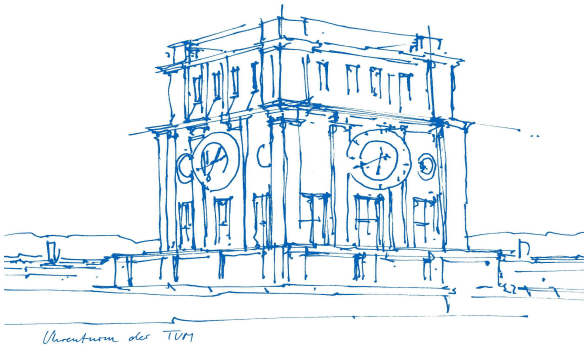
Kipoi: model zoo for genomics

Žiga Avsec

PhD candidate, Technical University of Munich

www.gagneurlab.in.tum.de

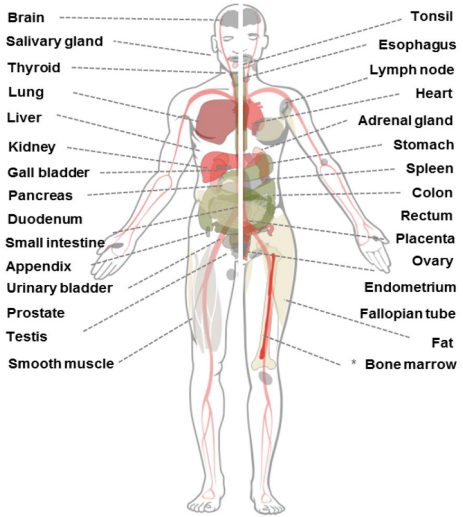
@gagneurlab, @KipoiZoo, @Avsecz



Genomics

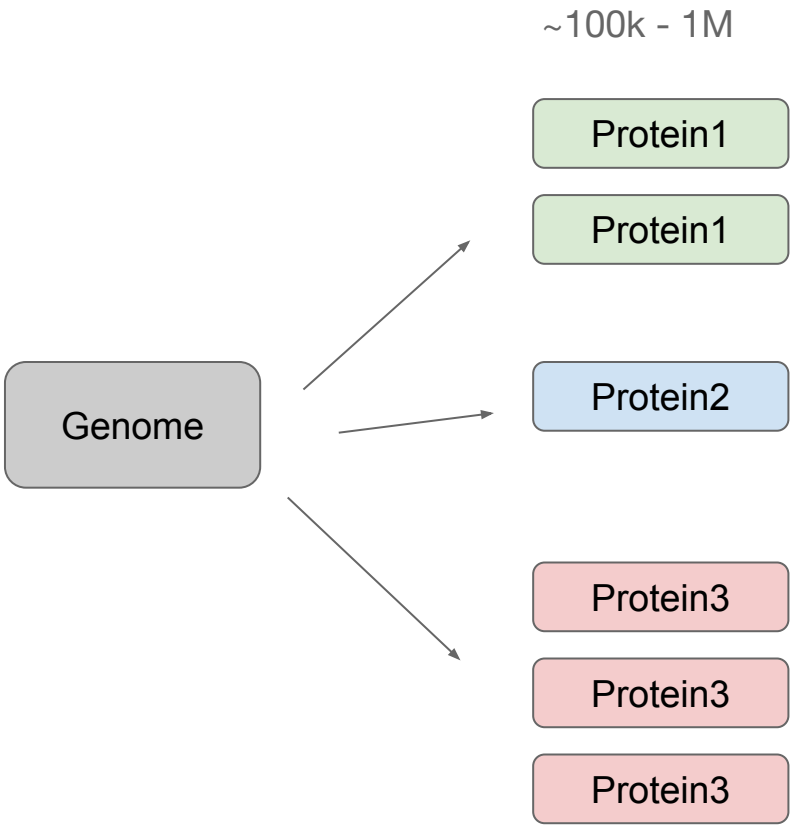
ACGTGTCAGTAGTTAAGCTAGTAGCTGATCGGTAACGTAGTGCACGTGTCAGTAGTTAAGCTAGTAGCTGATC

3 billion letters (x2) = 1 genome

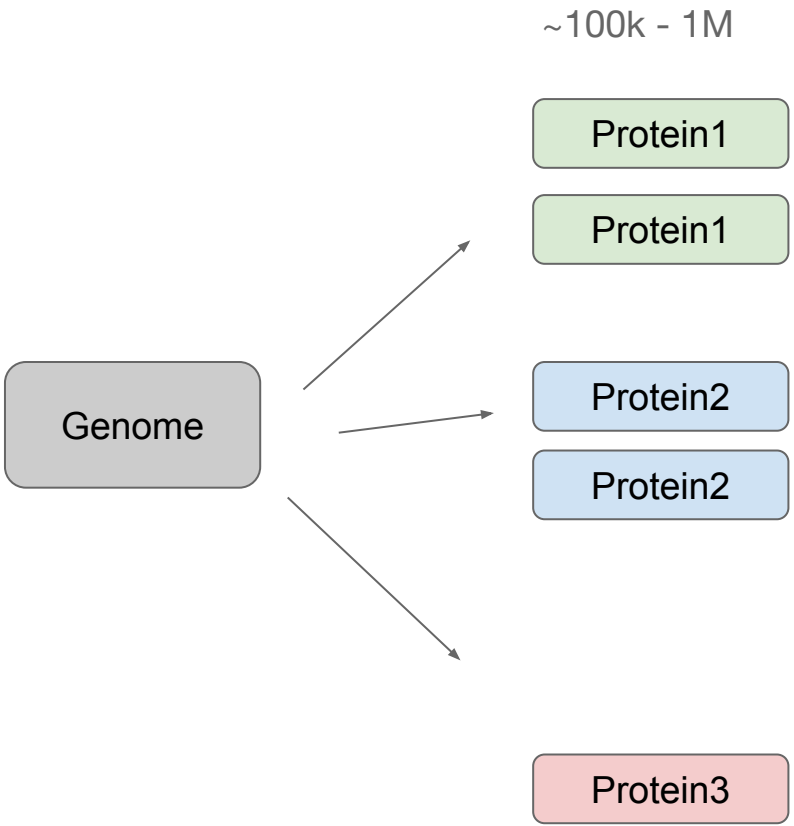


37 trillion cells

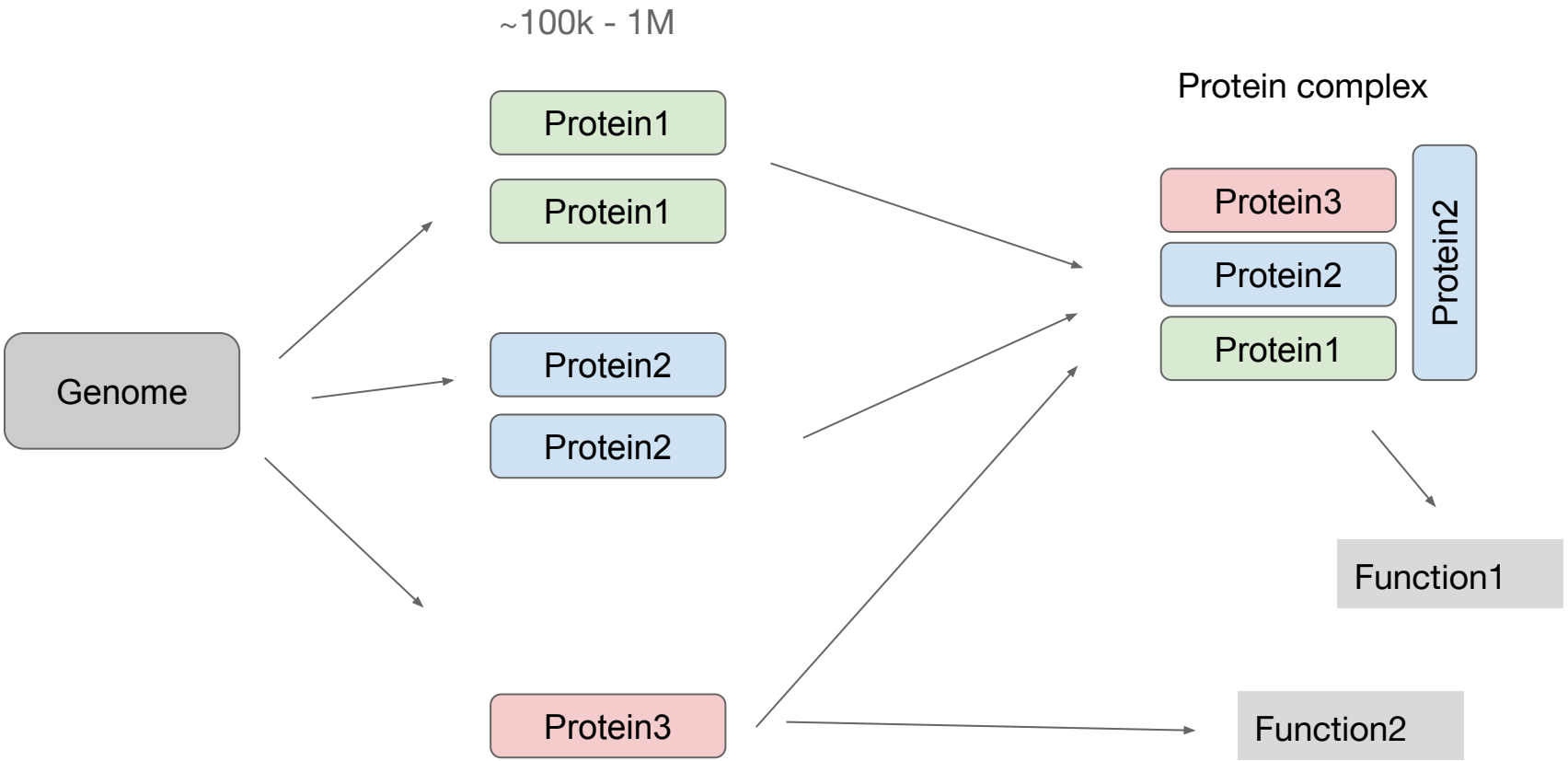
Proteins = main building blocks



Proteins = main building blocks



Proteins = main building blocks



How to make proteins from the genome?

Gene expression:
How information in DNA is read out

Protein (~100,000)

Translation

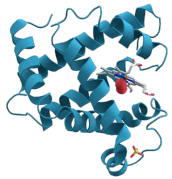
mature RNA

Splicing

precursor RNA

Transcription

DNA (2)



aucaugauacauagaucaugacuca

Intron

Exon

aucaugauauggauacgcauagaucaugacuca

atcgtatatatcatgatatggatacgcatagatcatgactcaggatacg

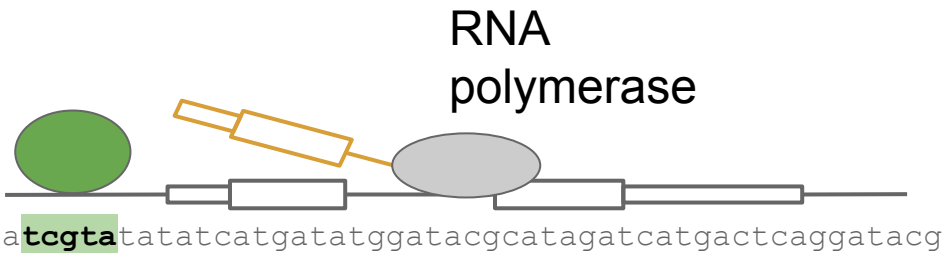
Gene (~20,000, 1% of DNA)

Gene expression:
How information in DNA is read out

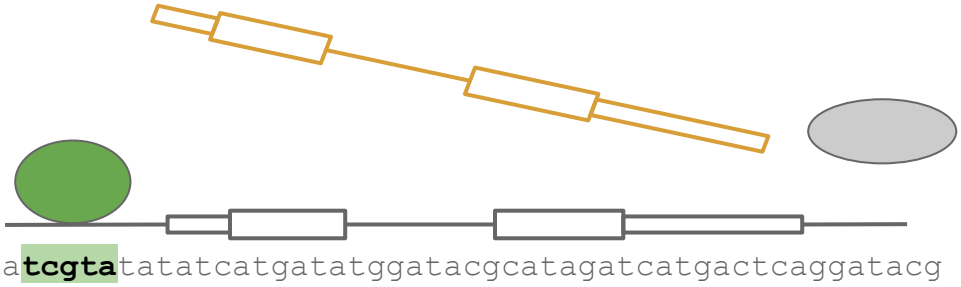
Transcription factor

Transcription Factor binding site

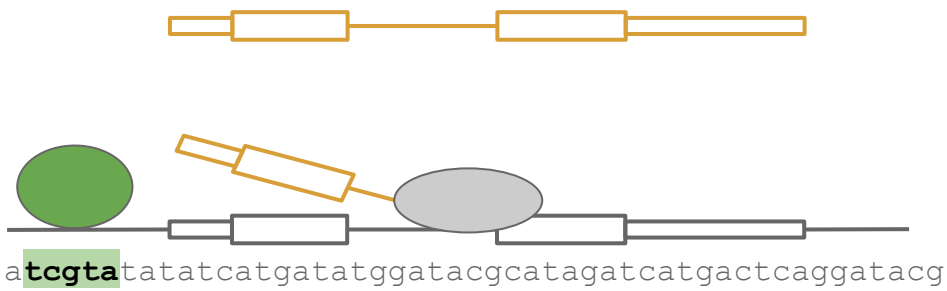
Gene expression:
How information in DNA is read out



Gene expression:
How information in DNA is read out



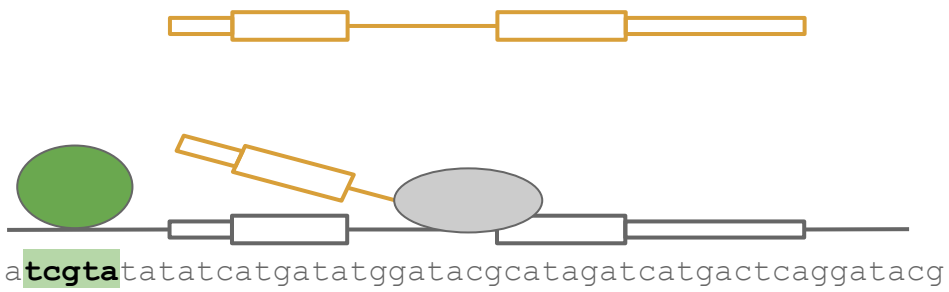
Gene expression:
How information in DNA is read out



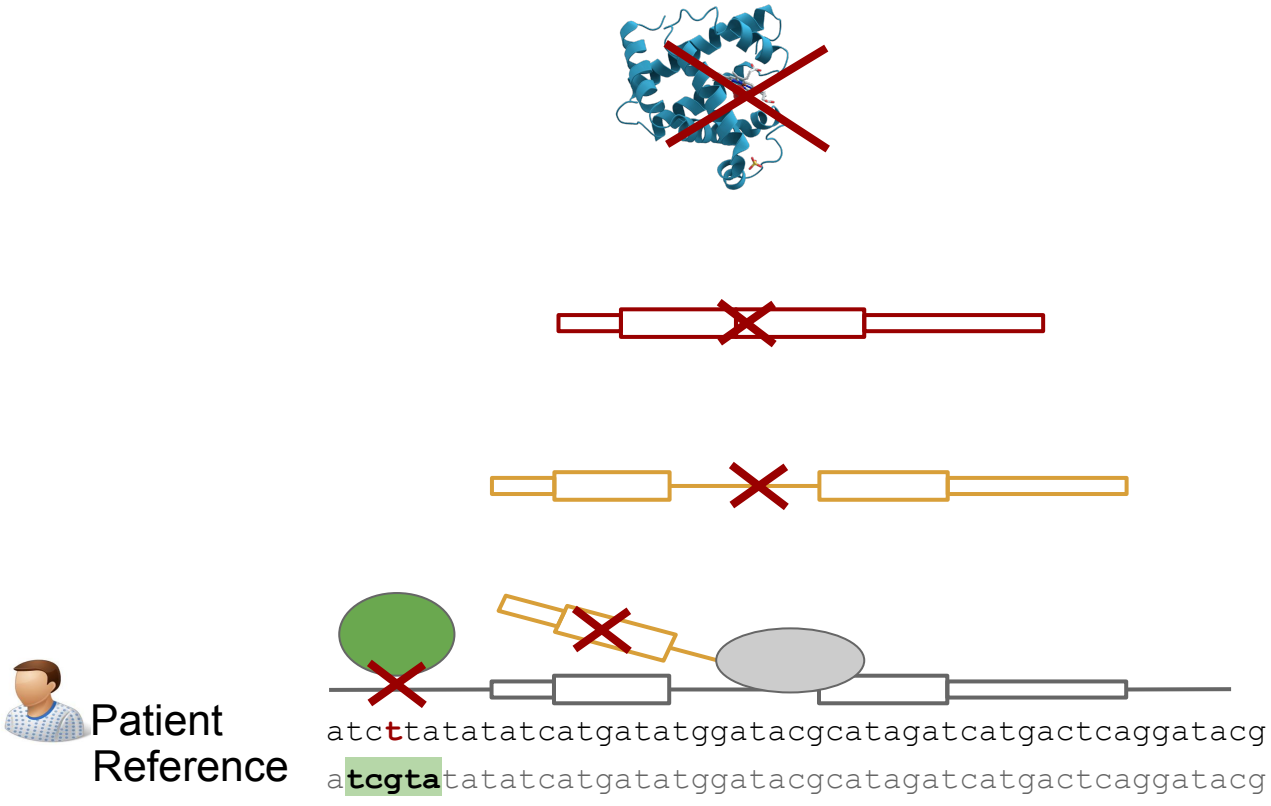
Gene expression:
How information in DNA is read out

The regulatory elements control:

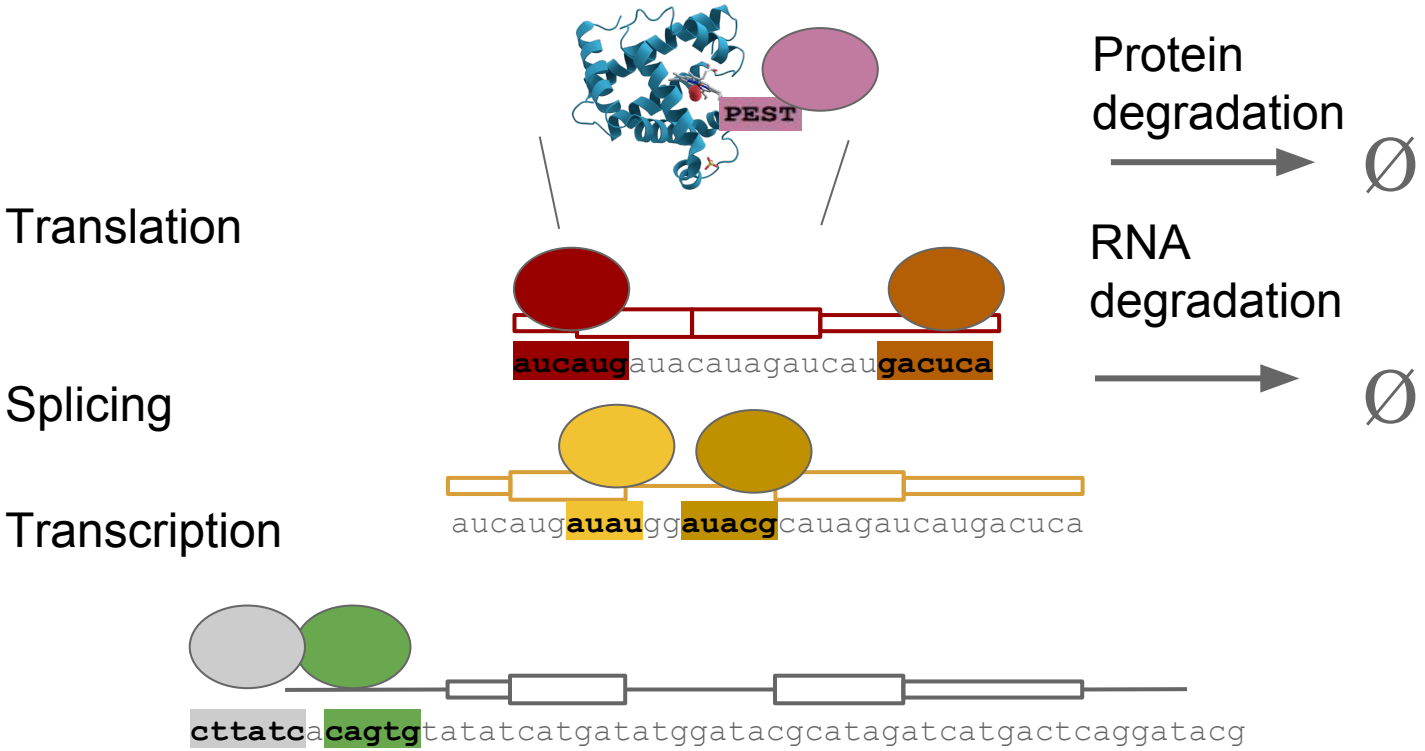
- The position of transcription initiation **(what)**
- The frequency of transcription **(how much)**



