Language Models in Biology

Biological sequences as text

DNA, RNA, Protein sequences → Text format

Tokenization strategies

K-mers, BPE, Character-level encoding

Pretraining objectives

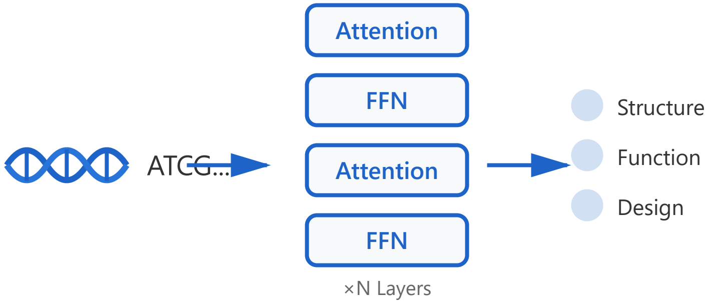
Masked LM, Next token prediction, Contrastive

Scale effects

Model size vs. performance trade-offs

Downstream tasks

Structure, Function, Design applications


ATCG...
Attention
FFN
Attention
FFN
×N Layers
Structure
Function
Design