Manifold Learning and Artificial Intelligence

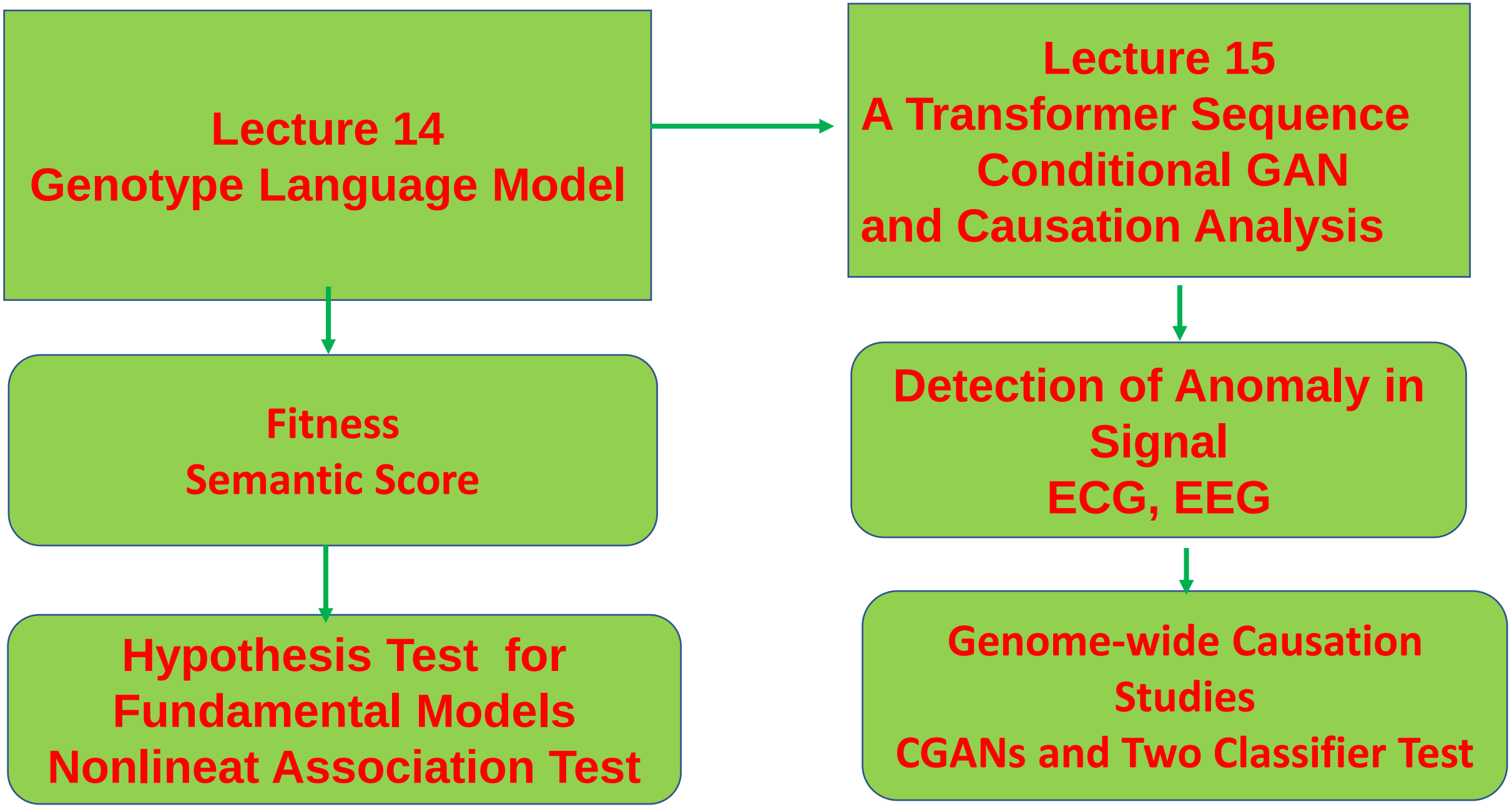
Lecture 14

A Fundamental Model for Genetic Studies of Complex Diseases

Momiao Xiong, University of Texas School of Public Health

- Time: 10:00 pm, US East Time, 05/6/2023
- 10:00 am, Beijing Time. 05/7/2023

Github Address: https://ai2healthcare.github.io/


Lecture 14
Genotype Language Model
Fitness
Semantic Score
Hypothesis Test for
Fundamental Models
Nonlineat Association Test
Lecture 15
A Transformer Sequence
Conditional GAN
and Causation Analysis
Detection of Anomaly in
Signal
ECG, EEG
Genome-wide Causation
Studies
CGANs and Two Classifier Test

14.1. Outlines

1. **Methods for analysis of mutation effect (using sequences only)**

- **Traditional, Embedding-Free method**
- **AI-based methods (Embedding-based):**
  - **(a) Natural language model, Transformer**
  - **(b) Variational autoencoder**
- **Genotype Language Model, Fundamental Model for Genetics**
  - **Token of genotypes**
  - **Architecture of Genotype Language Model**
  - **Loss Function and Training**
- **Score Function**
  - **Fitness**
  - **Semantic Score**

2. A General Framework for Hypothesis Testing in Fundamental Models

- **Null Hypothesis**
- **Test Statistics**
- **Distribution of Test Statistics**
- **Nonlinear Test**

References

Juan Rodriguez-Rivas et al. 2022. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. PNAS. 119 (4) e2113118119.

Hie et al. 2021, Learning the language of viral evolution and escape. Science, 371: 284-288

Gonzalo Benegas et al. 2023 (April 12). DNA language models are powerful zero-shot predictors of genome-wide variant effects. bioRxiv preprint doi: https://doi.org/10.1101/2022.08.22.504706.

Karim Beguir et al. 2023. Early computational detection of potential high-risk SARS-CoV-2 variants. Comput Biol Med. 2023 Mar;155:106618.