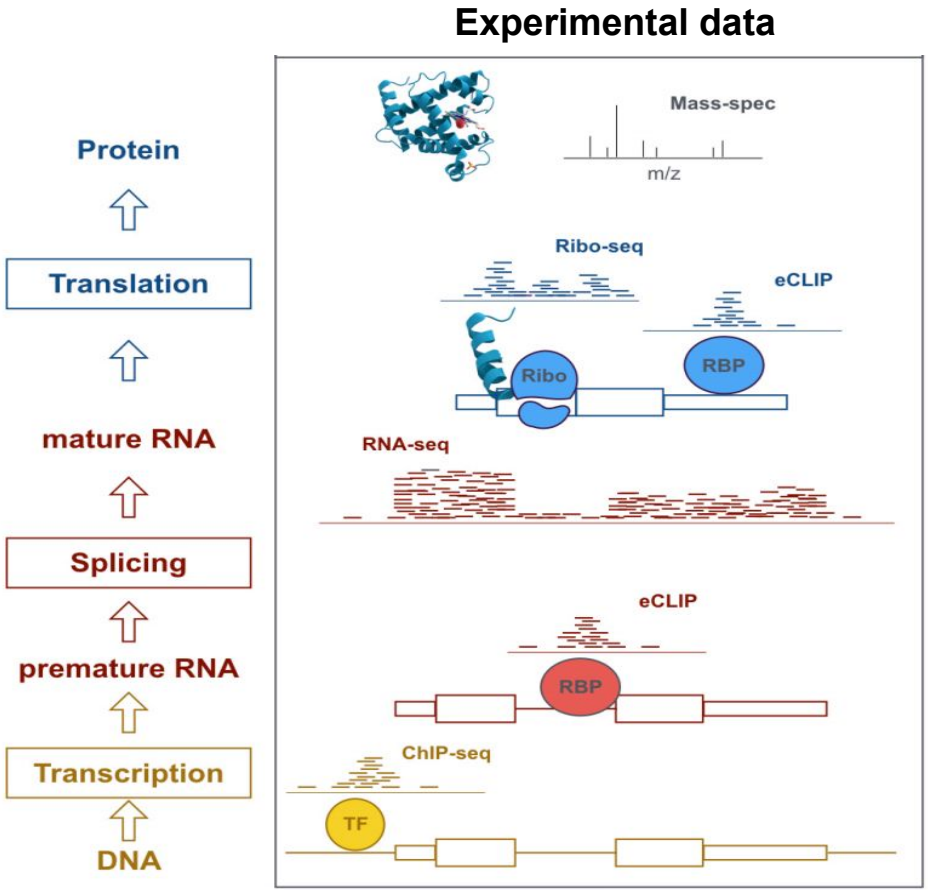
Genetic variants can disrupt regulatory elements



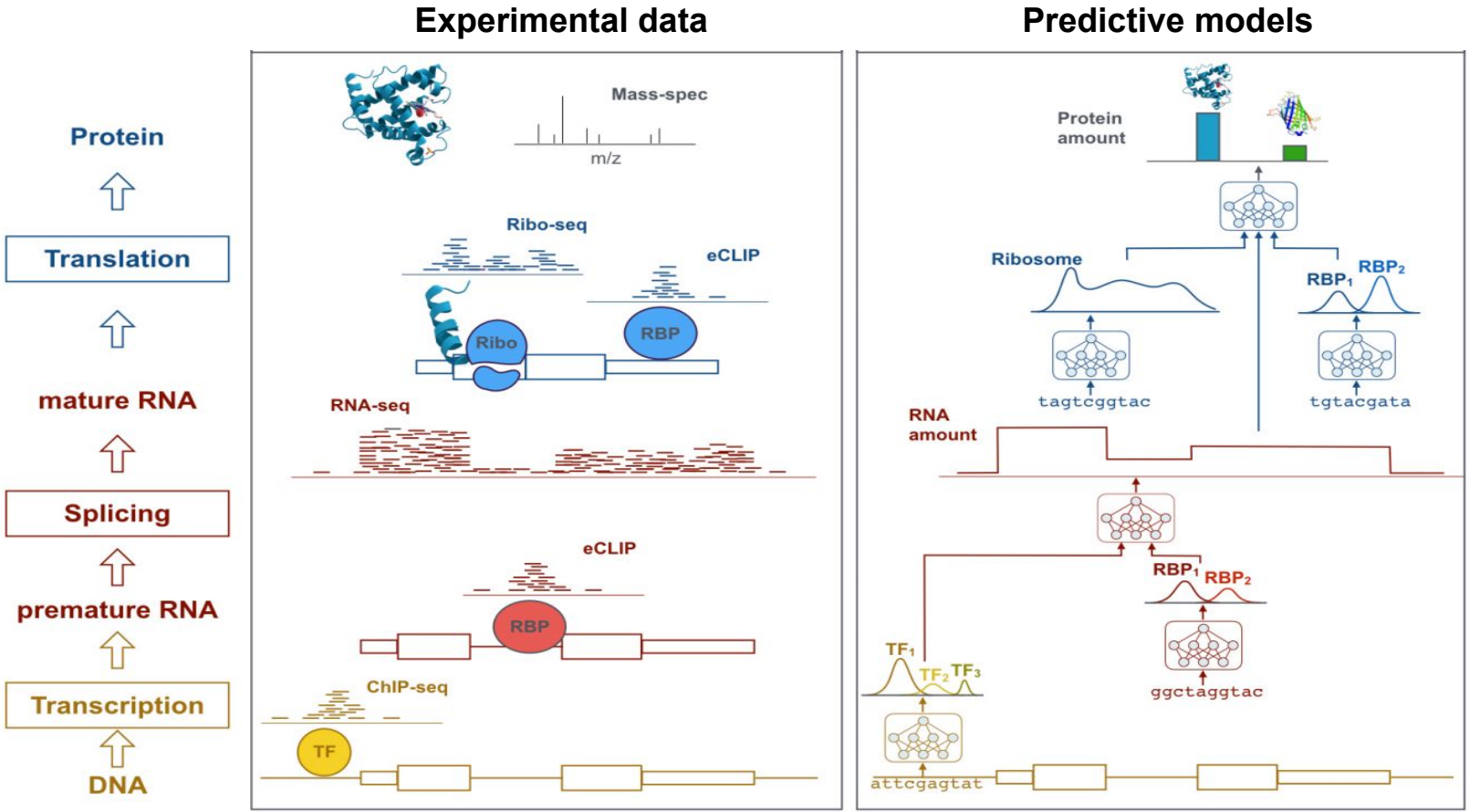
Thousands of regulatory elements across all steps of gene expression



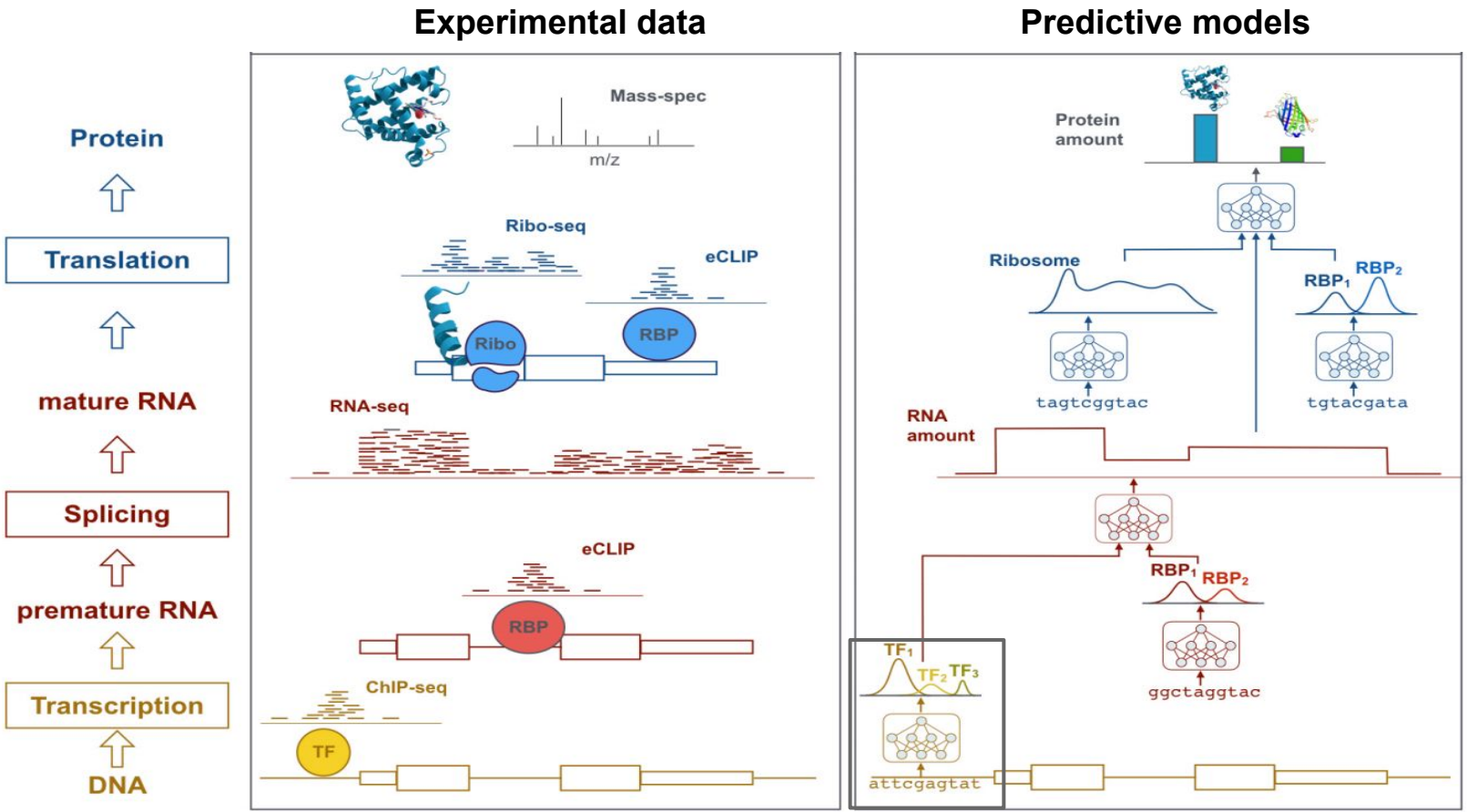
Measuring the regulatory steps via sequencing



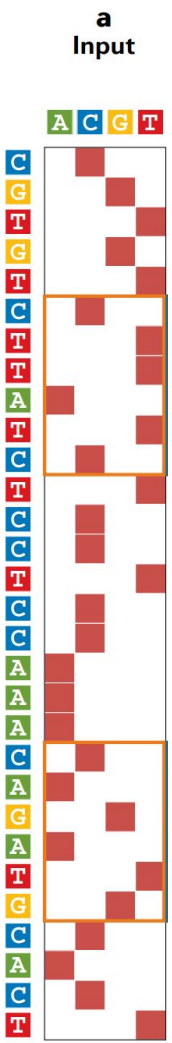
Learning the regulatory steps



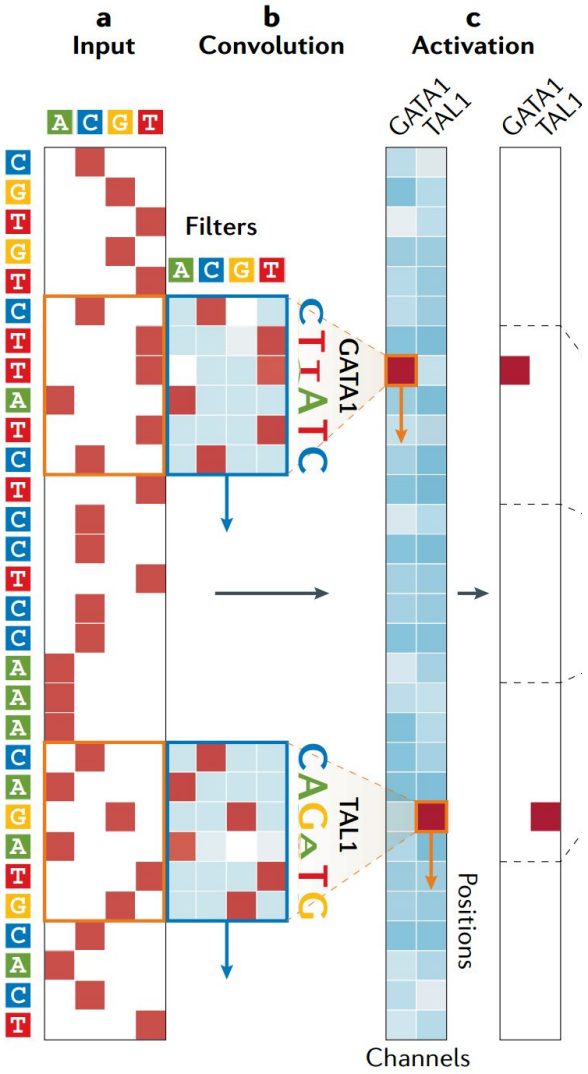
Learning the regulatory steps



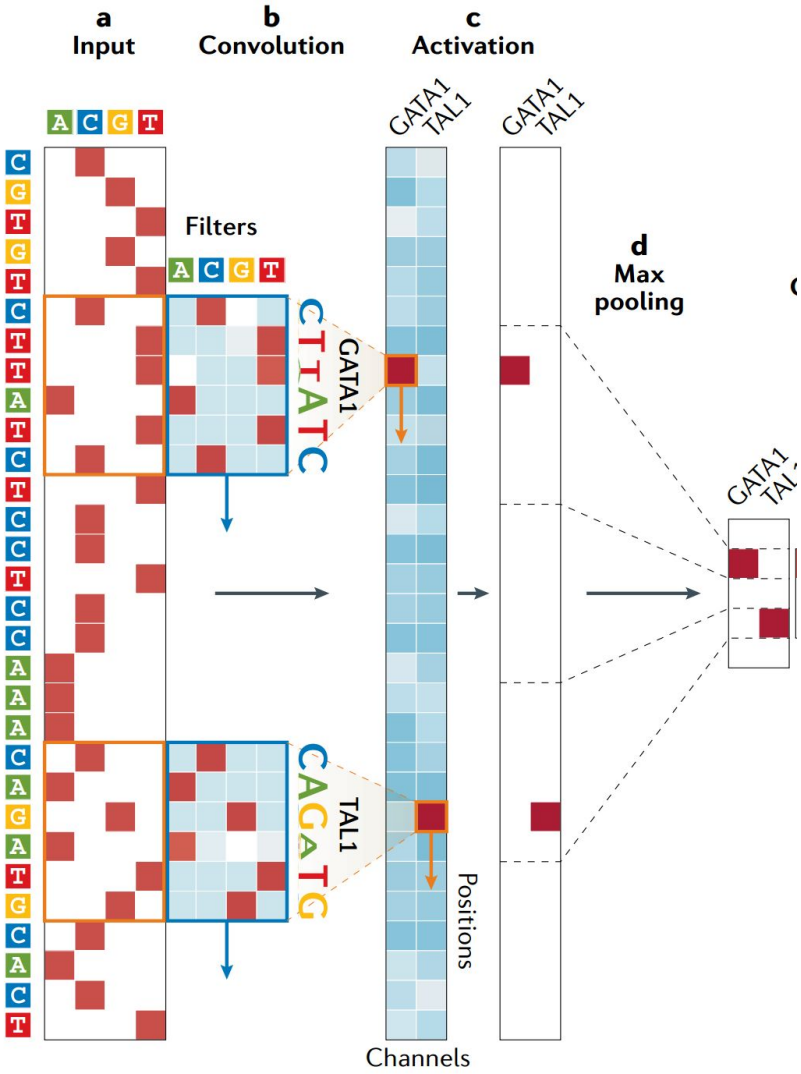
Detecting regulatory elements with convolutional neural networks


a
Input
A C G T
C
G
T
G
T
C
T
T
A
T
C
T
C
C
T
C
C
A
A
A
C
A
G
A
T
G
C
A
C
T

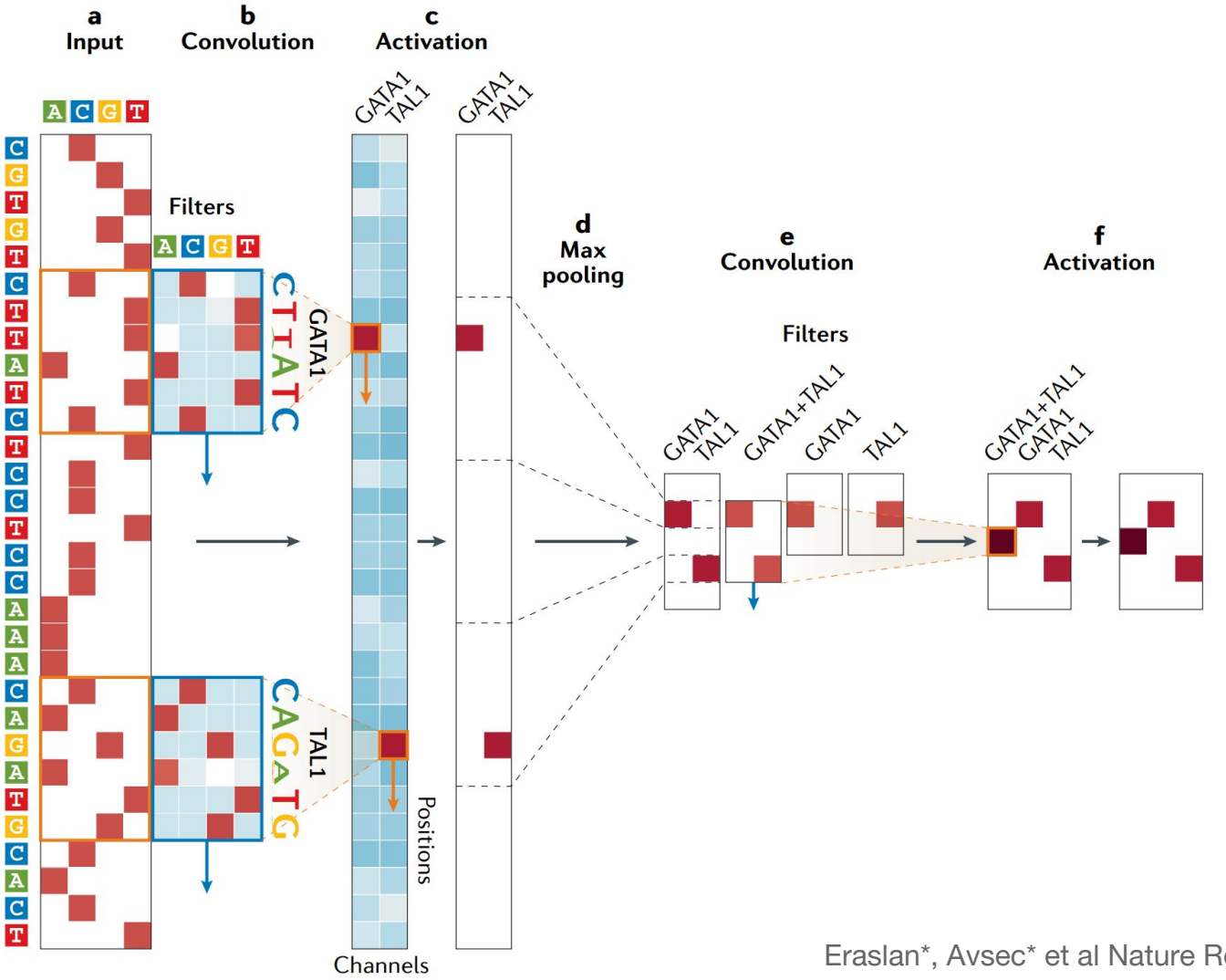
Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)



