LIBD

JOHNS HOPKINS UNIVERSITY

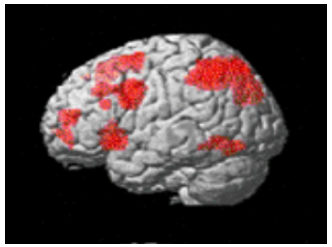
G-MIND: An End-to-End Multimodal Imaging-Genetics Framework for Biomarker Identification and Disease Classification

Sayan Ghosal, Qiang Chen, Giulio Pergola, Aaron L. Goldman, William Ulrich, Karen F. Berman, Giuseppe Blasi, Leonardo Fazio, Antonio Rampino, Alessandro Bertolino, Daniel R. Weinberger, Venkata S. Mattay, and Archana Venkataraman

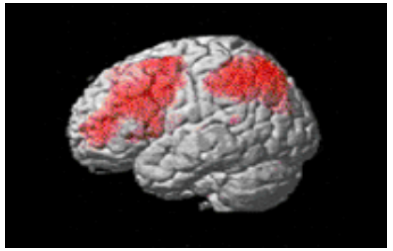
Multiview Representation of Schizophrenia

- Neuropsychiatric diseases are mainly characterized by atypical neural functioning (cognitive dysfunction, hallucination, etc.)

Schizophrenia patients

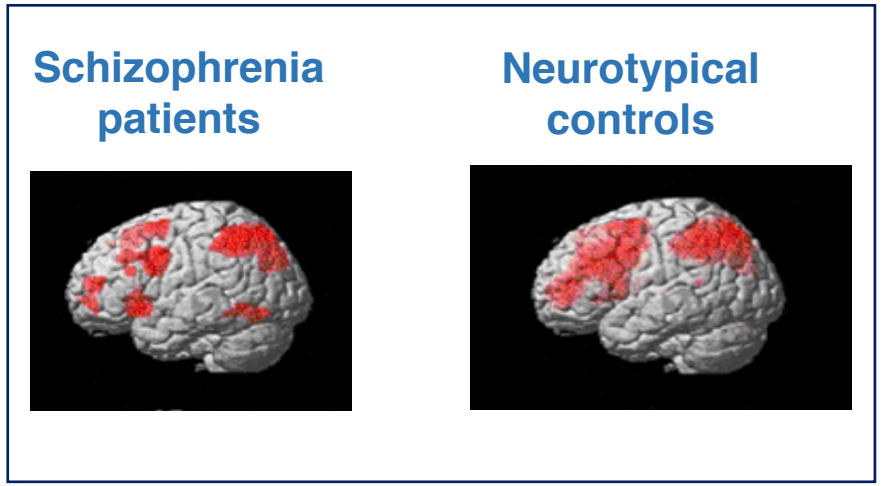


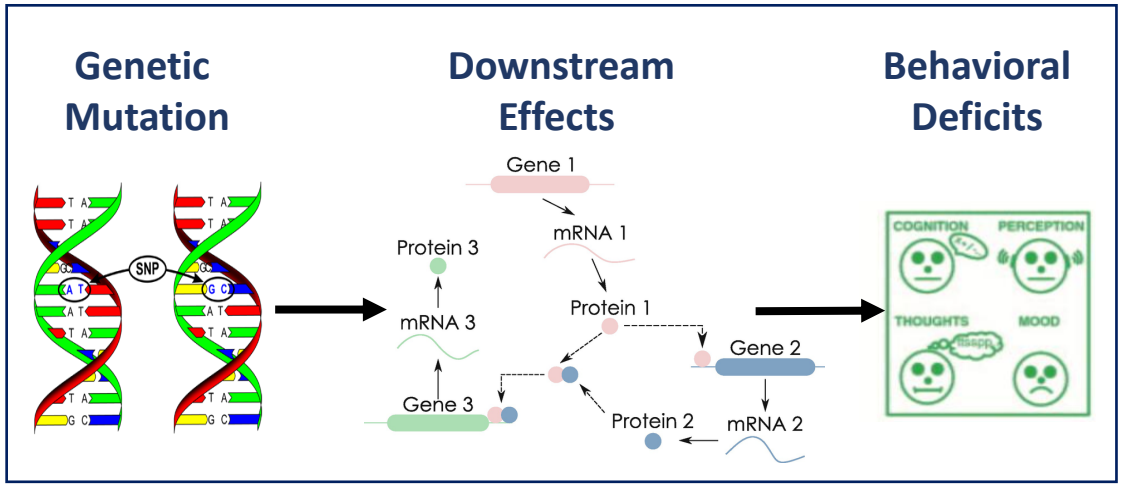
Neurotypical controls



Multiview Representation of Schizophrenia

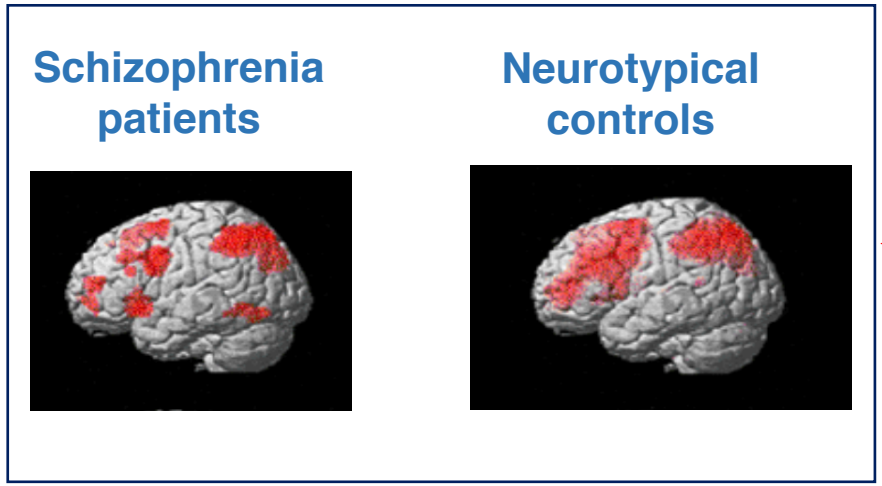
- Neuropsychiatric diseases are mainly characterized by atypical neural functioning (cognitive dysfunction, hallucination, etc.)
- Schizophrenia is highly heritable disease which means it has strong genetic underpinning.

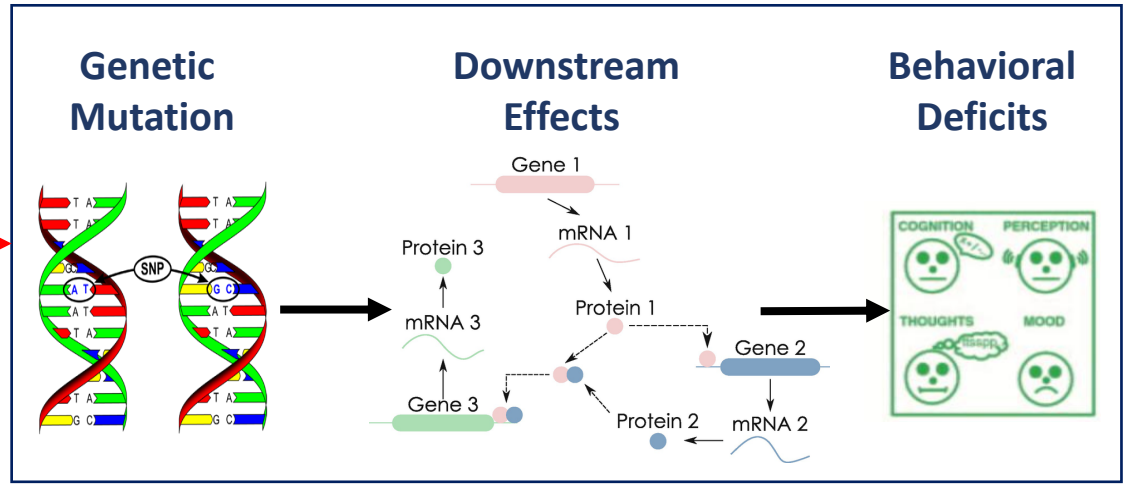




Multiview Representation of Schizophrenia

- Neuropsychiatric diseases are mainly characterized by atypical neural functioning (cognitive dysfunction, hallucination, etc.)
- Schizophrenia is highly heritable disease which means it has strong genetic underpinning.





Outline

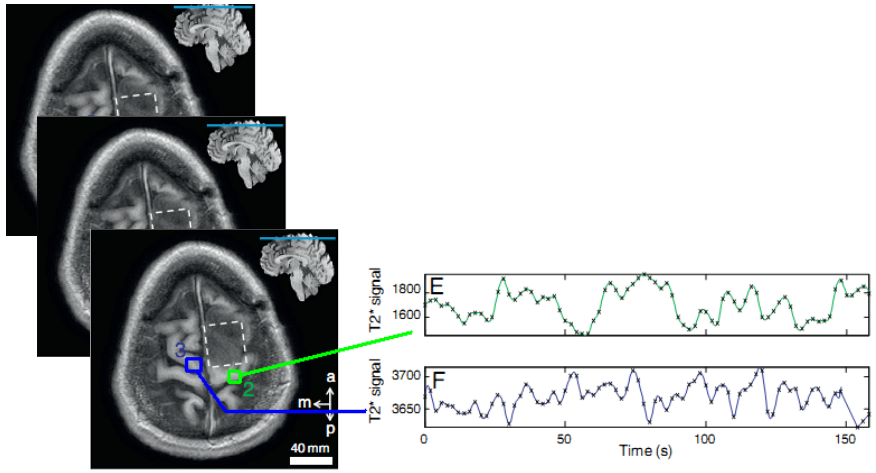
- Background
- GMIND: An End-to-End Model for Imaging-Genetics
- Experimental Results.
- Contributions

Tapping into the Brain

- **Functional Magnetic Resonance Imaging (fMRI)**

  → Captures different snapshots of brain activations across time as a BOLD signal

  → The experimental paradigm depends on the psychological question.

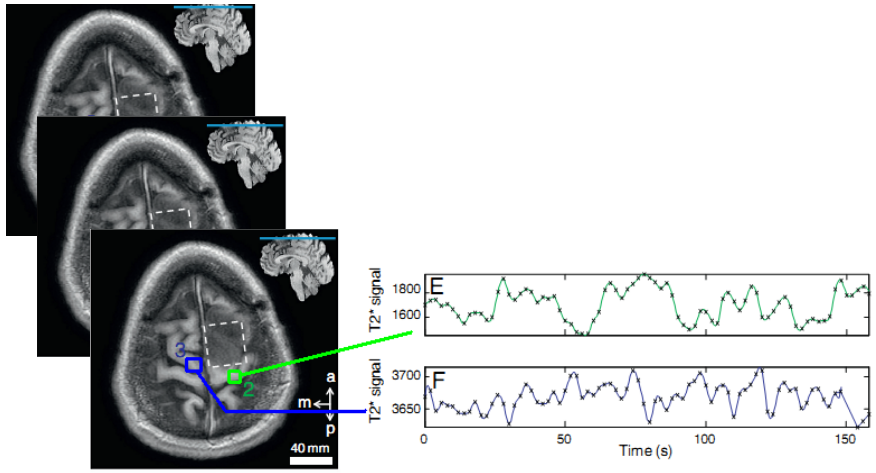


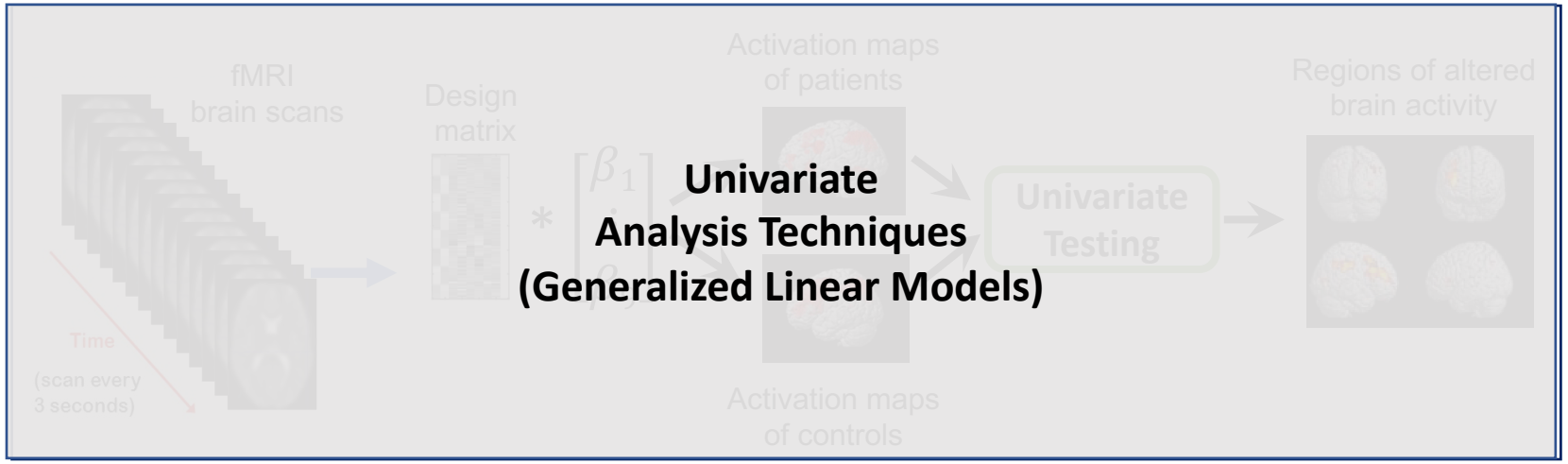
Tapping into the Brain

- **Functional Magnetic Resonance Imaging (fMRI)**

  → Captures different snapshots of brain activations across time as a BOLD signal

  → The experimental paradigm depends on the psychological question.



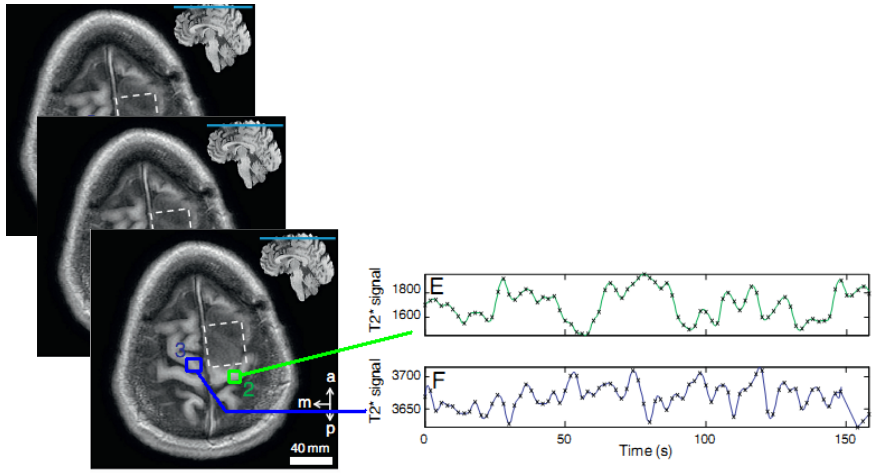


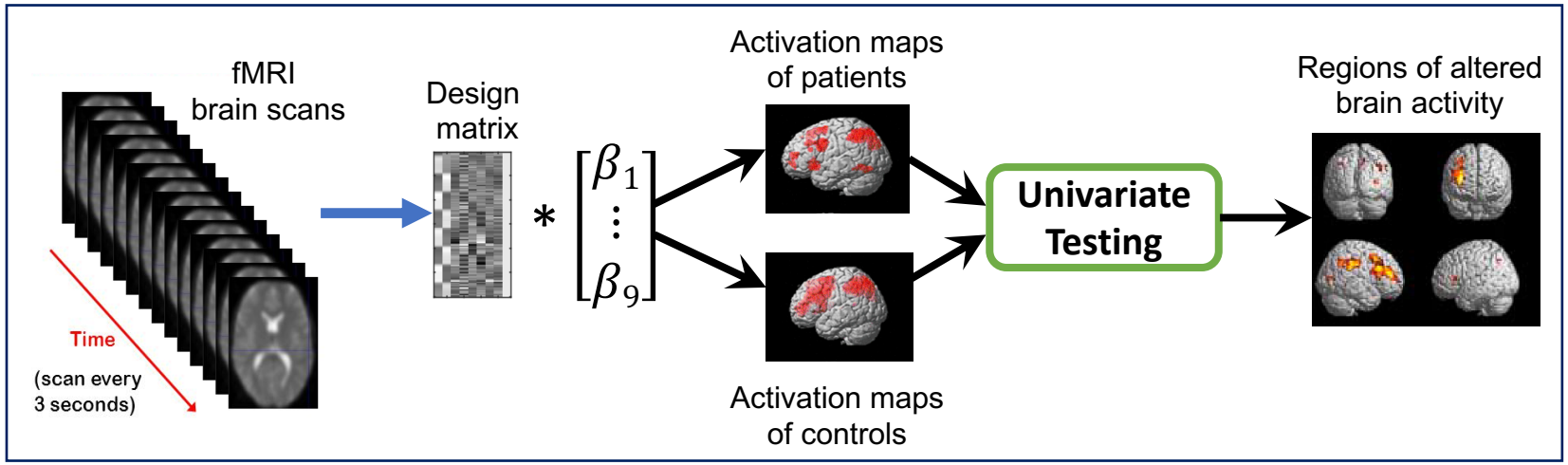
Tapping into the Brain

- **Functional Magnetic Resonance Imaging (fMRI)**

  → Captures different snapshots of brain activations across time as a BOLD signal

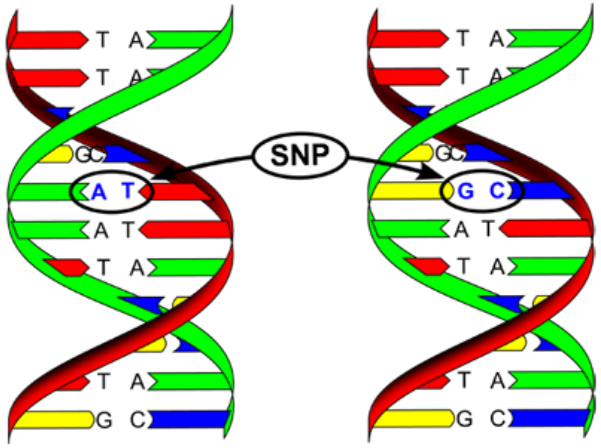
  → The experimental paradigm depends on the psychological question.