Xiaomin Li et al. 2022. TTS-CGAN: A Transformer Time-Series Conditional GAN for Biosignal Data Augmentation. arXiv:2206.13676 .

Zhao J, Boerwinkle E, Xiong MM. (2005) An entropy-based statistic for genome-wide association studies. Am J Hum Genet. 77:27-40.

Zhao J, Jin L, Xiong MM. (2006) Nonlinear tests for genome-wide association studies. Genetics. 174:1529-1538.

Goal of Analysis

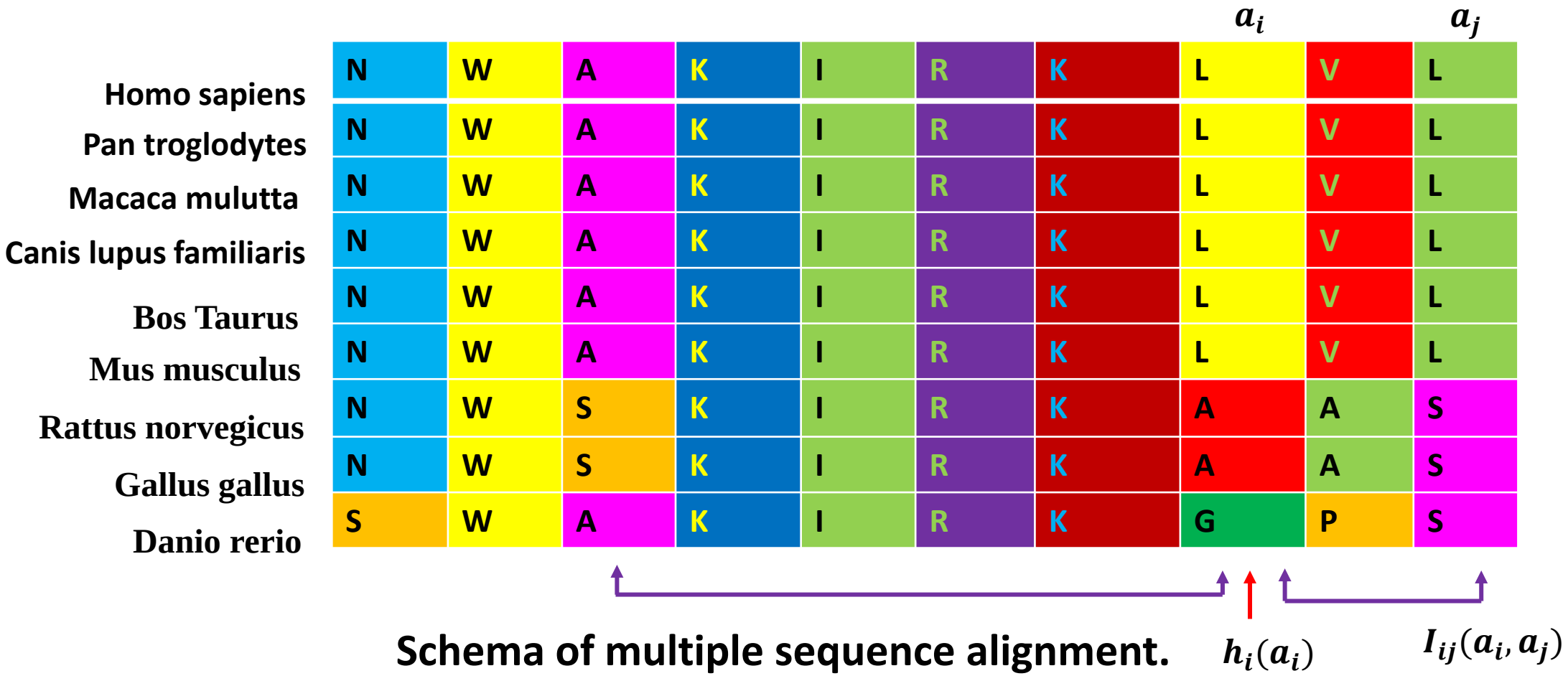
Intuitively, our goal is to identify mutations that induce high semantic change (e.g., a large impact on biological function) while being grammatically acceptable (e.g, biologically viable)

14.2. Mutability and Mutation Effect

- **Mutations**:
  point mutations, insertion/deletions, chromosome rearrangement
- **Mutation Phenotypes**: consequences of the mutation, functions .
- **Methods for mutation effect analysis:**

  Embedding Free, Multiple Sequence Alignment (MSA)
  Embedding-based

14.3. Embedding Free Methods (MSA)



Schema of multiple sequence alignment.

The Goal of MSA is to align the sequences which reflect evolutionary, functional, or structural relationship

14.3. 1. Mutation Effect Model

- **Independence**

  Assume an aminol acid (allele) sequence of length L:

  $a = a_1 \ldots a_L$.

Assume independence of a_i : The probability of a under independence model is

$\widehat{P}_{IND}(a_1 \ldots a_L) = \prod_{i=1}^{L} f_i(a_i)$,

where $f_i(a_i)$ is the empirical frequency of aminal acid (allele) a_i in the MSA.

