Identification of enhancer/gene (E/G) relationships: state-of-the-art methods

Sarah Djebali
IRSD, INSERM U1220, Toulouse

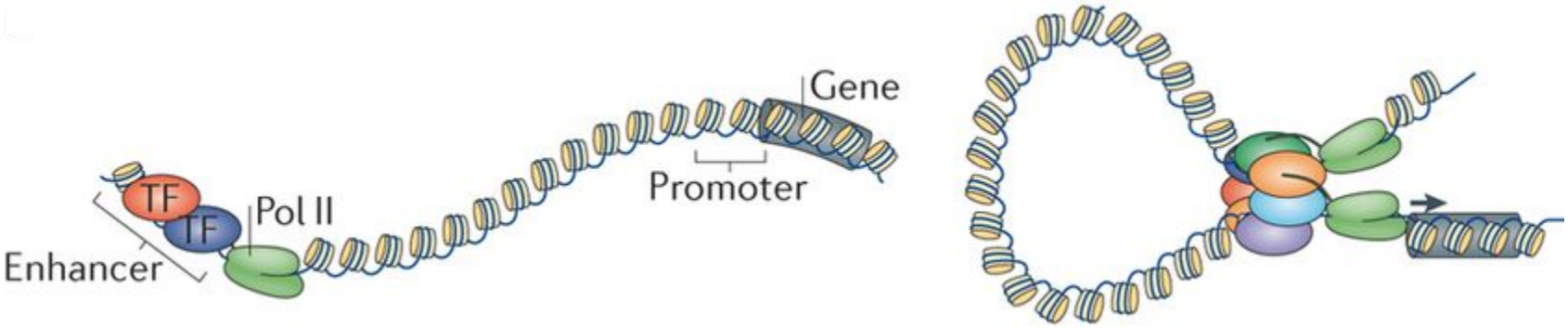
sarah.djebali@inserm.fr

Biopuces, INRAE - September 17th 2020

Outline

- Introduction
  - Definitions: enhancers, promoters and enhancer/gene (E/G) relationships
  - 3 broad approaches to identify them genome-wide in a cell type specific manner
- Focusing on methods by functional links
- The 4 most-promising methods
  - The random forest (RF) concept
  - Going over each of the 4 methods
- What's next

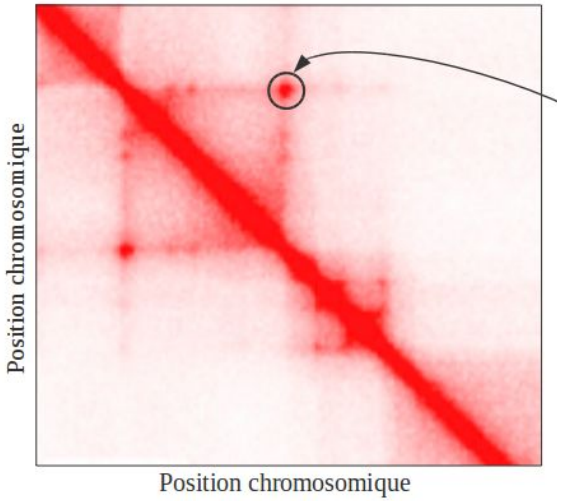
Enhancer and promoter regulatory elements, and the enhancer/gene (E/G) relationship



- Far away on the genome (**1D**) but physically close by (**3D**)
  - Sometimes several hundreds kb away (could be 1-2Mb)
- Enhancers can act from upstream or downstream of the gene
- An enhancer can activate several genes and a gene can be activated by several enhancers

How are enhancer/gene (E/G) relationships automatically identified ?

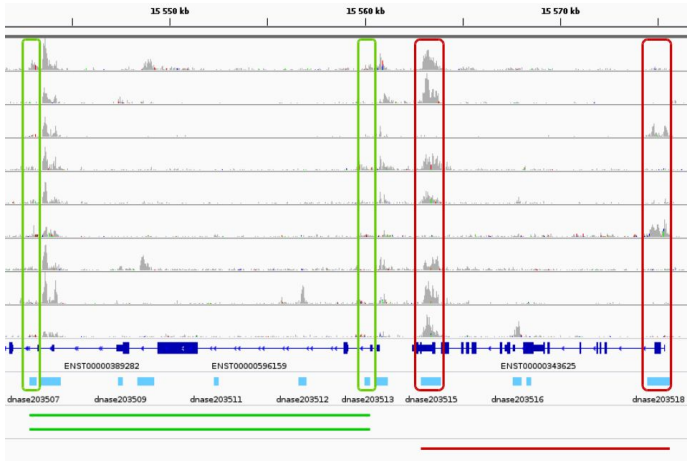
Spatial link (3D)



Ex : HiC, RNA pol II ChIA-PET, promoter capture HiC

Pb: costly and difficult to implement

Functional link (1D)

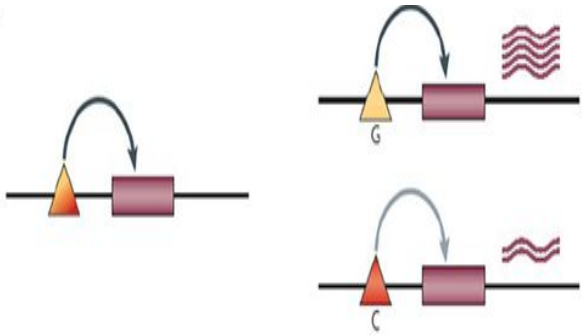


Ex : Correlation between DNA accessibility or expression at two regions across X cell types

Pb: no robust method

+ Comparative genomics
+ Genetic screening

Genetic link (1D)

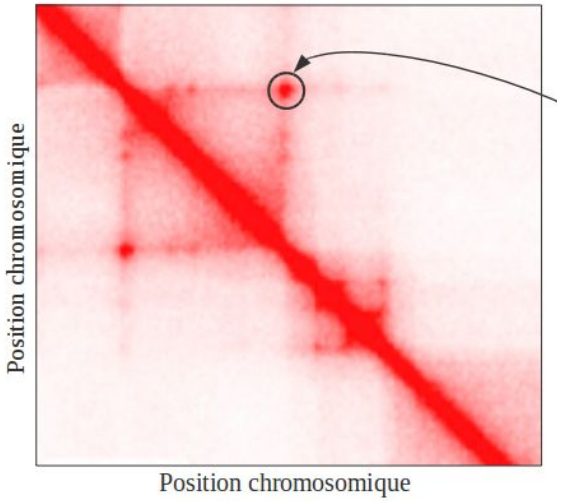


Ex : Expression QTL (eQTL), splicing QTL (sQTL), ...

Pb: costly (need genome-wide express for many individuals)

How are enhancer/gene (E/G) relationships automatically identified ?

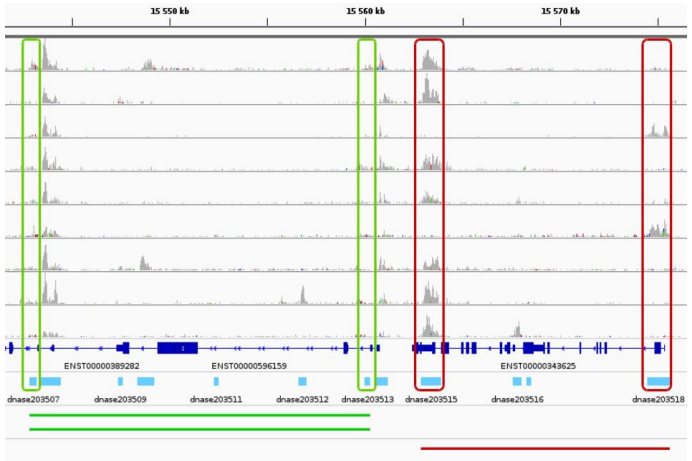
Spatial link (3D)



Ex : HiC, RNA pol II ChIA-PET, promoter capture HiC

Pb: costly and difficult to implement

Functional link (1D)

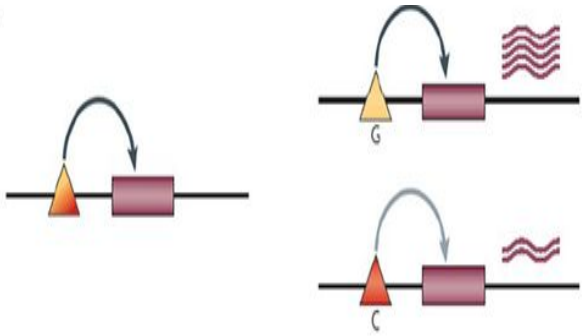


Ex : Correlation between DNA accessibility or expression at two regions across X cell types

Pb: no robust method

+ Comparative genomics
+ Genetic screening

Genetic link (1D)



Ex : Expression QTL (eQTL), splicing QTL (sQTL), ...

Pb: costly (need genome-wide express for many individuals)

Functional link (1D) methods

Broad category of functional link (1D) method	**(HT) functional 1D data taken as input for the prediction**	**Example of method / Type of method**
Non supervised / heuristic methods	- Few different data types - Very big number of different cell types	Correlation between chromatin accessibility at two regions separated by a distance of x across several cell types
Supervised machine learning methods	- Many different data types - A single cell type, the one for which the E/G prediction needs to be done	Training considering 3D relationships as ground truth and learning the combination of 1D data features that are associated with the true relationships

20 chronologically ordered methods from the literature (from 2011)

- For each of the 20 methods, provide:
  - Number (1-20)
  - Name
  - Broad class (unsupervised / supervised)
  - Brief description
  - Code repository (NA if not available)
  - Publication reference

#	Program name	Class	Description	Code website	Reference
1	Rodelsperger's method	Supervised	Random Forest using 4 features: distance, synteny, functional similarity and protein-protein interactome proximity, between the TF binding at enhancer and the target gene, and trained on 31 examples from the literature. Says whether a gene is the target of an enhancer less distant than 2Mb	NA	Rodelsperger et al, NAR, 2011
2	Histone mark activity to gene expression correlation across cell types	Unsupervised	Correlation between enhancer cluster activity (calculated from histone marks) and expression of gene at 5kb to 125kb distance across 9 cell lines	NA	Ernst et al, Nature, 2011
3	Enhancer to promoter activity correlation across cell types	Unsupervised	Iterative correlation between enhancer and promoter activities (calculated from histone marks or polII) across 19 mouse cell types, defining EPUs (no max distance but spearman correlation > 0.23)	NA	Shen et al, Nature, 2012
4	DNA accessibility pairwise correlation across cell types	Unsupervised	Correlation between promoter distal and promoter DHS peak accessibility across 79 cell types (at less than 500kb, correlation > 0.7)	NA	Thurman et al, Nature, 2012
5	DNA accessibility to gene expression correlation across cell types	Unsupervised	Correlation between promoter distal DHS peak accessibility and gene expression across 72 cell types (less than 100kb, permutation p-val < 0.05)	NA	Sheffield et al, Genome Research, 2013

#	Program name	Class	Description	Code website	Reference
6	SVM-MAP for methylation to expression relationship across cell types	Supervised	SVM trained on methylation signal at promoter and gene expression in 58 cell types, and applied to the same but at promoter distal sites	NA	Aran et al, Genome Biology, 2013
7	CAGE signal pairwise correlation	Unsupervised	Correlation between promoter (CAGE-directional) and enhancer (CAGE-bidirectional) CAGE peak signal across 808 cell types	NA	Andersson et al, Nature, 2014
8	PreSTIGE (predicting specific tissue interactions of genes and enhancers)	Unsupervised	Pairs cell type specific enhancers (H3K4me1 in 12 cell types) and cell type specific genes when not separated by a +100kb distal CTCF site	NA (only galaxy)	Corradin et al, Genome Research, 2014
9	IM-PET (integrated method for predicting enhancer targets)	Supervised	Random Forest using 4 features: distance, synteny, enhancer (CSI-ANN score) to promoter (FPKM) activity and enhancer TF to promoter expression correlations across 12 cell types, and trained on PolII ChIA-PET stringent connections with p300 signal exclusively at enhancer from 2 cell lines	http://tanlab4generegulation.org/IM-PET.html	He et al, PNAS, 2014
10	ELMER (Enhancer Linking by Methylation/ Expression Relationships)	Unsupervised	For cancer hypomethylated probes (vs normal) and 10 genes up and down of it, tests whether gene expression is higher in samples where methylation is lower (Mann-Whitney test for two extreme sets of samples)	https://bioconductor.org/packages/release/bioc/html/ELMER.html	Yao et al, Genome Biology, 2015

#	Program name	Class	Description	Code website	Reference
11	RIPPLE (Regulatory Interaction Prediction for Promoters and Long-range Enhancers)	Supervised	Minimal classifier based on training Random Forests (on each cell line) and Group Lasso-based Multi-task learning (on all cell lines) and using 5C data for positives, 23 epigenome datasets (8 histone marks, 13 TF ChIP-seq, DNAse-seq, RNA-seq) as features, a precomputed set of enhancers and promoters and a distance between 2.5kb and 1Mb	https://github.com/Roy-lab/RIPPLE	Roy et al, NAR, 2015
12	TargetFinder	Supervised	Gradient boosting in each cell type based on high-resolution HiC data (positives), and hundreds of epigenomic data around promoters, enhancers (known in advance) and the window between them + TF-gene functional similarity + synteny (as features) (20 times more negatives than positives and with same distance distribution). 10kb-2Mb distance	https://github.com/shwhalen/targetfinder	Whalen, Truty, Pollard, Nature Genetics, 2016
13	JEME (Joint Effect of Multiple Enhancers)	Supervised	1) Multiple linear regression to get all less than 1Mb possible enhancer/promoter interactions based on DNAse-seq in multiple cell types and 2) cell type specific interactions using Random Forests trained on polII ChIA-PET data (positives) and using 3 histone marks and DNAse-seq at promoter, enhancer and in the window between them as features	https://github.com/yiplabcuhk/JEME	Cao et al, Nature Genetics, 2017

#	Program name	Class	Description	Code website	Reference
14	PEP (Predicting Enhancer–Promoter interactions)	Supervised	Gradient boosting in each cell type based on high-resolution HiC data (positives and negatives, 1/20 ratio), and TFBS and sequences in predefined enhancers and promoters. 10kb-2Mb distance	https://github.com/ma-compbio/PEP	Yang et al, Bioinformatics, 2017
15	FOCS	Supervised	Multiple linear regression of chromatin signal (DNAse-seq or CAGE or GRO-seq) on the k closest enhancers of a promoter	https://github.com/Shamir-Lab/FOCS	Hait et al, GB, 2018
16	DeepTACT (Deep neural networks for chromatin conTACTs prediction)	Supervised	Bootstrapping deep learning method that integrates genome sequences and DNA accessibility t predict 3D contacts (from HiC used for training)	https://github.com/liwenran/DeepTACT	Li, Wong, Joang, NAR, 2019
17	ABC (Activity by Contact) model	Unsupervised	Heuristic method that computes the score of an enhancer/gene relationship by multiplying the activity of the enhancer (as defined by DNAse-seq and H3K27ac) by its contact with the gene (as defined by HiC or just distance) and normalizing it by the sum of ABC scores for all enhancers close to the gene	https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction	Fulco et al, Nature Genetics, 2019

#	Program name	Class	Description	Code website	Reference
18	3DPredictor	Supervised	Gradient boosting to make quantitative prediction of 3D structure (high-resolution HiC) based on CTCF ChIP-seq and RNA-seq data, and distance	https://github.com/labdevgen/3Dpredictor	Belokopytova et al, Genome Research, 2020
19	Average rank between DNA accessibility to gene expression correlation and distance methods	Unsupervised	Method that combines the DNA accessibility to gene expression correlation and the distance methods and provides the average rank between the two as a score	https://github.com/weng-lab/BENGI/tree/master/Scripts/Unsupervised-Methods	Moore et al, Genome Biology, 2020
20	EPIVAN (Promoter-Enhancer Interaction Predictor with pre-trained Vector and Attention based neural Networks)	Supervised	Attention based neural network with pre-trained vectors trained on known EPIs and the sequences of known E and P (as well as pretrained vectors)	https://github.com/hzy95/EPIVAN	Hong et al, Bioinformatics, 2020

Some observations about the methods

- From past to present:
  - Supervised more frequent than unsupervised
  - Code available more often (good!)
- Ground truth:
  - 3D data (polII ChIA-PET or prom capture HiC) for all meth
  - eQTL or/and genetic screening additionally for some meth
- But different ways of using it (un/supervised=after/before)
- Very different number of (cell types), distance and correlation thresholds for unsupervised methods
- Different ways of making +/- sets for supervised methods

