Paper overview:

Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture

Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble
arXiv preprint:1711.00137, 2017. 1, 2017

Michal Rozenwald

Higher School of Economics
Faculty of Computer Science

Feb 26, 2018

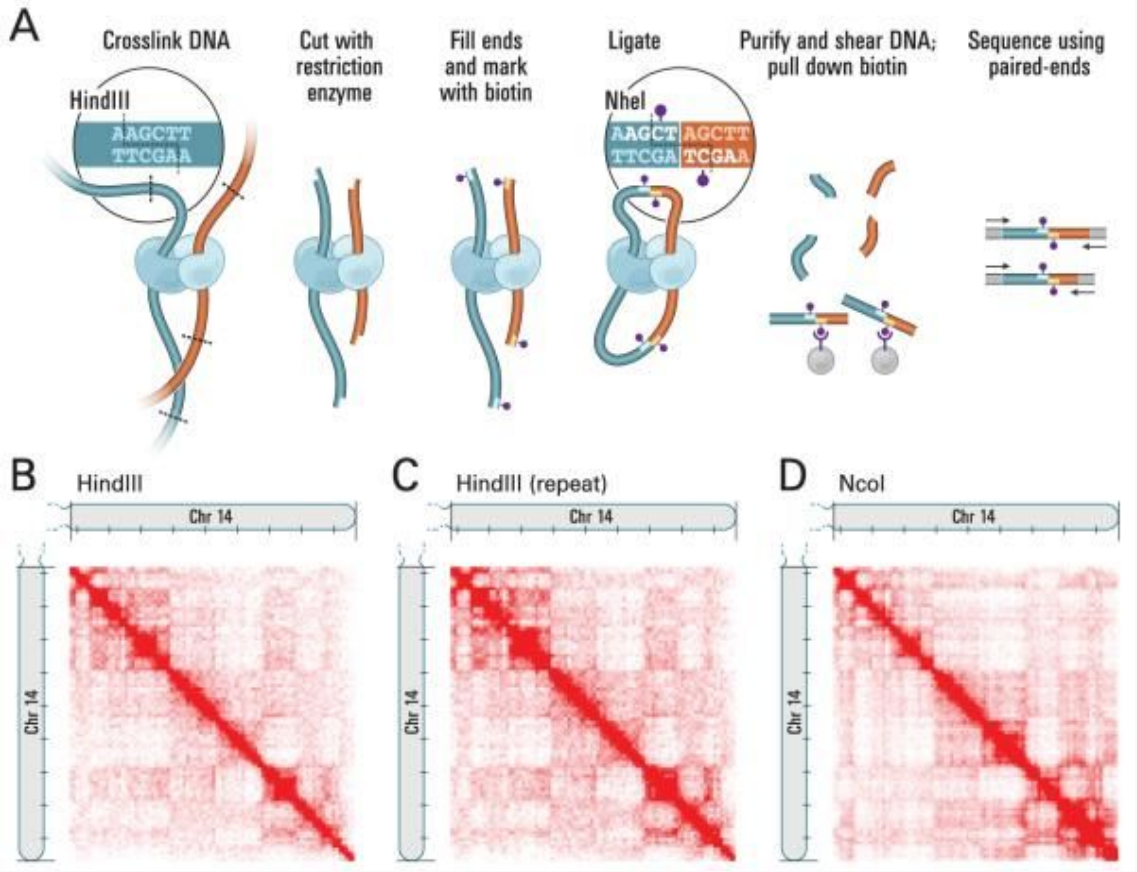
Plan

1. Motivation
2. Problem statement
3. Data
4. Rambutam Model
5. Results
6. Future work
7. Overall review

Bioinformatic Problems

- BIG data
- Many features
- Not as many samples
- Complex biology systems
- Hard to interpret features

HiC experiment



Lieberman-Aiden et al. 2009

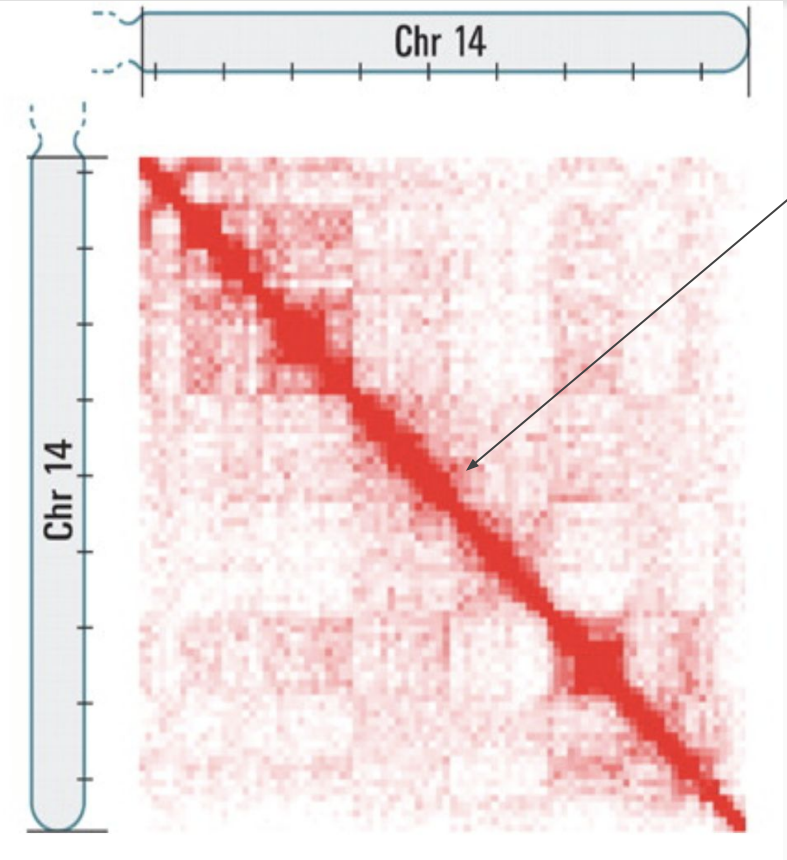
3D chromatin architecture

involved in:

- Gene regulation
- Replication timing
- Many cellular phenomena

---- Time and Cost Intensive

Model Prediction



Hi-C produces a genome-wide contact matrix.
Intrachromosomal interactions on chromosome 14.

Locus - locus connectivity

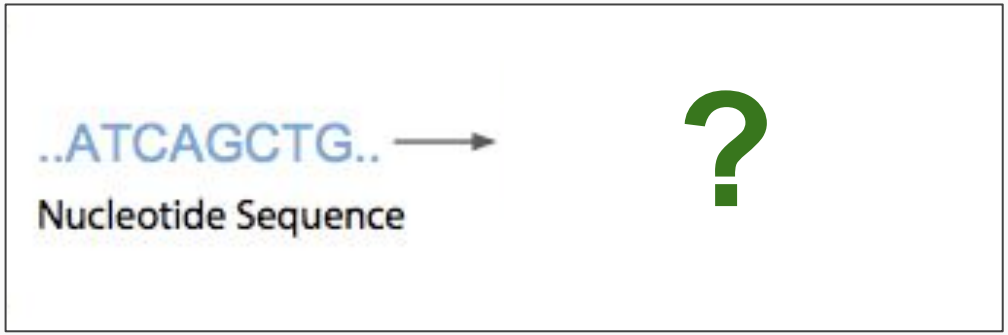
Each pixel represents all interactions between a 1Mb locus and another 1Mb locus;

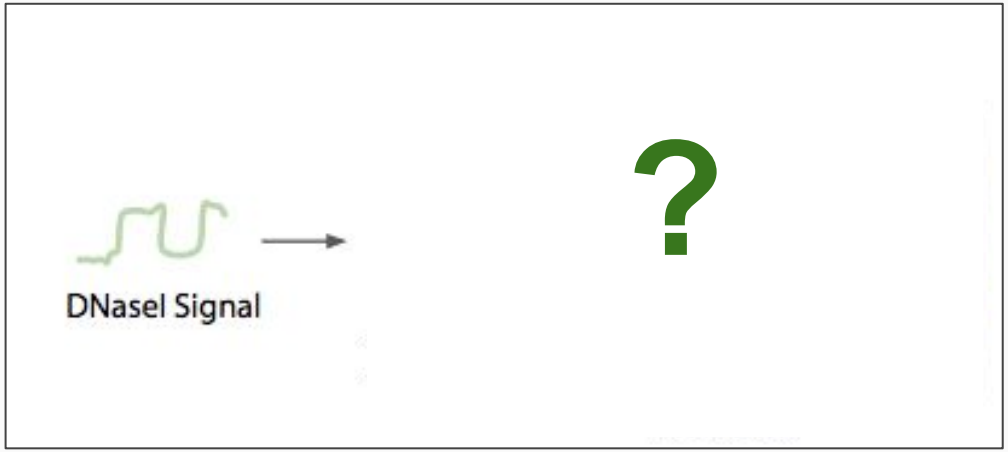
Intensity corresponds to the total number of reads (0-50).

Labels are binary:
+1 - indicates the locus pair has a q-value ≤ 1e-6

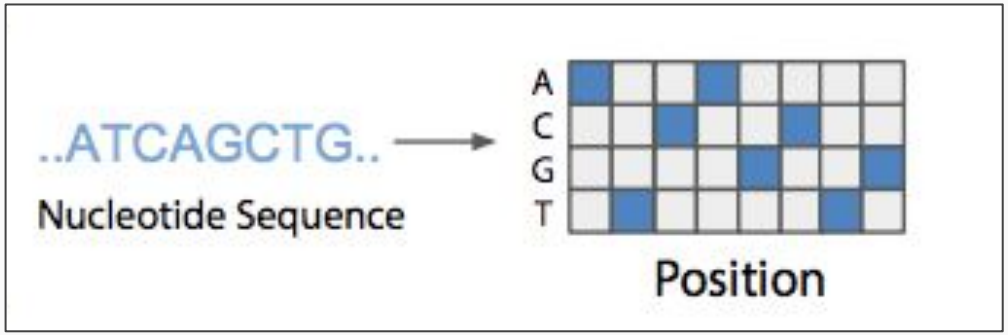
-1 - label corresponds to a q-value above that threshold.