The effect of an amino acid (Genotype) mutation $a_i \to b$ can be computed as

$$\Delta E_{IND}(i, b) = \log P_{IND}(a_1, \ldots a_i, \ldots, a_L) - \log P_{IND}(a_1, \ldots, b, \ldots, a_L) = \log f_i(a_i) - \log f_i(b)\ .$$

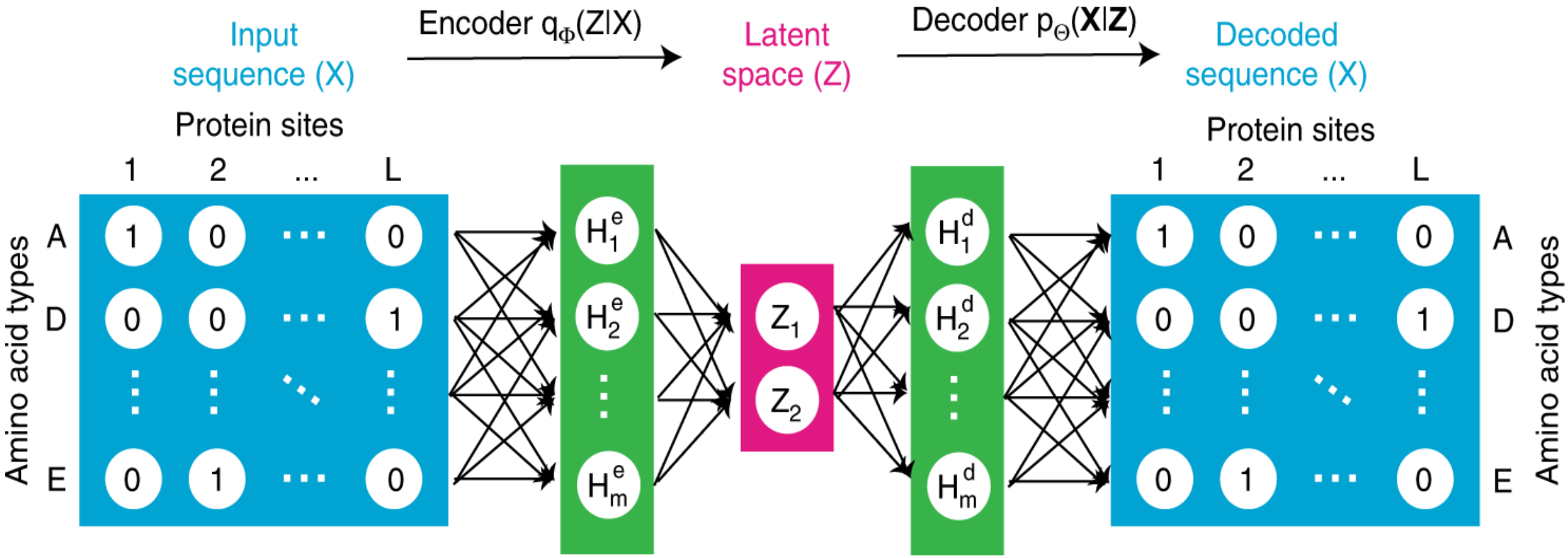
- **Epistatic Model**: **two-site coupling terms**

$$P_{DCA}(a_1, \ldots, a_L) = \frac{1}{Z} exp\left(\sum_{i=1}^{L} h(a_i) + \sum_{1 \le i < j \le L} J_{ij}(a_i, a_j)\right)$$

- **Calculation** :M. Ekeberg, T. Hartonen, E. Aurell, Fast pseudolikelihood maximization for directcoupling analysis of protein structure from many homologous amino-acid sequences. J. Comput. Phys. 276, 341–356 (2014).

MSA-based VAE

Deciphering protein evolution and fitness landscapes with latent space models
Nature Communications 10, 5644 (2019).



Deciphering protein evolution and fitness landscapes with latent space models

Code availability

The source code required to reproduce the results in this manuscript is freely available at https://github.com/xqding/PEVAE_Paper.

Suitable for complex disease

Methods: VAE

Demonstrate that latent space contain evolution information.

Let $S = (S_1, \ldots, S_L)$. Define a binary $21 \times L$ matrix X:

$$X_{ij} = \begin{cases} 1 & if\ S_j = i \\ 0 & otherwise \end{cases}$$

Let $P_\theta(X)$ be the marginal distribution of X.

Mutation Effect

Free Energy:

A free energy for a sequence $X(s)$ is defined as

$$\Delta G_{VAE}(x) = -\log P_\theta(x)$$

Mutation Effect:

The effect of mutations is defined as the changes in the free energy between a wide type sequence X and mutant sequence X'

Mutation effect $= \Delta G_{VAE}(X') - \Delta G_{VAE}(X)$.

$$\log P_\theta(X) = \log \int P_\theta(X,Z)dZ = \log \int q_\emptyset(Z|X)\frac{P_\theta(X,Z)}{q_\emptyset(Z|X)}dZ$$

Important Sampling

$$= \log E_{Z\sim q_\emptyset(Z|X)}\left[\frac{P_\theta(X,Z)}{q_\emptyset(Z|X)}\right] = \log \frac{1}{N}\sum_{i=1}^{N}\frac{P_\theta(X,Z^i)}{q_\emptyset(Z^i|X)}$$

$$Z^i \sim q_\emptyset(Z|X) \sim N(\mu, \Sigma)$$

$$\log P_\theta(X) \approx \sum_{i=1}^{N}\mathcal{L}\left(\theta, \emptyset, X^{(i)}\right)$$

$$p_\theta(X) = \int P_\theta(Z) P_\theta(X|Z) dZ$$

Encoder: $q_\emptyset(Z|X)$, Decoder: $p_\theta(X|Z)$

$log\, p_\theta(X) \geq \mathcal{L}(\theta, \emptyset, X)$, where

$$\mathcal{L}(\theta, \emptyset, X) = E_{q_\emptyset(Z|X)}[log\, p_\theta(X|Z)] - KL(q_\emptyset(Z|X) \| p_\theta(Z)) \ ,$$

Evidence Lower Bound (ELBO)

KL distance: $= E_{q_\emptyset(Z|X)}[\log \frac{q_\emptyset(Z|X)}{p_\theta(Z)}]$

The ELBO provides a general framework for VAE. The VAE consists of encoder and decoder. The posterior $q_\emptyset(Z|X)$ represents to encode the observed sequence X and maps variables X into latent variables Z and the conditional distribution $p_\theta(X|Z)$ represents to decode the latent variables Z back to the original variables X.