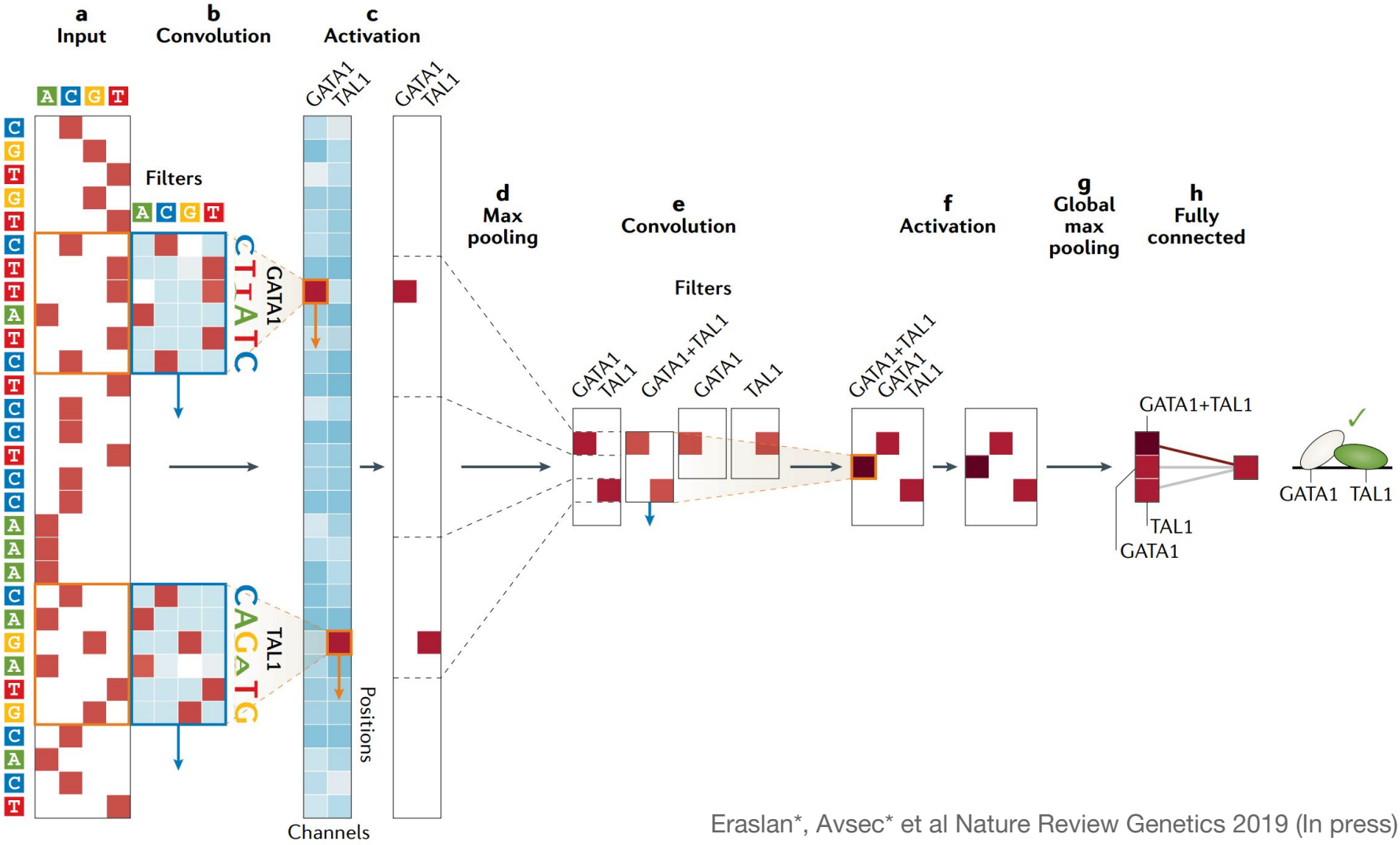
Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)



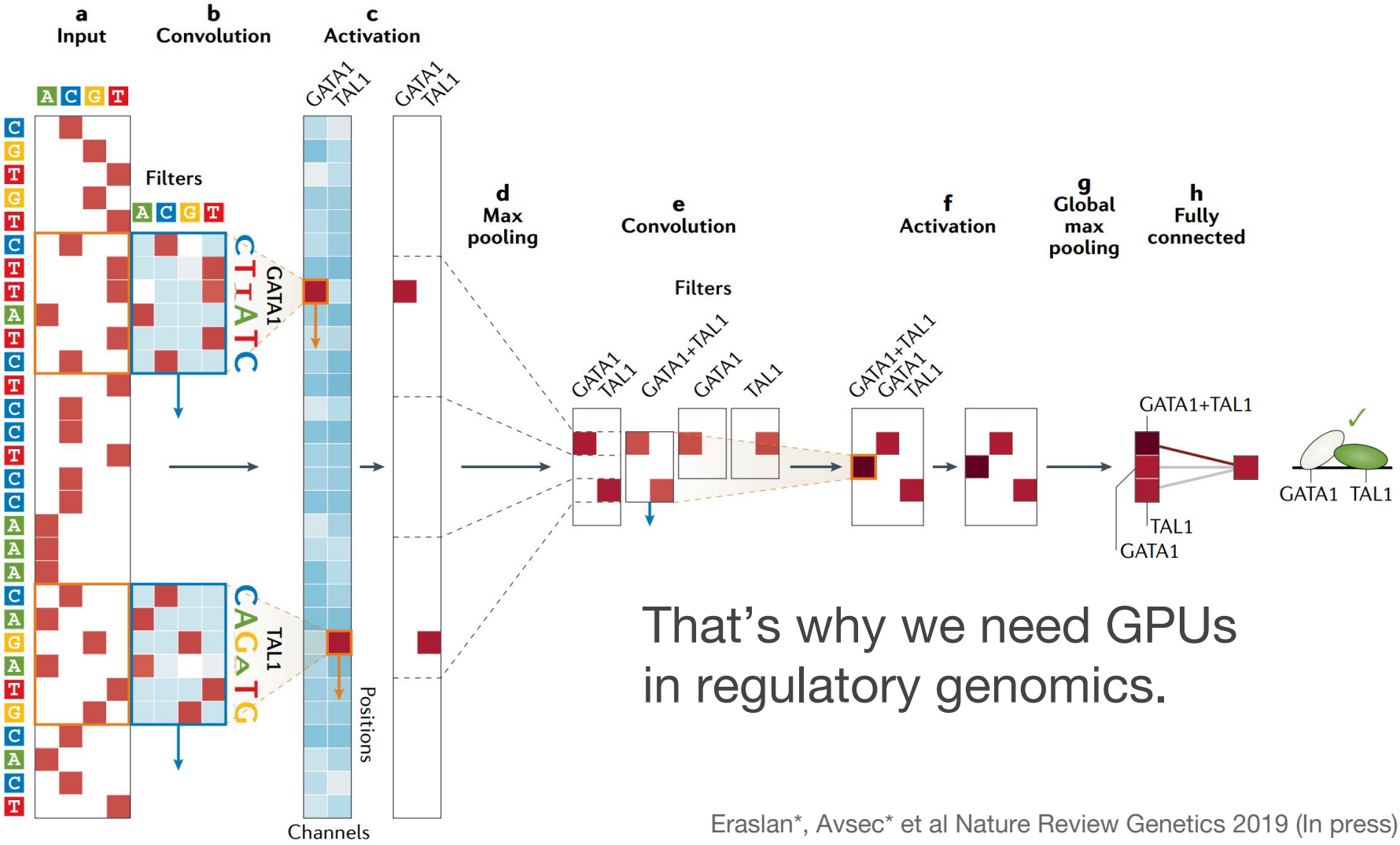
Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)



Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)



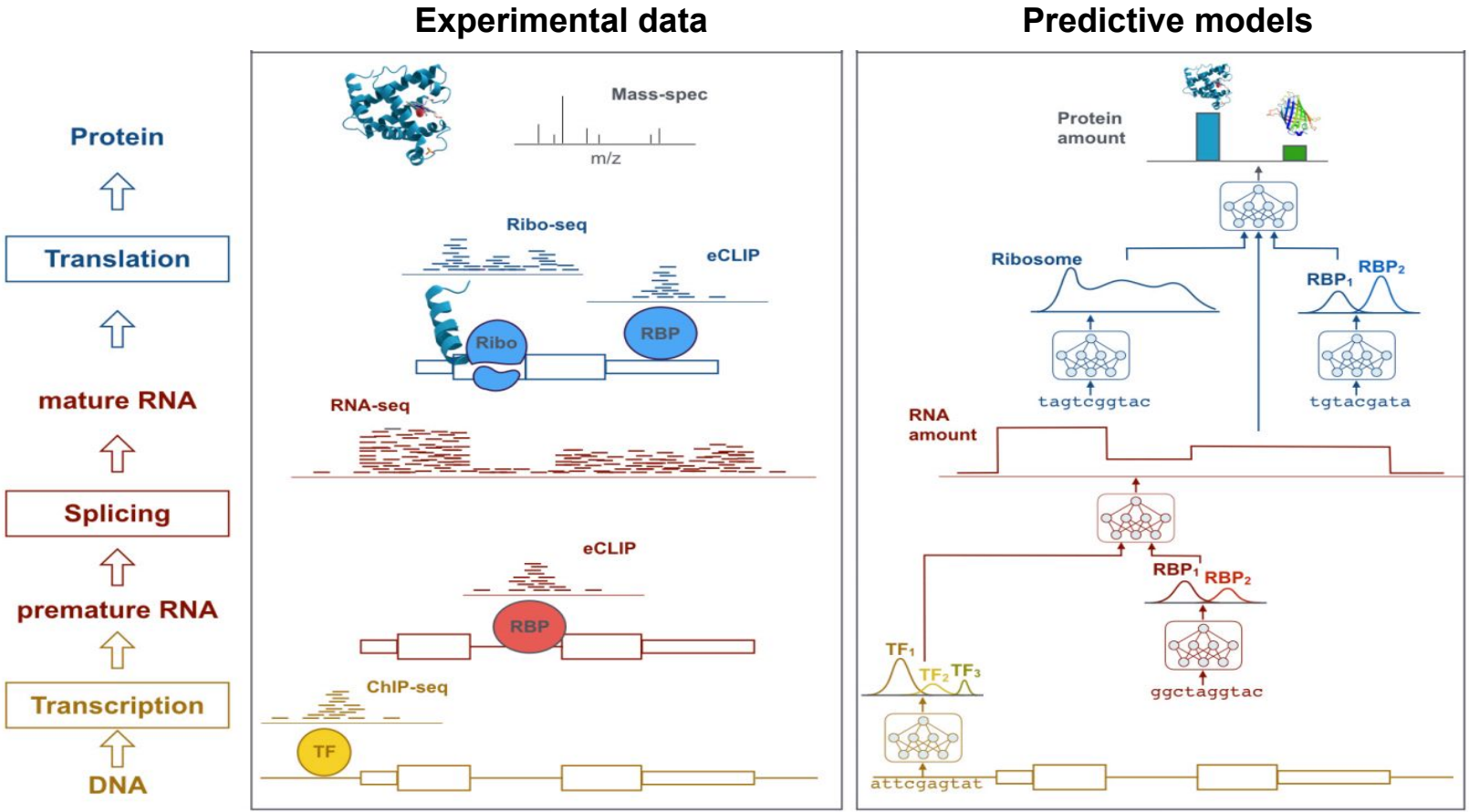
Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)



That's why we need GPUs in regulatory genomics.

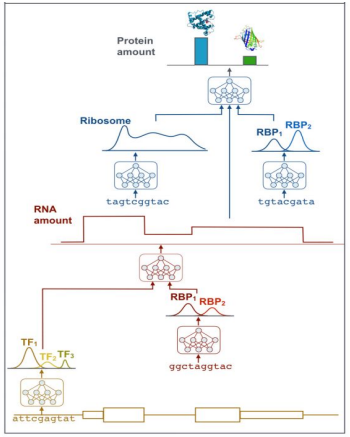
Eraslan*, Avsec* et al Nature Review Genetics 2019 (In press)

Learning the regulatory steps



List of published predictive models

- **Transcriptional regulation**
  - TF Binding
    - PWM Scanning (Jaspar, Cis-BP, MEME)
    - DeepBind
    - Improved DeepBind
    - FactorNet
    - GERV
    - DanQ
    - CKN-seq
  - Chromatin
    - DeepSEA
    - DeepChrome
    - Basenji
  - DNA methylation
    - CpGenie
    - DeepCpG
  - DNA Accessibility
    - Basset
  - TSS:
    - FIDDLE
  - Gene-Expression
    - Basenji
    - Expecto

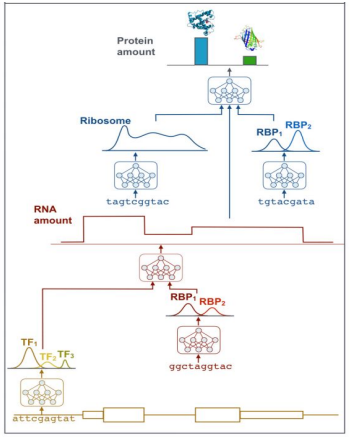


- **Post-transcriptional regulation**
  - RBP binding
    - iDeep
    - rbp_eclip (Avsec et al)
  - miRNA binding
    - TargetScan
    - deepMiRGene
  - Splicing
    - MaxEntScan 5', 3'
    - Labranchor
    - HAL
    - MMSplice
    - SpliceAI
  - mRNA half-life
    - Cheng et al 2017
  - Polyadenylation
    - APARENT
  - Translation
    - Optimus_5Prime
    - Cuperus et al 2017

See also: https://github.com/greenelab/deep-review

List of published predictive models

- **Transcriptional regulation**
  - TF Binding
    - PWM Scanning (Jaspar, Cis-BP, MEME)
    - DeepBind
    - Improved DeepBind
    - FactorNet
    - GERV
    - DanQ
    - CKN-seq
  - Chromatin
    - DeepSEA
    - DeepChrome
    - Basenji
  - DNA methylation
    - CpGenie
    - DeepCpG
  - DNA Accessibility
    - Basset
  - TSS:
    - FIDDLE
  - Gene-Expression
    - Basenji
    - Expecto



- **Post-transcriptional regulation**
  - RBP binding
    - iDeep
    - rbp_eclip (Avsec et al)
  - miRNA binding
    - TargetScan
    - deepMiRGene
  - Splicing
    - MaxEntScan 5', 3'
    - Labranchor
    - HAL
    - MMSplice
    - SpliceAI
  - mRNA half-life
    - Cheng et al 2017
  - Polyadenylation
    - APARENT
  - Translation
    - Optimus_5Prime
    - Cuperus et al 2017

See also: https://github.com/greenelab/deep-review

Can we easily apply these models to new data?

Can we easily re-use these models?

Lack of standardization leads to poor sharing and re-use of trained predictive models in genomics

Trained predictive models

Code repository

GitHub

Paper supplements

Methods

Supporting information

References

Author-maintained web page

Lack of standardization leads to poor sharing and re-use of trained predictive models in genomics

Trained predictive models

Code repository

GitHub

Paper supplements

Methods

Supporting information

References

Author-maintained web page

Data

Gene Expression Omnibus

PRIDE

ENA European Nucleotide Archive

....

Bioinformatics software

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

BIOCONDA®

The “**F**indable **A**ccessible **I**nteroperable **R**eusable” principle not only for data but also for trained predictive models

Challenges

- Making predictions end-to-end
  - `predict <model> -i input.data -o output.data`

Challenges

- Making predictions end-to-end
  - `predict <model> -i input.data -o output.data`

- Data heterogeneity (think one paper = one dataset)

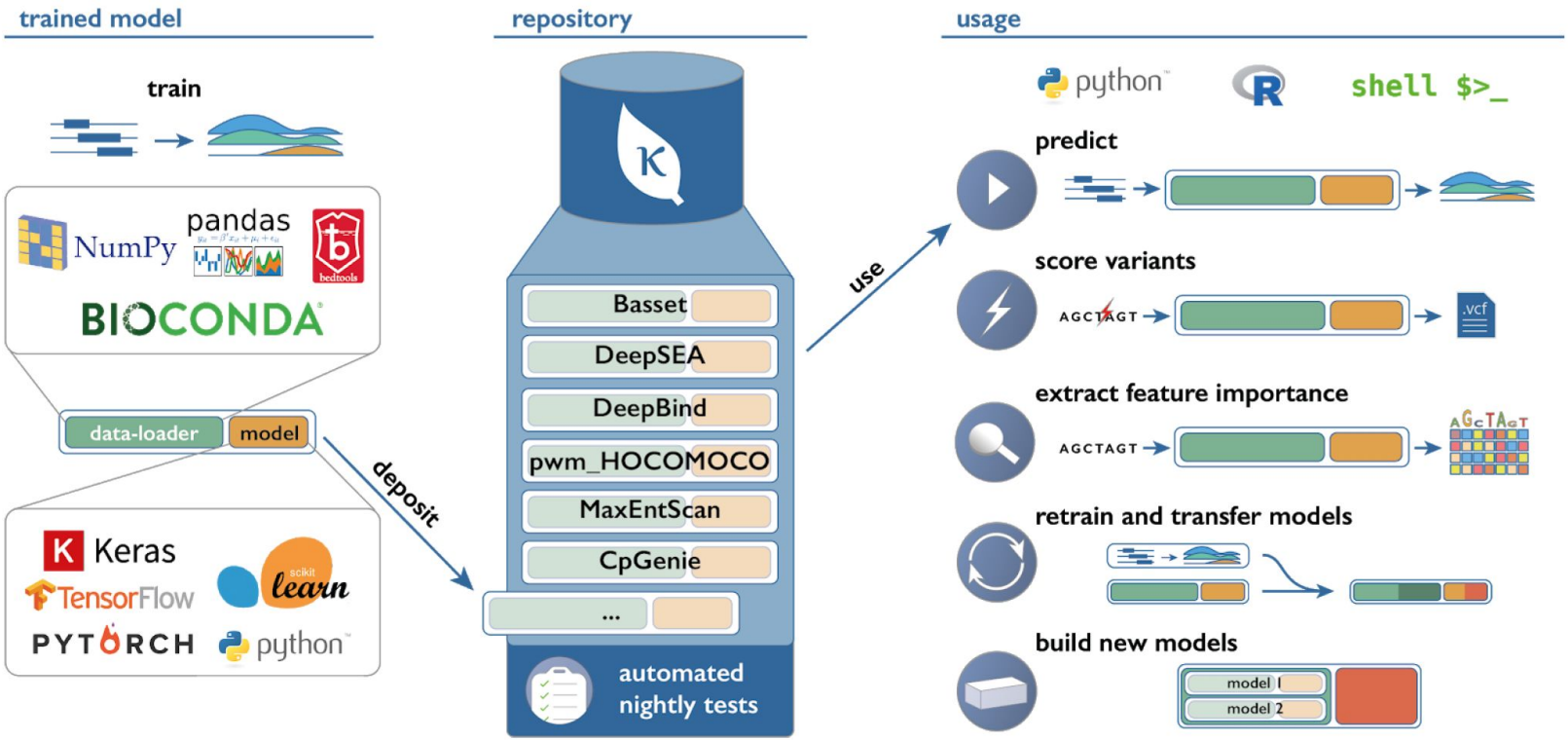
Challenges

- Making predictions end-to-end
  - `predict <model> -i input.data -o output.data`

- Data heterogeneity (think one paper = one dataset)

- Model heterogeneity (from deep learning frameworks to custom code)
  - Dependency issues

