Language Models in Biology

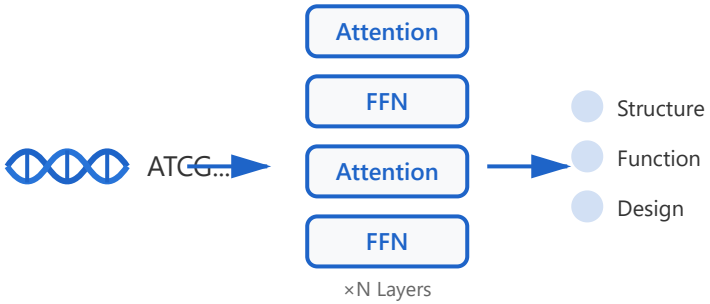
Biological sequences as text
DNA, RNA, Protein sequences → Text format

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



1. Biological Sequences as Text

DNA Sequences

DNA consists of four nucleotide bases (A, T, G, C) that can be naturally represented as text strings, making them directly compatible with language model architectures.

```
Original: ATCGATCGTAGCTAGCTA
Tokenized: A T C G A T C G T A G C T A G C T A
```

Protein Sequences

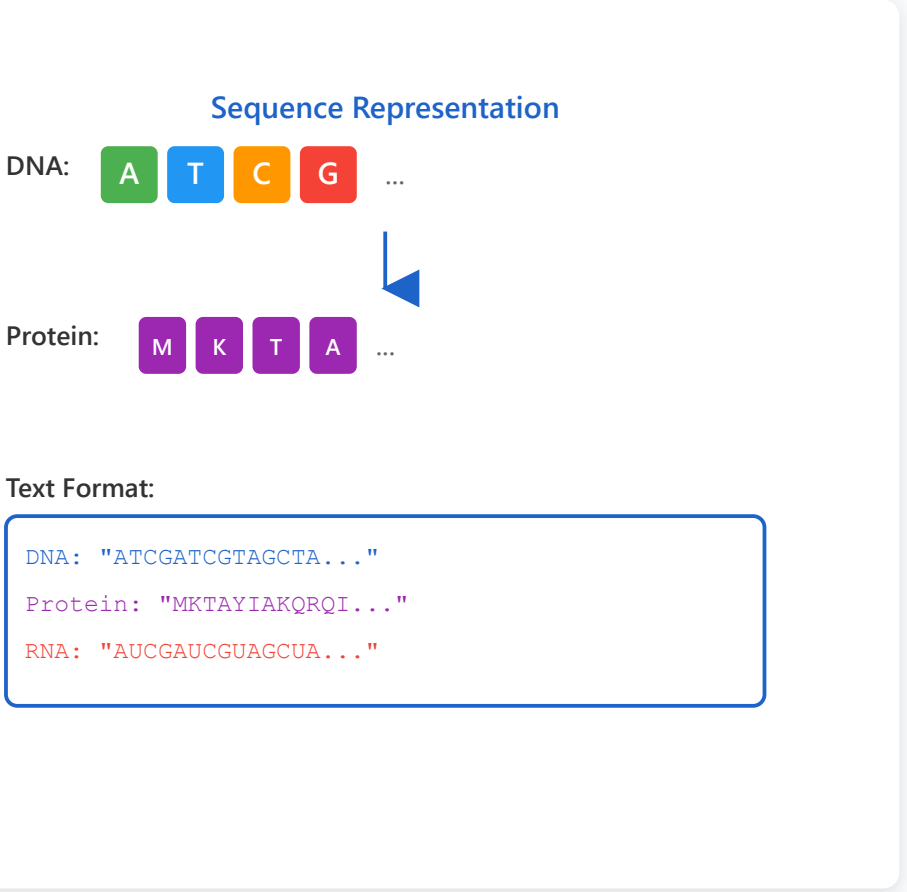
Proteins use 20 amino acids represented by single-letter codes, creating a natural alphabet for language modeling similar to human language.

```
Original: MKTAYIAKQRQISFVKSH
Vocab size: 20 amino acids + special tokens
```

RNA Sequences

RNA sequences (A, U, G, C) can be treated similarly to DNA, with additional structural information from secondary structure annotations.