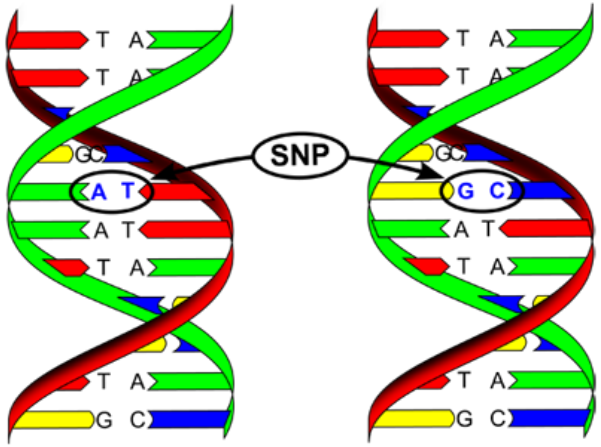
Genetic Variations

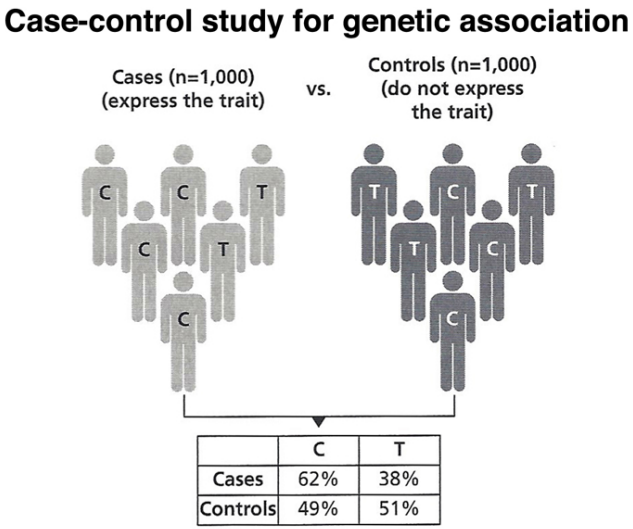
- **Single Nucleotide Polymorphism (SNP)**
  - → Captures variations of alleles in the DNA



Genetic Variations

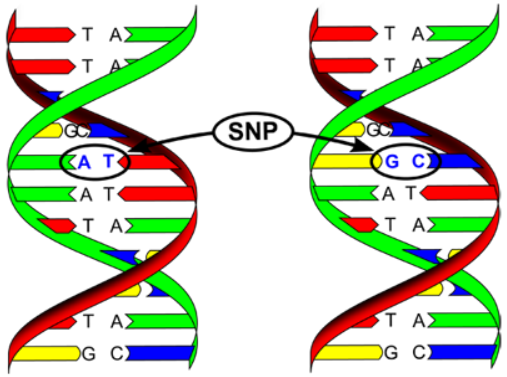
- **Single Nucleotide Polymorphism (SNP)**
  - → Captures variations of alleles in the DNA
- **The allelic configuration of SNPs provides information about the associated phenotype in a certain population**

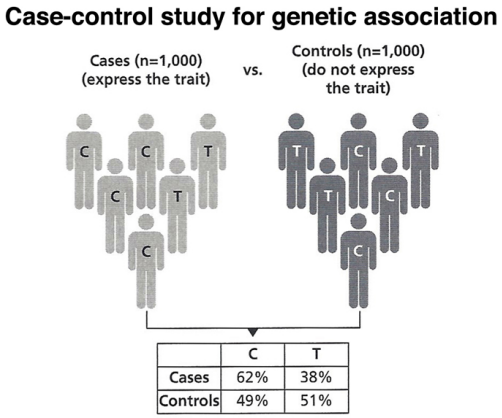


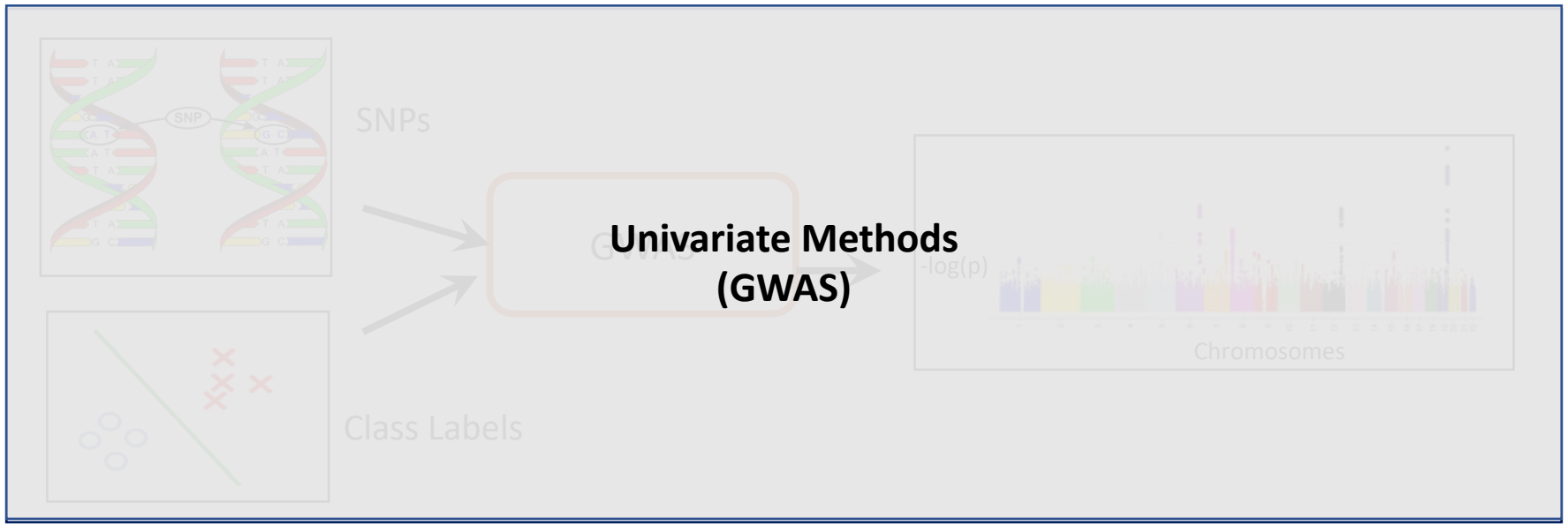


Genetic Variations

- **Single Nucleotide Polymorphism (SNP)**
  - → Captures variations of alleles in the DNA
- **The allelic configuration of SNPs provides information about the associated phenotype in a certain population**

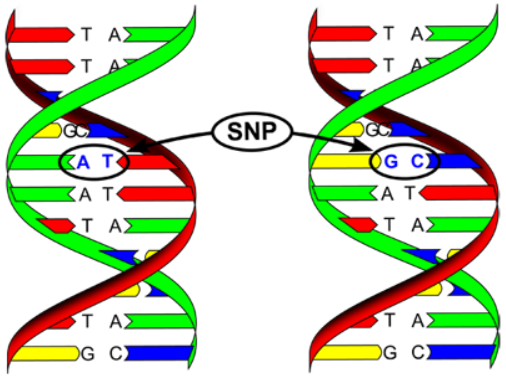


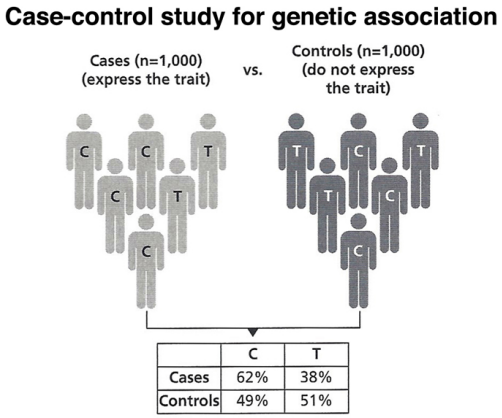


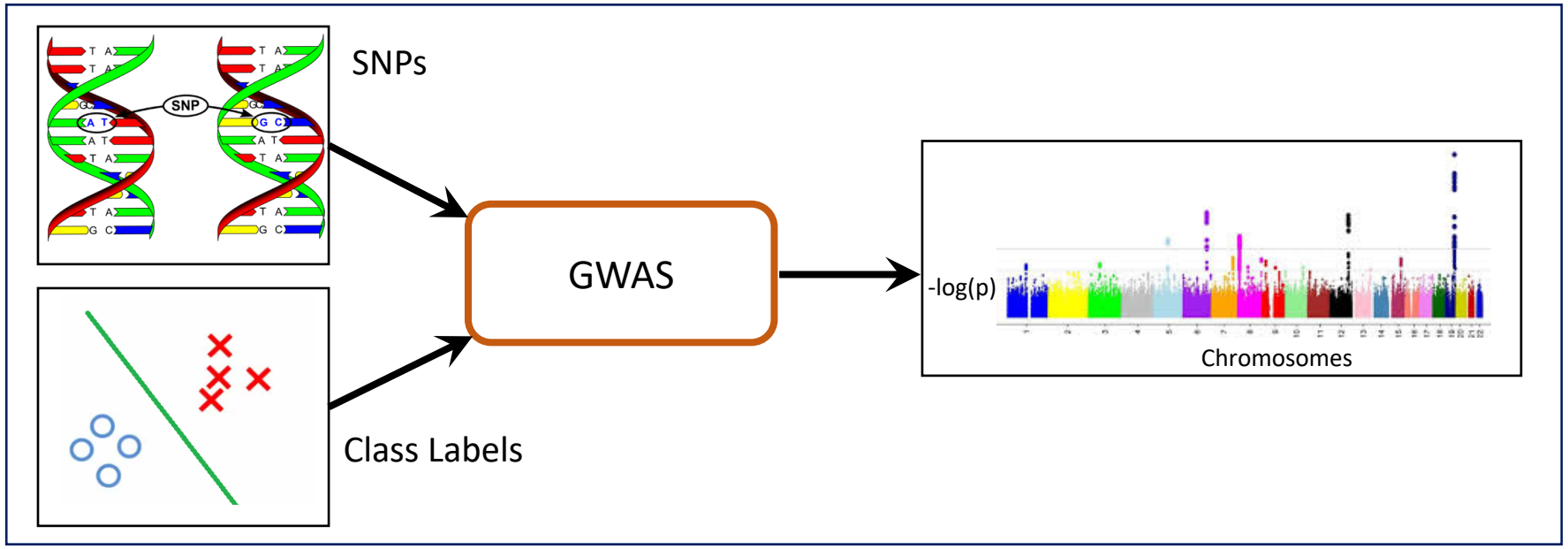


Genetic Variations

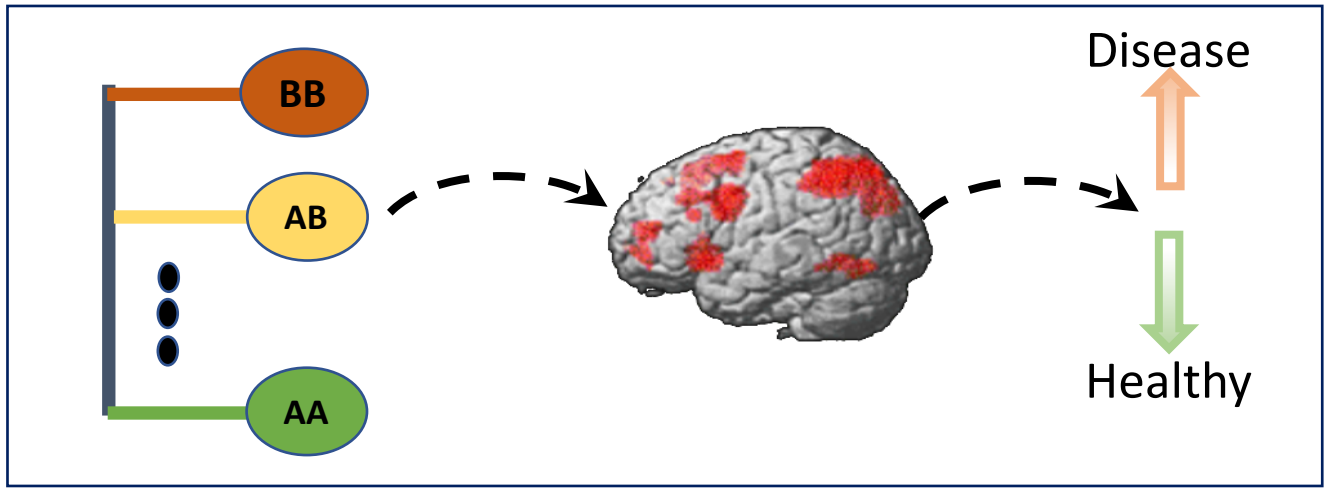
- **Single Nucleotide Polymorphism (SNP)**
  - → Captures variations of alleles in the DNA
- **The allelic configuration of SNPs provides information about the associated phenotype in a certain population**







Prior Works in Imaging-Genetics



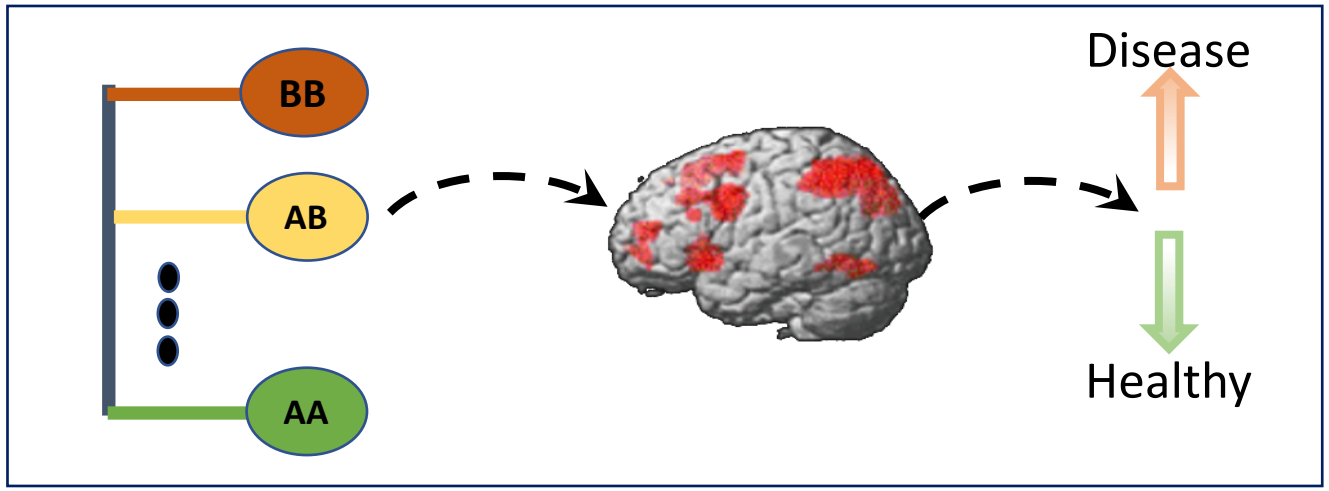
Cascaded Regression:

→ One-to-one mapping from gene to disease

→ Single modality is used

*H. Wang, et al., Bioinformatics (2012).

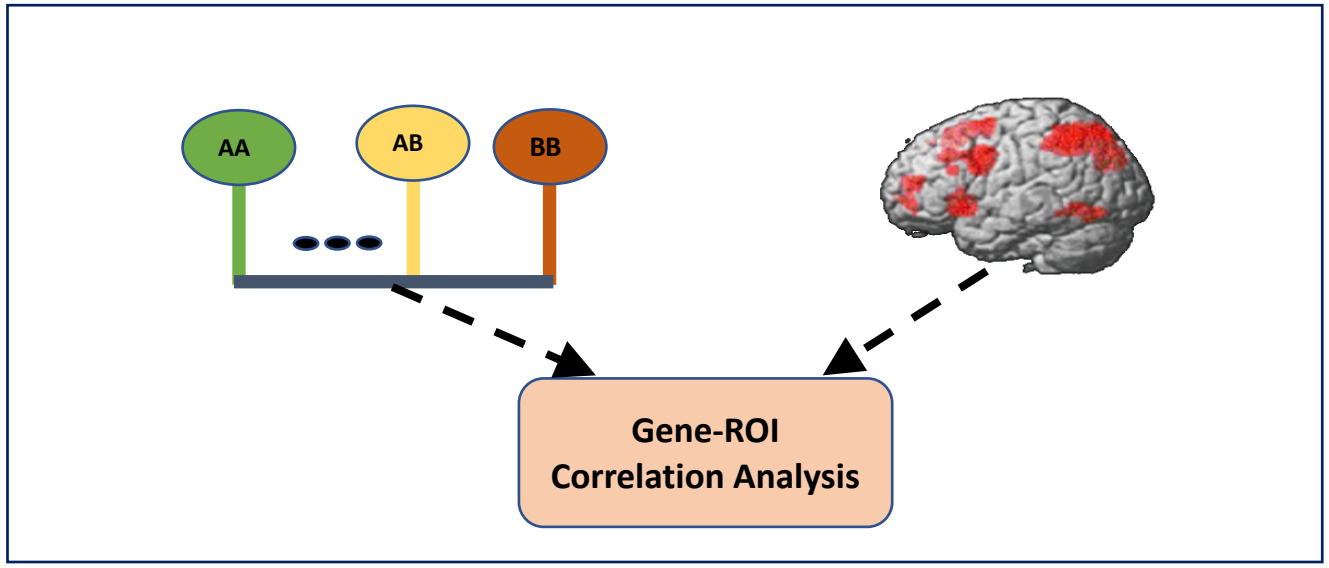
Prior Works in Imaging-Genetics



Cascaded Regression:

→ One-to-one mapping from gene to disease

→ Single modality is used



Correlation Based Analysis:

→ Maximizing the correlation between two modalities

→ Paired data is required

→ Does not incorporate patient heterogeneity.

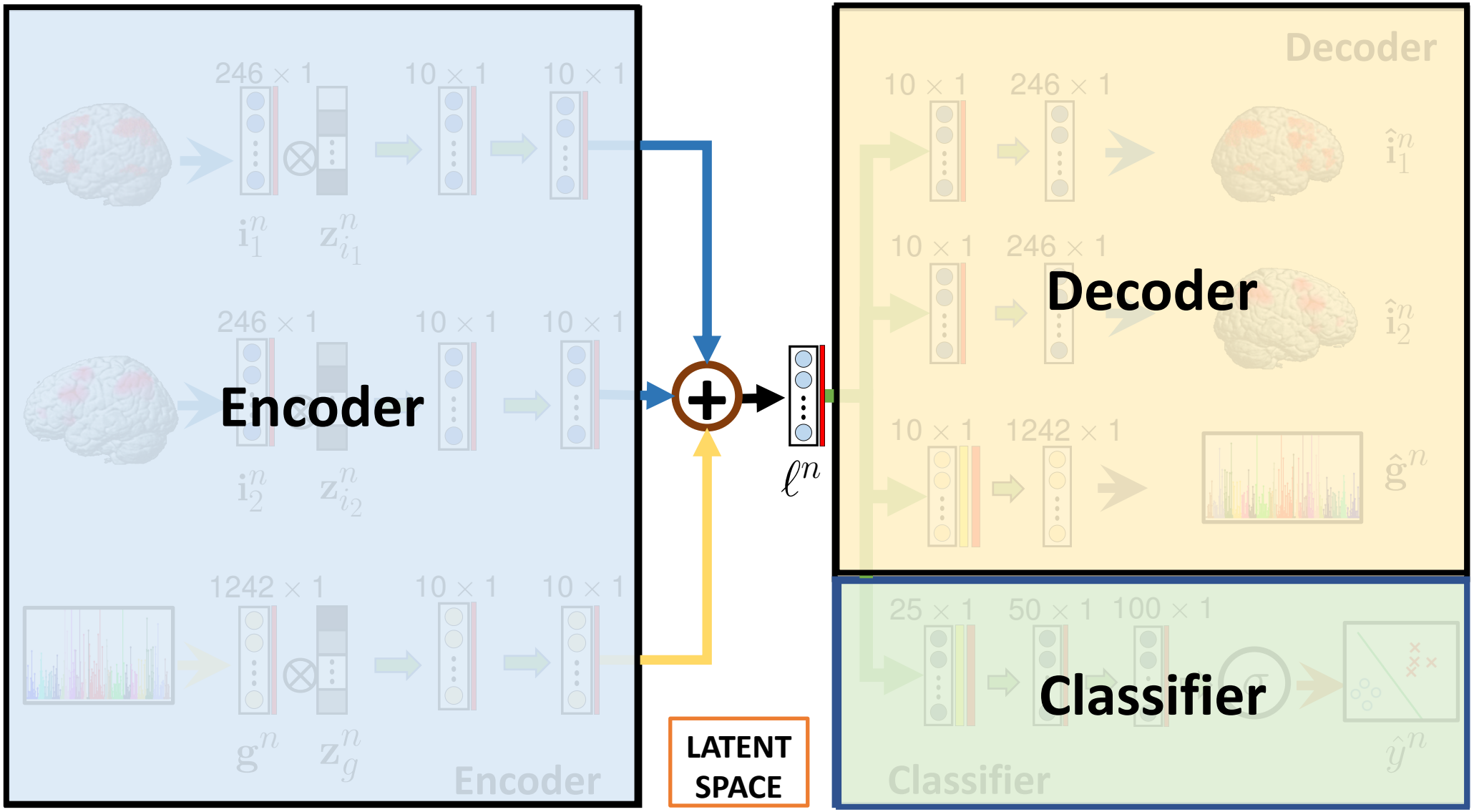
*H. Wang, et al., Bioinformatics (2012).