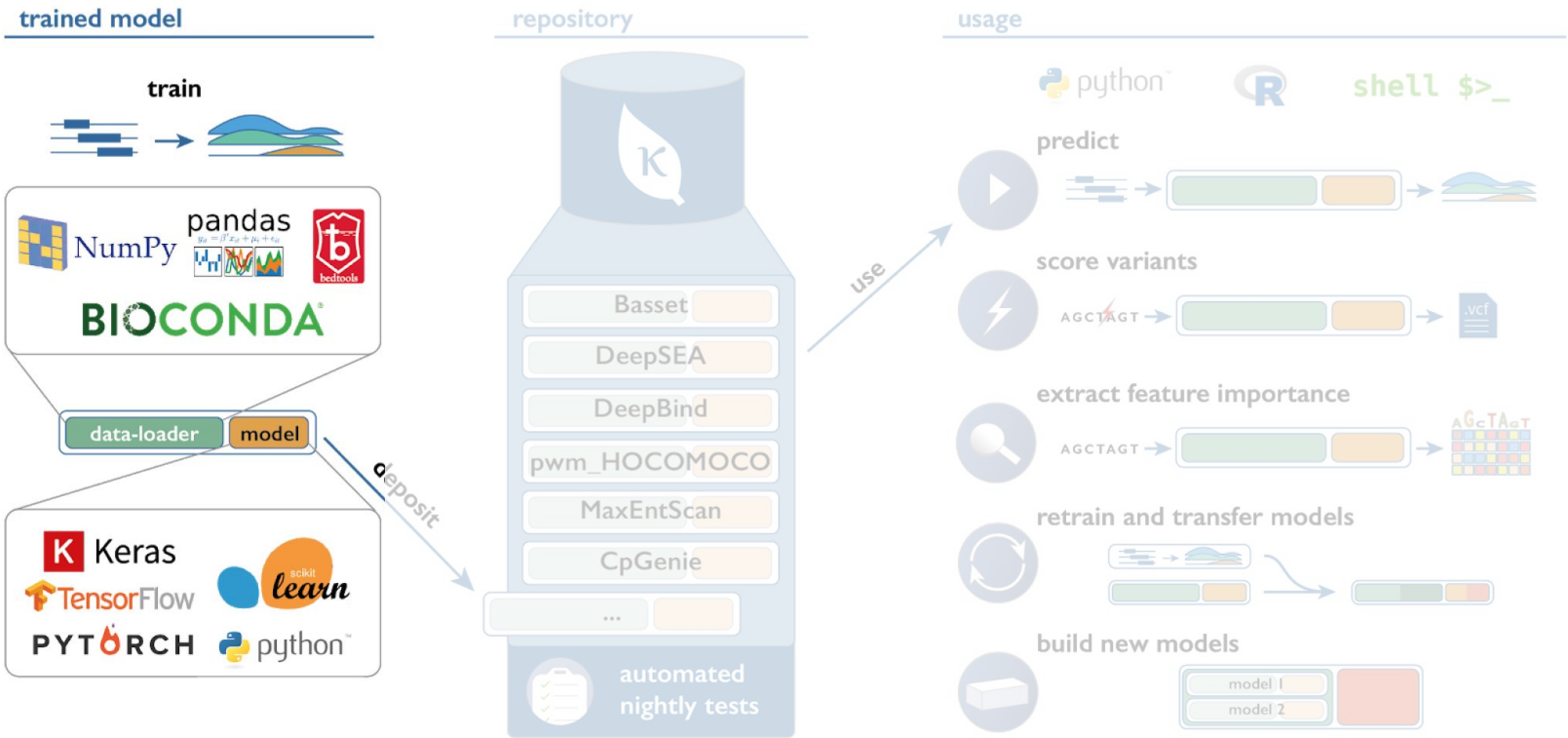
Kipoi.org [Kípi]



with A. Kundaje, Stanford, and O. Stegle, DKFZ

Avsec et al, Nature Biotechnology (In press)

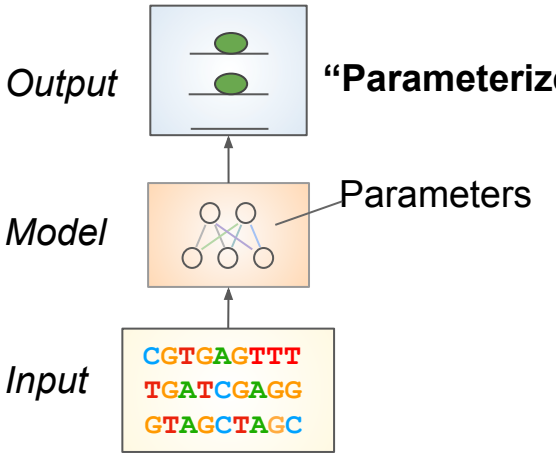
Trained model (model.yaml)



Model

data-loader | model

```
type: keras
args:
    weights:
        url: https://zenodo.org/record/1452399/files/model.weights.h5?download=1
        md5: 2a0ae0a29337eb8106d65e1baeda85d1
    arch:
        url: https://zenodo.org/record/1452399/files/model.arch?download=1
        md5: 6903bcab337a6753ad010f43f208df42
```



“Parameterized function”

Can be implemented using:

Keras

TensorFlow

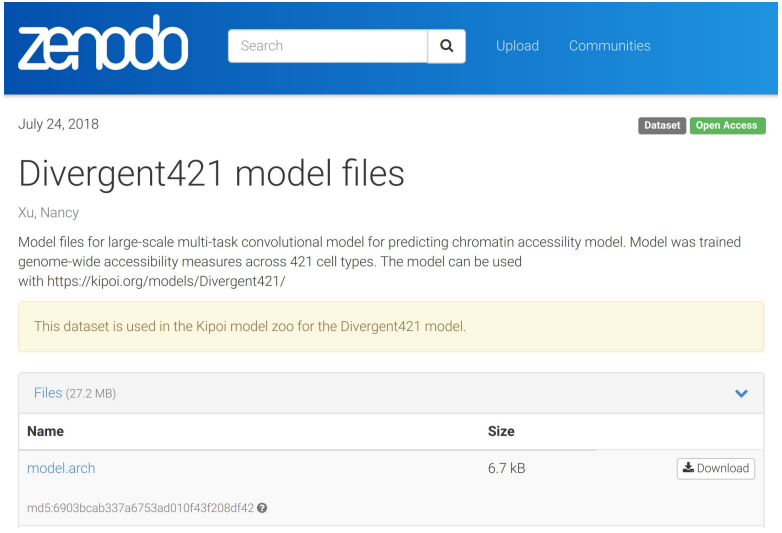
scikit learn

PYTORCH

python™

Model


```
type: keras
args:
    weights:
        url: https://zenodo.org/record/1452399/files/model.weights.h5?download=1
        md5: 2a0ae0a29337eb8106d65e1baeda85d1
    arch:
        url: https://zenodo.org/record/1452399/files/model.arch?download=1
        md5: 6903bcab337a6753ad010f43f208df42
```



Data-loader

data-loader | model

github.com/kipoi/kipoiseq

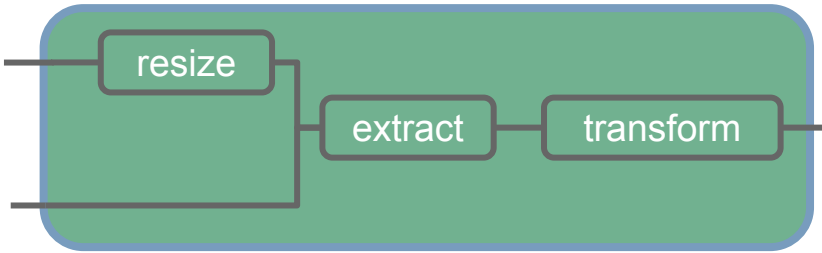
```
default_dataloader:
    defined_as: kipoiseq.dataloaders.SeqIntervalDl
    default_args:
        auto_resize_len: 1000
```

intervals.bed

```
chr1   1000   2000
chr2   5000   7000
```

genome.fa

```
>chr1
NNNNNNNNNNNN...
```




```
array([[[1, 0, 0, 0],
        [0, 1, 0, 0],
        [0, 0, 1, 0],
        [1, 0, 0, 0],
        [    …     ]],

       [[0, 1, 0, 0],
        [0, 0, 1, 0],
        [1, 0, 0, 0],
        [0, 0, 0, 1],
        [    …     ]]])
```

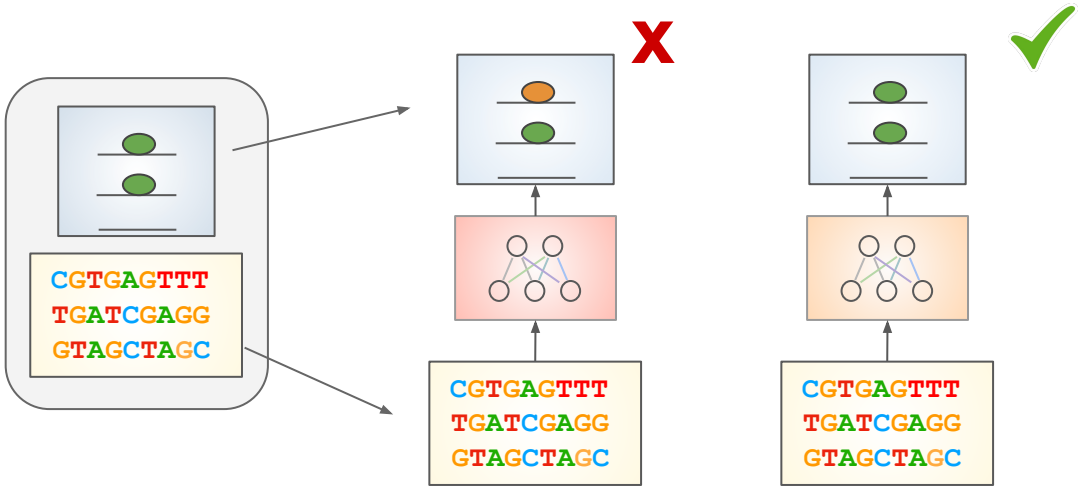
Dependencies

data-loader | model

```yaml
dependencies:
  conda:
    - python>=3.5
    - h5py
    - bioconda::pyfaidx
  pip:
    - tensorflow<=1.4.1
    - keras==1.2.2
    - kipoiseq
```

Test predictions

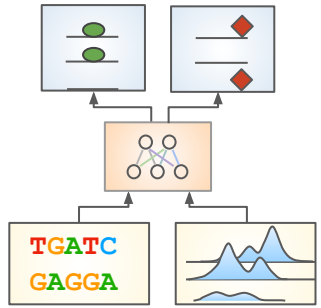
```
test:
  expect:
    url: https://s3.eu-central-1.amazonaws.…25f43/Divergent421/predictions.h5
    md5: 62da0ac731f323ea54ee6e30c38e0722
  precision_decimal: 5
```



Schema

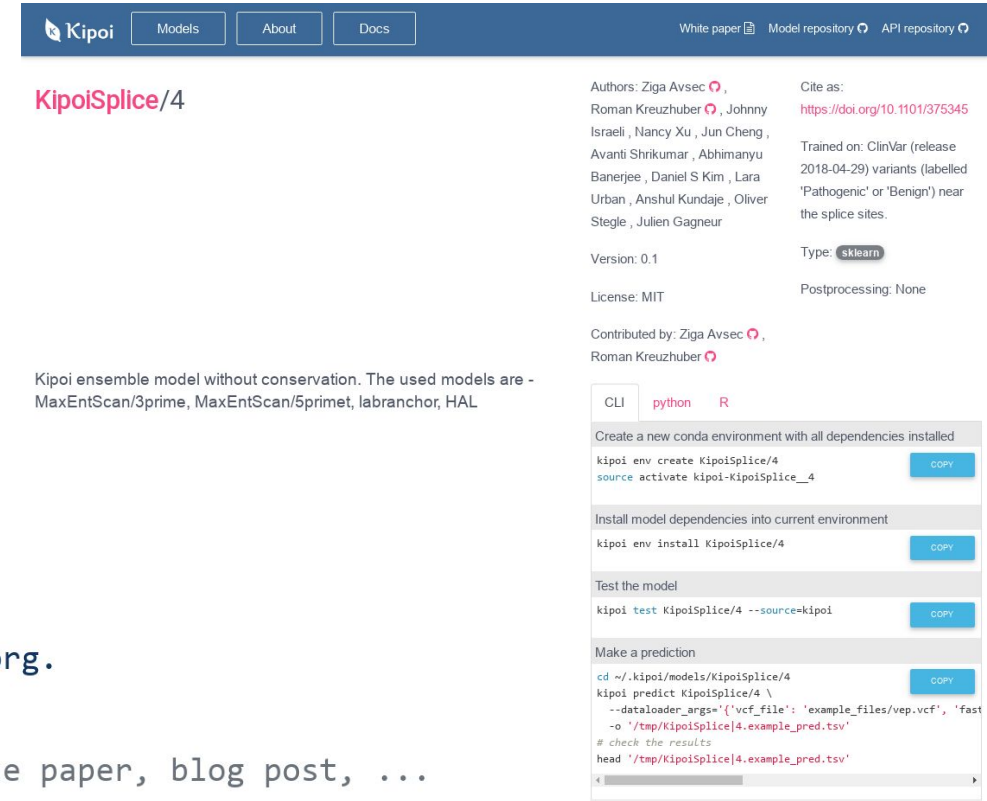
```
schema:
    inputs:
        shape: (1000,4)
        doc: "1000 base pair sequence of one-hot encoding ACGT"
    targets:
        shape: (421,)
        doc: "Binary 0/1 output for chromatin accessibility in…
        column_labels: task_names.txt
```

Supports multiple inputs/outputs

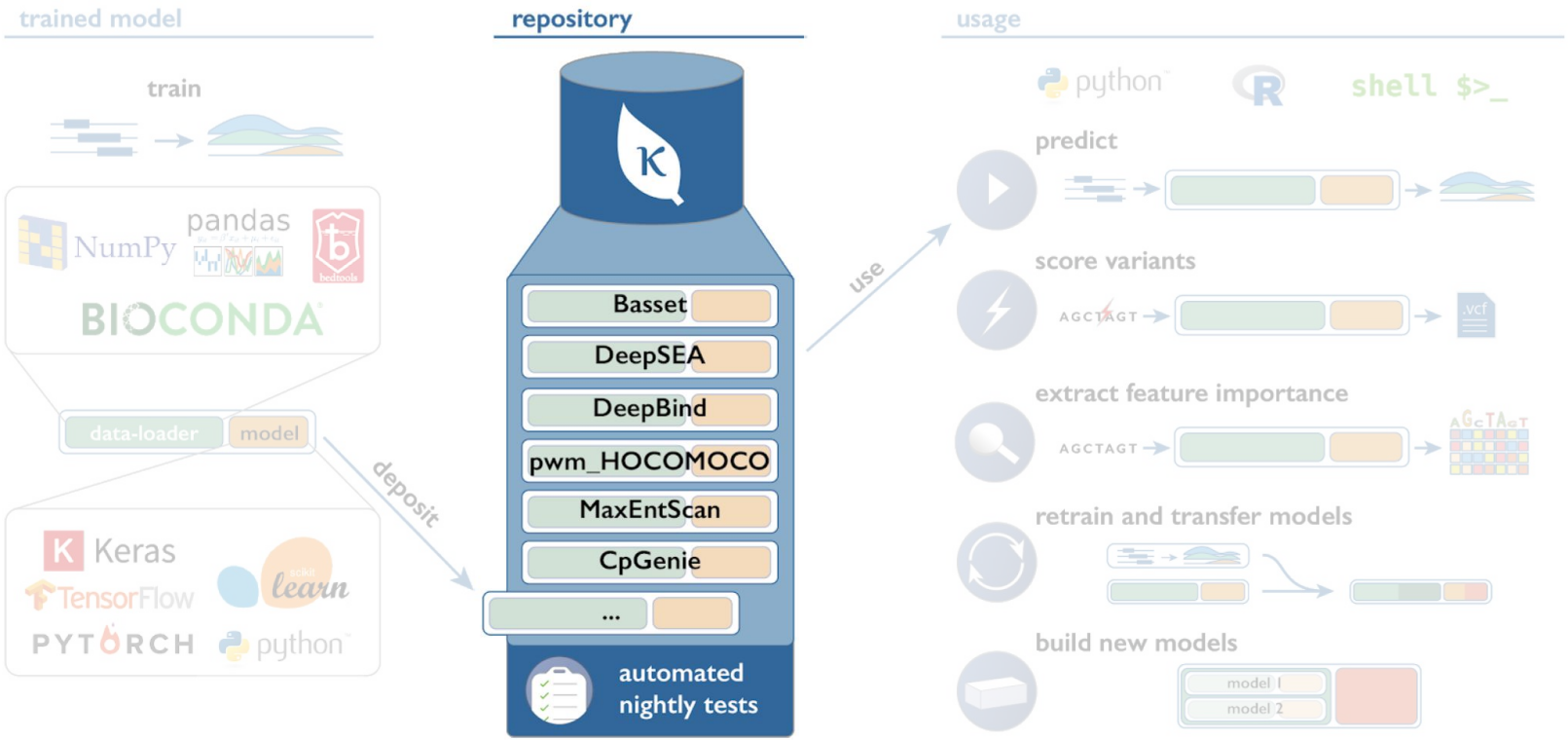


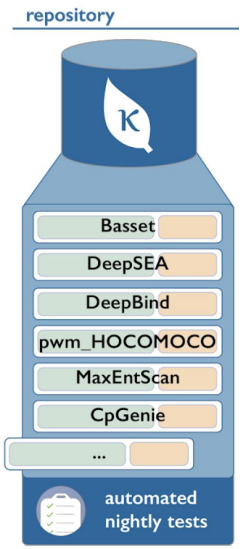
General information

```yaml
info:
  authors:
      - name: My Name
        github: myname
  name: MyKerasModel
  doc: >

    Description of the model shown on kipoi.org.

  cite_as: https://doi.org/x/y   # link to the paper, blog post, ...
  tags:  # under which category dos this model fall
    - DNA accessibility
```



Model repository





kipoi / **models**

Unwatch 18 | Unstar 43 | Fork 15

Code | Issues 28 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Model zoo for genomics http://kipoi.org

Edit

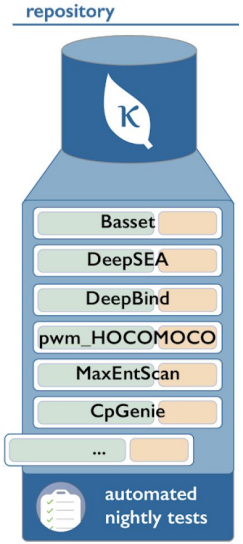
Manage topics

213 commits | **6** branches | **0** releases | **10** contributors | MIT

Branch: **master** | New pull request | Create new file | Upload files | Find file | Clone or download

s6juncheng Merge pull request #127 from kipoi/s6juncheng-patch-3 ··· Latest commit 23c1e96 2 days ago

.circleci	Update config.yml	7 months ago
Basenji	Update model.yaml	2 months ago
Basset	Update model.yaml	2 months ago
BassetGM12878_Demo	starting all the dependencies with bioconda packages (#93)	5 months ago
CleTimer	remove cite-as	28 days ago
CpGenie	starting all the dependencies with bioconda packages (#93)	5 months ago
DeepBind	map dl shema and model shema	a month ago
DeepCpG_DNA	starting all the dependencies with bioconda packages (#93)	5 months ago
DeepSEA	use linecache by default	8 days ago
Divergent421	added Divergent421 (#105)	4 months ago
FactorNet	add wgEncodeDukeMapabilityUniqueness35bp.chr22.bigWig for factornet	8 days ago



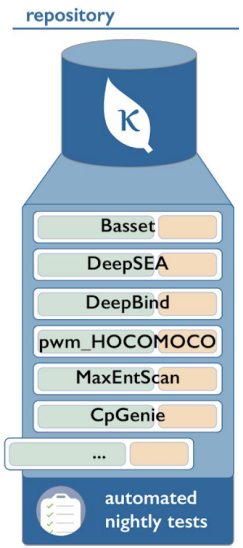
Branch: **kipoiseq_dl** ▾ **models** / **Divergent421** /

Create new file | Upload files | Find file | History

This branch is 57 commits ahead, 6 commits behind master. Pull request Compare

Avsecz remove tab — Latest commit `d59c7da` 2 days ago

..		
model.yaml	remove tab	2 days ago
task_names.txt	added task names to Divergent421	2 days ago



Branch: kipoiseq_dl ▾ **models** / **Divergent421** /

Create new file | Upload files | Find file | History

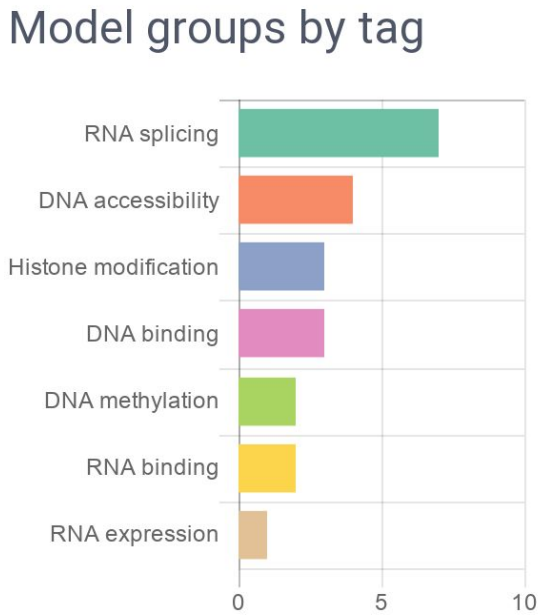
This branch is 57 commits ahead, 6 commits behind master. Pull request Compare