Maximizing ELBO to estimate the parameters $\theta, \emptyset$

14.4. Genotype Language Models

Intuitively, our goal is to identify variants that induce high semantic change (e.g., a large impact on biological function) while being grammatically acceptable (e.g, biologically viable)

14.4.1. Token

One genotype has two sites, each site has four letters: A, C, G, T.
Therefore, all possible number of genotypes is 4^2.

Figure 1 illustrates how a one-hot vector is used to tokenize each genotype

Consider a sequence of tokenized genotypes:

$$x = \begin{bmatrix} x_g^1 \\ \vdots \\ x_g^K \end{bmatrix}, x_g^i = one\ of\ x_{gj}^i, j = 1, \dots, 16.$$

$$\begin{bmatrix} AA \\ AC \\ CA \\ AG \\ GA \\ AT \\ TA \\ CC \\ CT \\ TC \\ CG \\ GC \\ TT \\ TG \\ GT \\ GG \end{bmatrix} \quad AA = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \cdots \quad GG = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Define a genotype vector:

$$g = \begin{bmatrix} g_1 \\ \vdots \\ g_K \end{bmatrix}$$

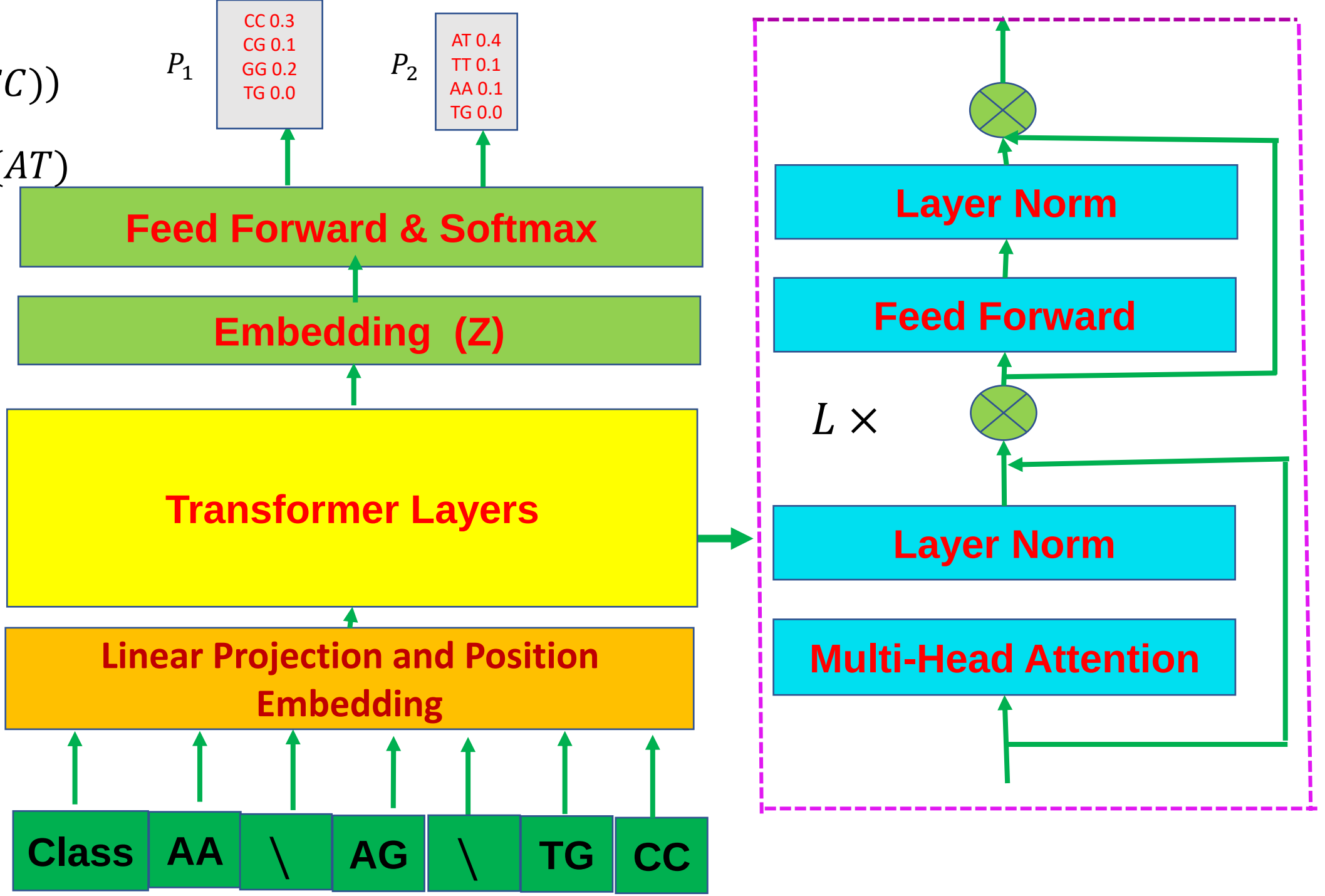
$$k = 16$$

Figure 1. Token for genotypes

$$L = \log P_1(CC))$$

$$+..+\log P_M(AT)$$

Loglikelihood over Masker positions



14.4.2. Model architecture

In this work, the input of the model consisted of the sequence characters corresponding to the genotype forming the variations in a gene. Each genotype is first tokenized, i.e., mapped to their index in the vocabulary containing the 16 genotypes, and then projected to an embedding space:

$$Z^0 = \begin{bmatrix} x_{class} \\ x_g^1 E_g \\ \vdots \\ x_g^K E_g \end{bmatrix} + E_{pos}, x_g^i \in R^m, E_g \in R^{m \times D}, E_{pos} \in R^{(K+1) \times D}, m = 16$$

The embeddings were then fed to the transformer model, consisting of a number of blocks, each composed of a self-attention operation followed by a position-wise multi-layer network.

