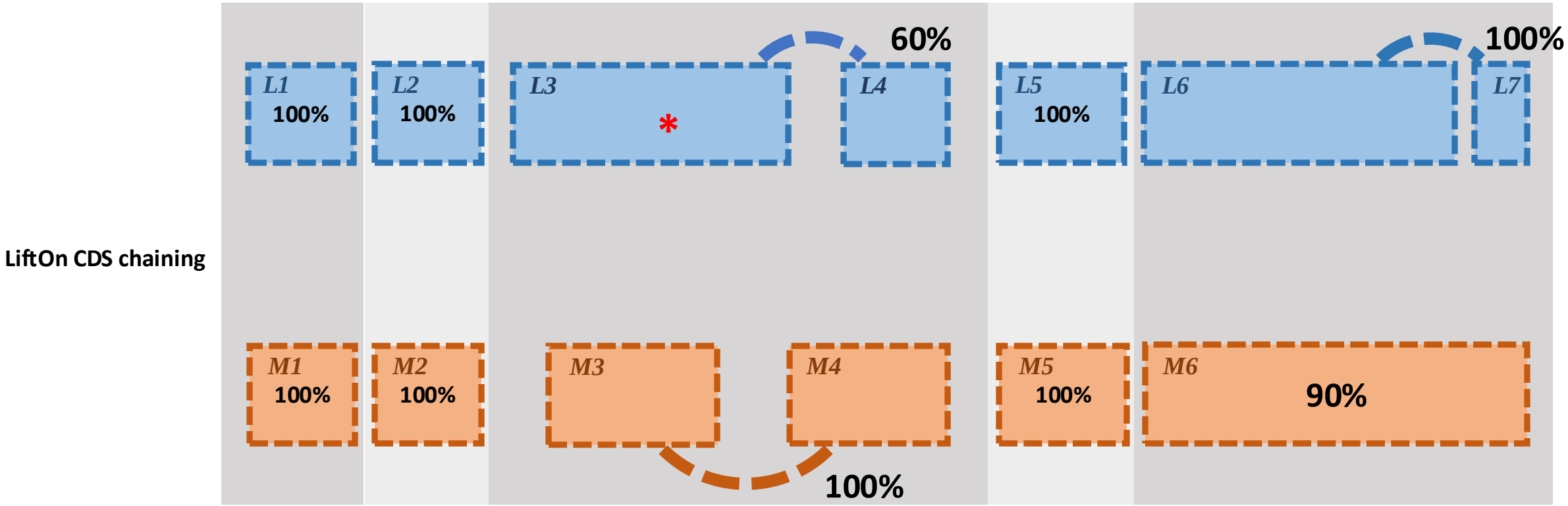
LiftOn: Protein-maximization algorithm

D **Step 3: group CDSs by “accumulated AA in the reference protein”**



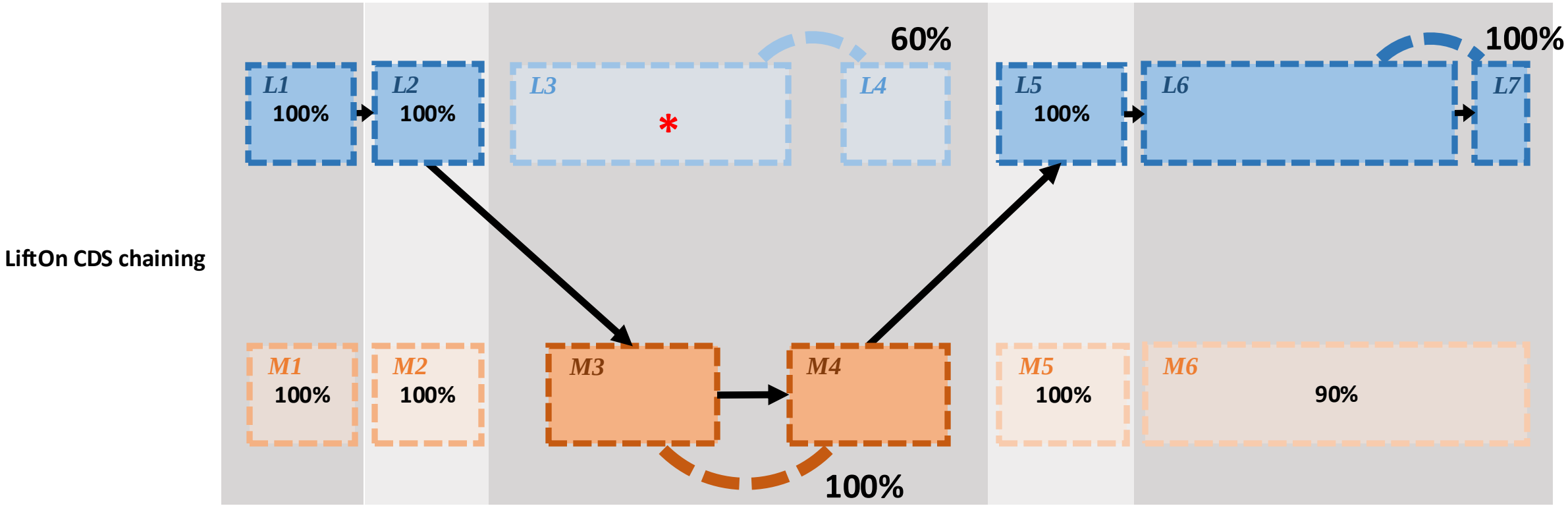
LiftOn : Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

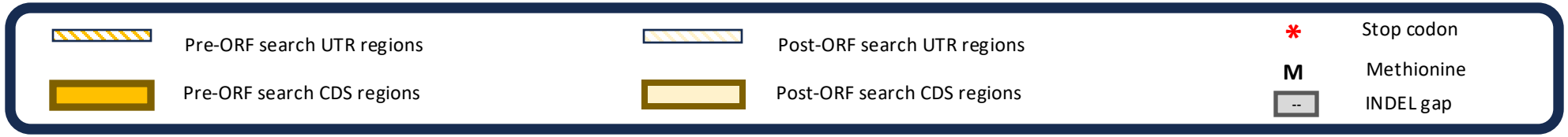


LiftOn : Protein-maximization algorithm

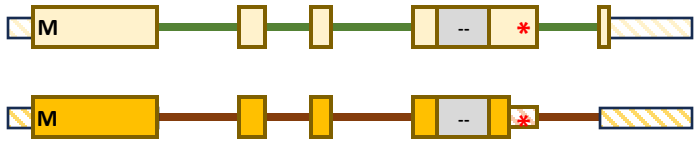
D **Step 3: group CDSs by “accumulated AA in the reference protein”**



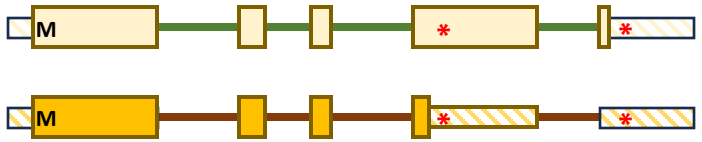
LiftOn: Protein-maximization algorithm



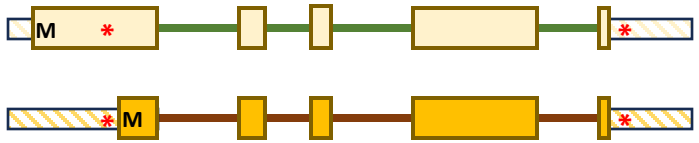
A Frameshift



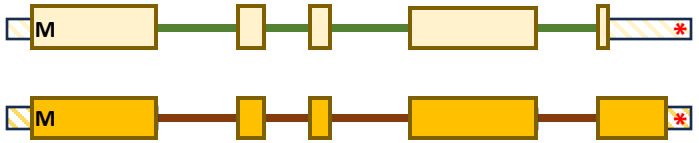
B Stop codon gain: Early translation stop



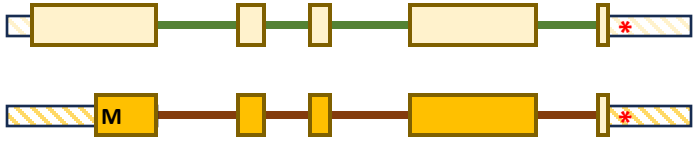
C Stop codon gain: Switching translation start



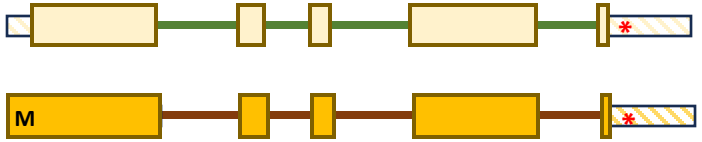
D Stop codon lost: Protein extension



E Start codon lost: Downstream start



F Start codon lost: Upstream start



LiftOn : Summary

Google search:

“LiftOn genome”

- LiftOn is a promising new tool to study comparative genomics
- LiftOn uses both DNA-DNA alignments (from Liftoff) & protein-DNA alignments (from miniprot) to map annotations between genome assemblies of the same or different species.
- LiftOn's protein-maximization algorithm improves the annotation of protein-coding genes in the T2T-CHM13 genome.
- LiftOn can map annotation between relatively distant species, at least as divergent as mouse and rat.

Chao, K. H., Heinz, J. M., Hoh, C., Mao, A., Shumate, A., Pertea, M., & Salzberg, S. L. (2024). Combining DNA and protein alignments to improve genome annotation with LiftOn. **bioRxiv**.

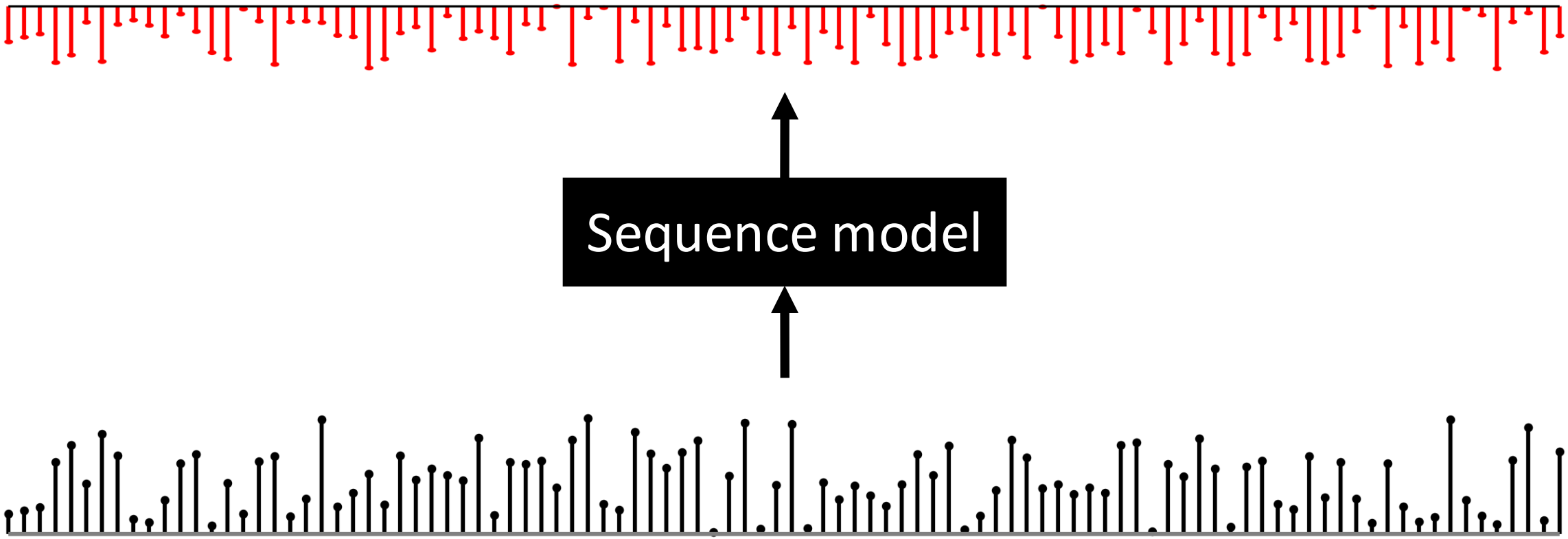
ccb.jhu.edu/lifton **github.com/Kuanhao-Chao/LiftOn**

Part II

Deep learning splice site prediction

- OpenSpliceAI: Pytorch reimplementation of SpliceAI
- Splam: splice junction recognizer

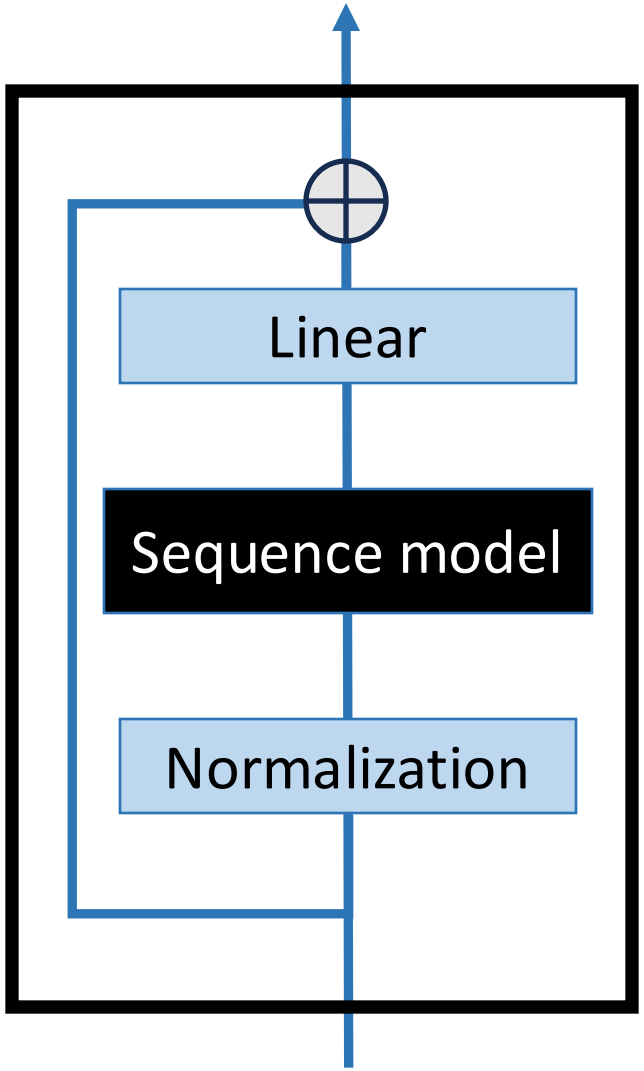
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s



Sequence models map a sequence to a sequence

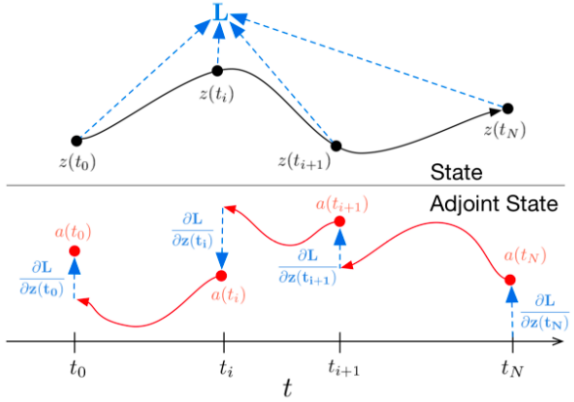
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

(batch, length, dim)

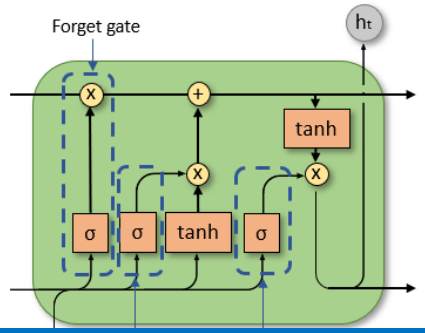


(batch, length, dim)

Neural ODEs

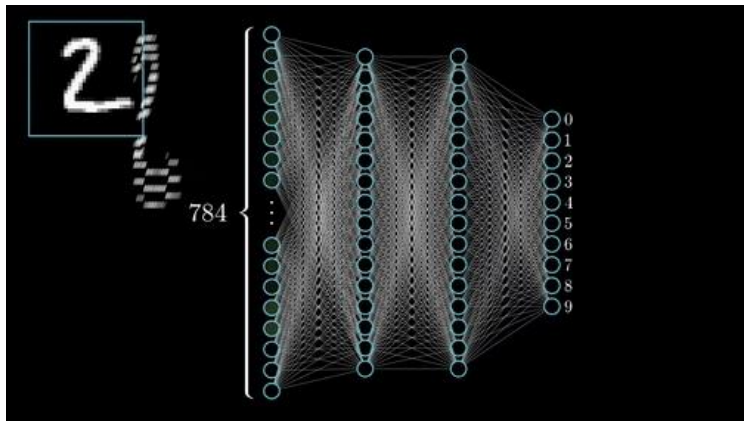


RNN



Future work

CNNs



Transformers

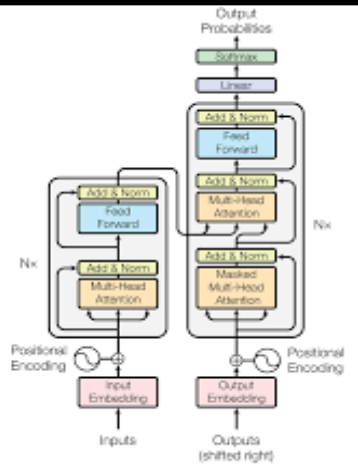
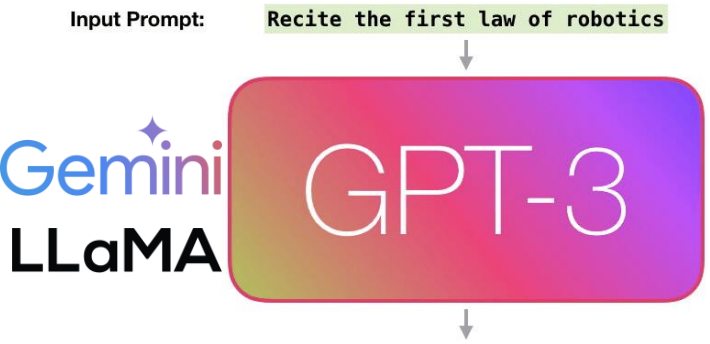
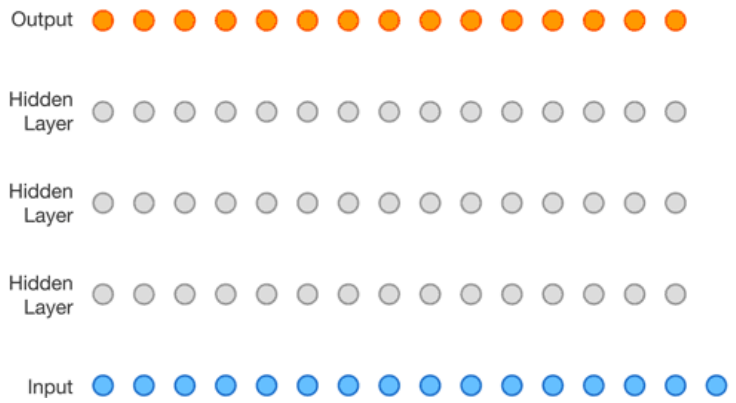


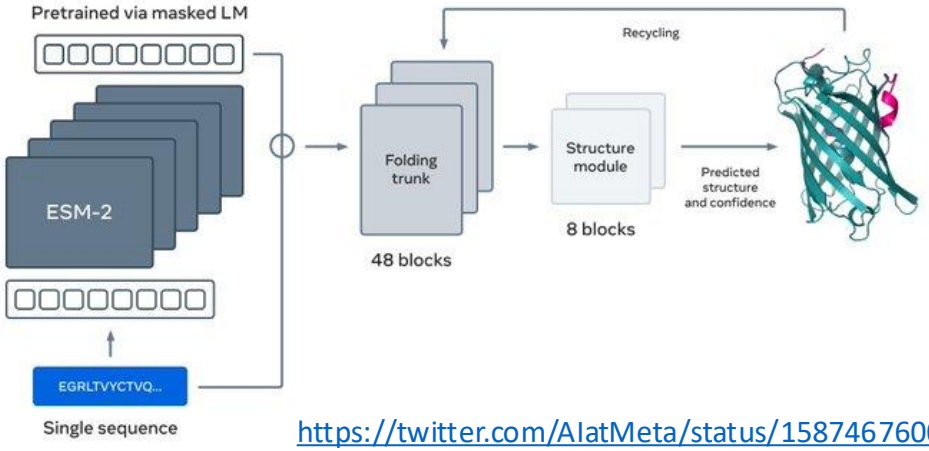
Figure 1: The Transformer - model architecture.