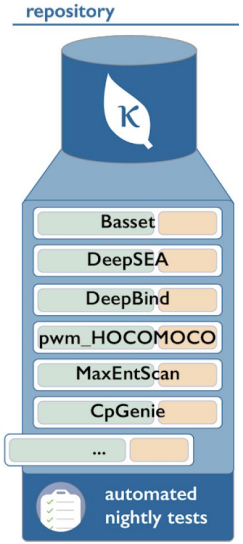
Avsecz remove tab — Latest commit d59c7da 2 days ago

..		
model.yaml	remove tab	2 days ago
task_names.txt	added task names to Divergent421	2 days ago

model groups: 20

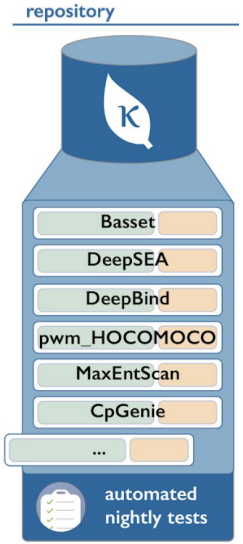
of models: 2073





Testing models

- Install new conda environment
- Test the pipeline: data -> dataloader -> model
- Test the predictions match

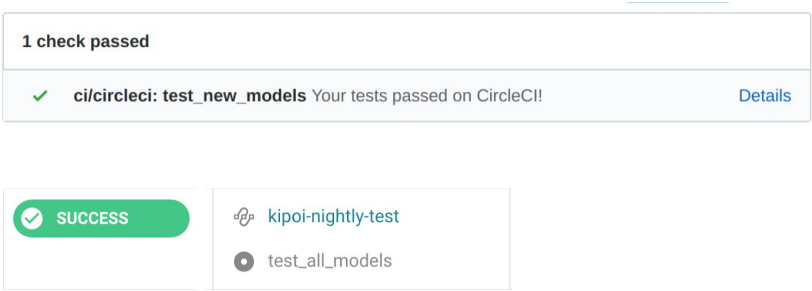


Testing models

- Install new conda environment
- Test the pipeline: data -> dataloader -> model
- Test the predictions match

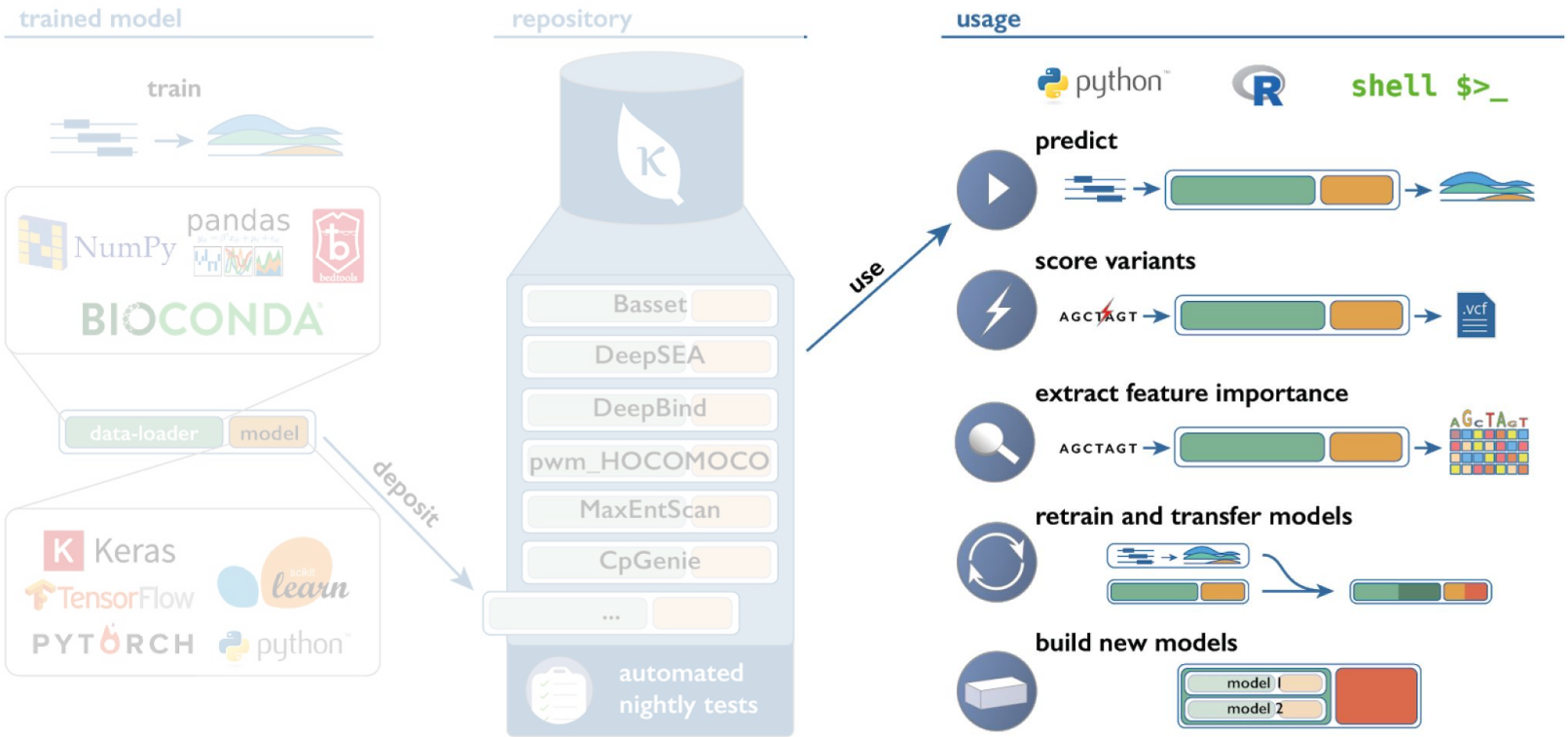
When?

- Pull-request
- Nightly (all model groups)



circleci

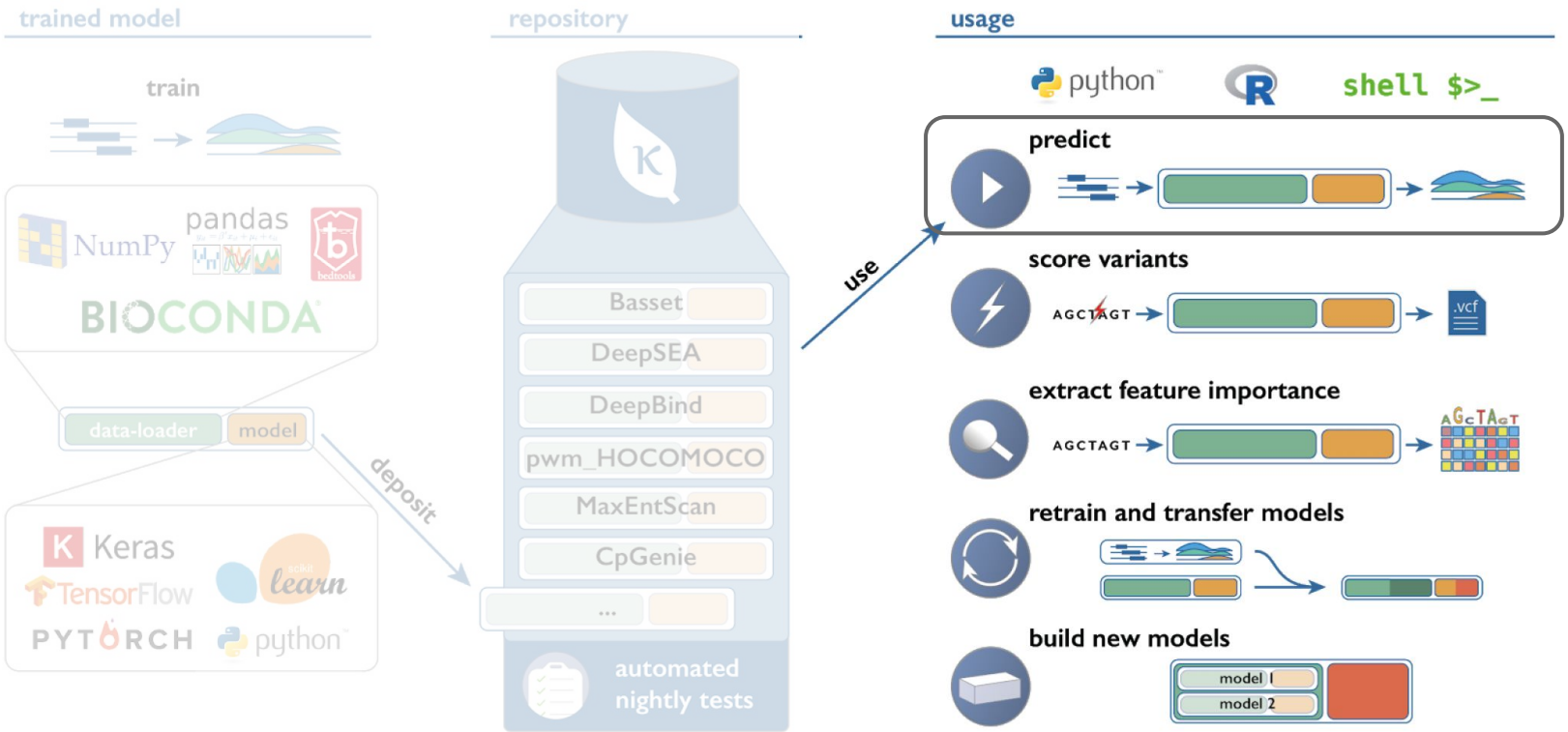
Using models



For the impatient:
30 seconds introduction to Kipoi

```python
import kipoi

kipoi.list_models() # list available models

model = kipoi.get_model("Basset") # load the model

model = kipoi.get_model(  # load the model from a past commit
    "https://github.com/kipoi/models/tree/<commit>/<model>",
    source='github-permalink'
)

# main attributes
model.model # wrapped model (say keras.models.Model)
model.default_dataloader # dataloader
model.info # description, authors, paper link, ...

# main methods
model.predict_on_batch(x) # implemented by all the models regardless of the framework
model.pipeline.predict(dict(fasta_file="hg19.fa",
                            intervals_file="intervals.bed"))
# runs: raw files -[dataloader]-> numpy arrays -[model]-> predictions
```

Case study 1: Benchmarking models

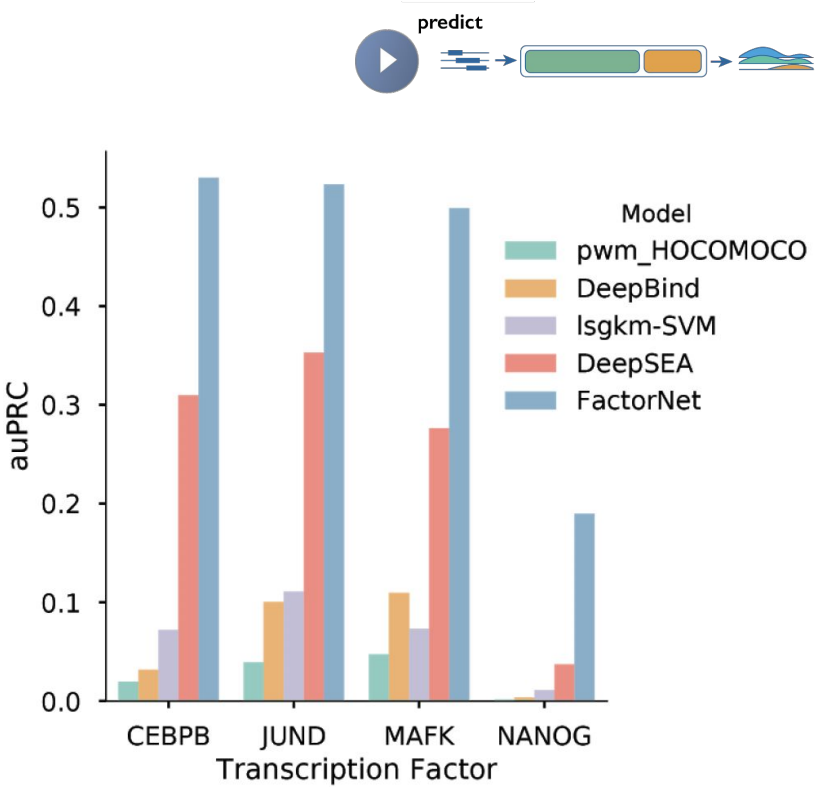


Benchmarking alternative models

Model	Publication	Type	Framework	Input seq. length
pwm_HOCOMOCO	Kulakovskiy et al 2015	Position weight matrix scan	Keras	101bp
DeepBind	Alipanahi et al 2015	Convolutional neural network	Keras	101bp
lsgkm-SVM	Ghandi et al 2014	Support vector machine	LS-GKM	101bp
DeepSEA	Zhou *et al 2015*	Convolutional neural network	PyTorch	1000bp
FactorNet	Quang *et al* 2017	Convolutional and recurrent neural network	Keras	1002bp

Benchmarking alternative models

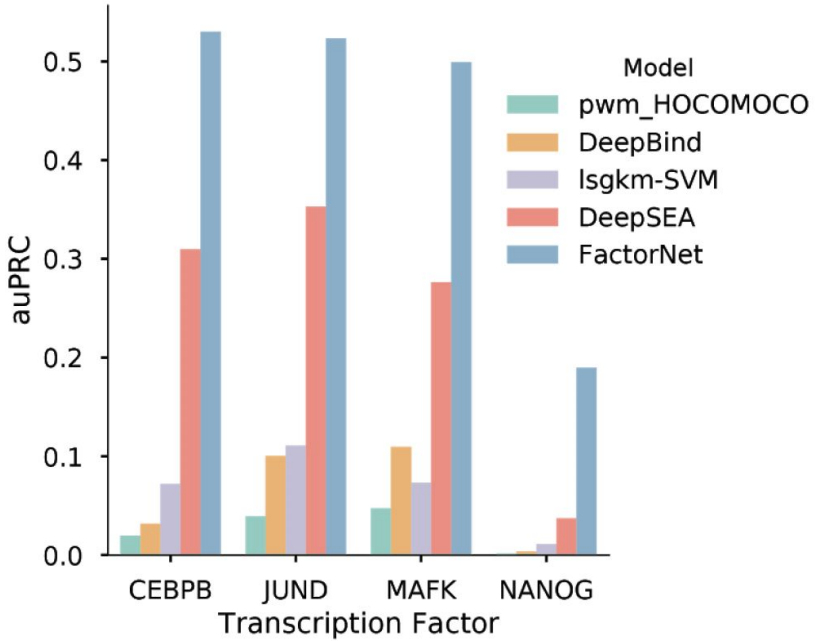
Model	Publication	Type	Framework	Input seq. length
pwm_HOCOMOCO	Kulakovskiy et al 2015	Position weight matrix scan	Keras	101bp
DeepBind	Alipanahi et al 2015	Convolutional neural network	Keras	101bp
lsgkm-SVM	Ghandi et al 2014	Support vector machine	LS-GKM	101bp
DeepSEA	Zhou *et al 2015*	Convolutional neural network	PyTorch	1000bp
FactorNet	Quang *et al* 2017	Convolutional and recurrent neural network	Keras	1002bp



Benchmarking alternative models

Model	Publication	Type	Framework	Input seq. length
pwm_HOCOMOCO	Kulakovskiy et al 2015	Position weight matrix scan	Keras	101bp
DeepBind	Alipanahi et al 2015	Convolutional neural network	Keras	101bp
lsgkm-SVM	Ghandi et al 2014	Support vector machine	LS-GKM	101bp
DeepSEA	Zhou *et al 2015*	Convolutional neural network	PyTorch	1000bp
FactorNet	Quang *et al* 2017	Convolutional and recurrent neural network	Keras	1002bp

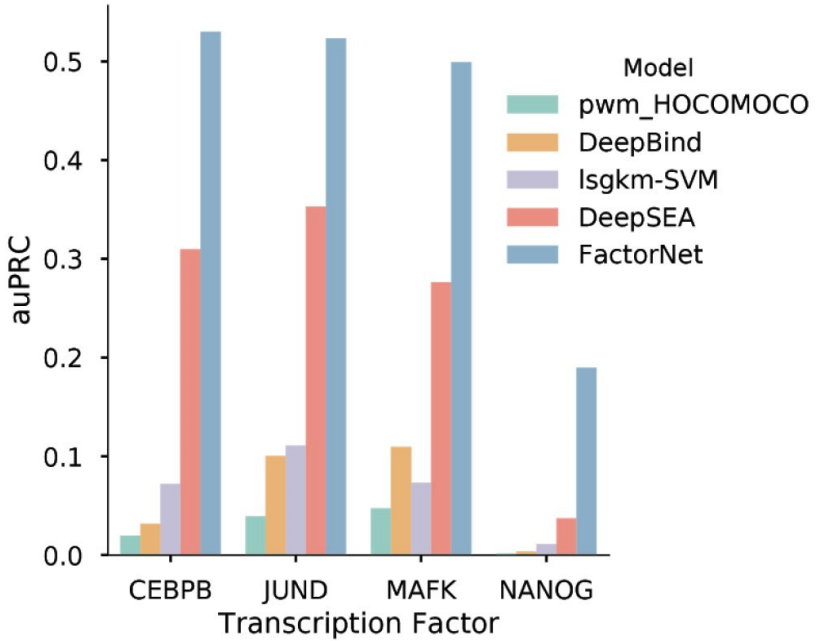
```
# Create and activate a new conda environment with
# all model dependencies installed
kipoi env create <Model>
source activate kipoi-<Model>
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
```



Benchmarking alternative models

Model	Publication	Type	Framework	Input seq. length
pwm_HOCOMOCO	Kulakovskiy et al 2015	Position weight matrix scan	Keras	101bp
DeepBind	Alipanahi et al 2015	Convolutional neural network	Keras	101bp
lsgkm-SVM	Ghandi et al 2014	Support vector machine	LS-GKM	101bp
DeepSEA	Zhou *et al 2015*	Convolutional neural network	PyTorch	1000bp
FactorNet	Quang *et al* 2017	Convolutional and recurrent neural network	Keras	1002bp

```
# Create and activate a new conda environment with
# all model dependencies installed
kipoi env create <Model>
source activate kipoi-<Model>
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
```



<- Works very nicely with workflow-management tools like Snakemake

Why not a single command?

```
predict <model> -i input.data -o output.data
```

```
# Create and activate a new conda environment with
# all model dependencies installed
kipoi env create <Model>
source activate kipoi-<Model>
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
```

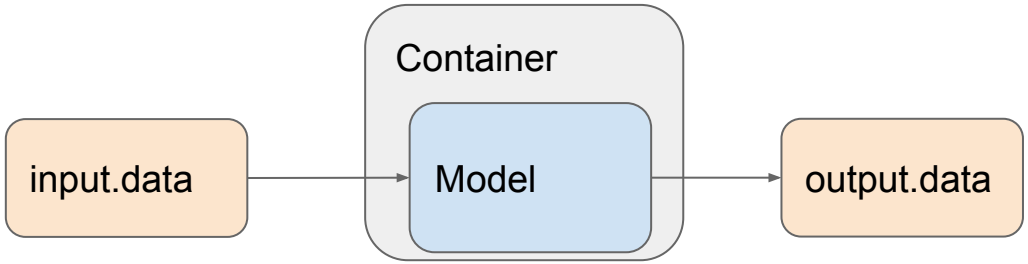
Why not a single command?

```
predict <model> -i input.data -o output.data
```

```
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
  --singularity
```

Why not a single command?