* Pearlson GD, Liu J, et al., *Front Genet*. 2015

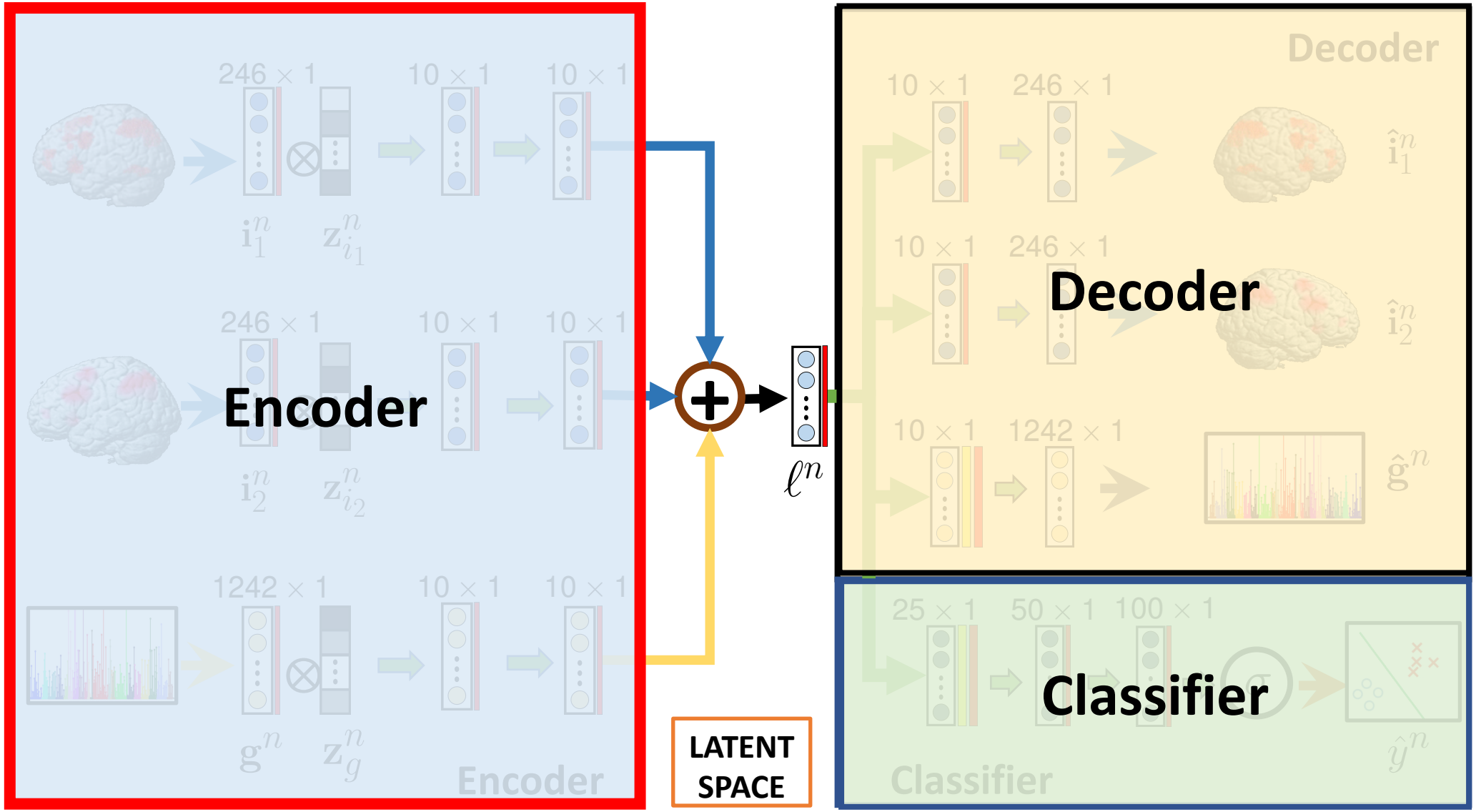
Outline

- Background
- GMIND: An End-to-End Model for Imaging-Genetics
- Experimental Results.
- Contributions

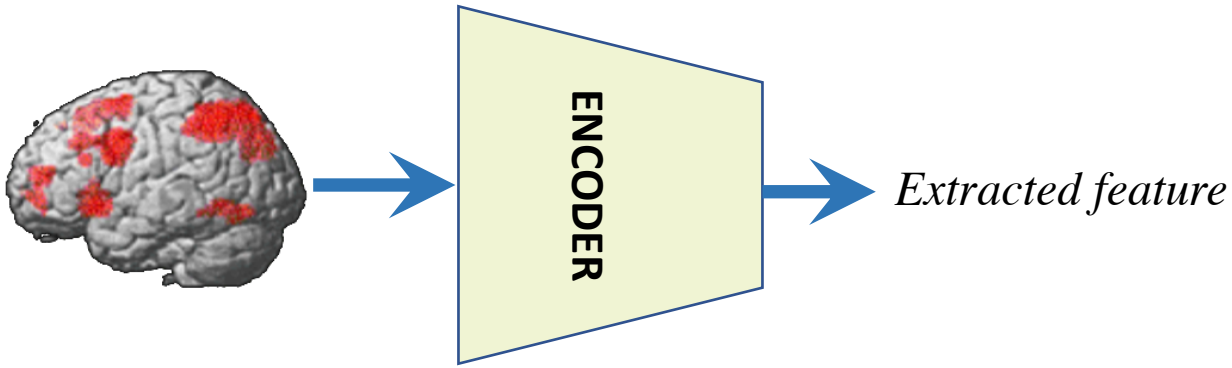
GMIND: An End-to-End Multimodal Framework



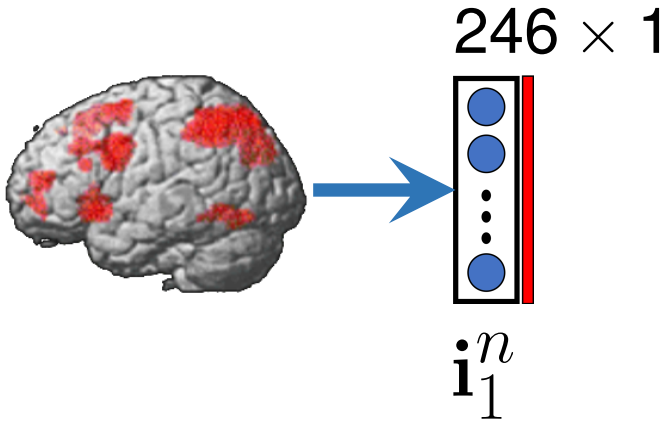
GMIND: End-to-End Multimodal Framework



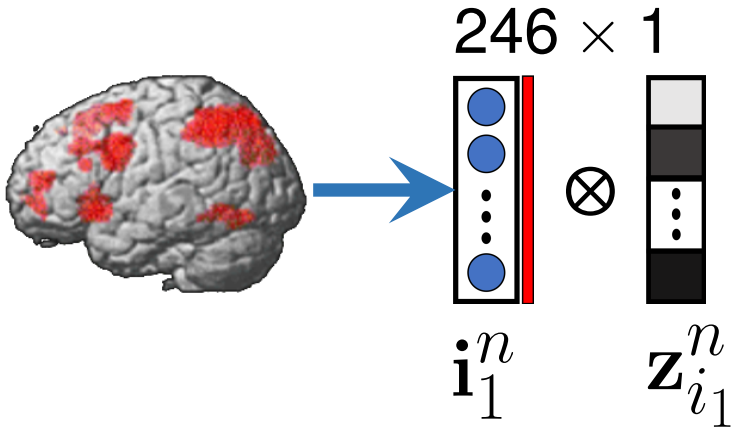
Feature Importance via Concrete Dropout

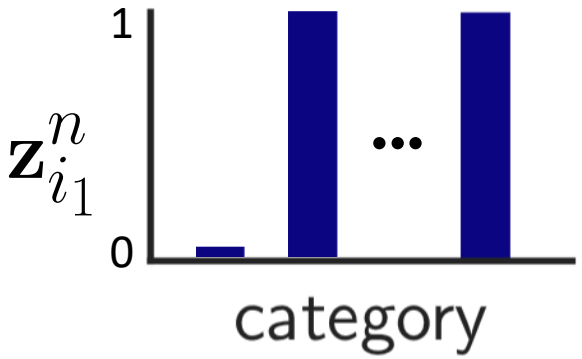


Feature Importance via Concrete Dropout

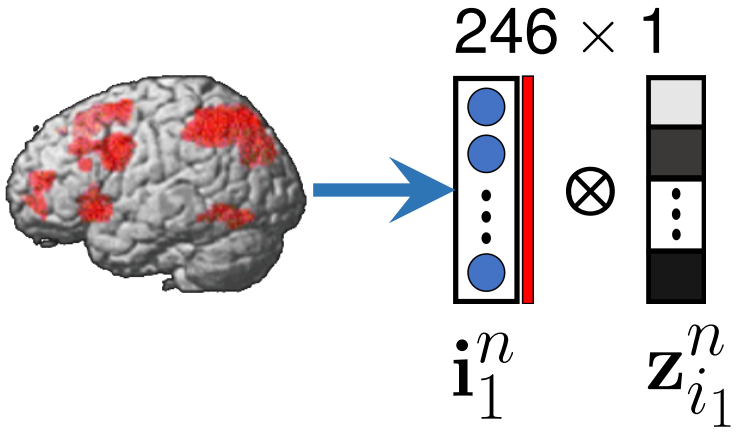


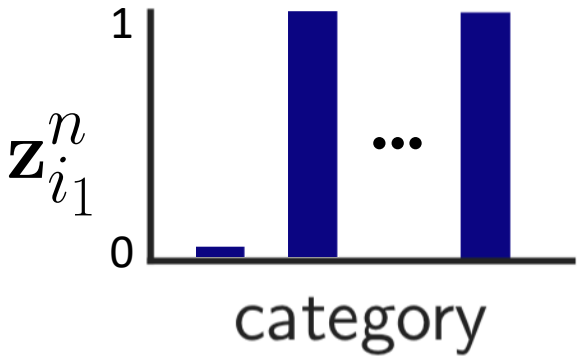
Feature Importance via Concrete Dropout





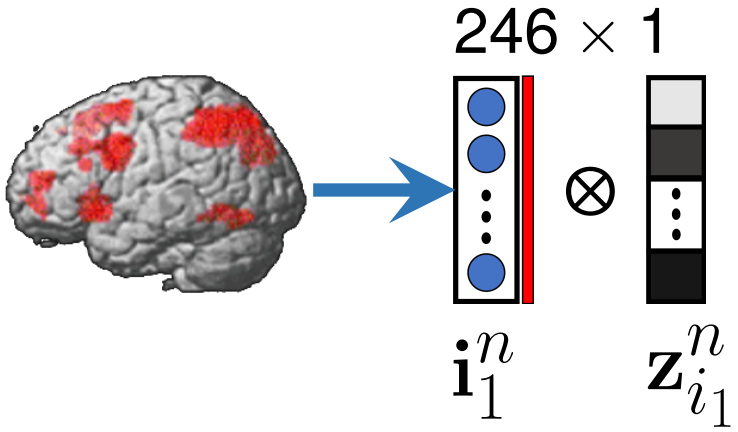
Feature Importance via Concrete Dropout

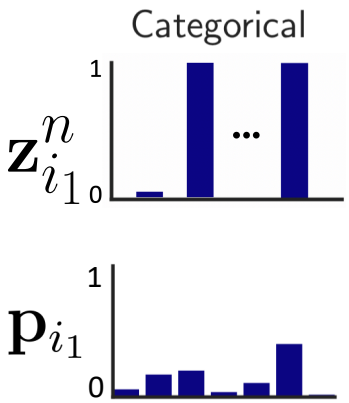




$$\mathbf{z}_{i_1}^n = \sigma\left(\frac{\log(\mathbf{p}_{i_1}) - \log(1-\mathbf{p}_{i_1}) + \log(\mathbf{u}_{i_1}^n) - \log(1-\mathbf{u}_{i_1}^n)}{t}\right)$$

Feature Importance via Concrete Dropout

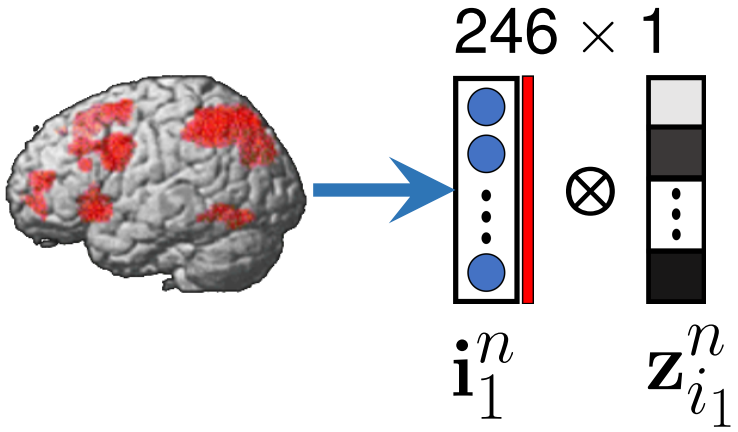


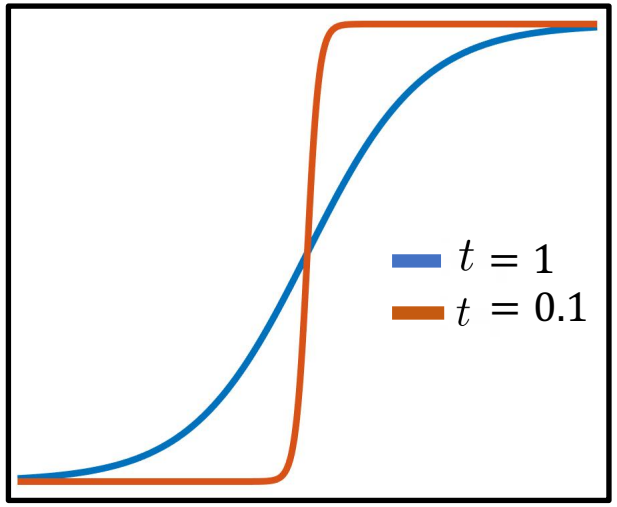


Probability of Retention **Uniform Random Variable**

$$\mathbf{z}_{i_1}^n = \sigma\left(\frac{\log(\mathbf{p}_{i_1}) - \log(1 - \boxed{\mathbf{p}_{i_1}}) + \log(\mathbf{u}_{i_1}^n) - \log(1 - \boxed{\mathbf{u}_{i_1}^n})}{t}\right)$$

Feature Importance via Concrete Dropout



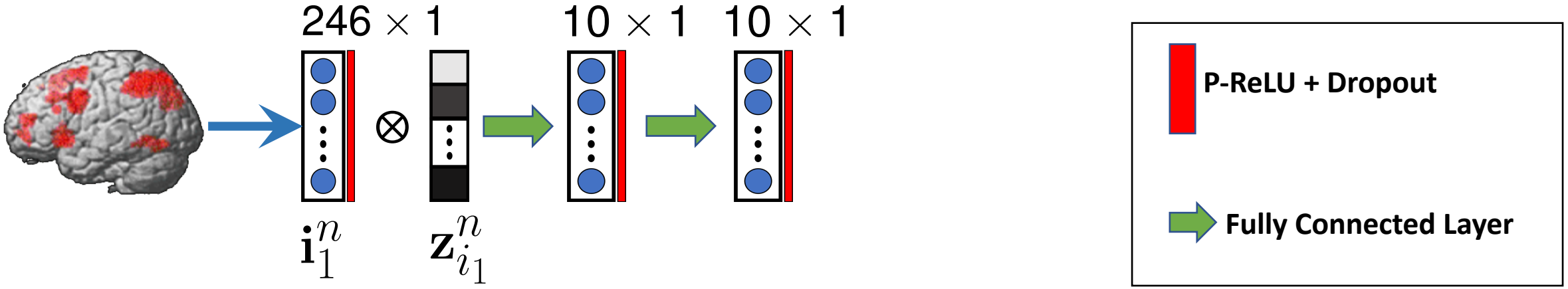


Probability of Retention **Uniform Random Variable**

$$\mathbf{z}_{i_1}^n = \sigma\left(\frac{\log(\mathbf{p}_{i_1}) - \log(1 - \boxed{\mathbf{p}_{i_1}}) + \log(\mathbf{u}_{i_1}^n) - \log(1 - \boxed{\mathbf{u}_{i_1}^n})}{\boxed{t}}\right)$$

