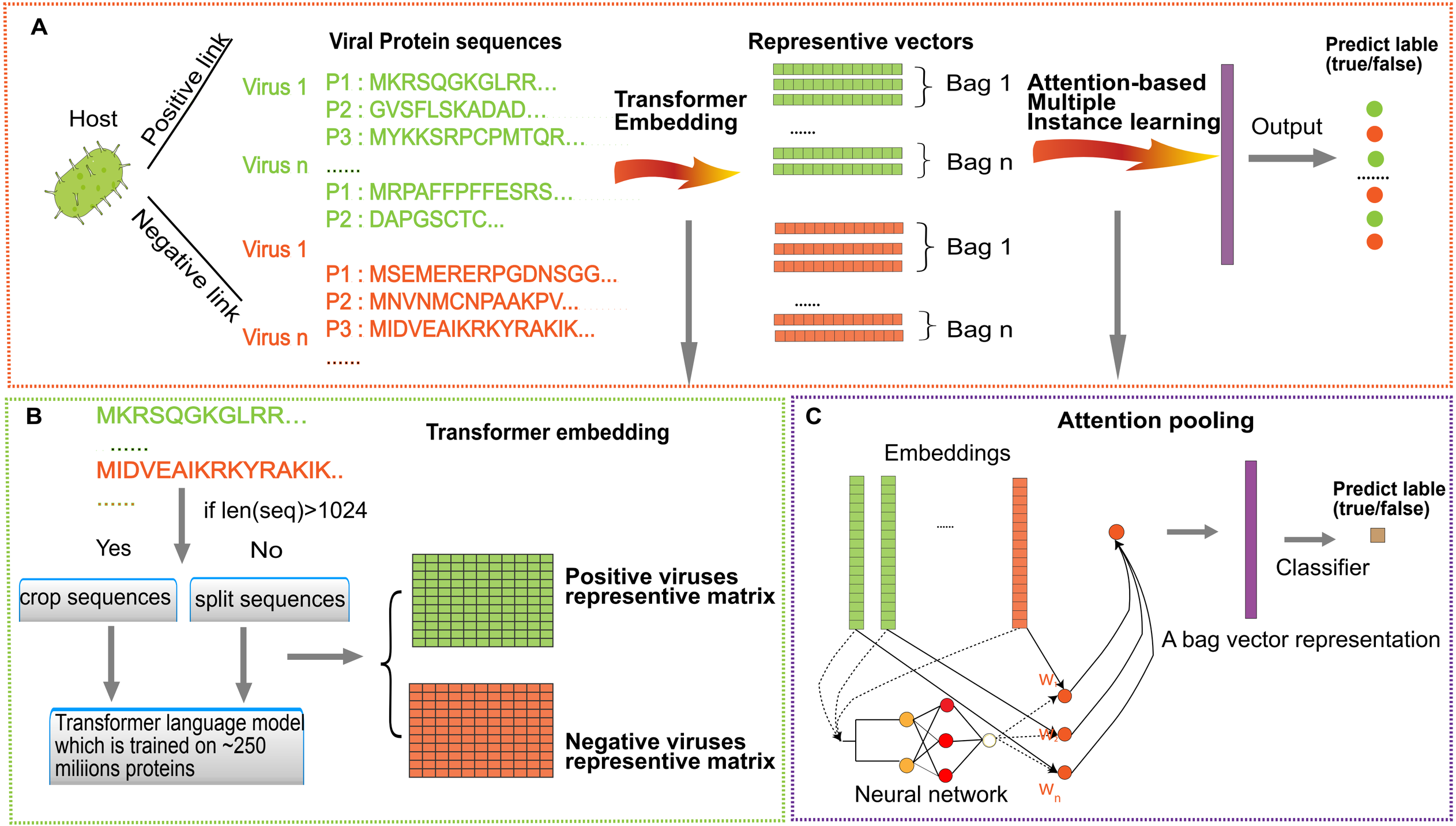
A
Host
Positive link
Negative link
Viral Protein sequences
Virus 1
P1 : MKRSQGKGLRR…
P2 : GVSFLSKADAD…
P3 : MYKKSRPCPMTQR…
Virus n
.....
P1 : MRPAFFPFFESRS…
P2 : DAPGSCTC...
Virus 1
P1 : MSEMERERPGDNSGG...
P2 : MNVNMCNPAAKPV...
P3 : MIDVEAIKRKYRAKIK...
Virus n
.....
Transformer Embedding
Representive vectors
Bag 1
.....
Bag n
Bag 1
.....
Bag n
Attention-based Multiple Instance learning
Output
Predict lable (true/false)
.....
B
MKRSQGKGLRR…
.....
MIDVEAIKRKYRAKIK..
.....
if len(seq)>1024
Yes
No
crop sequences
split sequences
Transformer language model which is trained on ~250 miliions proteins
Transformer embedding
Positive viruses representive matrix
Negative viruses representive matrix
C
Attention pooling
Embeddings
.....
Neural network
W_1
W_2
W_n
A bag vector representation
Classifier
Predict lable (true/false)