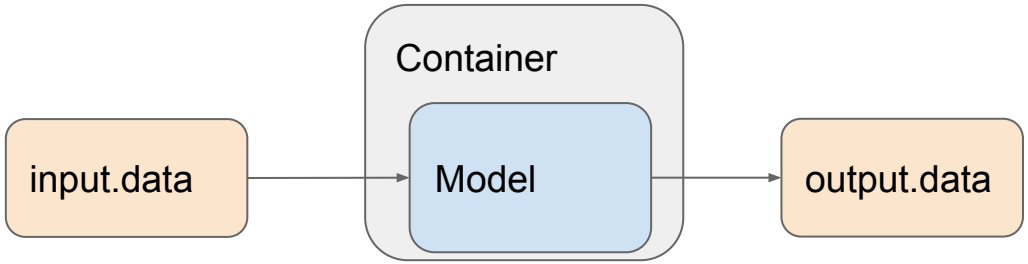
```
predict <model> -i input.data -o output.data
```




```
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
  --singularity
```

Why not a single command?

```
predict <model> -i input.data -o output.data
```



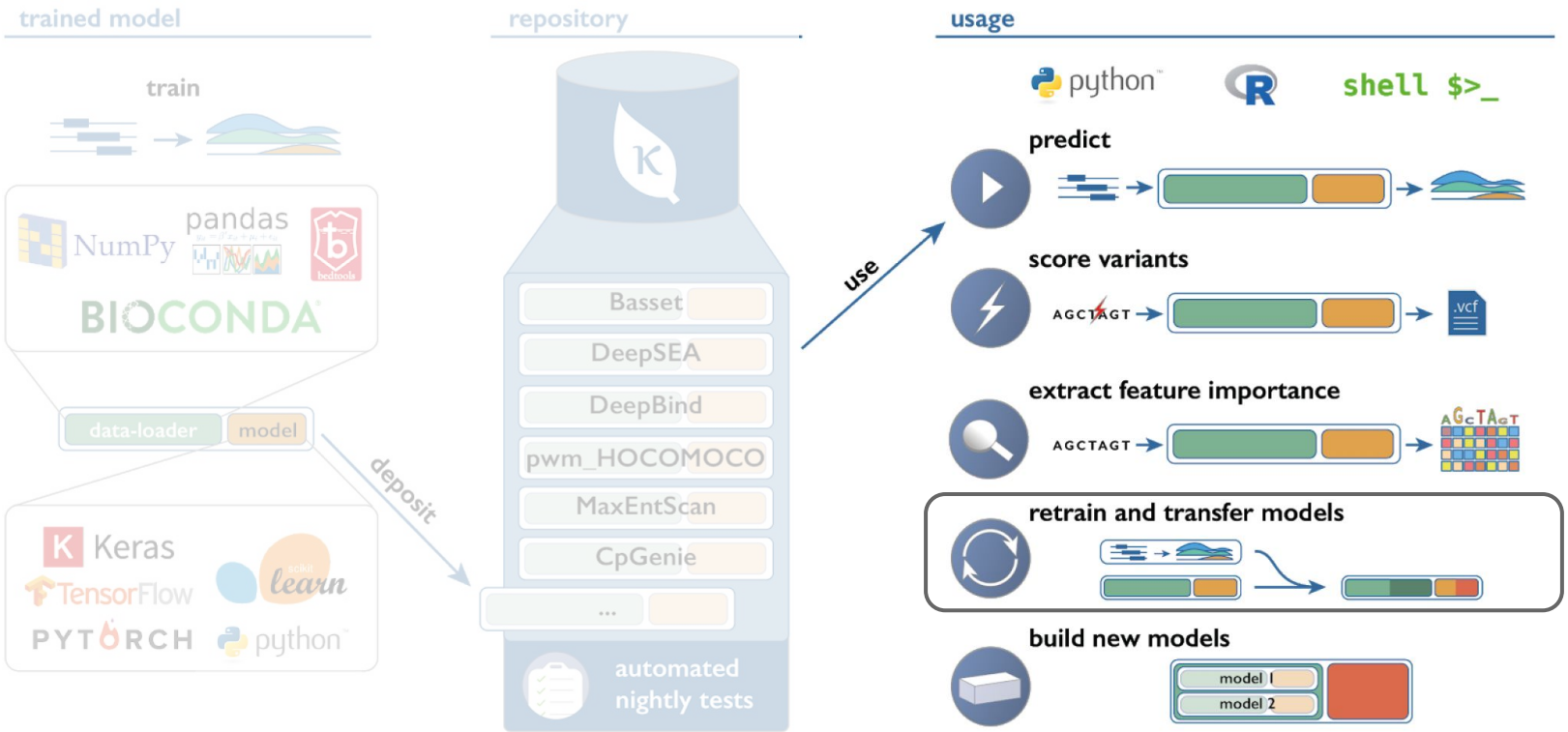

```
# Run model prediction
kipoi predict <Model> \
  --dataloader_args='{
      "intervals_file": "intervals.bed",
      "fasta_file": "hg38.fa"}' \
  -o '<Model>.preds.h5'
  --singularity
```

In-progress:

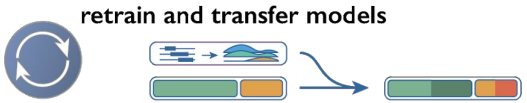
- `--docker`
- `--docker-gpu`

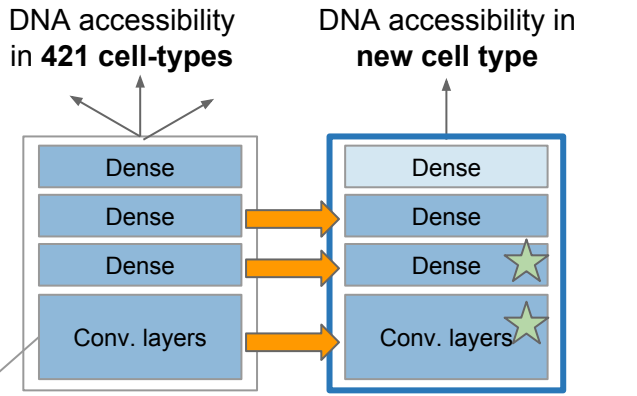
<- Container will be available on NGC

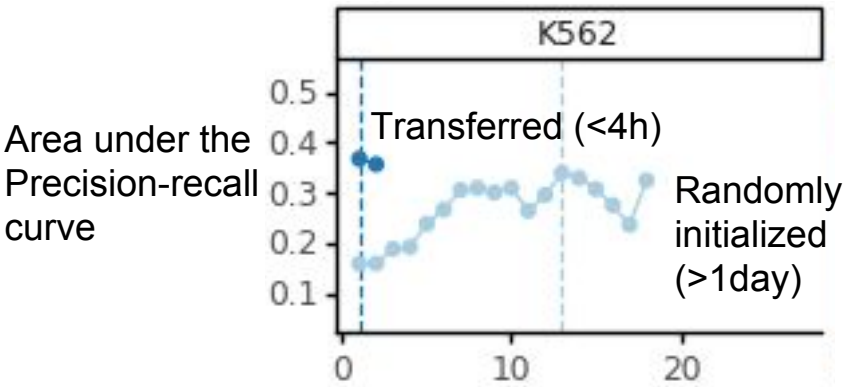
Transfer learning



Transfer learning: Adapting existing models to new tasks



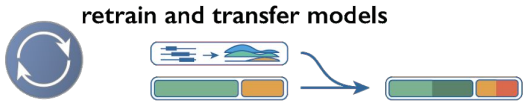




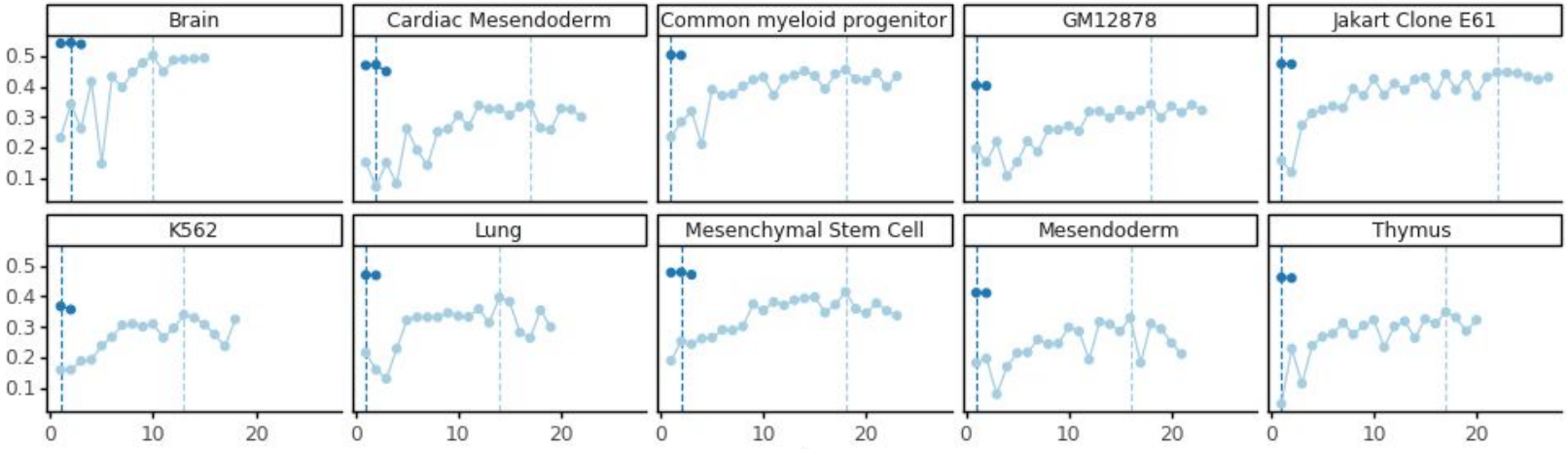
Takes a few days to train
(`Divergent421` model in Kipoi)

See also Kelley et al. Gen. res. 2016

Transfer learning: Adapting existing models to new tasks

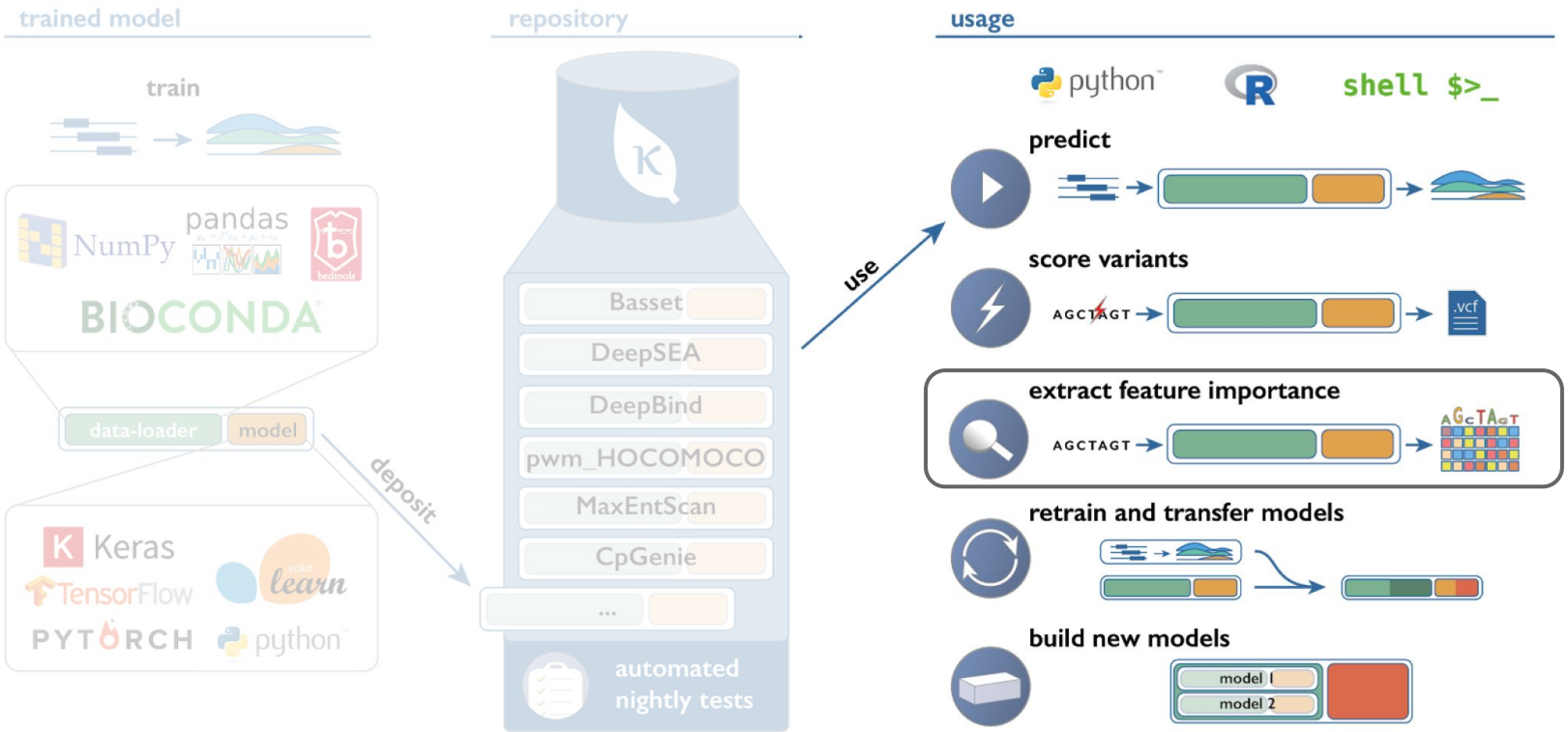


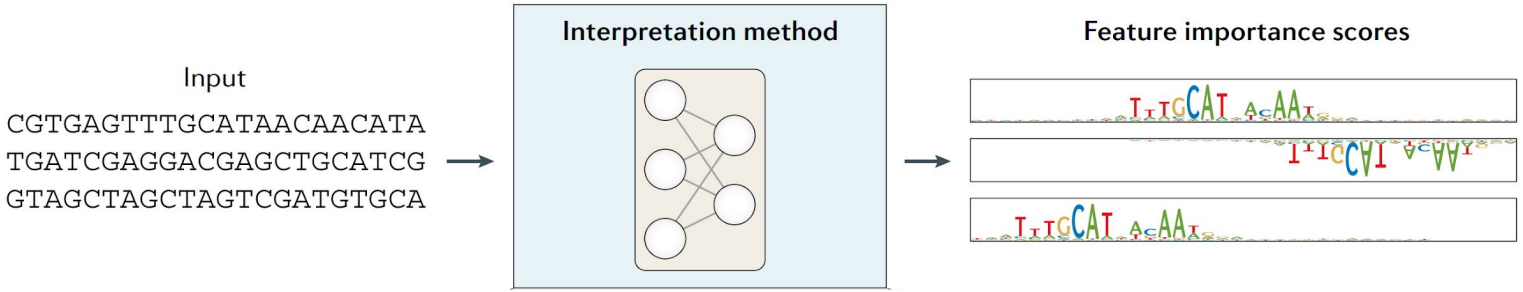
Area under the Precision-recall curve



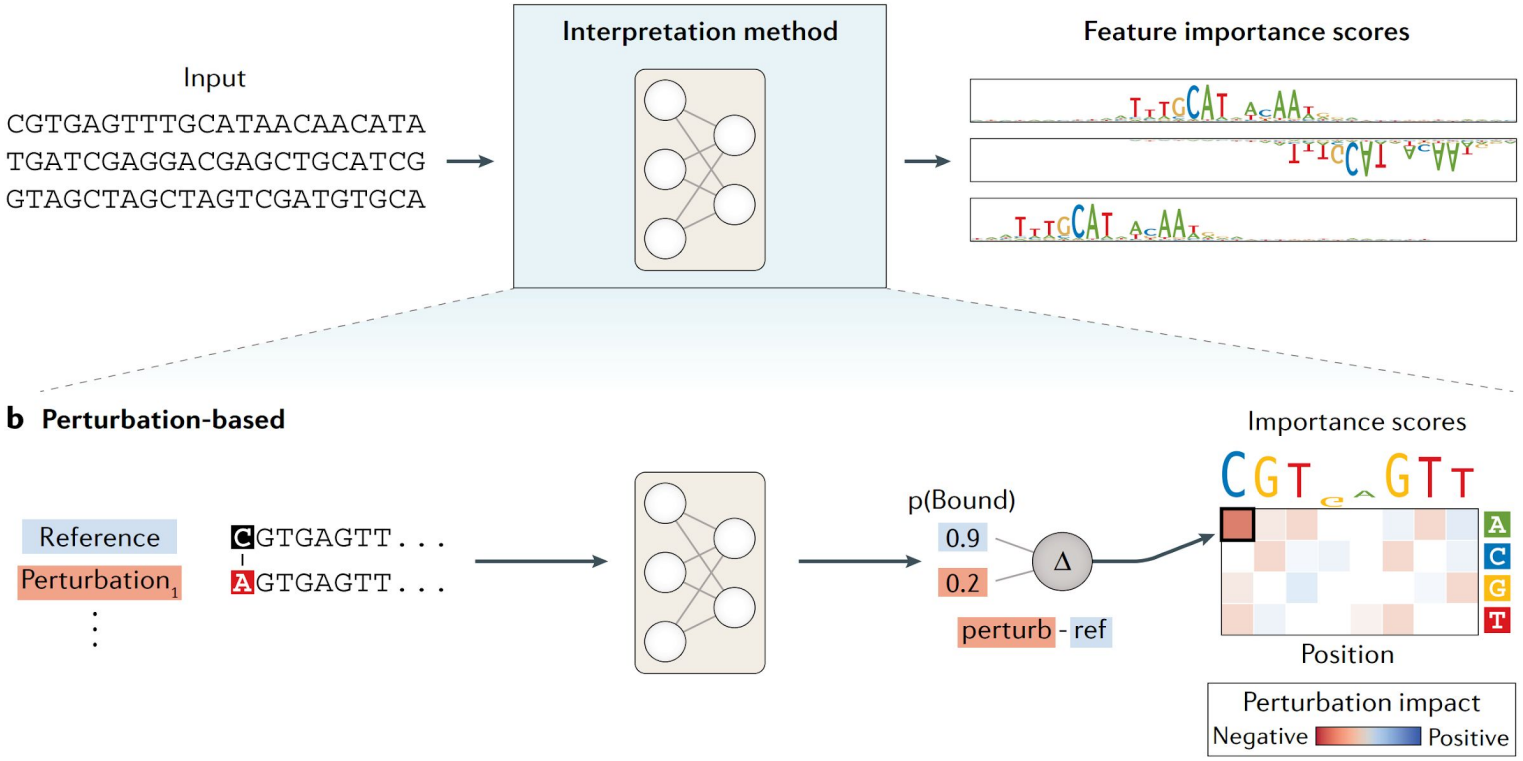
See also Kelley et al. Gen. res. 2016

Interpreting models

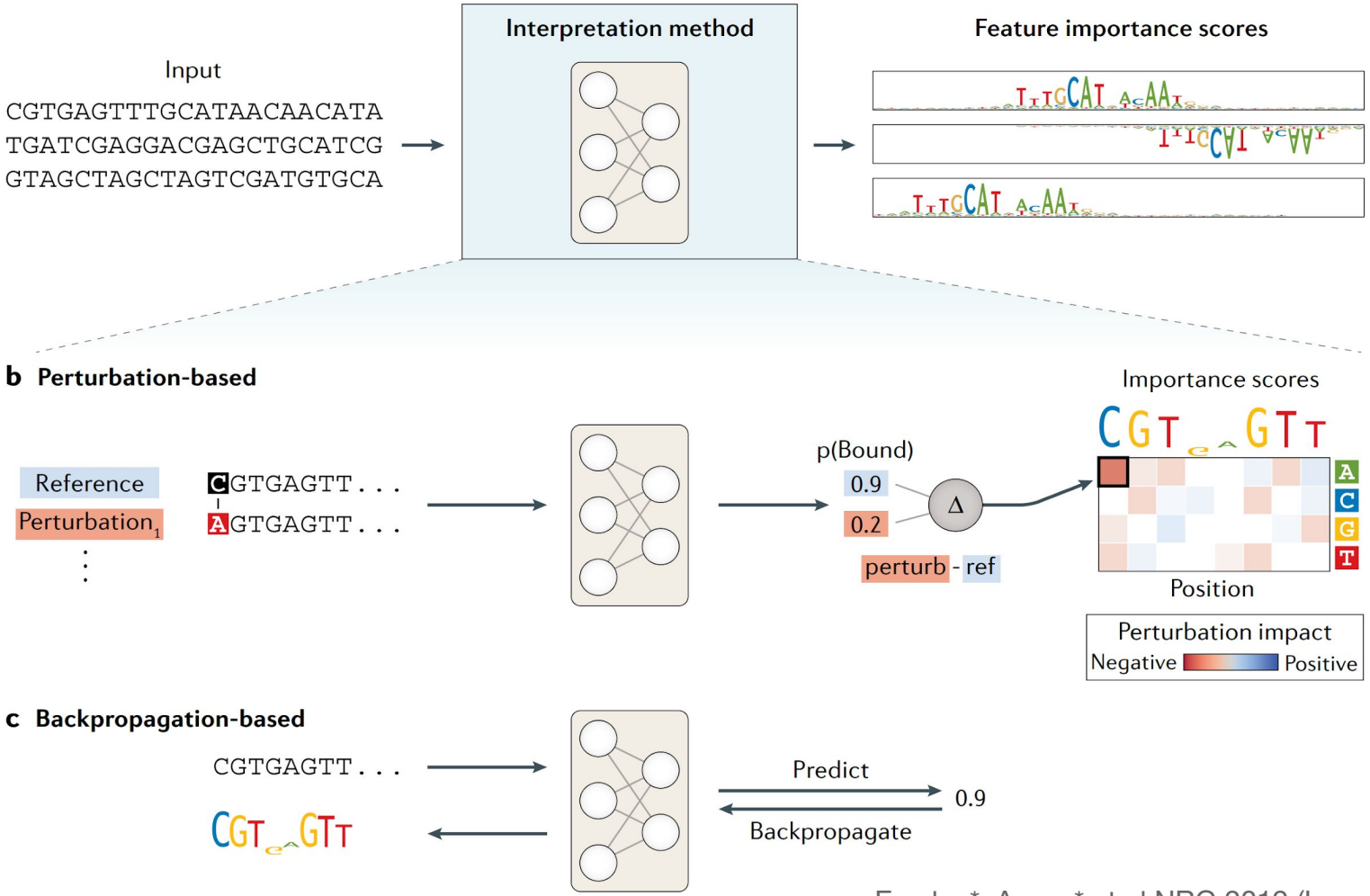




Eraslan*, Avsec* et al NRG 2019 (In press)