Extent of relaxation from Bernoulli

Feature Importance via Concrete Dropout



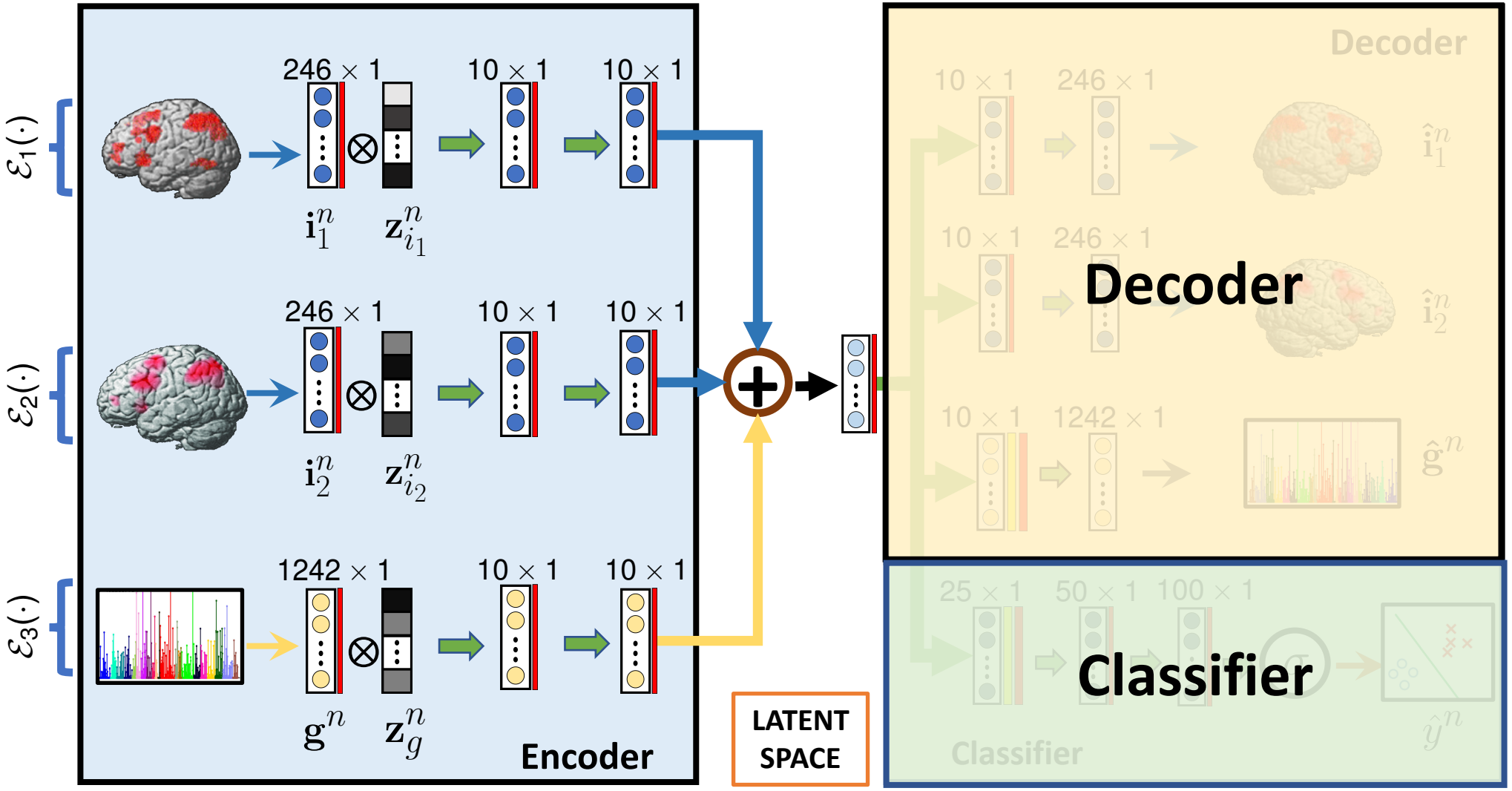
Probability of Retention

Uniform Random Variable

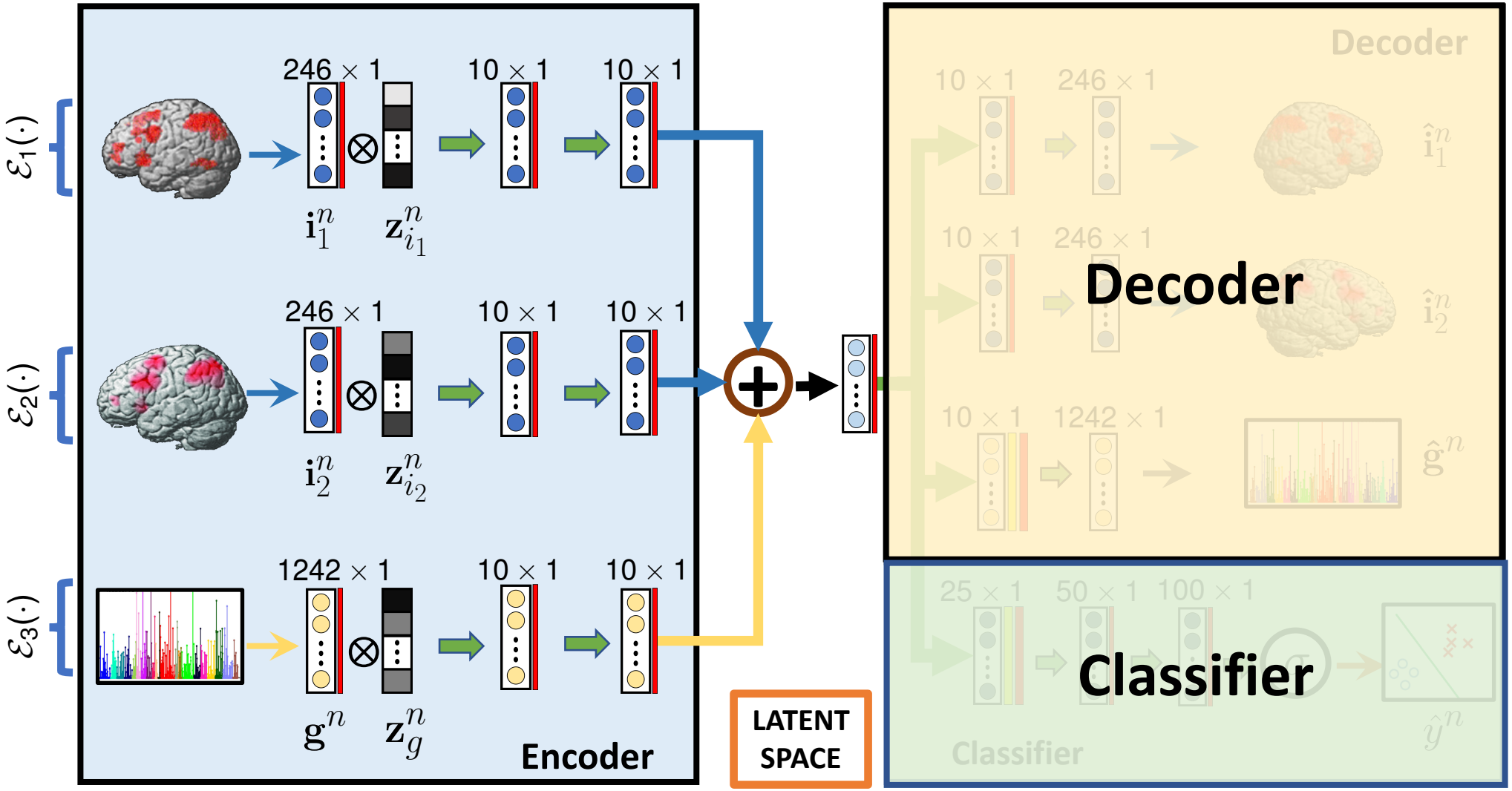
$$\mathbf{z}_{i_1}^n = \sigma\left(\frac{\log(\mathbf{p}_{i_1}) - \log(1 - \boxed{\mathbf{p}_{i_1}}) + \log(\mathbf{u}_{i_1}^n) - \log(1 - \boxed{\mathbf{u}_{i_1}^n})}{\boxed{t}}\right)$$

Extent of relaxation from Bernoulli

GMIND: End-to-End Multimodal Framework

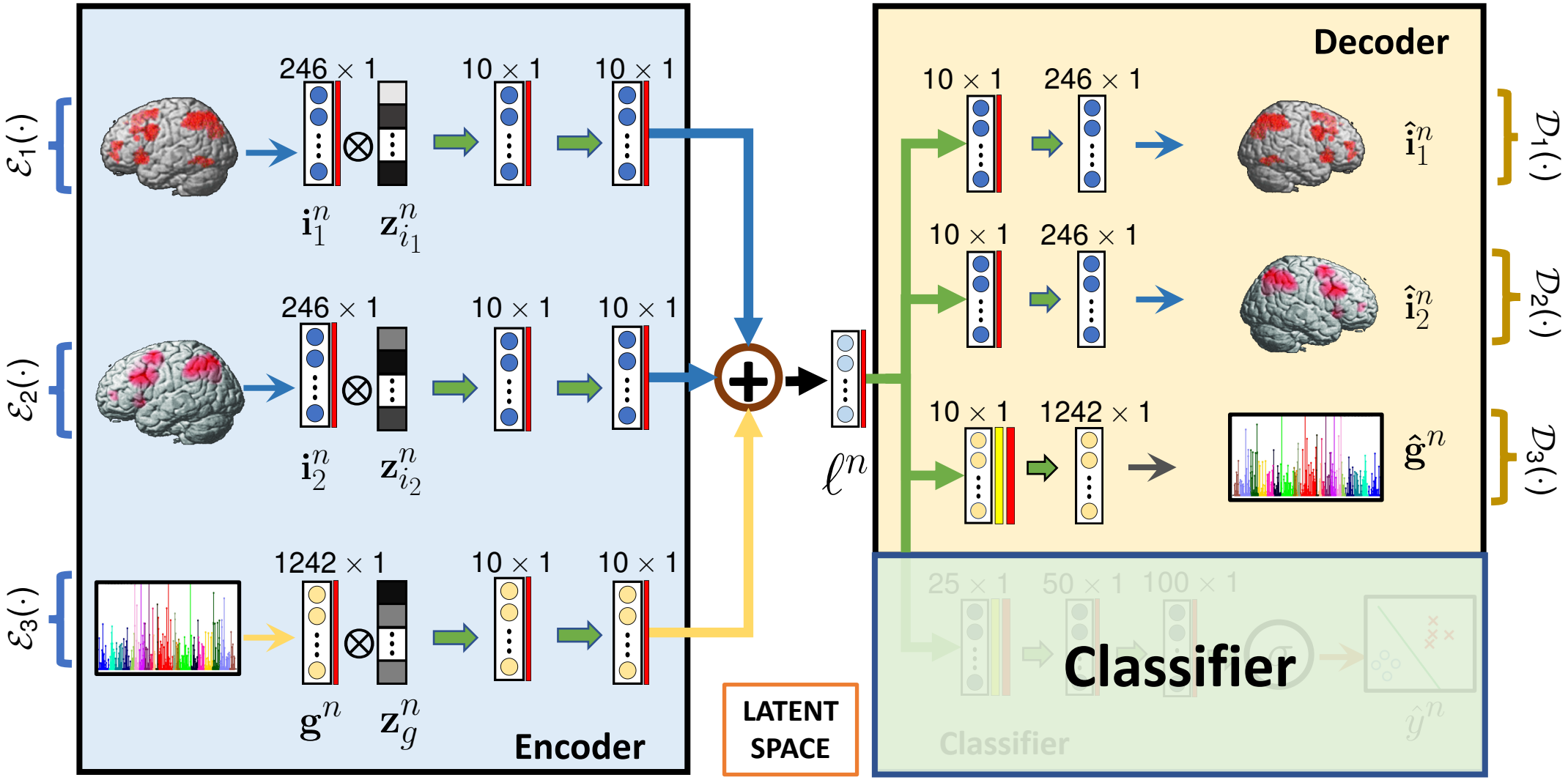


GMIND: End-to-End Multimodal Framework

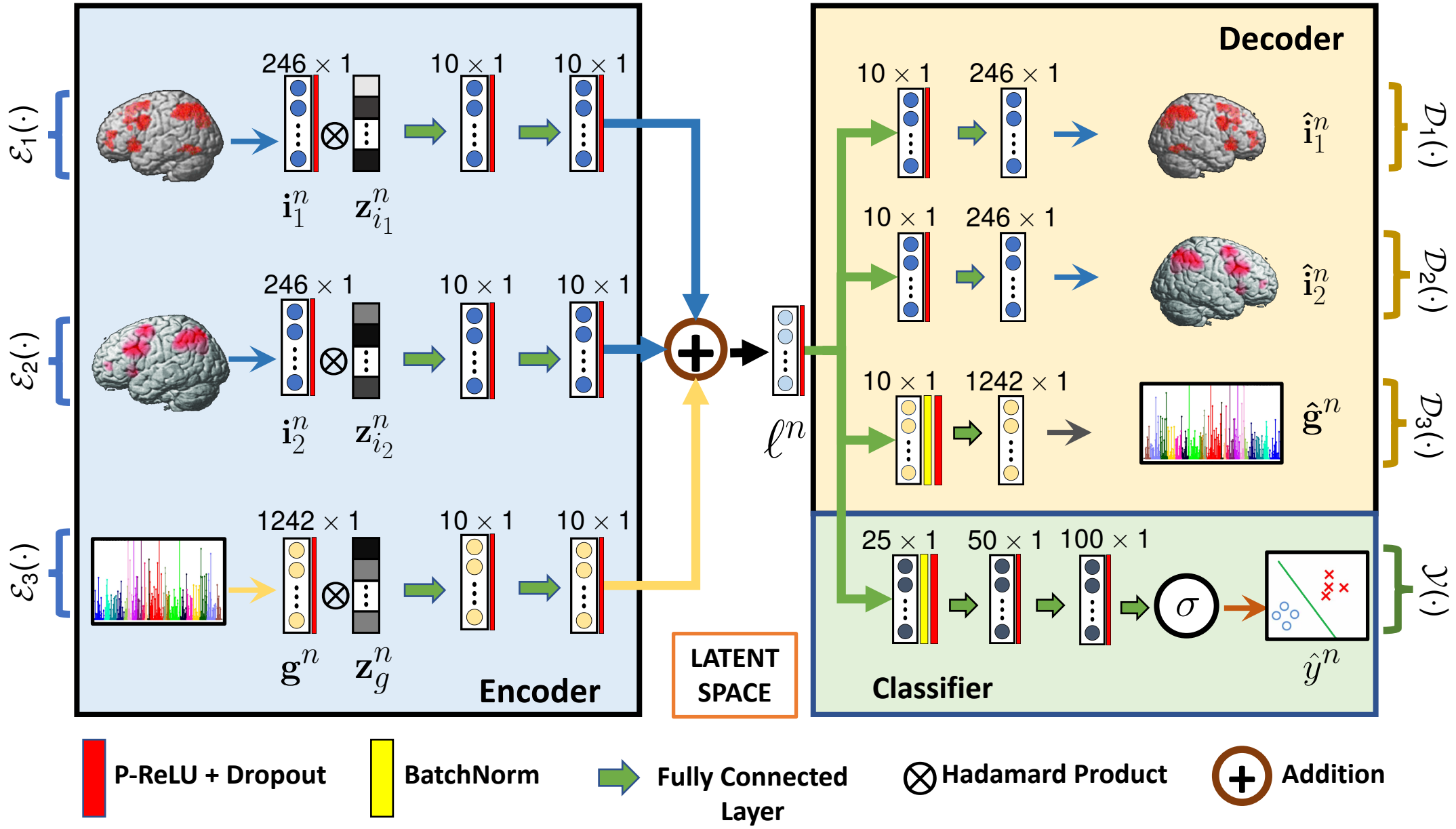


$$\ell^n = \frac{1}{M_n}\left(\mathcal{E}_1(\mathbf{i}_1^n, \mathbf{z}_{i_1}^n) + \mathcal{E}_2(\mathbf{i}_2^n, \mathbf{z}_{i_2}^n) + \mathcal{E}_g(\mathbf{g}^n, \mathbf{z}_g^n)\right)$$

GMIND: End-to-End Multimodal Framework



GMIND: End-to-End Mutimodal Framework



GMIND: End-to-End Mutimodal Framework

$$\mathcal{L}(\mathbf{i}_1, \mathbf{i}_2, \mathbf{g}) = \sum_{n=1}^{N_1} ||\mathbf{i}_1^n - \mathcal{D}_1(\ell^n)||_2^2 + \lambda_2 \sum_{n=1}^{N_2} ||\mathbf{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 + \lambda_3 \sum_{n=1}^{N_g} ||\mathbf{g}^n - \mathcal{D}_3(\ell^n)||_2^2$$

$$-\lambda_4 \sum_{n=1}^{N} (y^n \log(\hat{y}^n) + (1 - y^n) \log(1 - \hat{y}^n)) + \lambda_5 \sum_{m=1}^{3} \sum_{k} KL(Ber(q)||Ber(p_{mk})$$

GMIND: End-to-End Mutimodal Framework

$$\mathcal{L}(\mathbf{i}_1, \mathbf{i}_2, \mathbf{g}) = \underbrace{\sum_{n=1}^{N_1} ||\mathbf{i}_1^n - \mathcal{D}_1(\ell^n)||_2^2 + \lambda_2 \sum_{n=1}^{N_2} ||\mathbf{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 + \lambda_3 \sum_{n=1}^{N_g} ||\mathbf{g}^n - \mathcal{D}_3(\ell^n)||_2^2}_{\textbf{Autoencoder Loss}}$$

$$-\lambda_4 \sum_{n=1}^{N} (y^n \log(\hat{y}^n) + (1 - y^n) \log(1 - \hat{y}^n)) + \lambda_5 \sum_{m=1}^{3} \sum_{k} KL(Ber(q)||Ber(p_{mk})$$

GMIND: End-to-End Mutimodal Framework

$$\mathcal{L}(\mathbf{i}_1, \mathbf{i}_2, \mathbf{g}) = \underbrace{\sum_{n=1}^{N_1} ||\mathbf{i}_1^n - \mathcal{D}_1(\ell^n)||_2^2 + \lambda_2 \sum_{n=1}^{N_2} ||\mathbf{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 + \lambda_3 \sum_{n=1}^{N_g} ||\mathbf{g}^n - \mathcal{D}_3(\ell^n)||_2^2}_{\textbf{Autoencoder Loss}}$$

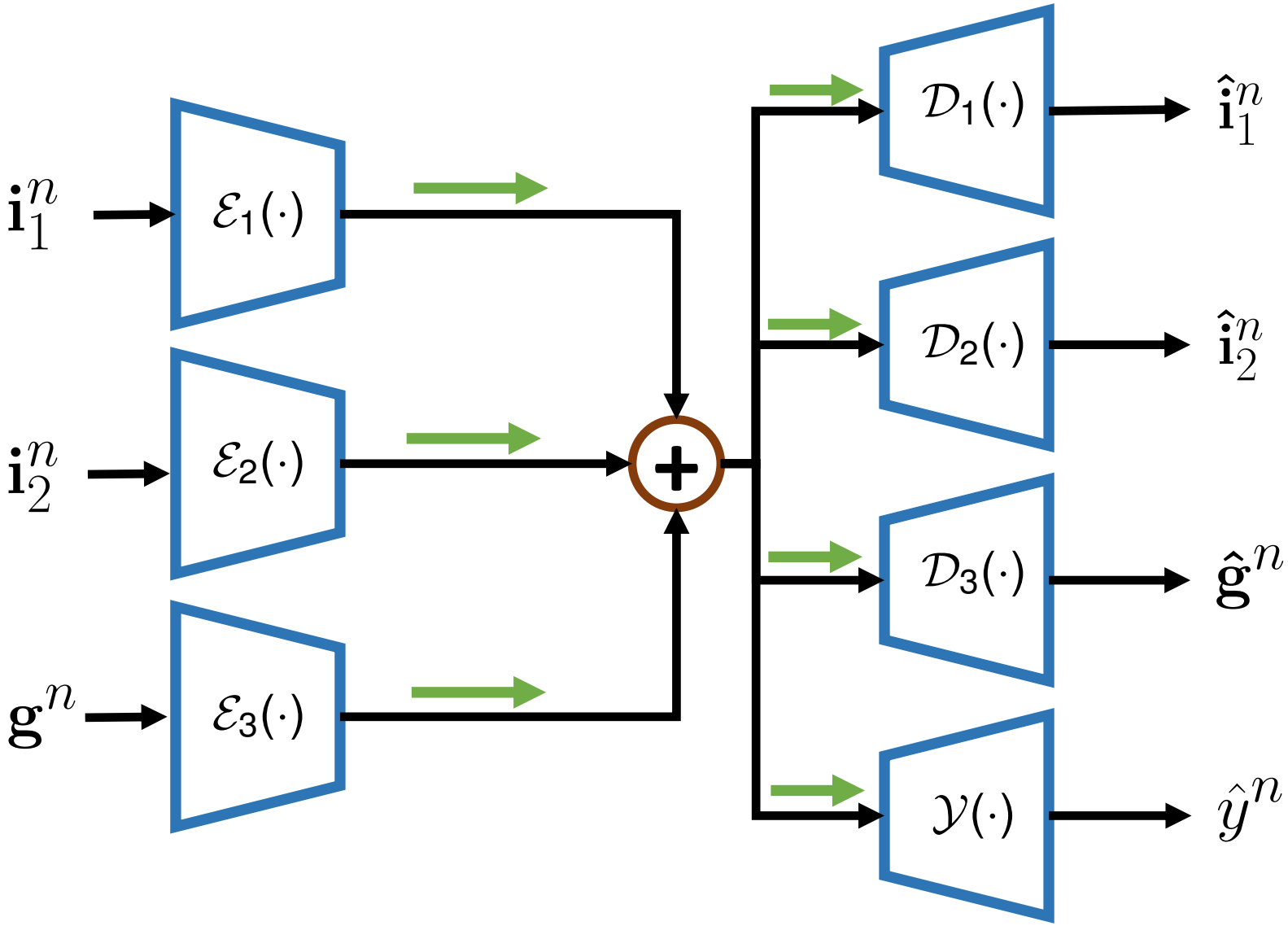
$$\underbrace{-\lambda_4 \sum_{n=1}^{N} (y^n \log(\hat{y}^n) + (1 - y^n) \log(1 - \hat{y}^n))}_{\textbf{Cross Entropy Loss}} + \lambda_5 \sum_{m=1}^{3} \sum_{k} KL(Ber(q)||Ber(p_{mk})$$