Self-attention modules explicitly construct pairwise interactions between all positions in the sequence which enable them to build complex representations that incorporate context from across the sequence. A positional encoding must be added to the embedding of each token to distinguish its position in the sequence.

14.4.3. Training

- **Data Sources:**

(1) Allen Ancient DNA Resource (AADR):

https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data

On this page you can download a merged dataset consisting of genotypes for thousands of ancient and present-day individuals at up to 1.23 million positions in the genome (in hg19 coordinates).

(2) UK Biobank data

Genome-wide genotyping was performed on all UK Biobank participants using the UK Biobank Axiom Array. Approximately 850,000 variants were directly measured, with > 90million variants imputed using the Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels.

UK Biobank data access guide

https://biobank.ndph.ox.ac.uk/ukb/exinfo.cgi?src=AccessingData

(3) Database of Genotypes and Phenotypes (dbGaP)

https://ncbiinsights.ncbi.nlm.nih.gov/tag/dbgap/

dbGaP contains more than 500 NGS case–control studies

(4) The 1000 Genomes Project

(5) Genome of the Netherlands Consortium

DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task
Honglin Xiong et al. 2023.

14.4.4. Loss Function

Each input sequence was corrupted by replacing a fraction of the genotypes with a special mask token [MASK]. The network was then trained to predict the missing tokens [MASK] from the corrupted sequence.

for each sequence , we randomly sampled a set of indices x , for which the genotype tokens are replaced by a mask token, resulting in a corrupted sequence $\tilde{x}$. During pre-training, the set M was defined such that **15%** of the genotypes in the sequence are corrupted:

When corrupted, a genotype has **a 10% chance to be replaced** by another randomly selected genotype, an **80% chance of being masked** and **10% chance of being unchanged.**

During fine-tuning these probabilities do not change, however, only 2.5% of the genotypes in the sequence are corrupted.

Let M be the set of masked tokens, x be a sequence of genotypes and $\tilde{x}$ be a corrupted sequences by mask tokens.

Define **Output pf last layer in transformer.**

$$Z = [Z_{clas}, Z_1, \dots, Z_K], Z_i \in R^D, Z \in R^{D\times(K+1)}$$

$$a_i = WZ_i + b, W \in R^{m\times D}, b \in R^m, m = 16$$

$$a_i = \begin{bmatrix} a_i^1 \\ \vdots \\ a_i^{16} \end{bmatrix}, P_i = \begin{bmatrix} p_i^1 \\ \vdots \\ P_i^{16} \end{bmatrix}, \mathrm{i} \in M$$

$$P_\theta(x_i|\tilde{x}) = P_i = softmax(WZ_i + b) \qquad P_i^j = \frac{\exp(a_i^j)}{\sum_{j'=1}^{16}\exp(a_i^{j'})}$$

The training objective corresponds to the negative log-likelihood of the true sequence $x_i^j, i \in M$ $(j^{th}\ genotype)$ at the corrupted positions:

$$min_\theta\ L_\theta(\tilde{x}|x) = -\sum_{i\in M}\log P_\theta(x_i|\tilde{x}) \qquad (1)$$

14.4.5. Score Calculations (Inference)

- **Notations**

Once fine-tuned, the model was used to compute the semantic change and the log-likelihood to characterise gene.

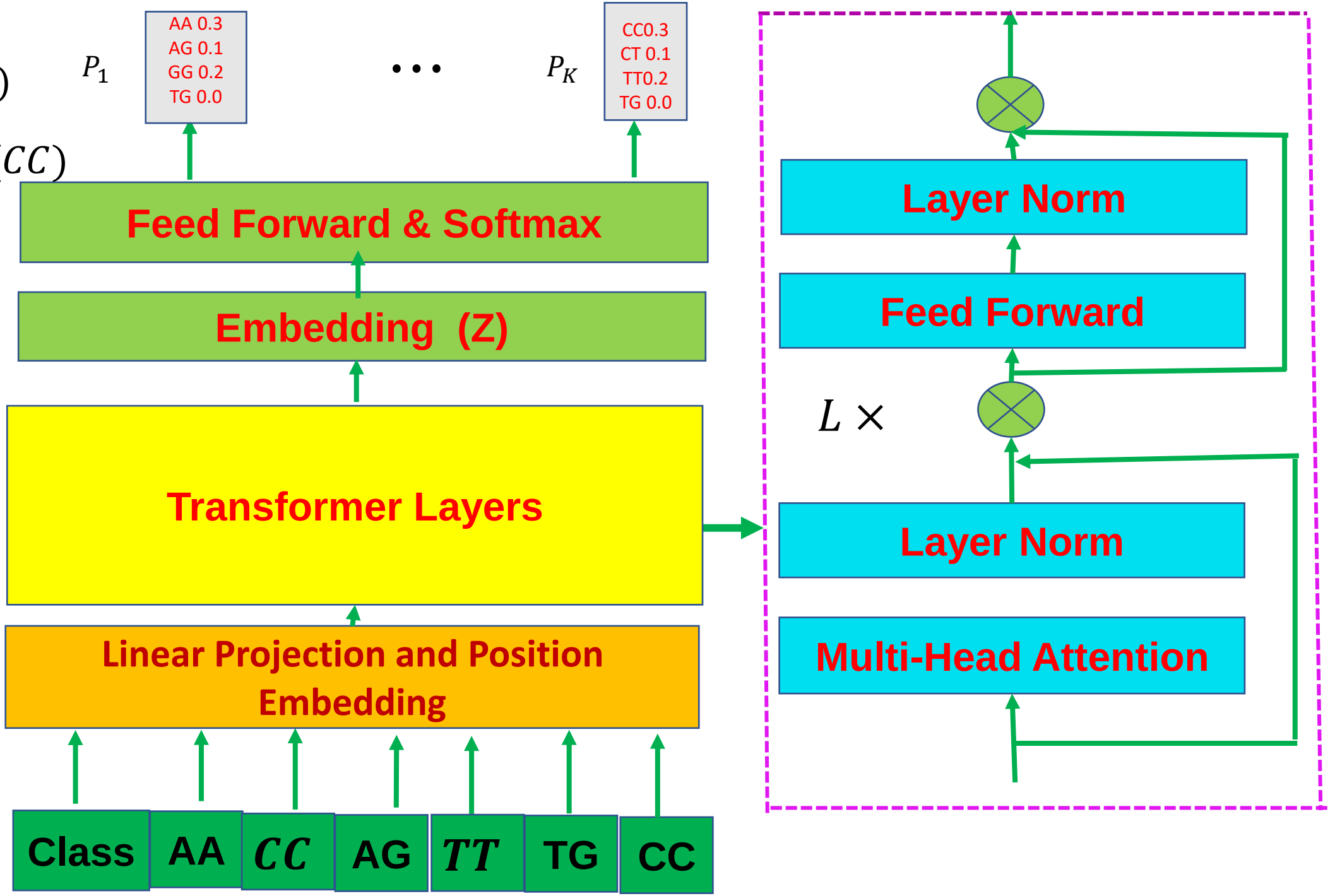
Formally, an input sequence was represented by a sequence of tokens defined as $x = \begin{bmatrix} x_g^1 \\ \vdots \\ x_g^K \end{bmatrix}$ where K is the number of tokens and $x_g^i \in \chi$ where χ that contains the genotype tokens and other tokens such as class and mask tokens. In this work, a class token was appended to all sequences before feeding them to the network, as such x_g^1 represents the class token, while $x_g^2, \ldots x_g^K$ represents the genotypes, or masked genotypes, in the gene. The sequence x is passed through attention layers.

