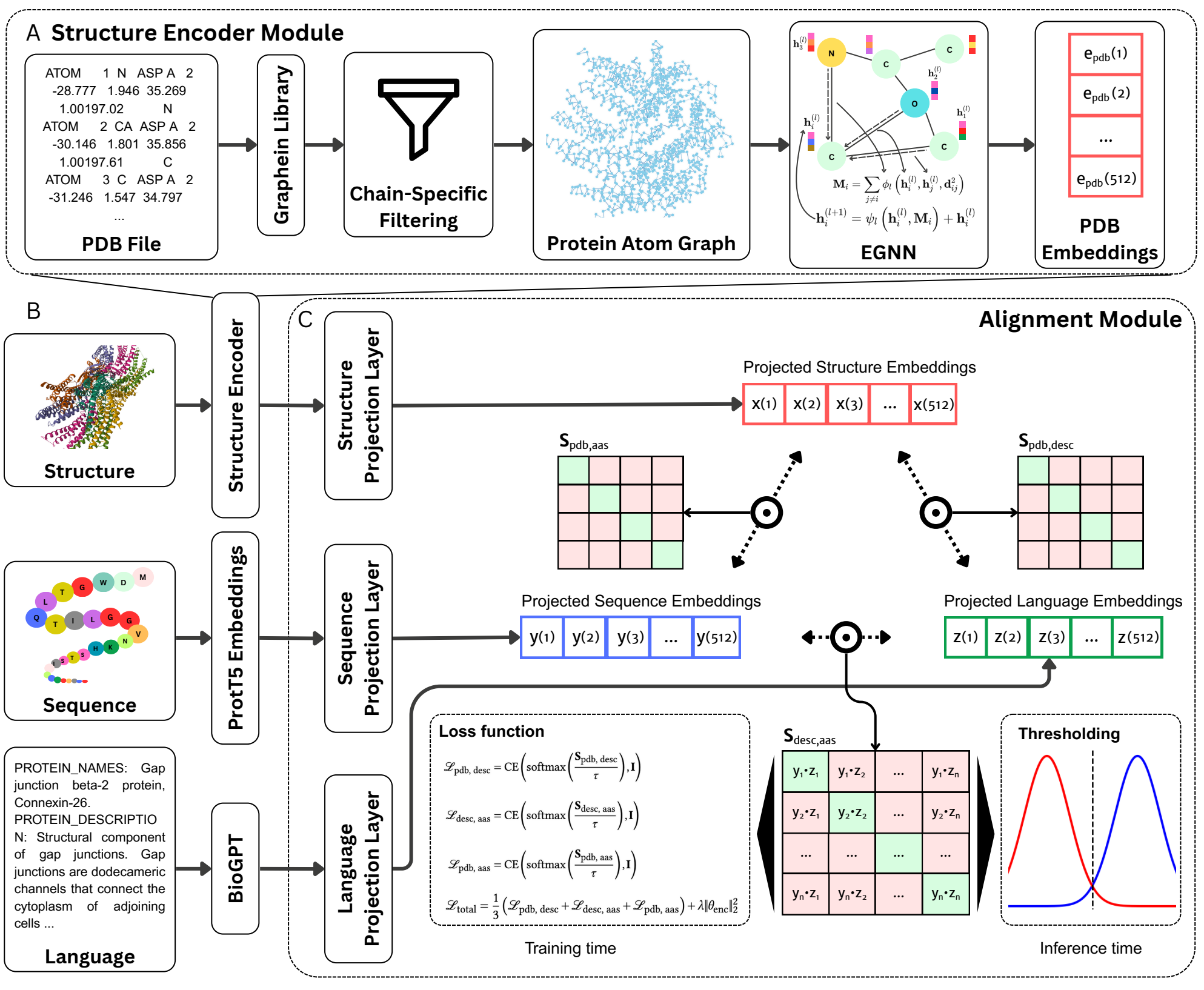
A Structure Encoder Module
ATOM 1 N ASP A 2
-28.777 1.946 35.269
1.00197.02 N
ATOM 2 CA ASP A 2
-30.146 1.801 35.856
1.00197.61 C
ATOM 3 C ASP A 2
-31.246 1.547 34.797
...
PDB File
Graphein Library
Chain-Specific Filtering
Protein Atom Graph
EGNN
$\mathbf{M}_i = \sum_{j \neq i} \phi_l\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{d}_{ij}^2\right)$
$\mathbf{h}_i^{(l+1)} = \psi_l\left(\mathbf{h}_i^{(l)}, \mathbf{M}_i\right) + \mathbf{h}_i^{(l)}$
$e_{pdb}(1)$
$e_{pdb}(2)$
...
$e_{pdb}(512)$
PDB Embeddings
B
Structure
Sequence
PROTEIN_NAMES: Gap junction beta-2 protein, Connexin-26.
PROTEIN_DESCRIPTIO
N: Structural component of gap junctions. Gap junctions are dodecameric channels that connect the cytoplasm of adjoining cells ...
Language
Structure Encoder
ProtT5 Embeddings
BioGPT
C
Alignment Module
Structure Projection Layer
Sequence Projection Layer
Language Projection Layer
Projected Structure Embeddings
x(1) x(2) x(3) ... x(512)
$S_{pdb,aas}$
$S_{pdb,desc}$
Projected Sequence Embeddings
y(1) y(2) y(3) ... y(512)
Projected Language Embeddings
z(1) z(2) z(3) ... z(512)
Loss function
$\mathcal{L}_{pdb,\,desc} = CE\left(\text{softmax}\left(\frac{S_{pdb,\,desc}}{\tau}\right), I\right)$
$\mathcal{L}_{desc,\,aas} = CE\left(\text{softmax}\left(\frac{S_{desc,\,aas}}{\tau}\right), I\right)$
$\mathcal{L}_{pdb,\,aas} = CE\left(\text{softmax}\left(\frac{S_{pdb,\,aas}}{\tau}\right), I\right)$
$\mathcal{L}_{total} = \frac{1}{3}\left(\mathcal{L}_{pdb,\,desc} + \mathcal{L}_{desc,\,aas} + \mathcal{L}_{pdb,\,aas}\right) + \lambda\|\theta_{enc}\|_2^2$
Training time
$S_{desc,aas}$
$y_1 \cdot z_1$ $y_1 \cdot z_2$ ... $y_1 \cdot z_n$
$y_2 \cdot z_1$ $y_2 \cdot z_2$ ... $y_2 \cdot z_n$
... ...
$y_n \cdot z_1$ $y_n \cdot z_2$ ... $y_n \cdot z_n$
Thresholding
Inference time