GMIND: End-to-End Mutimodal Framework

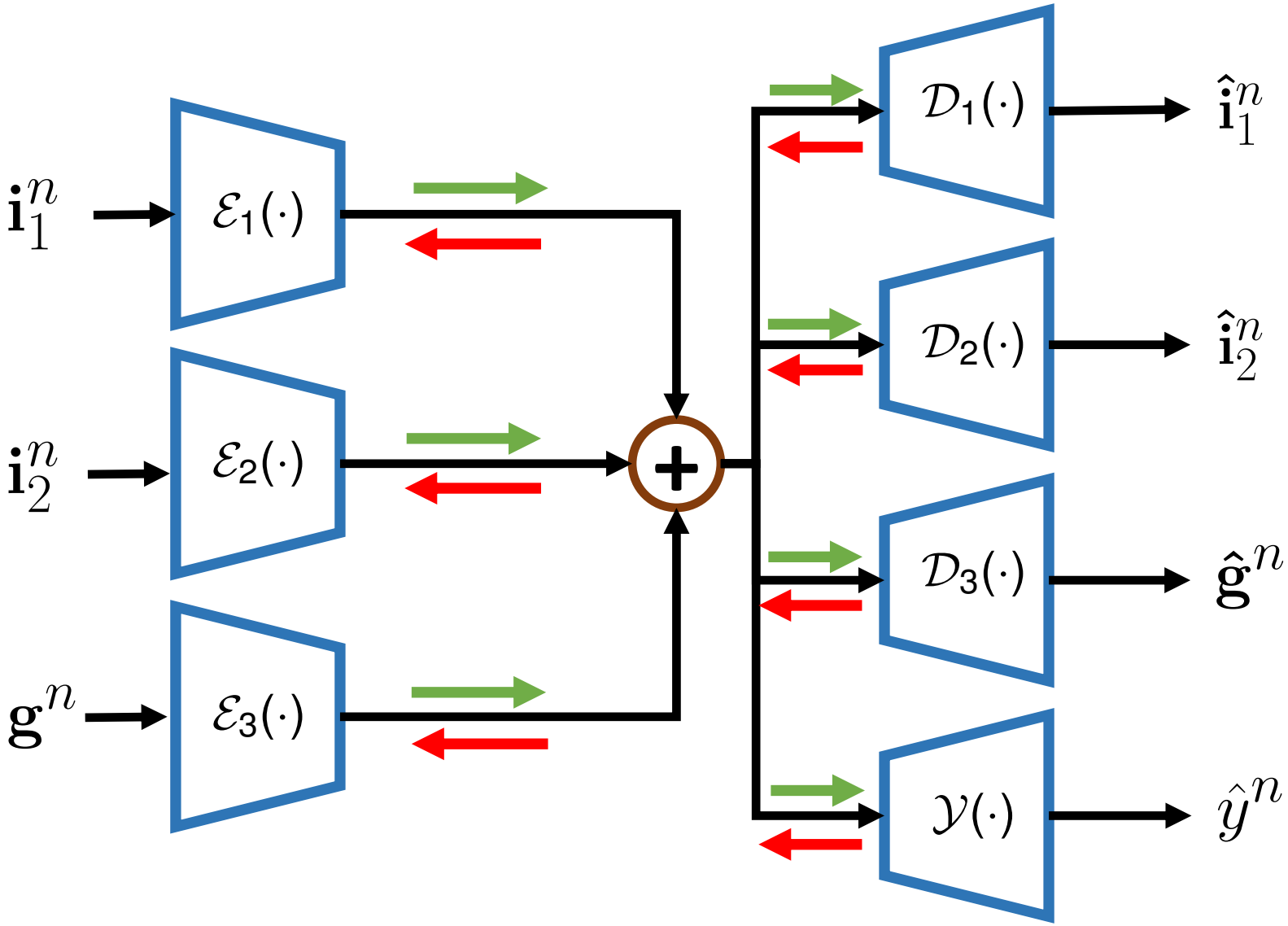
$$\mathcal{L}(\mathbf{i}_1, \mathbf{i}_2, \mathbf{g}) = \underbrace{\sum_{n=1}^{N_1} ||\mathbf{i}_1^n - \mathcal{D}_1(\ell^n)||_2^2 + \lambda_2 \sum_{n=1}^{N_2} ||\mathbf{i}_2^n - \mathcal{D}_2(\ell^n)||_2^2 + \lambda_3 \sum_{n=1}^{N_g} ||\mathbf{g}^n - \mathcal{D}_3(\ell^n)||_2^2}_{\text{Autoencoder Loss}}$$

$$\underbrace{-\lambda_4 \sum_{n=1}^{N} (y^n \log(\hat{y}^n) + (1 - y^n) \log(1 - \hat{y}^n))}_{\text{Cross Entropy Loss}} + \underbrace{\lambda_5 \sum_{m=1}^{3} \sum_{k} KL(Ber(q)||Ber(p_{mk})}_{\text{Sparsity Penalty on Feature Importance Maps}}$$

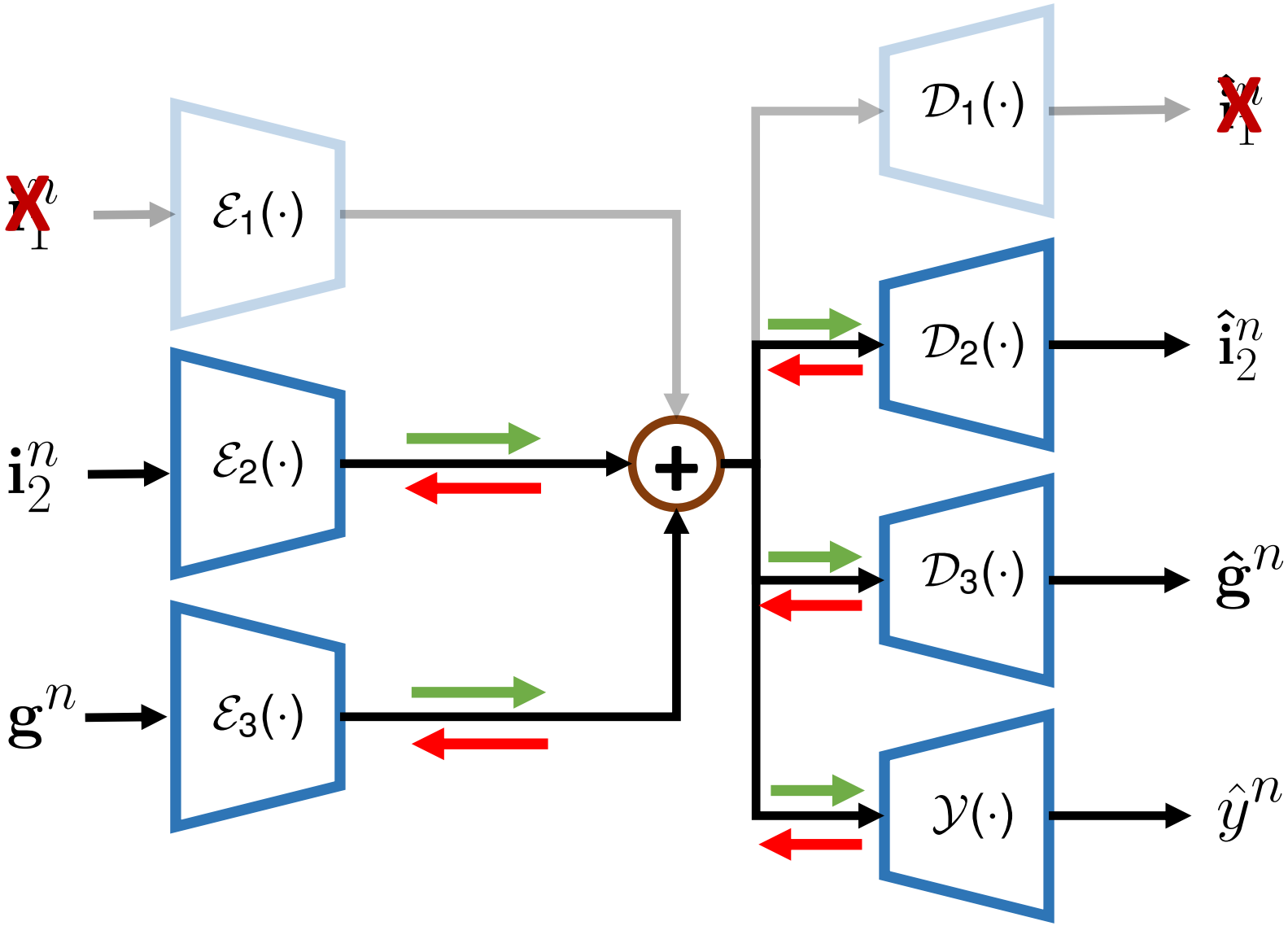
Handling Missing Data



Handling Missing Data



Handling Missing Data



Outline

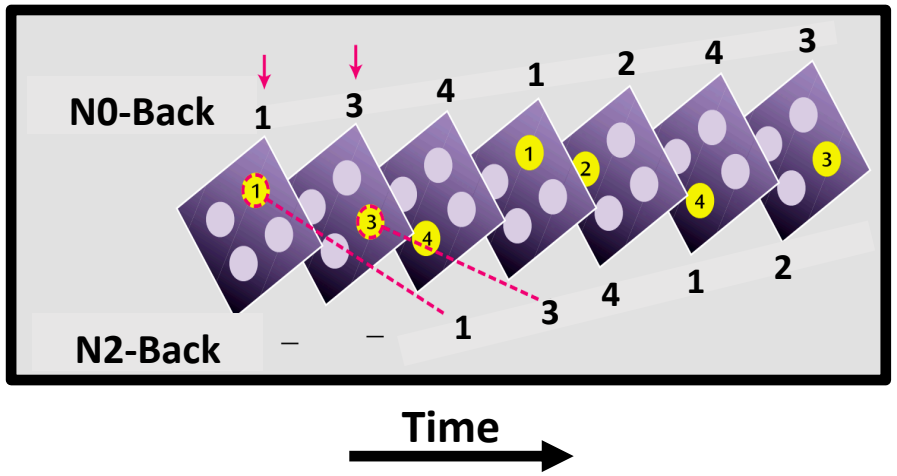
- Background
- GMIND: An End-to-End Model for Imaging-Genetics
- Experimental Results
- Contributions

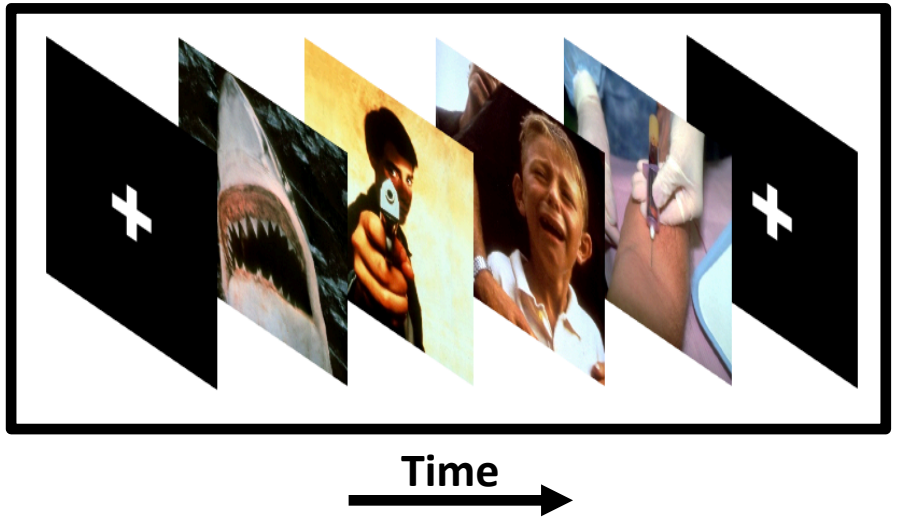
Experimental Data

Two task fMRI datasets:

→ N-back working memory task

→ SDMT episodic memory task





Experimental Data

Two task fMRI datasets:

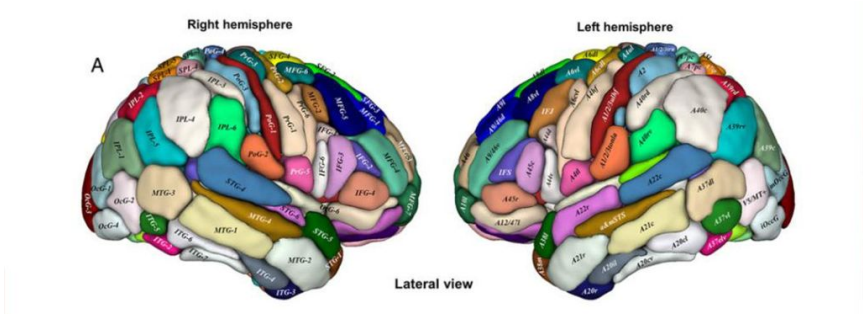
→ N-back working memory task

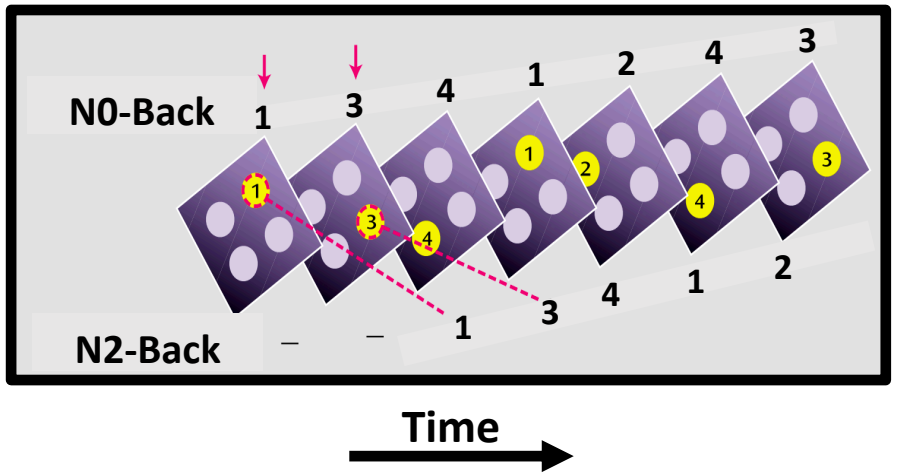
→ SDMT episodic memory task

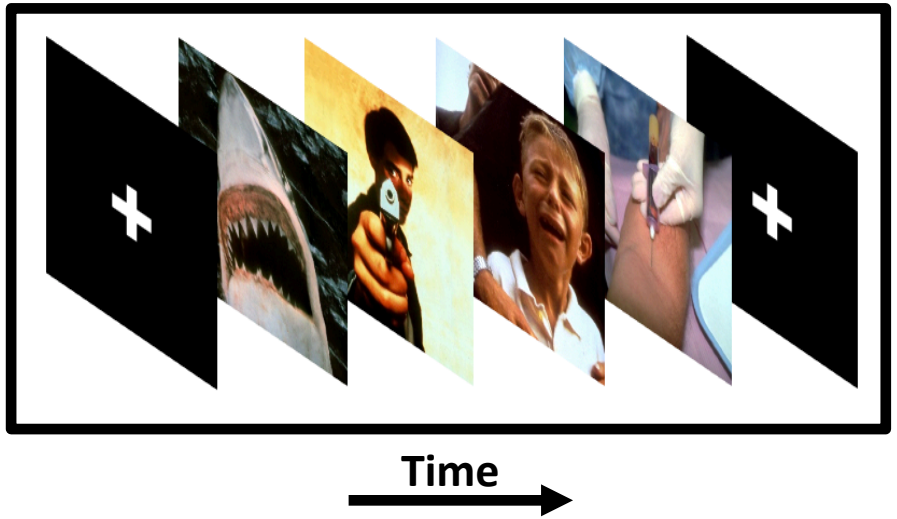
Brain Parcellation:

→ Brainnetome atlas parcellate the brain in 246 region.

→ The contrast map is averaged across the ROIs.





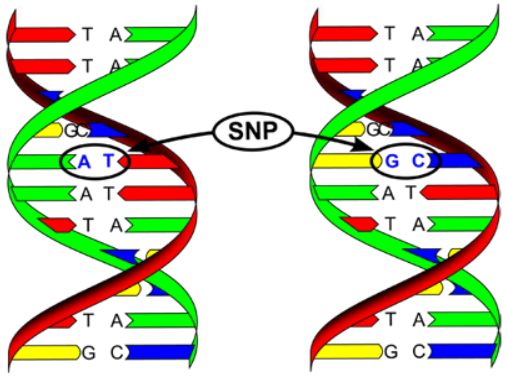


Experimental Data

Genetic Data: SNP

→ Genotyping was done using Illumina Bead Chips (510K/ 610K/660K/2.5M)

→ We use the PGC result to select 1242 independent SNPs thresholded at $P < 10^{-4}$

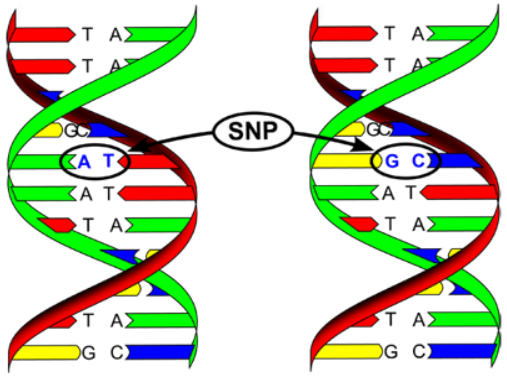


Experimental Data

Genetic Data: SNP

→ Genotyping was done using Illumina Bead Chips (510K/ 610K/660K/2.5M)

→ We use the PGC result to select 1242 independent SNPs thresholded at $P < 10^{-4}$



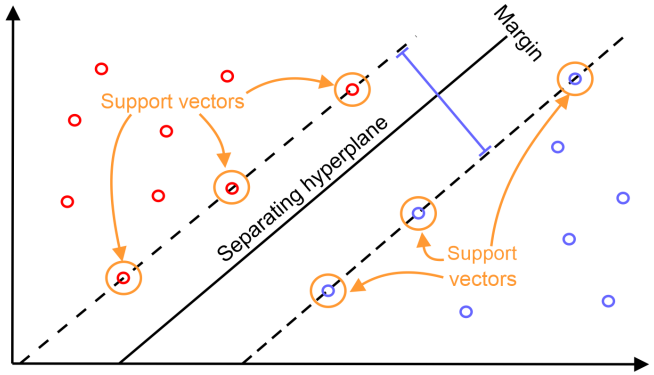
Locations:

- Lieber Institute for Brain Development (LIBD)
- University of Bari, Italy (BARI)

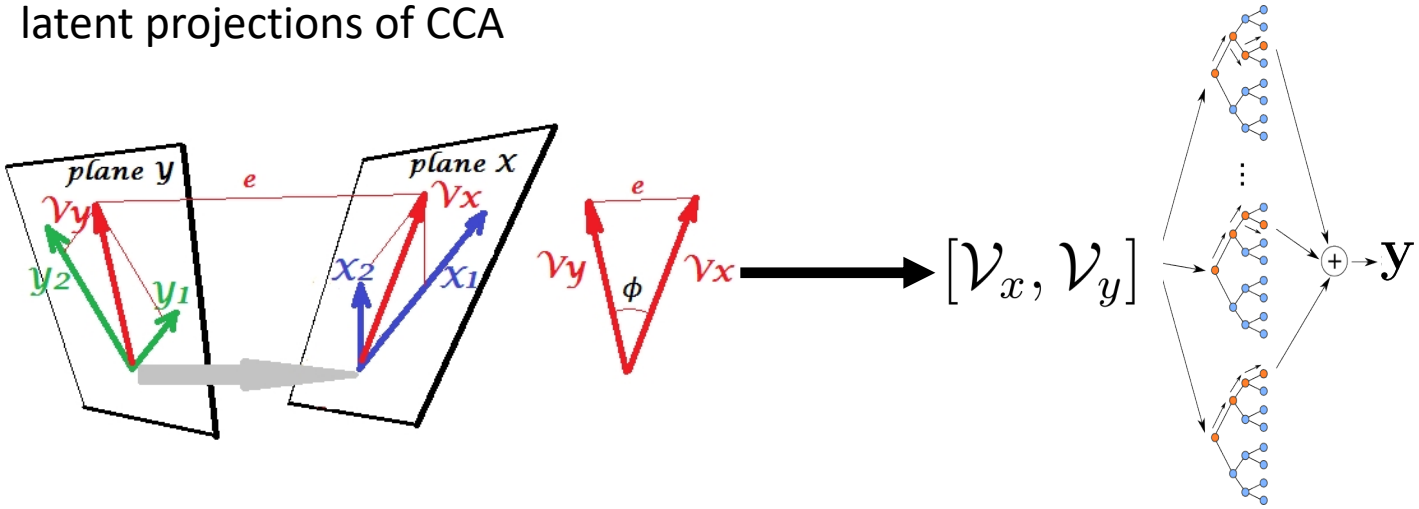
Institution	Modalities		
	N-Back	SDMT	SNP
LIBD	160	110	210
BARI	97	---	97

Baselines

1. **Support Vector Machines:** We construct a linear SVM classifier after concatenating all the data modalities, $\mathcal{I} = [\mathbf{i}_1^T, \mathbf{i}_2^T, \mathbf{g}^T]^T$

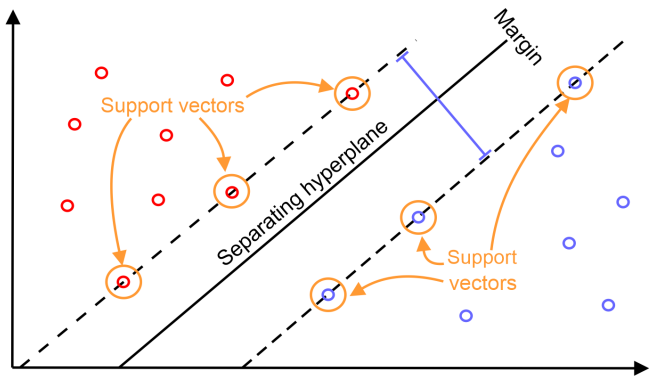


2. **CCA + RF:** We build a random forest classifier using the latent projections of CCA