The prerequisites of a good method

Prerequisite name	Prerequisite description
CODE	has a freely available code that can be run on UNIX and that is not dedicated to certain kinds of samples (e.g. cancer)
CTSPEC	able to predict in a particular cell type
MULTI	able to predict multi-multi relationships
CONSIST	able to use the same input data for predicting enhancers, promoters and E/G

#	Method name	Class	Reason for eliminating
1	Rodelsperger	Supervised	CODE, CTSPEC, MULTI
2	Hist mark-to-gene expr corr	Unsupervised	CODE, CTSPEC, MULTI
3	Enh-to-prom activity corr	Unsupervised	CODE, CTSPEC
4	DNA access corr	Unsupervised	CODE, CTSPEC
5	DNA access-to-expr corr	Unsupervised	CODE, CTSPEC
6	SVM-MAP for methyl-to-expr corr	Supervised	CTSPEC
7	CAGE corr	Unsupervised	CODE, CTSPEC
8	PreSTIGE	Unsupervised	CODE
9	IM-PET	Supervised	**NA**
10	ELMER	Unsupervised	CODE
11	RIPPLE	Supervised	CODE, CONSIST
12	TargetFinder	Supervised	CONSIST
13	JEME	Supervised	**NA**
14	PEP	Supervised	CONSIST
15	FOCS	Supervised	CTSPEC
16	DeepTACT	Supervised	**NA**
17	ABC model	Unsupervised	**NA**
18	3DPredictor	Supervised	CONSIST
19	AVG rank between DNA access-to-expr and dist	Unsupervised	CTSPEC
20	EPIVAN	Supervised	CONSIST

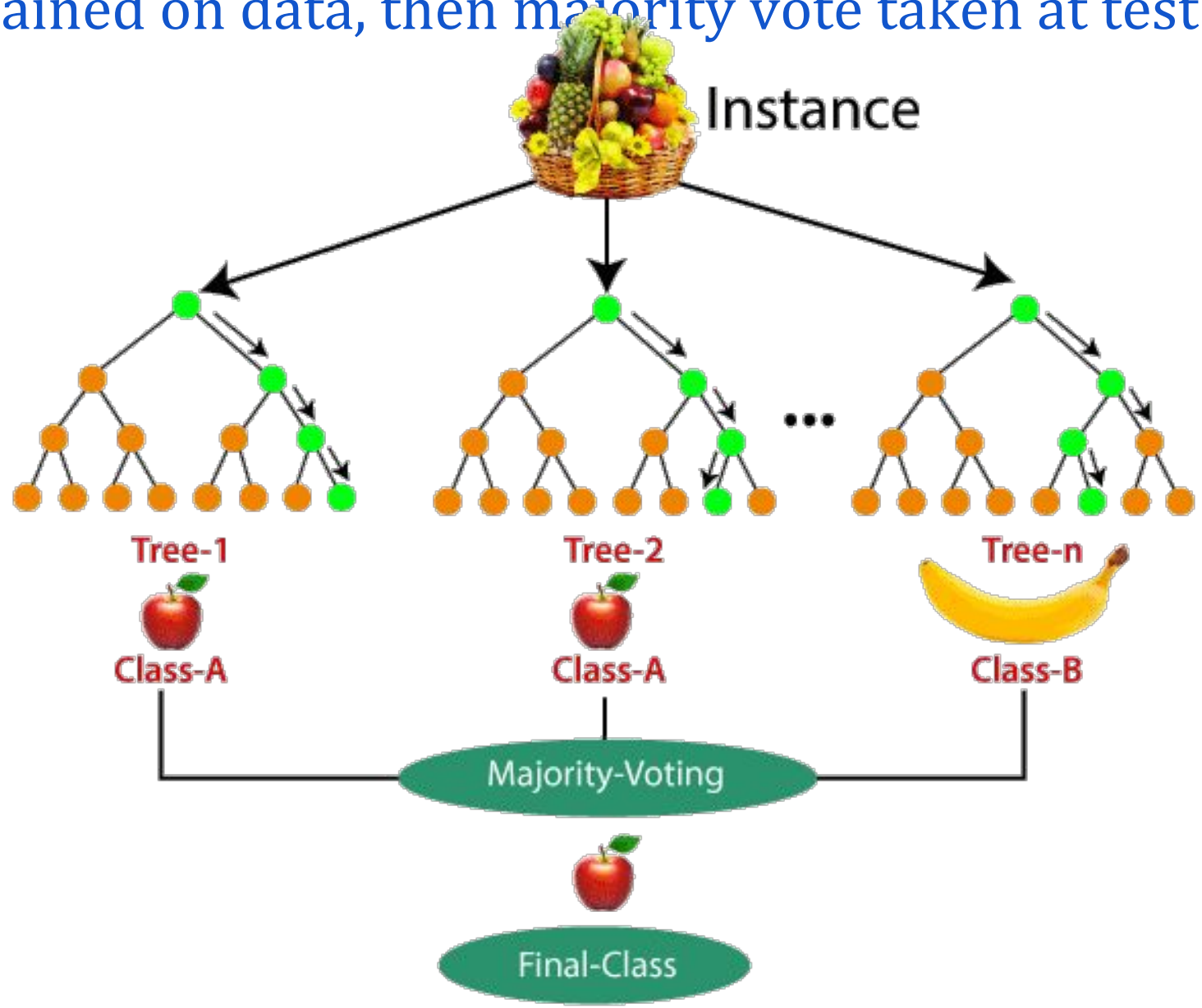
4 methods satisfying all 4 prerequisites (9, 13, 16, 17)

Method name	Class	Underlying statistical model	Publication date
IM-PET (integrated method for predicting enhancer targets)	Supervised	Random Forests	2014
JEME (Joint Effect of Multiple Enhancers)	Supervised	Multiple linear regression and Random Forests	2017
DeepTACT (Deep neural networks for chromatin conTACTs prediction)	Supervised	Bootstrapping deep learning method	2019
ABC (Activity by Contact) model	Unsupervised	Heuristic model	2019

Recall (sensitivity) and precision of predictive methods

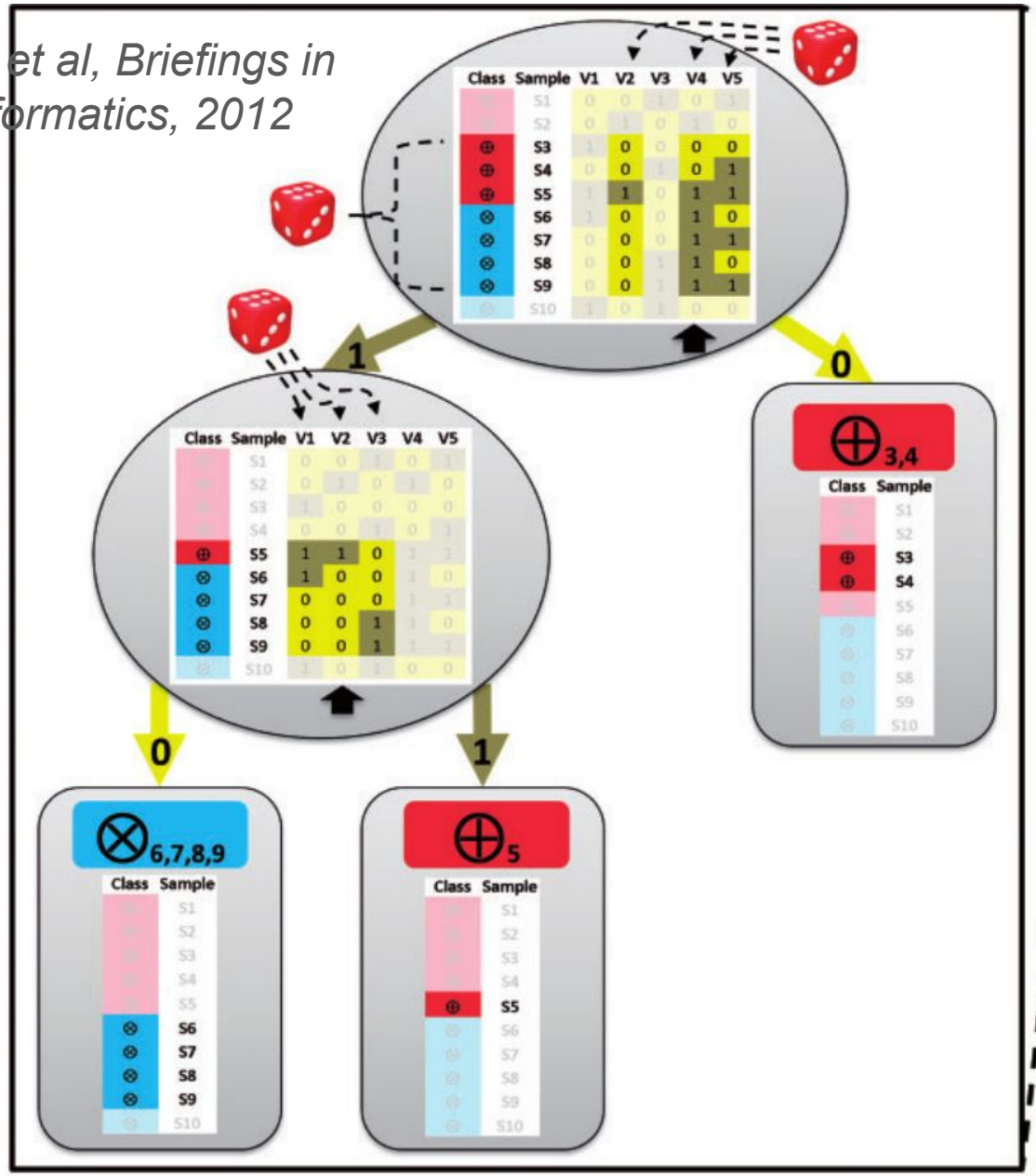
- Recall = Sensitivity = % of true connections (relationships) predicted by the method
- Precision = % of predicted connections that are true
- Always find a compromise between the two
- For a given predictive method that provides a score associated to each prediction, make the score vary to obtain several values of (recall, precision)
  - Precision recall curve
  - Method with greatest area under the curve?

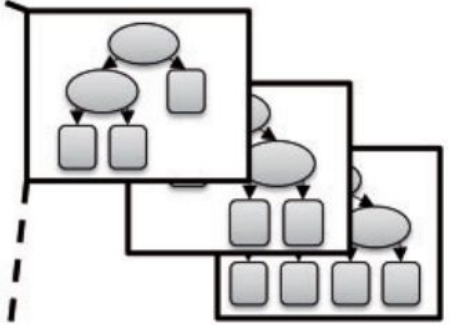
Random Forest (RF) classifier: several decision trees trained on data, then majority vote taken at test step



Training of an individual tree from a RF model

Touw et al, Briefings in Bioinformatics, 2012

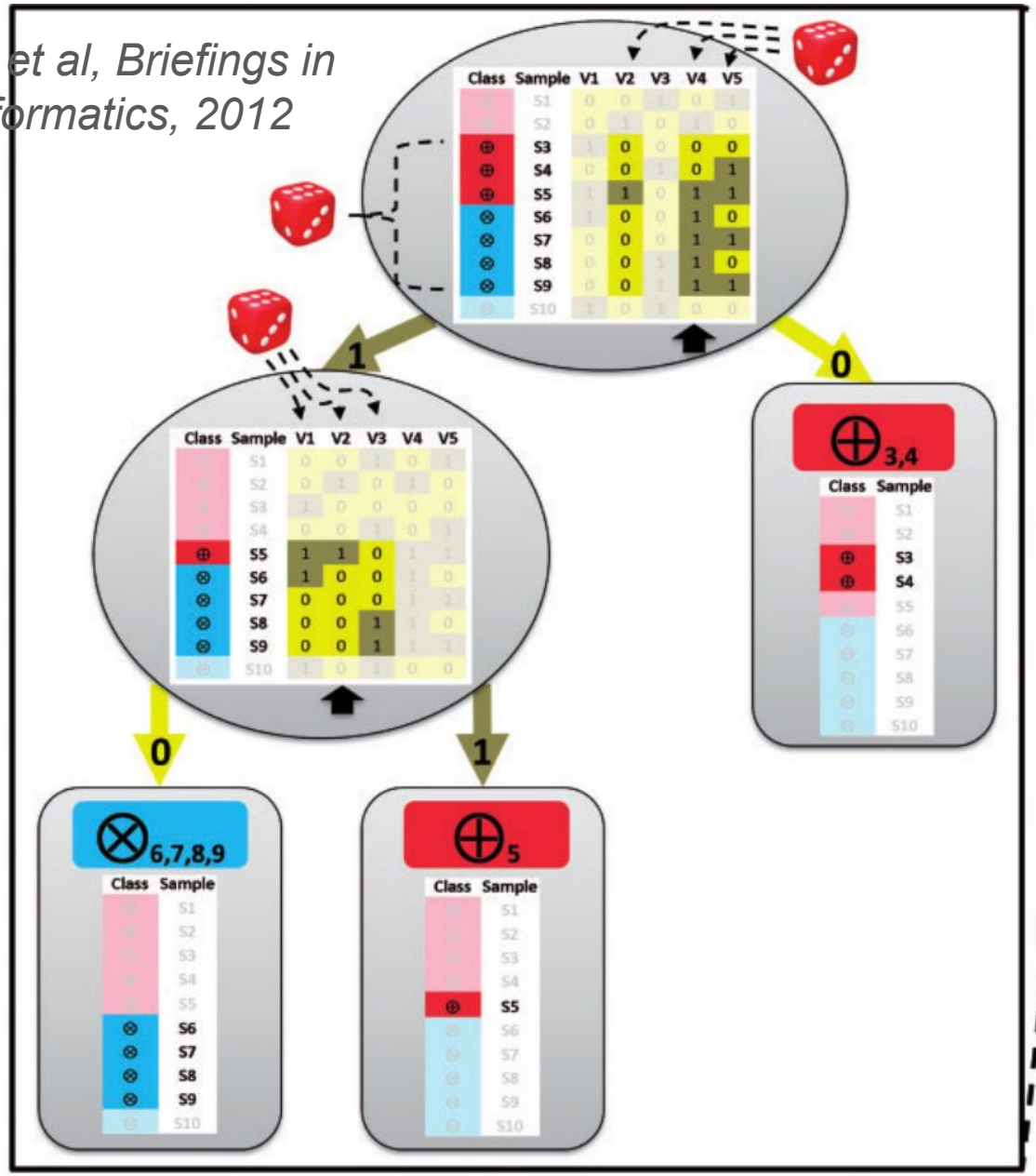


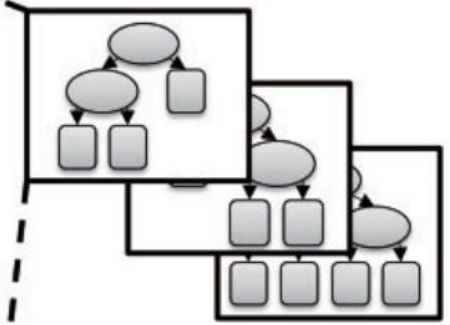


Training data = data matrix in the ellipses (samples S1 to S10 are individuals, that belong to 2 classes, encircled cross for healthy & encircled plus sign for ill, with measurements for variables V1 to V5)