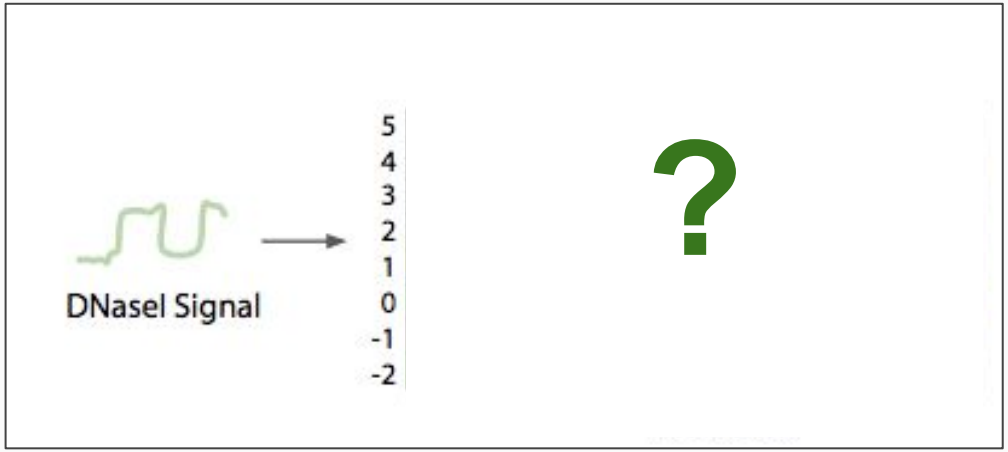
Input Data



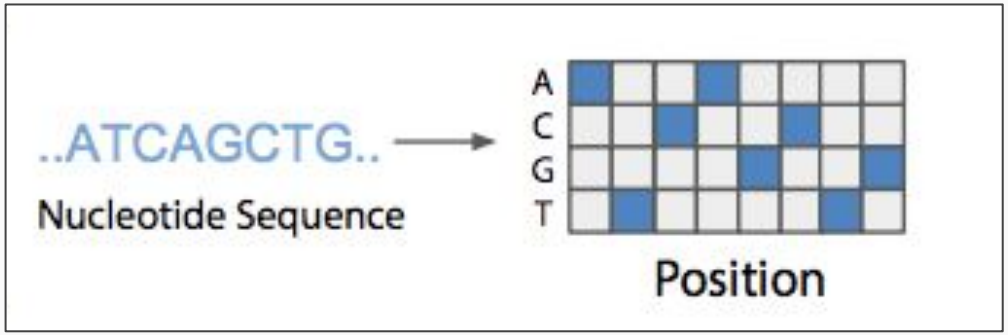


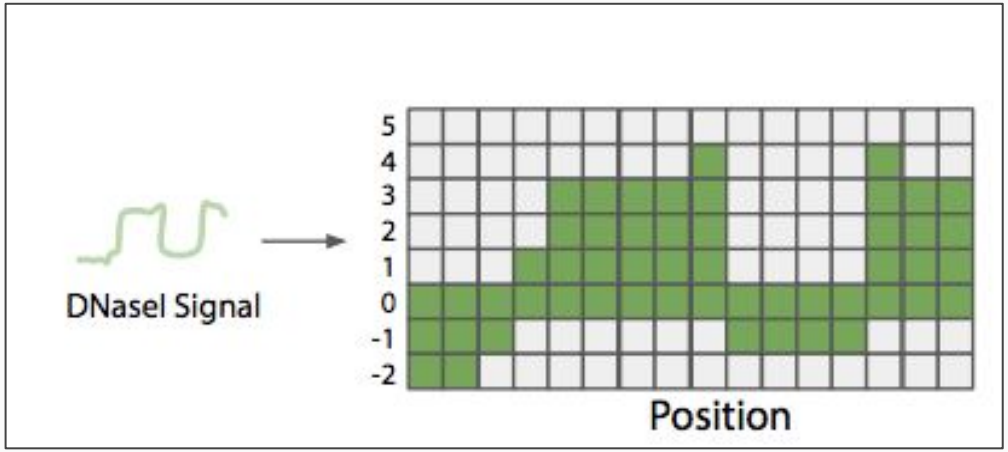
Model Input


..ATCAGCTG..
Nucleotide Sequence
A
C
G
T
Position

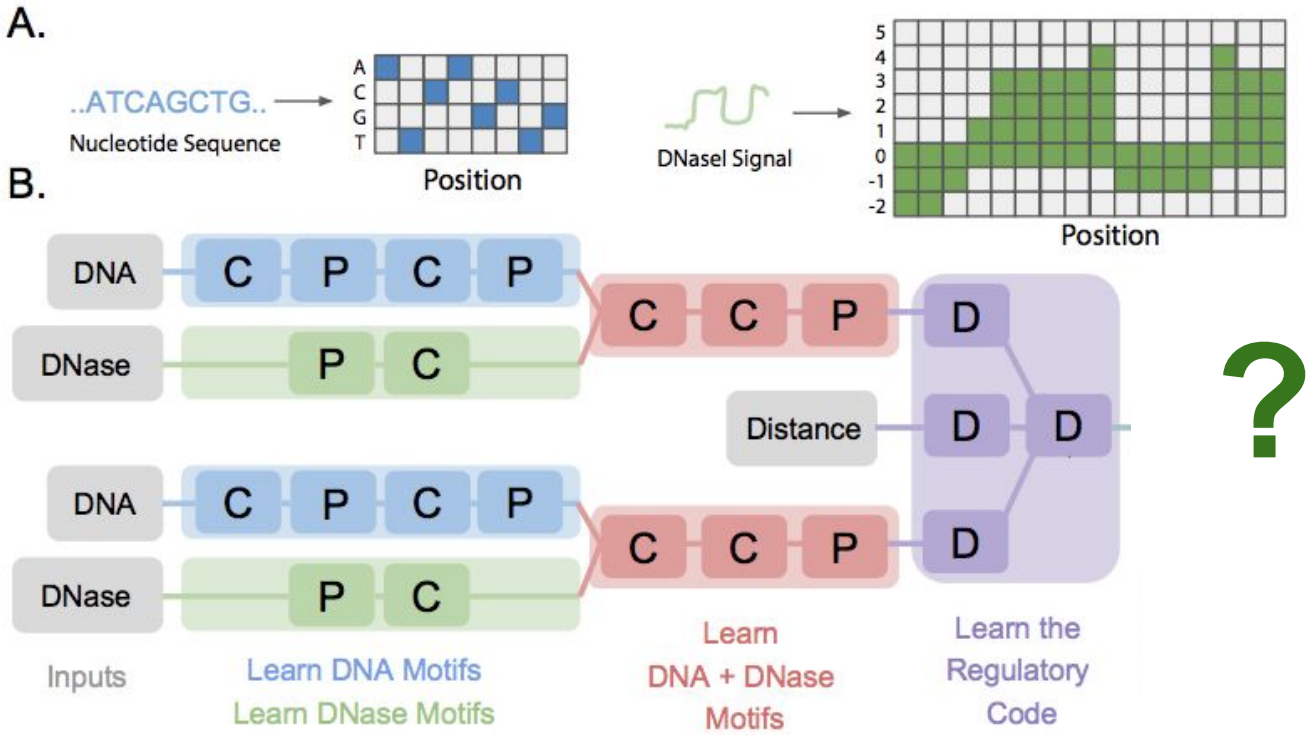

DNasel Signal
5
4
3
2
1
0
-1
-2
?