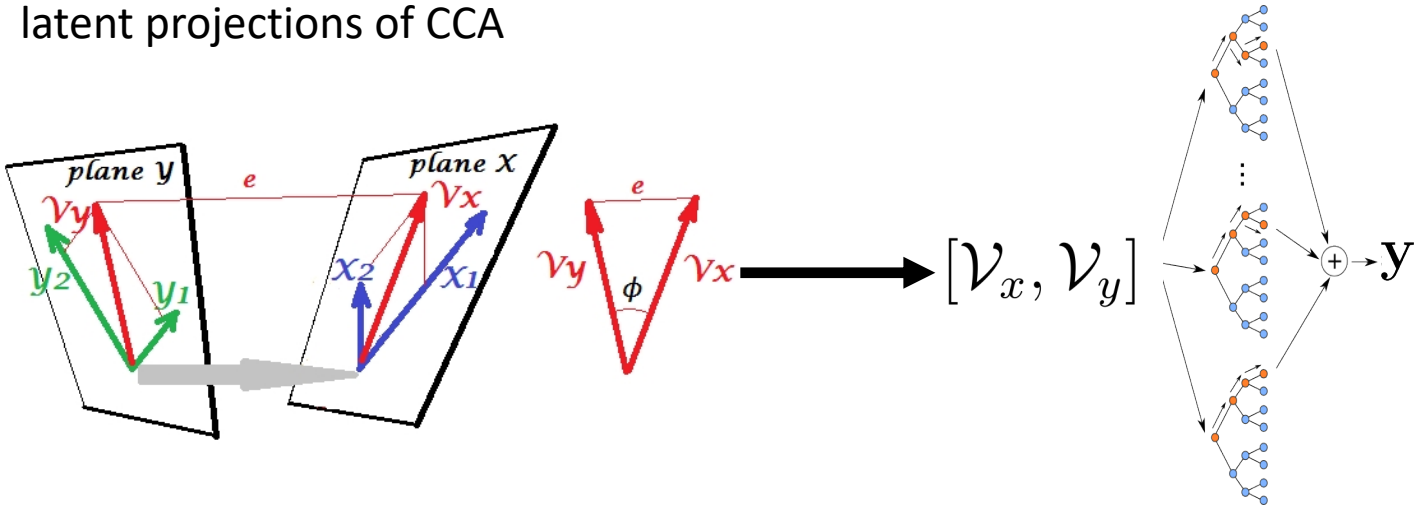


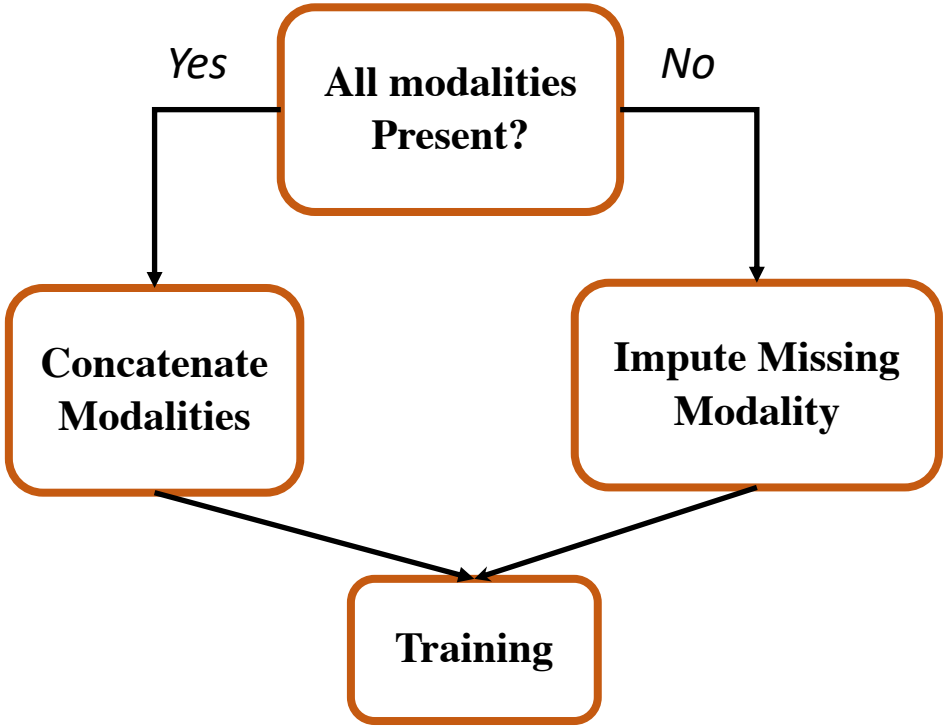
Baselines

1. **Support Vector Machines:** We construct a linear SVM classifier after concatenating all the data modalities, $\mathcal{I} = [\mathbf{i}_1^T, \mathbf{i}_2^T, \mathbf{g}^T]^T$



2. **CCA + RF:** We build a random forest classifier using the latent projections of CCA

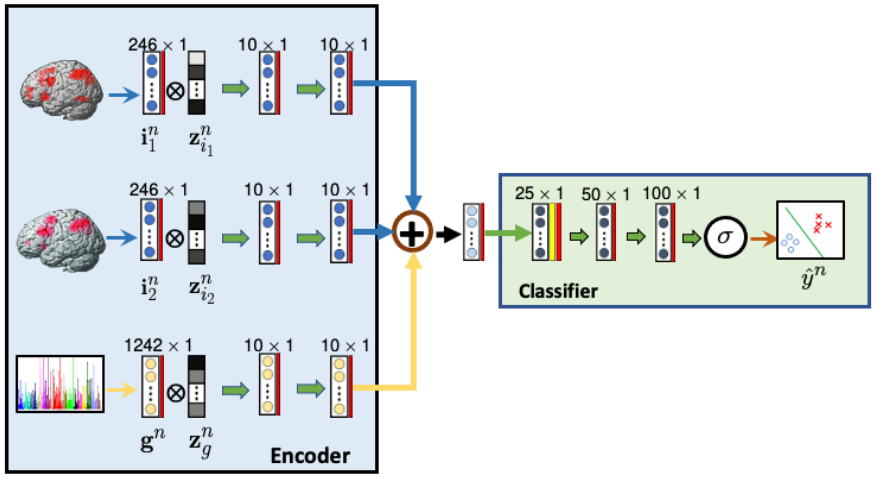




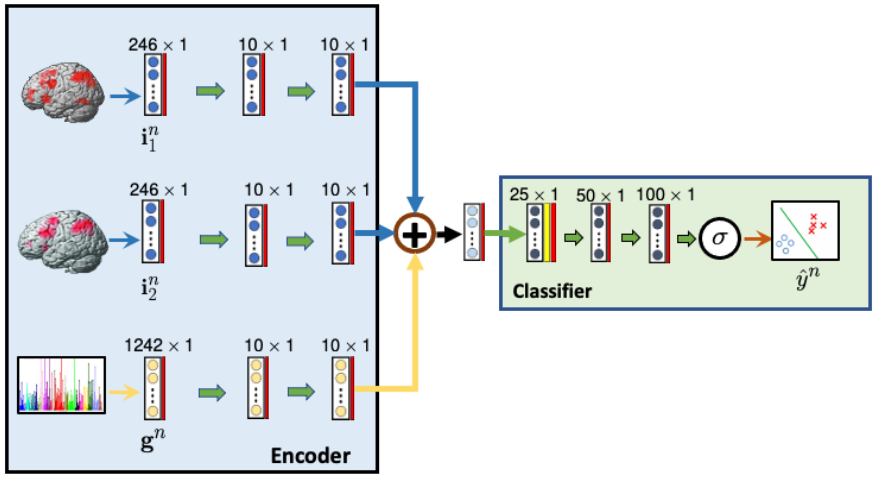
Handling Missing Data

Ablation Study

1. **Encoder + Dropout:** We compare our model to another ANN architecture where we only used the encoder, the classifier, and the learnable dropout layer.



2. **Encoder Only:** We compare our model to an ANN architecture based on the encoder and the classifier of G-MIND.



Quantitative Results

Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

10 fold cross validation result on LIBD data which includes missing data modalities

Quantitative Results

Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

10 fold cross validation result on LIBD data which includes missing data modalities

Quantitative Results

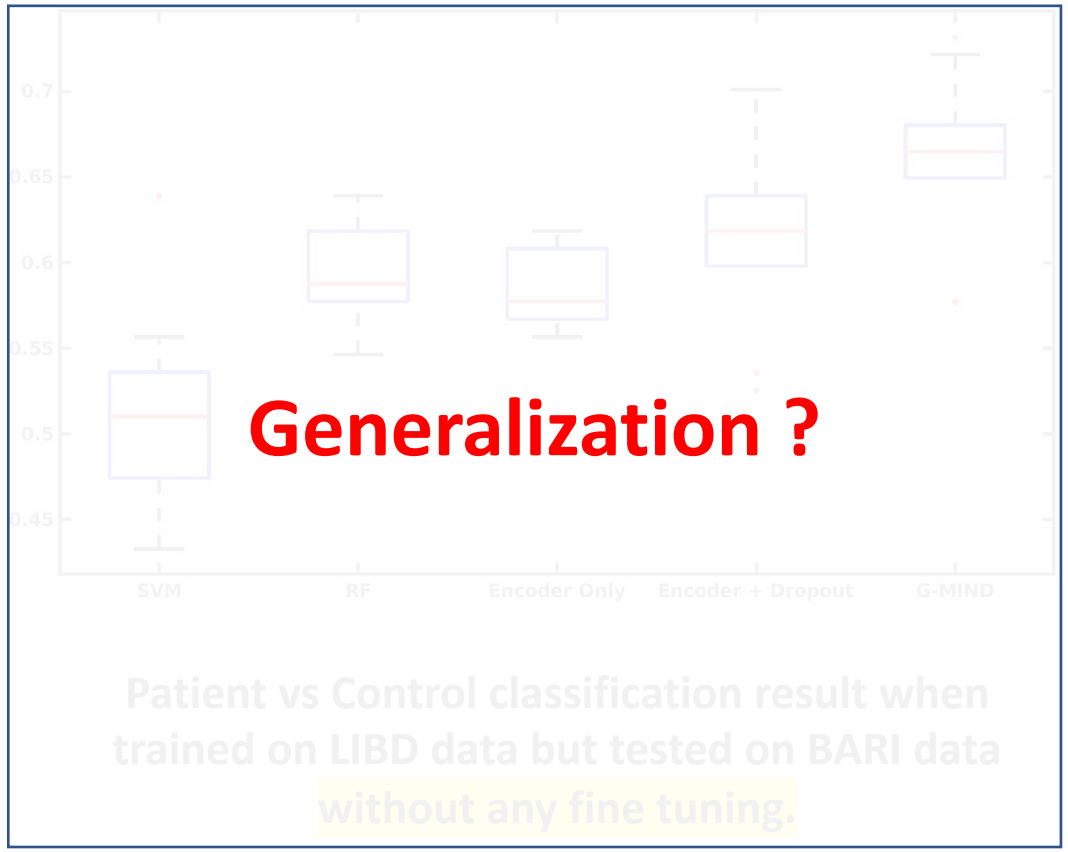
Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

10 fold cross validation result on LIBD data which includes missing data modalities

Quantitative Results

Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

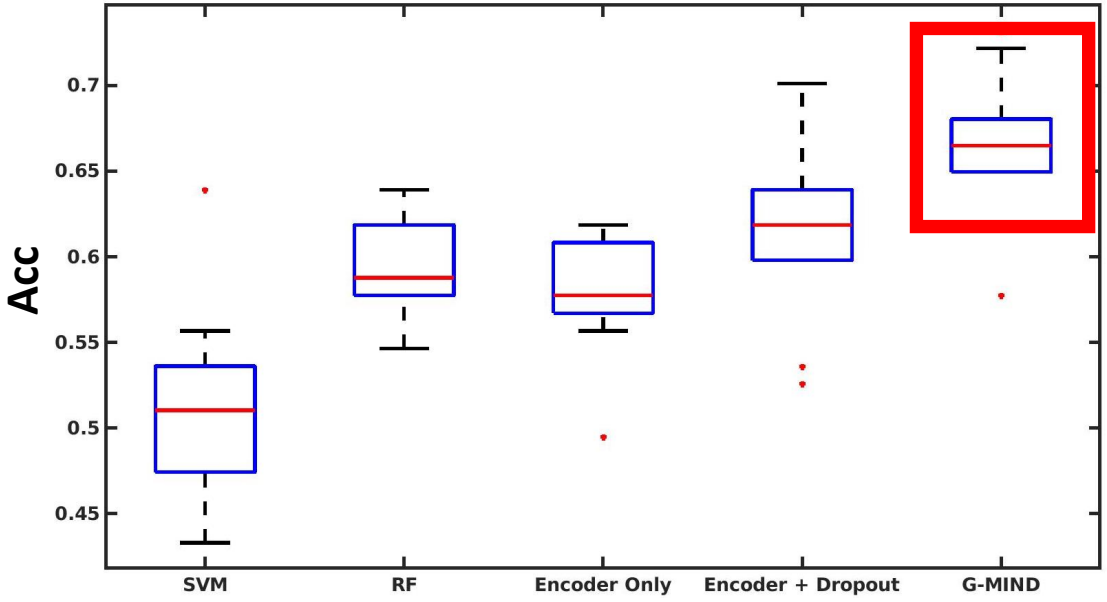
10 fold cross validation result on LIBD data which includes missing data modalities



Quantitative Results

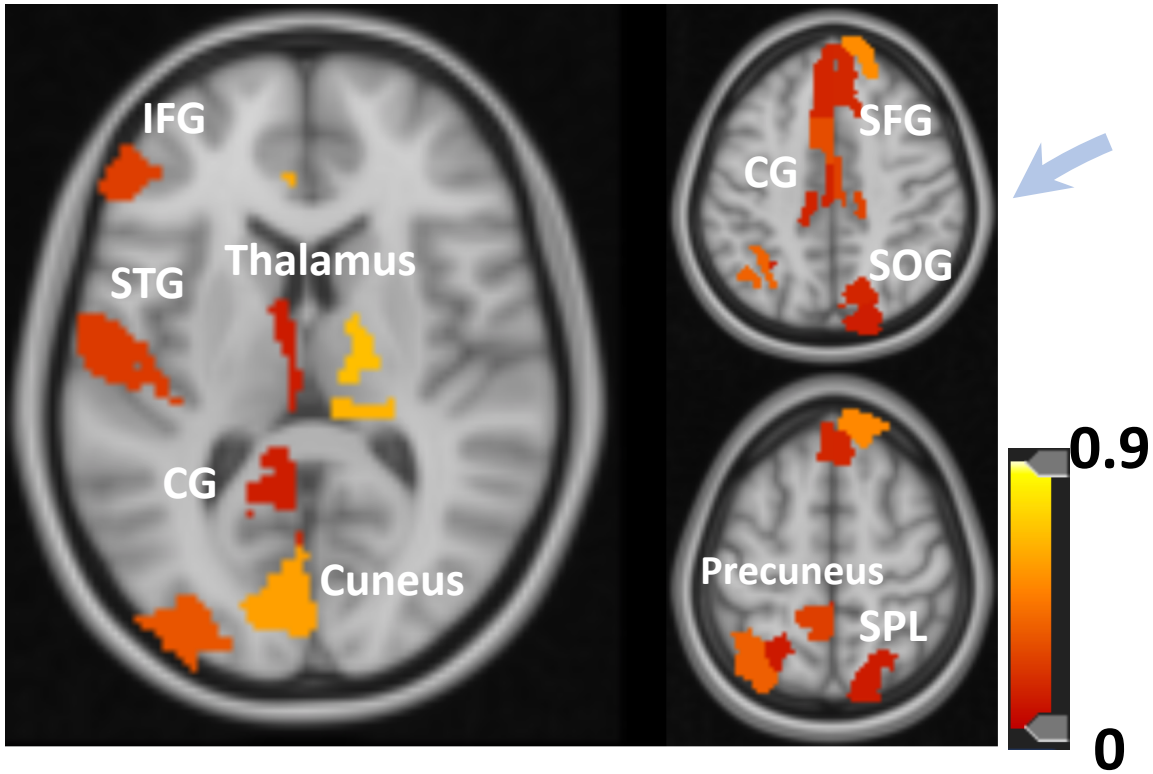
Method \ Perf	Sens	Spec	Acc	Auc
SVM	0.66	0.47	0.58	0.55
CCA+RF	0.15	**0.92**	0.51	0.56
Encoder Only	0.57	0.57	0.57	0.59
Encoder + Dropout	0.61	0.56	0.59	0.62
G-MIND	**0.75**	0.58	**0.67**	**0.68**

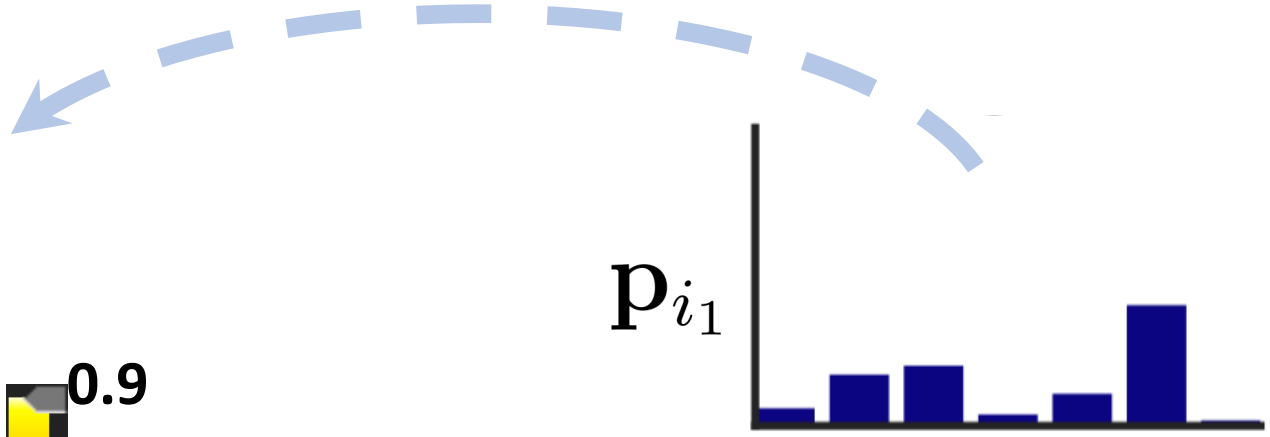
10 fold cross validation result on LIBD data which includes missing data modalities



Patient vs Control classification result when trained on LIBD data but tested on BARI data without any fine tuning.

