SilentMethyl: a multimodal framework for CpG methylation prediction and local variant-effect inference

Samuel Zhang

Mentors: Prof. Gil Alterovitz (Harvard Medical School), Dr. Shaojun Pei (Brigham and Women's Hospital)

MIT PRIMES Conference, October 2026

Table of Contents

1 Introduction

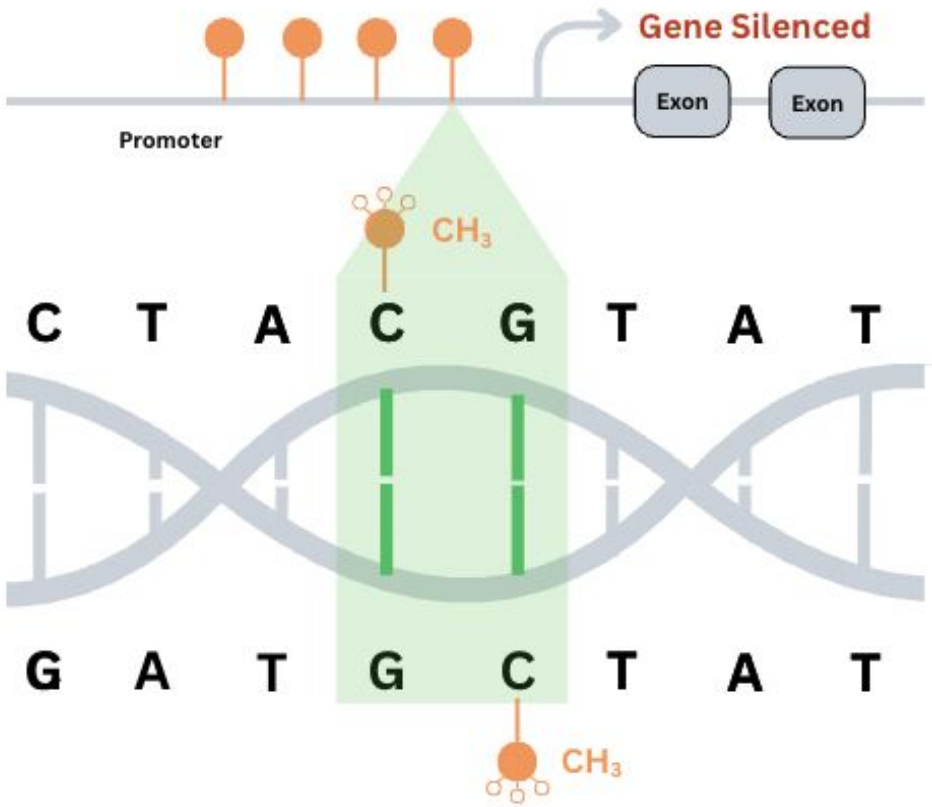
2 Previous Works