- Fixed vocabulary size makes tokenization straightforward
- Sequential nature enables transfer of NLP techniques
- Enables pre-training on massive unlabeled sequence databases



2. Tokenization Strategies

Character-level Encoding

Each nucleotide or amino acid is treated as a single token. Simple and direct, but may miss important patterns spanning multiple positions.

```
Input: ATCGATCG
Tokens: [A][T][C][G][A][T][C][G]
```

K-mer Tokenization

Sequences are split into overlapping or non-overlapping subsequences of length k. Captures local patterns and motifs effectively.

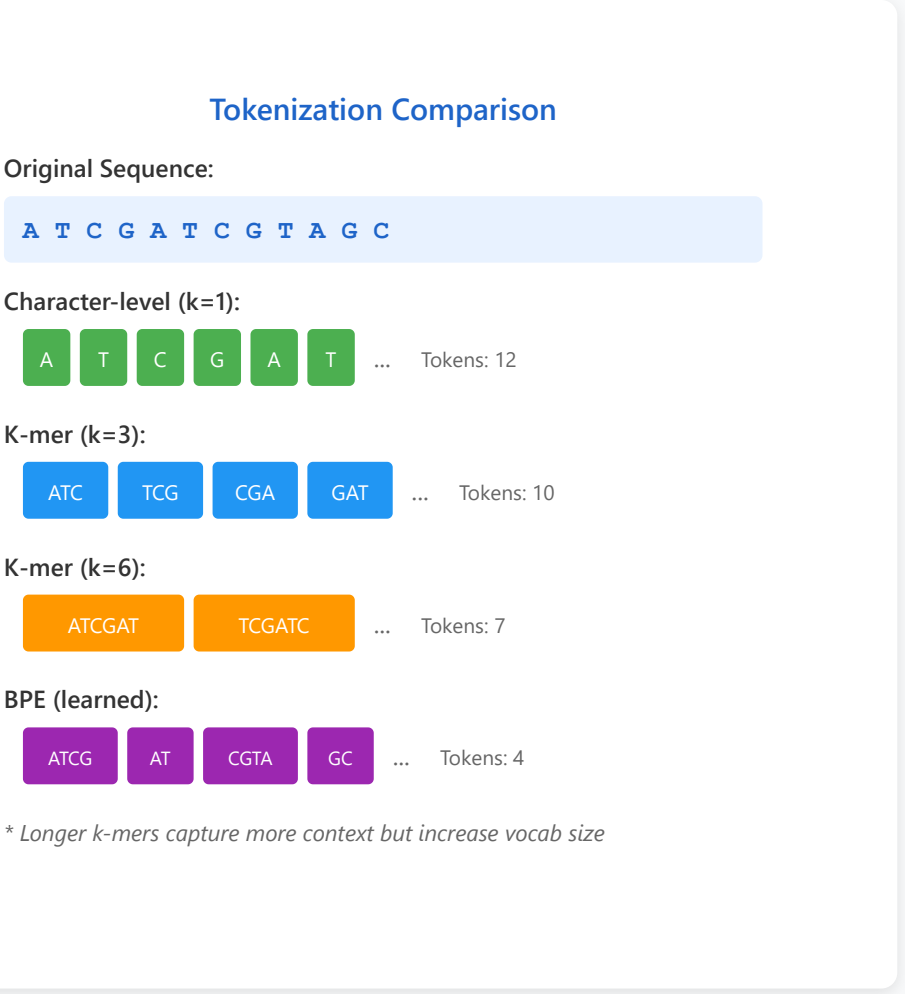
```
Input: ATCGATCG (k=3)
Tokens: [ATC][TCG][CGA][GAT][ATC][TCG]
```