Define $Z = (Z_1, \ldots, Z_K) as$ **the output of the last attention layer** where Z_i is the sequence embedding vector at position i .

$$L = \log P_1(AA)$$

$$+..+\log P_K(CC)$$

Loglikelihood over positions in the sequence.



- **Computing Fitness**

  Step 1: **Pretraining**

  Step 2: **Fine-tune**

  The last attention layer output Z is transformed by a feed-forward layer and a softmax activation into a vector of probabilities over tokens at each position

  $a_i = WZ_i + b, \qquad W \in R^{m \times D}, Z_i \in R^D, b \in R^m$ where m is the number of genotypes and some tokens,

  $P_i == softmax(WZ_i + b) = softmax(a_i)$

Fitness is defined as the log-likelihood of a variant $l(x)$ and is computed from these probabilities.

Let $\boldsymbol{l(x_n^i)}$ be a log probability of the individual n being the genotype $x_{j_n}^i$

at the i^{th} position of the genome, which is **defined as fitness ($E\left[l\left(x_n^i\right)\right] = V^{\prime}$**

This quantity $L(x_n^i)$ measures the likelihood of observing genotypes $x_{j_n}^i$ in the i^{th} position of the gene according to the model. It measures the fitness.

$$a^i = \begin{bmatrix} a_1^i \\ \vdots \\ a_m^i \end{bmatrix} = WZ^i + b, \qquad a_j^i = W_{j.}Z^i + b_j$$

$$P^i = \begin{bmatrix} P_1^i \\ \vdots \\ P_m^i \end{bmatrix} = softmax(a^i) = softmax(WZ^i + b), p_j^i = \frac{\exp(W_{j.}Z^i + b_j)}{\sum_{j'} \exp(W_{j'.}Z^i + b_{j'})}$$

$$\log P\left(x_n^i = x_{j_n}^i | x\right) = \log P_{j_n}^i = W_{j_n.}Z_n^i + b_{j_n} - \log \sum_{j_n'} \exp(W_{j_n'.}Z_n^i + b_{j_n'})$$

$$l(x^i) = \sum_{n=1}^{N} \log l(x_n^i) = \sum_{n=1}^{N} \log P\left(x_n^i = x_{j_n}^i | x\right)$$

$$= \sum_{n=1}^{N} \left[W_{j_n.}Z_n^i + b_{j_n} - \log \sum_{j_n'} \exp(W_{j_n'.}Z_n^i + b_{j_n'}) \right] = \sum_{n=1}^{N} W_{j_n.}Z_n^i + \sum_{n=1}^{N} b_{j_n} - \sum_{n=1}^{N} \log \sum_{j_n'} \exp(W_{j_n'.}Z_n^i + b_{j_n'})$$

14.4.6. Cases – Control Studies

- **Null Hypothesis:**

H_0: There is no difference in fitness between cases and controls. $\boldsymbol{H_0: V^A = V^C}$

H_a: Presence of difference in fitness between cases and controls. $\boldsymbol{H_a: V^A \neq V^C}$

- **Notations and Fitness in Cases and Controls.**

n_A: Number of cases
n_C: Number of controls
$l(x_n^i)$: fitness of the individual n in cases at the i^{th} position in a gene with a genotype $x_{j_n}^i$.
$l(y_n^i)$: fitness of the individual n in controls at the i^{th} position in a gene with a genotype $y_{j_n}^i$.

$$l\left(x_n^i\right) = \log P\left(x_n^i = x_{j_n}^i | x\right) = W_{j_n.}^A Z_{An}^i + b_{j_n}^A - \log \sum_{j_n'} \exp(W_{j_n'.}^A Z_{An}^i + b_{j'n}^A)$$

$$l\left(y_n^i\right) = \log P\left(y_n^i = y_{j_n}^i | y\right) = W_{j_n.}^C Z_{Cn}^i + b_{j_n}^C - \log \sum_{j_n'} \exp(W_{j_n'.}^C Z_{Cn}^i + b_{j_n'}^C)$$