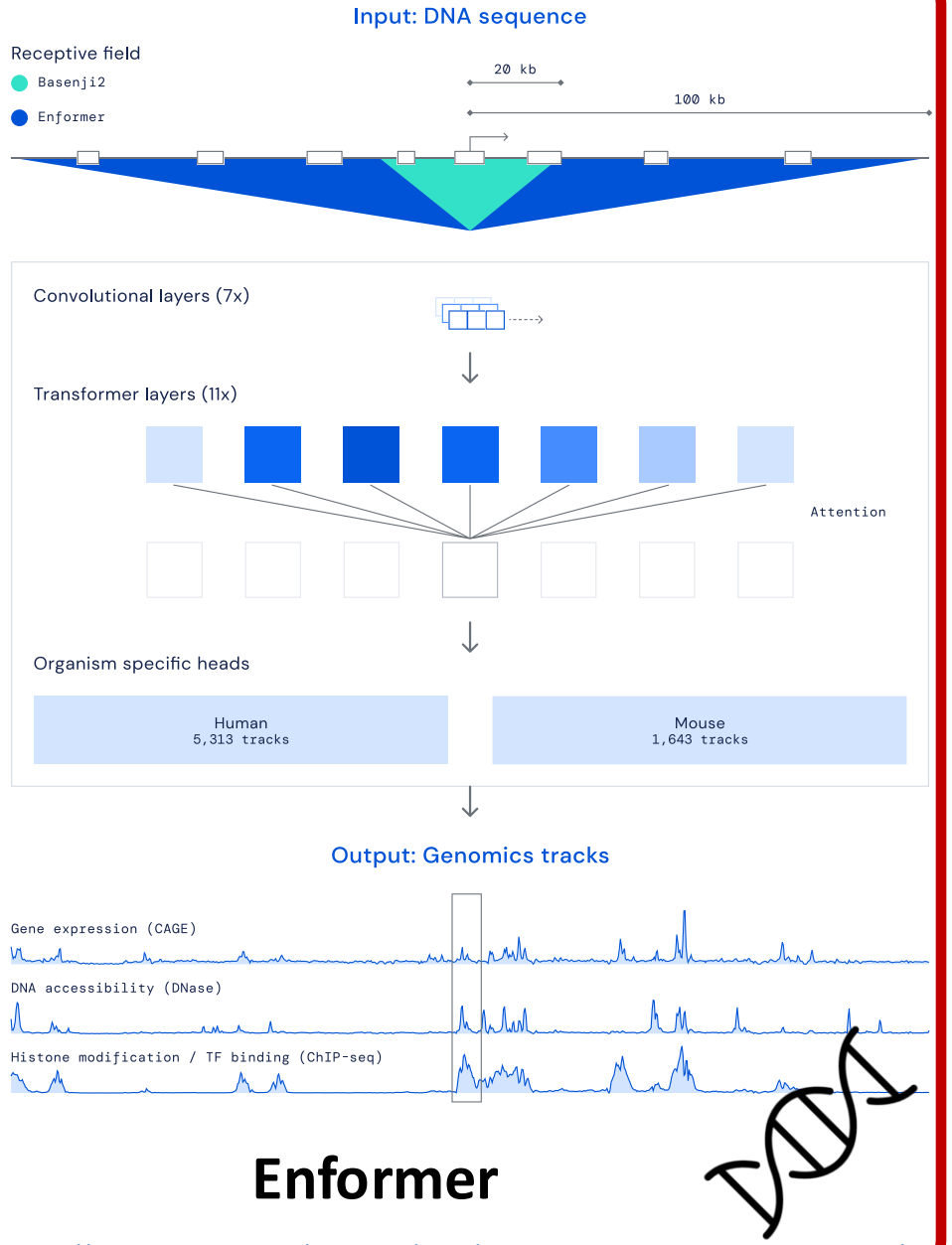
https://jalammar.github.io/how-gpt3-works-visualizations-animations/

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



ESMFold

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

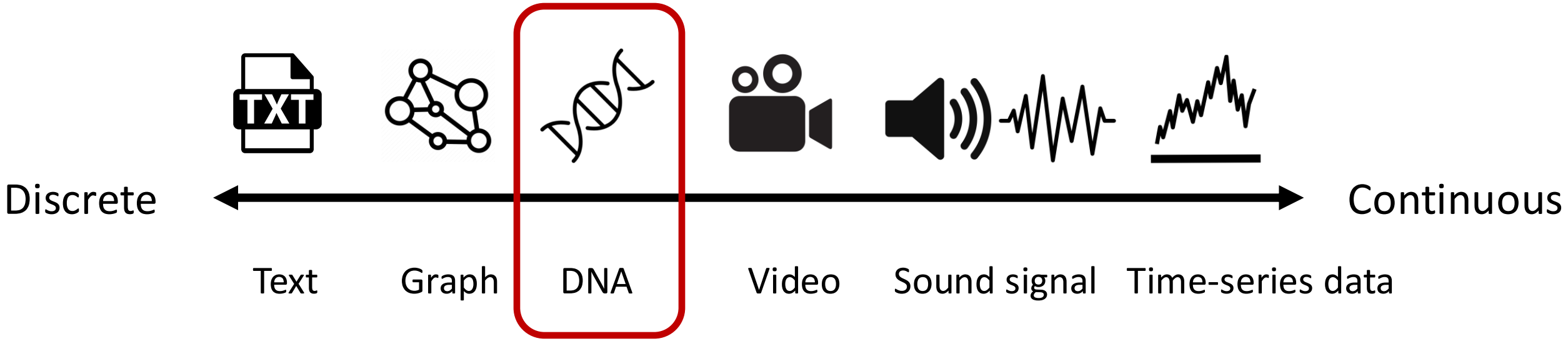


Enformer

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

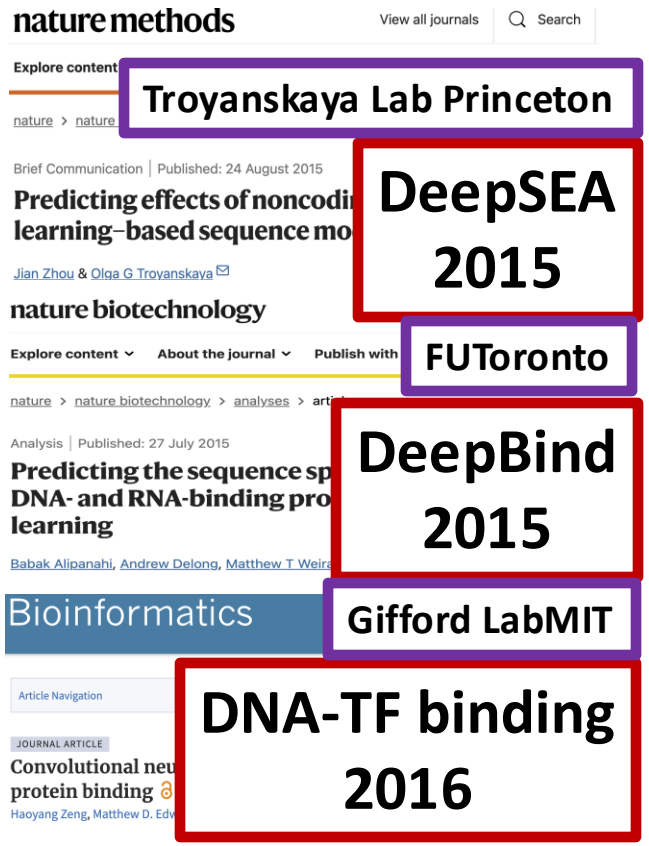
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

Spectrum of Sequential Data



DeepMind Calico illumina® Troyanskaya Laboratory

Why Convolutional Neural Network to DNA ?



Troyanskaya Lab Princeton

DeepSEA 2015

FUToronto

DeepBind 2015

Gifford LabMIT

DNA-TF binding 2016

Calico

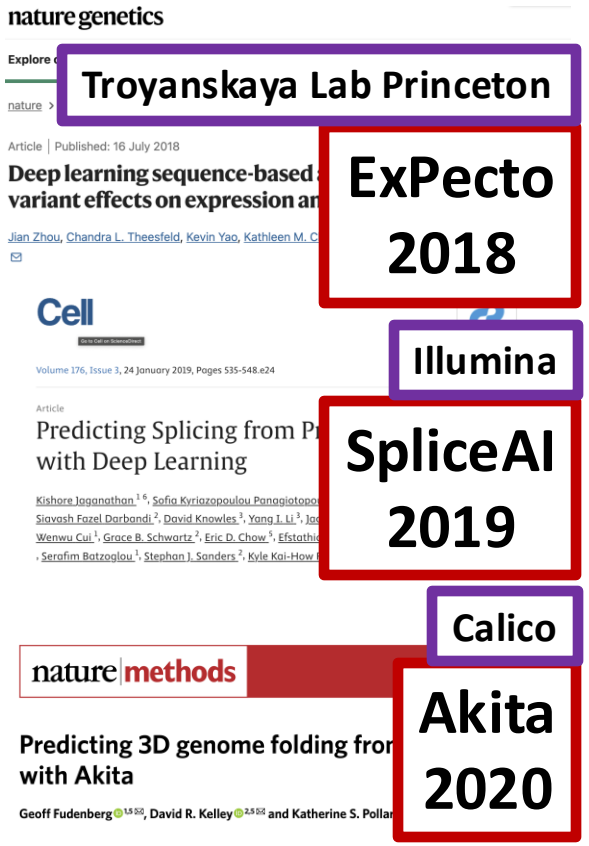
Basset 2016

Calico

Basenji 2018

Google Health

DeepVariant 2018



Troyanskaya Lab Princeton

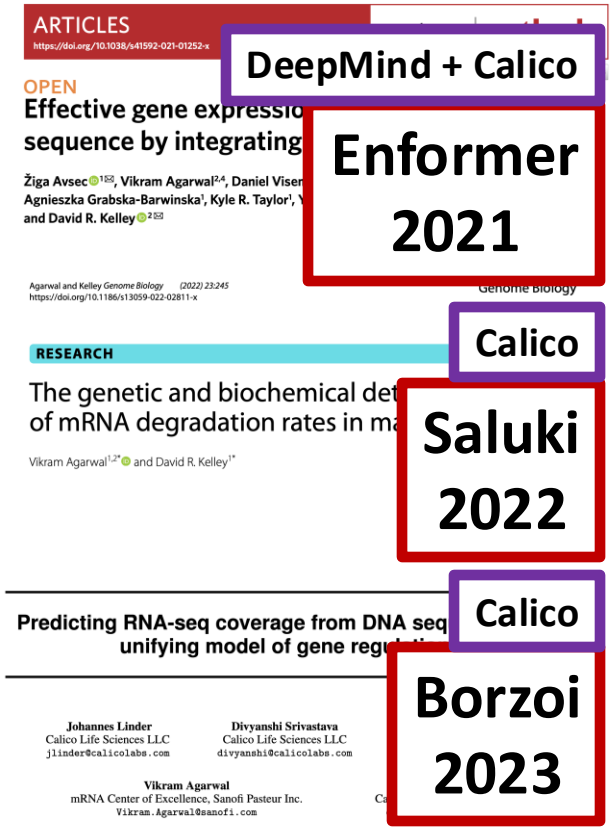
ExPecto 2018

Illumina

SpliceAI 2019

Calico

Akita 2020



DeepMind + Calico

Enformer 2021

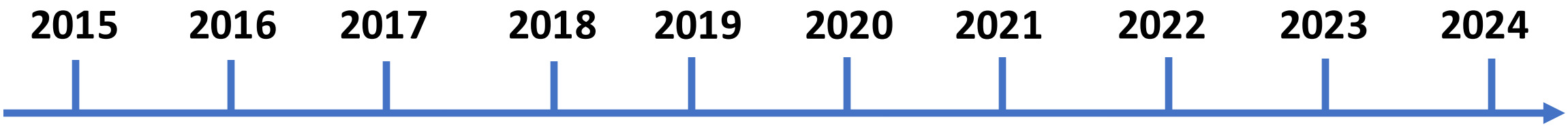
Calico

Saluki 2022

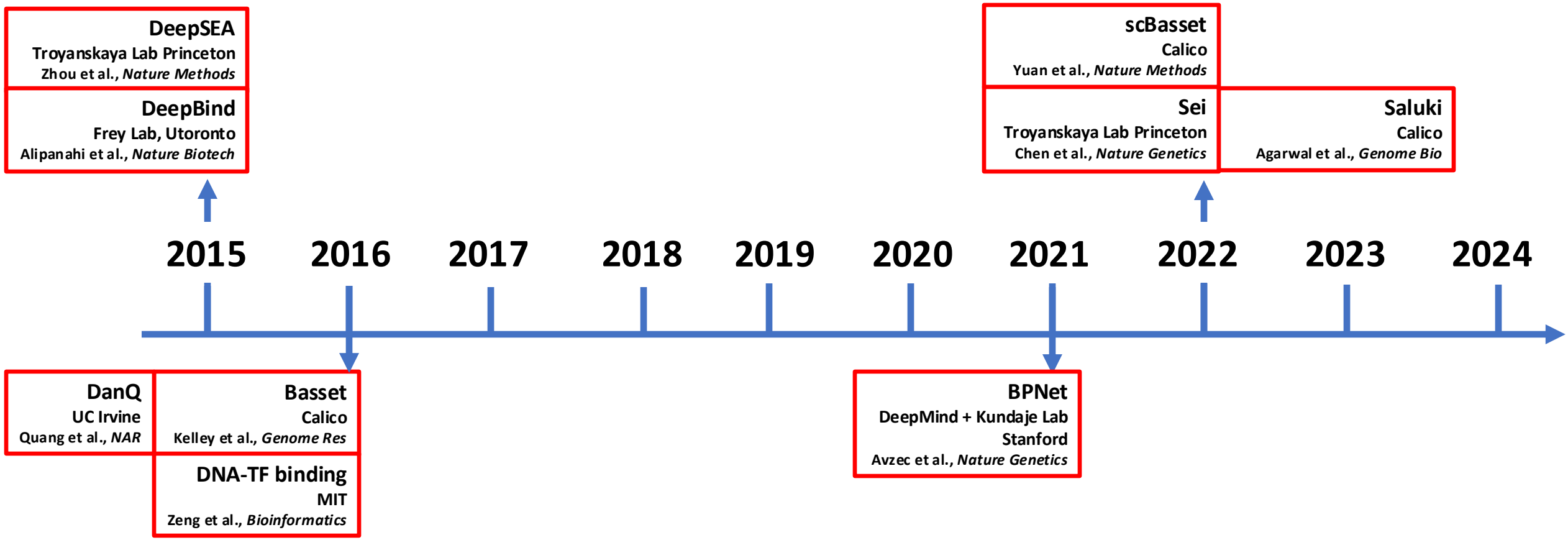
Calico

Borzoi 2023

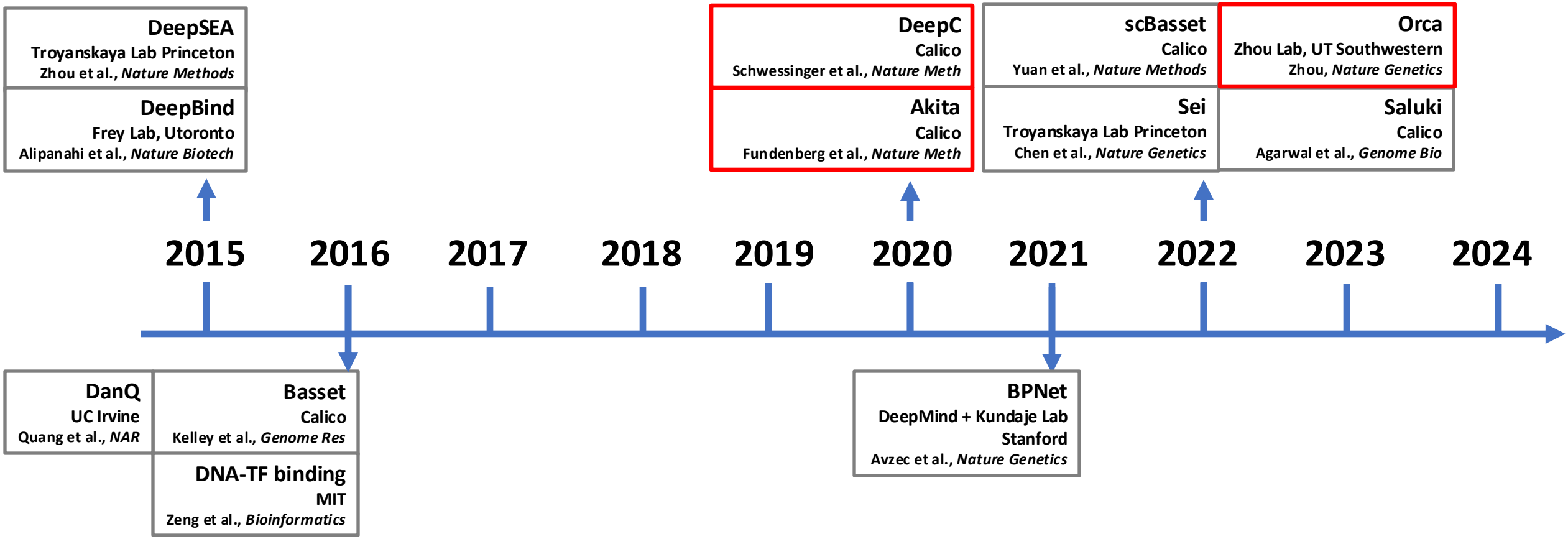
Why Convolutional Neural Network to DNA ?



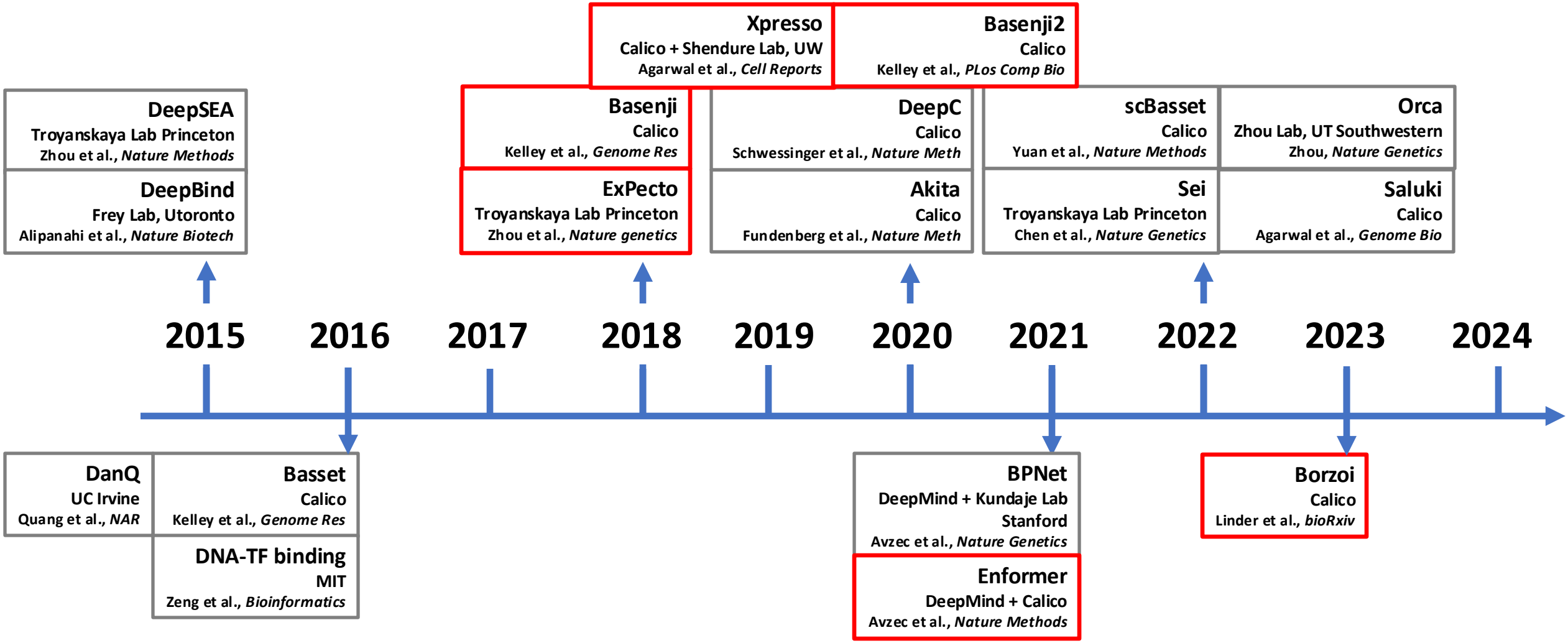
Why Convolutional Neural Network to DNA ?



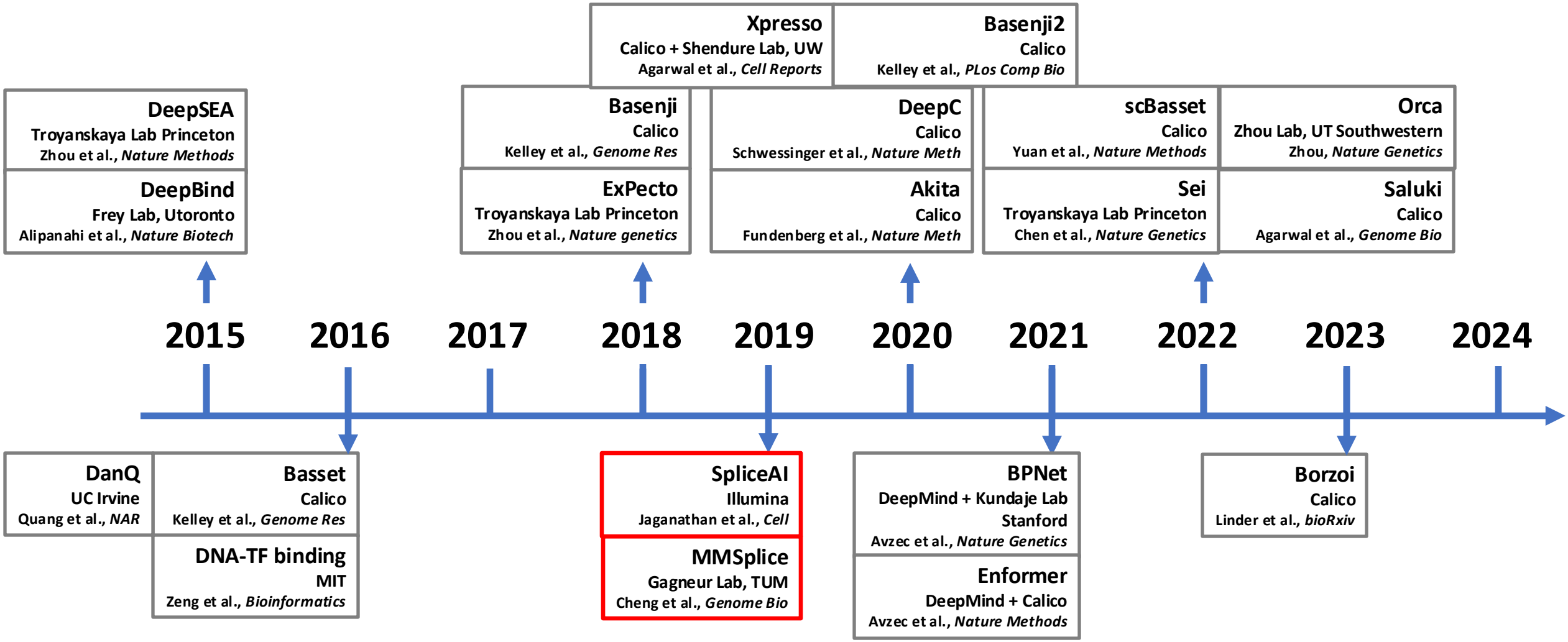
Why Convolutional Neural Network to DNA ?



Why Convolutional Neural Network to DNA ?



Why Convolutional Neural Network to DNA ?



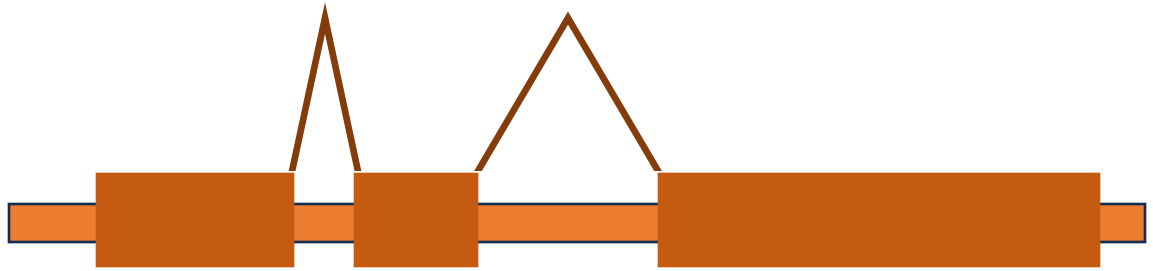

Steven Salzberg


Mihaela Pertea


Anqi Liu

OpenSpliceAI : splice site predictor

Pre-mRNA



Chao, K. H., Mao, A., Liu, Anqi, Salzberg, S. L., & Pertea, M. (**2024**). OpenSpliceAI. 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!