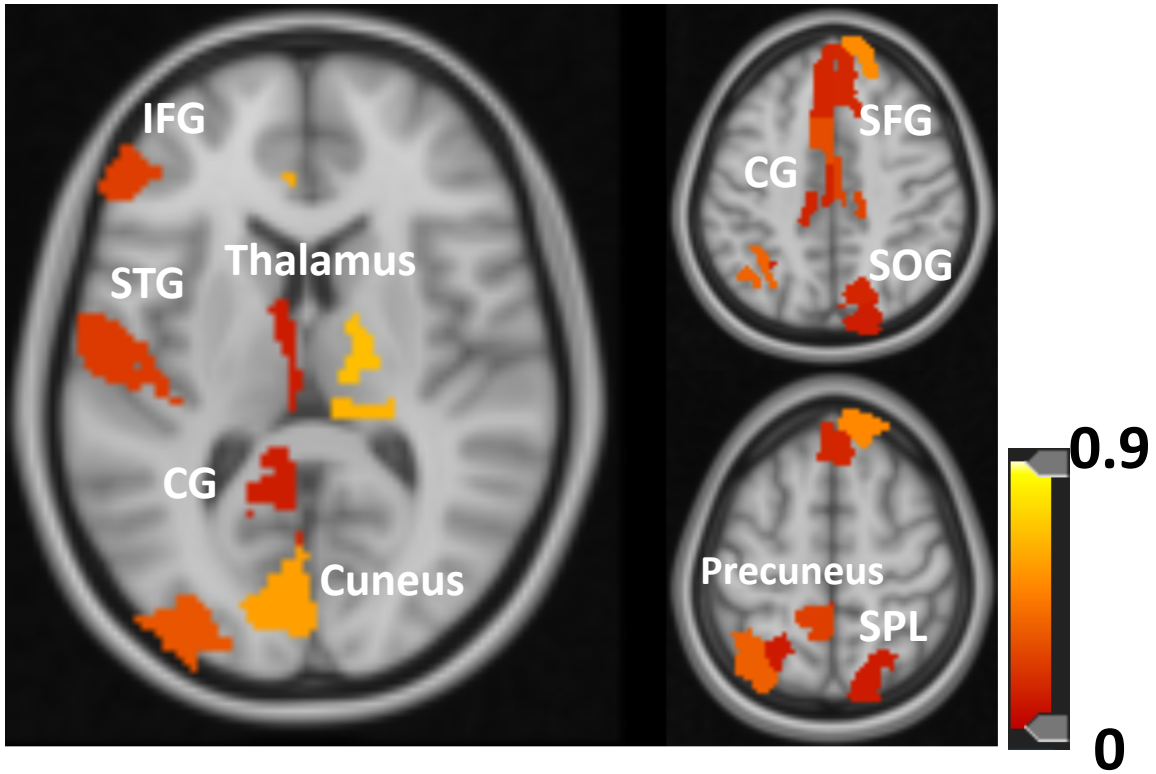
Imaging Biomarkers (N-back)



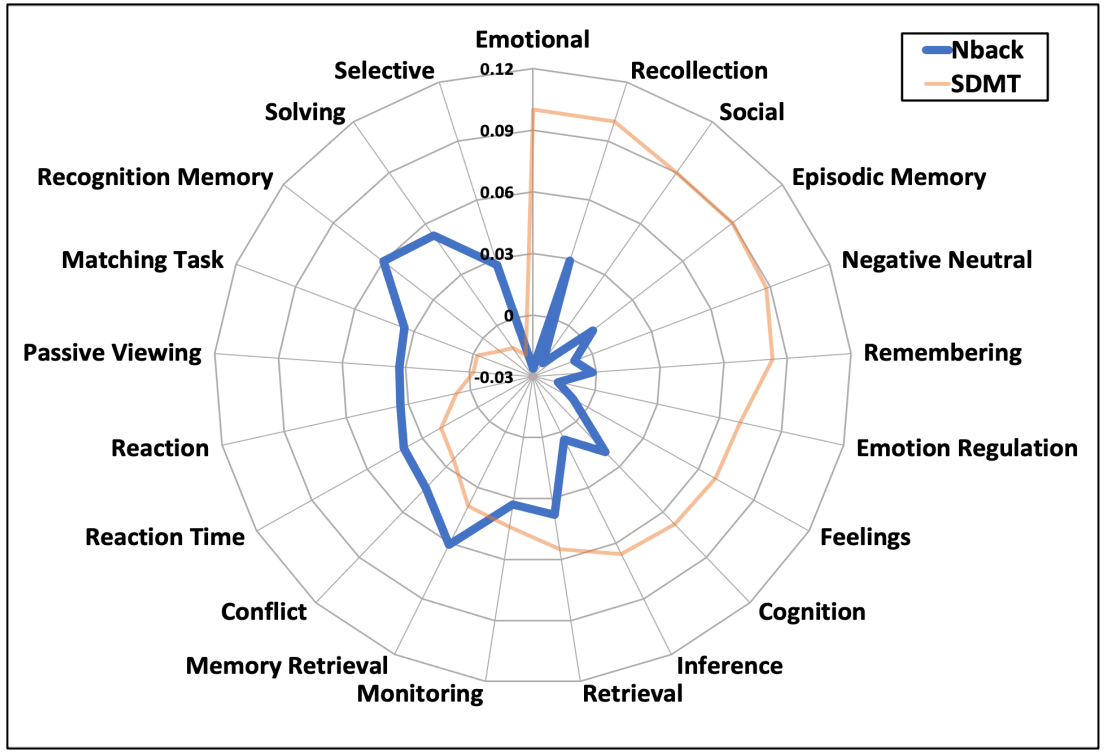


The representative set of brain regions as captured by the probability map for N-back task

Imaging Biomarkers (N-back)

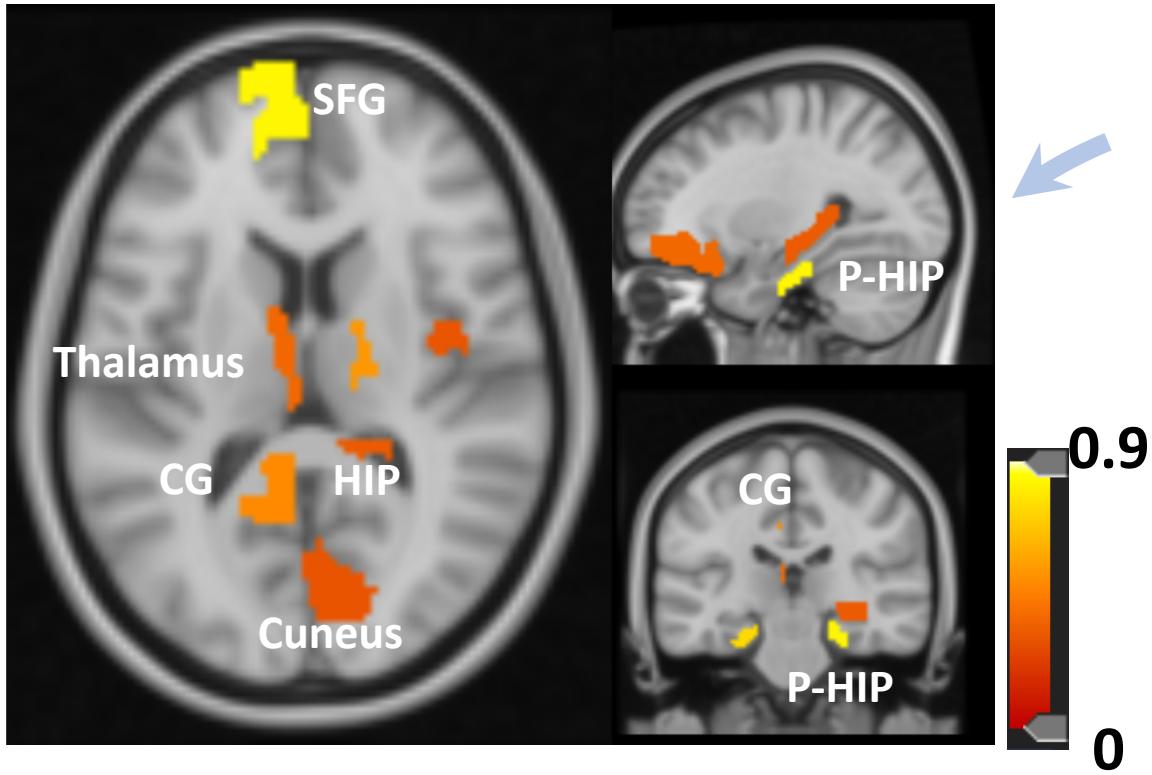


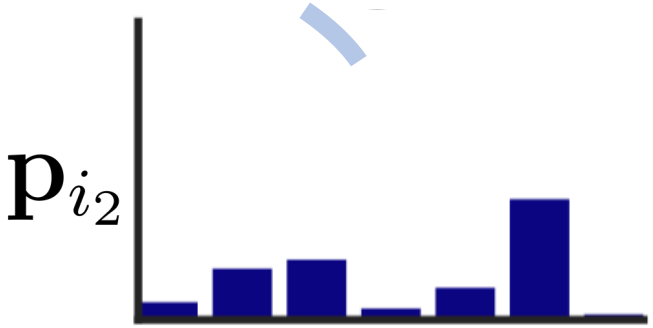
The representative set of brain regions as captured by the probability map for N-back task



Higher order cognitive states underlying the imaging biomarkers for N-back task

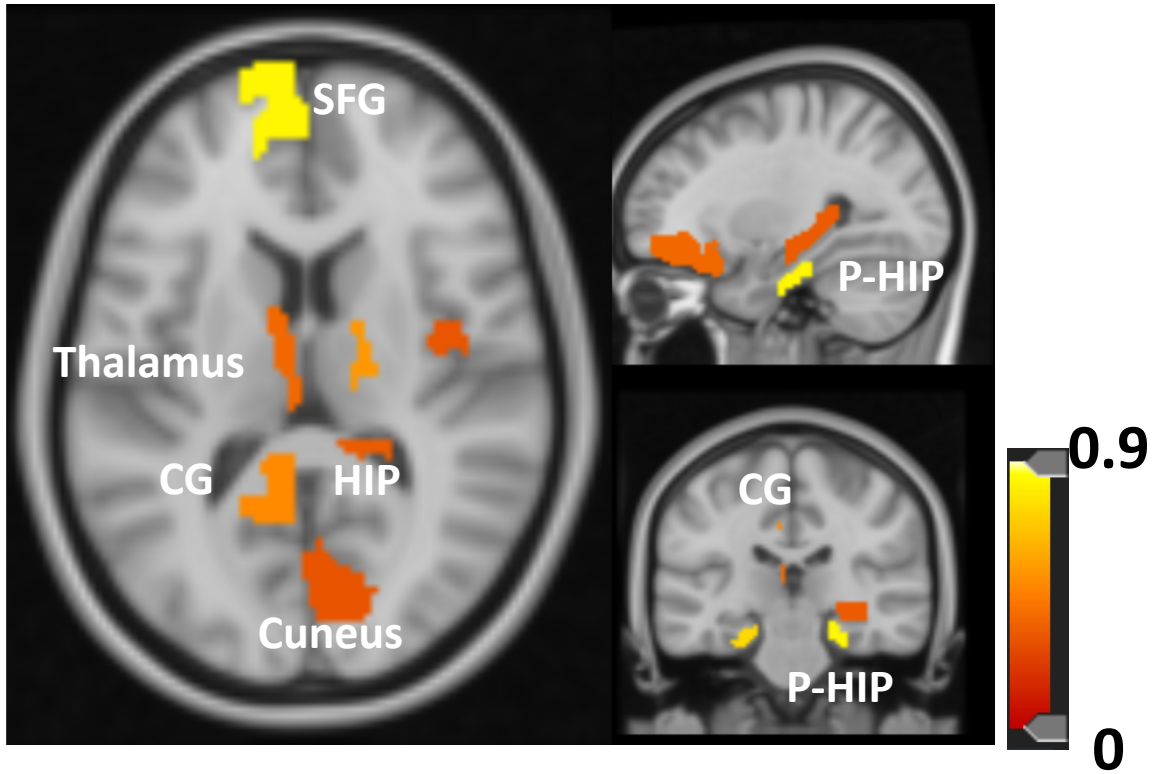
Imaging Biomarkers (SDMT)



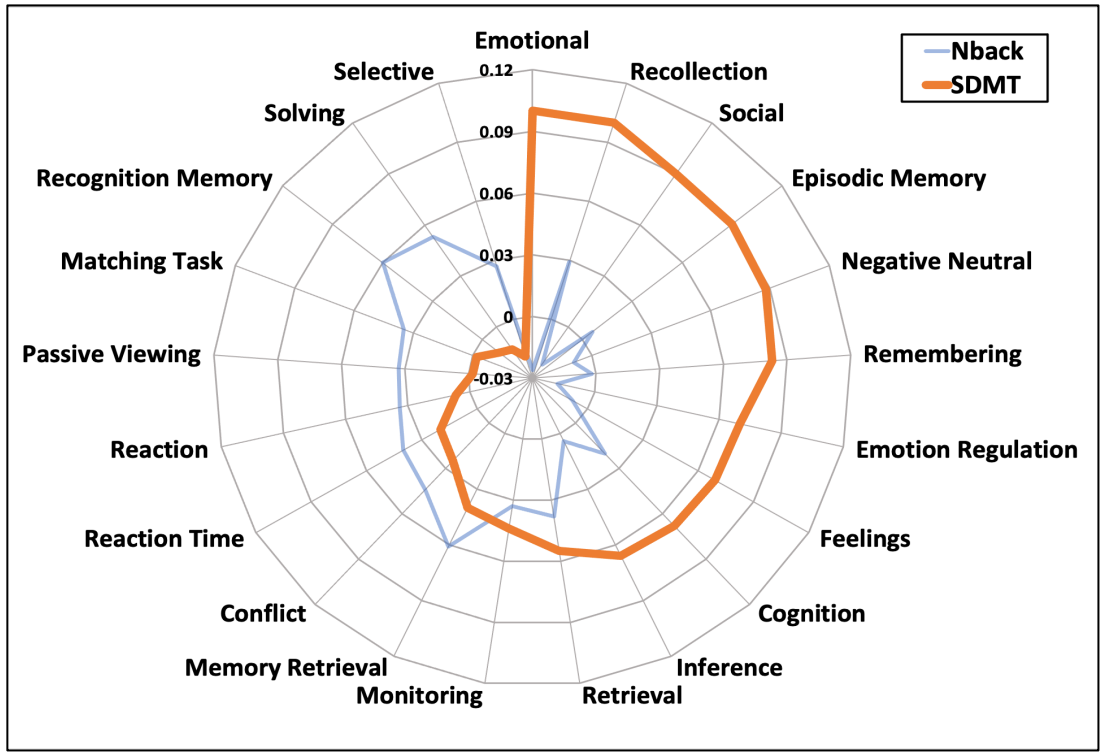


The representative set of brain regions as captured by the probability map for SDMT task

Imaging Biomarkers (SDMT)

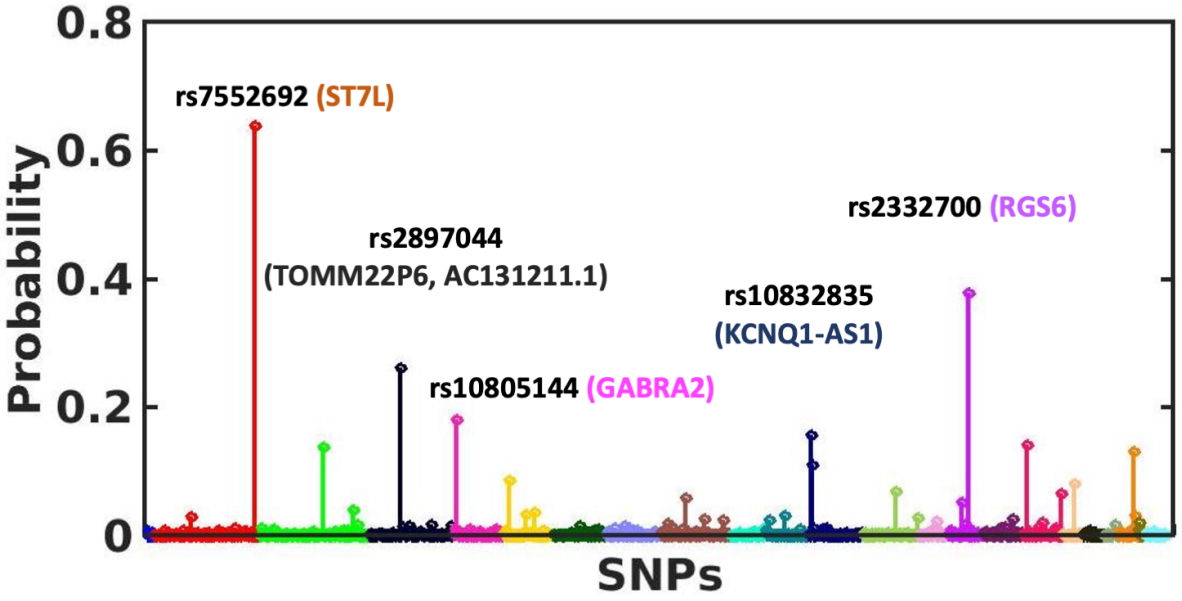


The representative set of brain regions as captured by the probability map for SDMT task



Higher order cognitive states underlying the imaging biomarkers for SDMT task

Genetic Biomarkers

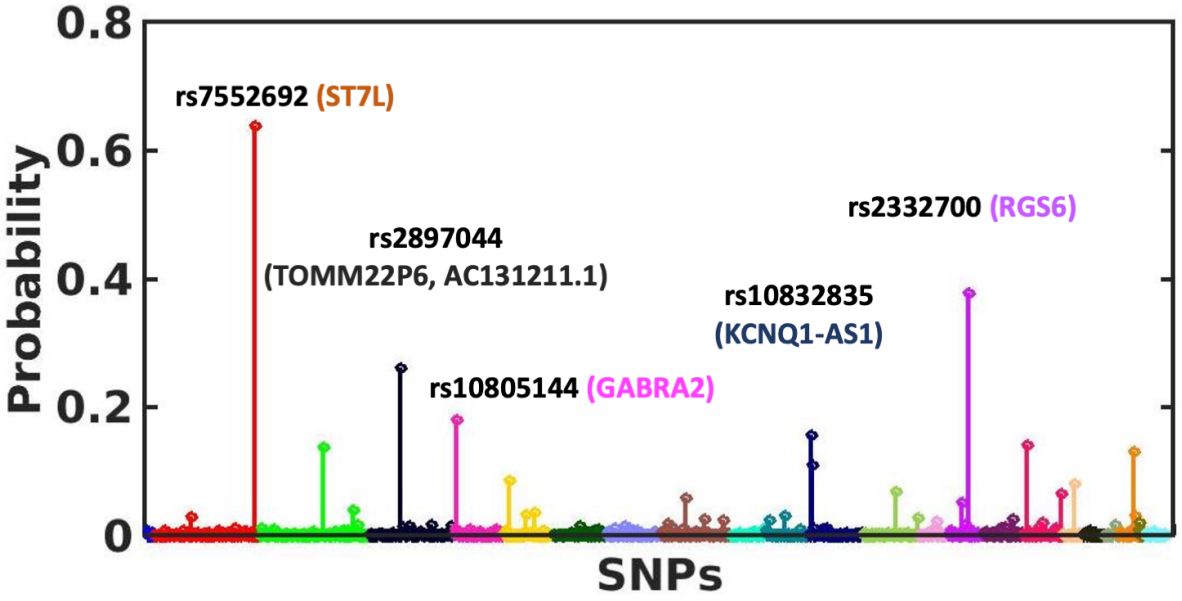


The median importance map of all the SNP across and their overlap-ping genes across the 10 folds

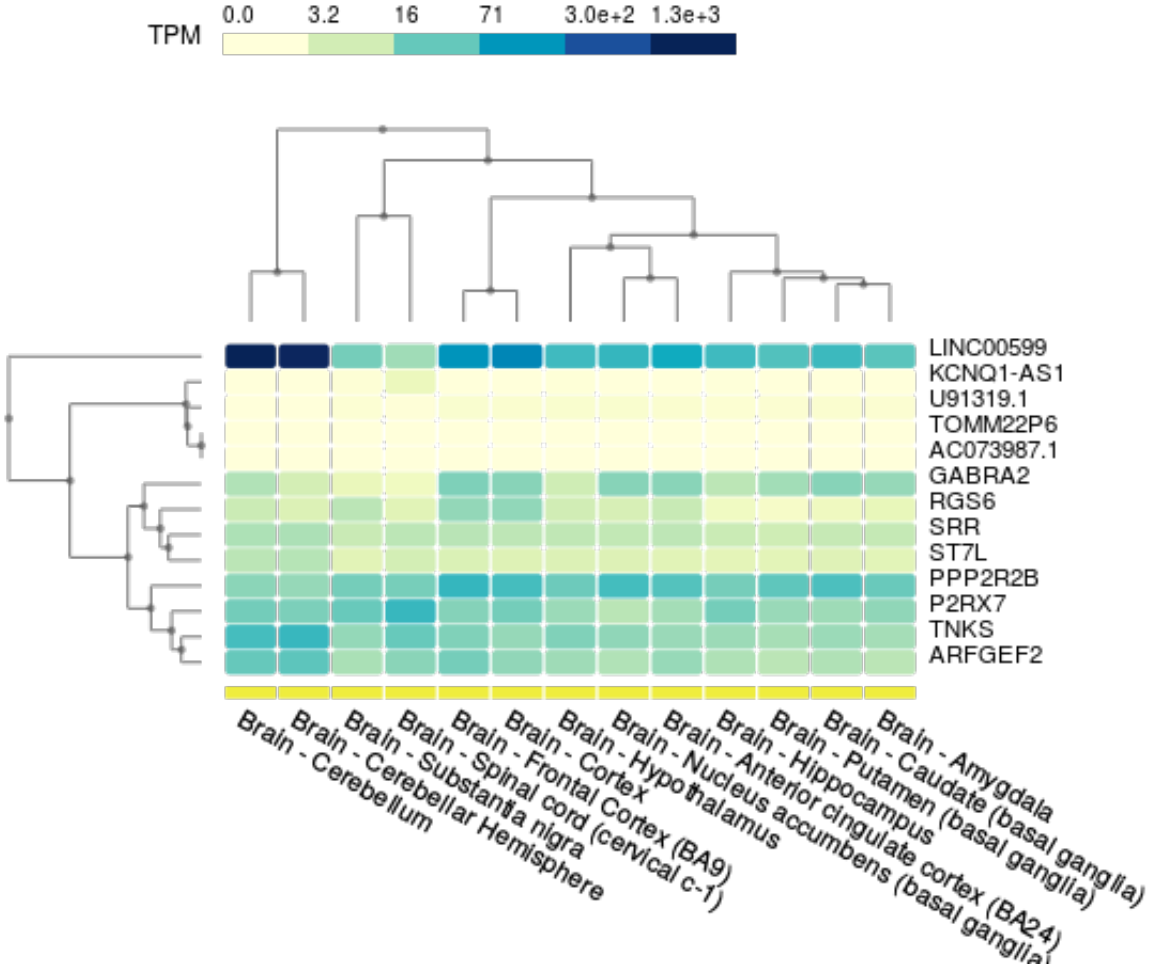
Biological Processes	FDR
Central nervous system development	0.005
→ Nervous system development	0.001
→ System development.	0.005
Generation of neurons	0.005
→ Neurogenesis	0.004
Regulation of calcium ion transport into cytosol	0.04
→ Regulation of sequestering of calcium ion	0.008