- **Define average fitness in cases and controls for marker *i*:**

$$\bar{l}_A = \frac{1}{n_A} \sum_{n=1}^{n_A} l\left(x_n^i\right)$$

$$\bar{l}_C = \frac{1}{n_C} \sum_{n=1}^{n_C} l\left(y_n^i\right)$$

- **Define covariance matrix under the null hypothesis:**

$$\Lambda = var(\bar{l}_A - \bar{l}_C) = \frac{1}{n_A}\text{var}(l\left(x_n^i\right)) + \frac{1}{n_C}\text{var}(l\left(y_n^i\right)) = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\sigma^2$$

$$\hat{\sigma}^2 = \text{var}(l\left(x_n^i\right) = \text{var}(l\left(y_n^i\right) = \frac{1}{n_A + n_C - 2}\left[\sum_{n=1}^{n_A}\left(l\left(x_n^i\right) - \bar{l}_A\right)^2 + \sum_{n=1}^{n_C}\left(l\left(y_n^i\right) - \bar{l}_C\right)^2\right]$$

- **Association Tests**

Single Marker

$$T_s = \frac{n_A n_C}{n_A + n_C} \frac{\left(\bar{l}_A - \bar{l}_C\right)^2}{\hat{\sigma}^2} \qquad \bar{l}_A \sim N\left(V^A, \frac{1}{n_A}\hat{\sigma}^2\right), \bar{l}_C \sim N(V^C, \frac{1}{n_C}\hat{\sigma}^2) \qquad \textbf{(3)}$$

Distribution

Under the null hypothesis $T_s \sim \chi^2_{(1)}$

Multiple Markers oa a Gene

Define

$$l(x_n) = \begin{bmatrix} l(x_n^1) \\ \vdots \\ l(x_n^K) \end{bmatrix}, l(y_n) = \begin{bmatrix} l(y_n^1) \\ \vdots \\ l(y_n^K) \end{bmatrix}, \bar{l}_A = \begin{bmatrix} \bar{l}_A^1 \\ \vdots \\ \bar{l}_A^K \end{bmatrix}, \bar{l}_C = \begin{bmatrix} \bar{l}_C^1 \\ \vdots \\ \bar{l}_C^K \end{bmatrix}$$

- **Estimation of Covariance Matrix**

$$\xi = \bar{l}_A - \bar{l}_C, \Lambda = cov(\xi, \xi) = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\Sigma \qquad \hat{\Lambda} = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)S$$

$$S = \frac{1}{n_A + n_C - 2}\left[\sum_{n=1}^{n_A}\left(l(x_n) - \bar{l}_A\right)\left(l(x_n) - \bar{l}_A\right)^T + \left(l(y_n) - \bar{l}_C\right)\left(l(y_n) - \bar{l}_C\right)^T\right]$$

- **Test Association**

$$V^A = E[l(x_n)] = \begin{bmatrix} V_1^A \\ \vdots \\ V_K^A \end{bmatrix}, V^C = E[l(y_n)] = \begin{bmatrix} V_1^A \\ \vdots \\ V_K^A \end{bmatrix}$$

$$T_M = \left(\bar{l}_A - \bar{l}_C\right)^T \hat{\Lambda}^{-1}\left(\bar{l}_A - \bar{l}_C\right)$$

(4

$$\bar{l}_A \sim N\left(V^A, \frac{1}{n_A}\Sigma\right), \bar{l}_C \sim N\left(V^C, \frac{1}{n_C}\Sigma\right), \bar{l}_A - \bar{l}_C \sim N(0, \Lambda)$$

Under the null hypothesis, $T_M \sim \chi^2_{(K)}$

- **QTL**

$$y_n = \mu + \sum_{i=1}^{K} l\left(x_n^i\right)\beta_n + \varepsilon_n, n = 1, \dots, N$$

y_n: A quantitative trait of individual i

14.4.7. Semantic Embedding and Mutation Effect

Notations

$Z_{An}^{i} \in R^{H}$: Embedding vector of individual n in cases with genotype in position i

$Z_{Cn}^{i} \in R^{H}$: Embedding vector of individual n in controls with genotype in position i

$$\bar{Z}_{A}^{i} = \frac{1}{n_A}\sum_{n=1}^{n_A} Z_{An}^{i},\ \bar{Z}_{C}^{i} = \frac{1}{n_C}\sum_{n=1}^{n_C} Z_{Cn}^{i} \qquad \mu_A = \begin{bmatrix} \mu_A^1 \\ \vdots \\ \mu_A^H \end{bmatrix} = E\left[Z_{An}^{i}\right], \mu_C = \begin{bmatrix} \mu_C^1 \\ \vdots \\ \mu_C^H \end{bmatrix}$$

$$\xi = \bar{Z}_{A}^{i} - \bar{Z}_{C}^{i}, Var(\xi) = \Lambda = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\Sigma \qquad \Sigma = Cov(Z_{An}^{i}, Z_{An}^{i})$$

$$\widehat{\Lambda} = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\mathrm{S}, \mathrm{S} = \frac{1}{n_A + n_C - 2}\left[\sum_{n=1}^{n_A}\left(Z_{An}^{i} - \bar{Z}_{A}^{i}\right)\left(Z_{An}^{i} - \bar{Z}_{A}^{i}\right)^{T} + \sum_{n=1}^{n_C}\left(Z_{Cn}^{i} - \bar{Z}_{C}^{i}\right)\left(Z_{Cn}^{i} - \bar{Z}_{C}^{i}\right)^{T}\right]$$

- **Test Statistic**

Single Marker

- **Null Hypothesis** $H_0: \mu_A = \mu_C$

H_0: There is no difference in embedding of genotype in position i between cases and controls

H_a: Presence of difference in embedding of genotype in position i between cases and controls

$$T_s = \left(\bar{Z}_A^i - \bar{Z}_C^i\right)^T \hat{\Lambda}^{-1} \left(\bar{Z}_A^i - \bar{Z}_C^i\right)$$

$$\bar{Z}_A^i \sim N\left(\mu_A, \frac{1}{n_A}\Sigma\right), \bar{Z}_C^i \sim N\left(\mu_C, \frac{1}{n_C}\Sigma\right)$$