Model Input

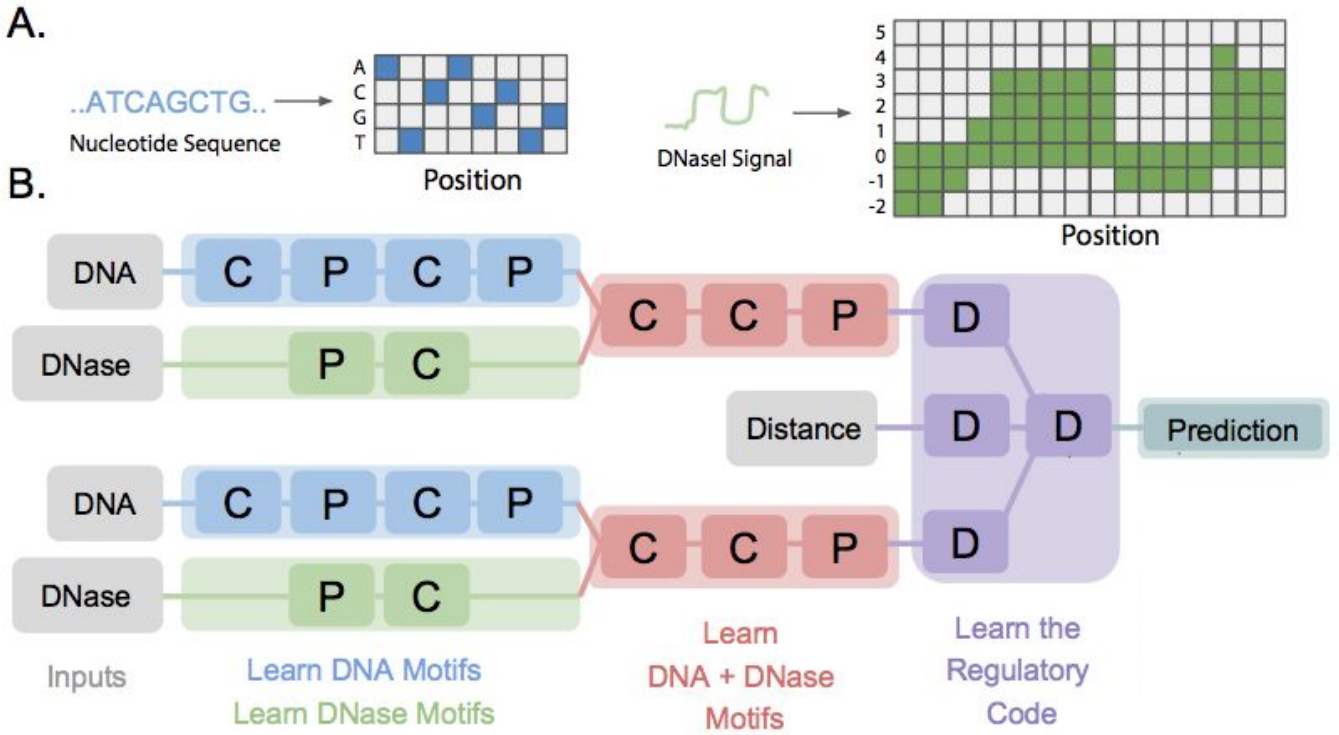

..ATCAGCTG..
Nucleotide Sequence
A
C
G
T
Position


DNasel Signal
5
4
3
2
1
0
-1
-2
Position

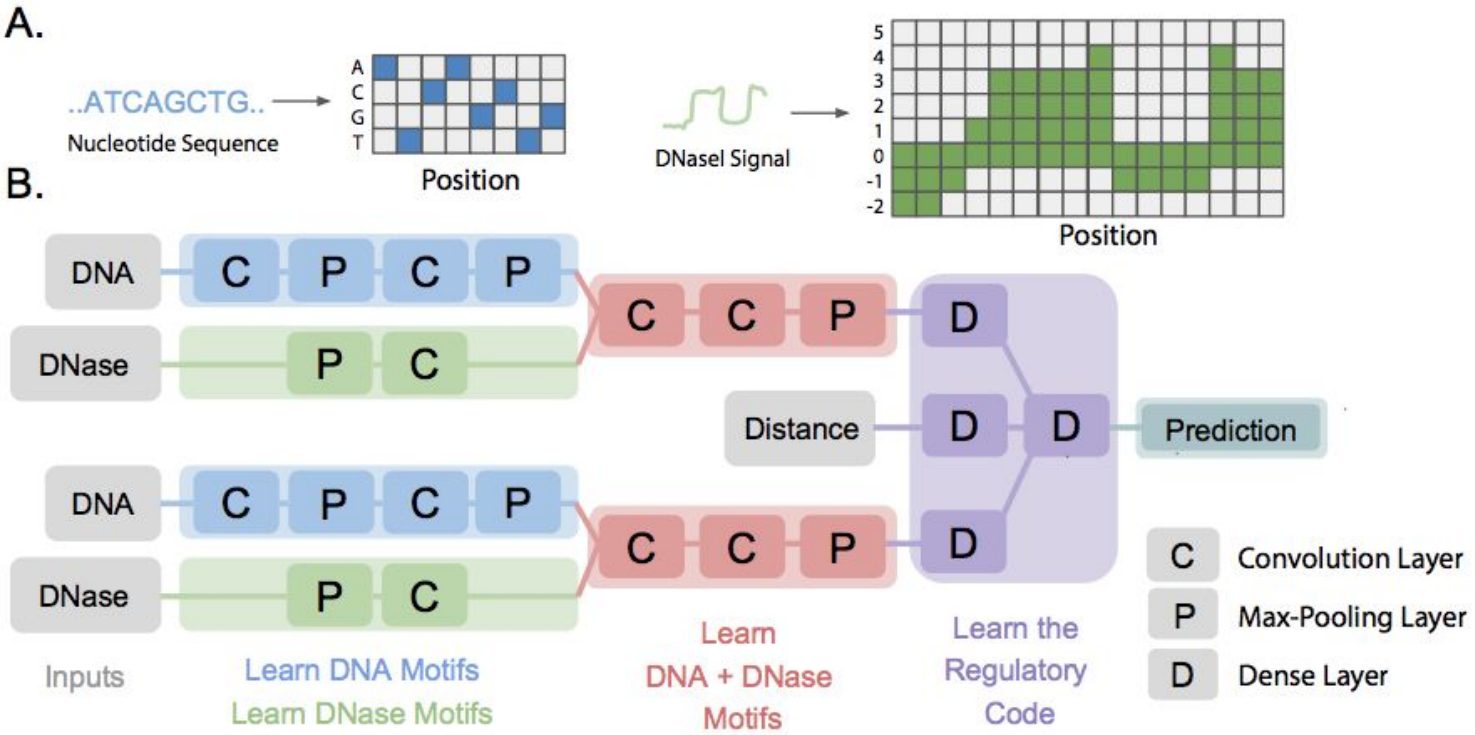
Model Topology



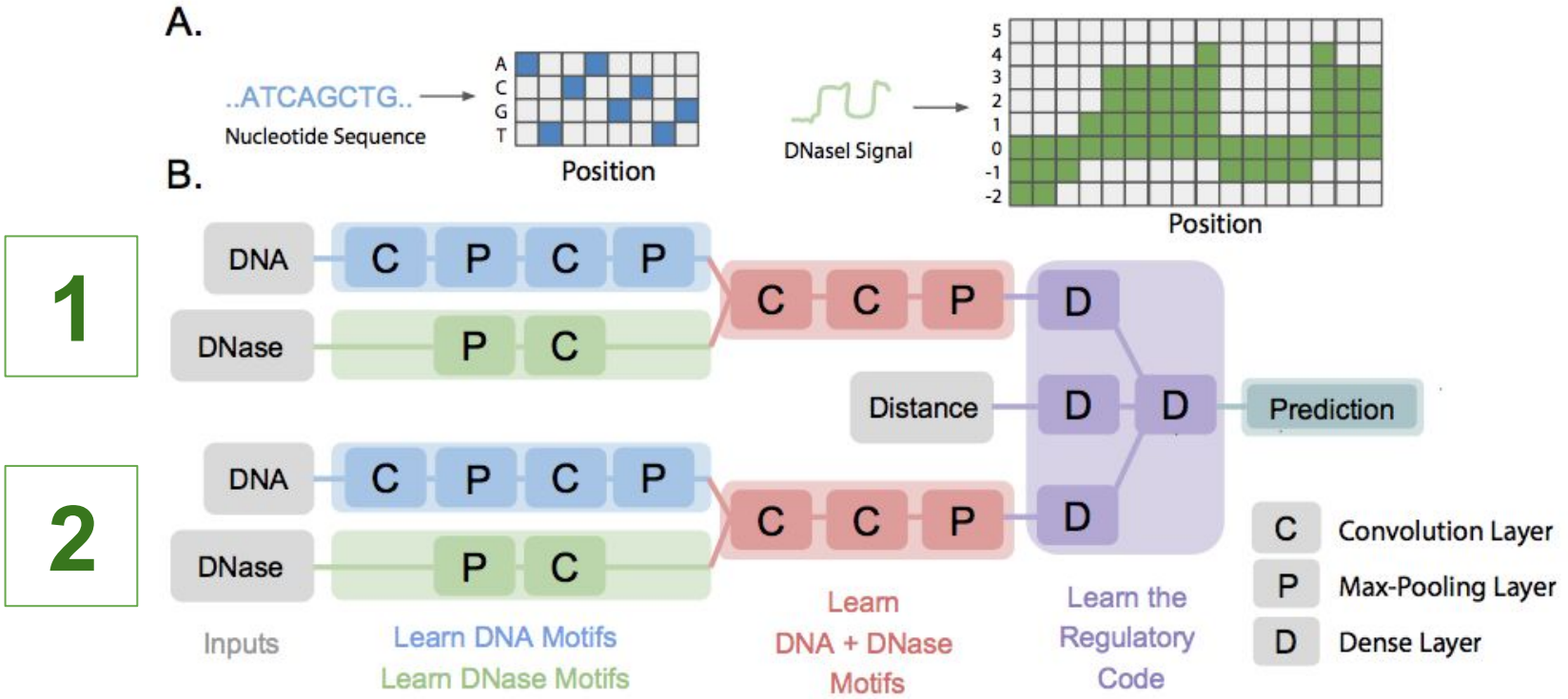
Model Topology



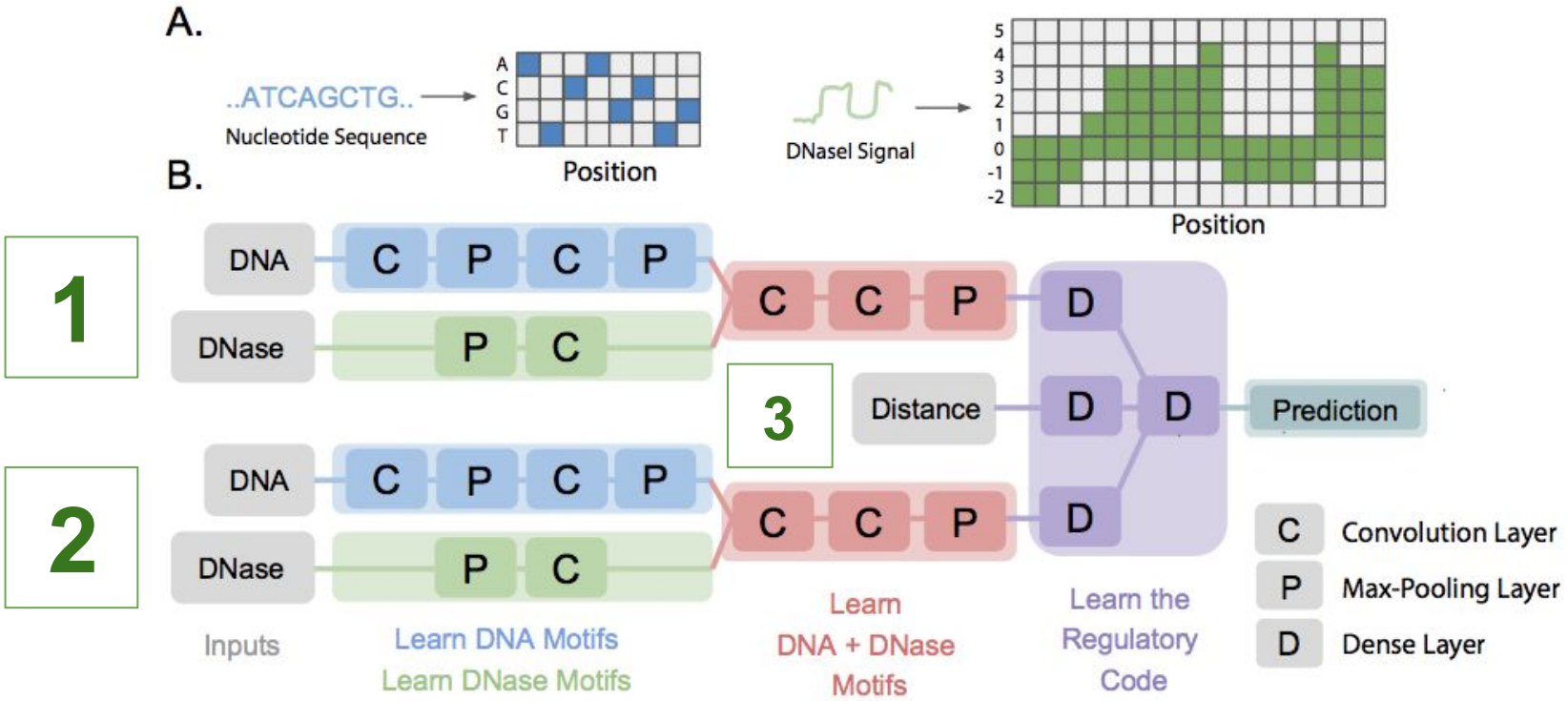
Model Topology



Model Topology

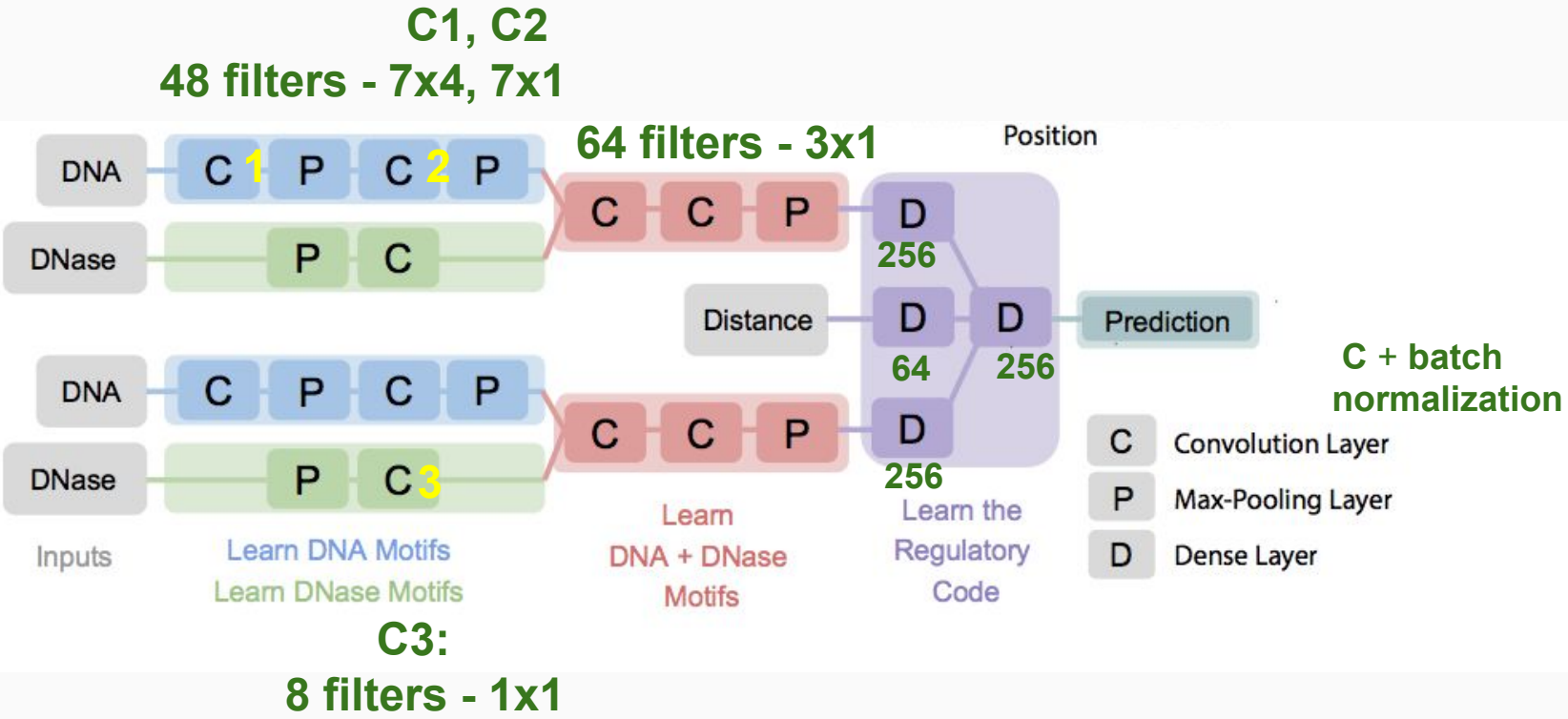


Model Topology



Model Topology

Network details:



Model Topology

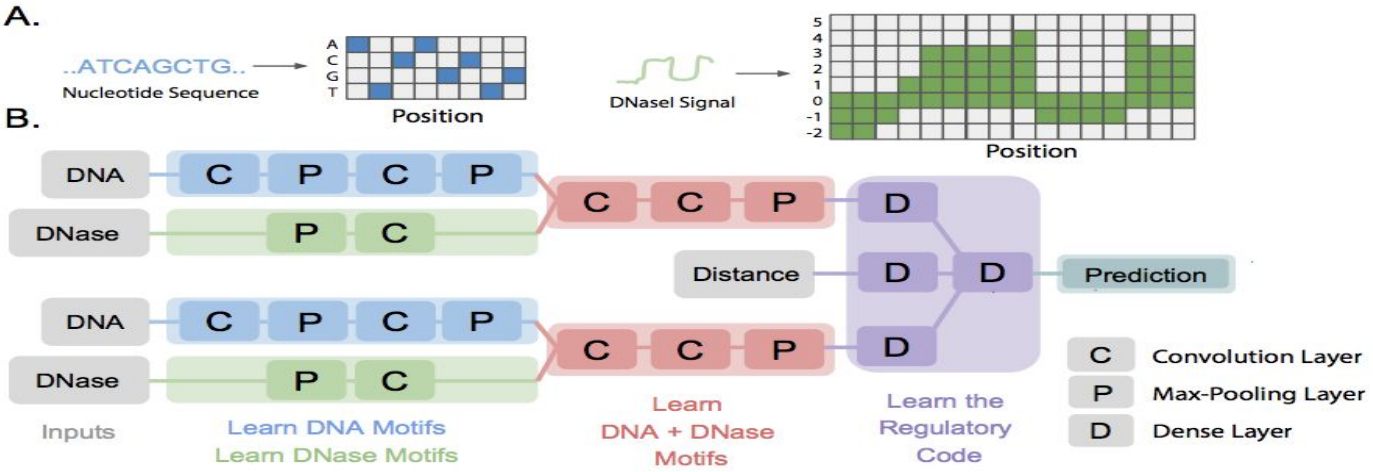


Figure 1: **The inputs to and architecture of the Rambutan network.** (a) Schematic showing how nucleotide sequence and DNase fold-change are encoded for input into the network. (b) In Rambutan, inputs flow from left to right, starting with four sources of input data and ending with a final prediction. Each symbol represents a layer in the network, colored by type: a "C" represents a convolutional layer, a "P" represents a max pooling layer, and a "D" represents a dense inner product layer. All layers are followed by a batch normalization layer then a ReLU activation, except for the final prediction layer which is just a two-node dense layer with a softmax activation for binary prediction. The blue symbols indicate layers that learn DNA-specific patterns, the green symbols indicate layers that learn DNase-specific patterns, red symbols represent layers that learn patterns comprised of both DNA and a DNase components, and purple layers learn a form of regulatory code that interprets how these patterns are predictive of a contact. Feature maps learned from the blue and green layers are concatenated before being fed into the red layers.