The enriched biological processes obtained via GO enrichment analysis

Genetic Biomarkers

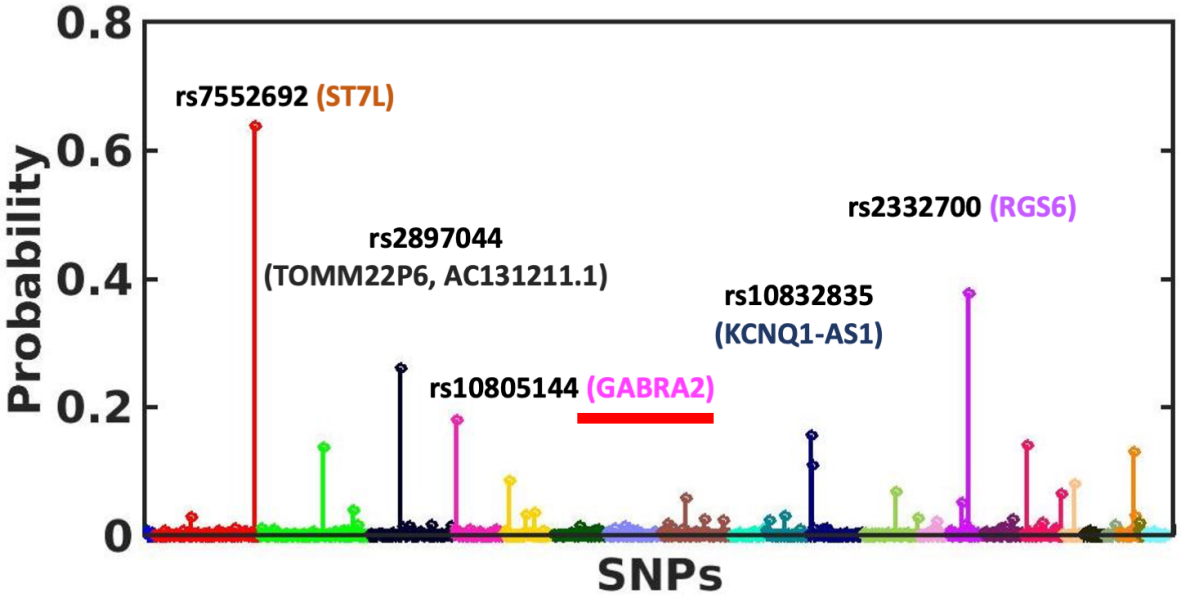


The median importance map of all the SNP across and their overlap-ping genes across the 10 folds

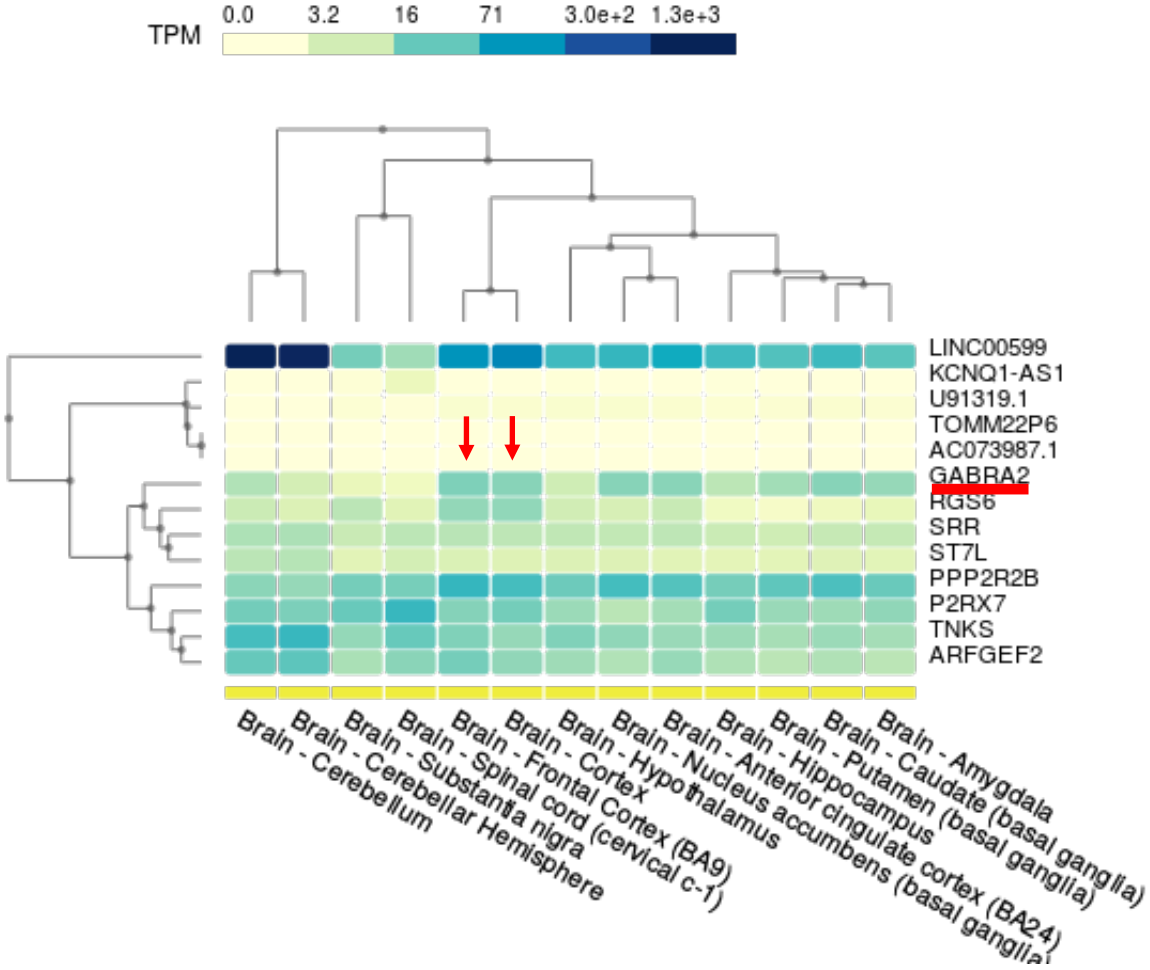


The gene expression pattern of the selected set of genes in different brain tissues.

Genetic Biomarkers

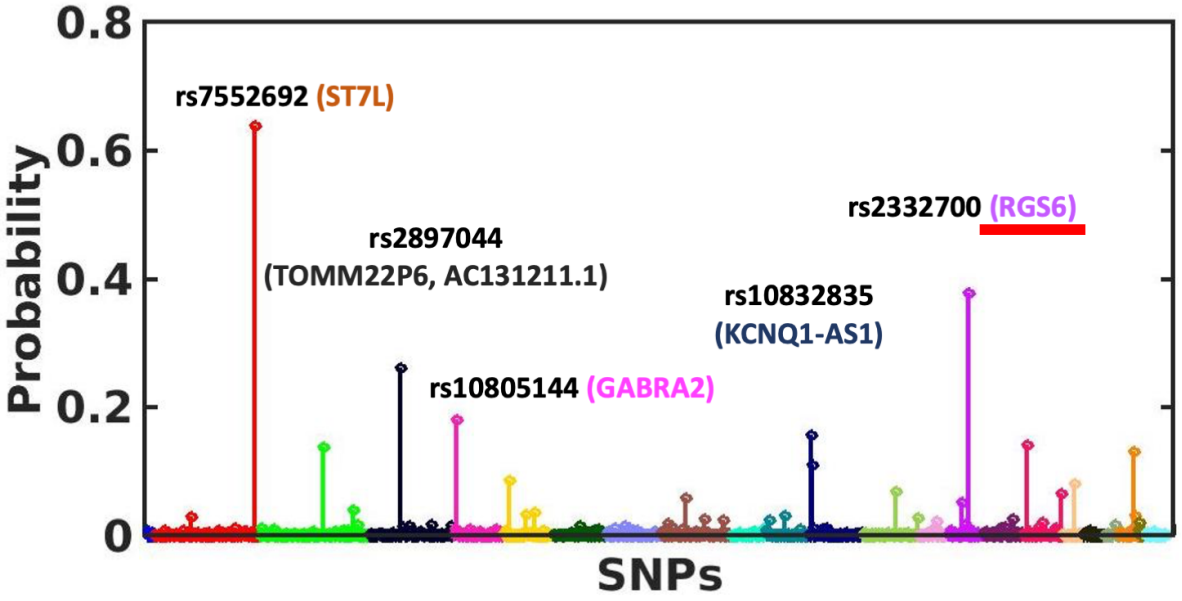


The median importance map of all the SNP across and their overlap-ping genes across the 10 folds

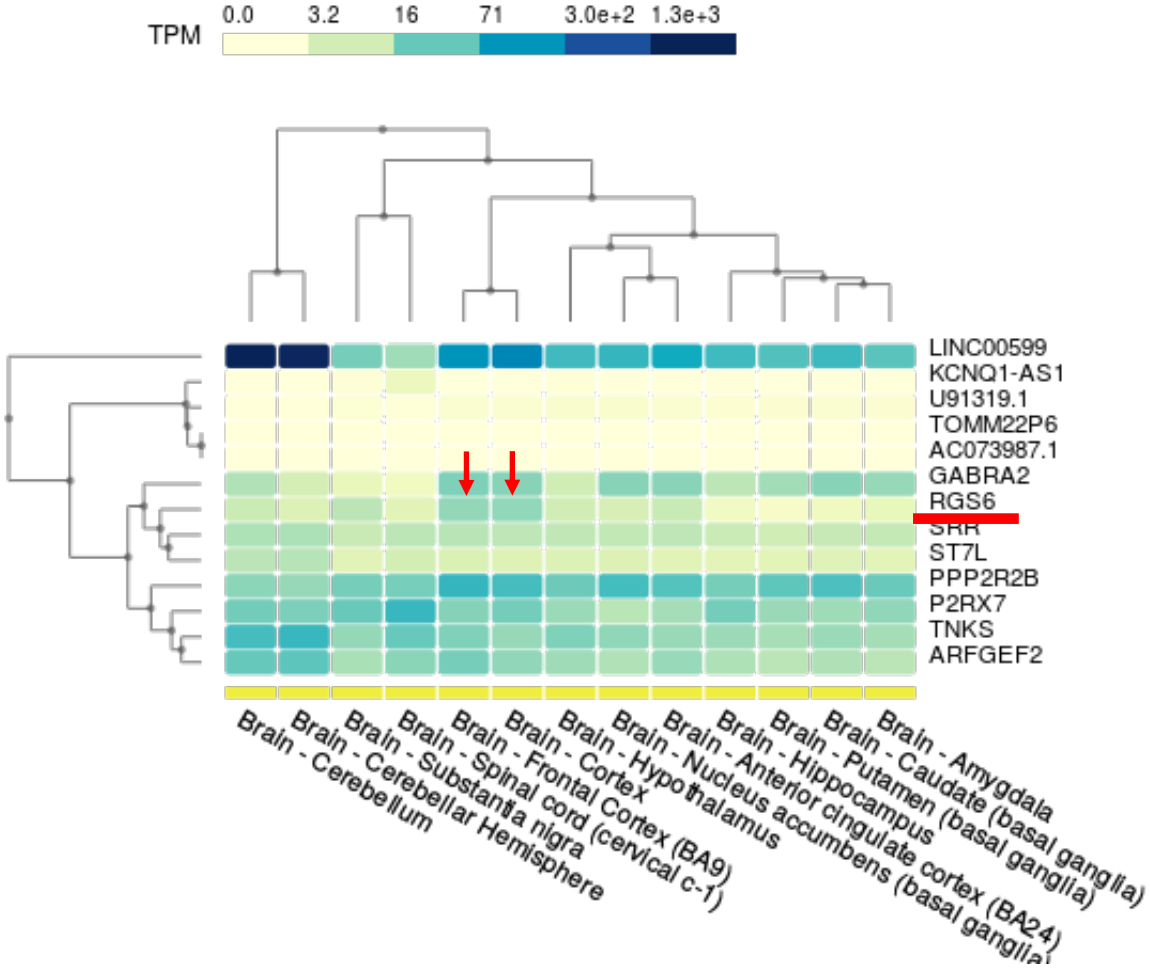


The gene expression pattern of the selected set of genes in different brain tissues.

Genetic Biomarkers



The median importance map of all the SNP across and their overlap-ping genes across the 10 folds



The gene expression pattern of the selected set of genes in different brain tissues.

Outline

- Background
- GMIND: An End-to-End Model for Imaging-Genetics
- Experimental Results
- Contributions

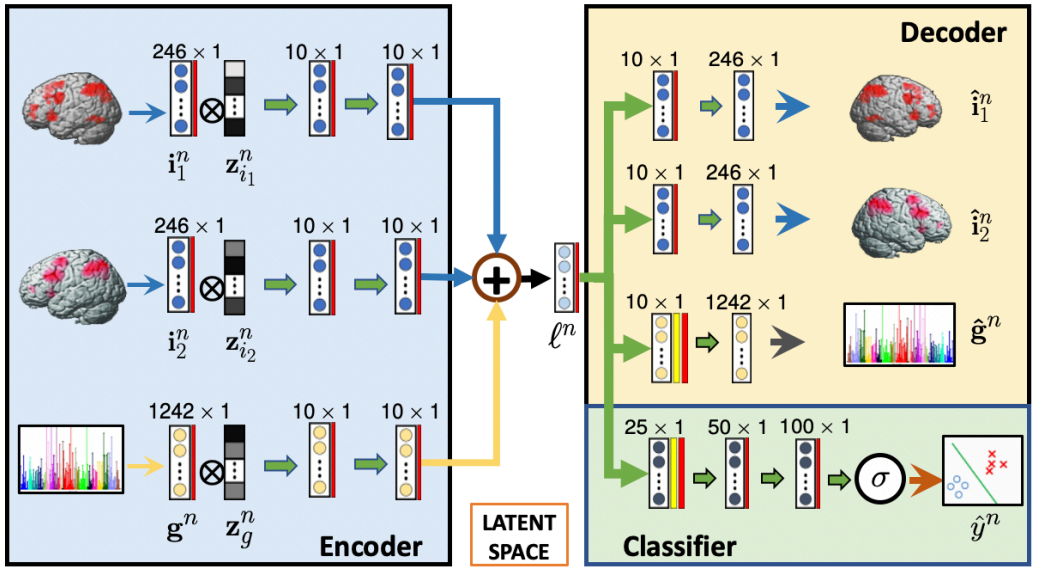
Acknowledgements

This work has generously been supported by

- **National Science Foundation** CRCNS award 182275 and CAREER award 1845430
- **Lieber Institute for Brain Development**
- **University of Bari Aldo Moro, Italy**

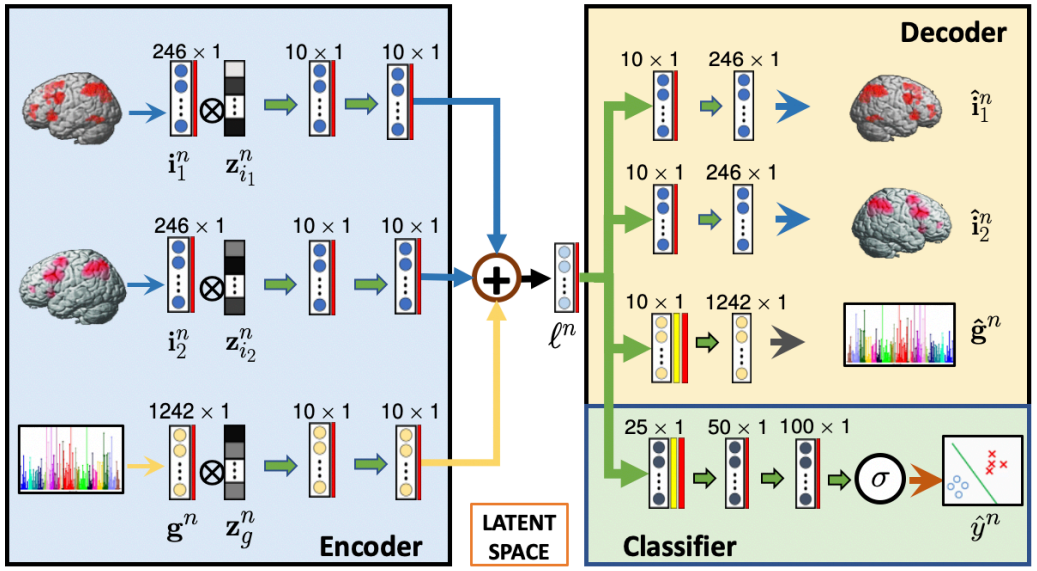
Contributions

- An End-to-End framework to integrate imaging and genetic data modalities



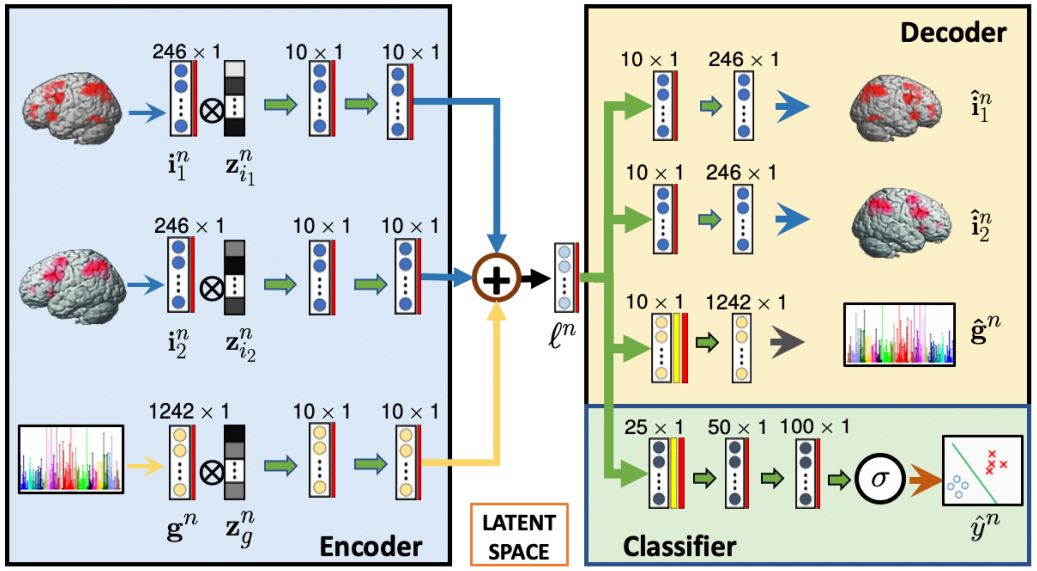
Contributions

- An End-to-End framework to integrate imaging and genetic data modalities
- It can identify imaging and genetic biomarkers via learnable dropout mask



Contributions

- An End-to-End framework to integrate imaging and genetic data modalities
- It can identify imaging and genetic biomarkers via learnable dropout mask
- The latent space captures an overcomplete representation of the data which provides robustness against missing data



Contributions

- An End-to-End framework to integrate imaging and genetic data modalities
- It can identify imaging and genetic biomarkers via learnable dropout mask
- The latent space captures an overcomplete representation of the data which provides robustness against missing data
- The cross-site generalization shows its ability to identify a robust set of biomarkers