Training of an individual tree from a RF model

Touw et al, Briefings in Bioinformatics, 2012

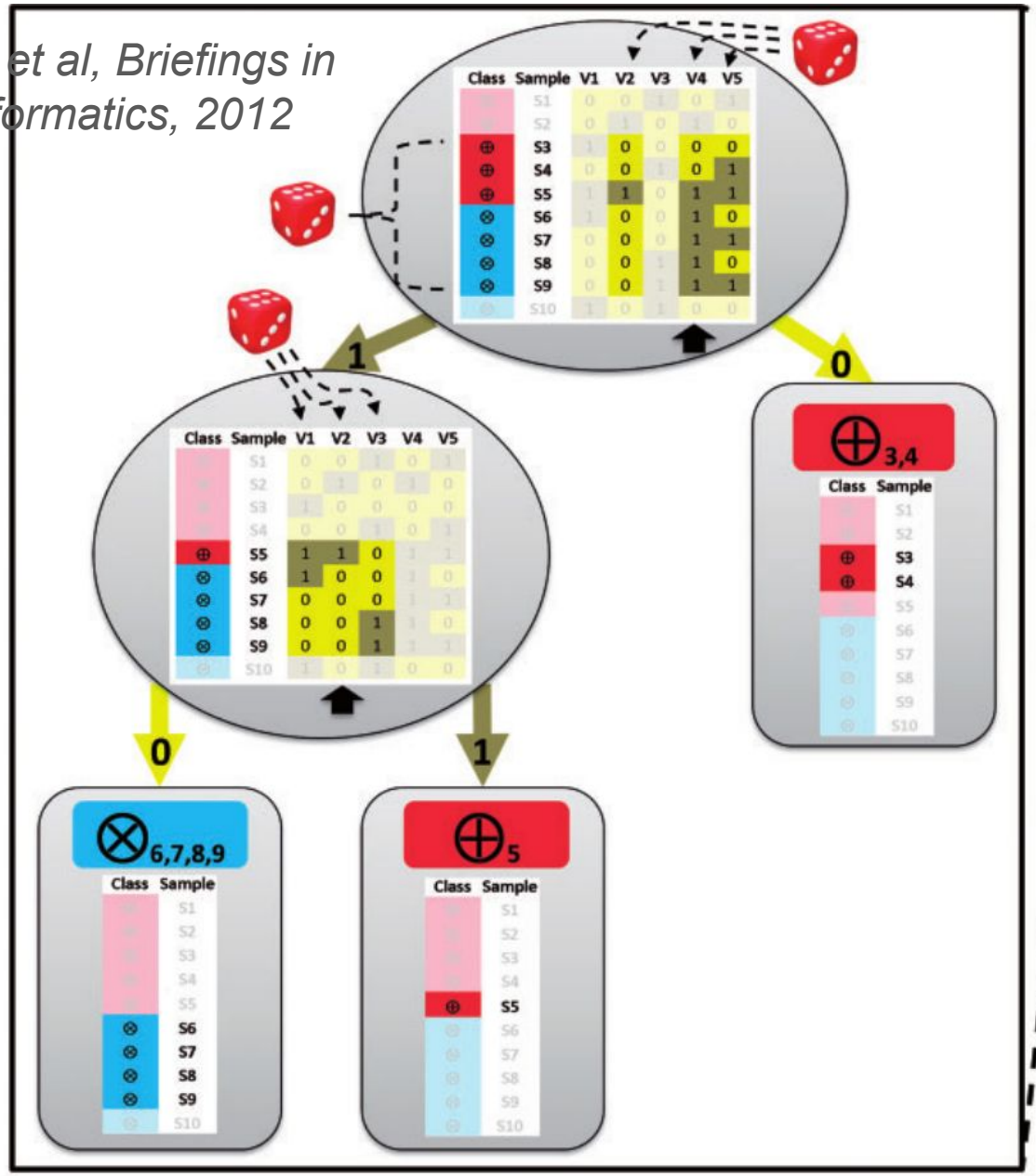


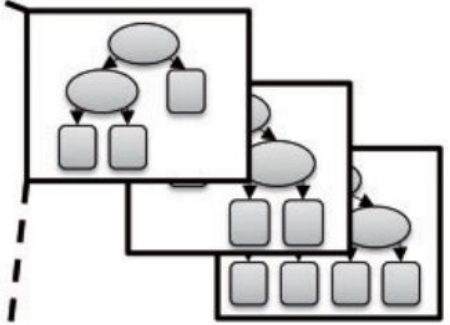


A bootstrap set is created by sampling samples from the data at random and with replacement until it contains as many samples as there are in the data set

Training of an individual tree from a RF model

Touw et al, Briefings in Bioinformatics, 2012

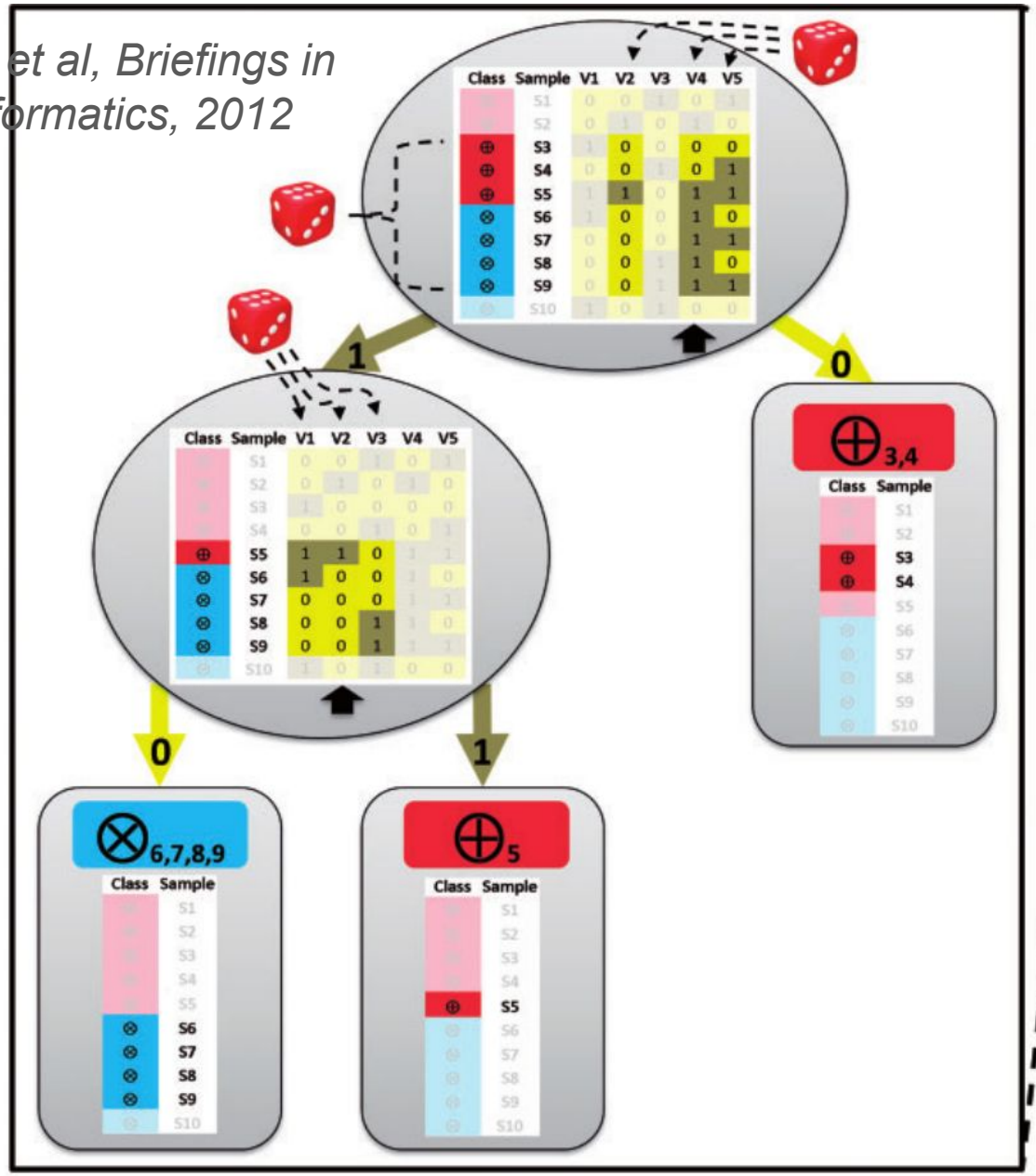


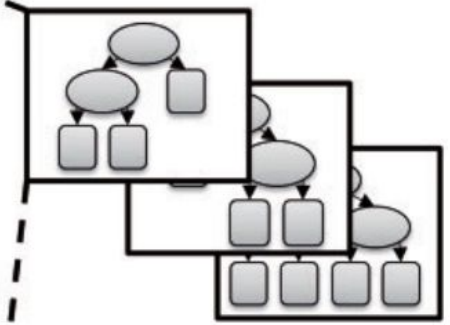


For every node (ellipse), a few variables are randomly selected and evaluated for their ability to split the data. The variable with the largest decrease in impurity is chosen to define the splitting rule

Training of an individual tree from a RF model

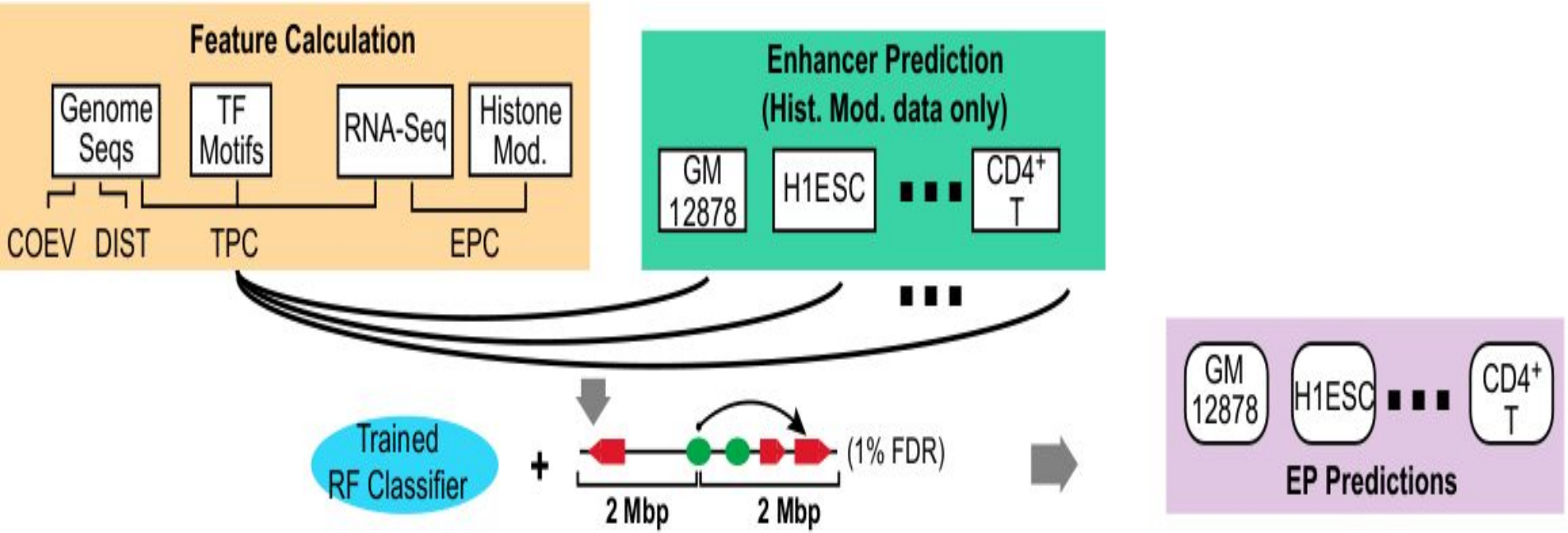
Touw et al, Briefings in Bioinformatics, 2012





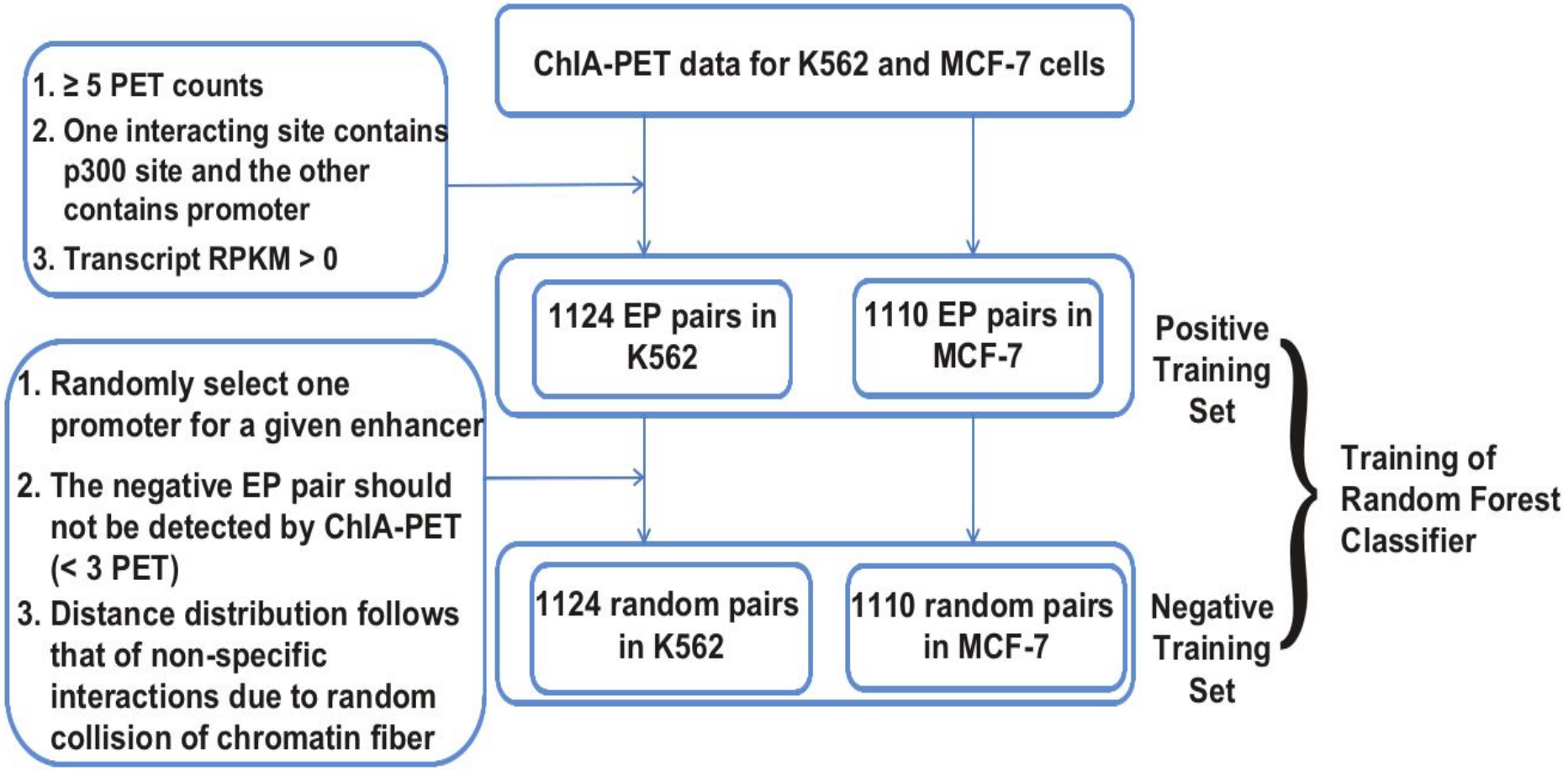
This process is repeated until the nodes are pure (so called leaves; indicated with round-edged boxes): they contain samples of the same class (encircled cross or plus signs)

IM-PET
(Integrated Method for Predicting Enhancer Targets)

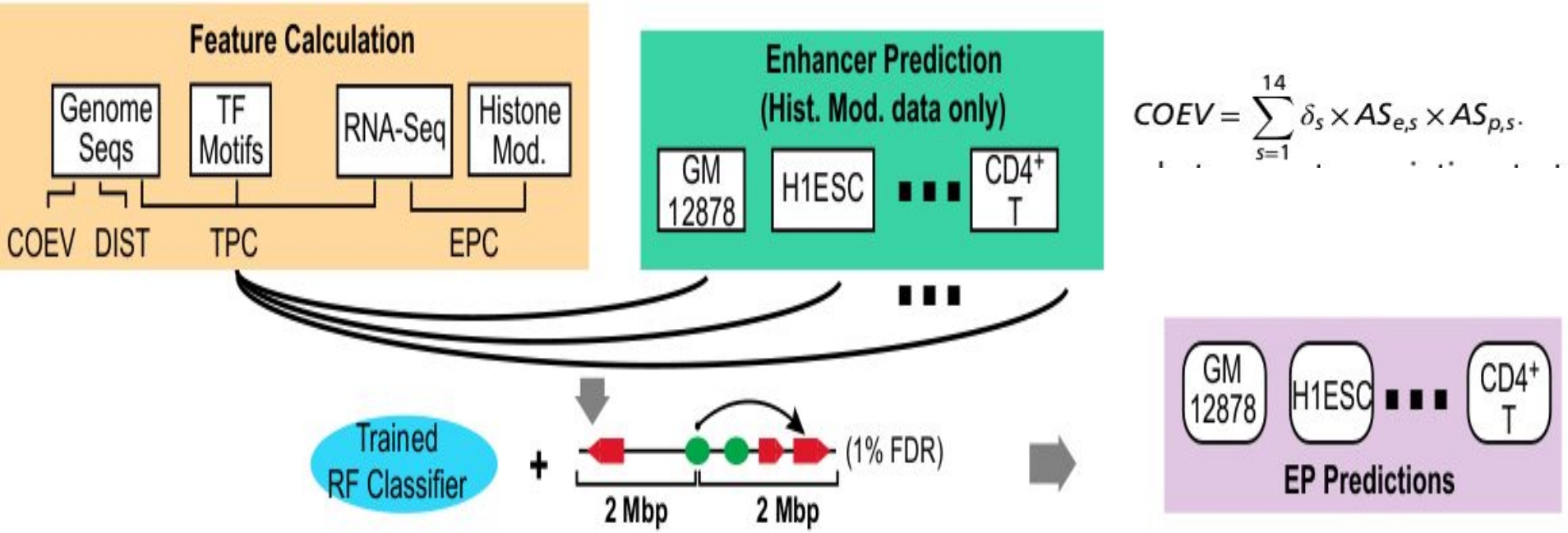


- RF training done on K562 and MCF7 cell lines for which polII ChIA-PET data is available, in combination with 3 histone marks and p300 for enhancers and RNAseq for promoters
- Negative set made using chromatin fiber equation (k reflects efficiency of cross-linking reaction) $\boldsymbol{f(s) = k \times s^{-3/2} \times e^{-1400/s^2}}$