Byte Pair Encoding (BPE)

Data-driven approach that learns common subword units from the training corpus. Balances vocabulary size with sequence length.

```
Learns frequent patterns:
"ATG" → start codon
"TAA" → stop codon
```

- Choice affects model's ability to capture biological motifs
- K-mer size impacts computational efficiency and context
- BPE can discover biologically meaningful units



3. Pretraining Objectives

Masked Language Modeling (MLM)

Random tokens are masked, and the model learns to predict them using bidirectional context. Similar to BERT, enables learning rich representations.

```
Input: ATCG[MASK]TCGTA
Target: Predict 'A' using context
Model: ESM, ProtBERT
```

Next Token Prediction

Autoregressive training where the model predicts the next token given all previous tokens. Similar to GPT architecture, useful for generation tasks.

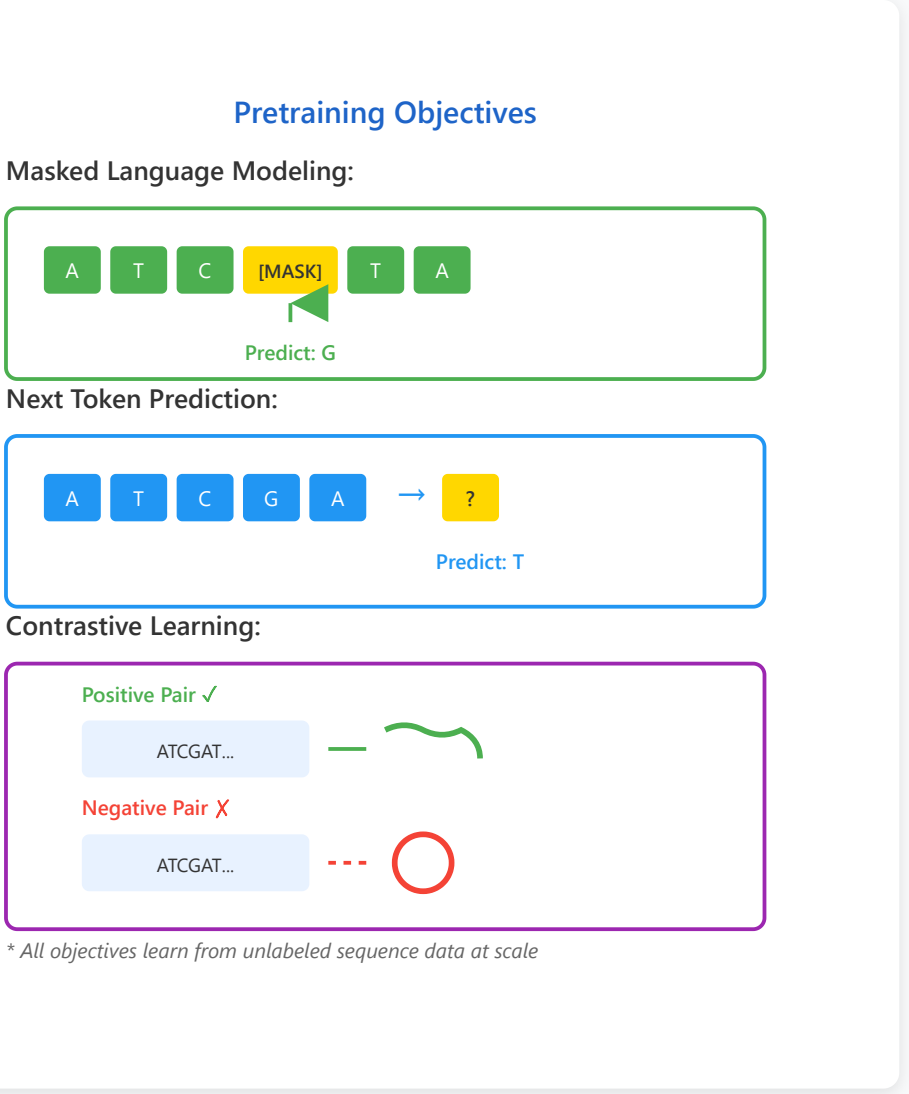
```
Input: ATCGATCG
Target: Predict next 'T'
Model: ProGen, ProtGPT2
```