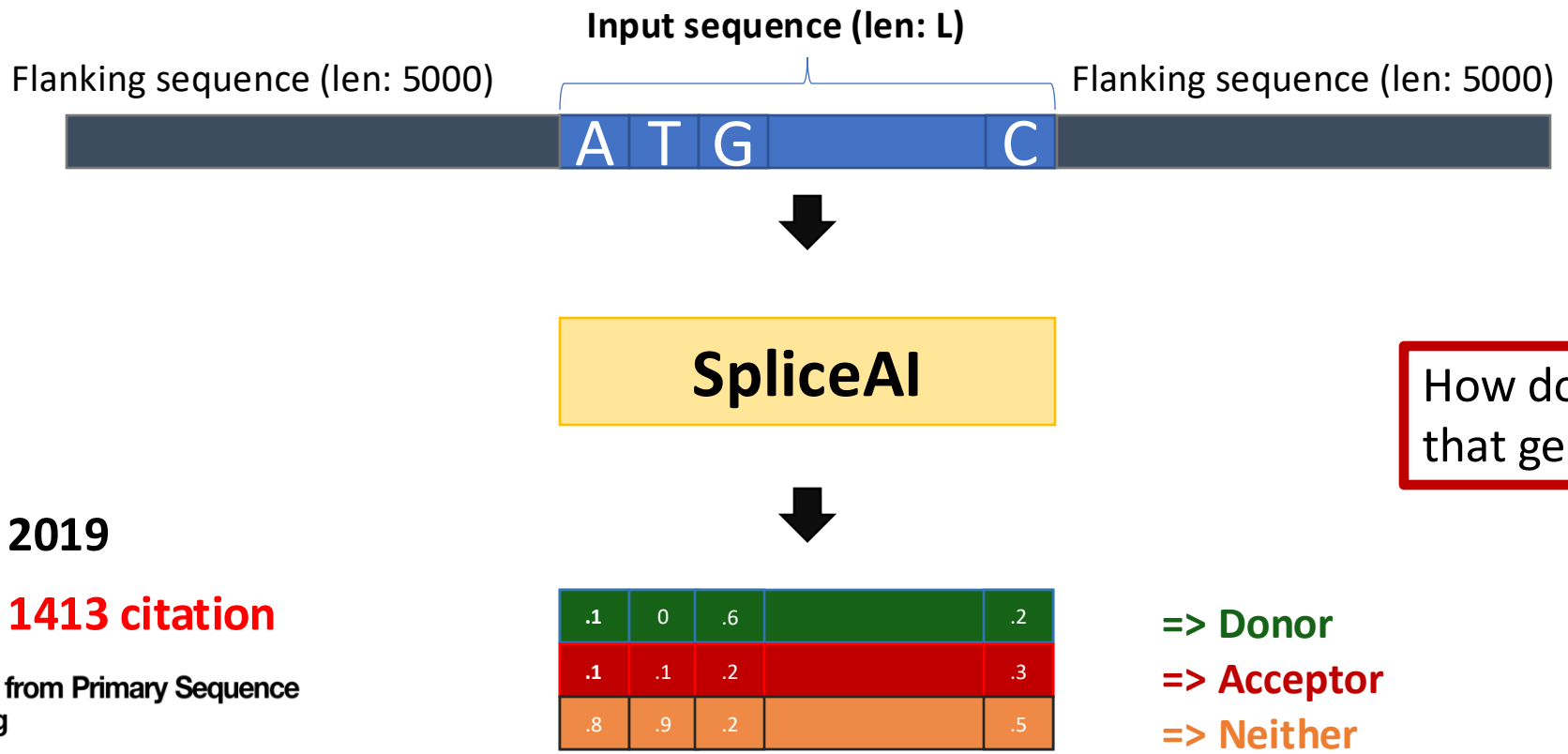
X AGACTCAGCCCCCGGAGACTTAGTTAGAGGAAGAAAAAGGTAGGACAGAAGAAAAAGGCAGGACATACAAGGTGCTGGCCCAGGGCGG

Y 000000000000000000000002000000010000000020000000000010000000000000000000000000000000000

Donor: 2 Acceptor: 1 Neither: 0

SpliceAI: splite site predictor

$$\begin{bmatrix} A \\ C \\ G \\ T \\ N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} Acceptor \\ Donor \\ Neither \\ Padding \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$



Dimension:

X: $(L+10000) * 4$

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

Y: $L * 3$

2019

1413 citation

Cell

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 **0000000200001000000020000010000020001000002000000010000000**

$$X = \left[\text{NN … NNATGTCGTGTCGAGTTG} \right]$$

$$Y = \left[\text{0000000200} \right]$$

SpliceAI: data preprocessing

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

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 00000002000001000000020000010000020001000002000000010000000

X=[NN … NNATGTCGTGTCGAGTTG
TCGTGTCGAGTTGTCGTGTTCAG]

Y=[000000020 0
0001000000]

SpliceAI: data preprocessing

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

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 0000000200001000000020000010000020001000002000000010000000

X=[NN … NNATGTCGTGTCGAGTTG
TCGTGTCGAGTTGTCGTGTTCAG
TTGTCGTGTTCAGGTCAGTCAGG]

Y=[0000000200
0001000000
2000001000]

SpliceAI: data preprocessing

X NN … NNATGTCGTGTCGAGTTGTCGTGTTCAGGTCAGTCAGGTCAGTAAGTAGAGCTCANN … NN

Y 0000000200001000000020000010000020001000002000000010000000

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

$$\mathbf{X} = \left[\begin{array}{l} \text{NN … NNATGTCGTGTCGAGTTG} \\ \text{TCGTGTCGAGTTGTCGTGTTCAG} \\ \text{TTGTCGTGTTCAGGTCAGTCAGG} \\ \vdots \\ \text{AAGTAGAGCTCANNNNN … NN} \end{array} \right]$$

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

$$\mathbf{Y} = \left[\begin{array}{l} \text{0000000200} \\ \text{0001000000} \\ \text{2000001000} \\ \vdots \\ \text{0000000000} \end{array} \right]$$

SpliceAI: data preprocessing

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

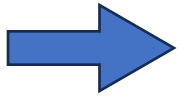
Raw gene DNA sequence

Gene 1  $L = 33200$

Gene 2  $L = 14600$

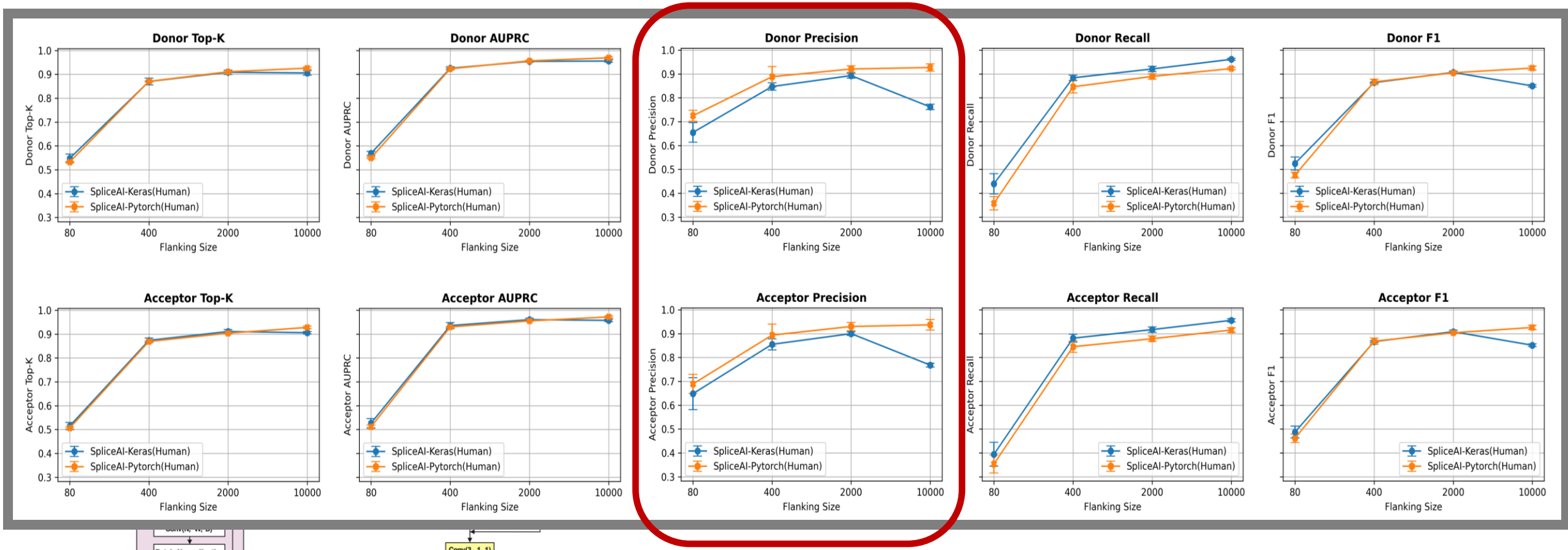
...

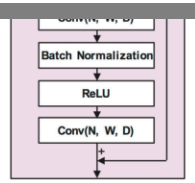
Gene n  $L = 25000$

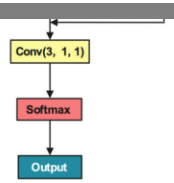


X	Y
[7, 15000, 4]	[7, 5000, 3]
[3, 15000, 4]	[3, 5000, 3]
...	
[5, 15000, 4]	[5, 5000, 3]

OpenSpliceAI : better than SpliceAI!