Eraslan*, Avsec* et al NRG 2019 (In press)



Eraslan*, Avsec* et al NRG 2019 (In press)

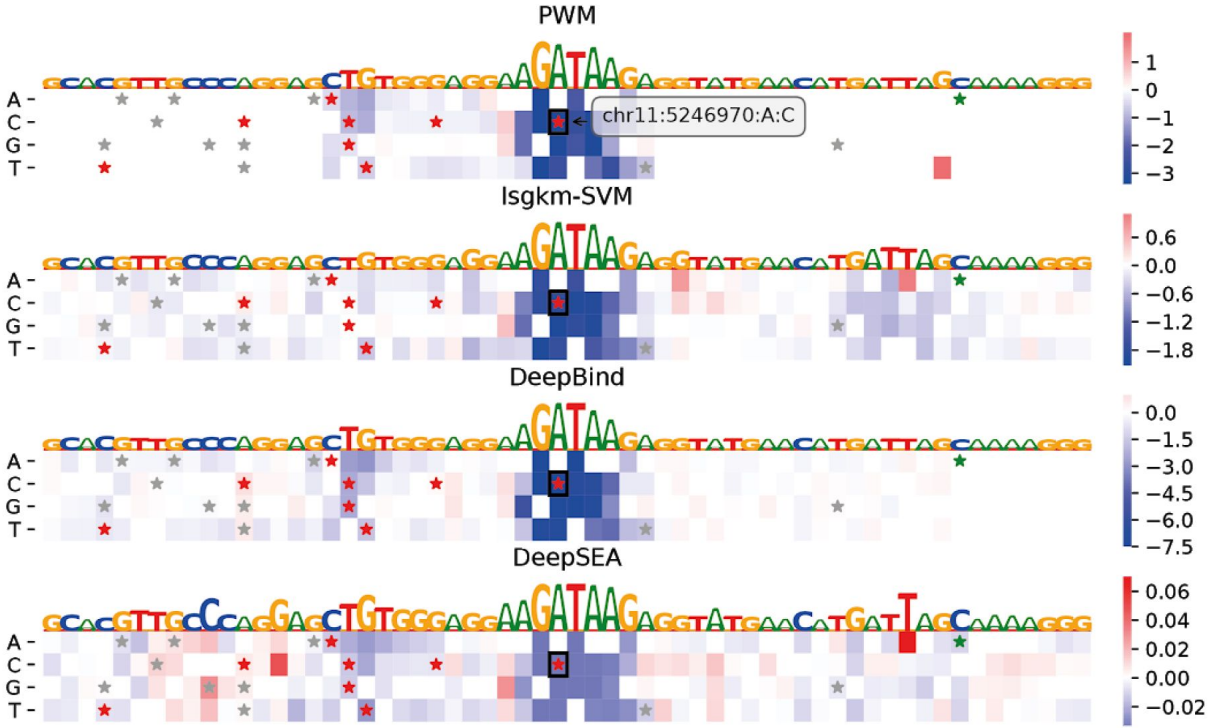
kipoi-interpret

```
# Python
import kipoi
from kipoi_interpret.importance_scores.gradient import GradientXInput

model = kipoi.get_model("model”)

imp_score = GradientXInput(model)

scores = imp_score.score(seqs)
```

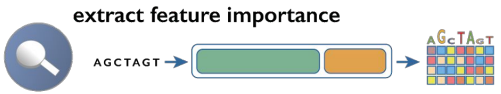
```
# CLI
kipoi interpret create_mutation_map \
  <Model> \
  --dataloader_args='{
      “intervals_file”: “intervals.bed”,
      “fasta_file”: “hg38.fa”}' \
  -o 'mmap.h5'
```

Feature importance score plugin

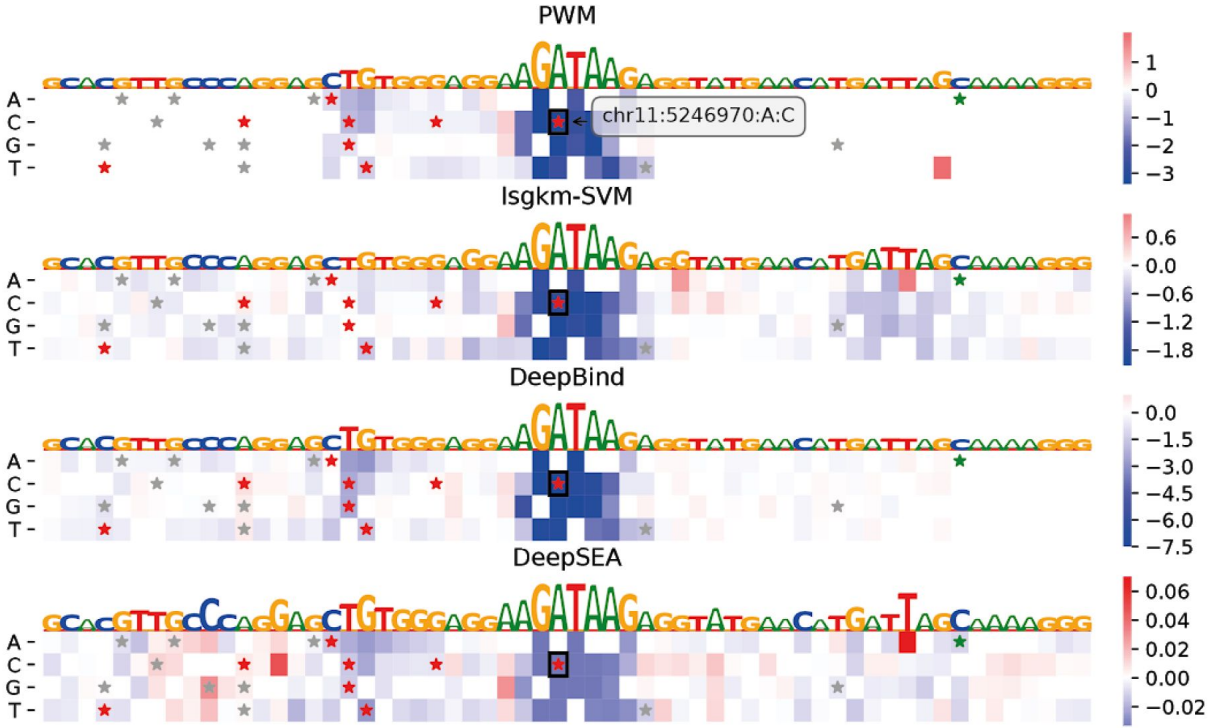
- Variant rs35703285, near the beta globin gene HBB, is pathogenic (ClinVar) and linked to Beta thalassemia



Feature importance score plugin



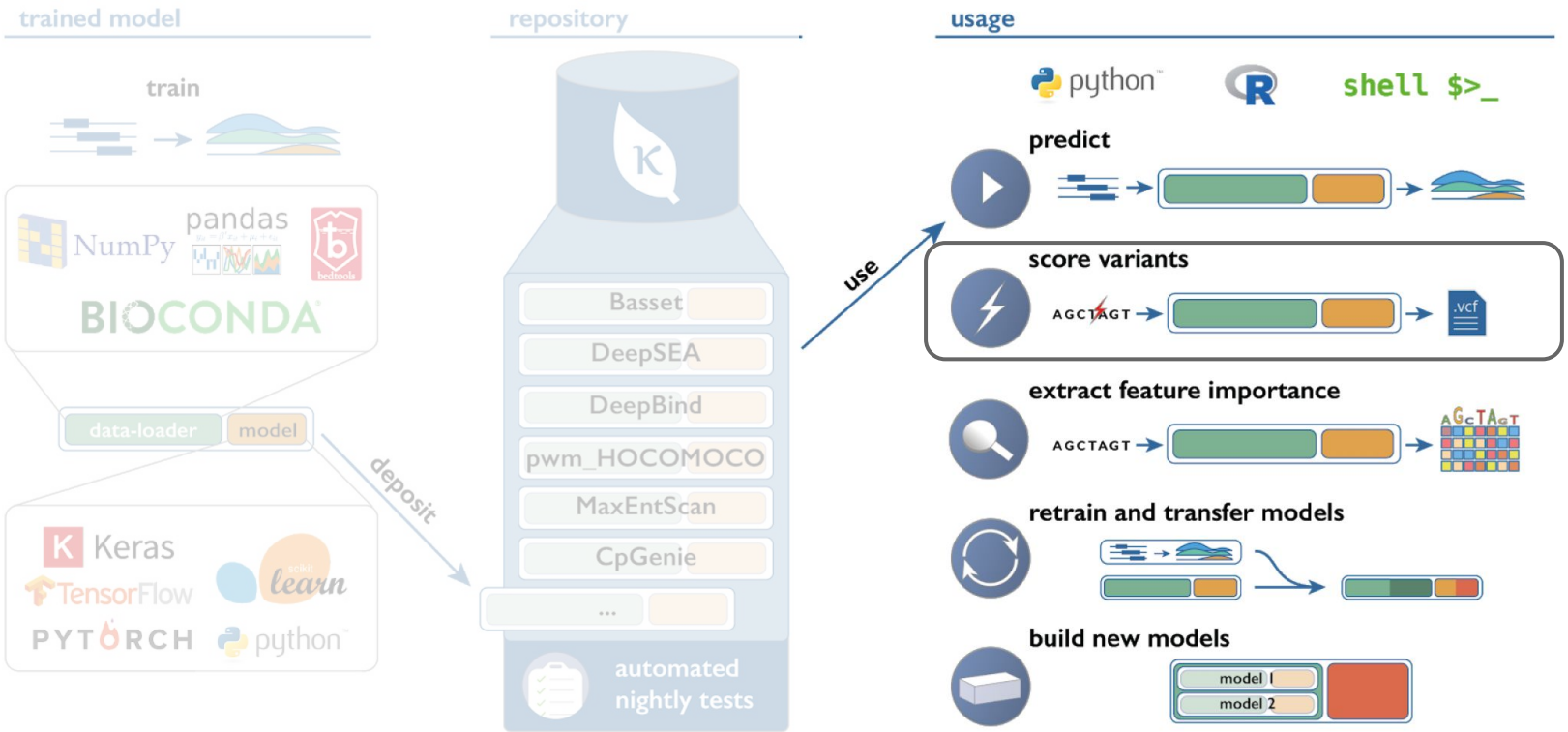
- Variant rs35703285, near the beta globin gene HBB, is pathogenic (ClinVar) and linked to Beta thalassemia



Methods

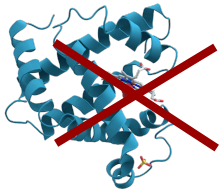
- ISM
- grad
- input*grad
- DeepLift

Scoring genetic variants

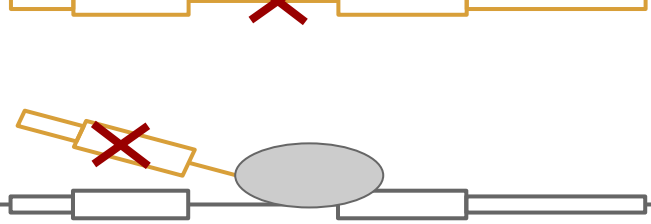


Scoring genetic variants

- Each one of us carries ca 1,000,000 variants



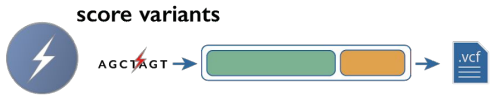
```
#CHROM   POS        ID REF ALT
           …
chr22    41320486   .   G   T
           …
```



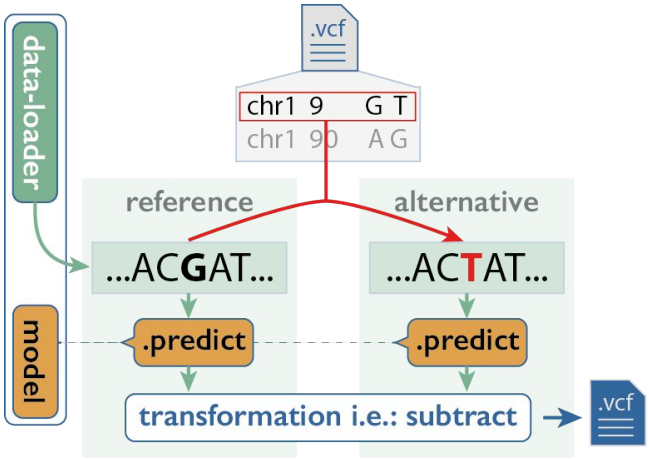
Patient
Reference

```
atcttatatatcatgatatggatacgcatagatcatgactcaggatacg
atcgtatatatcatgatatggatacgcatagatcatgactcaggatacg
```

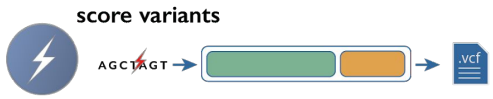
Variant effect prediction plugin



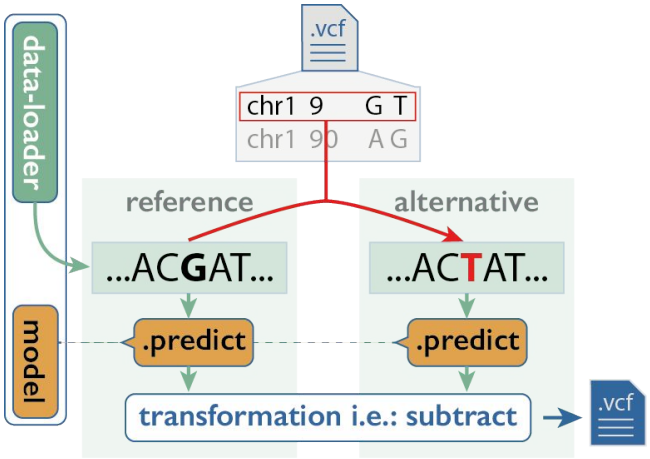
- In-silico mutagenesis



Variant effect prediction plugin



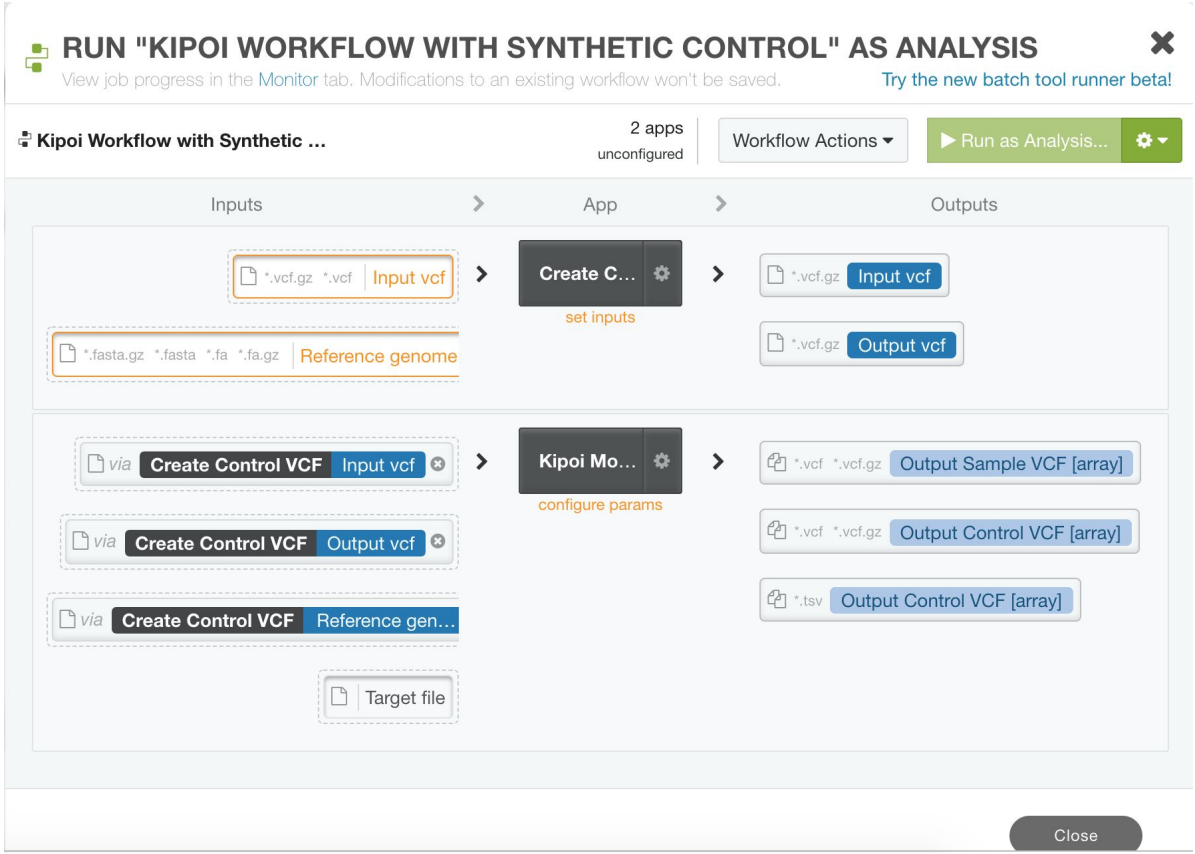
- In-silico mutagenesis



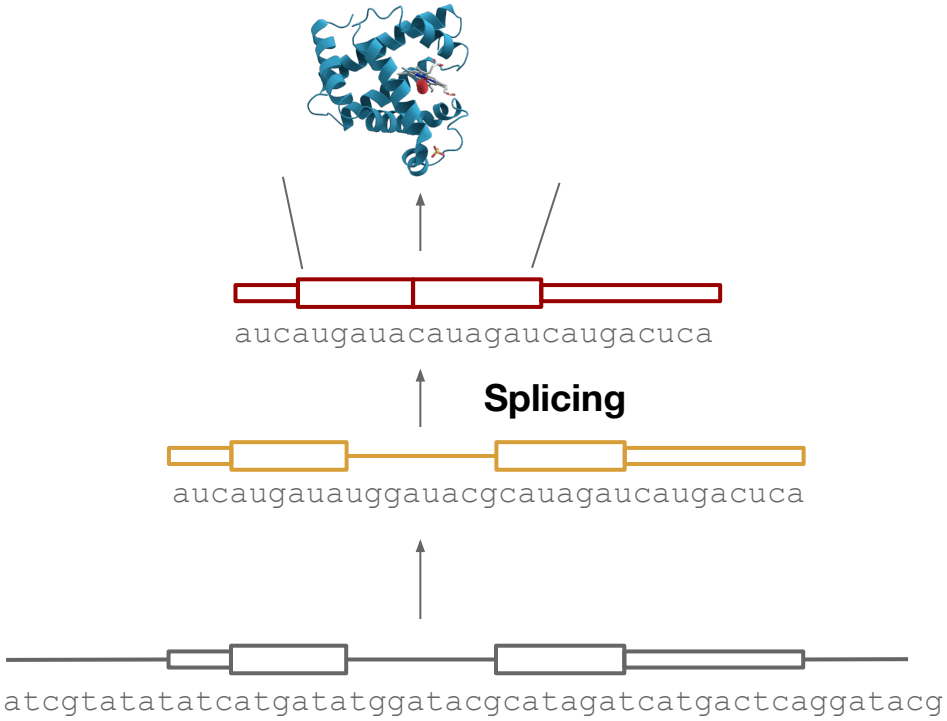
- Supported by 12/20 model groups, runnable on VCF files

```
# Annotate VCF file with variant scores
kipoi veff score_variants <Model> \
  --dataloader_args='{
     "fasta_file": "hg38.fa"}' \
  --vcf_path 'input.vcf' \
  -o 'annotated.vcf'
```