Contrastive Learning

Learns by contrasting positive pairs (e.g., sequence and structure) against negative pairs. Effective for multimodal alignment.

```
Positive: (Sequence, Structure)
Negative: (Sequence, Random Structure)
Model: ESM-IF, ProteinCLIP
```

- MLM: Best for understanding tasks (classification, prediction)
- Next token: Best for generation and design tasks
- Contrastive: Best for multimodal tasks and alignment



** All objectives learn from unlabeled sequence data at scale*

4. Scale Effects

Model Size Scaling

Larger models (more parameters) generally achieve better performance on downstream tasks, following similar scaling laws as in natural language models.

```
ESM-2: 8M → 150M → 650M → 3B → 15B params
Performance improves consistently with size
```

Data Scaling

Training on larger sequence databases (UniProt, GenBank) provides richer representations. Models benefit from evolutionary diversity in training data.

```
ESM-2: Trained on 250M+ sequences
ProtT5: Trained on UniRef50 (45M sequences)
```

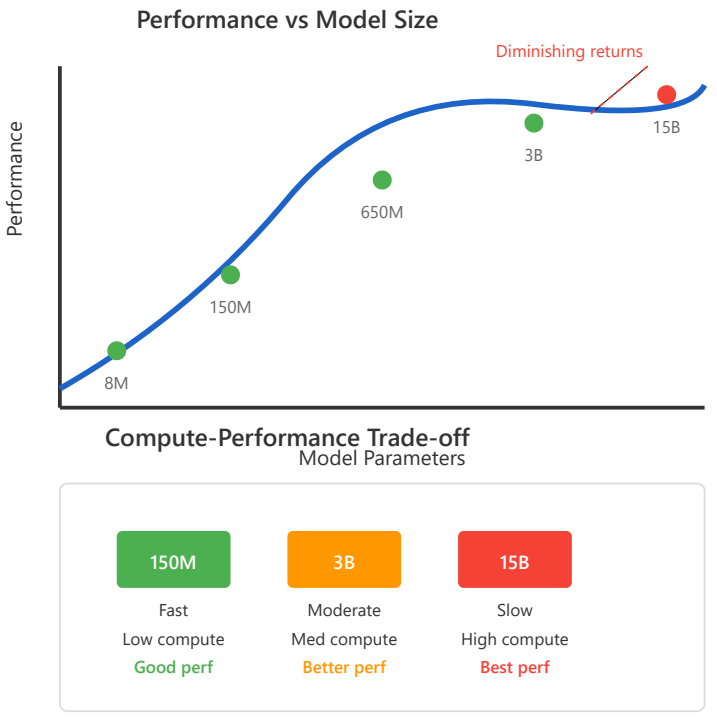
Compute-Performance Trade-offs

Larger models require more computational resources but provide diminishing returns. Need to balance accuracy gains with practical deployment constraints.

```
15B model: 100x compute of 150M model
Performance gain: ~15-20% on benchmarks
```

- Scaling laws similar to NLP models apply to biological sequences
- Emergent capabilities appear at certain scale thresholds
- Model selection depends on task complexity and resources

Scaling Effects



5. Downstream Tasks

Structure Prediction

Predicting 3D protein structures from sequences. Models learn structural constraints from sequence patterns. AlphaFold2 and ESMFold achieve near-experimental accuracy.

```
Input: Protein sequence
Output: 3D coordinates of all atoms
Applications: Drug design, protein engineering
```