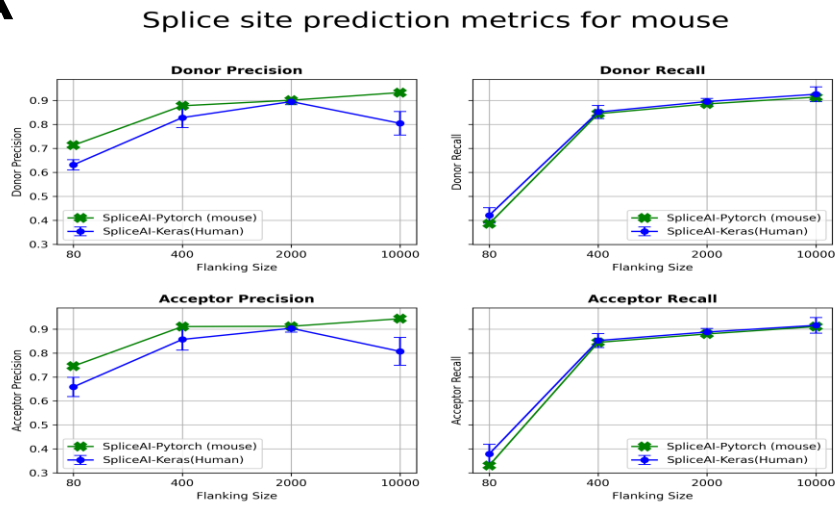




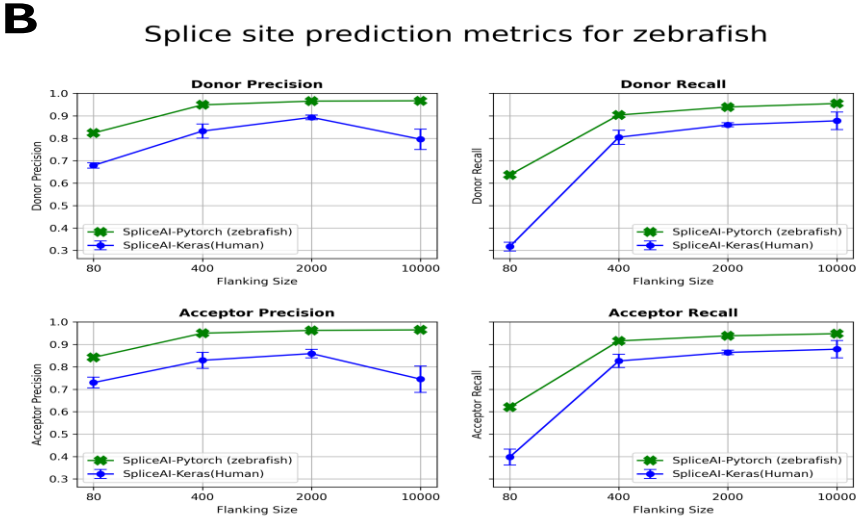


OpenSpliceAI : retrain on different species

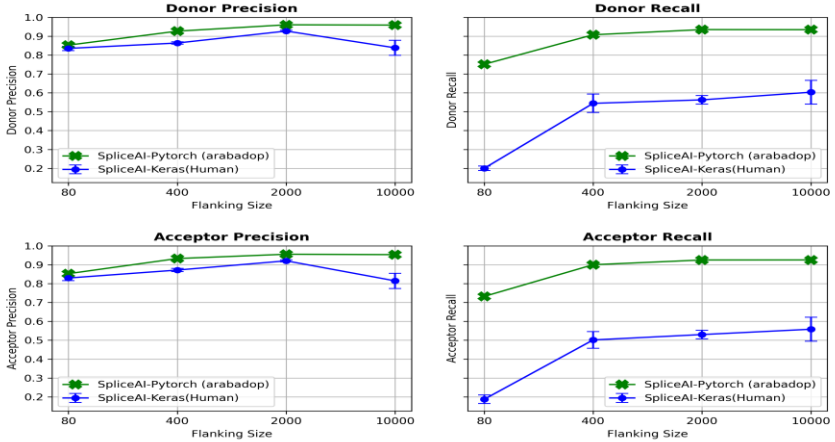
A



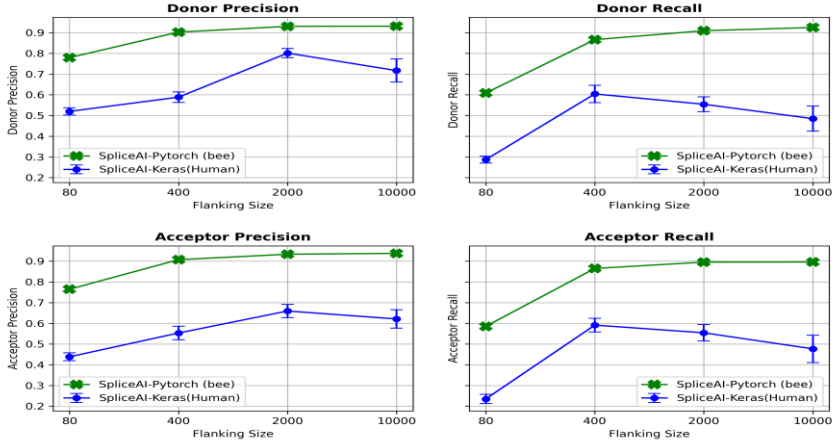
B



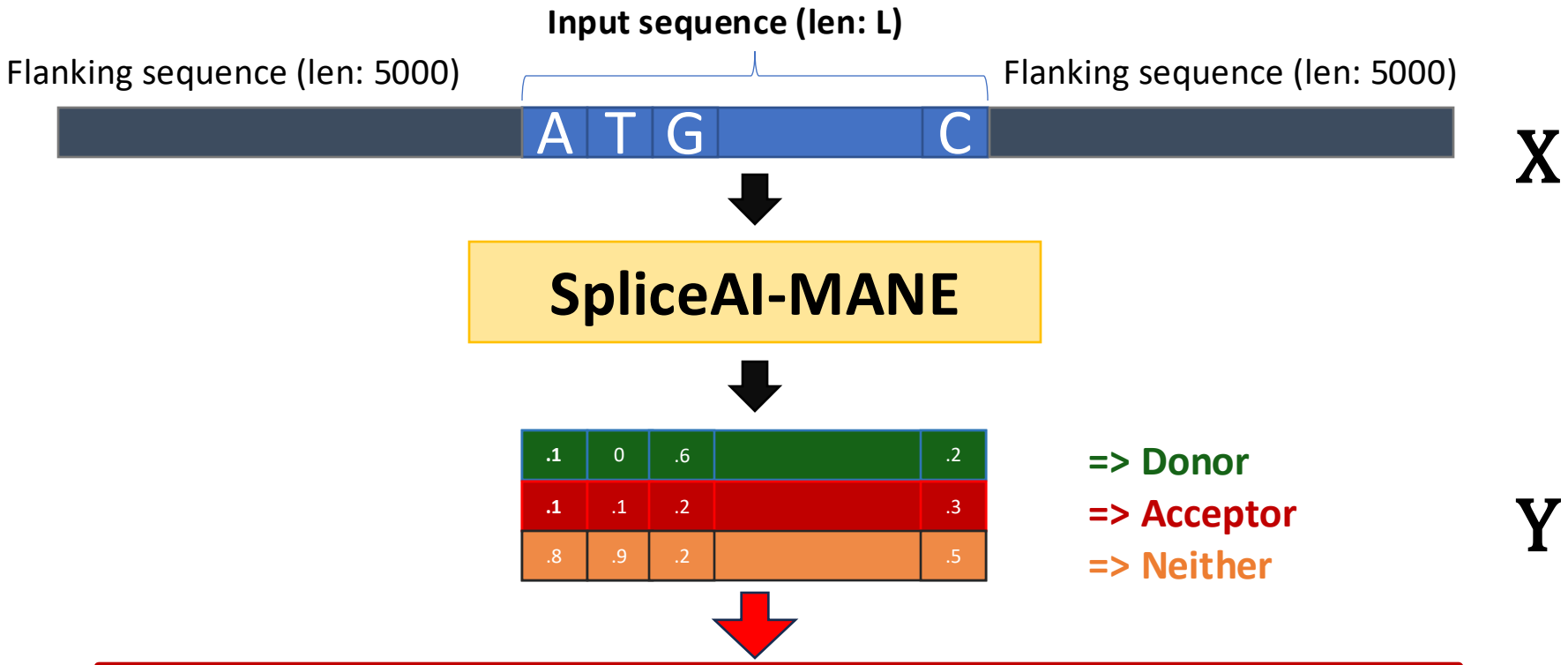
C



D



OpenSpliceAI : Calibration



X

Y

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

OpenSpliceAI : Calibration

Model

0.99

OpenSpliceAI : Calibration

Model

0.00

0 0.2 0.4 0.6 0.8 1.0

Model predicted probability

OpenSpliceAI : new concept – Calibration

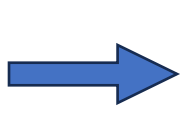
Model

0.6 0.9

0 0.2 0.4 0.6 0.8 1.0

Model predicted probability

OpenSpliceAI : new concept – Calibration



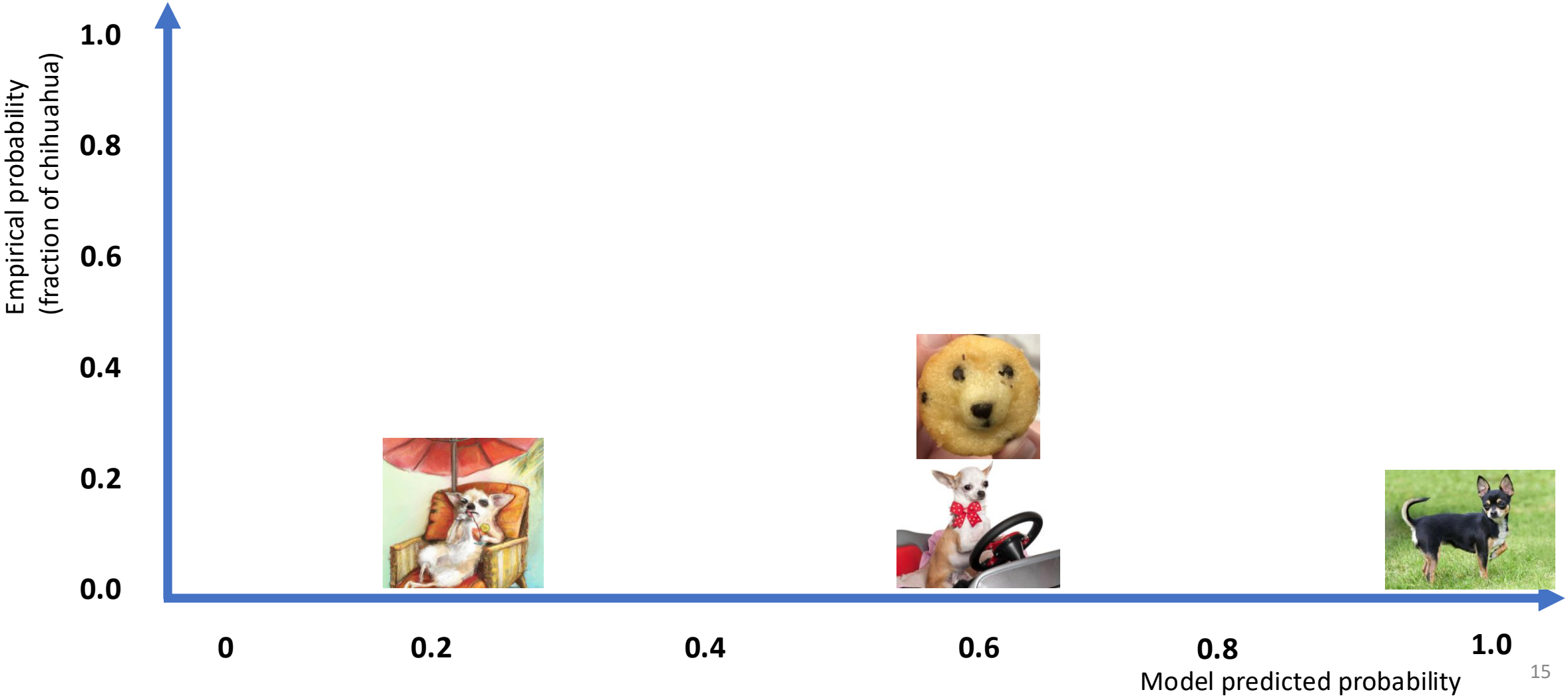
Model

0.69

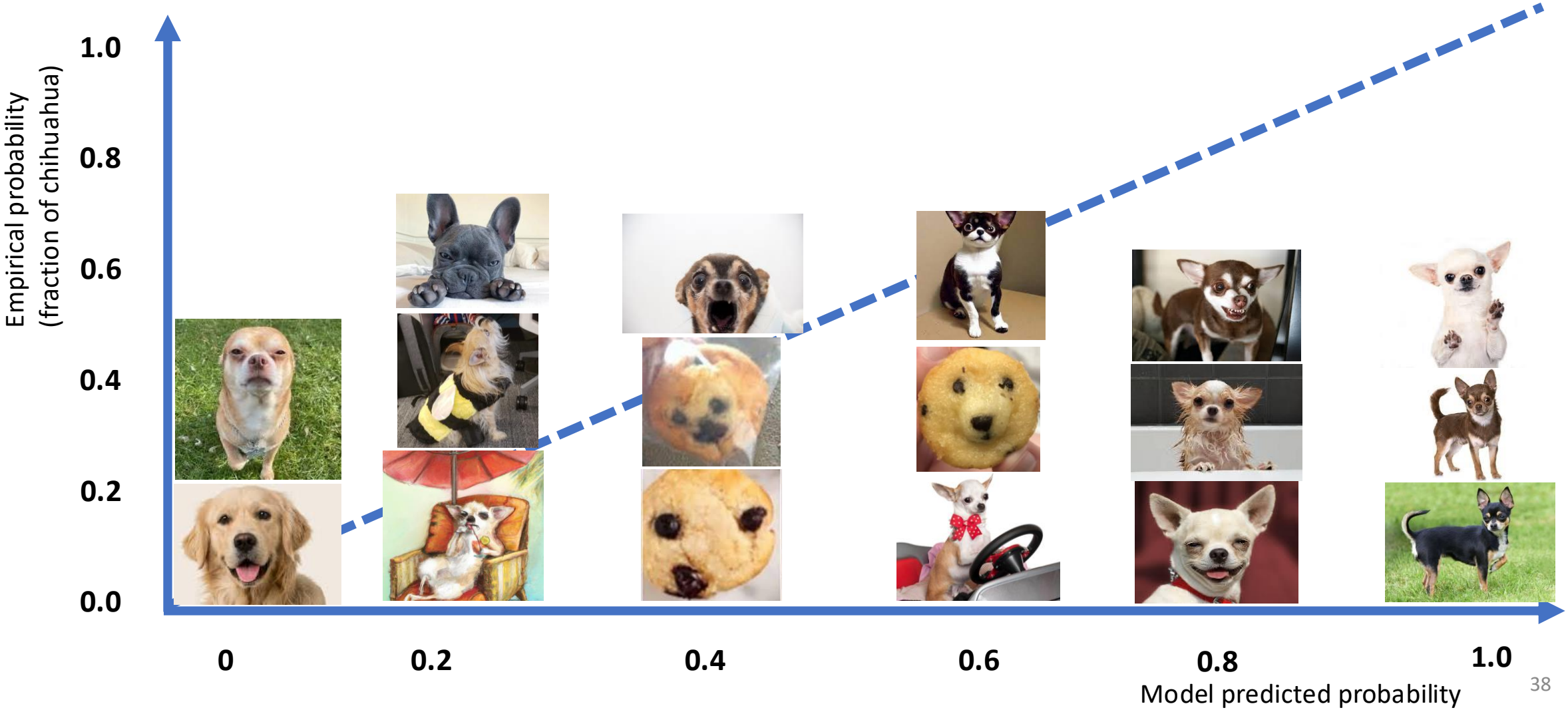
0 0.2 0.4 0.6 0.8 1.0

Model predicted probability

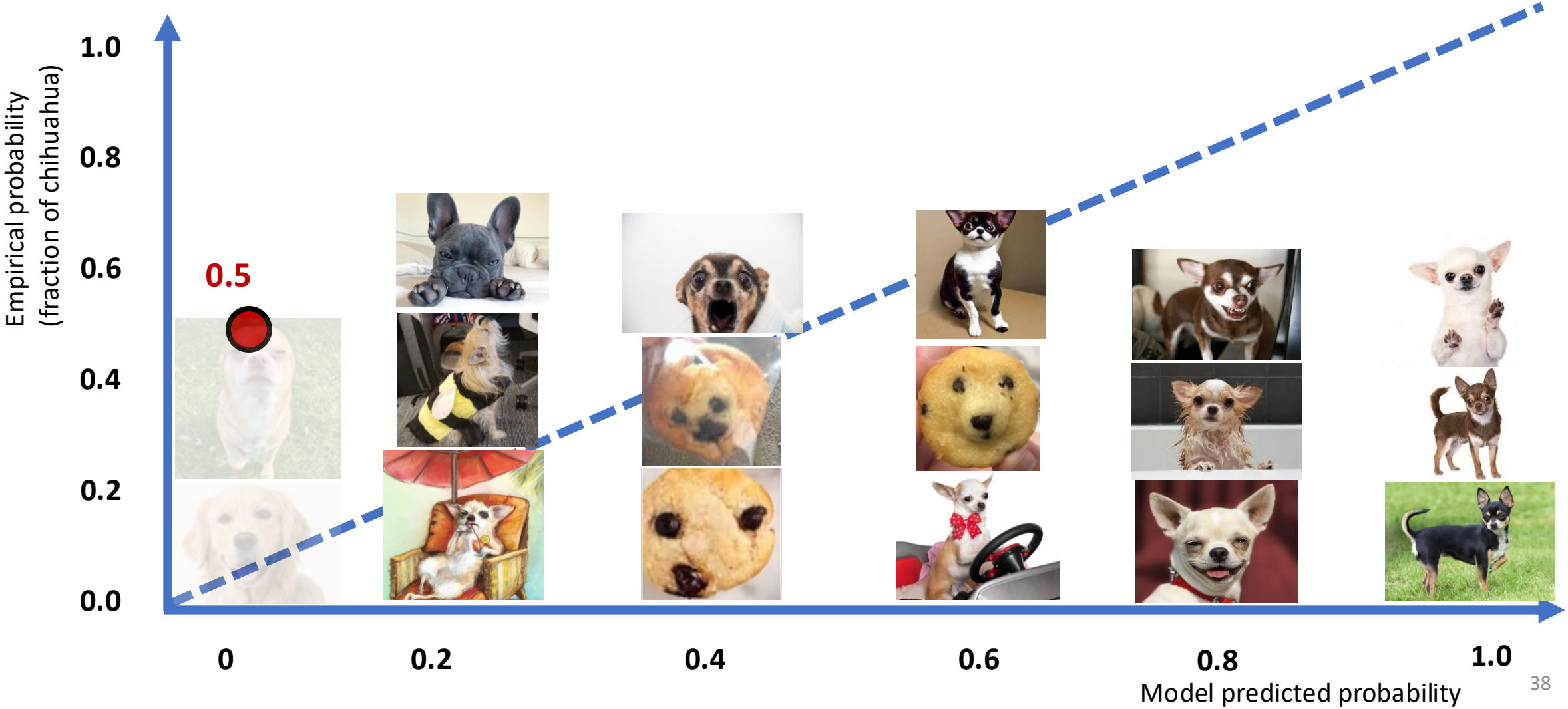
OpenSpliceAI : new concept – Calibration