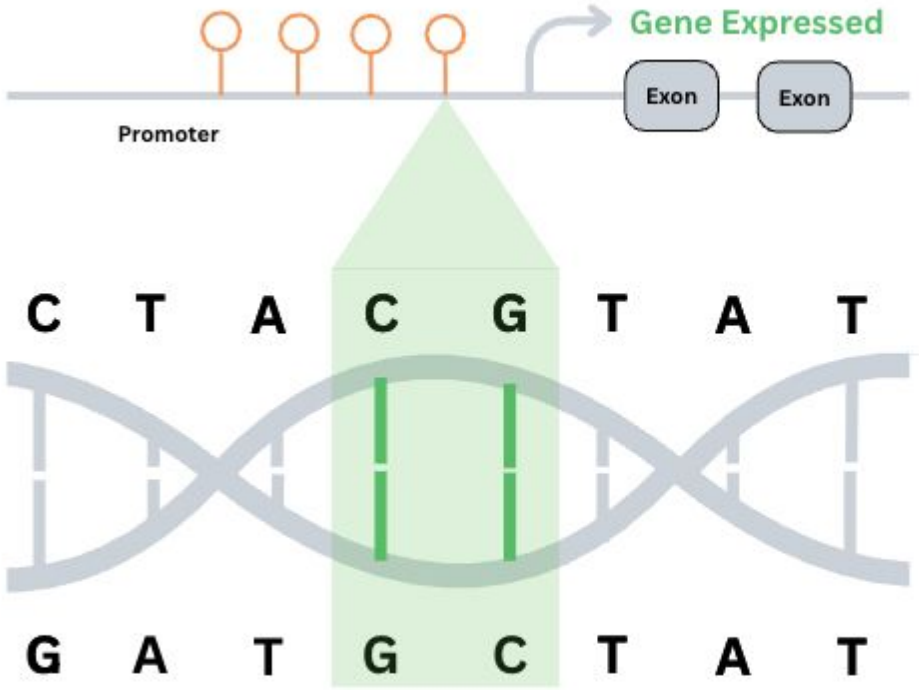
3 Methods

4 Results

5 Conclusions

DNA Methylation



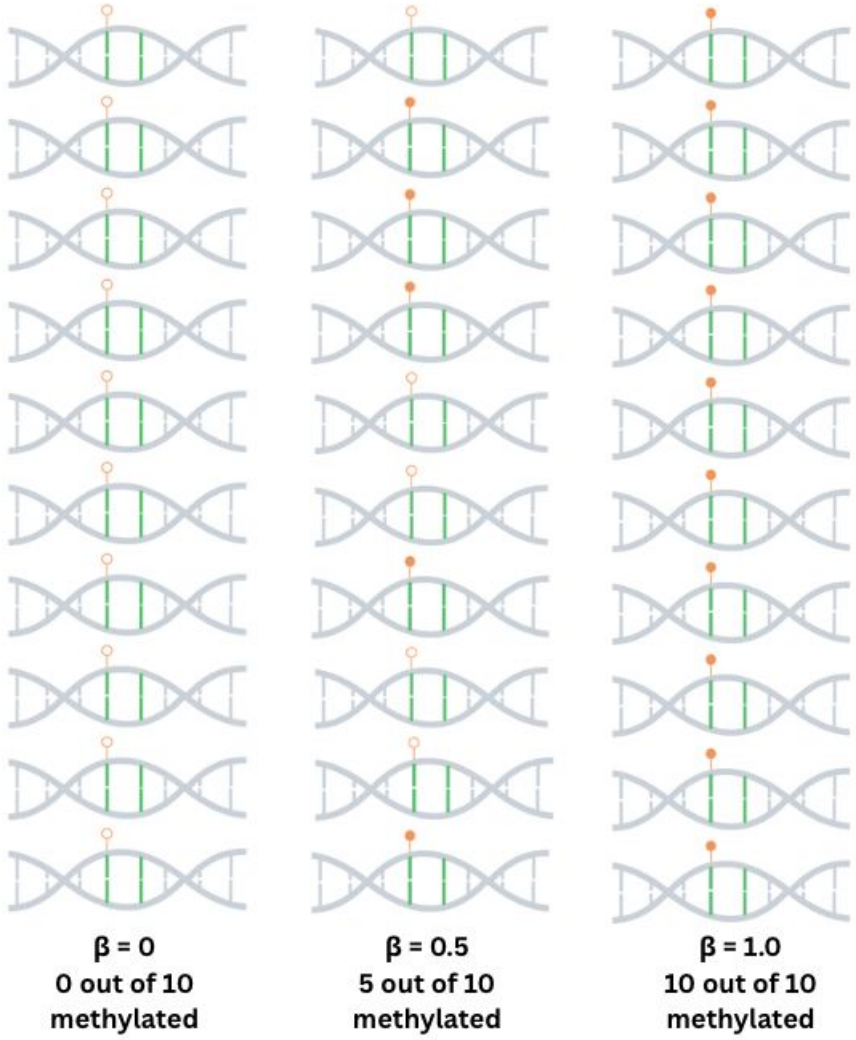


Measuring methylation

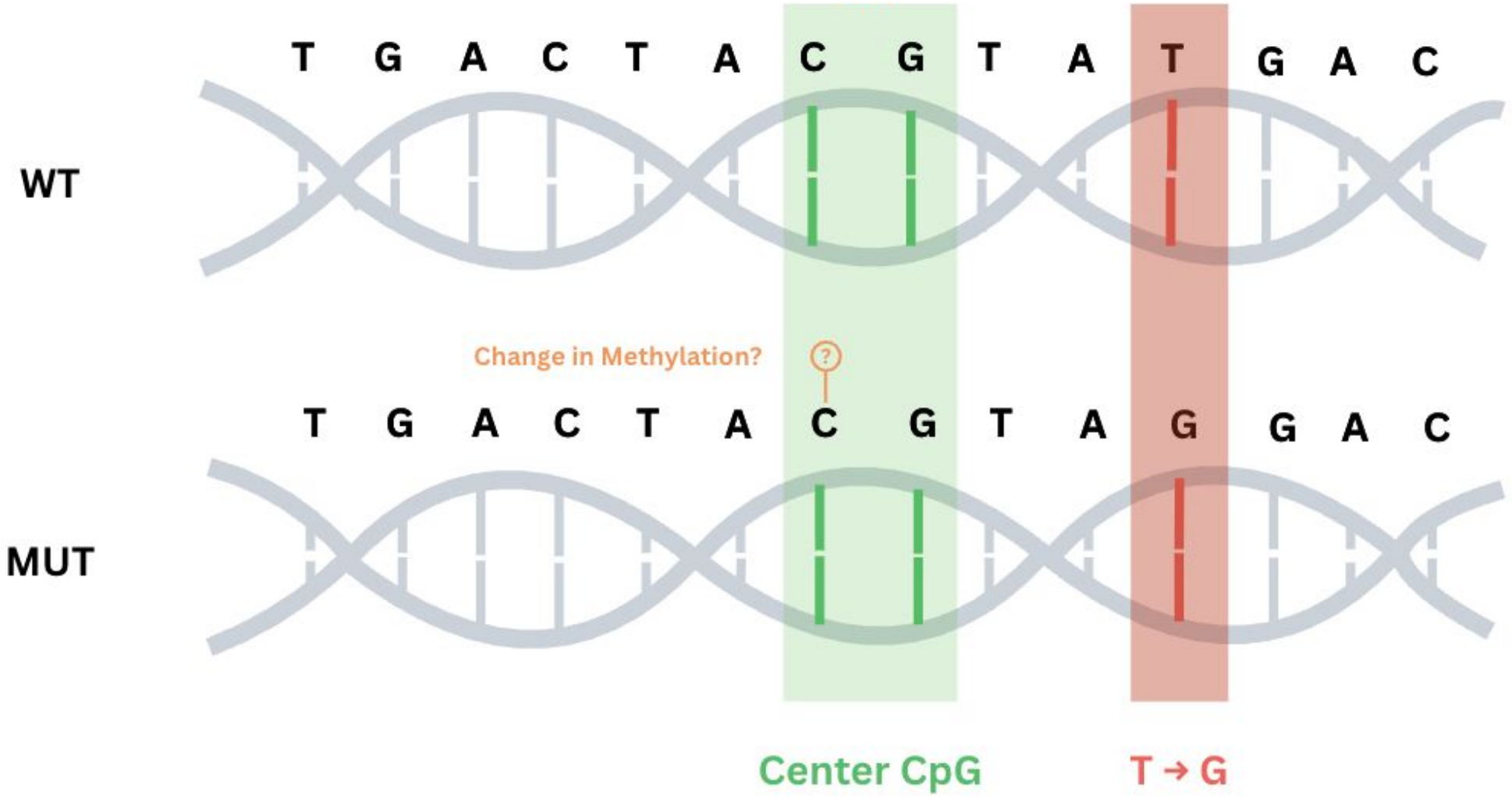
- Methylation levels describe how much of the measured DNA is methylated at a site.
- A β value ranges from 0 to 1.
- Different CpGs can have very different methylation levels.