Selection of EP (Enhancer/Promoter) pairs and RF classifier training



IM-PET (Integrated Method for Predicting Enhancer Targets)

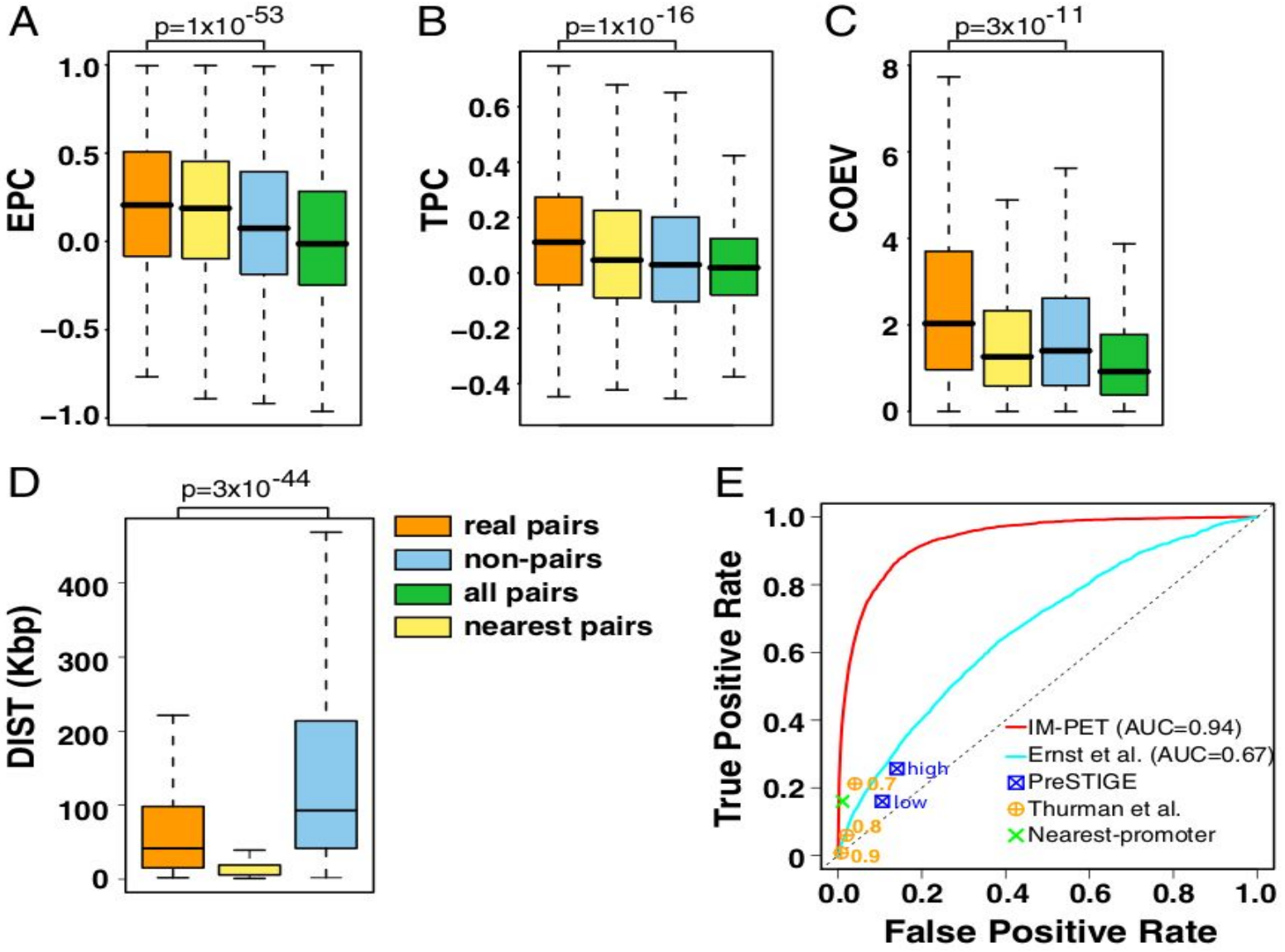


$$COEV = \sum_{s=1}^{14} \delta_s \times AS_{e,s} \times AS_{p,s}.$$

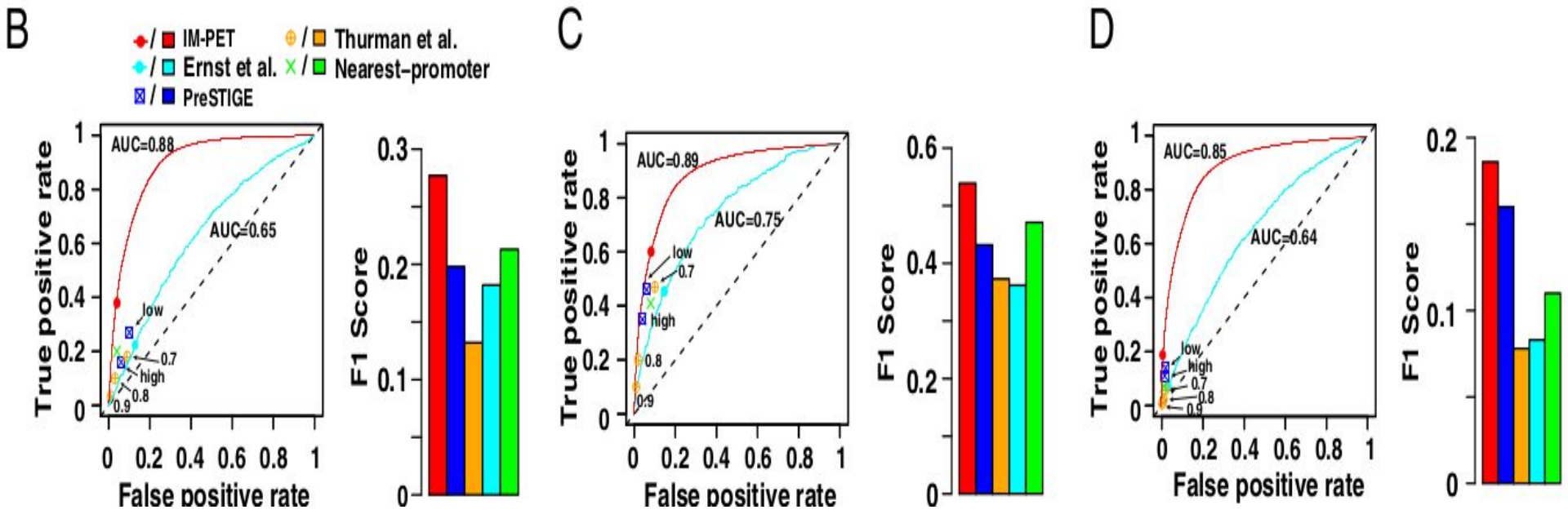
4 discriminative variables/features used:

- Enhancer and target promoter activity profile correlation (EPC)
- TF and target promoter expression correlation (TPC)
- Coevolution of enhancer and target promoter (COEV)
- Distance constraint between enhancer and target promoter (DIS)

Discriminative features and performance evaluation by cross-validation



ROC curves & F1 score using additional polII ChIA-PET (B), deep HiC EP pairs (C) and eQTL-gene pairs (D)

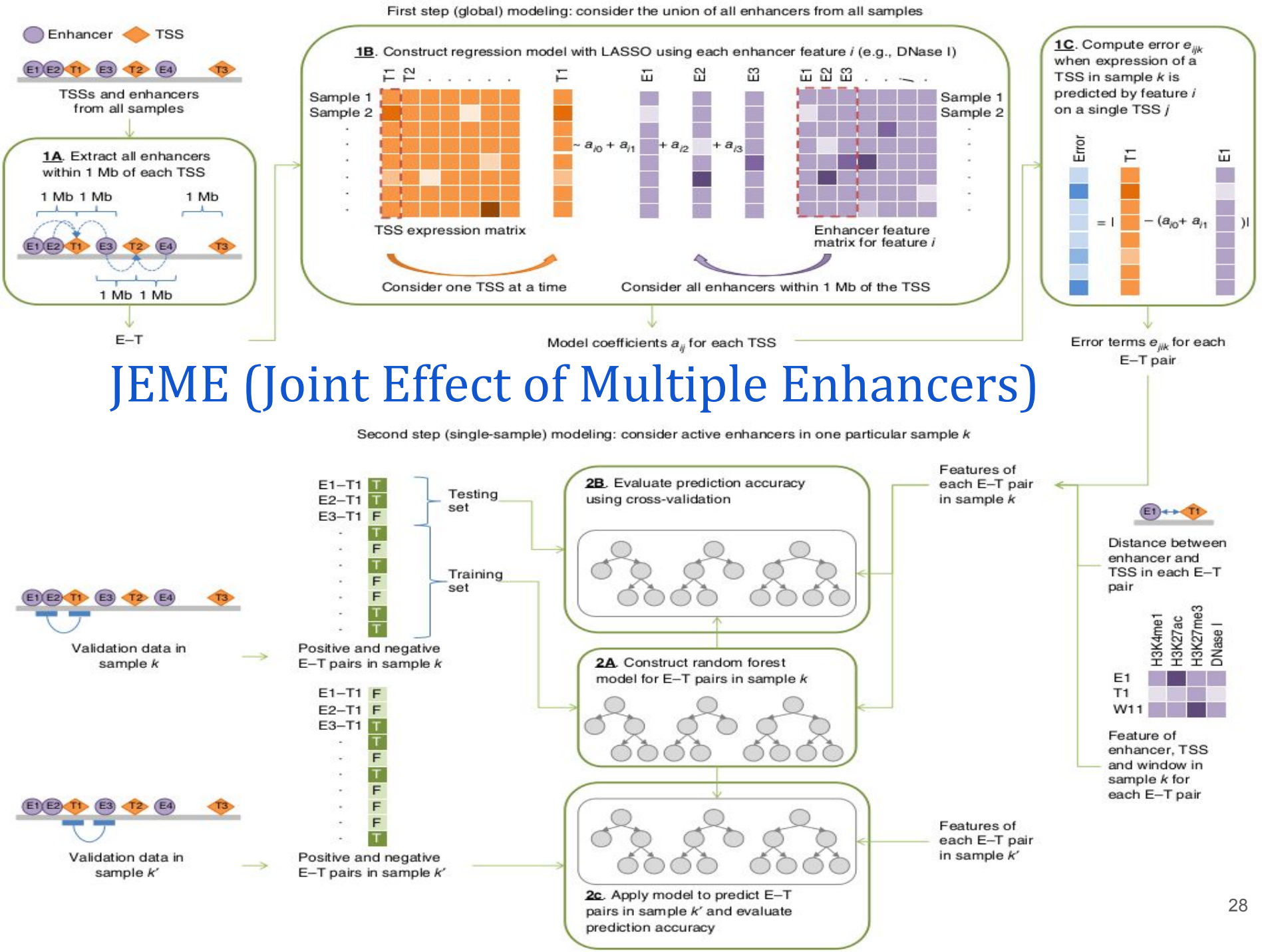


Predictions in 12 cell lines are compared to:

(B) polII ChiA-PET from 3 cell lines (K562, MCF7, and CD4 + T cells)