Parameters

Parameters:

ADAM optimizer:
- learning rate: **0.01**
- beta1: 0.9
- beta2: 0.999
- epsilon: 10^−8
- decay factor: 1 − 10^−8

Batch normalization:
- epsilon: 0.001
- momentum: 0.9

Loss: Logistic

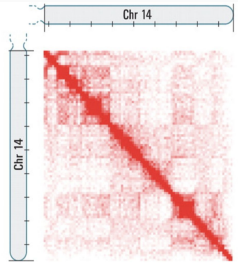
Data

Data: OPEN SOURCE DATA

1. **Human genome** (TAGCTTGAC….) - hg19 reference
- the UCSC Human Genome Browser
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/

2. **DNase data** - Roadmap **Epigenomics** Consortium
for the cell types: GM12878, K562, IMR90, NHEK, HMEC, and HUVEC
for validation: 47 other human cell types
http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/

3. **Raw Hi-C contact maps** for the same cell types
the Gene Expression Omnibus
https://www.ncbi.nlm. nih.gov/geo/query/acc.cgi?acc=GSE63525



Data: OPEN SOURCE DATA

Statistical significance for Hi-C contacts was **calculated by Fit-Hi-C** on the:

1 kb resolution contact map for **GM12878** (over **82 million** high confidence contacts)
5 kb resolution contact maps for **GM12878, K562, IMR90, NHEK, HMEC, and HUVEC (table 1)**

Table 1: **Number of contacts and non-contacts across the entire genome for each of the six cell types at 5 kb resolution.** GM12878 stands out because it has been sequenced much more deeply than the other cell types.