Definition (β value)

$$\beta = \frac{\#\text{ methylated copies}}{\#\text{ copies measured}} \in [0, 1]$$



What changes when one DNA base changes?



Why is it difficult to measure variant effects on methylation?

- Experimentally testing large numbers of variants is difficult.
- Matched variant and methylation measurements are limited.
- Tissue and regulatory context also influence methylation.

Table of Contents

1 Introduction

2 Previous Works

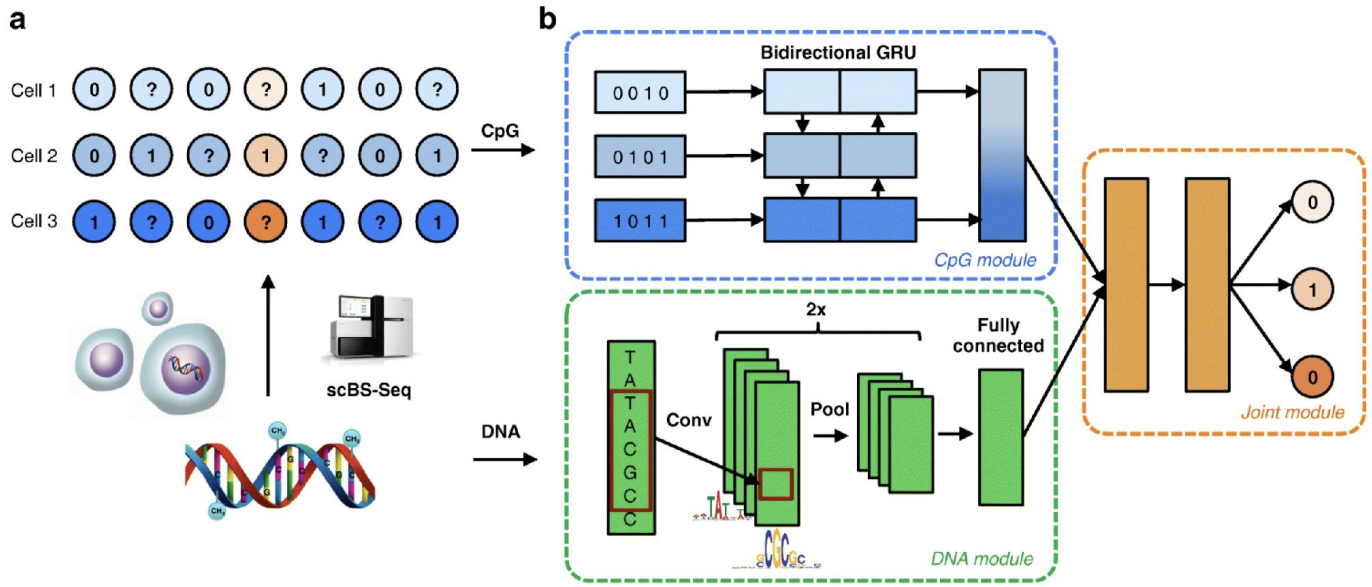
3 Methods

4 Results

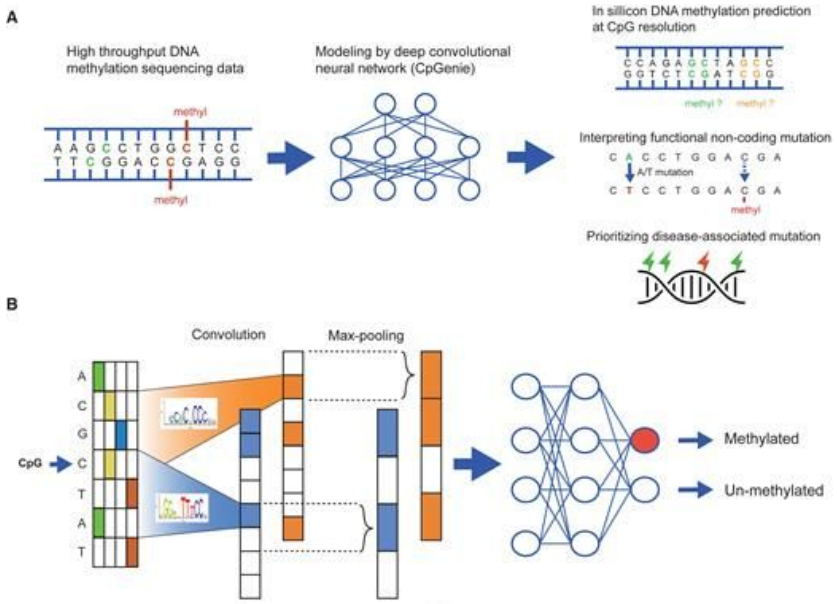
5 Conclusions

Existing models predict methylation from DNA

- **DeepCpG** uses DNA sequence and methylation at neighboring CpGs.
- **CpGenie** predicts methylation from DNA sequence and evaluates predicted variant effects.
- These methods show that local information can help predict methylation.



Angermueller et al. (2017)



Zeng et al. (2017)

Other methods include tissue/regulatory data

- **INTERACT** focuses on methylation prediction in brain tissue.
- **DeepMethylation** combines DNA sequence with reference tissue information.
- **Methven** combines DNA sequence with chromatin accessibility.