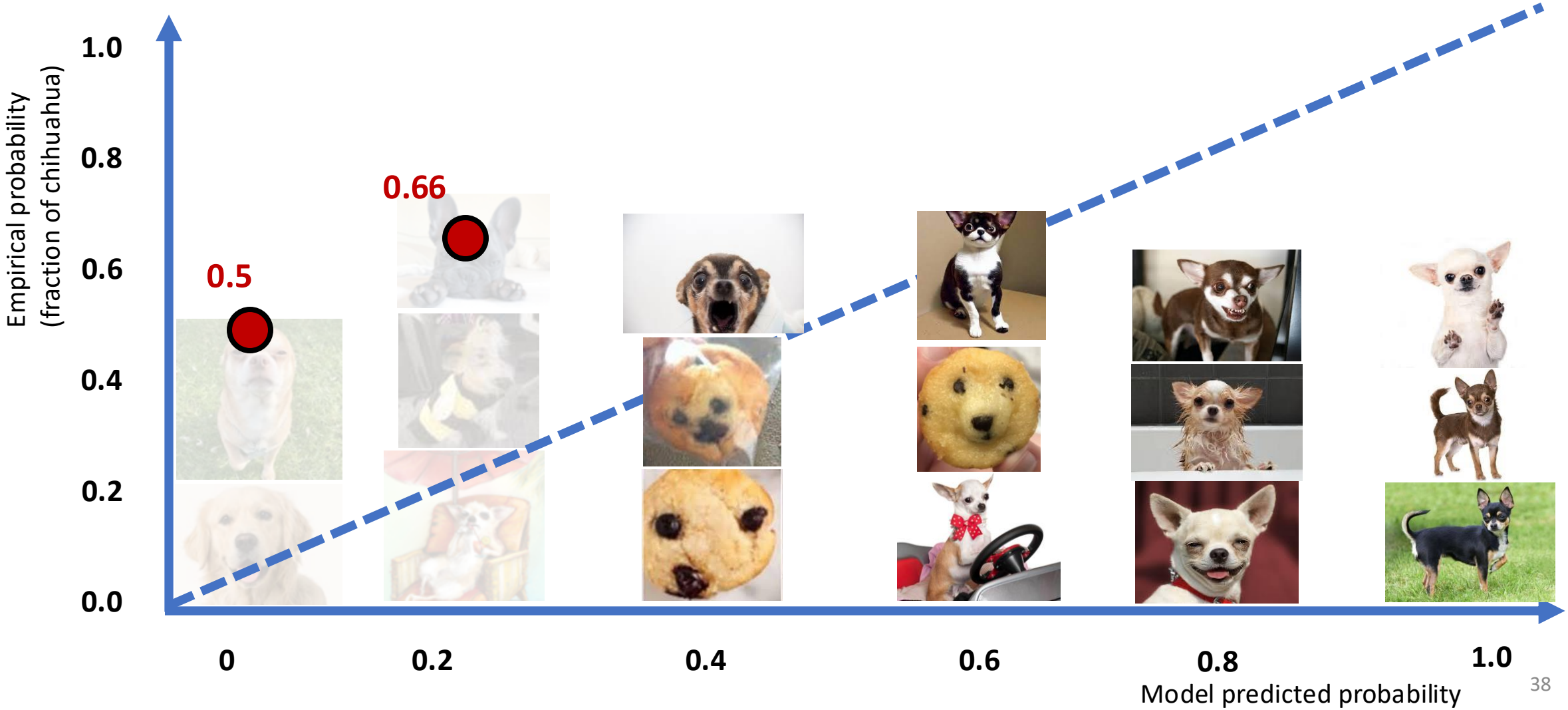
OpenSpliceAI : Calibration



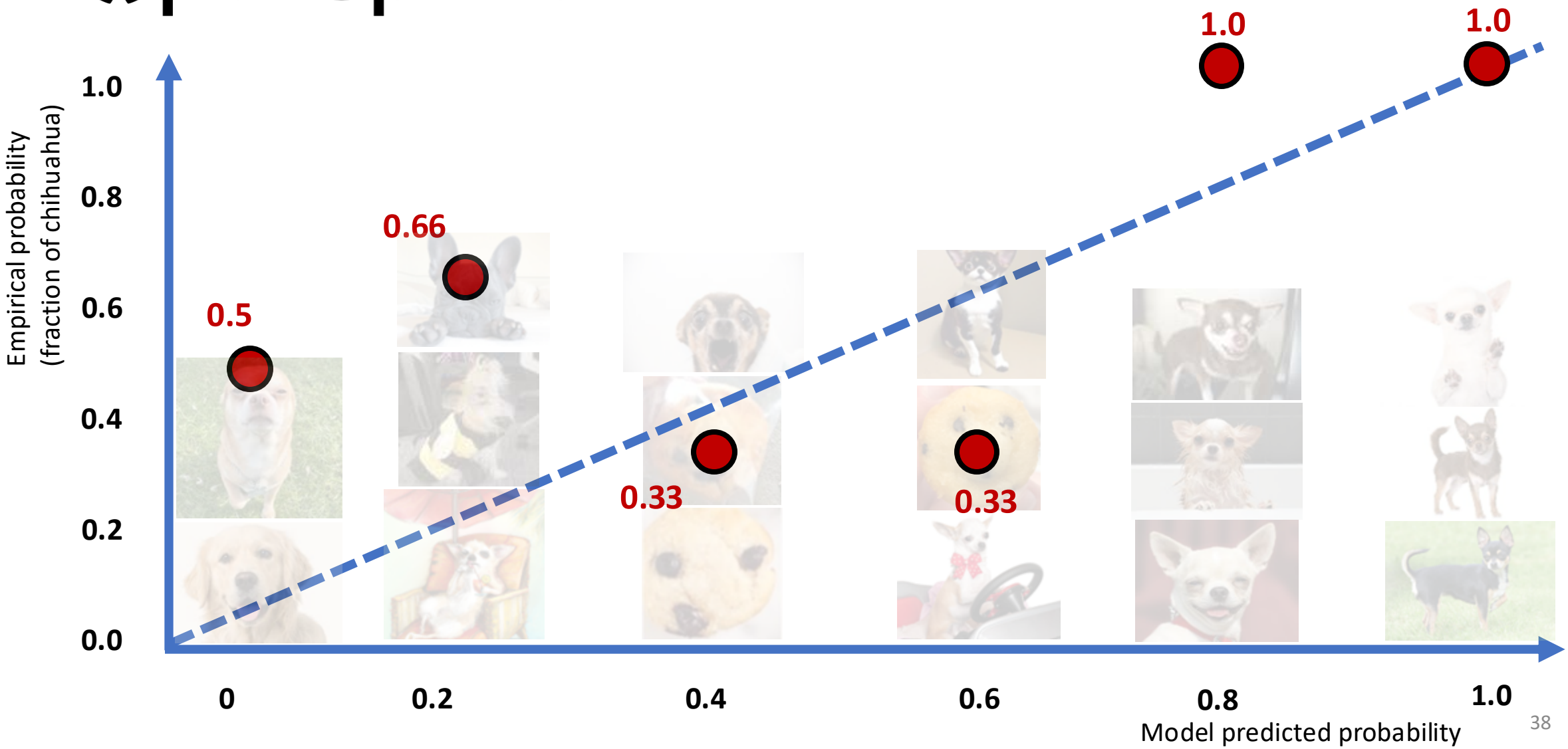
OpenSpliceAI : Calibration



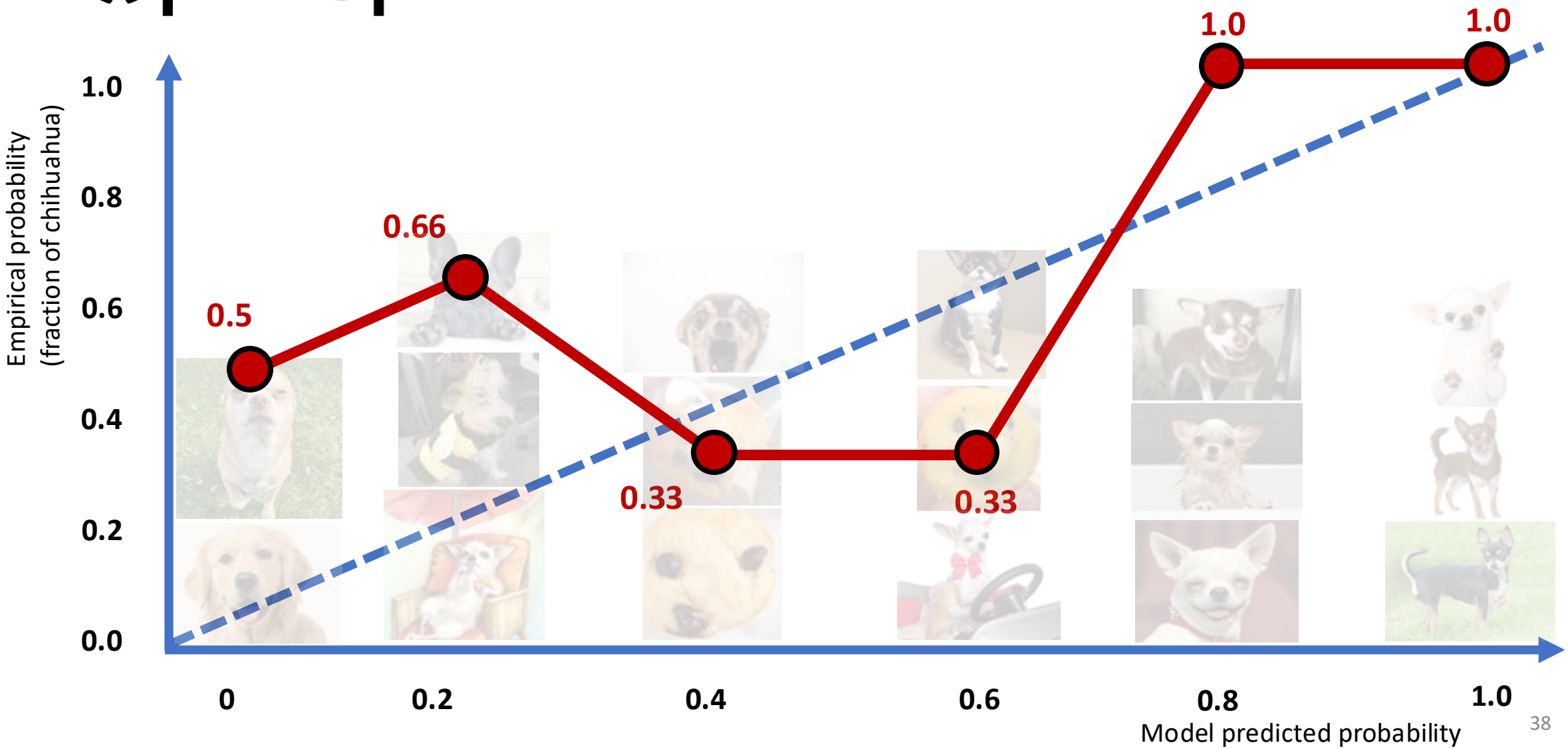
OpenSpliceAI : Calibration



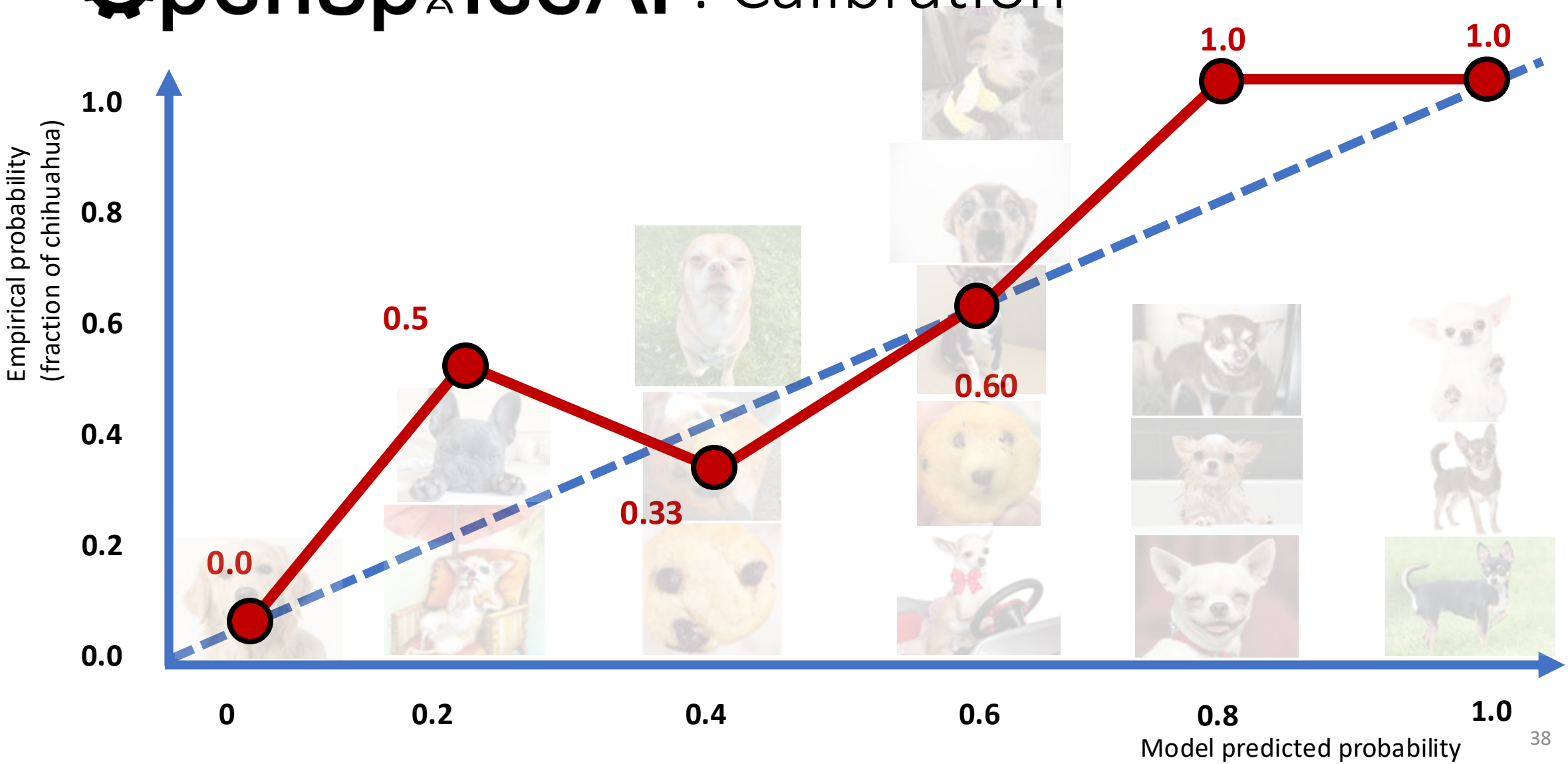
OpenSpliceAI : Calibration



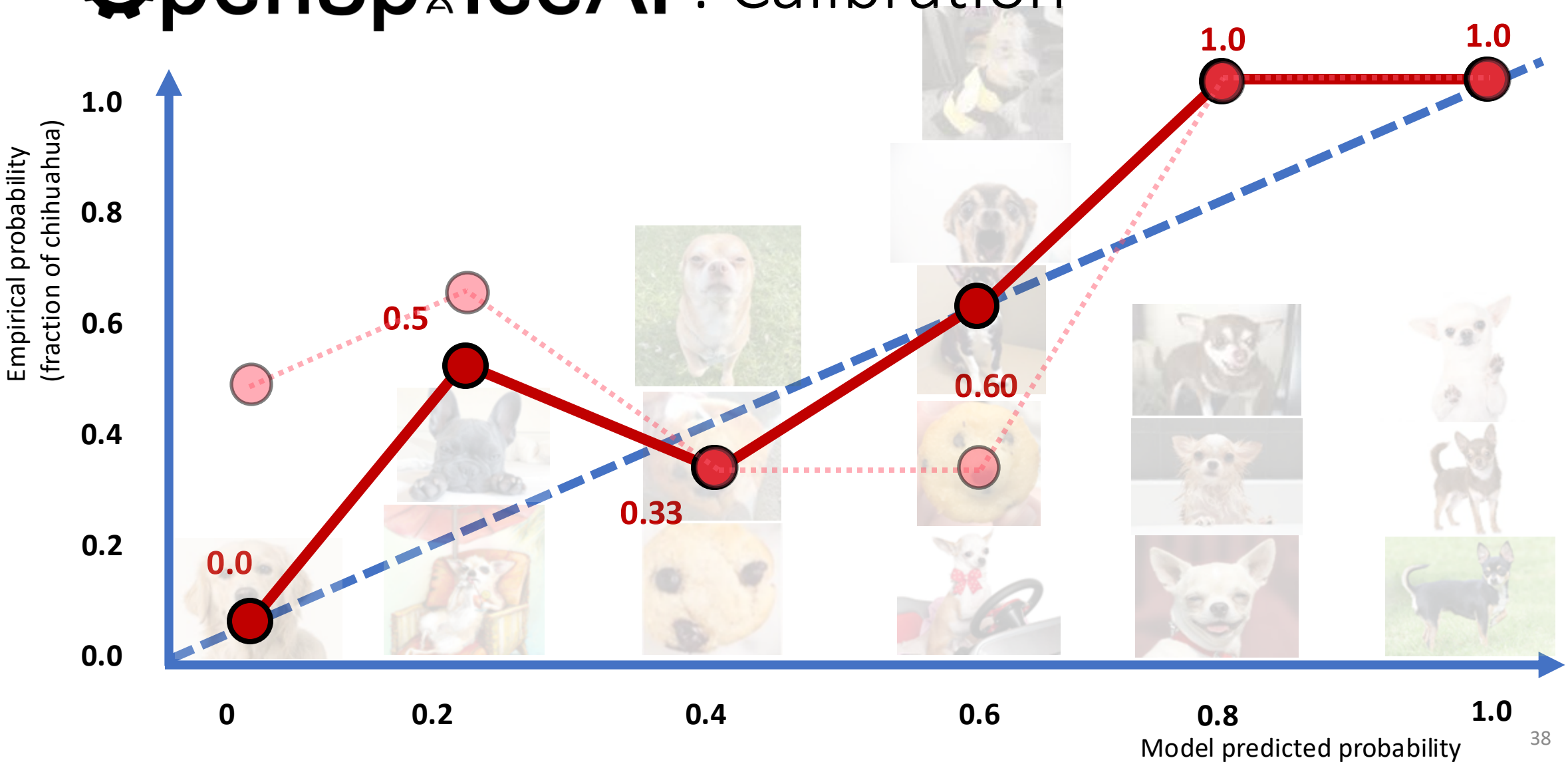
OpenSpliceAI : Calibration



OpenSpliceAI : Calibration



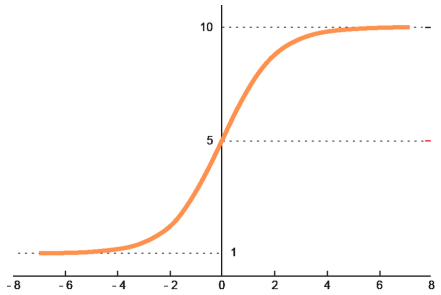
OpenSpliceAI : Calibration

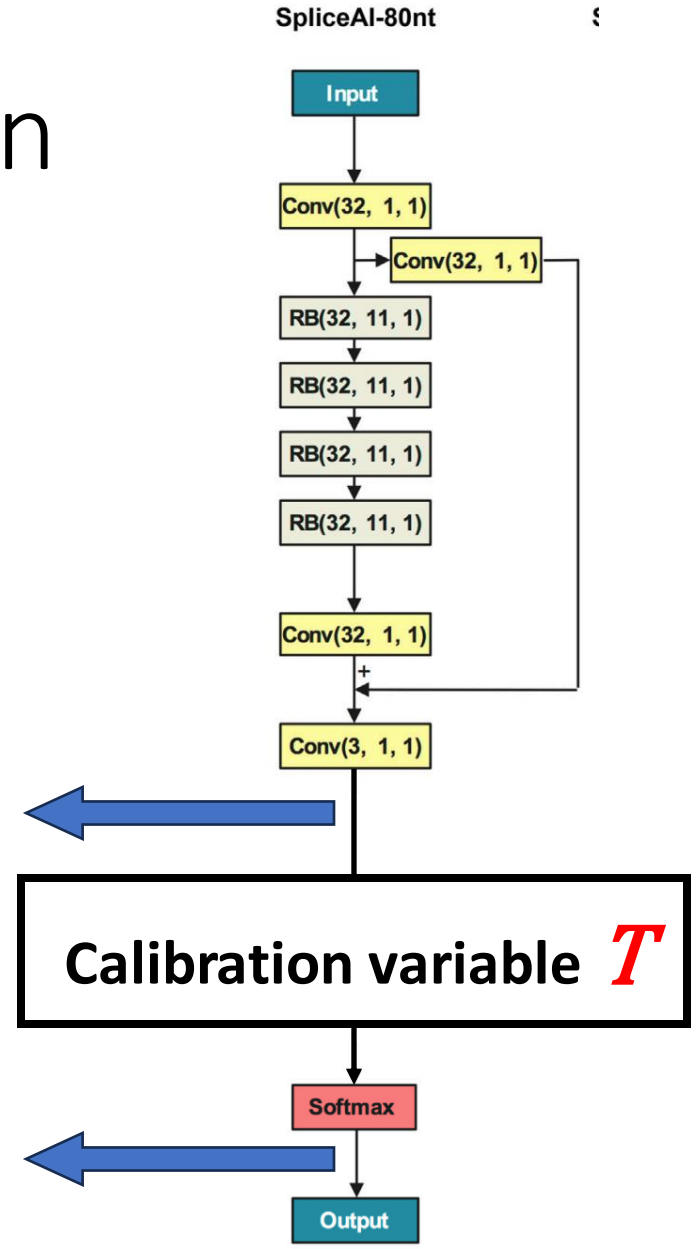


SpliceAI-toolkit : Calibration

- Platt scaling (Temperature scaling)

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





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

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

OpenSpliceAI : Calibration

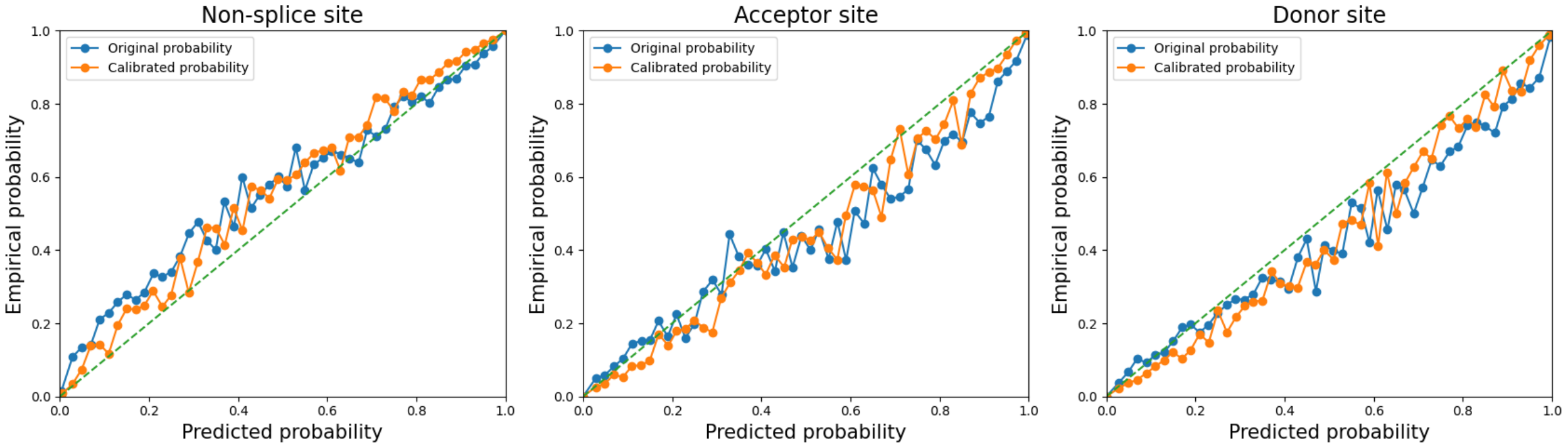
$$conf(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i \qquad acc(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} 1(\hat{y}_i = y_i)$$

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

Optimal temperature: 1.28049

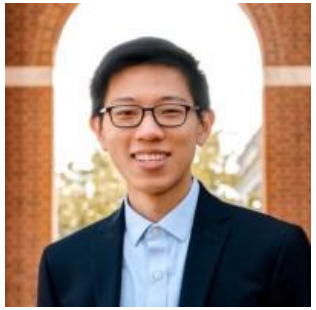
After calibration - **NLL**: 0.11896934, **ECE** : 0.00000079

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



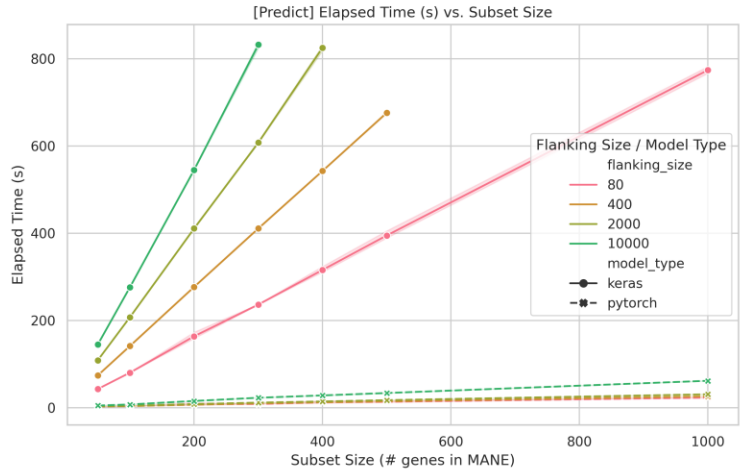
Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

OpenSpliceAI : benchmark

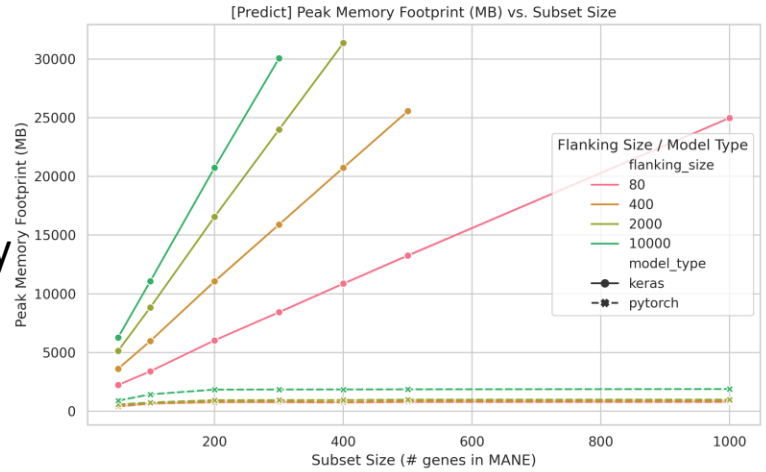


Alan Mao

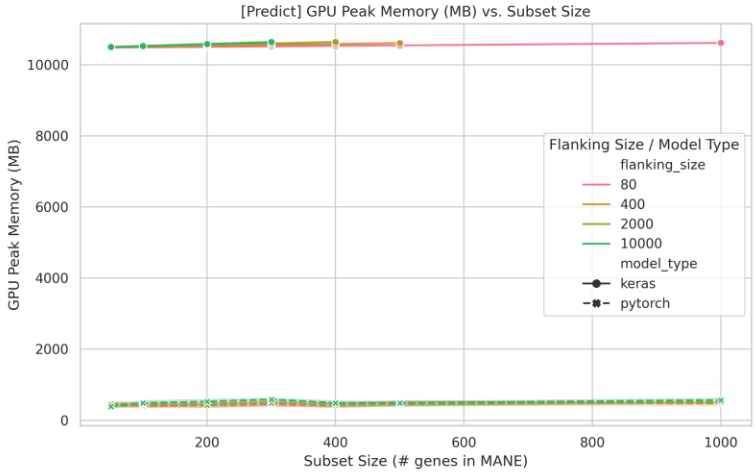
System time



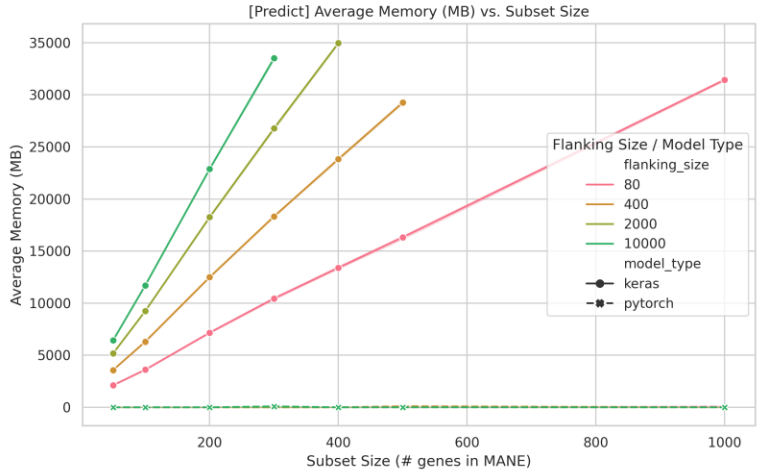
GPU
Peak memory



Memory



GPU
Avg memory



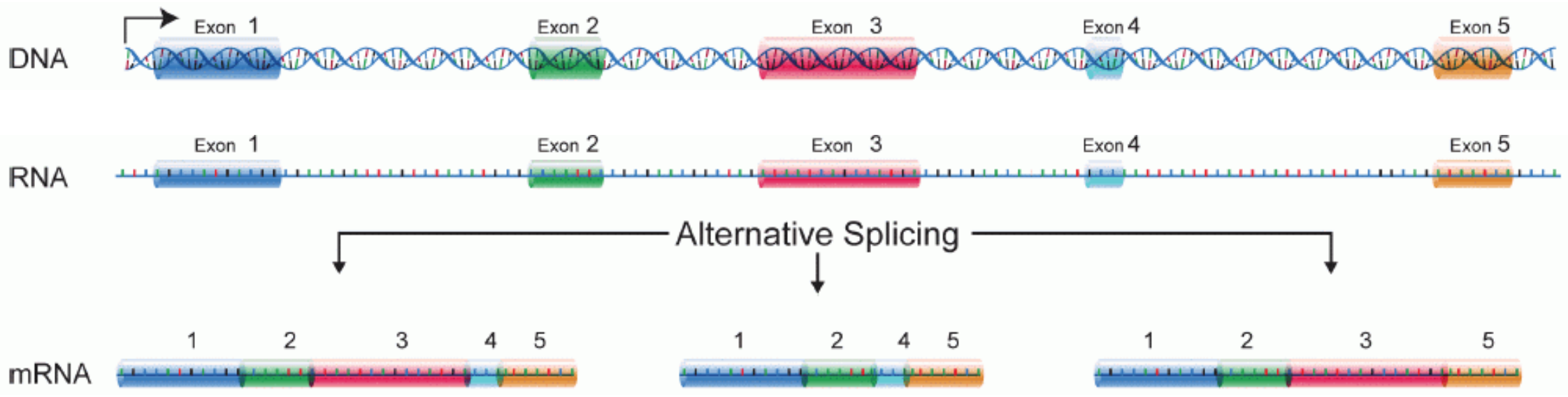
OpenSpliceAI : Summary

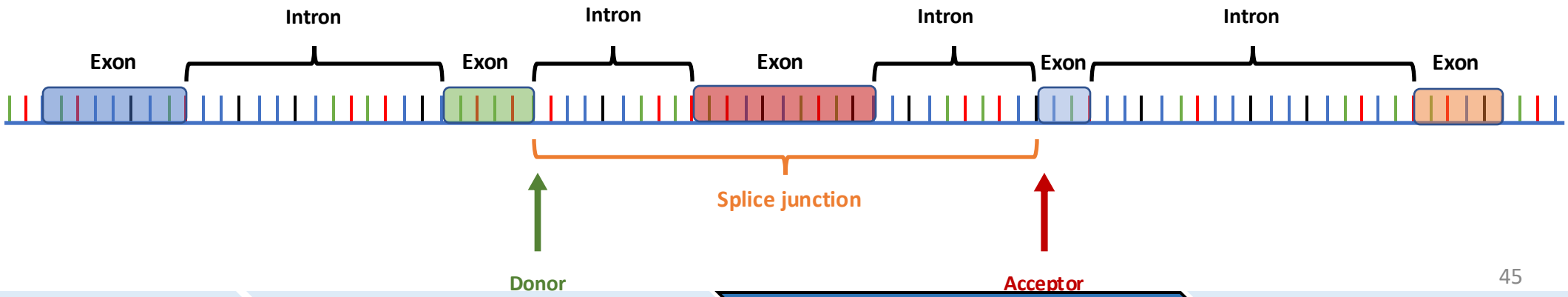
1. Data preprocessing: sliding window chunking
2. Easy-to-run framework to train your own SpliceAI
3. Pretrained SpliceAI-MANE
4. Pretrained SpliceAI on different species
5. Predict genetic variants' effect on splice sites
6. Model calibration: temperature scaling

Problem solved ?

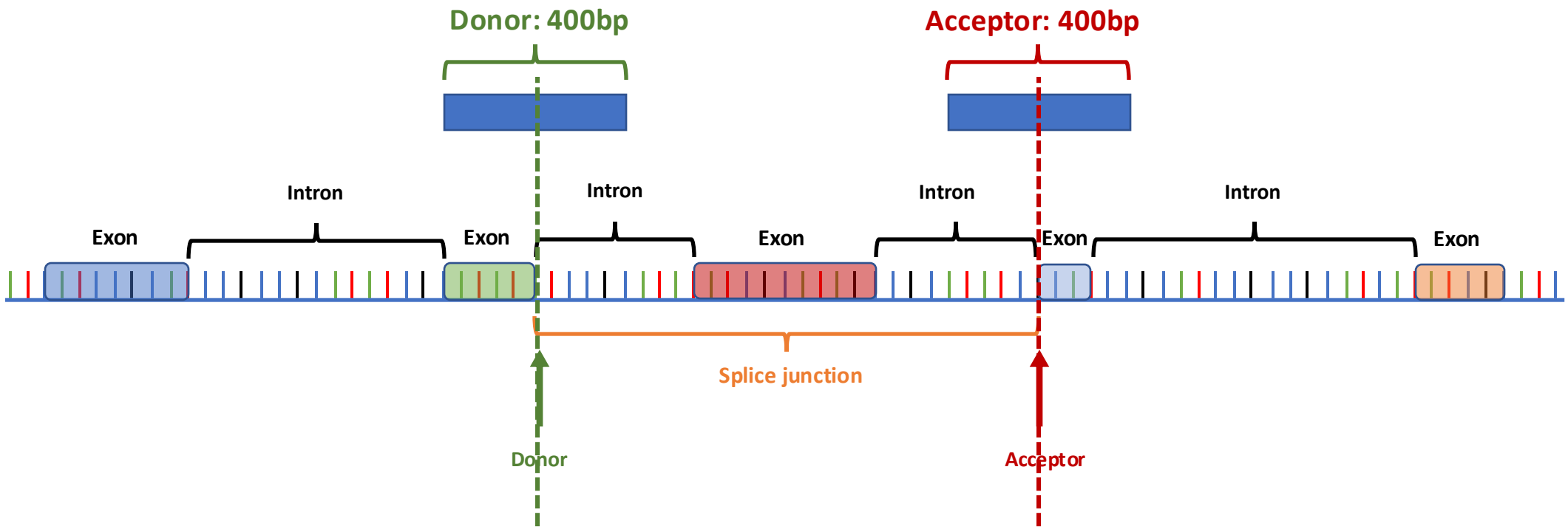
Chao, K. H., Mao, A., Liu, Anqi, Salzberg, S. L., & Pertea, M. (**2024**). OpenSpliceAI. Manuscript in preparation. https://ccb.jhu.edu/spliceai-toolkit/ **(in preparation)**

Is canonical labelling approach correct?



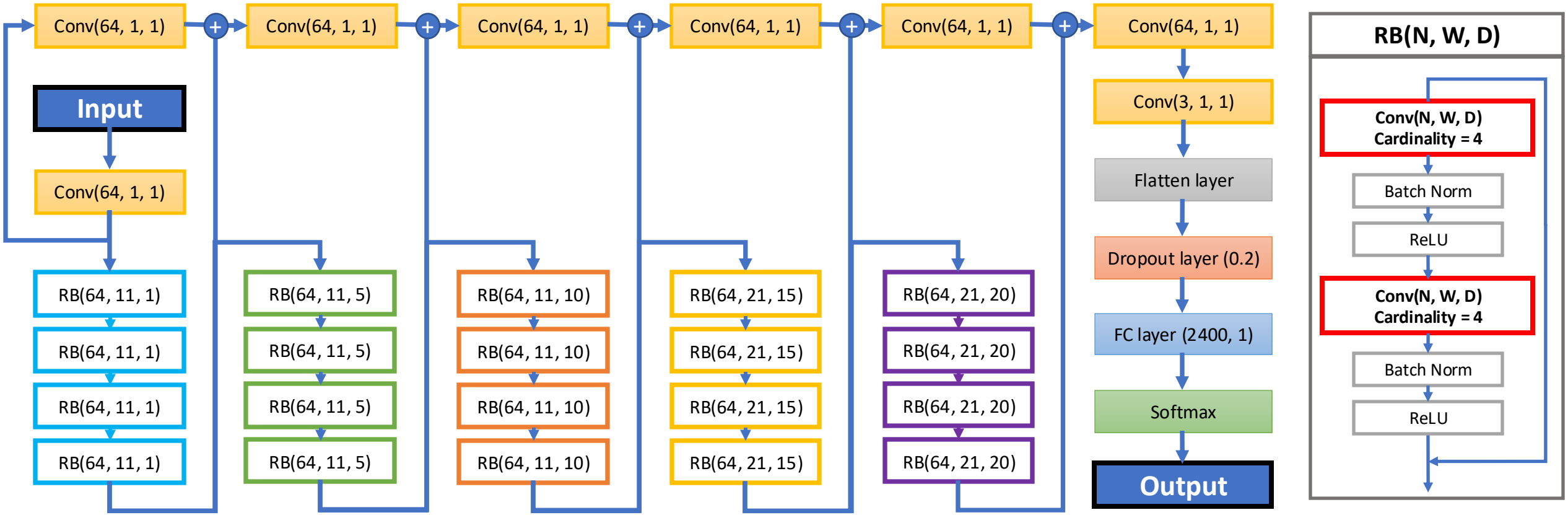


SPLAM : Data Processing



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

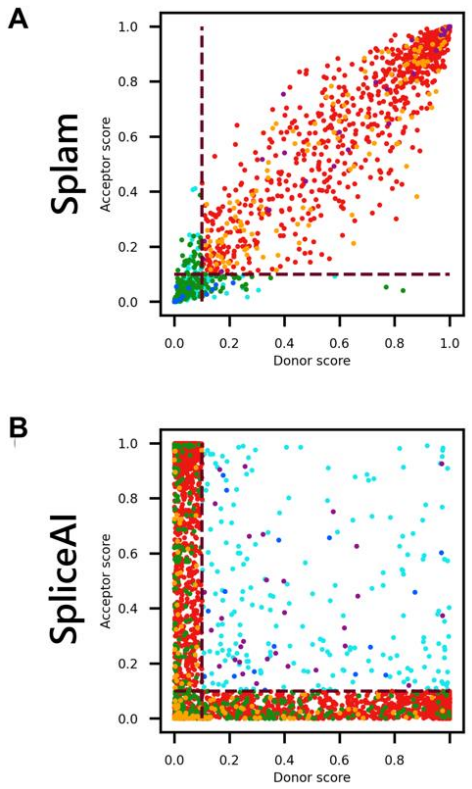
SPLAM : Splam Model Architecture



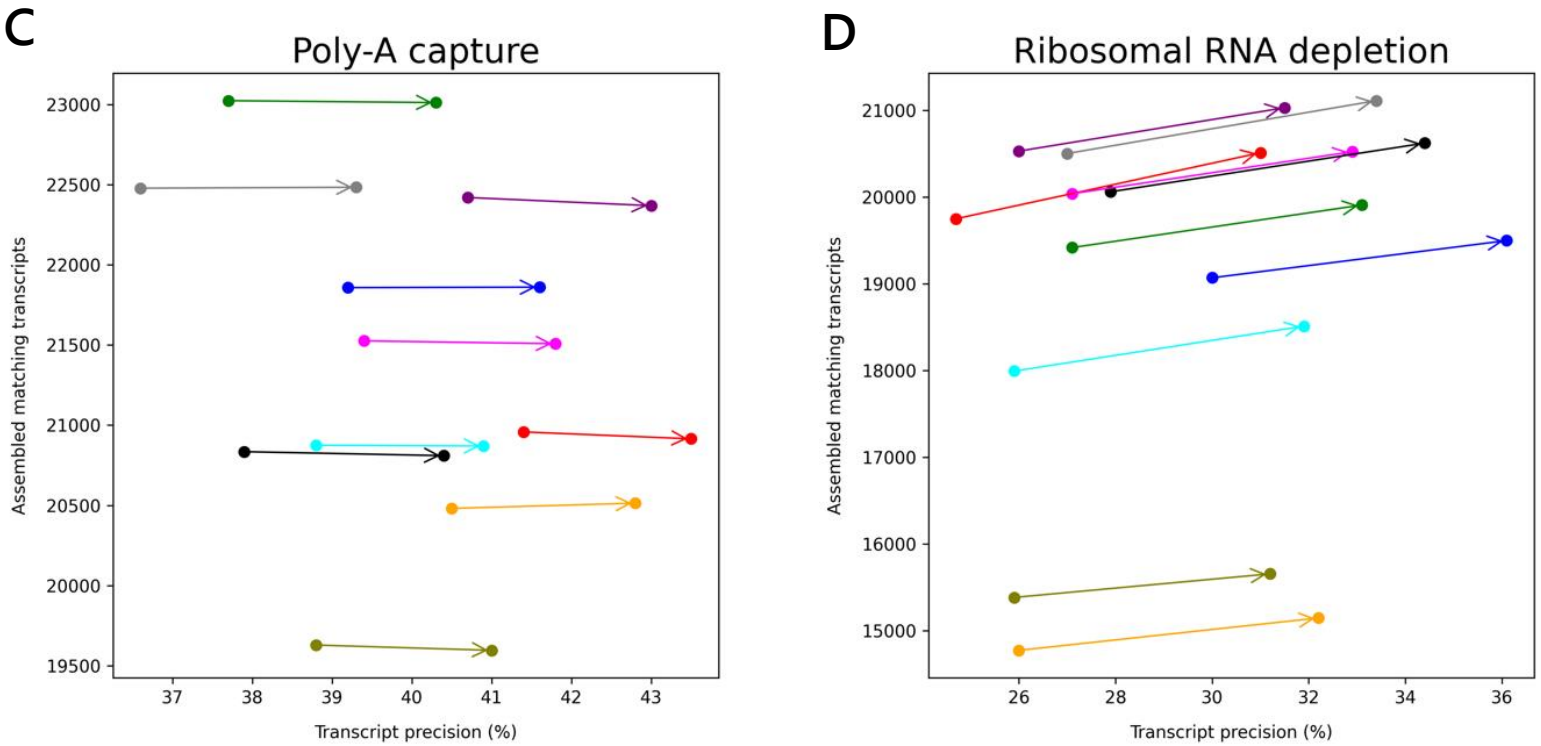
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. ***Genome Biology*** **in press.** **https://ccb.jhu.edu/splam/**