Cell Type	Positives	Negatives
GM12878	66,164,353	32,908,102
K562	24,969,080	69,952,576
IMR90	27,732,346	67,097,536
NHEK	13,434,985	85,470,959
HMEC	7,574,730	91,104,864
HUVEC	13,155,281	85,872,471

Data: OPEN SOURCE DATA

Statistical significance for Hi-C contacts was **calculated by Fit-Hi-C** on the:

1 kb resolution contact map for **GM12878** (over **82 million** high confidence contacts)
5 kb resolution contact maps for **GM12878, K562, IMR90, NHEK, HMEC, and HUVEC (table 1)**

Table 1: **Number of contacts and non-contacts across the entire genome for each of the six cell types at 5 kb resolution.** GM12878 stands out because it has been sequenced much more deeply than the other cell types.

Cell Type	Positives	Negatives
GM12878	66,164,353	32,908,102
K562	24,969,080	69,952,576
IMR90	27,732,346	67,097,536
NHEK	13,434,985	85,470,959
HMEC	7,574,730	91,104,864
HUVEC	13,155,281	85,872,471

~66 + 33

million

~7 + 91

Data: METABRIC

Total:

contact map contains **>2 trillion** locus pairs

Training set:

12,500 mini batches, exposing it to 12.8 million samples

1 kb resolution GM12878 contact map contains

5 kb resolution contact maps for GM12878, K562, IMR90, NHEK, HMEC, and HUVEC

Experiments

Results: 1 kb resolution predictions

1 kb resolution predictions for GM12878 (only available)