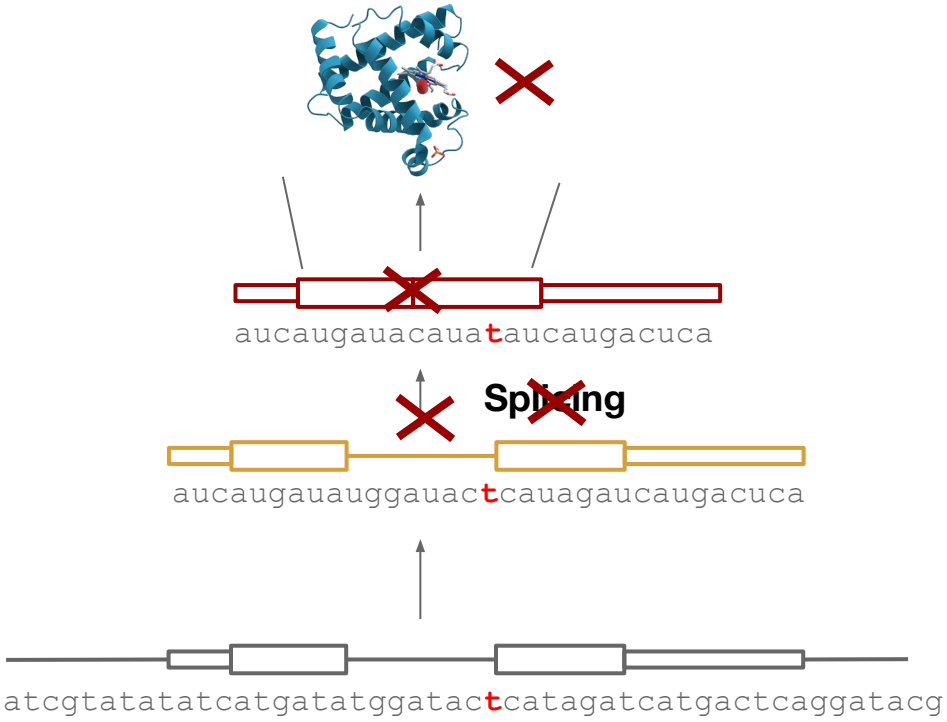
Kipoi variant scoring as a DNAnexus applet



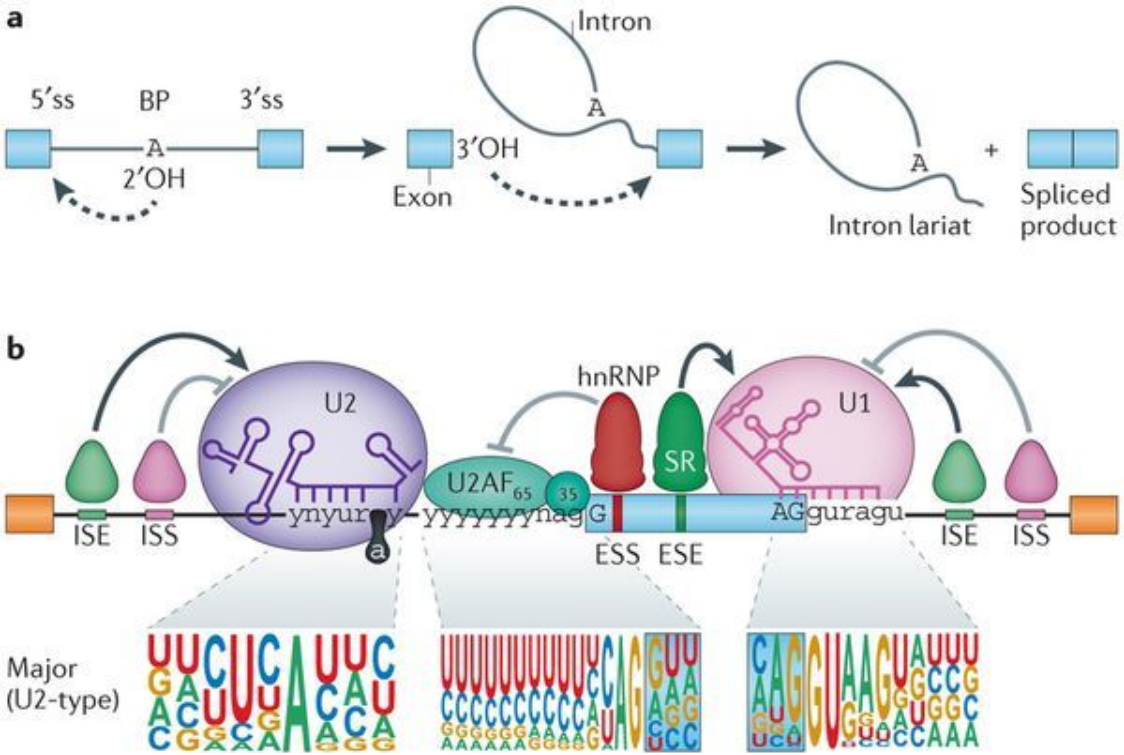
Splicing: an essential step of protein production



Splicing: an essential step of protein production

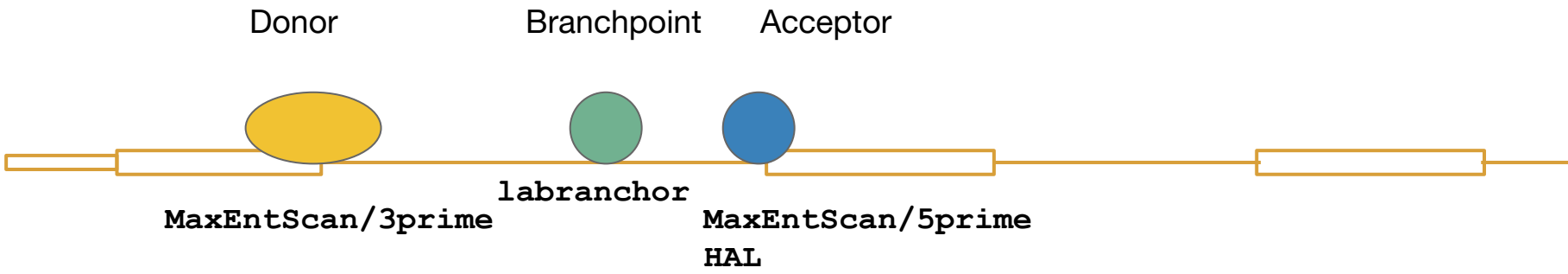


Splicing is a complex, multi-step process

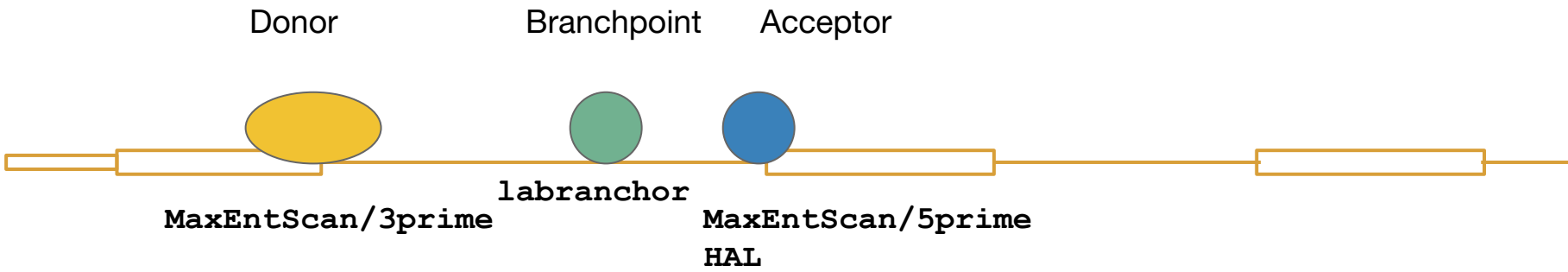


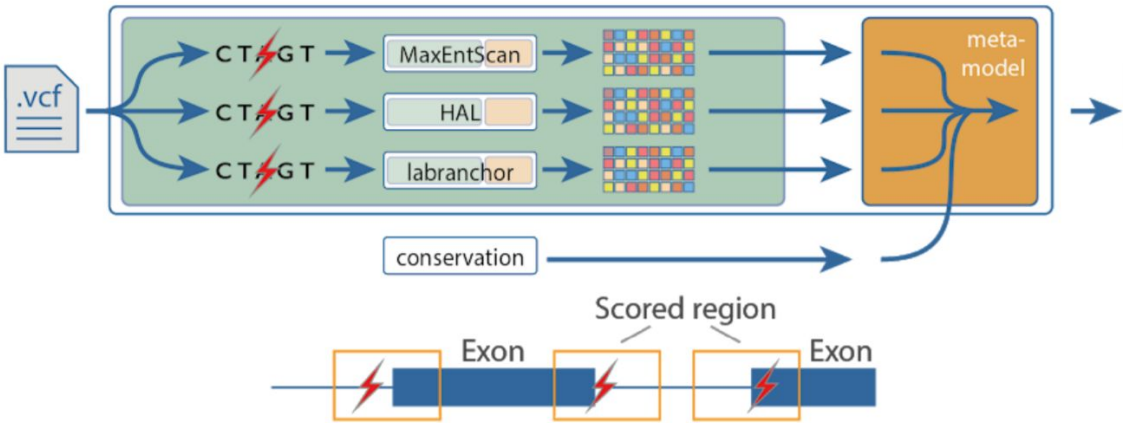
Scotti & Swanson, 2016 NRG

Different models model different regions



Different models model different regions

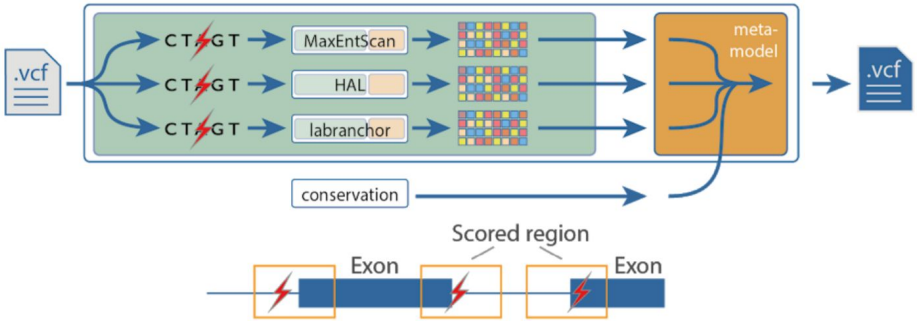


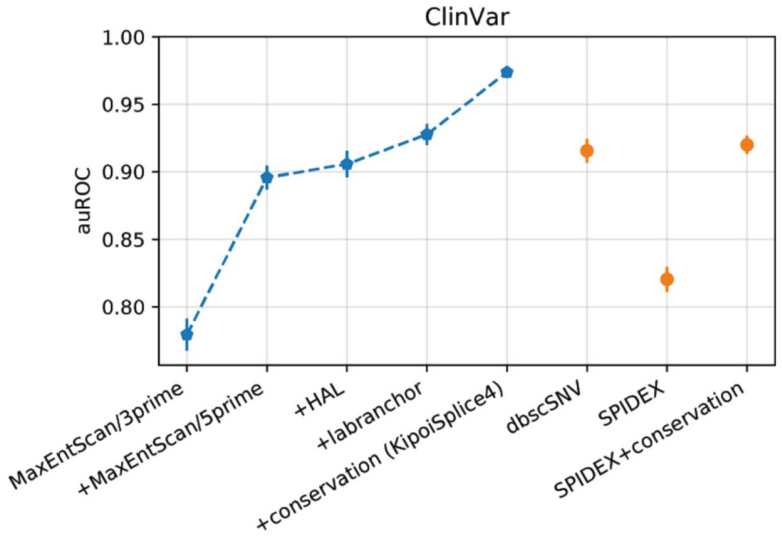


Binary classification:

- Pathogenic (ClinVar)
- Benign (ClinVar)

Ensemble model predicting pathogenic variants near splice sites





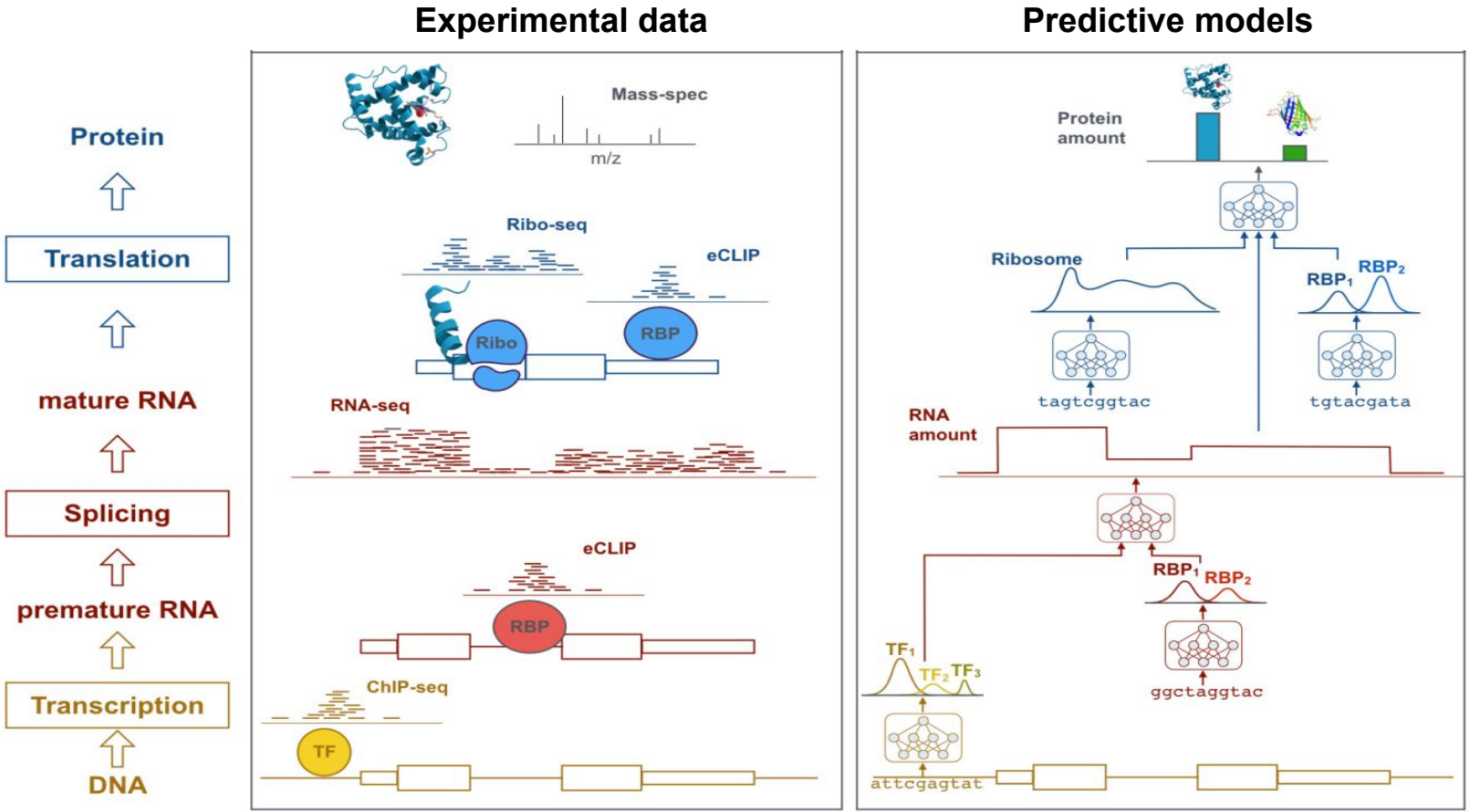
Kipoi models:

KipoiSplice/4
KipoiSplice/4cons
MMSplice

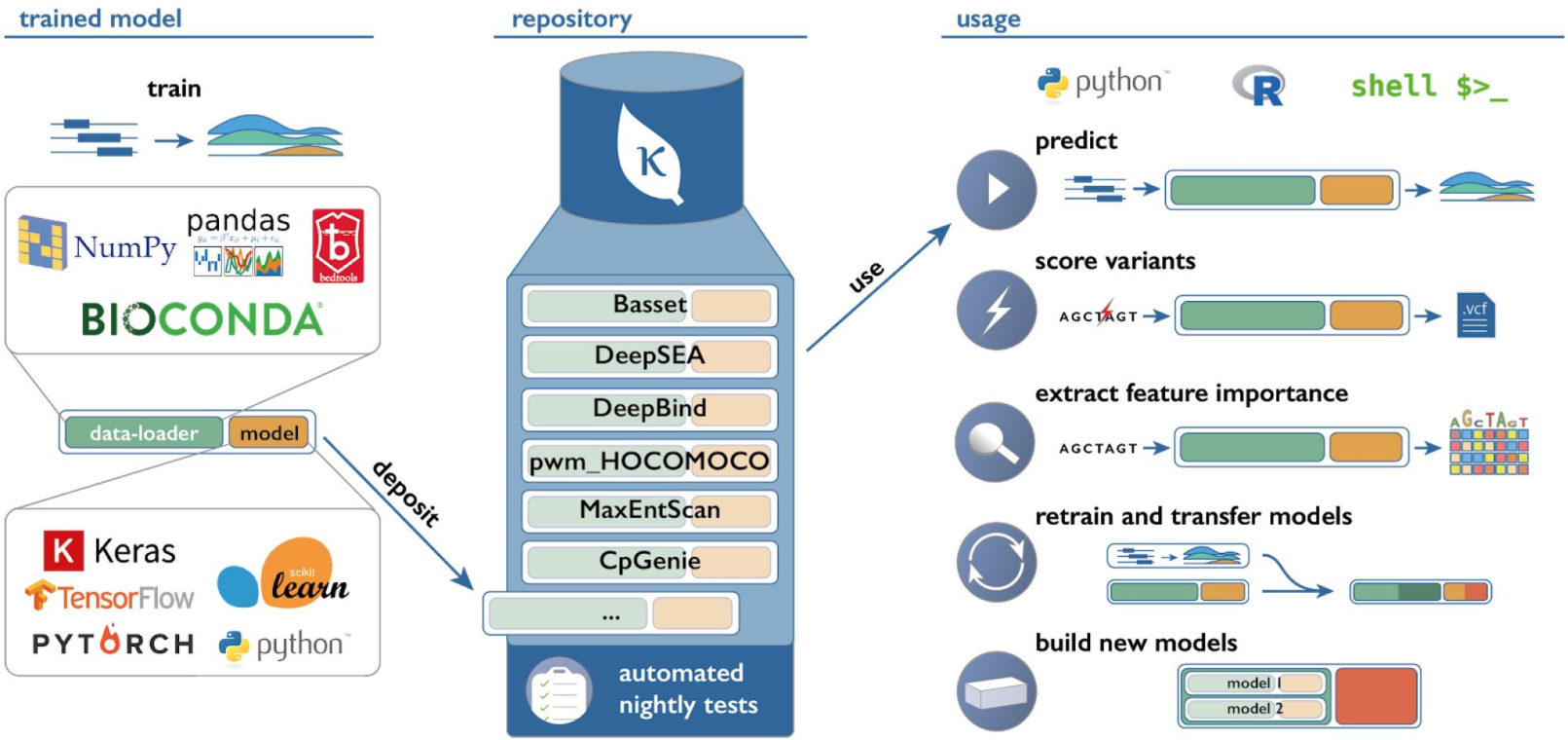
See also MMSplice from Cheng et al., Genome Biology, CAGI Splicing challenge 2018 winner

Summary

Learning the regulatory steps



Kipoi.org [Kípi]



with A. Kundaje, Stanford, and O. Stegle, DKFZ

Avsec et al, Nature Biotechnology (In press)

Kipoi.org
@KipoiZoo

Acknowledgements

Gagneur lab
TU Munich
- Julien Gagneur
- Jun Cheng

Kundaje lab
Stanford
- Anshul Kundaje
- Avanti Shrikumar
- Nancy Xu
- Abhimanyu Banerjee
- Chuan Sheng Foo

Stegle lab
Cambridge EMBL-EBI
- Oliver Stegle
- **Roman Kreuzhuber**
- **Thorsten Beider**
- Lara Urban

Nvidia
- Jonny Israeli
- Fernanda Foertter
- Gary Dunn
- Adam Simpson

DNA Nexus
- Jason Chin
- Maria Simbirsky
- Andrew Carroll

Roman Kreuzhuber

Thorsten Beider

Kipoi: model zoo for genomics

Žiga Avsec

PhD candidate, Technical University of Munich

www.gagneurlab.in.tum.de

@gagneurlab, @KipoiZoo, @Avsecz