SPLAM : deep-learning splice site predictor

Score stability



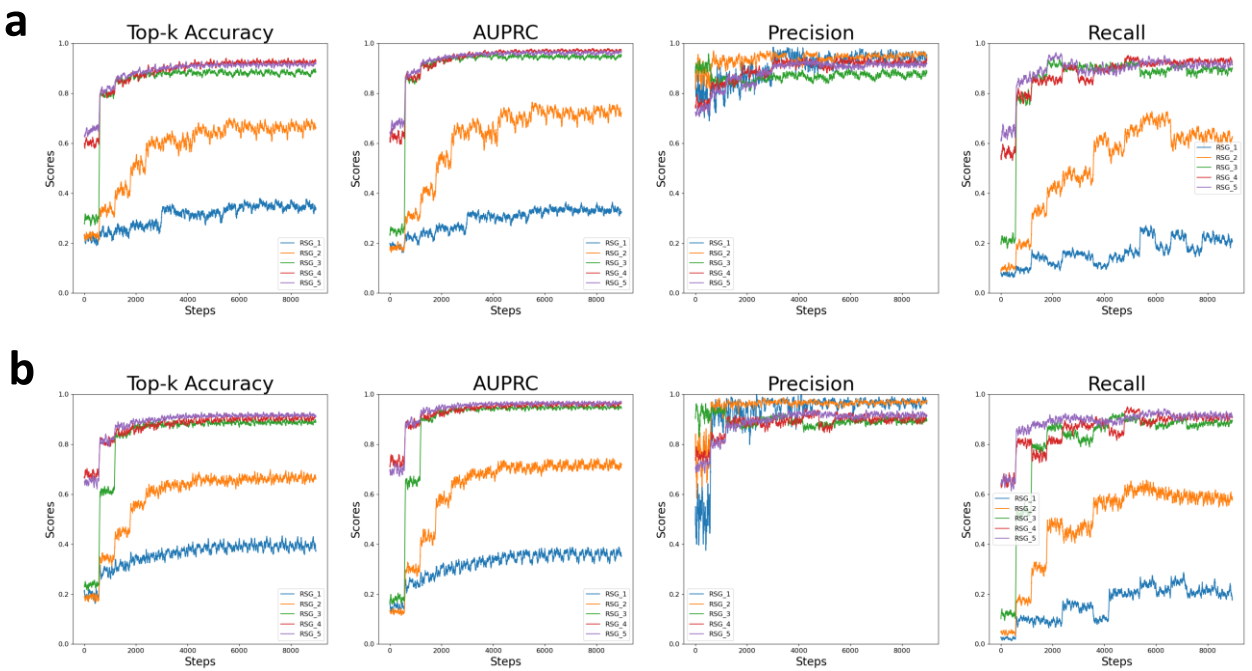
Transcriptome assembly improvement



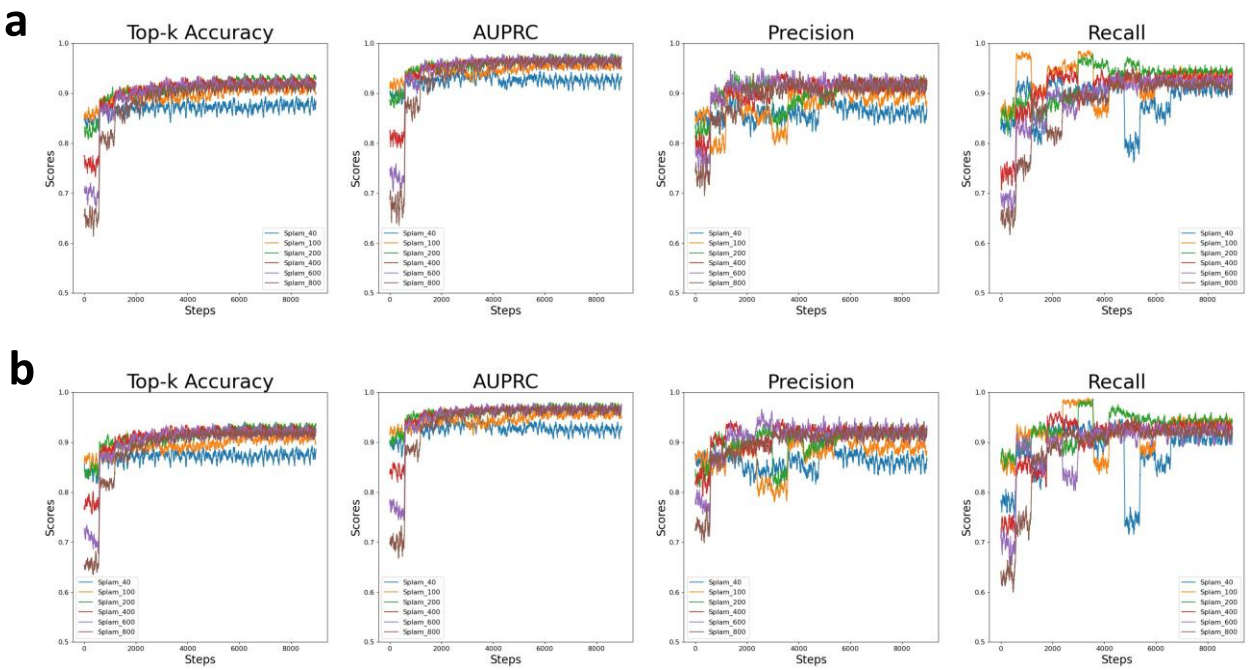
Chao, K. H., Mao, A., Salzberg, S. L., & Pertea, M. (2023). Splam: a deep-learning-based splice site predictor that improves spliced alignments. ***Genome Biology*** **in press.** **https://ccb.jhu.edu/splam/**

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. ***Genome Biology*** **in press.** **https://ccb.jhu.edu/splam/**

SPLAM : Summary

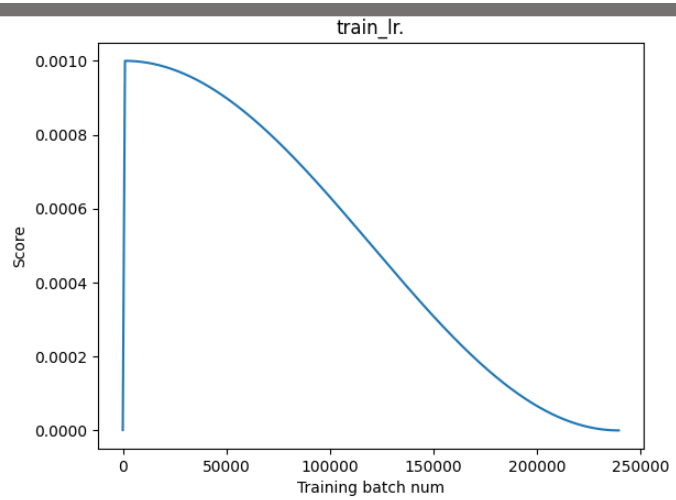
- Better than Splic
- Generalize to nc

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

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

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?

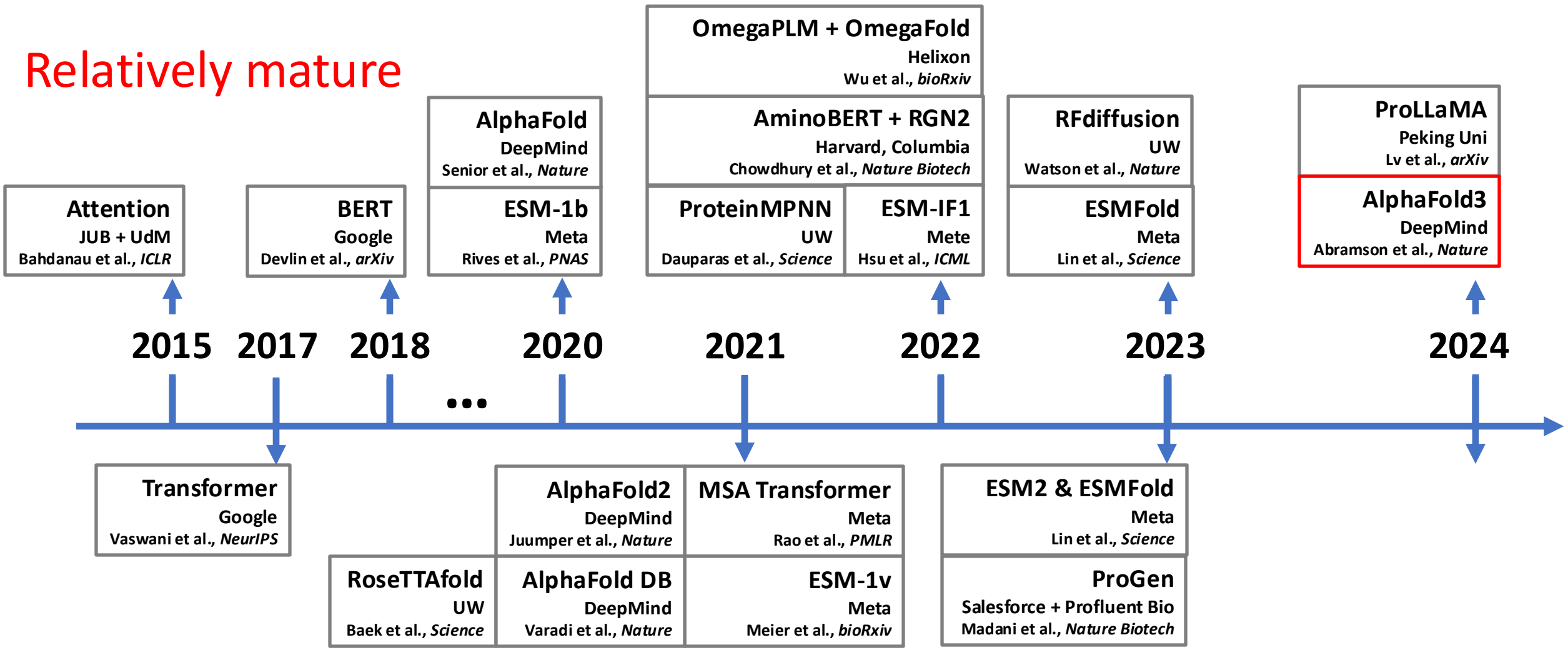
CNN or ?

Future sequence models in genomics?

CNN or **/and Transformer**?

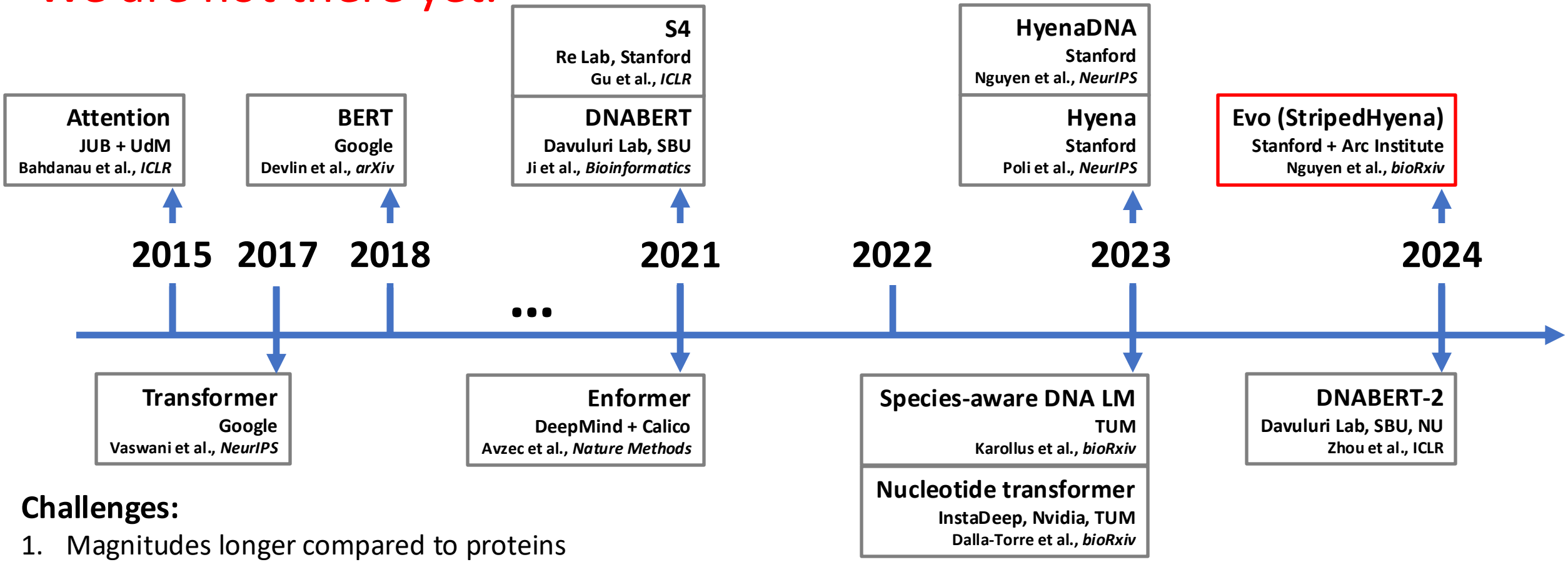
Future? – Protein transformer-based models

Relatively mature



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

Part III

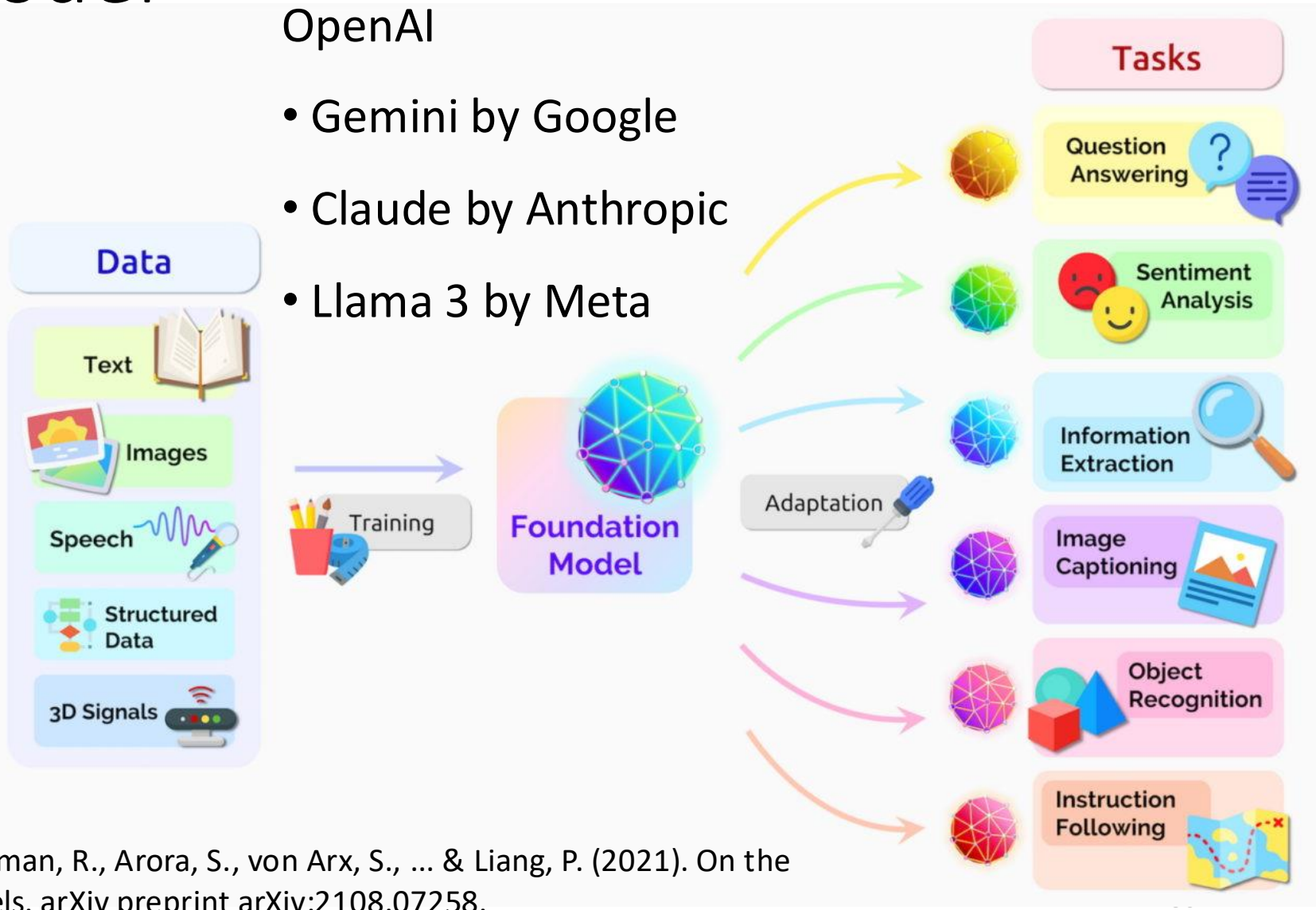
Fungi Language Model +

Gene Expression Prediction

Foundation model

- Stanford researchers called transformers “foundation models” in an August 2021 paper because they see them driving a paradigm shift in AI.

- GPT-3, GPT-4 by OpenAI
- Gemini by Google
- Claude by Anthropic
- Llama 3 by Meta

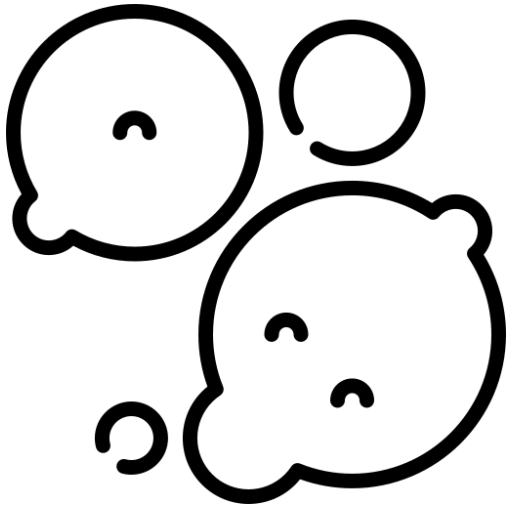


Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Foundation model

- **Versatility**: wide range of downstream tasks
- **Transfer learning**: learn general representation of data. Task-specific is limited
- **Efficiency**: computational efficiency of fine-tuning models
- **Generalization**: “zero-shot” or “few-shot”
- **Emergent abilities**:
  - basic arithmetic
  - simple programming tasks
  - summarization, translation, or question-answering.

Goals



- Interpretable fungi LLM:
  - Strains / Species diversity
  - coding vs noncoding regions
  - repeat regions
- Predicting ChIP-exo, histone marks, and RNA-Seq
- Does fine-tuning a pretrained language model outperform training a new model from scratch?

Rossi, M. J., Kuntala, P. K., Lai, W. K., Yamada, N., Badjatia, N., Mittal, C., ... & Pugh, B. F. (2021). A high-resolution protein architecture of the budding yeast genome. *Nature*, *592*(7853), 309-314.

Fungi Language Model

- Q1: To what evolutionary distance should we include in our LM?
- Q2: What is the quality of the annotation? Coding vs non-coding regions
- Q3: How repetitive are the genomes?

Selected Genomes for LM

Fungi diverged from other life around 1.5 billion years ago

Same species, Different strains

Order level

Kingdom level

Dataset 1

R64 reference yeast

Dataset 2

80 strains of yeasts

Dataset 3

165 Saccharomycetales

Dataset 4

More distant fungi

Q1: Diversity of strains?

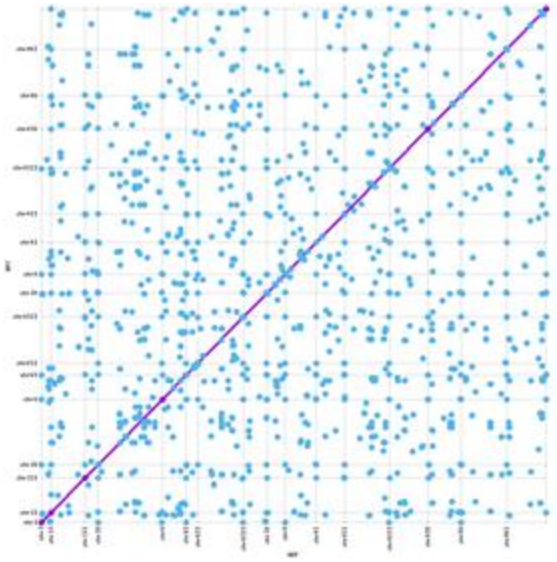
Q2: Diversity of species?

Q3: Even more diverse?

