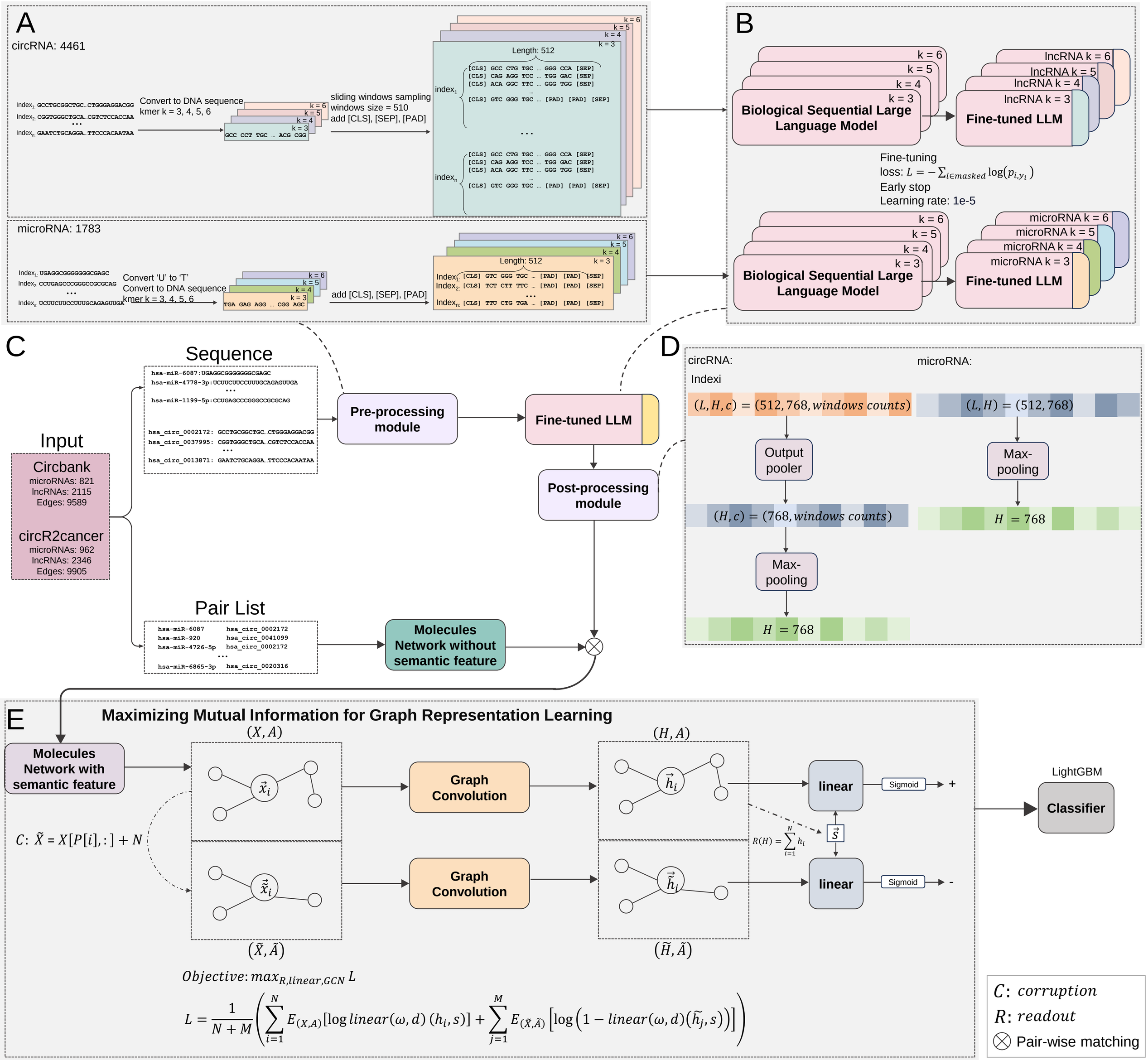
A
circRNA: 4461
Convert to DNA sequence
kmer k = 3, 4, 5, 6
sliding windows sampling
windows size = 510
add [CLS], [SEP], [PAD]
Length: 512
microRNA: 1783
Convert 'U' to 'T'
Convert to DNA sequence
kmer k = 3, 4, 5, 6
add [CLS], [SEP], [PAD]
B
Biological Sequential Large Language Model
Fine-tuned LLM
lncRNA k = 6
lncRNA k = 5
lncRNA k = 4
lncRNA k = 3
Fine-tuning
loss: $L = -\sum_{i \in masked} \log(p_{i,y_i})$
Early stop
Learning rate: 1e-5
microRNA k = 6
microRNA k = 5
microRNA k = 4
microRNA k = 3
C
Sequence
Input
Circbank
microRNAs: 821
lncRNAs: 2115
Edges: 9589
circR2cancer
microRNAs: 962
lncRNAs: 2346
Edges: 9905
Pre-processing module
Fine-tuned LLM
Post-processing module
Pair List
Molecules Network without semantic feature
D
circRNA:
Indexi
microRNA:
$(L, H, c) = (512, 768, windows\ counts)$
$(L, H) = (512, 768)$
Output pooler
Max-pooling
$(H, c) = (768, windows\ counts)$
$H = 768$
Max-pooling
$H = 768$
E
Maximizing Mutual Information for Graph Representation Learning
Molecules Network with semantic feature
(X, A)
C: $\tilde{X} = X[P[i], :] + N$
Graph Convolution
Graph Convolution
(H, A)
$(\tilde{X}, \tilde{A})$
$(\tilde{H}, \tilde{A})$
$R(H) = \sum_{i=1}^{N} h_i$
linear
linear
Sigmoid
Sigmoid
+
-
LightGBM
Classifier
$Objective: max_{R,linear,GCN}\ L$
$L = \frac{1}{N+M}\left(\sum_{i=1}^{N} E_{(X,A)}[\log linear(\omega, d)(h_i, s)] + \sum_{j=1}^{M} E_{(\tilde{X},\tilde{A})}\left[\log\left(1 - linear(\omega, d)(\tilde{h}_j, s)\right)\right]\right)$
C: corruption
R: readout
Pair-wise matching