Function Prediction

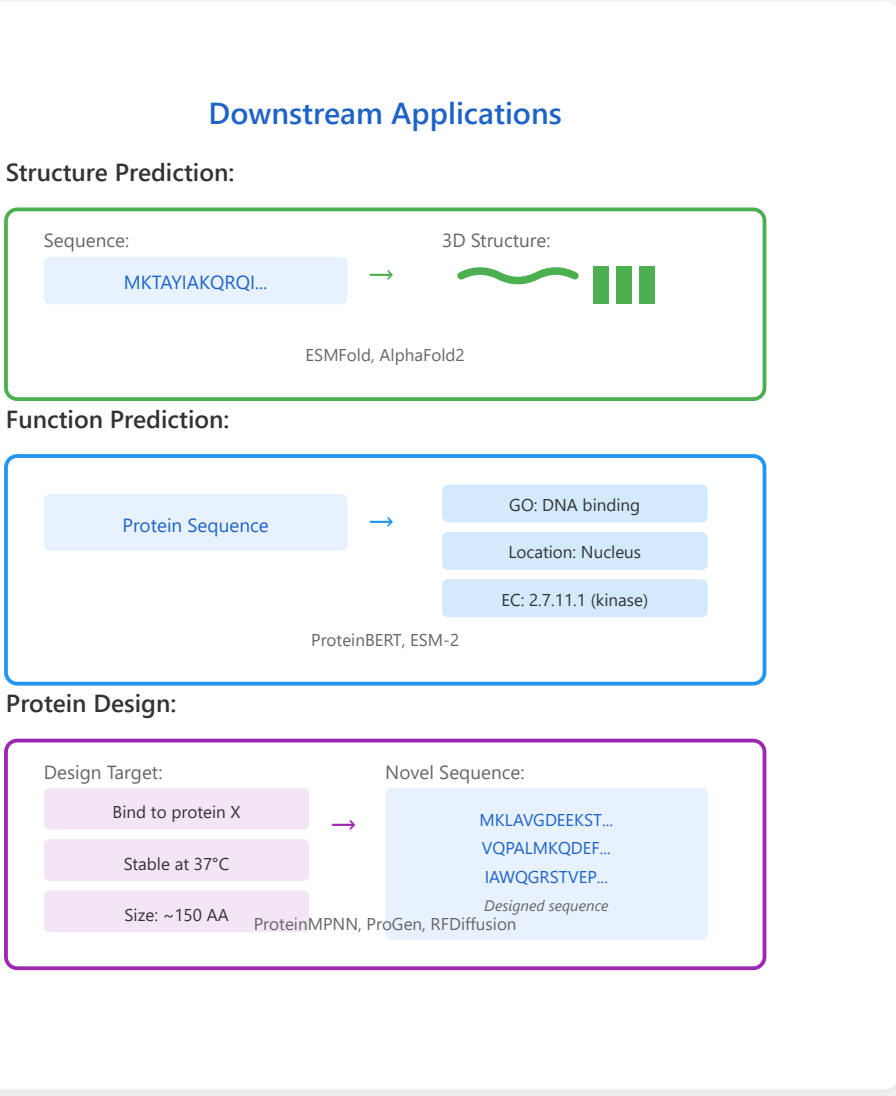
Predicting protein functions, subcellular localization, interactions, and enzymatic activity from sequence representations.

```
Tasks: GO term prediction, EC number
Active site identification
Protein-protein interaction prediction
```

Protein Design

Generating novel sequences with desired properties. Includes de novo design, optimization of existing proteins, and inverse folding (structure to sequence).

```
Input: Desired function/structure
Output: Novel protein sequence
Models: ProteinMPNN, ESM-IF, RFDiffusion
```



- Fine-tuning pretrained models dramatically improves performance
- Zero-shot capabilities emerge from large-scale pretraining

- Multimodal models combine sequence, structure, and function