Genome evaluation – Distance between species/strains

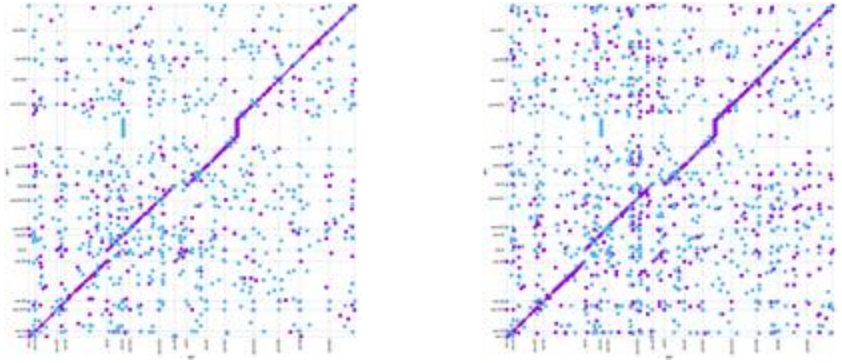
R64 Reference Yeast

Mummerplot

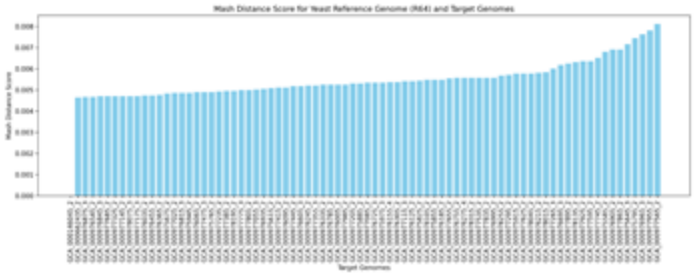


80 strains of yeasts

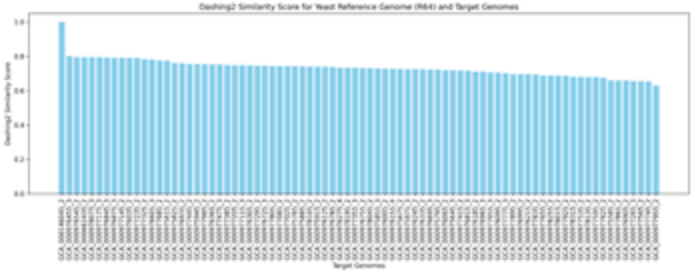
Mummerplot



Mash

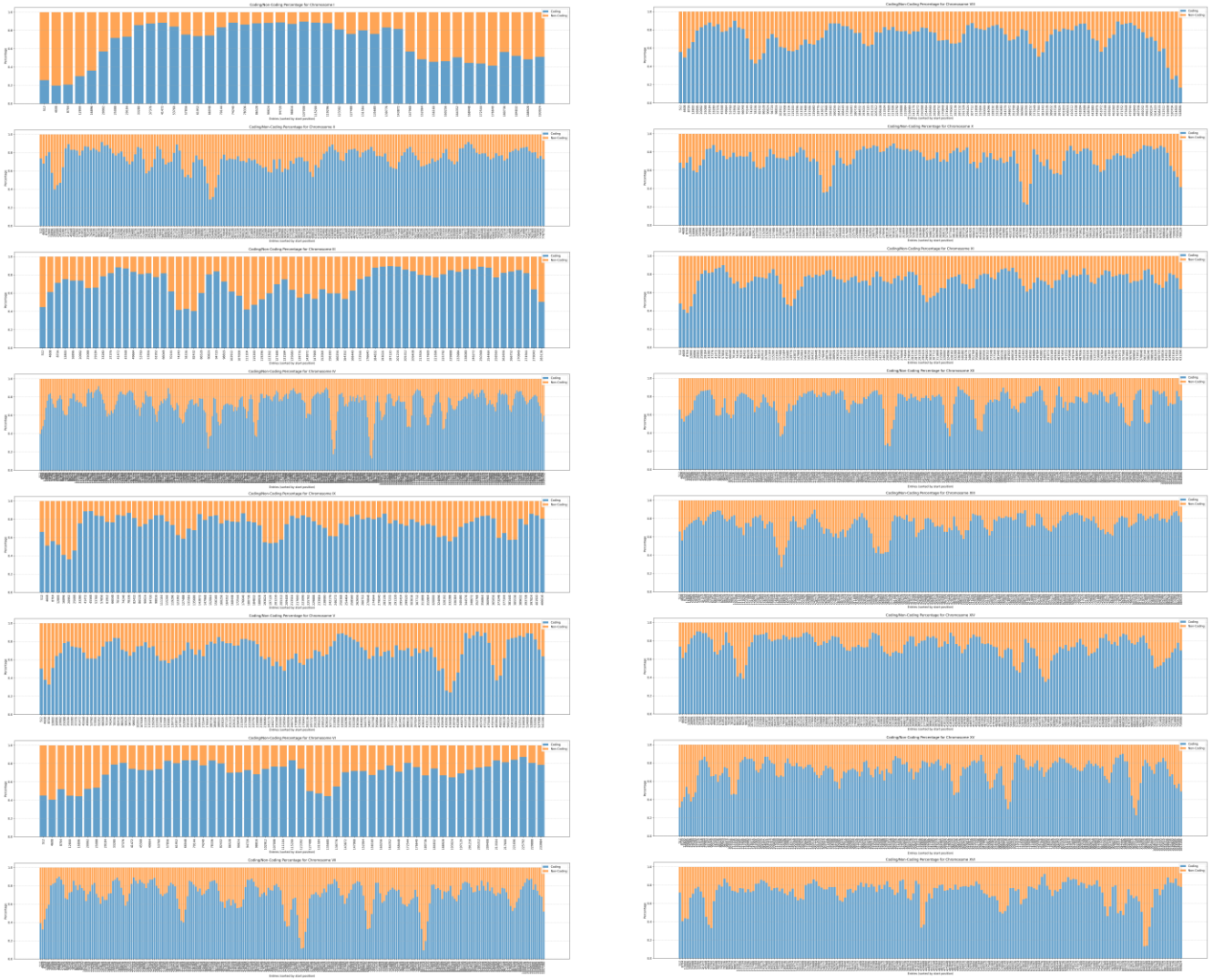


Dashing2



165 Saccharomycetales

Genome evaluation – coding / noncoding regions



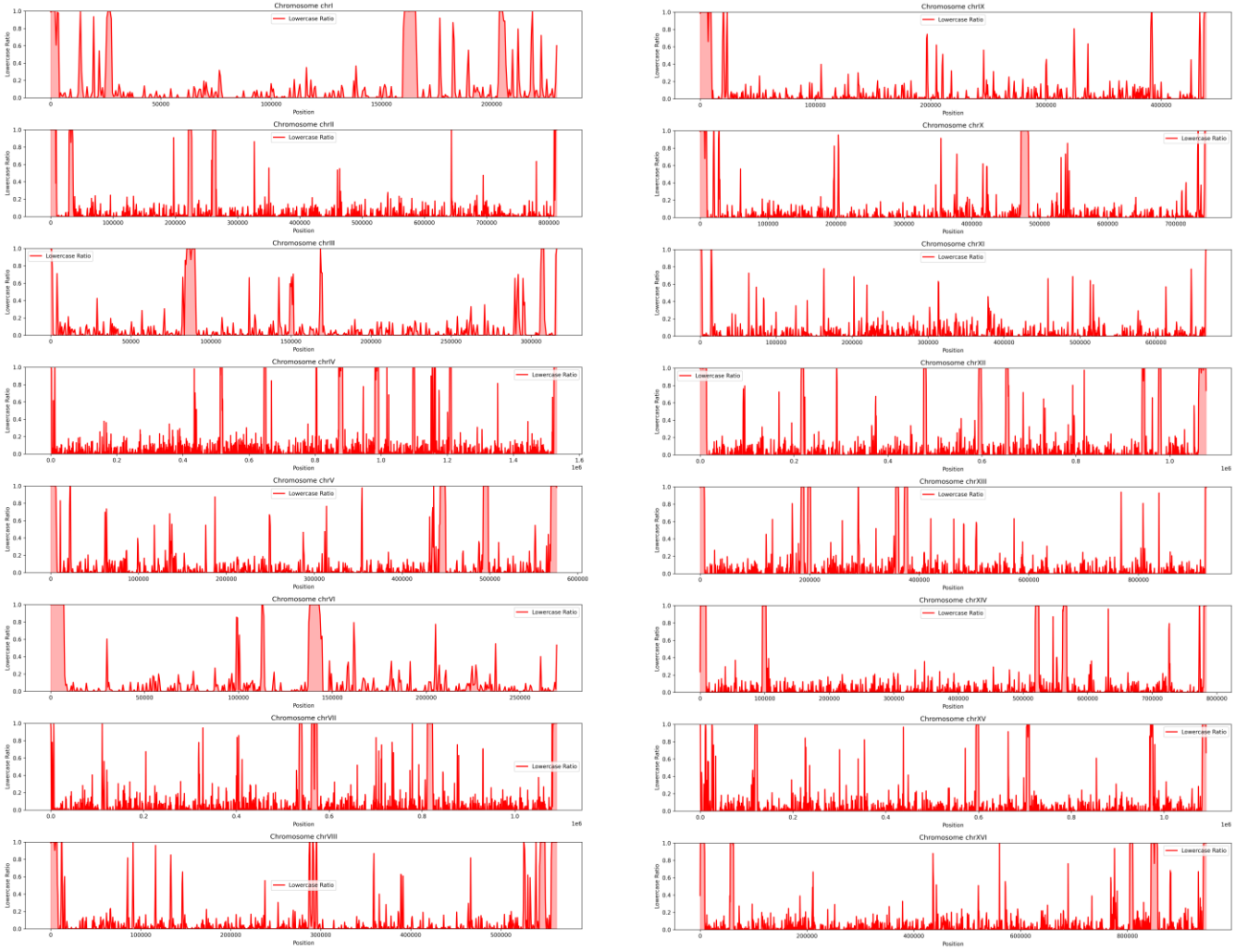
R64 Reference Yeast

Saccharomyces cerevisiae:
72.46%

Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences, 120(44), e2311219120.

Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z. Y., Miller, Z. R., ... & Kuleshov, V. (2024). Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model. bioRxiv, 2024-06.

Genome evaluation – repeat regions



R64 Reference Yeast

Saccharomyces cerevisiae:
7.39%

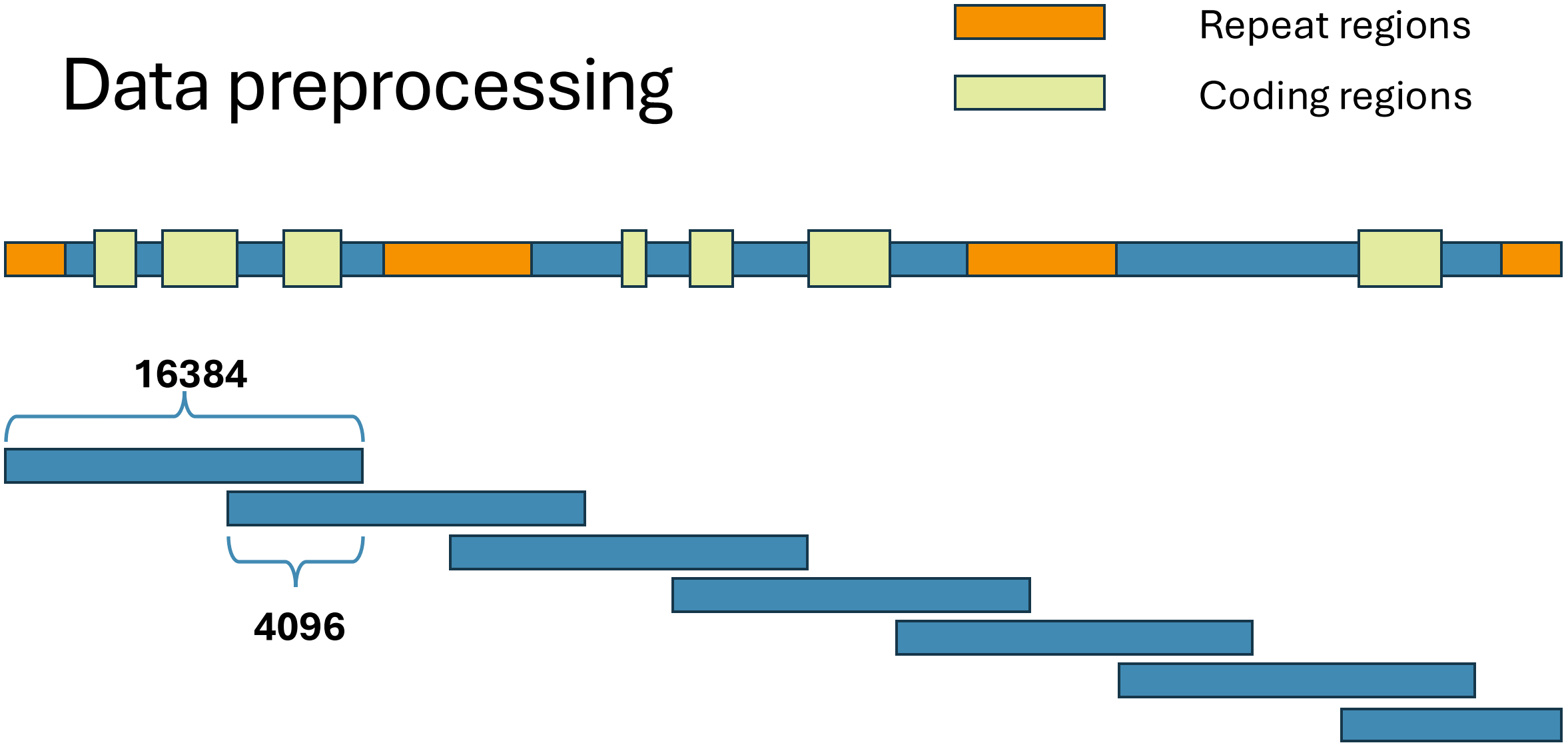
Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences, 120(44), e2311219120.

Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z. Y., Miller, Z. R., ... & Kuleshov, V. (2024). Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model. bioRxiv, 2024-06.