Table of Contents

1 Introduction

2 Previous Works

3 Methods

4 Results

5 Conclusions

Data Sources

- Methylation targets came from 97 solid tissue normal breast samples
- Each target was the median methylation at a CpG across samples
- Inputs included a 1,000-bp DNA sequence, reference chromatin measurements, and conservation scores.

Algorithm (target and input construction)

Require: $\beta_{i,s}$ for CpG i and sample s, O_i = samples with a measurement at CpG i

1: **for** each CpG i **do**

2: $\beta_i \leftarrow \text{median}_{s \in O_i} \beta_{i,s}, \qquad y_i \leftarrow \mathbf{1}[\beta_i > 0.5]$

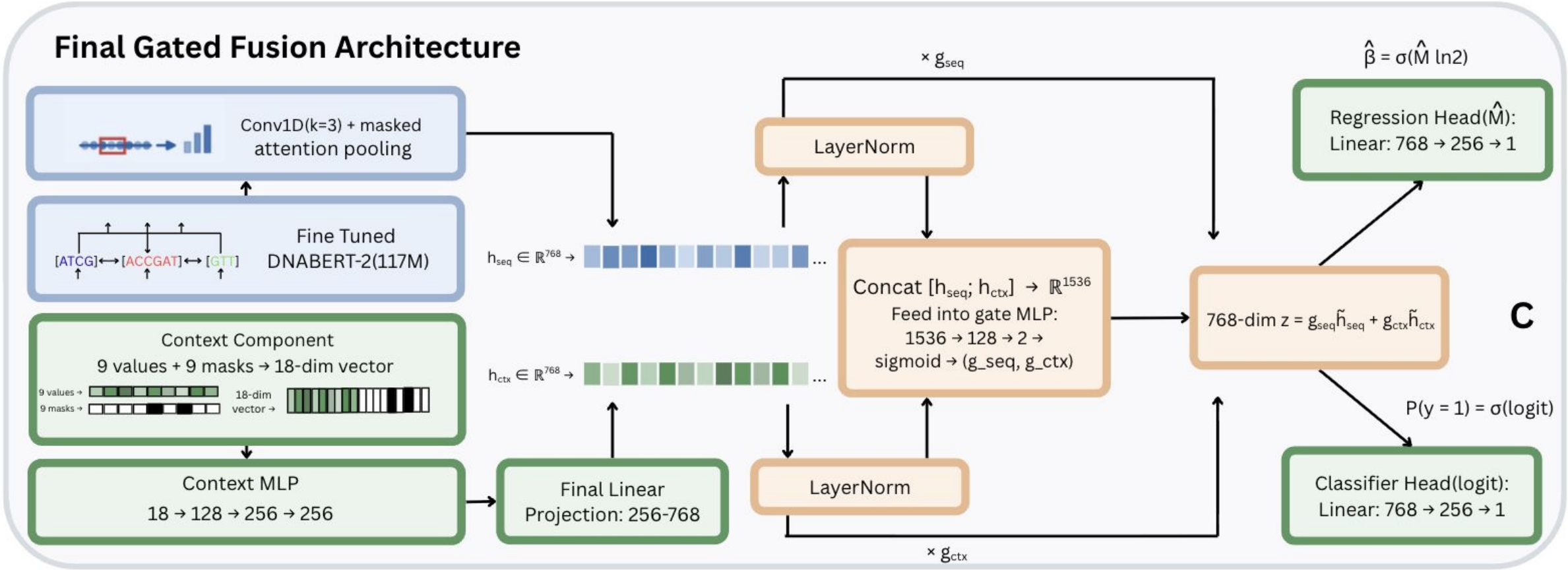
3: $\widetilde{\beta}_i \leftarrow \text{clip}(\beta_i, 10^{-4}, 1 - 10^{-4}), \qquad M_i \leftarrow \log_2\left(\frac{\widetilde{\beta}_i}{1 - \widetilde{\beta}_i}\right)$

4: $x_i \in \{\text{A}, \text{C}, \text{G}, \text{T}\}^{1000}, \qquad c_i \in \mathbb{R}^{18}$

5: **end for**

6: $D_{\text{test}} \leftarrow \{i : \text{chr}_i \in \{8, 9\}\}, \quad D_{\text{val}} \leftarrow \{i : \text{chr}_i \in \{10, 11\}\}, \quad D_{\text{train}} \leftarrow \{i : \text{chr}_i \notin \{8, 9, 10, 11\}\}$

Model Architecture



Evaluating baseline methylation prediction

- Training, validation, and testing used different chromosomes(test chr8-9, val chr10-11).
- The main test set contained 26,570 CpGs.
- We compared fusion with sequence-only, context-only, CpGenie, and DeepCpG models.
- We repeated testing across training seeds and three additional chromosome splits.

Algorithm (held-out evaluation)

Require: models $f^{(k)}$, $k \in \{\text{fusion, sequence, CpGenie, DeepCpG}\}$

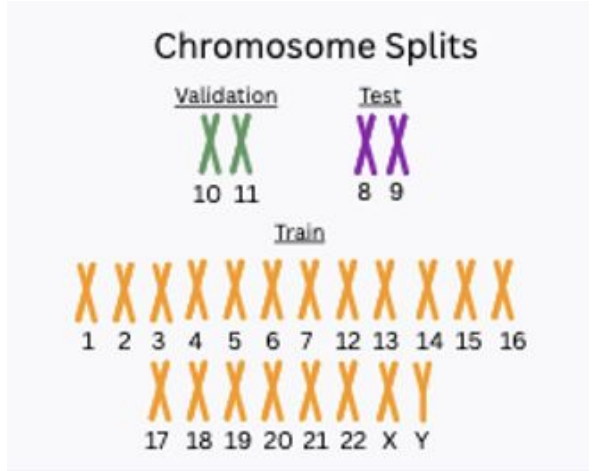
1: **for** each k **do**

2: $\quad \widehat{\beta}_i \leftarrow f^{(k)}(x_i, c_i), \quad i \in D_{\text{test}}$