Cross-chromosome validation

	Training:	Validate:	Test:
Chromosomes:	1–20	21	22

Results: 1 kb resolution predictions

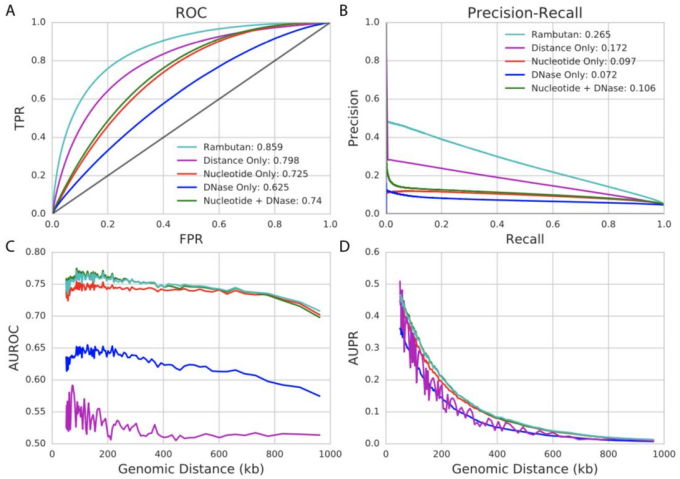


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

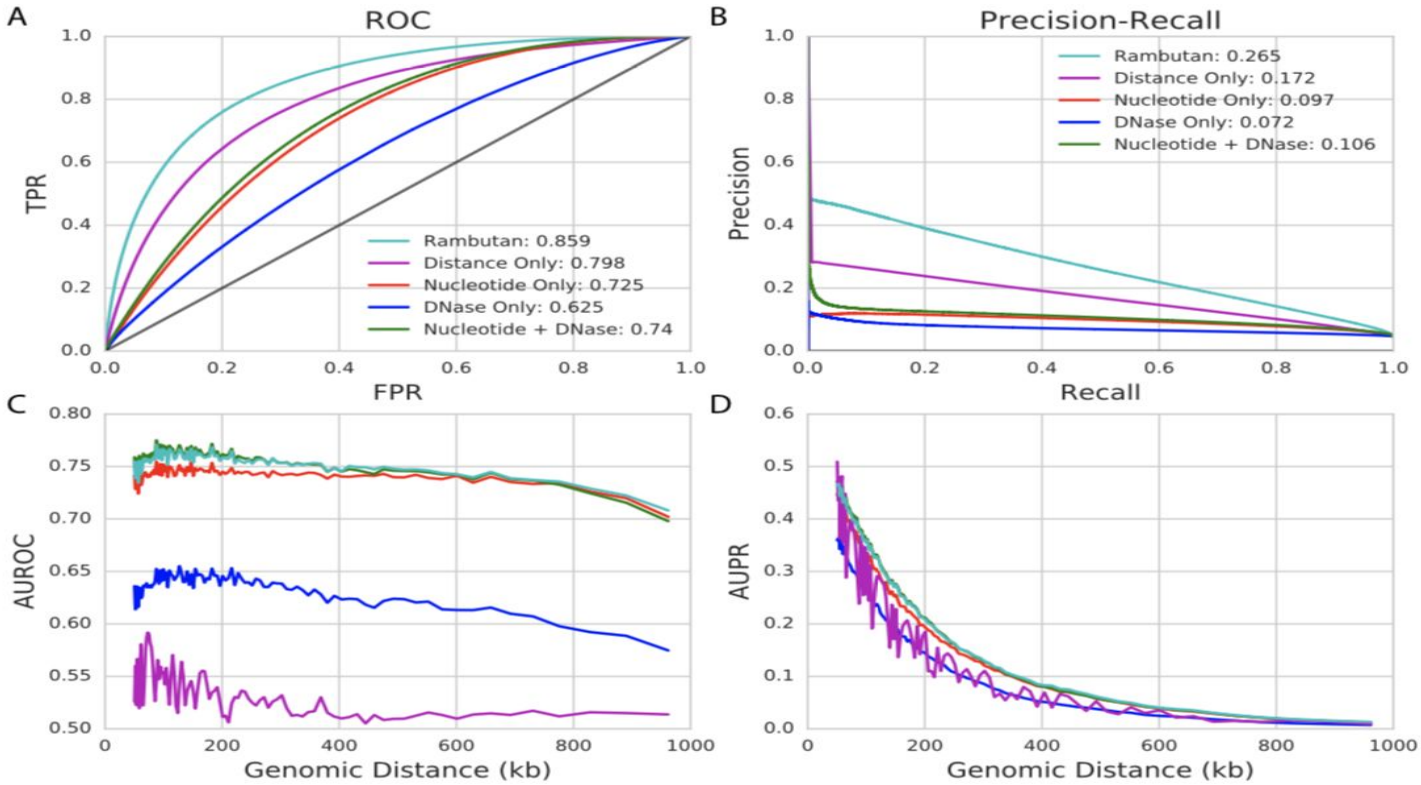


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

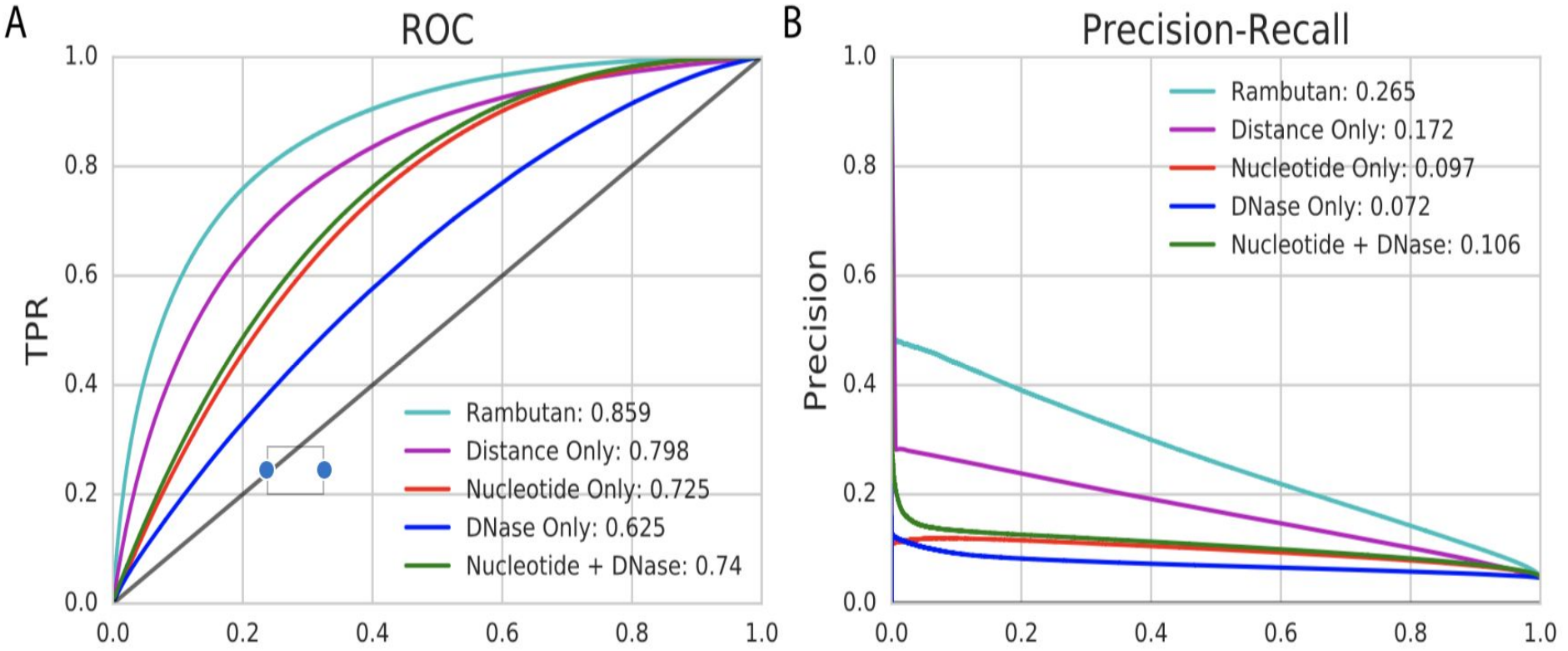


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

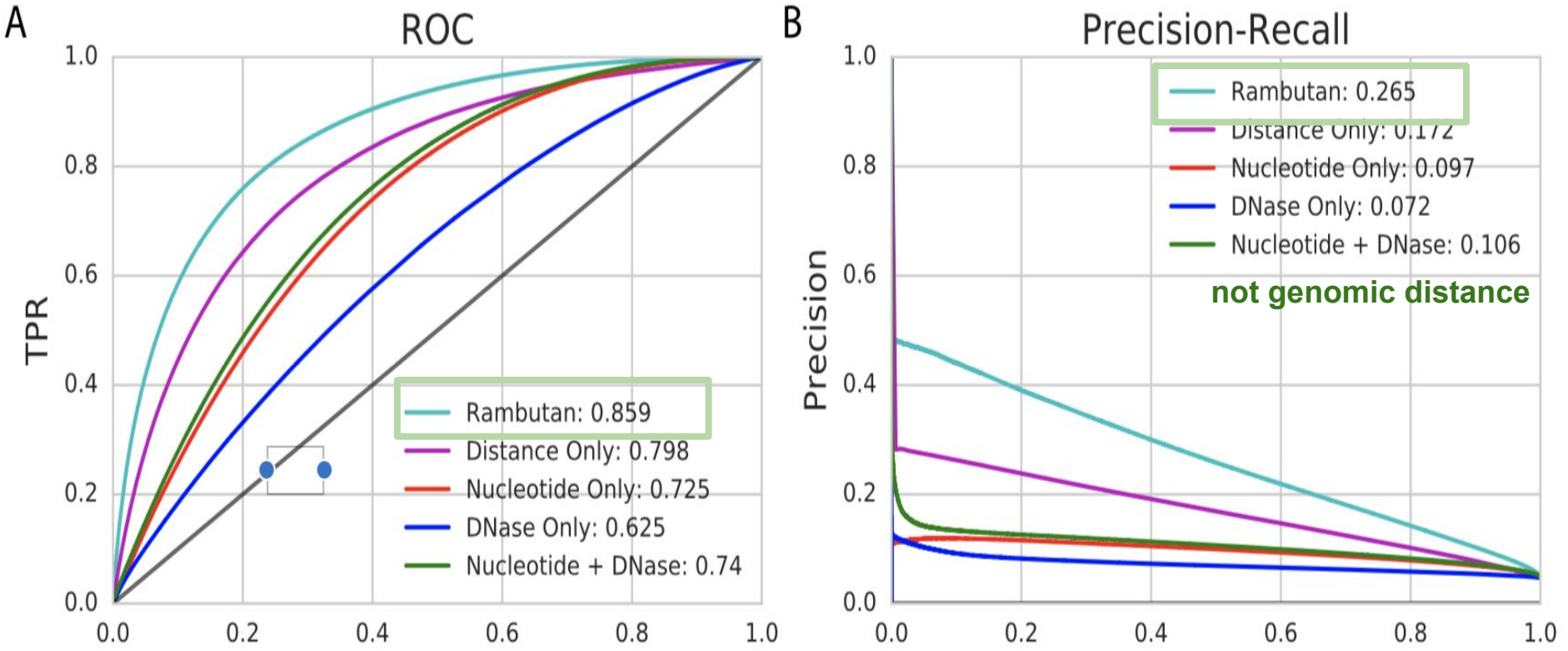
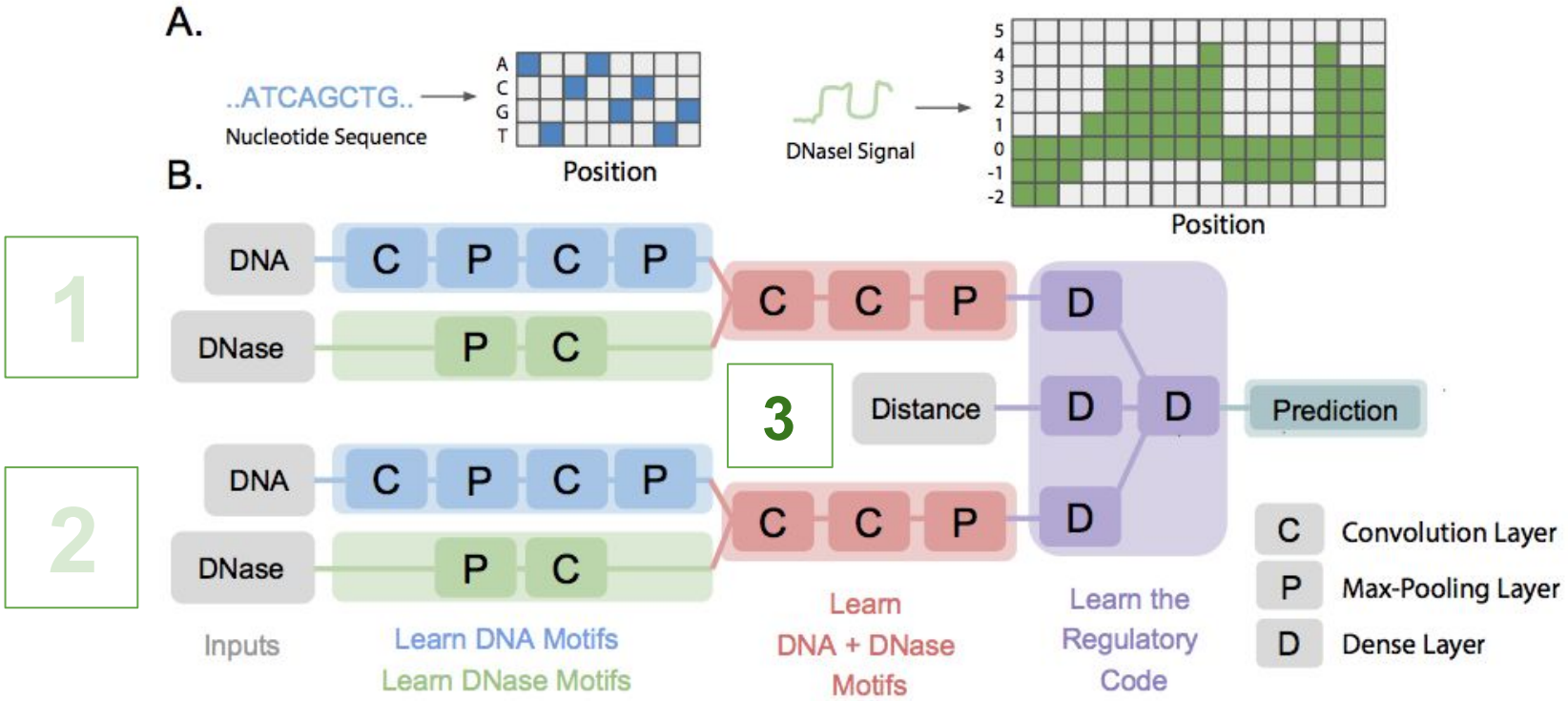