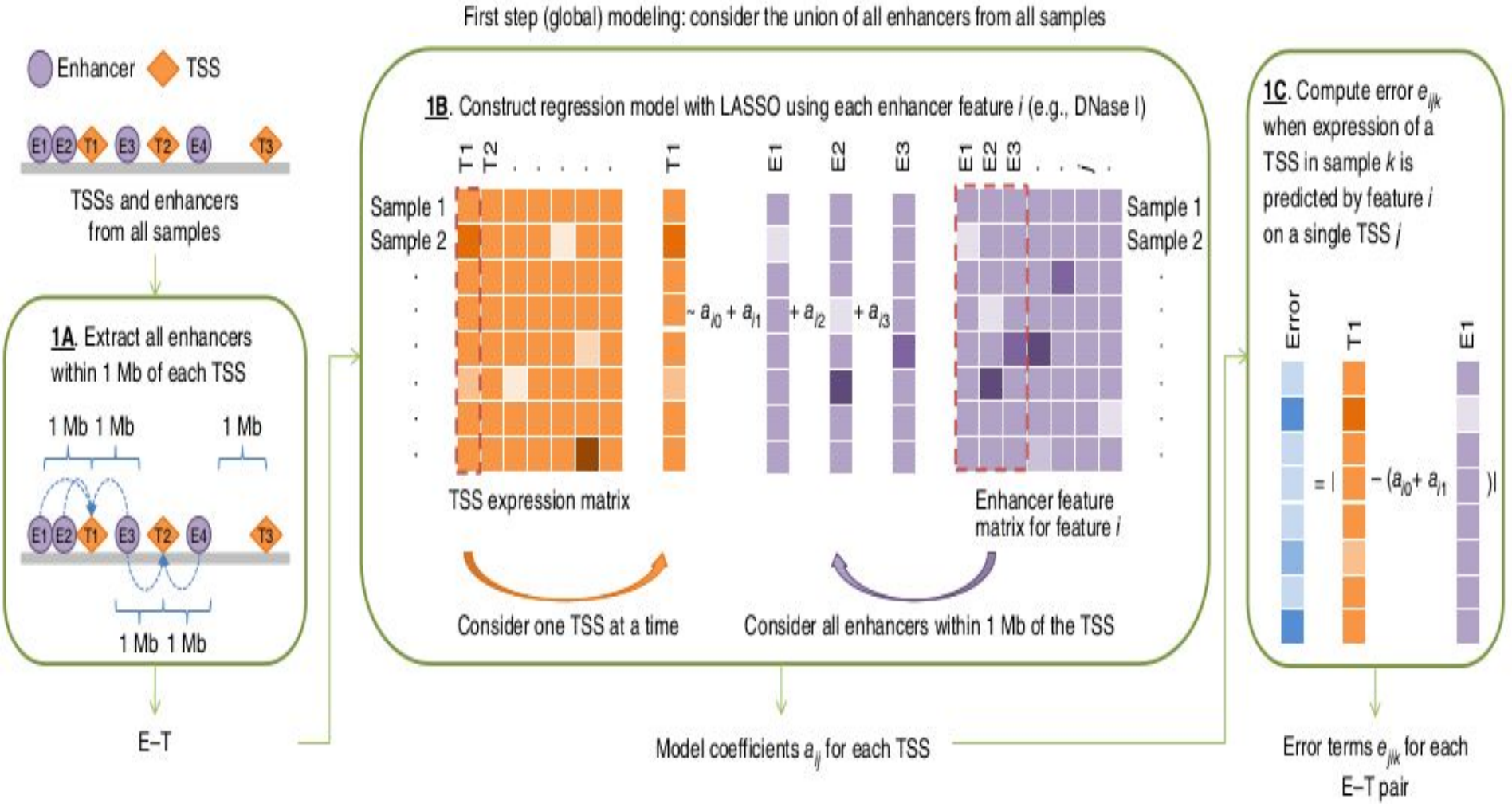
(C) deep HiC from IMR90 cell line

(D) eQTL from GM12878 and HepG2 cell lines

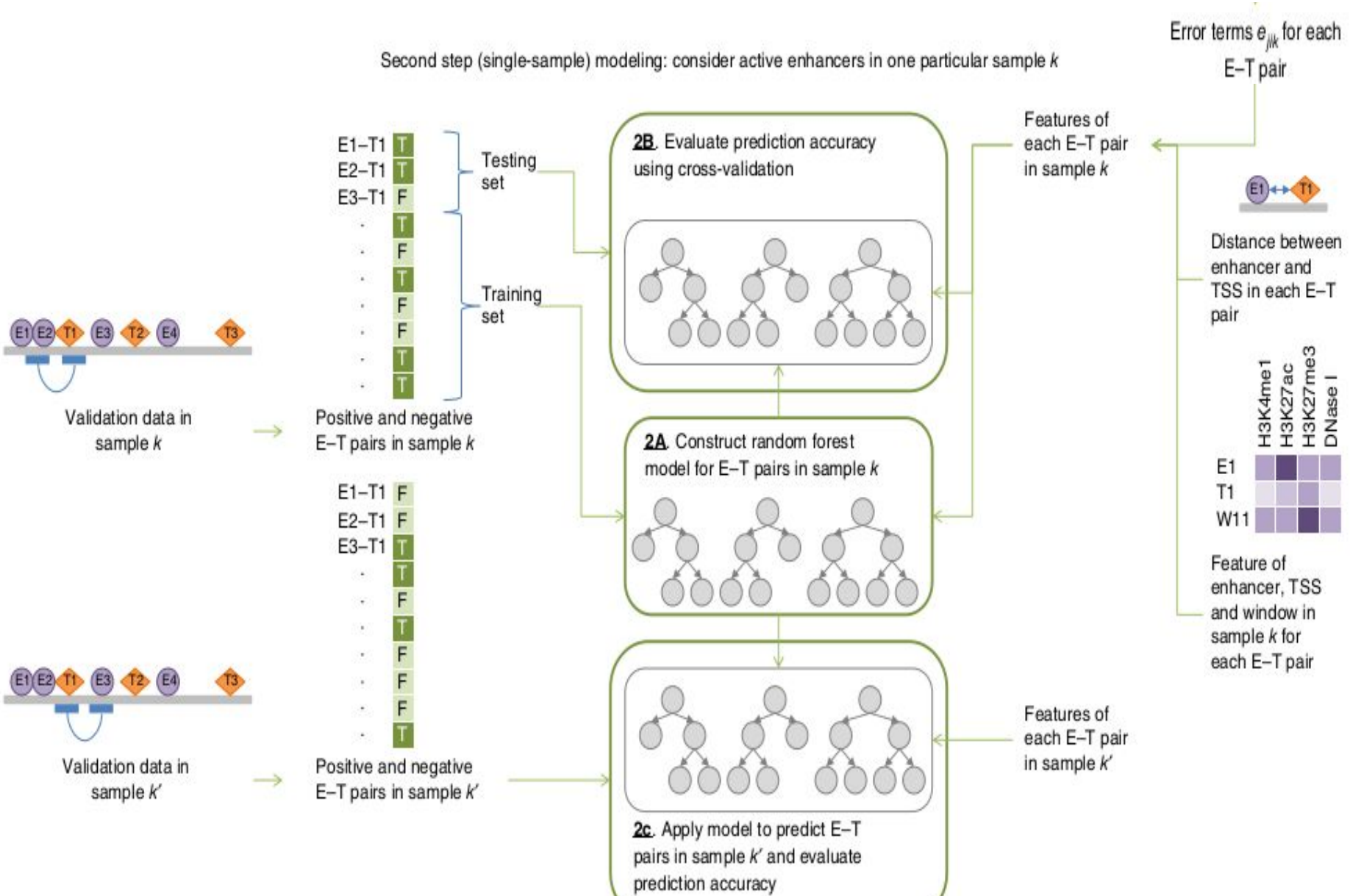


JEME (Joint Effect of Multiple Enhancers)

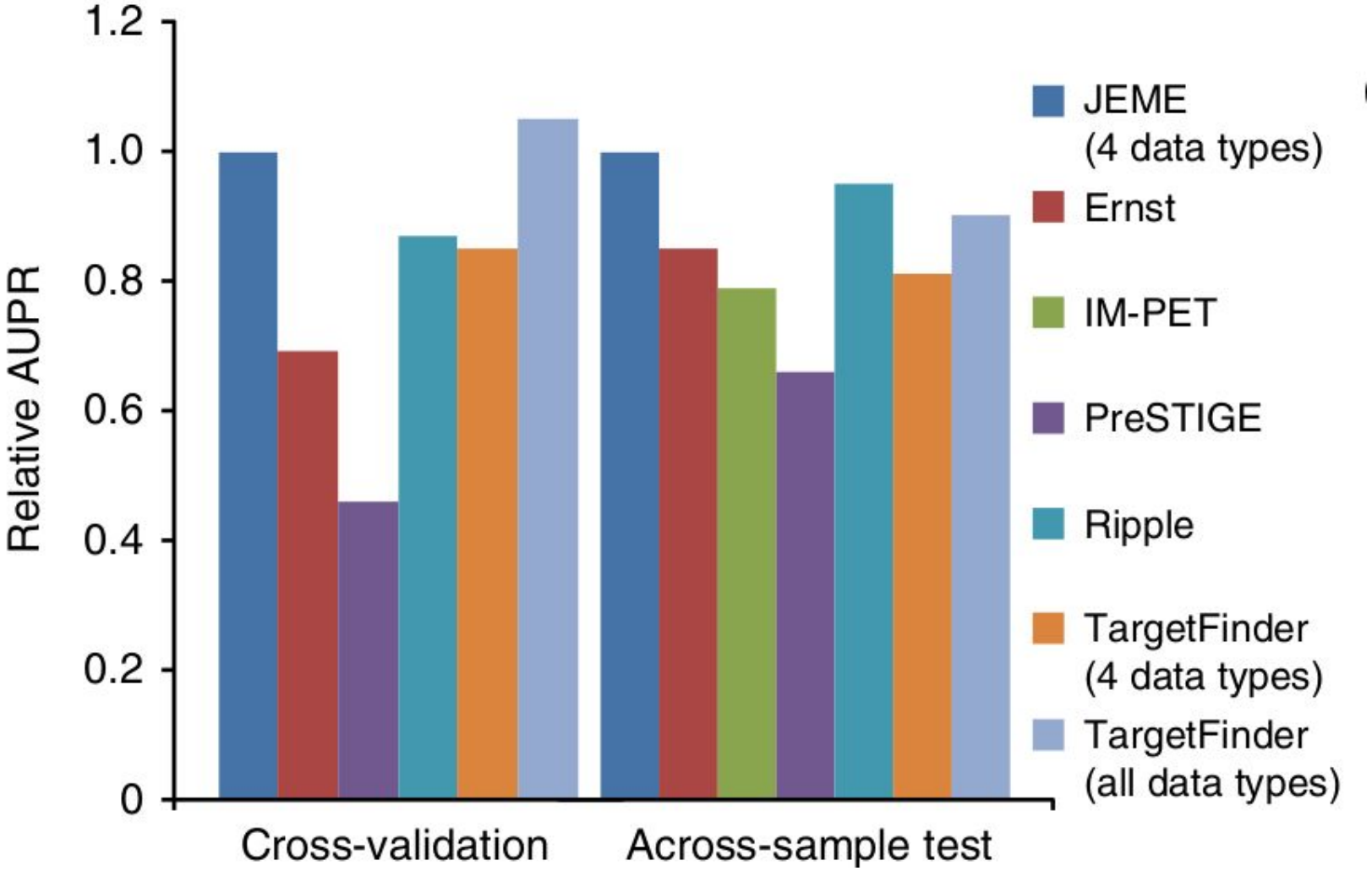
1st step: prediction of all possible EP pairs in all samples using multiple linear regression



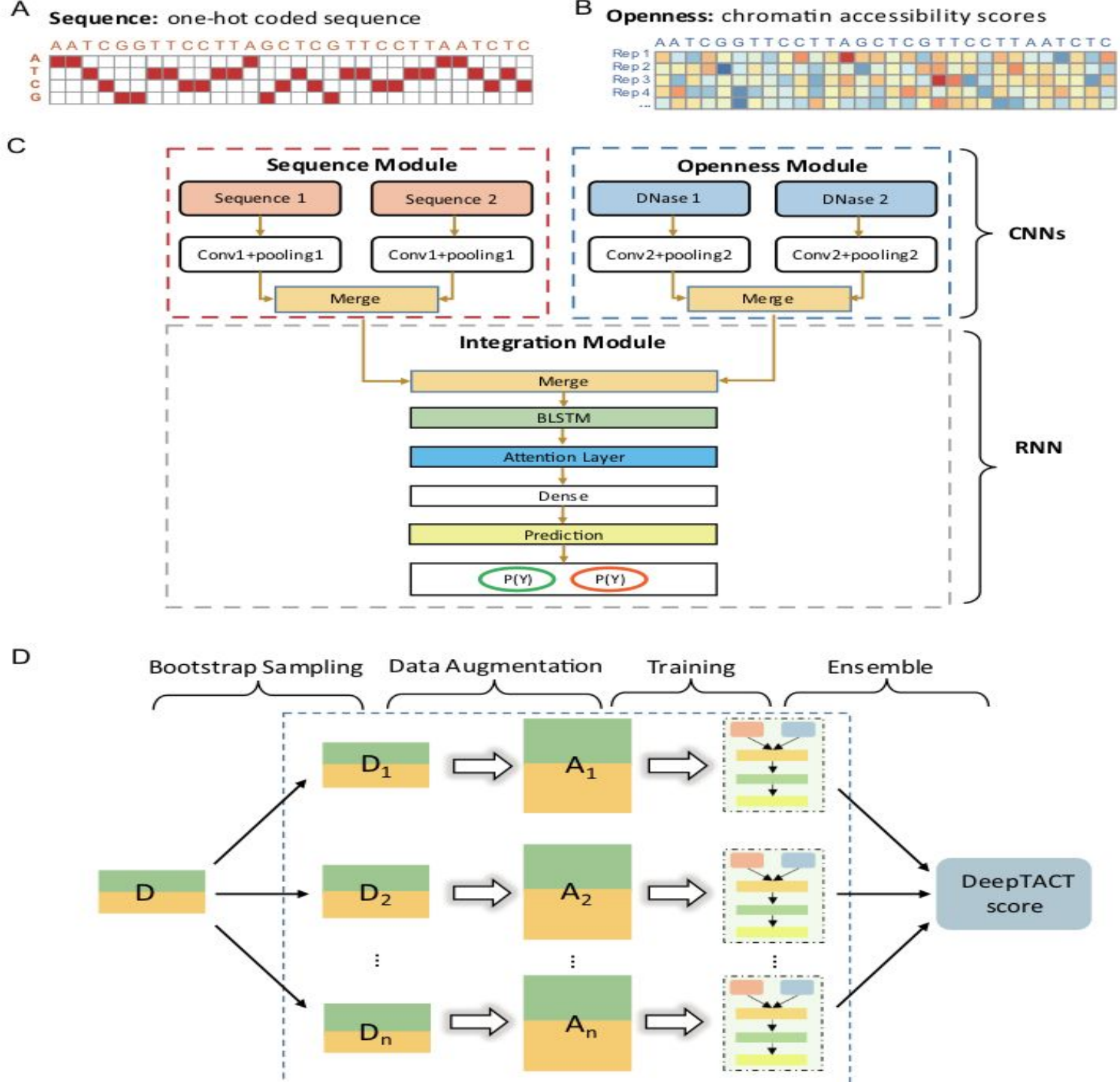
2nd step: prediction of EP pairs in a particular cell type using RF trained on polII ChIA-PET data



Performance of the E-T prediction methods
(training in GM12878, validation in GM12878 (CV) or K562

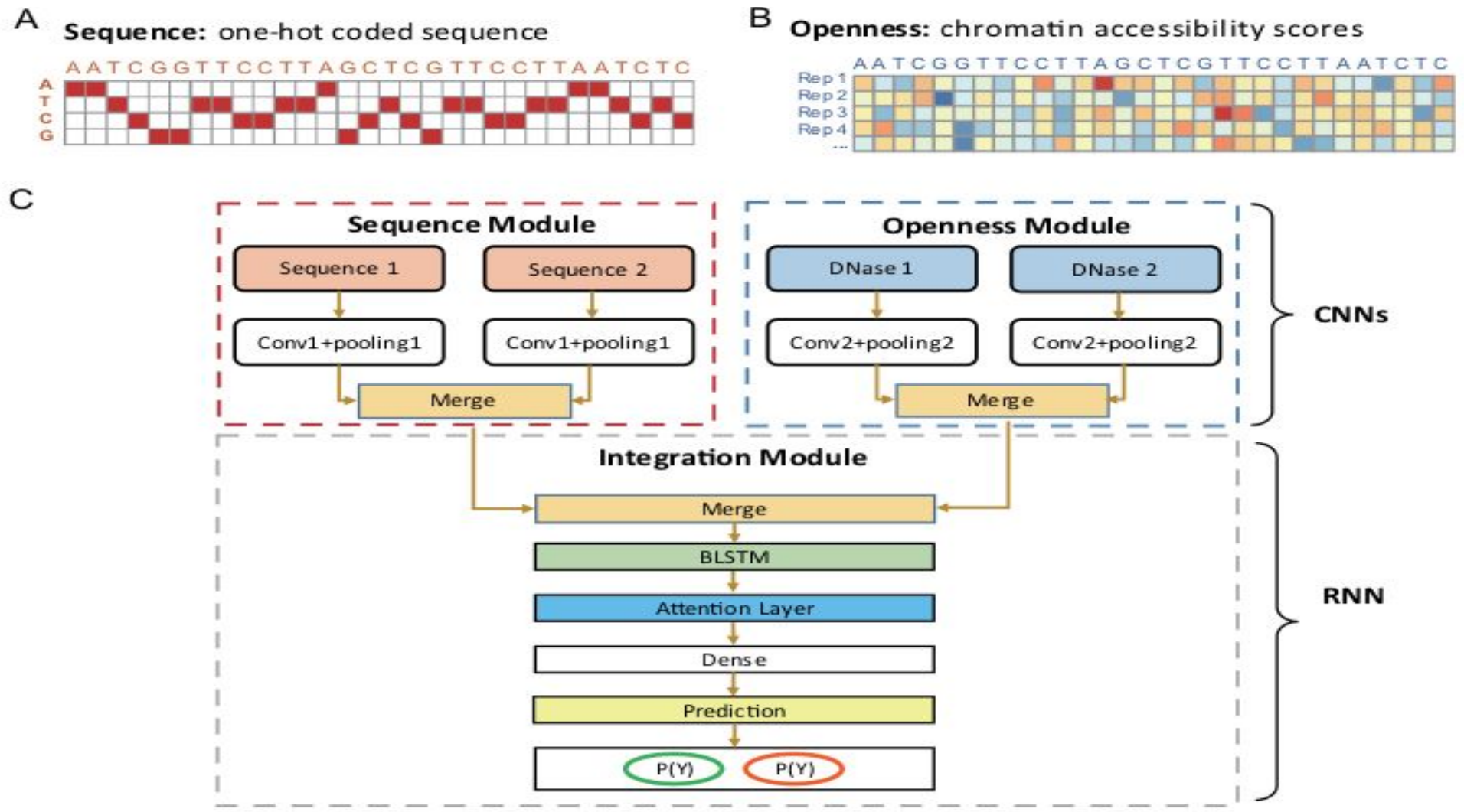


Relative AUPR = AUPR relative to a naive model



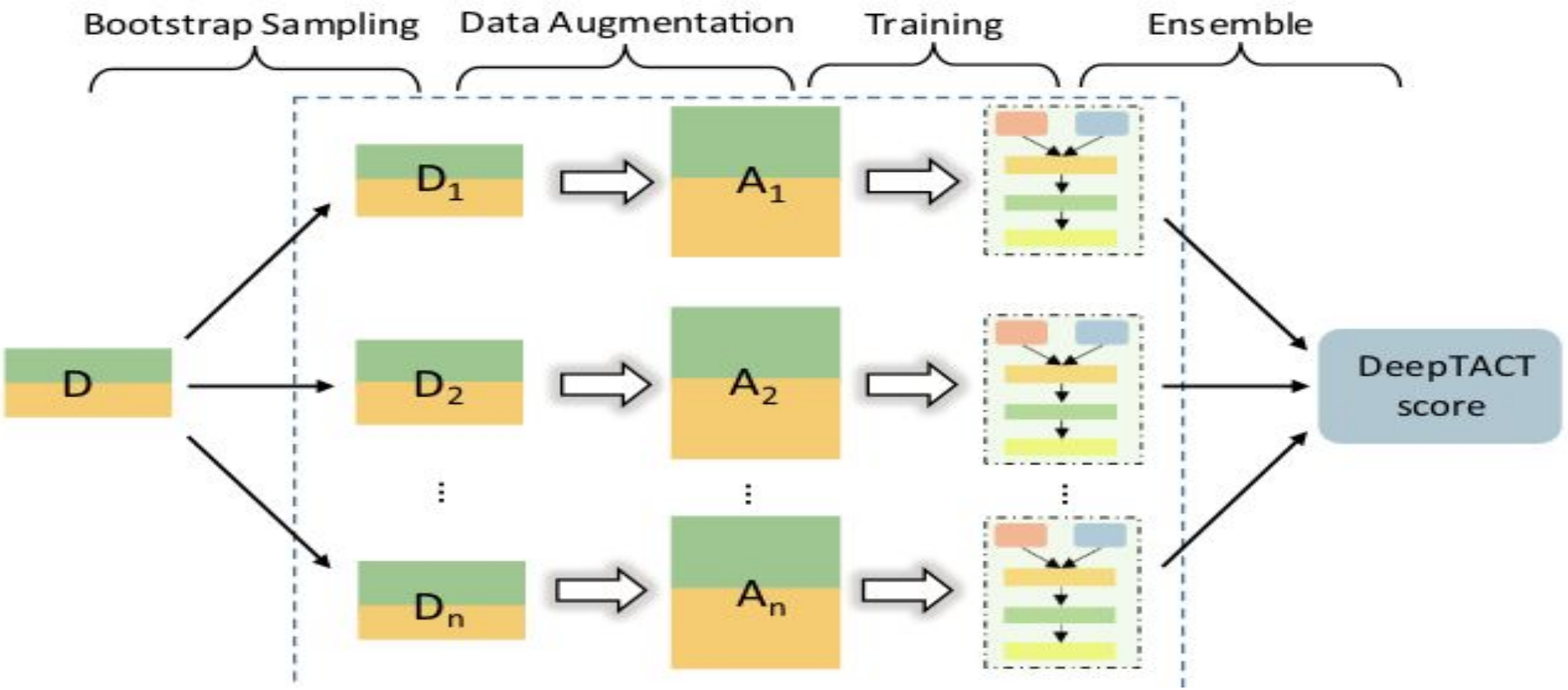
DeepTACT (Deep neural networks for chromatin conTACT prediction)