3: $\quad \text{MAE}_\beta \leftarrow \frac{1}{|D_{\text{test}}|} \sum_i |\widehat{\beta}_i - \beta_i|, \qquad \text{RMSE}_\beta \leftarrow \sqrt{\frac{1}{|D_{\text{test}}|} \sum_i (\widehat{\beta}_i - \beta_i)^2}$

4: **end for**

5: repeat with test chromosomes $\{12, 18\}$, $\{4, 20\}$, $\{13, 15, 21\}$



Estimating and ranking variant effects

Algorithm (paired variant scoring)

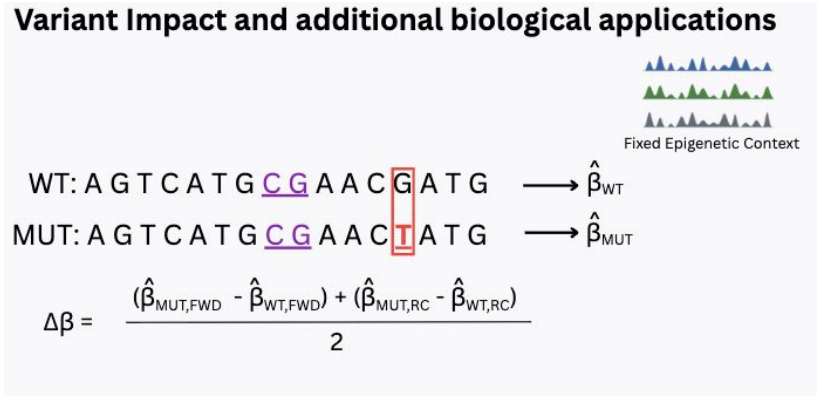
Require: fusion model f, sequences $x^{\text{REF}}, x^{\text{ALT}}$, context c

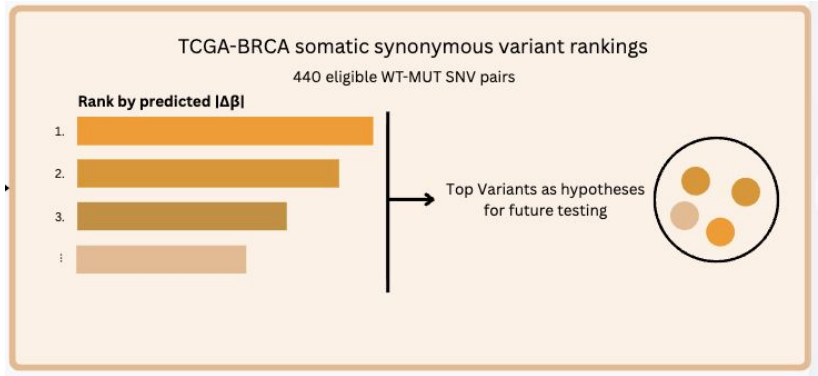
1: $\Delta_o \leftarrow f(x_o^{\text{ALT}}, c) - f(x_o^{\text{REF}}, c), \quad o \in \{\text{FWD}, \text{RC}\}$

2: $\widehat{\Delta\beta} \leftarrow \dfrac{\Delta_{\text{FWD}} + \Delta_{\text{RC}}}{2}$

3: **return** variants sorted by $|\widehat{\Delta\beta}|$

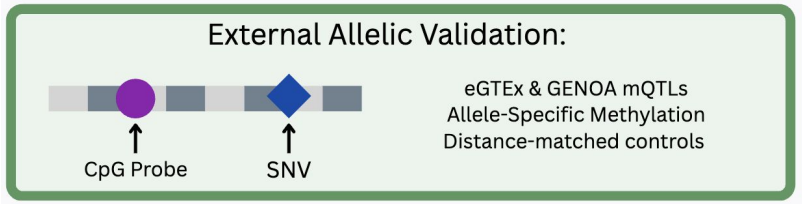
Additional application: We ranked 440 synonymous variants by predicted change magnitude and also examined STK11 variants.





Evaluating variant effects using external data

- **mQTLs:** Variants linked to methylation differences across individuals (eGTEx, GENOA)
- **Allele-specific methylation:** Methylation differences between alleles within a sample (Do-Tycko, Rosenski-Dor-Kaplan).
- We tested whether predicted directions and relative rankings matched independent measurements.



Algorithm (external concordance)

Require: pairs $i = 1, \dots, n$ with $\Delta\widehat{M}_i$ and $\Delta_{\text{obs},i}$, significant pairs P_{sig}, control pairs P_{ctrl}

1: $\rho_{\text{signed}} \leftarrow \rho\big(\Delta\widehat{M},\ \Delta_{\text{obs}}\big)$

2: $\text{Dir} \leftarrow \frac{1}{n}\sum_i \mathbf{1}\big[\text{sign}(\Delta\widehat{M}_i) = \text{sign}(\Delta_{\text{obs},i})\big]$

3: $\text{AUROC}(s) \leftarrow \frac{1}{|P_{\text{sig}}|\,|P_{\text{ctrl}}|}\sum_{i\in P_{\text{sig}}}\sum_{j\in P_{\text{ctrl}}}\left(\mathbf{1}[s_i > s_j] + \tfrac{1}{2}\mathbf{1}[s_i = s_j]\right)$