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

Distance input role



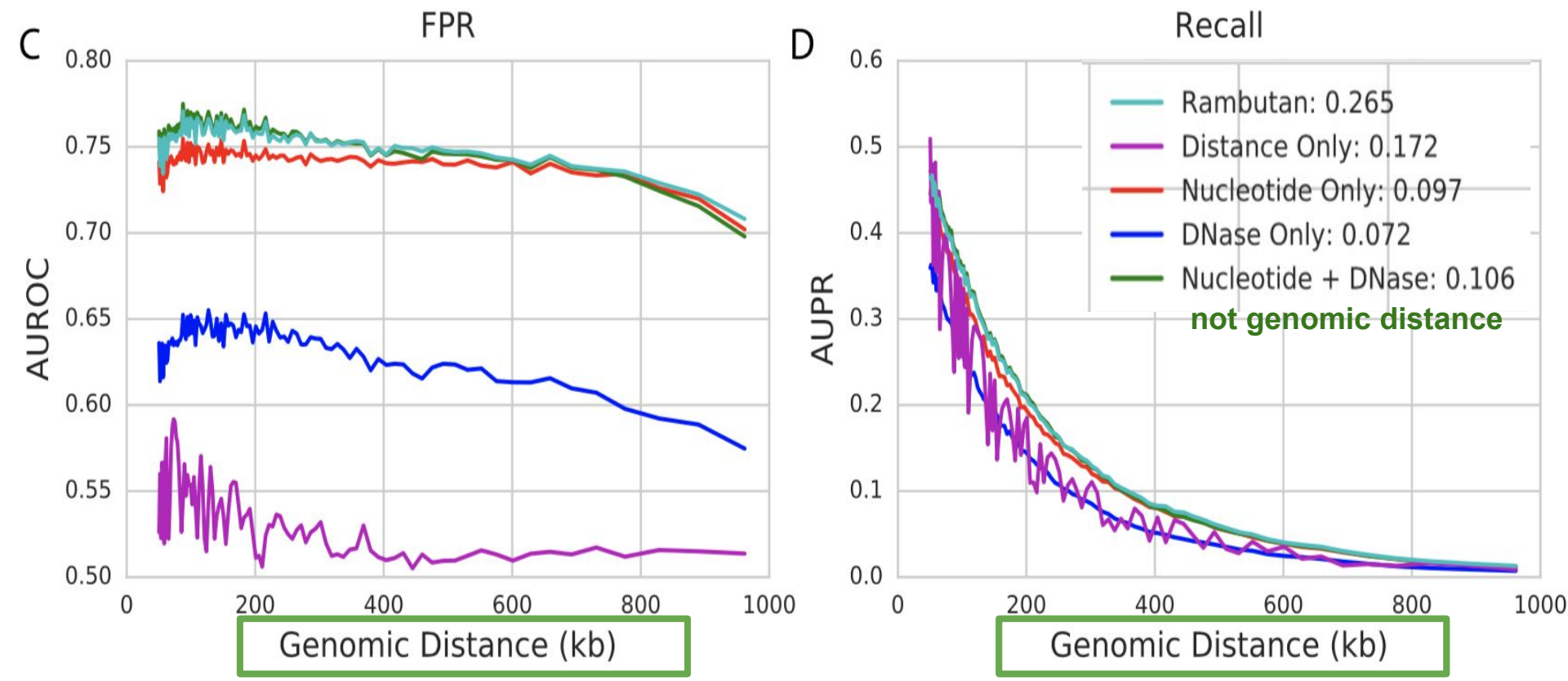


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

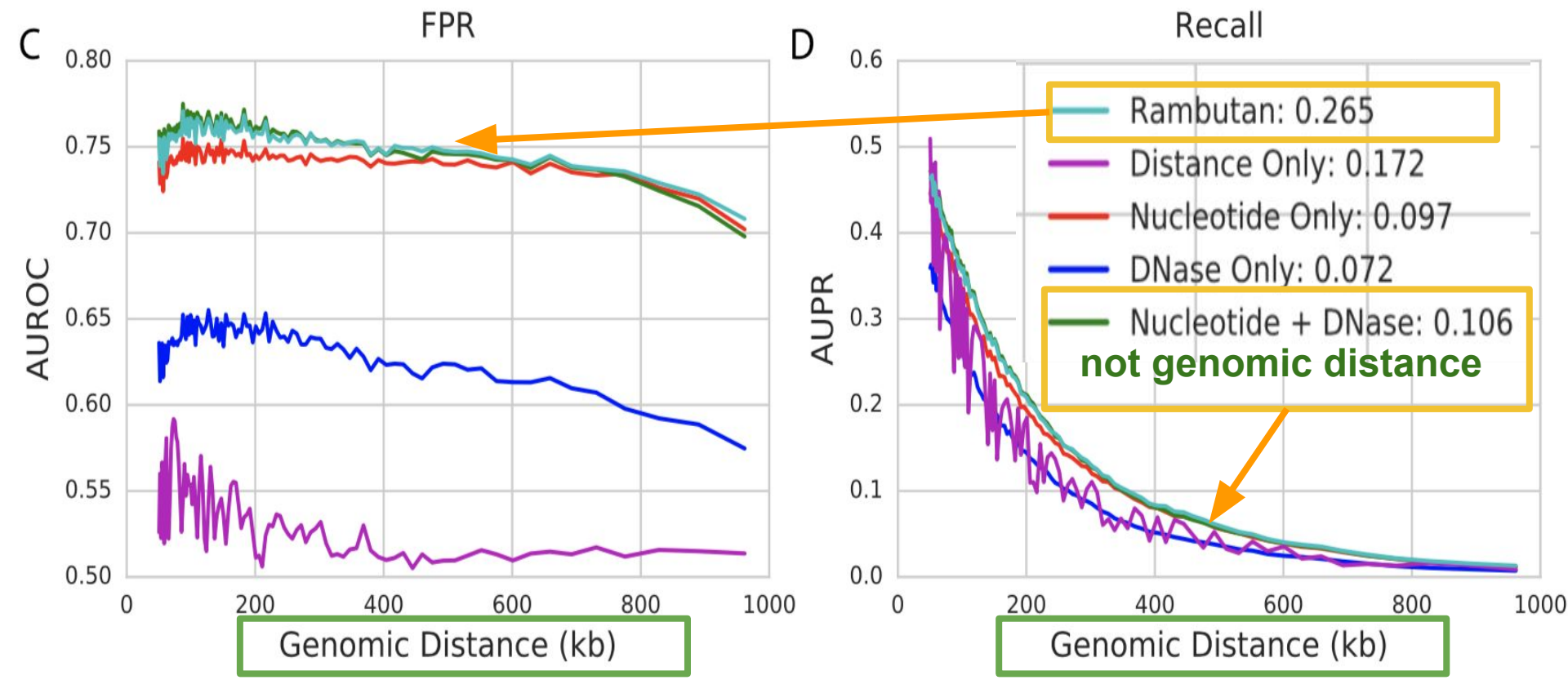


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

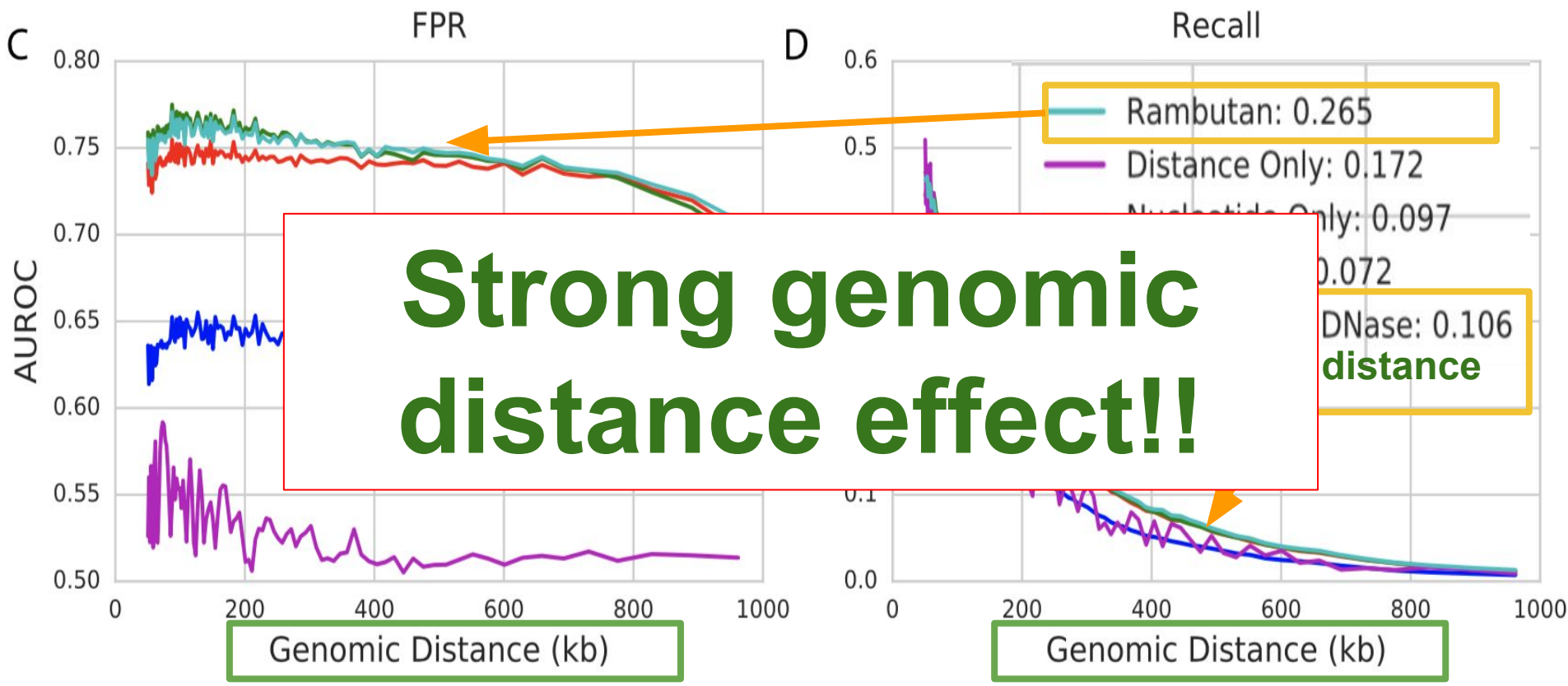


Figure 2: **Performance of the Rambutan model at 1 kb resolution.** (a and b) ROC curves and PR curves for the full Rambutan model are shown and compared to those from other baselines. The area under each of these curves is shown in the legend. (c and d) The area under the ROC curve and area under the PR curve is shown as a function of genomic distance. Since the sparsity of contacts increases as a function of genomic distance the measurements at further distances would be very imprecise. We handle this by instead using percentiles such that each point contains 1% of all true contacts when ordered by genomic distance.

Results: Rambutan makes cell type specific predictions

Statistical significance for Hi-C contacts was calculated by Fit-Hi-C on the:

1 kb resolution contact map for **GM12878**

5 kb resolution contact maps for **GM12878, K562, IMR90, NHEK, HMEC, and HUVEC** (**other cell types**)

Table 1: **Number of contacts and non-contacts across the entire genome for each of the six cell types at 5 kb resolution.** GM12878 stands out because it has been sequenced much more deeply than the other cell types.