Convolutional Neural Networks for learning sequence and openness on each side, recurrent neural network for integration

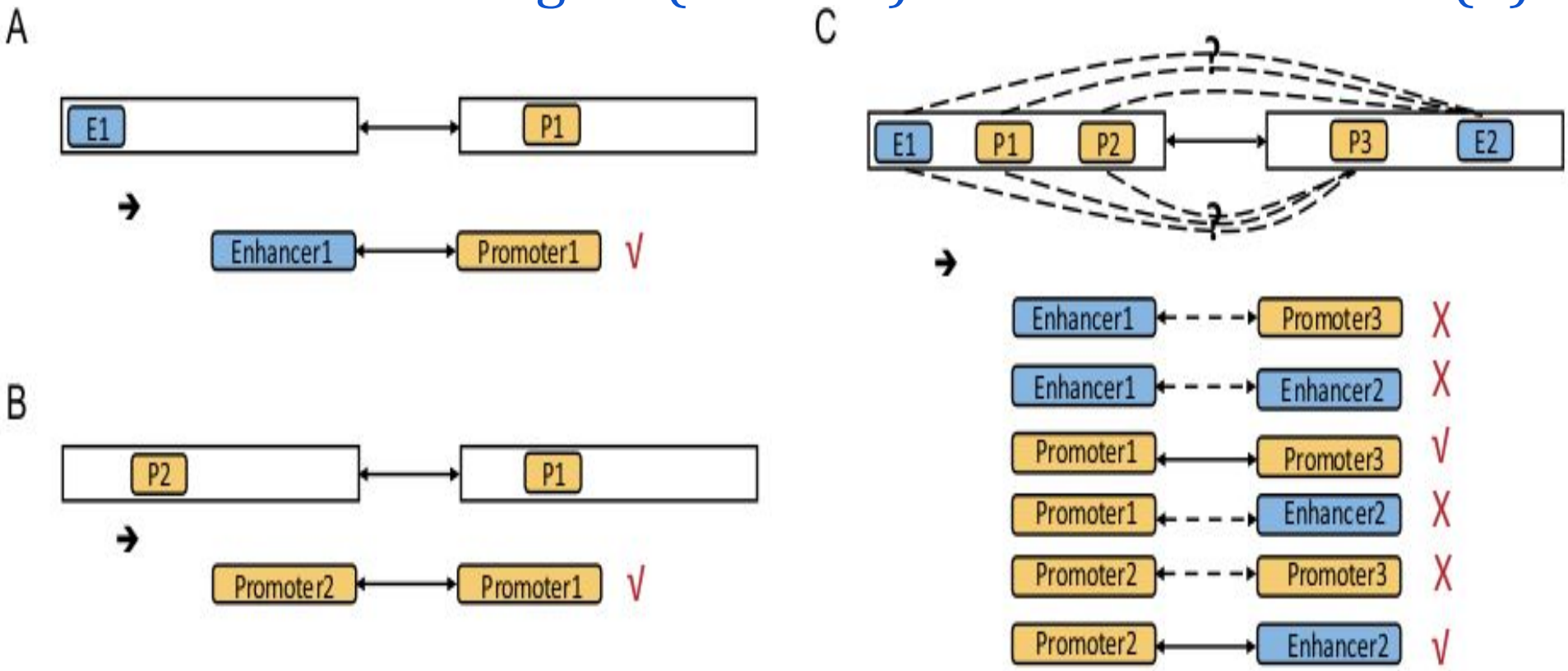


→ Trained on promoter capture HiC data

Ensembl strategy based on bootstrapping technique to overcome the instability of the deep neural network



Pairs of interacting regions containing a single regulatory element in each region (A and B) or several elements (C)

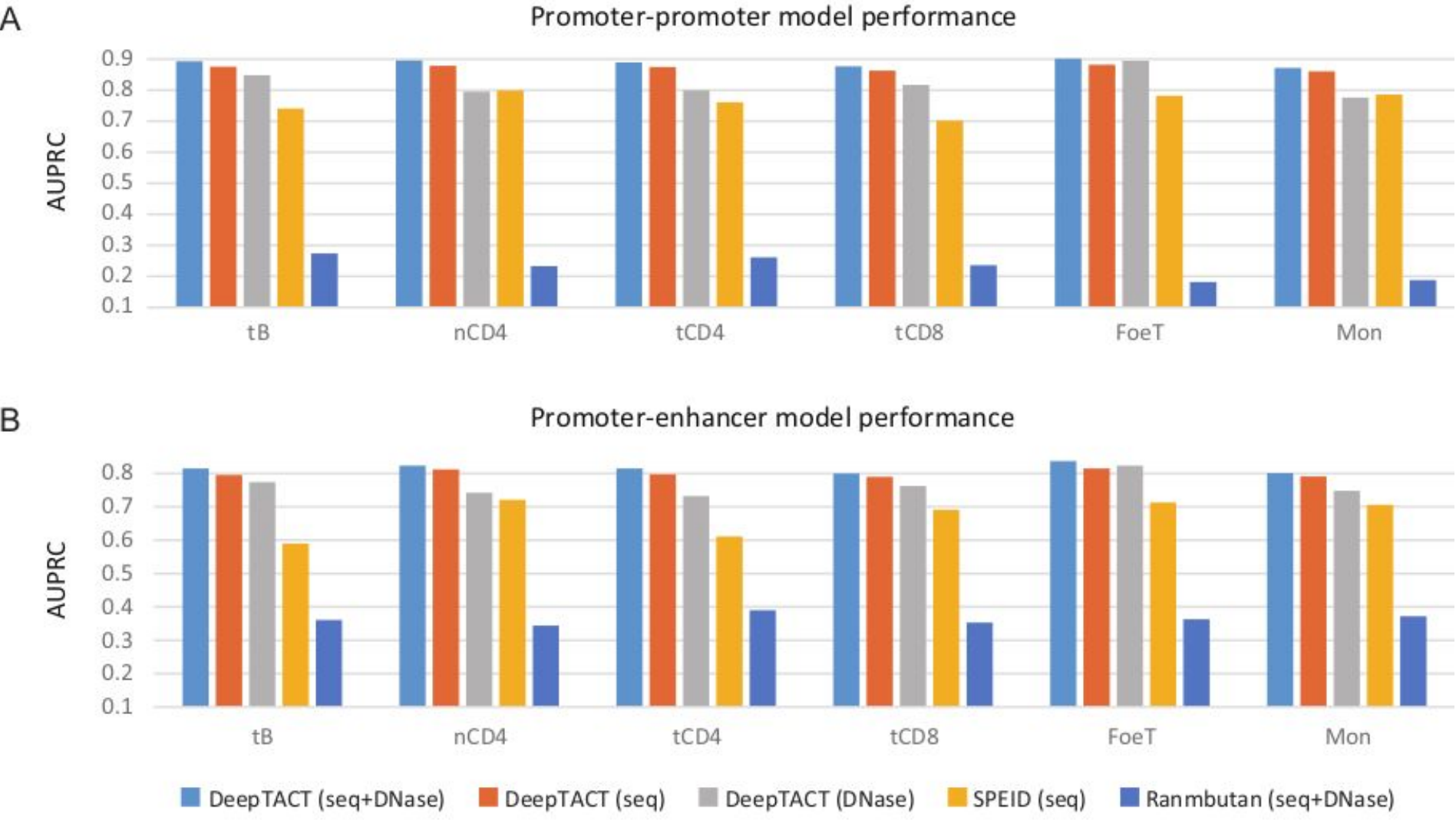


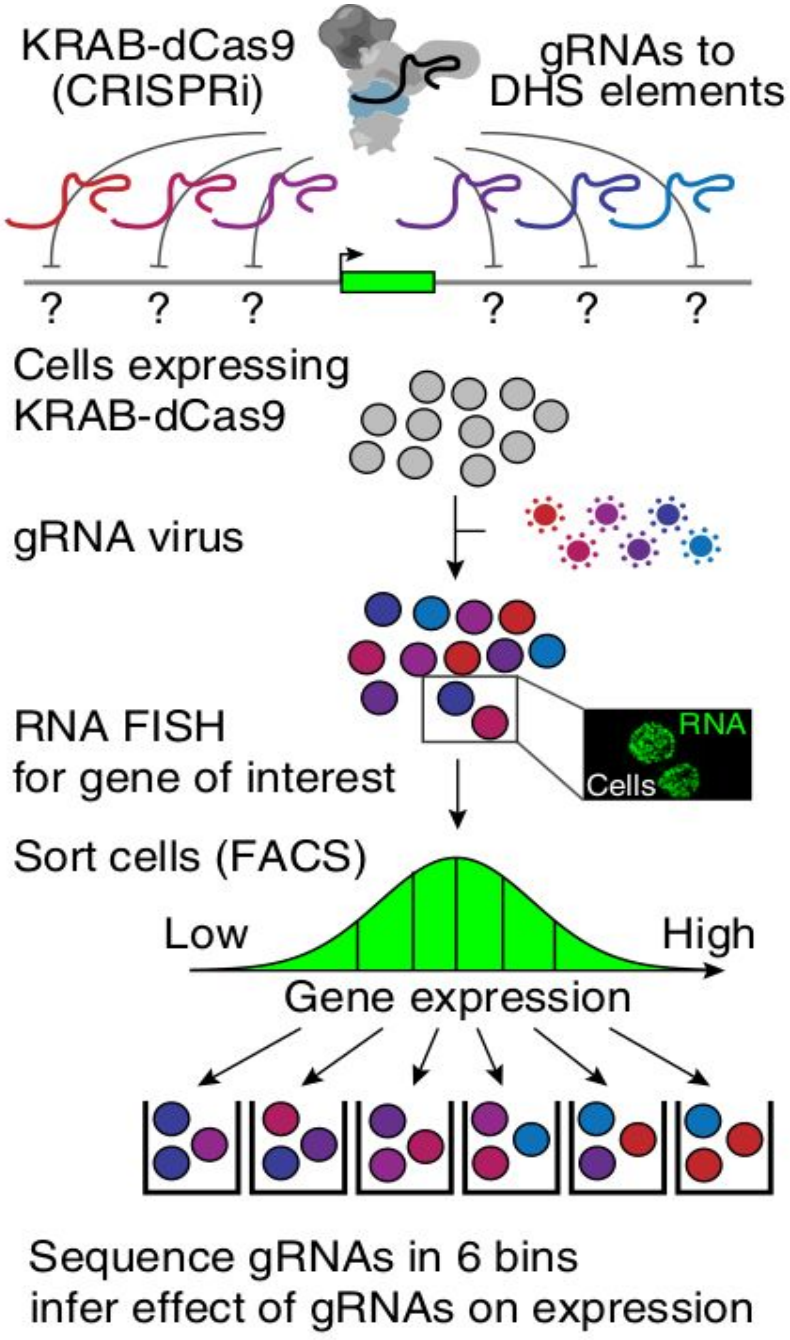
- Training is done on the the single regulatory region types of connections but prediction is done on all types of connections
- Enhancers = 65, 432 FANTOM5 permissive enhancers (all cell types) extended by 2kb on each side (from their middle)
- Promoters are 1kb regions surrounding ensembl TSS

DeepTACT characteristics

- The input for the predictive model is the sequences of two regulatory elements represented with a one-hot encoding strategy and their chromatin accessibility scores derived from DNase-seq experiments of a given cell type. Based on this input, the model will compute the predictive score of whether the two regulatory elements have 3D contact
- Separately predicts promoter-promoter and promoter-enhancer relationships
- Sees itself as a way to improve the resolution of HiC data like HiCplus, Epitensor and 3DEpiLoop, but claims to be more resolutive (1kb) and/or able to use fewer datasets as input

Performance evaluation of DeepTACT on 6 cell types





The CRISPRi-FlowFISH technique to identify open regions with an effect on cis genes (Fulco et al, Nature Genetics, 2019)

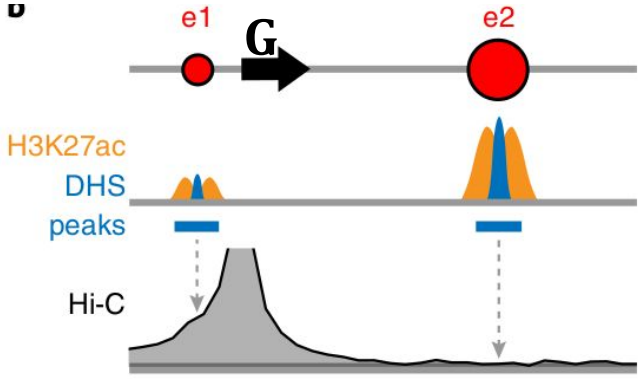
A large enhancer perturbation dataset

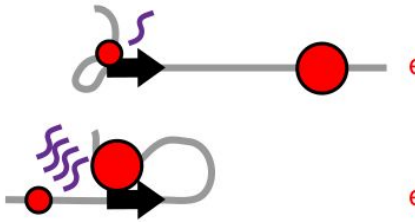
- Use CRISPRi-FlowFISH in K562 (erythroleukemia cells)
- 4,662 candidate regulatory element (CRE)-gene pairs tested
- Screens done for 30 genes in 5 gx regions (1.1-4Mb)
- Tested all DHS elements in K562 at 450 kb of the tested genes (108-277 elements per gene, 884 unique elements)
- Selected genes are either tissue-specific (GATA1) or ubiquitous (RAB7A) and were selected to have FlowFISH probe sets that are specific and with enough stat power
- Elements over the gene are excluded because recruitment of KRAB-dCas9 in a gene body interfere with transcription

Global summary statistics results

- Individual enhancers regulate up to 5 (tested) genes
- Individual genes regulated by up to 14 distal (tested) elements
- Some enhancers skip over proximal genes to regulate more distal genes
- Out of 3,863 distal element-gene (DEG) pairs tested, 141 have significant effect on gene expression at FDR < 0.05
- Decrease in expression in 77% of cases (109/141) and increase in 23% of cases, with absolute effect sizes 3-93% (median 22%)
- To assess several predictors, use 109 experimentally validated DE-G pairs as true positive and 3,754 non regulatory connections as true negative (precision-recall plots)