Table of Contents

1 Introduction

2 Previous Works

3 Methods

4 Results

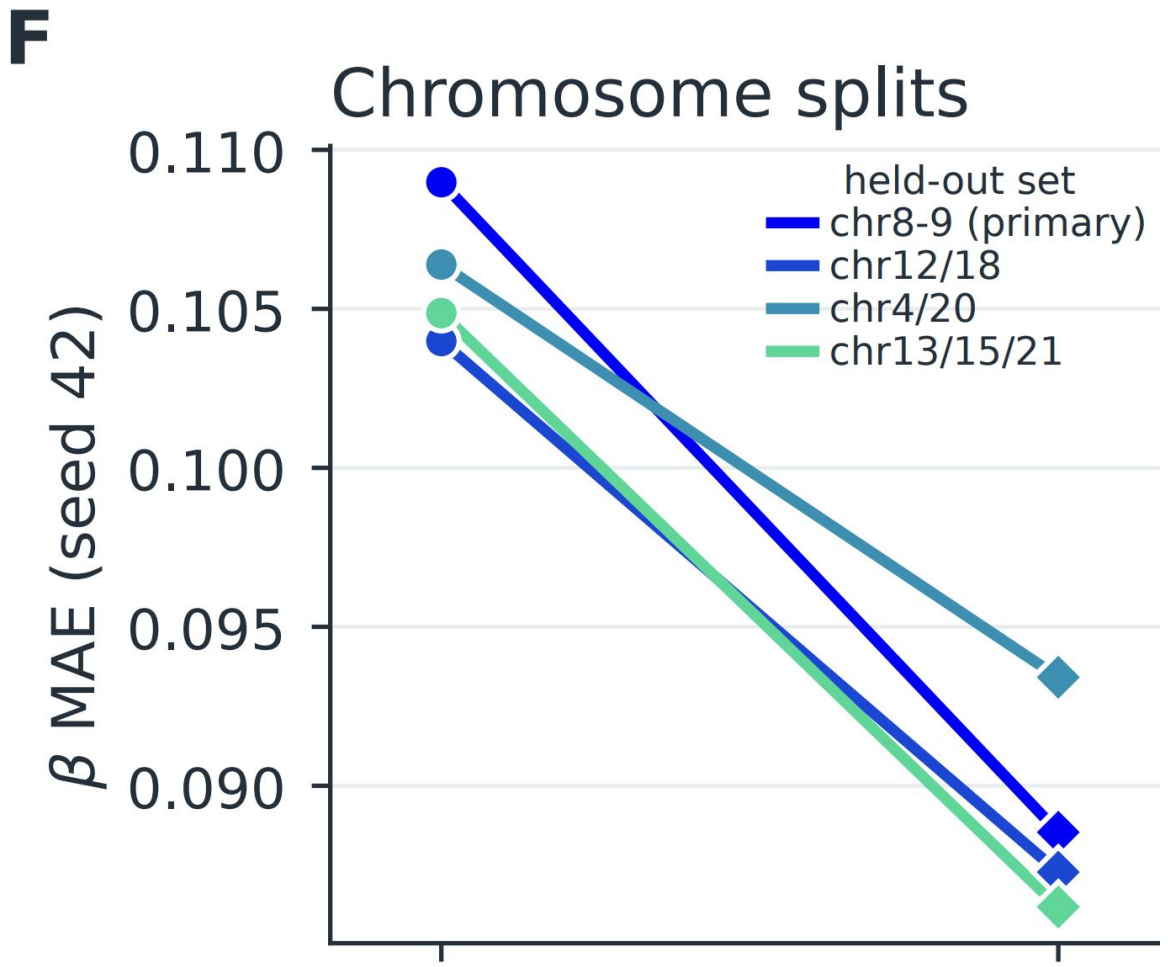
5 Conclusions

Fusion improved baseline methylation prediction

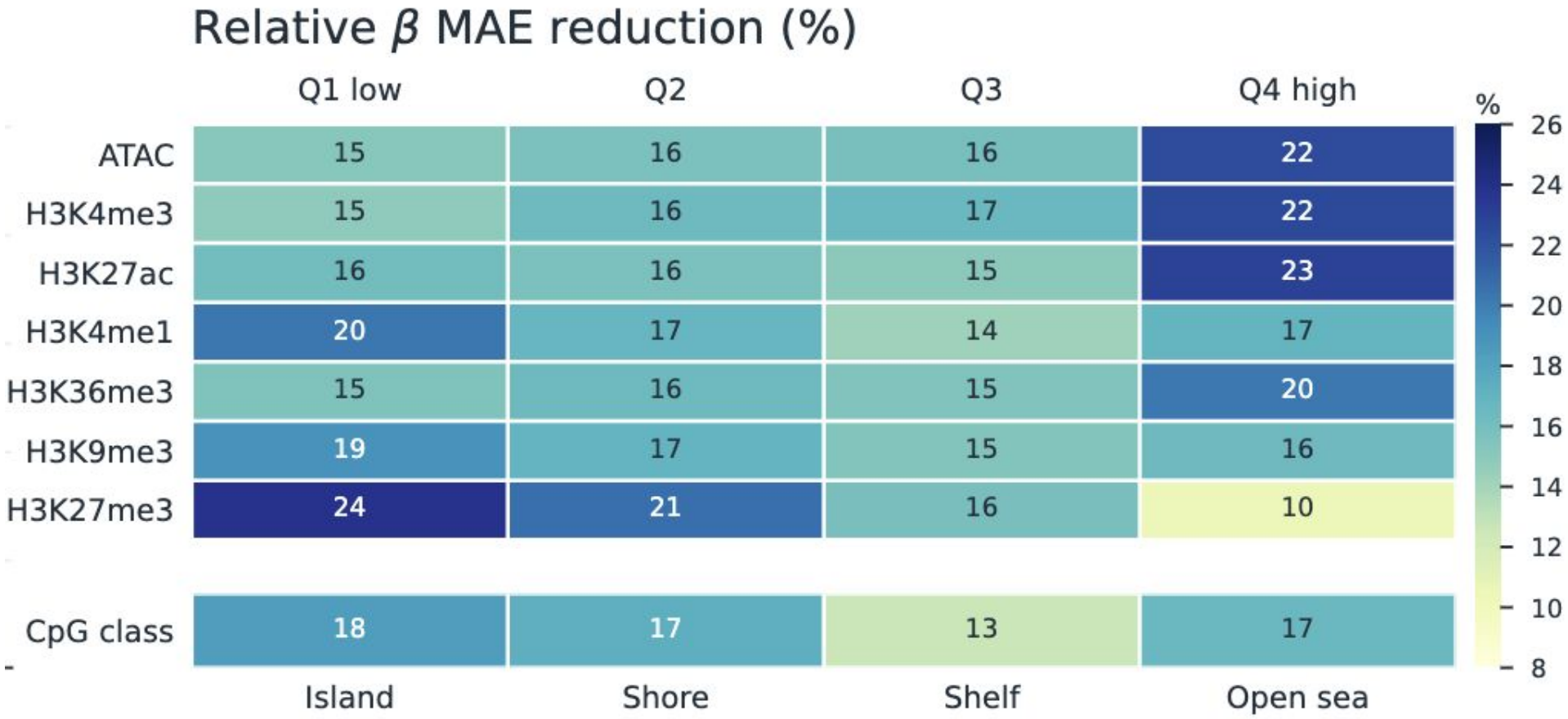
Model	β MAE	M MAE	ROC-AUC
Composition ridge	0.1954	1.9600	0.8748
k-mer ridge	0.1565	1.6866	0.9190
CpGenie	0.1281 ± 0.0013	1.3858 ± 0.0103	0.9406 ± 0.0006
DeepCpG	0.1253 ± 0.0005	1.3412 ± 0.0112	0.9437 ± 0.0011
Context	0.1022 ± 0.00003	1.1763 ± 0.0005	0.9658 ± 0.00001
Sequence	0.1099 ± 0.0008	1.1941 ± 0.0070	0.9569 ± 0.0007
Fusion	0.0914 ± 0.0037	1.0341 ± 0.0289	0.9744 ± 0.0025