Cell Type	Positives	Negatives
GM12878	66,164,353	32,908,102
K562	24,969,080	69,952,576
IMR90	27,732,346	67,097,536
NHEK	13,434,985	85,470,959
HMEC	7,574,730	91,104,864
HUVEC	13,155,281	85,872,471

Results: Rambutan makes cell type specific predictions

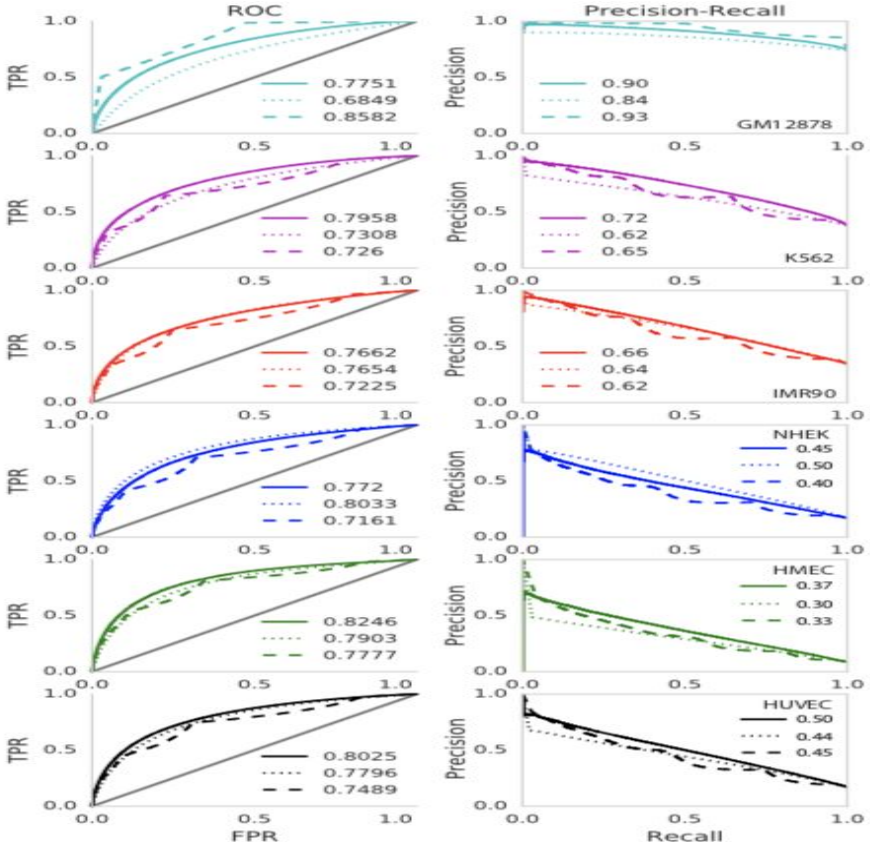


Figure 3: **Performance of the Rambutan model across cell types.** Each row corresponds to a different cell type. Solid lines correspond to Rambutan's predictions, dotted lines correspond to using genomic distance as a predictor, and dashed lines correspond to using GM12878's contact map as a predictor.

Results: Rambutan makes cell type specific predictions

Using the model trained in GM12878:
predict **1 kb** resolution contact maps for each of these cell types,
convert the signal to 5 kb resolution

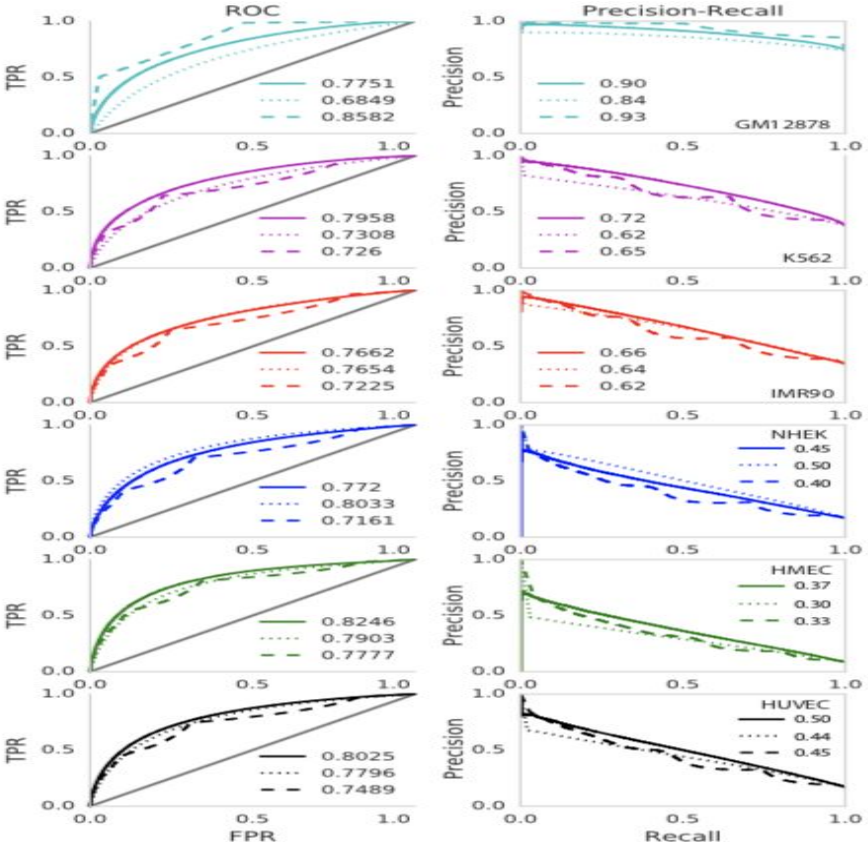


Figure 3: **Performance of the Rambutan model across cell types.** Each row corresponds to a different cell type. Solid lines correspond to Rambutan's predictions, dotted lines correspond to using genomic distance as a predictor, and dashed lines correspond to using GM12878's contact map as a predictor.

Results: Rambutan makes cell type specific predictions

Using the model trained in GM12878:
predict **1 kb** resolution contact maps for each of these cell types,
convert the signal to 5 kb resolution

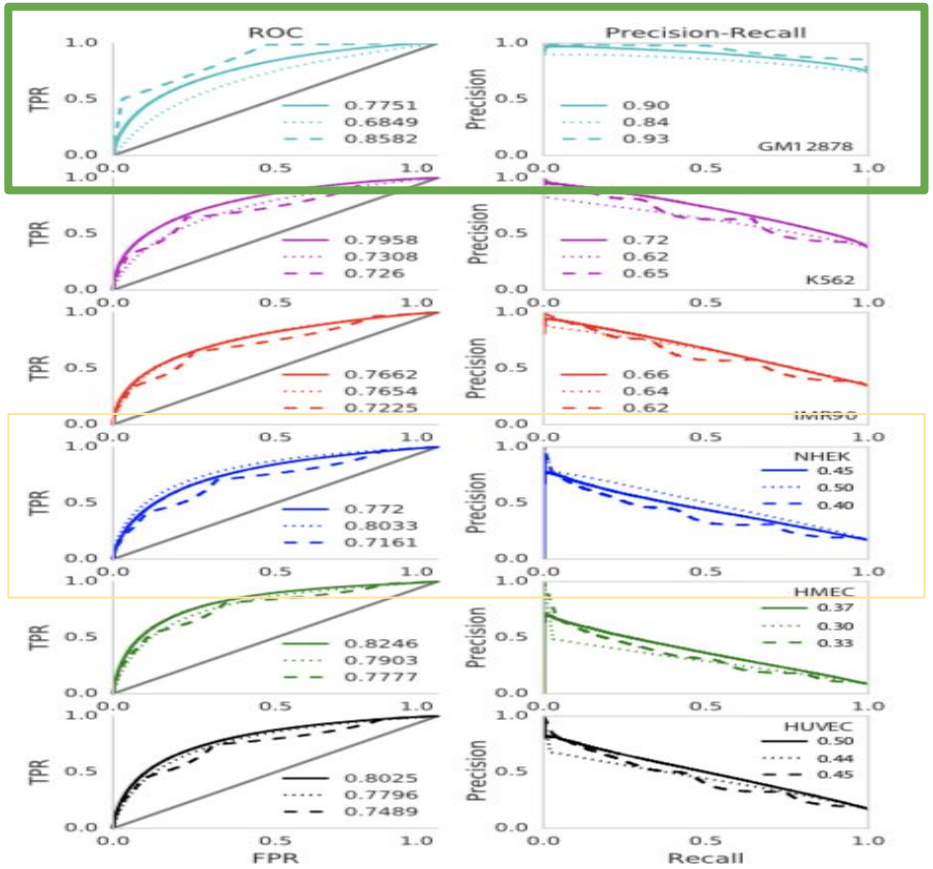


Figure 3: **Performance of the Rambutan model across cell types.** Each row corresponds to a different cell type. Solid lines correspond to Rambutan's predictions, dotted lines correspond to using genomic distance as a predictor, and dashed lines correspond to using GM12878's contact map as a predictor.

Results

Results

Simpler models such as logistic regression and a random forest classifier performed **no better** than using **genomic distance alone**.

The experiments confirm:

- Rambutan achieved a test set **AUROC**: 0.846, **AUPR**: 0.373
- **Nucleotide sequence** and **DNase** were leveraged **together for better** performance than using either input separately
- Rambutan performs well at all **genomic distances**, with AUROC not decreasing significantly as distance increases
- Trained on GM12878 of 1 kb resolution contact maps **predicted** 5 kb resolution contact maps for **other cell types**

Interesting, working approach.

“Full Rambutan model strongly outperforms each of the other methods that we investigated.”

Ideas for future work:

- ➔ Use another or more features as input
- ➔ Copmere the results with existed methods of HiC prediction
- ➔ New approaches for interpretation
- ➔ Comparison with existing other bio markets
- ➔ Use deeper NN
- ➔ Use RNN

Questions?

Thank you

Reference

Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture

Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble
arXiv preprint:1711.00137, 2017. 1, 2017

https://www.biorxiv.org/content/early/2017/01/30/103614

Comprehensive mapping of long-range interactions reveals folding principles of the human genome.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoy- annopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J.
Science, 326(5950), 289–293. (2009).

Additional Slides

Changing resolution

Rambutan makes predictions at **1 kb** resolution that can be processed to obtain **5 kb resolution** contact maps.

Each pair of 5 kb loci corresponds to 25 pairs of 1 kb loci.

The procedure involves using either the maximum Rambutan prediction or the minimum Fit-Hi-C p-value among those 25, depending on which type of data is being analyzed.

The max function treats the values within that square as not being independent from each other, and was found empirically to perform the best among several other methods.

Insulation Score

The **insulation score** measures the extent to which a given genomic locus exhibits a pattern of local contacts **indicative** of the **boundary of a TAD** (Crane et al., 2015). Calculating the score involves running a square down the **diagonal** of the contact map and **summing** the number of **contacts within that square** for each position in the chromosome. This score is then converted into a **log** fold enrichment over the average number of contacts:

$$IS_i = \log(x_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right), \qquad (1)$$

where xi is the **sum of contacts** within the **square surrounding locus i**

We calculate the insulation score using a 1 Mb square for both the **original Hi-C maps** and for the **Rambutan predictions**. Second calculation sums the probability of each pair of loci being in contact instead of the number of contacts.

Thanks!

michal.rozenwald@gmail.com