RepeatModeler + RepeatMasker

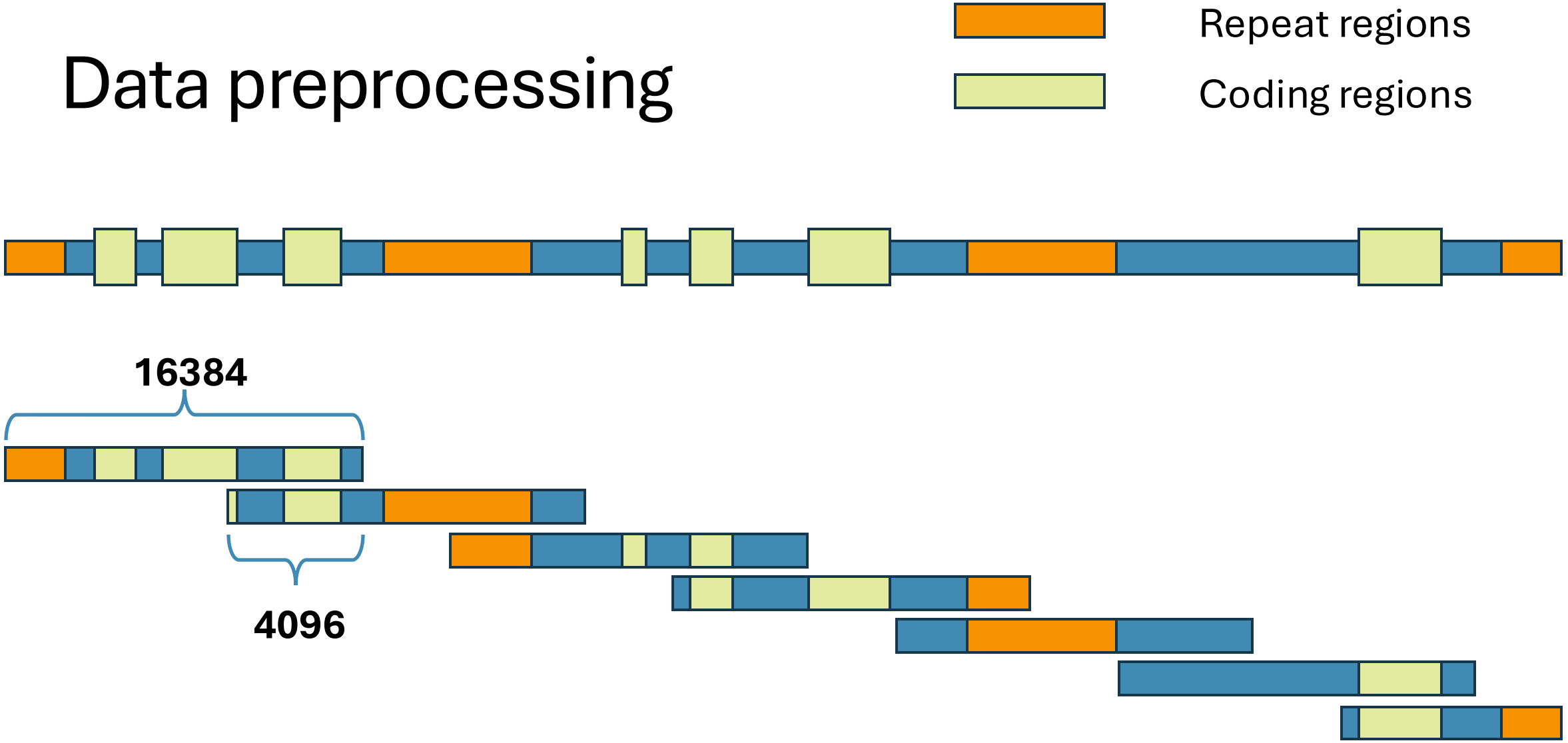
Data preprocessing

Data preprocessing



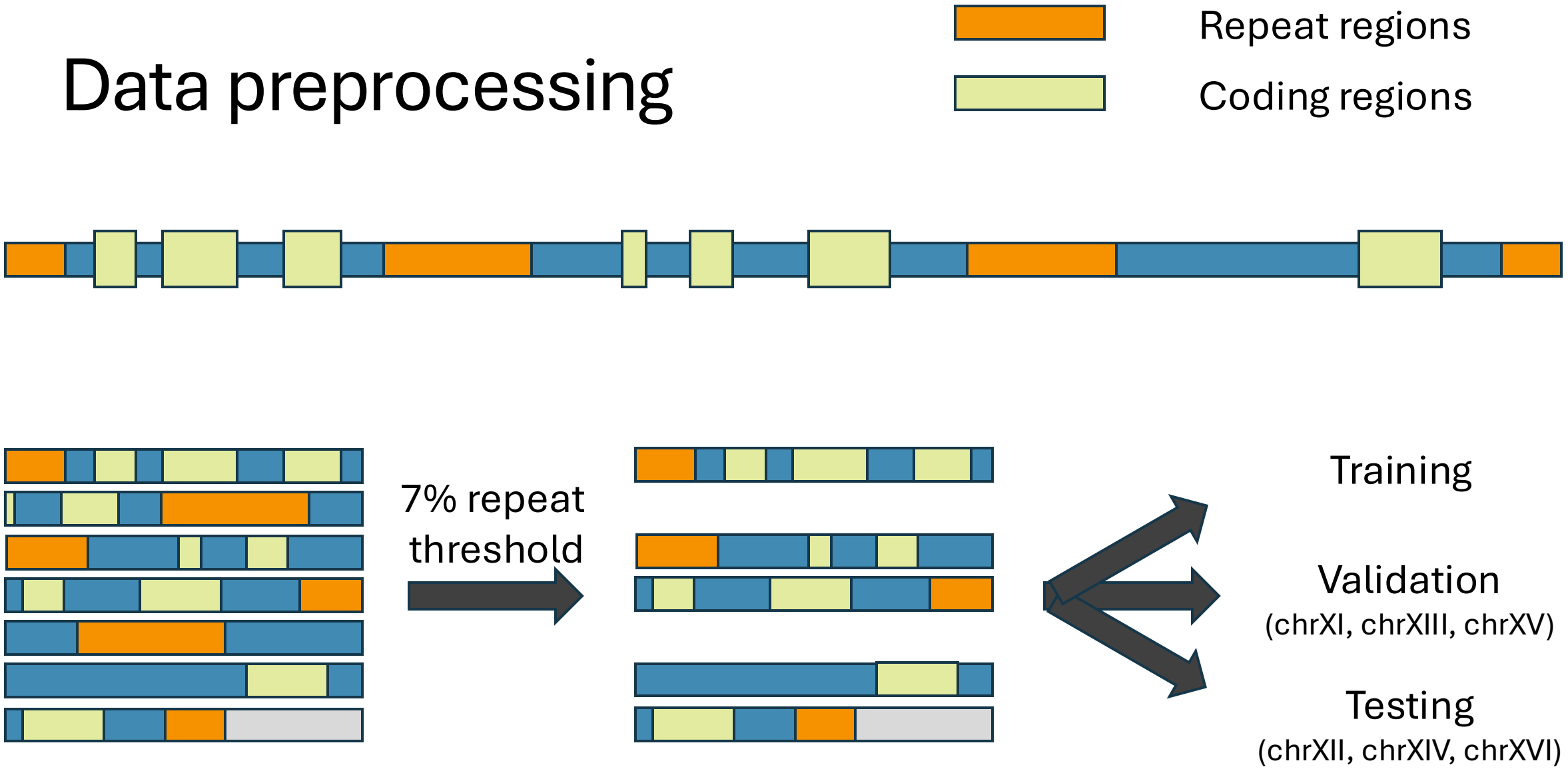
~ 7 genes per window

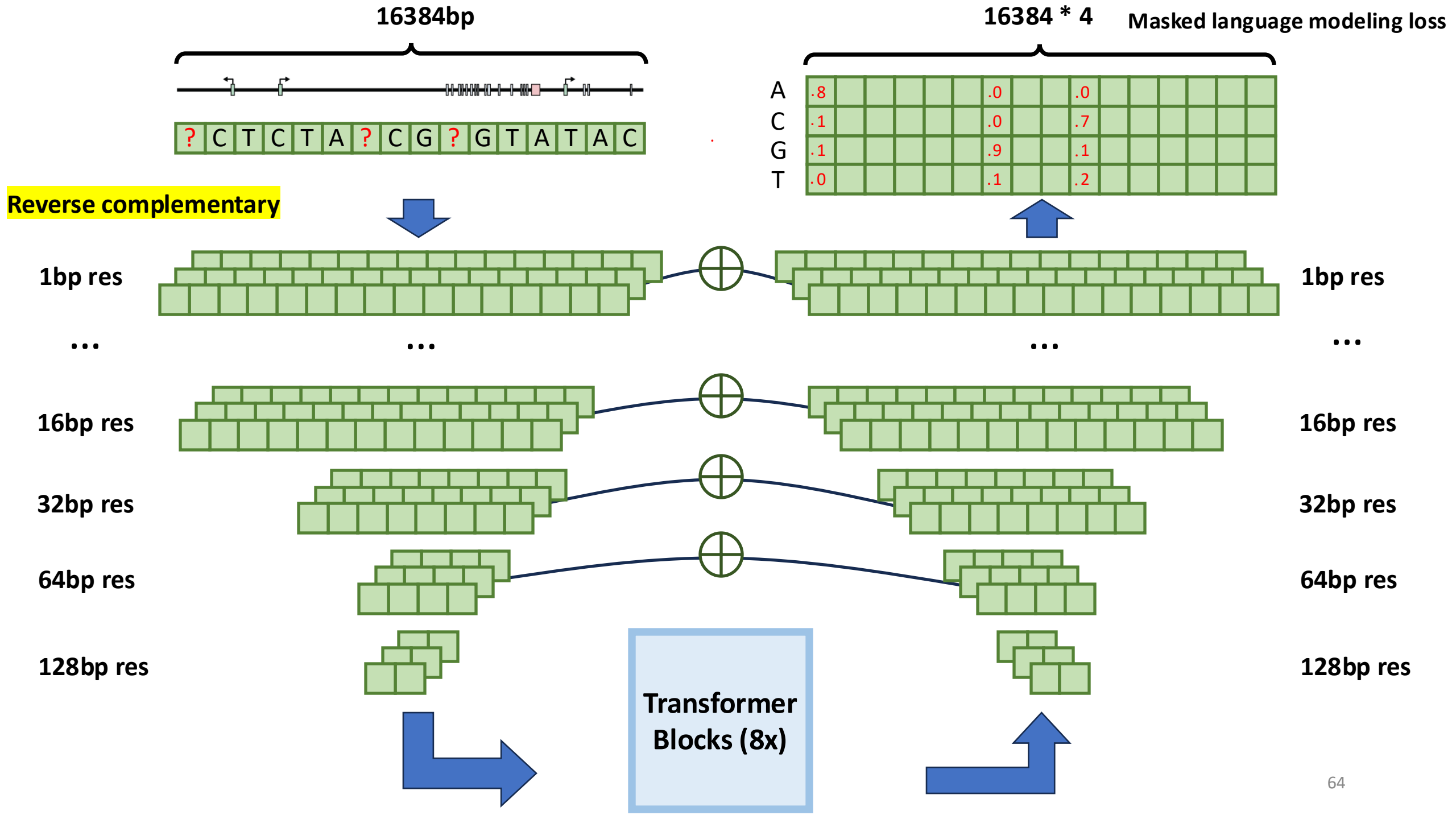
Data preprocessing



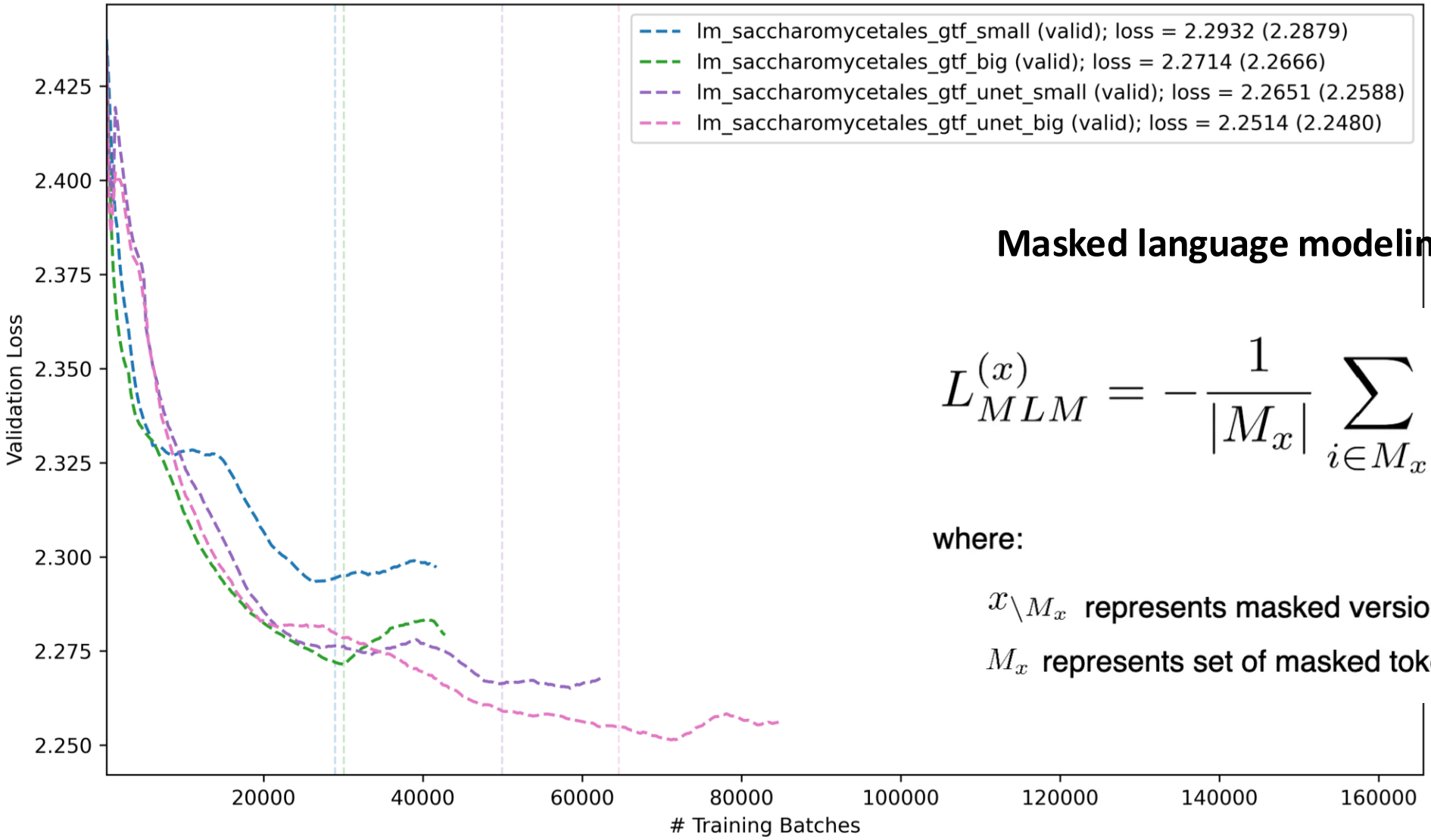
~ 7 genes per window

Data preprocessing




16384bp
16384 * 4
Masked language modeling loss
? C T C T A ? C G ? G T A T A C
A
C
G
T
.8 .0 .0
.1 .0 .7
.1 .9 .1
.0 .1 .2
Reverse complementary
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Language model performance – different models



Total params: 320,708 (1.22 MB)
Total params: 3,642,116 (13.89 MB)
Total params: 13,665,828 (52.13 MB)
Total params: 71,790,564 (273.86 MB)

Masked language modeling loss

$$L_{MLM}^{(x)} = -\frac{1}{|M_x|} \sum_{i \in M_x} logP(x_i / x_{\backslash M_x})$$

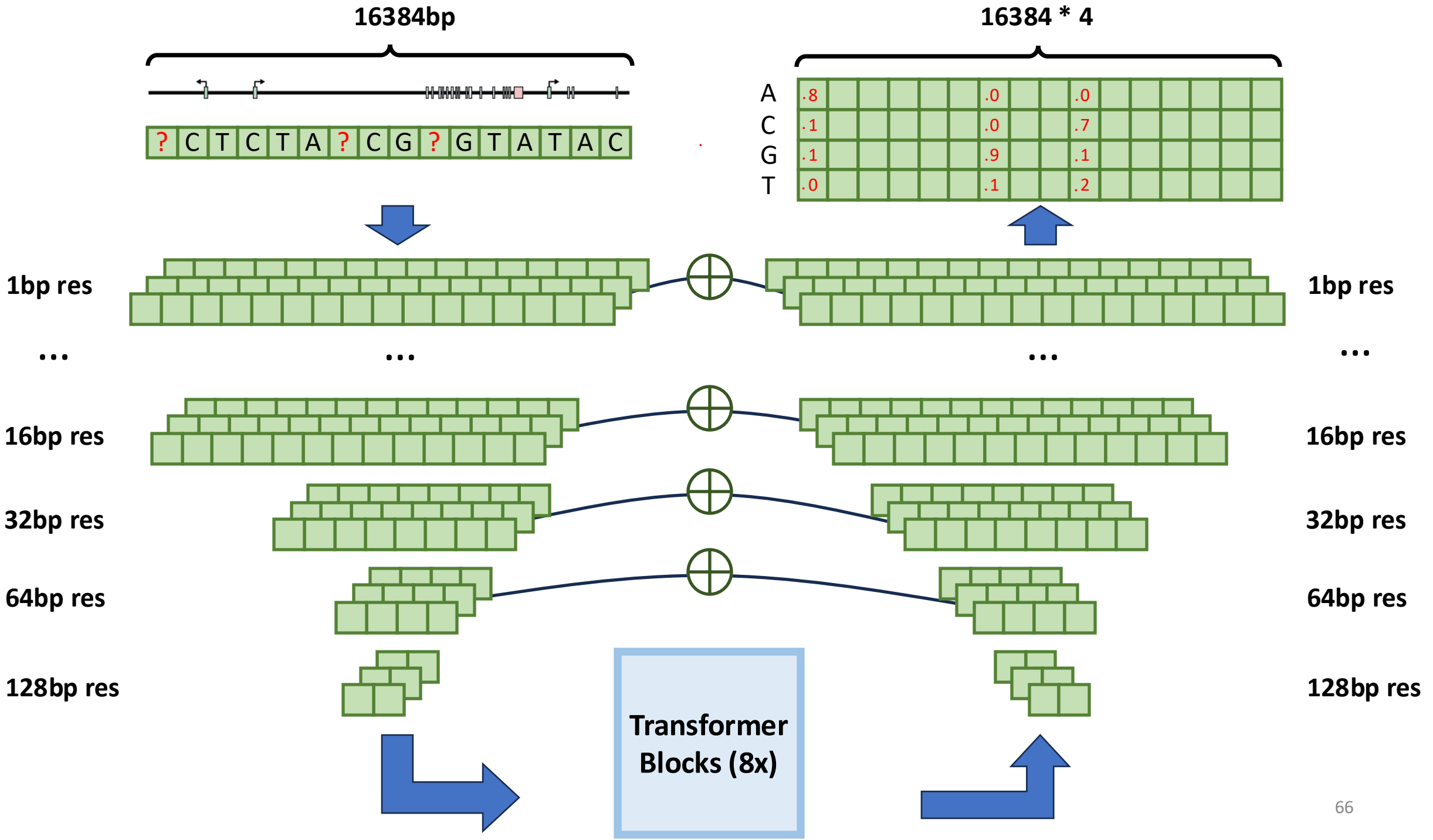
where:

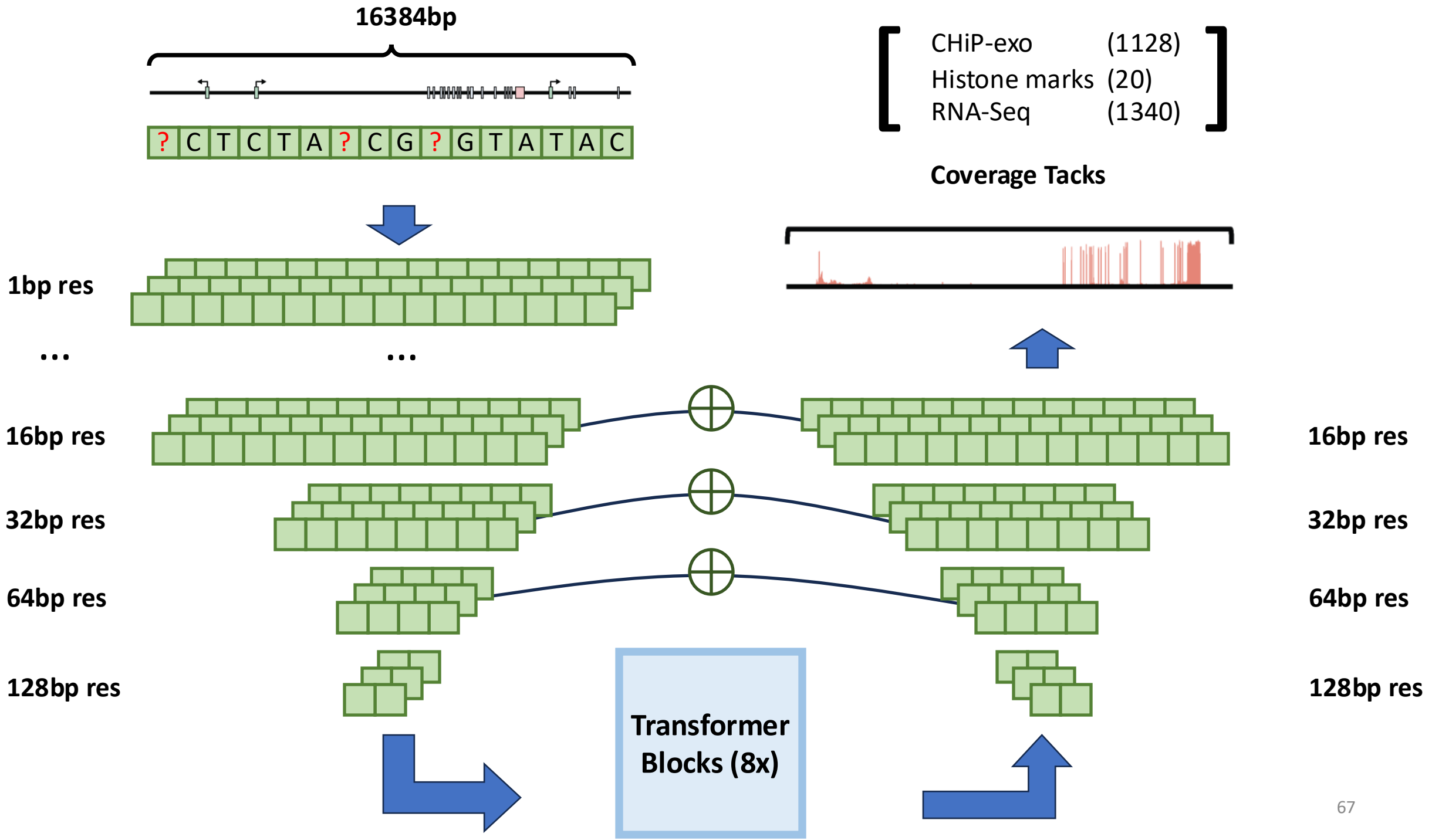
$x_{\backslash M_x}$ represents masked version of x

M_x represents set of masked token positions in x

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., & Kelley, D. R. (2023). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Biorxiv, 2023-08.

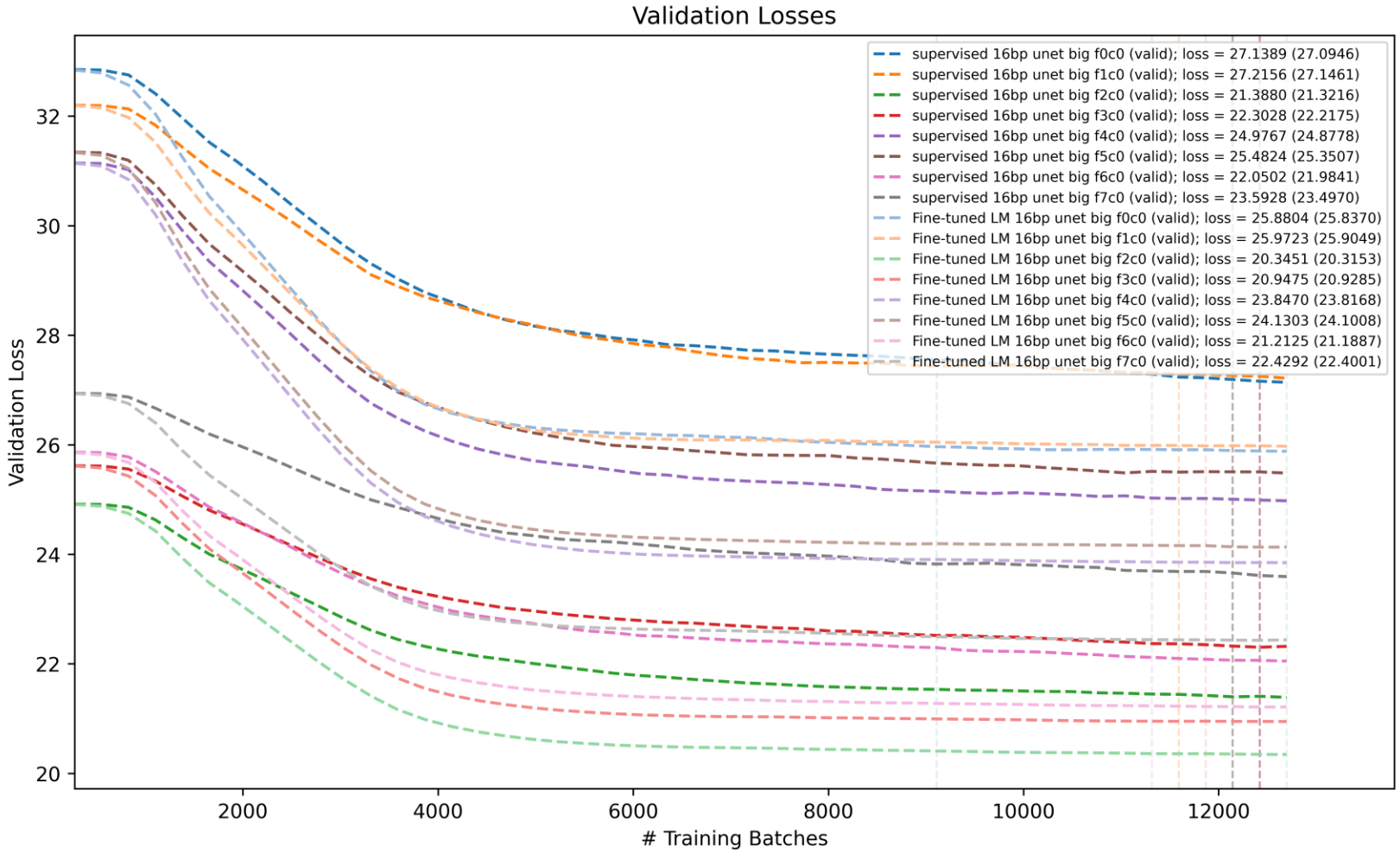
Supervised ChiP-exo, histone marks, RNA-Seq prediction


16384bp
16384 * 4
? C T C T A ? C G ? G T A T A C
A
C
G
T
.8 .0 .0
.1 .0 .7
.1 .9 .1
.0 .1 .2
1bp res
16bp res
32bp res
64bp res
128bp res
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

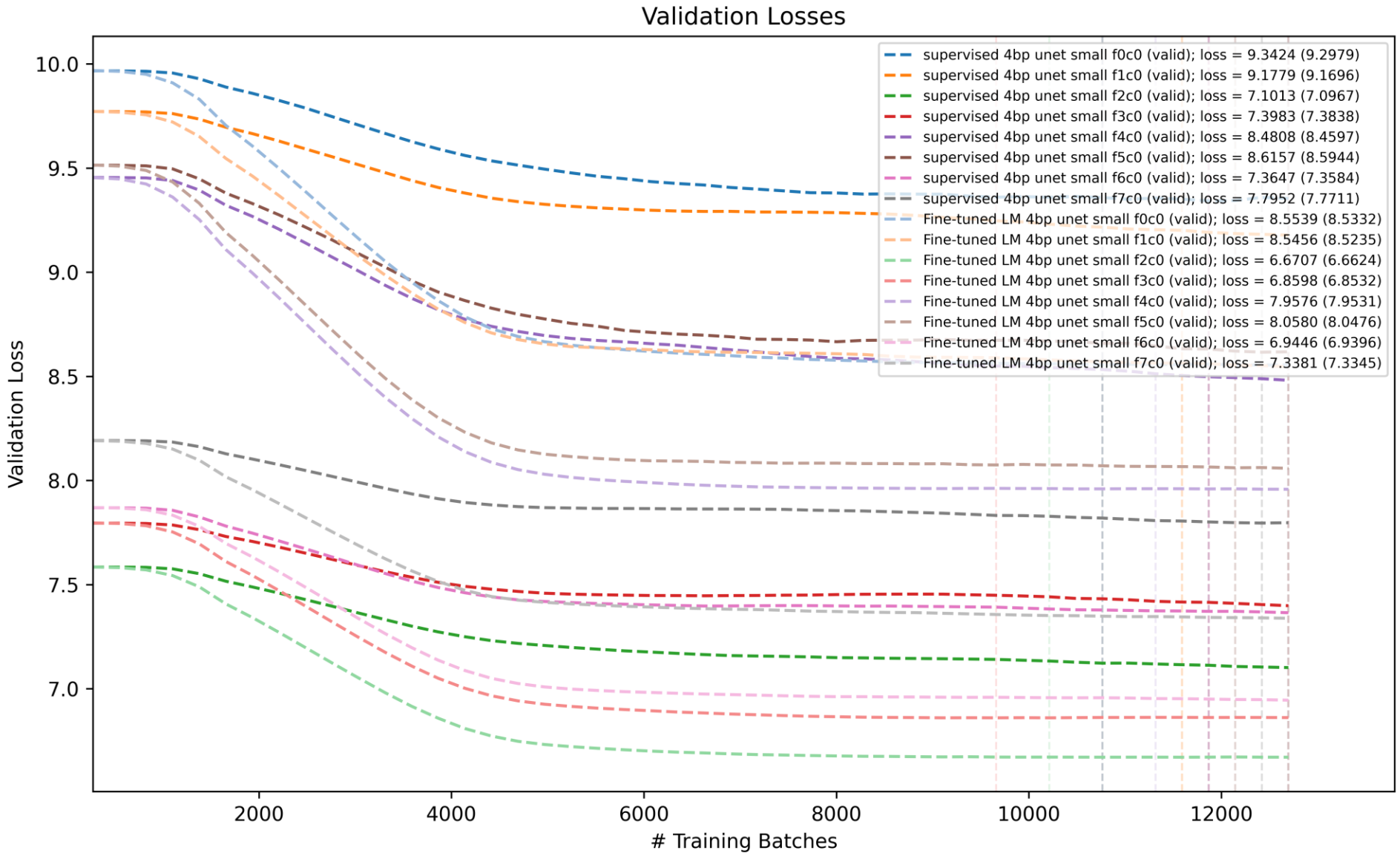

16384bp
? C T C T A ? C G ? G T A T A C
CHiP-exo (1128)
Histone marks (20)
RNA-Seq (1340)
Coverage Tacks
1bp res
...
...
16bp res
32bp res
64bp res
128bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Fine-tuning vs Training from Scratch (16 bp resolution)

- Divide genome into 8 folds.
- Train 8 models with distinct validation and test folds.



Fine-tuning vs Training from Scratch (4 bp resolution)



Fungi LM: Summary

1. Fungi language model: The Saccharomycetales order is a good evolutionary distance, offering good species diversity.
2. Repeat & coding down-weight masking are important
3. Under the exact model architecture, pretrained LM weights & fine-tuning can outperform training a model from scratch.

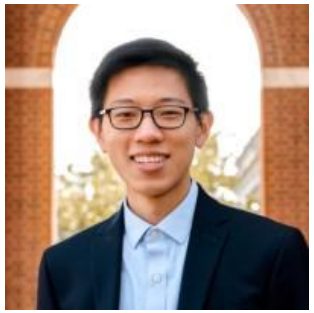
khchao.com @KuanHaoChao Kuanhao-Chao

Acknowledgement


Steven Salzberg


Mihaela Pertea


Alaina Shumate


Alan Mao

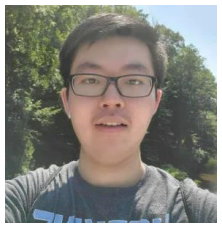

Jakob Heinz


Celine Hoh

Salzberg Lab

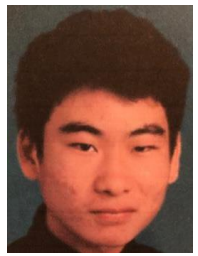






Pertea Lab



khchao.com @KuanHaoChao Kuanhao-Chao

Acknowledgement




Johannes Linder Majed Mohamed Magzoub David Kelley Sean Hackett

Kelley Lab & Calico Computing Team

Great mentors, collaborators and good friends!