Consistent improvement across chromosome splits

- Fusion outperformed sequence alone in all three additional chromosome splits.
- These tests checked whether the result depended on one particular data split.



The effectiveness of context varies

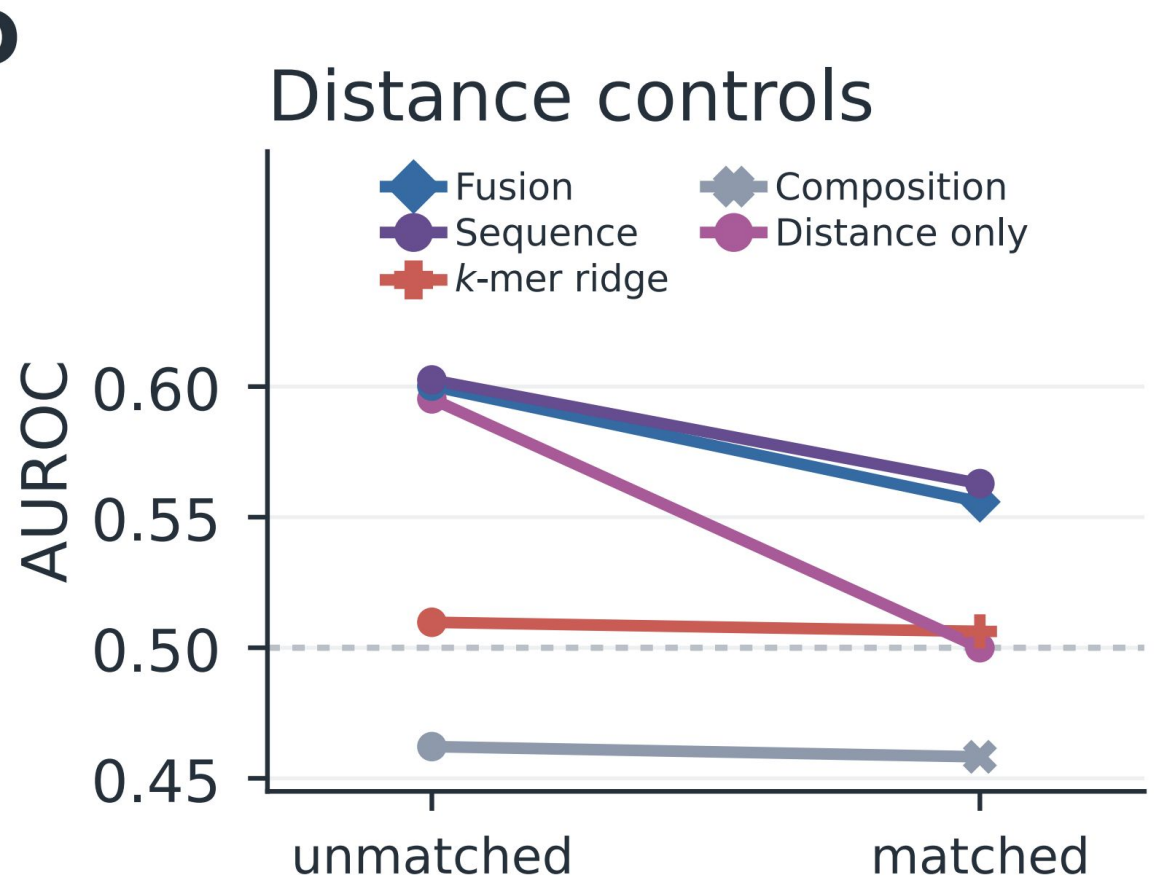


External data show modest consistency with predicted effects

External dataset	Comparisons	Signed Spearman ρ (95% CI)	Correct direction (95% CI)
eGTEx breast mQTLs	418 pairs	0.236 (0.098–0.392)	59.6% (54.2–65.1%)
GENOA blood mQTLs	4,037 pairs	0.149 (0.112–0.185)	55.6% (53.8–57.2%)
Do–Tycko ASM	722 SNPs	0.241 (0.180–0.299)	59.0% (55.0–63.0%)