The Activity-By-Contact (ABC) model

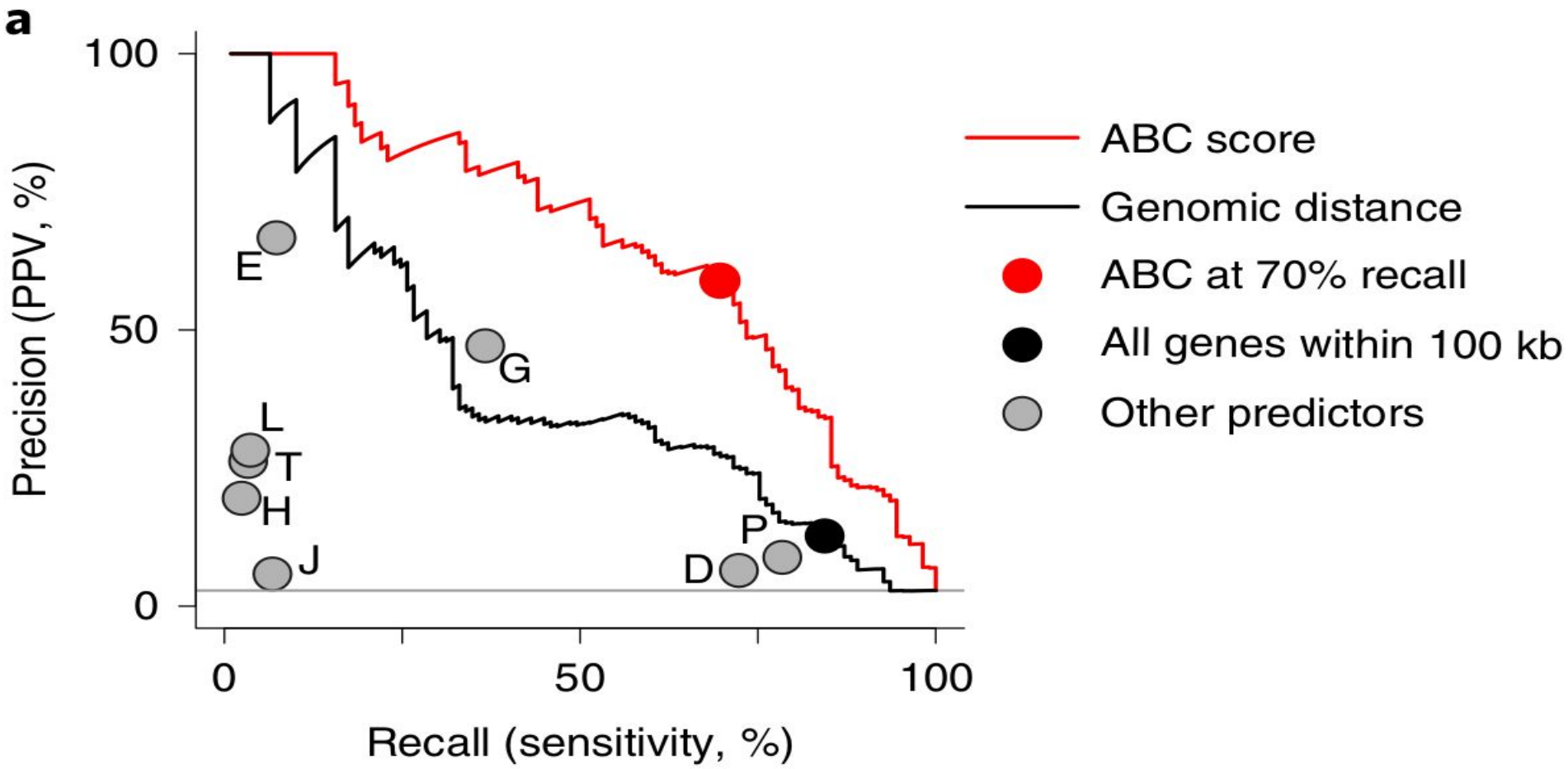




	Activity × contact = A × C					ABC score
e1:	1	×	12	=	12	0.75
e2:	4	×	1	=	4	0.25

$$\text{ABC score}_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{\text{all elements e within 5 Mb of } G} A_e \times C_{e,G}}$$

- A_E = Activity = geometric mean of the read counts of the DHS and the H3K27ac ChIP-seq at enhancer E
- $C_{E,G}$ = Contact = KR-normalized HiC contact frequency between E and the promoter of gene G at 5 kb resolution



- G: element assigned to the TSS of the closest expressed gene
- E: assign each expressed gene to the closest DE
- D: element assigned to the promoters in the same HiC contact domain
- L: element assigned to the promoters at the opposite of HiC loops
- P: assign based on RNA polII ChIA-PET loops
- T: genes predicted by the algorithm TargetFinder (machine learning)
- J: genes predicted by the algorithm JEME (machine learning)

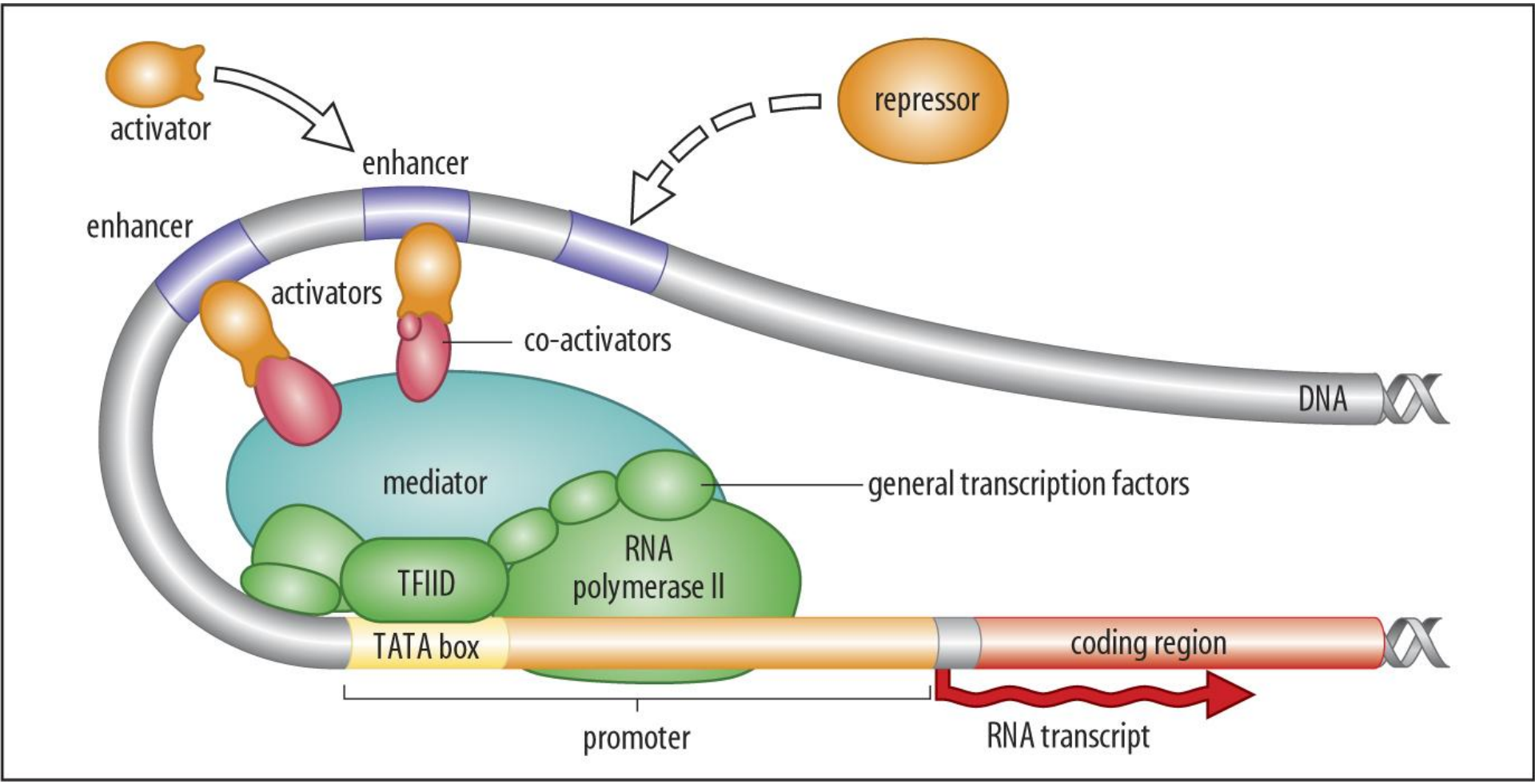
Summary

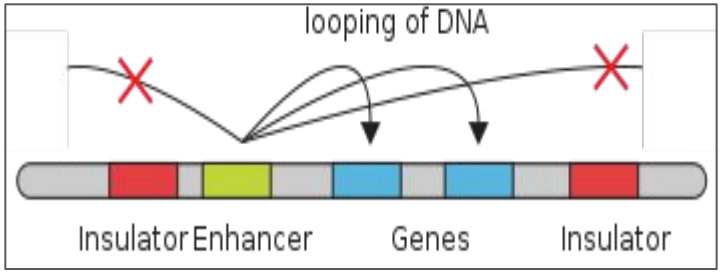
Method name	Underlying statistical model	Data used to make the positive set in a given cell type	Features used for testing in a cell type	How is/are negative set(s) made?	How is the method evaluated ?
IM-PET	RF	polII ChIA-PET + p300 + (enhancers predicted by CSI-ANN based on 3 histone marks) + RNA-seq for promoters	3 histone marks in 12 cell types + TFBS + evolution + distance	Random but based on chromatin fiber equation	5-fold cross validation + additional ChIA-PET + HiC + eQTL
JEME	Multiple linear regression and RF	polII ChIA-PET + chromHMM enhancer states	3 histone marks + DNAse-seq + RNA-seq	4 different ways	5-fold cross validation + across cell type validation
DeepTACT	Deep neural network	Promoter capture HiC + DNAse-seq + FANTOM5 permissive set of enhancers	DNAse-seq	Random with same distance distribution as positive set	Cross validation + ChiA-PET + eQTL
ABC model	Heuristic rules based on existing knowledge of the field	NA	Chromatin accessibility, H3K27ac ChiP-seq, (HiC)	NA	Compare to genetic screening data (30 genes, 109 positives, 1 cell line)

What's next ?

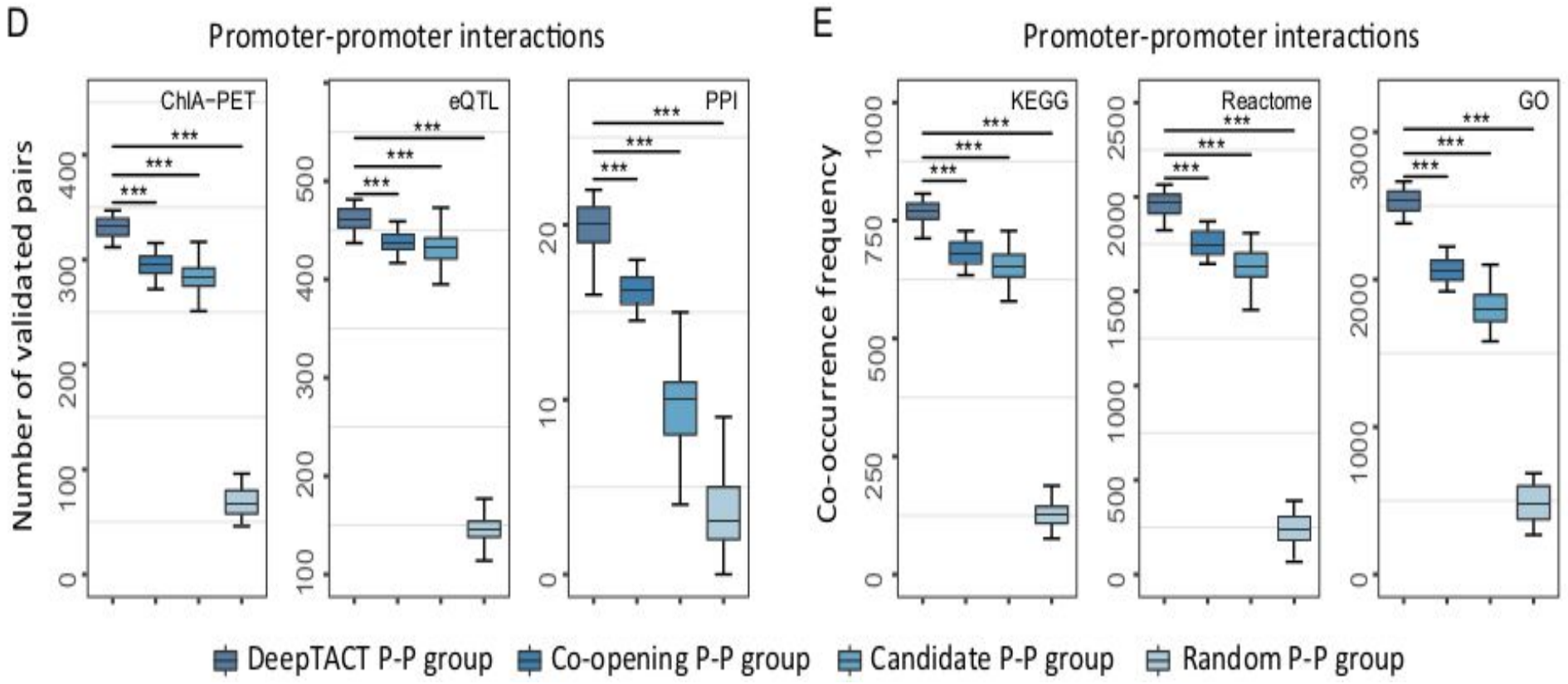
- Try the 4 selected methods on real data (small and big)
- Plan the evaluation
  - Choose the reference sets
  - Get the necessary input data for each method
  - Determine the evaluation metrics
- Evaluate the 4 methods on each reference set
  - Or evaluate the underlying statistical models ?
- Determine the best approach
  - Devise one that uses as few input data types as possible

Additional slides


activator
repressor
enhancer
enhancer
activators
co-activators
DNA
mediator
general transcription factors
RNA
polymerase II
TFIID
TATA box
coding region
promoter
RNA transcript

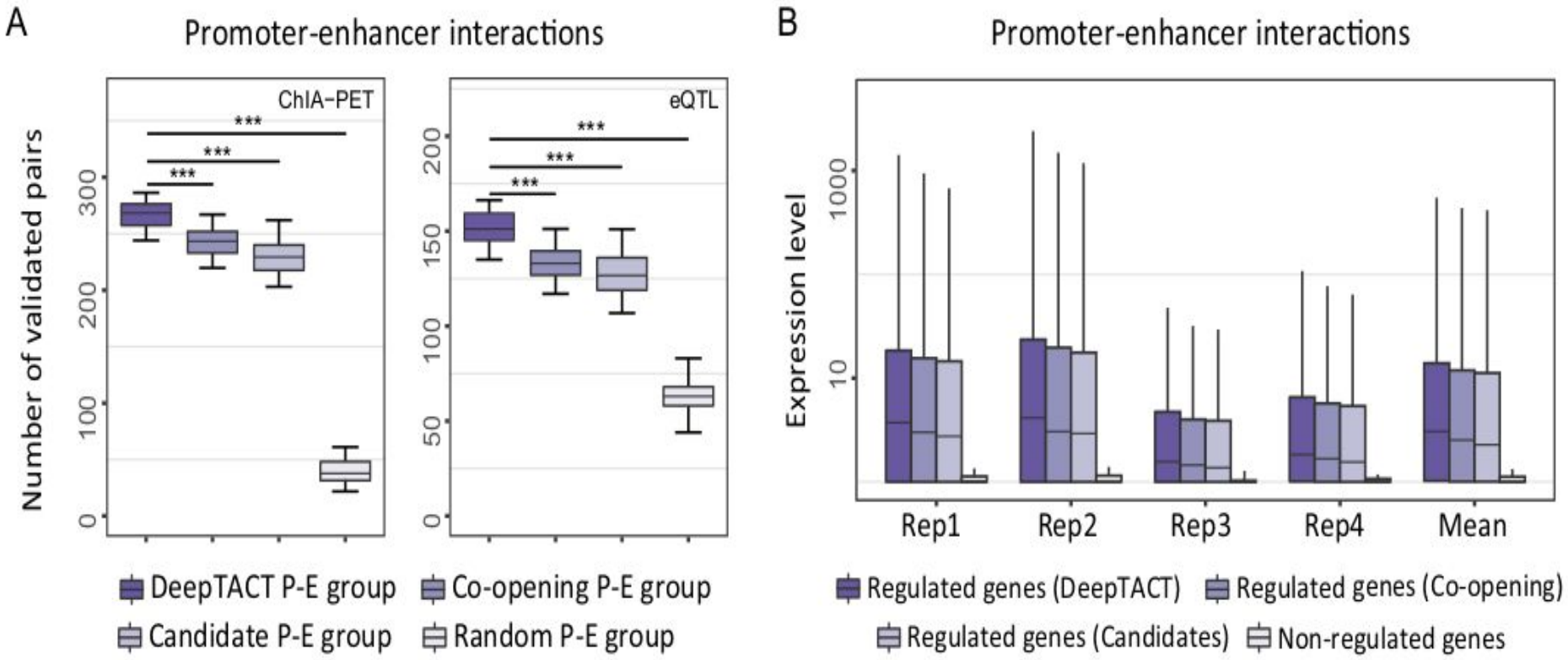

looping of DNA
Insulator
Enhancer
Genes
Insulator

DeepTACT provides finer mapping of promoter–promoter interactions from PCHi-C data.