Under the null hypothesis, $T_s \sim \chi^2_{(H)}$

Multiple Markers or a Gene

$$\bar{Z}_{An} = \frac{1}{K-1}\sum_{i=2}^{K} Z_{An}^i, \bar{Z}_{Cn} = \frac{1}{K-1}\sum_{i=2}^{K} Z_{Cn}^i$$

$$\bar{Z}_A = \frac{1}{n_A}\sum_{n=1}^{n_A} \bar{Z}_{An}, \qquad \bar{Z}_C = \frac{1}{n_C}\sum_{n=1}^{n_C} \bar{Z}_{cn}$$

$$Var(\bar{Z}_A - \bar{Z}_C) = \Lambda = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\Sigma$$

$$\widehat{\Lambda} = \left(\frac{1}{n_A} + \frac{1}{n_C}\right)\mathrm{S},$$

$$\mathrm{S} = \frac{1}{n_A + n_C - 2}\left[\sum_{n=1}^{n_A}(\bar{Z}_{An} - \bar{Z}_A)(\bar{Z}_{An} - \bar{Z}_A)^T + \sum_{n=1}^{n_C}(\bar{Z}_{Cn} - \bar{Z}_C)(\bar{Z}_{Cn} - \bar{Z}_C)^T\right]$$

- **Null Hypothesis**

H_0: There is no difference in the total embedding of the genotype in a genomic region between cases and controls.

H_a: Presence of difference in the total embeddings of the genotypes in a genomic region between cases and controls.

$$\Sigma_A = Cov(Z^i_{An}, Z^i_{An}), \Sigma_C = Cov(Z^i_{Cn}, Z^i_{Cn})$$

Define test statistics

$$\boldsymbol{T_m = (\bar{Z}_A - \bar{Z}_C)^T \widehat{\Lambda}^{-1}(\bar{Z}_A - \bar{Z}_C)} \qquad \bar{Z}_A \sim N\left(\mu_A, \frac{1}{n_A}\Sigma_A\right), \bar{Z}_C \sim N(\mu_C, \frac{1}{n_C}\Sigma_C)$$

Under the null hypothesis, $T_m \sim \chi^2_{(H)}$

Justification for Tests

$X^A, \mu_{0,}^A \Sigma_0^A$
X^C, μ_0^C, Σ_0^C

$$Z^A \approx f(\mu_0^A) + B(X^A - \mu_0^A) \qquad Z^A, \mu_E^A, \Sigma_E^A$$

$$Z^C \approx f(\mu_0^C) + D(X^C - \mu_0^C) \qquad Z^C, \mu_E^C, \Sigma_E^C$$

$$\mu_E^A = E[Z^A] \approx f(\mu_0^A), \Sigma_E^A = Cov(Z^A) \approx B\Sigma_0^A B^T, B = \frac{\partial f}{\partial (X^A)^T}$$

$$\mu_E^C = E[Z^C] \approx, f(\mu_0^C), \Sigma_E^C = Cov(Z^C) \approx D\Sigma_0^C D^T, D = \frac{\partial f}{\partial (X^C)^T}$$

$$\bar{Z}^A \sim N\left(\mu_E^A, \frac{1}{n_A}\Sigma_E^A\right), Z^C \sim N\left(\mu_E^C, \frac{1}{n_C}\Sigma_E^C\right), T_E = \left(\bar{Z}^A - \bar{Z}^C\right)^T \Lambda^{-1}\left(\bar{Z}^A - \bar{Z}^C\right)$$

$$\Lambda = \frac{1}{n_A} B\Sigma_0^A B^T + \frac{1}{n_C} D\Sigma_0^C D^T, \text{ Under } H_0, T_E \sim \chi^2_{(H)}$$

Power Calculation of the Nonlinear Tests

- Under alternative hypothesis H_a,

$$T_E = \left(\bar{Z}^A - \bar{Z}^C\right)^T \Lambda^{-1} \left(\bar{Z}^A - \bar{Z}^C\right) \sim \text{Noncentral } \chi^2_{(H)} \text{ with}$$

Noncentrality λ

$$\lambda = \left(f\left(\mu_0^A\right) - f\left(\mu_0^C\right)\right)^T \Lambda^{-1} \left(f\left(\mu_0^A\right) - f\left(\mu_0^C\right)\right)$$

$$f\left(\mu_0^A\right) - f\left(\mu_0^C\right) \approx D\left(\mu_0^A - \mu_0^C\right) + \frac{1}{2}\begin{bmatrix} \left(\mu_0^A - \mu_0^C\right)^T H_1 \left(\mu_0^A - \mu_0^C\right) \\ \vdots \\ \left(\mu_0^A - \mu_0^C\right)^T H_H \left(\mu_0^A - \mu_0^C\right) \end{bmatrix}$$

Justification for Tests

- **Transformer models can universally approximate arbitrary continuous sequence-to-sequence functions**

  Yun et al. 2020, ARE TRANSFORMERS UNIVERSAL APPROXIMATORS OF SEQUENCE-TO-SEQUENCE FUNCTIONS?

  APPROXIMATION ABILITY OF TRANSFORMER NETWORKS FOR FUNCTIONS WITH VARIOUS SMOOTHNESS OF BESOV SPACES: ERROR ANALYSIS AND TOKEN EXTRACTION, ICLR 2023.

  Shi et al. 2021; SparseBERT: Rethinking the Importance Analysis in Self-attention

- **Embedding is a nonlinear sequence-to-sequence function. Hypothesis testing on embedding is a nonlinear hypothesis test.**

  Zhao J, Jin L, Xiong MM. (2006) Nonlinear tests for genome-wide association studies. Genetics. 174:1529-1538.