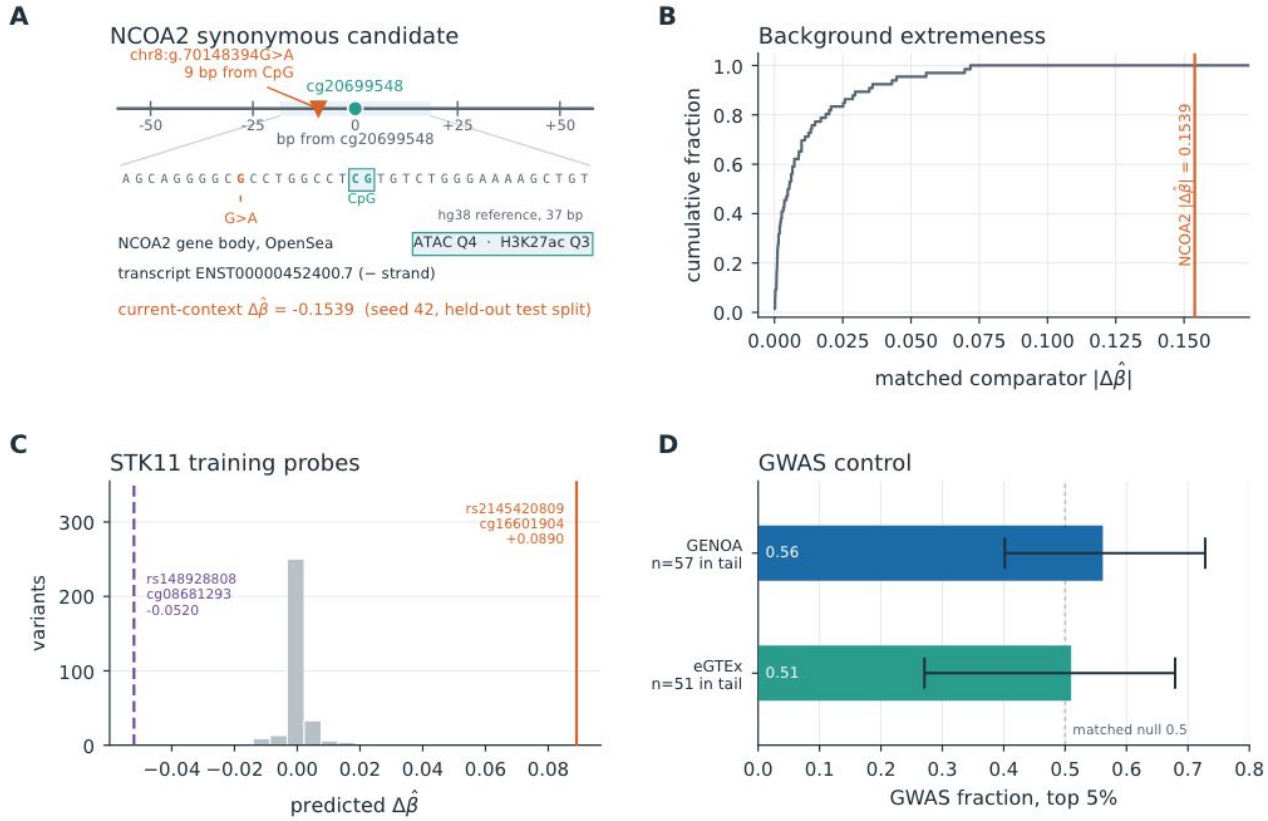
Distance explains parts of the apparent performance

- Significant mQTLs tended to be closer to their CpGs than controls.
- Distance alone performed almost as well as fusion before matching.
- After distance matching, fusion AUROC fell from 0.600 to 0.556, but remained above chance.



Prioritizing variants for experimental testing

- **NCOA2:** Top synonymous candidate, beating out 66 matched variants.
- **STK11:** The model prioritizes a pathogenic coding variant (rs148928808).
- **GWAS:** The model did not noticeably favor GWAS-annotated variants.



Limitations

- Training data came from “solid tissue normal” breast tissue collected adjacent to tumors.
- Reference context was not matched to individual samples.
- Predictions cover local sequence windows and a limited set of CpGs.
- External validation does not establish causal effects of individual variants.

Table of Contents

1 Introduction

2 Previous Works

3 Methods

4 Results

5 Conclusions

Main Findings

- Combining sequence and reference context improved baseline methylation prediction.
- External data showed modest consistency with predicted variant effects.
- SilentMethyl can help prioritize variants for experimental testing.

Future Work

- Test the model in independent breast cohorts and additional tissues.
- Expand predictions to more CpGs and better-matched reference data.
- Experimentally test prioritized variants and their effects on methylation and gene regulation.

Acknowledgements

I would like to thank the MIT PRIMES Program and the PRIMES bioinformatics mentors for their guidance and support throughout this project:

- Dr. Gil Alterovitz
- Dr. Shaojun Pei
- Dr. Ning Xie

End of Presentation

THANK YOU!