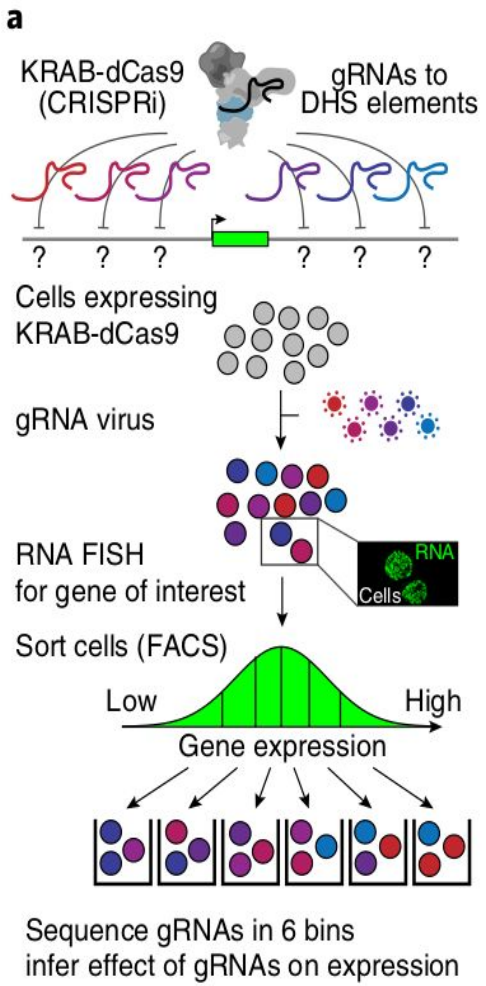


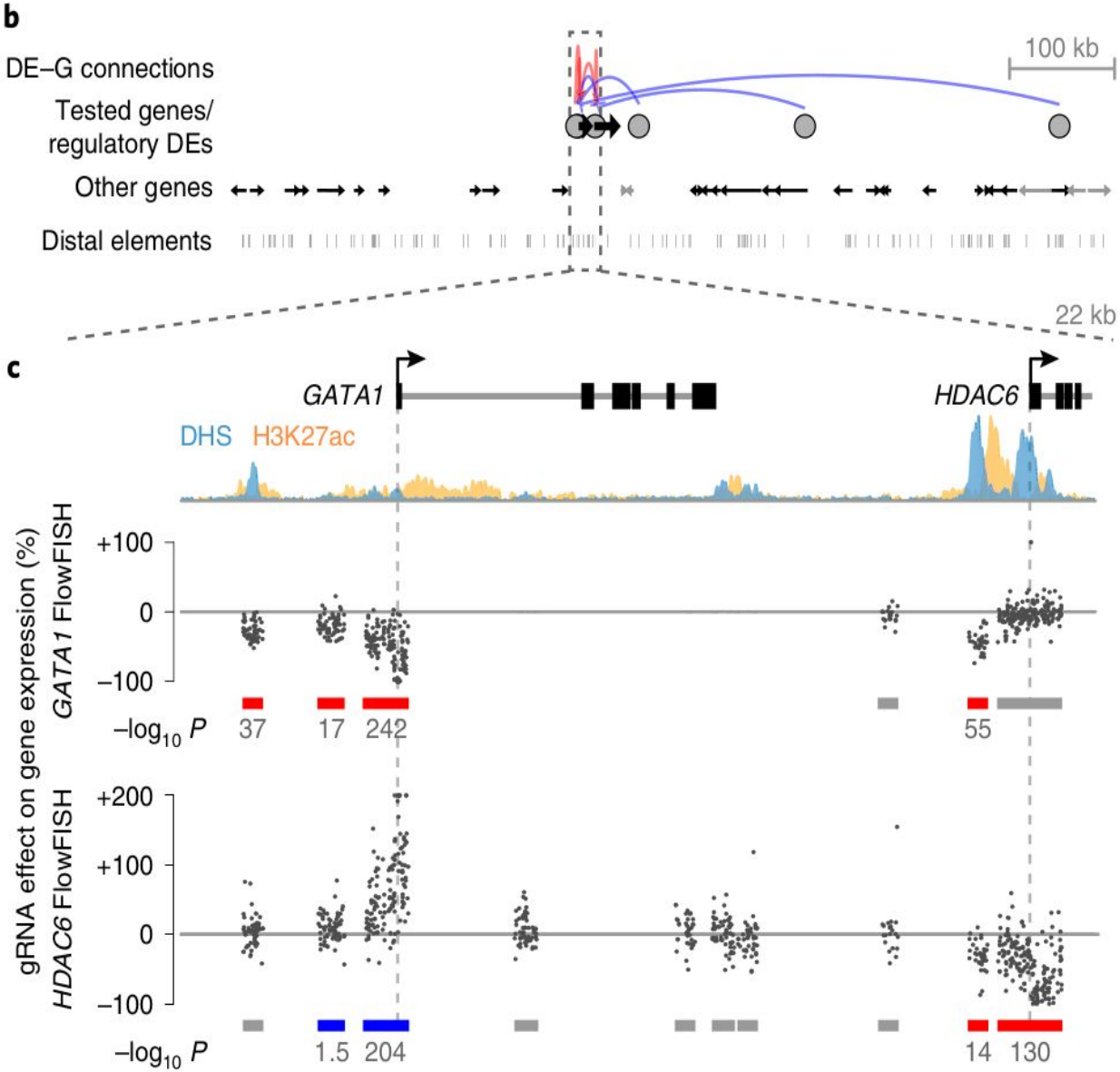
- Co-opening interactions = random sampling from the significant co-opening interactions (based on pearson correlation of openness across bioreplicates)
- Candidate interactions = all possible combinations of regulatory elements from promoter capture hic data
- Random interactions = random sampling from all possible combinations of

DeepTACT provides finer mapping of promoter–enhancer interactions from PCHi-C data.



The Activity-By-Contact (ABC) model, Fulco et al, Nature Genetics, 2019

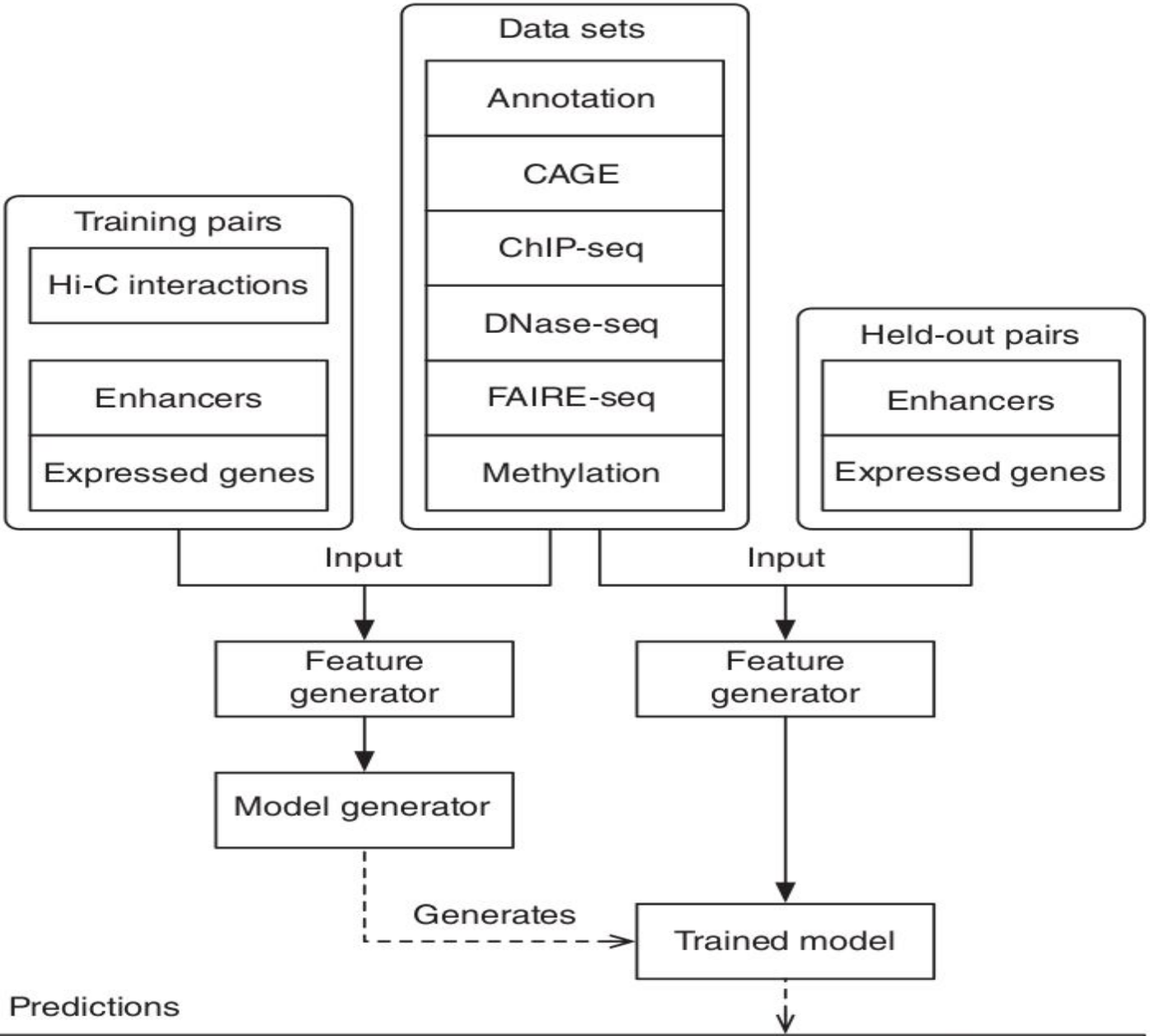




The CRISPRi-FlowFISH technique

To measure the effects of candidate elements on the expression of a gene of interest:

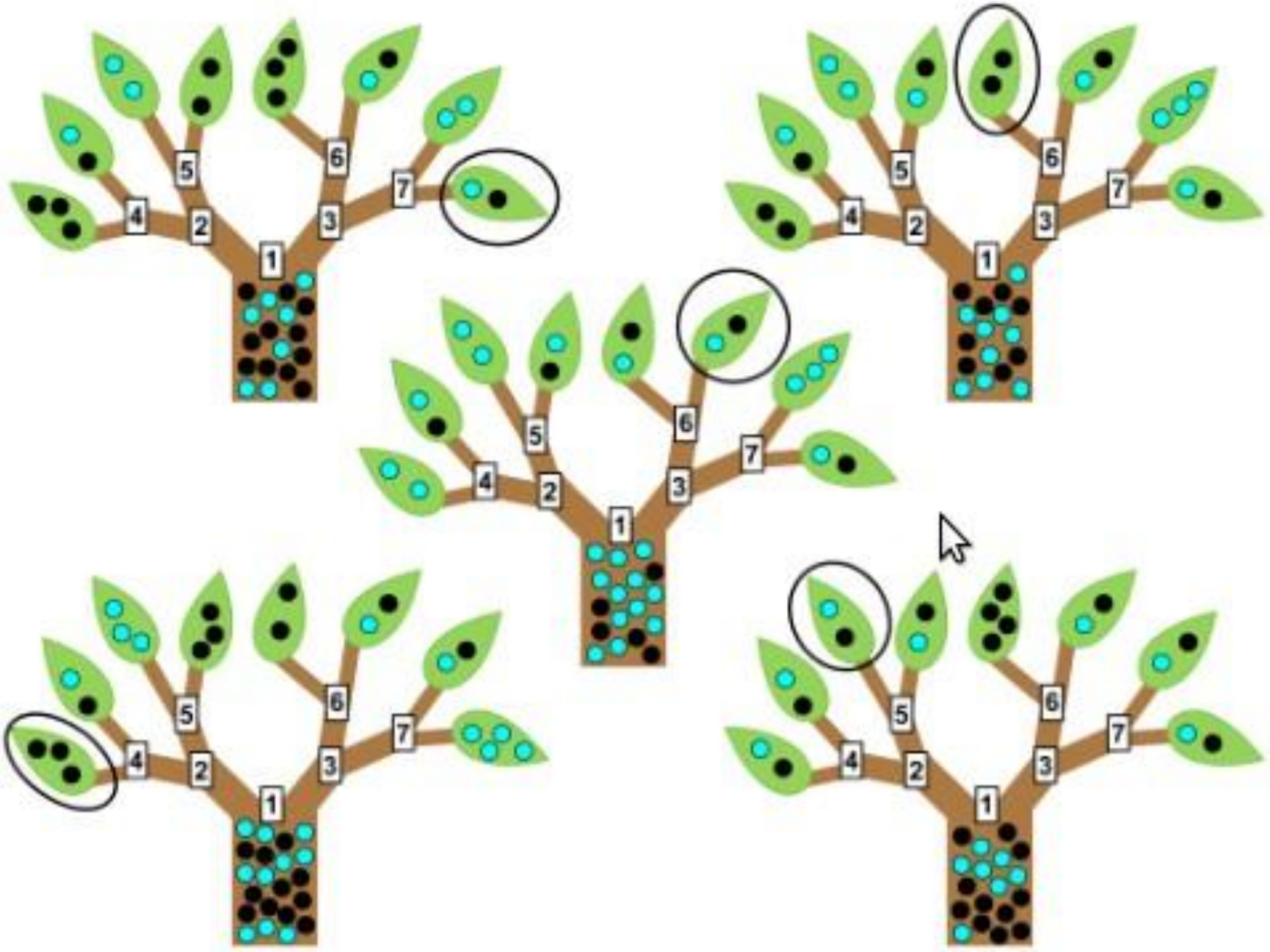
- Use RNA FISH to quantitatively label single cells according to their expression of an RNA of interest
- Sort labelled cells with FACS into six bins based on RNA abundance
- Use high-throughput sequencing to determine the abundance of each gRNA in each bin
- Use this information to infer the effect of each gRNA (i.e DHS) on gene expression (compare to 100s of negative CTRL gRNAs in the same screen to assess significance)

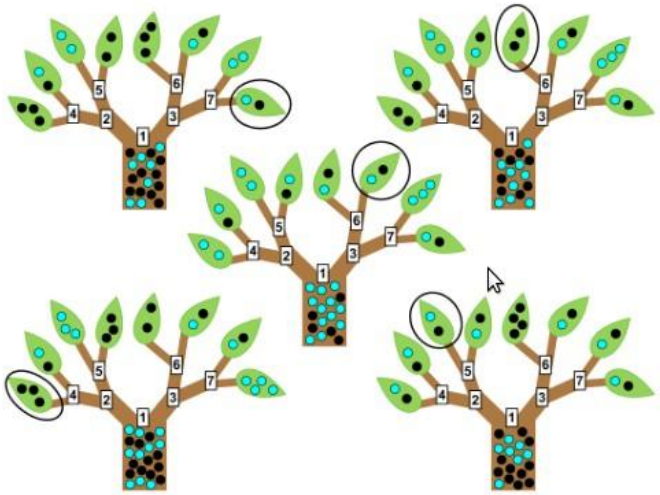


Predictions

Enhancer	Gene	Score	Score >0.5?	Prediction
chr. 1: 150–200	ENSG001	0.93	Yes	Interaction
chr. 2: 400–900	ENSG002	0.48	No	No interaction

TargetFinder, Whalen and Pollard, Nature genetics, 2015


1
2
3
4
5
6
7



Par données 3D

Vraie?	Relation
oui	E1-G1
non	E2-G2
oui	E3-G1
...	...

Données 1D

- Accessibilité ADN
- Distance E-G
- Similarité profil phylogénétique
- Présence de FT*
- Hybridation FT*

C_1 avec 1D et 3D	Apprentissage et validation croisée	Performance du modèle sur C_1
C_2 avec 1D et 3D	Application sur C_2 du modèle appris sur C_1	Performance croisée sur C_2

*FT : Facteur de Transcription