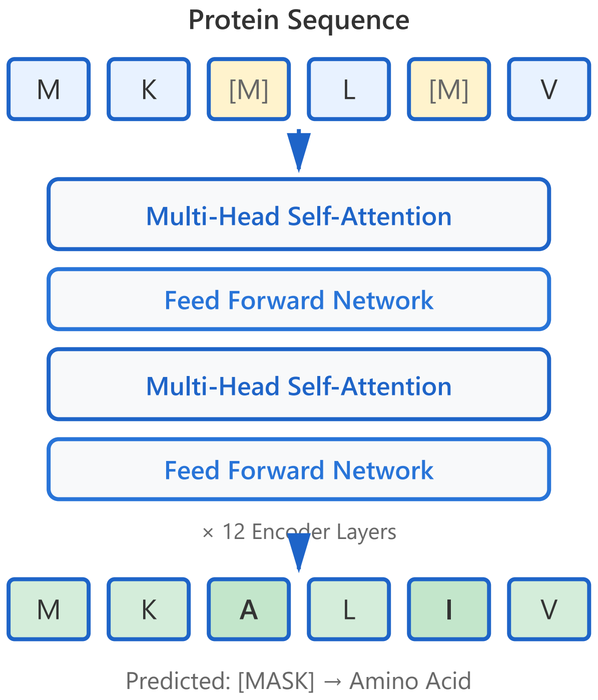
BERT for Proteins



ProtBERT architecture

12-layer bidirectional encoder

Masked language modeling

15% random masking strategy

Attention patterns

Learns residue interactions

Structural insights

Captures 3D contact maps

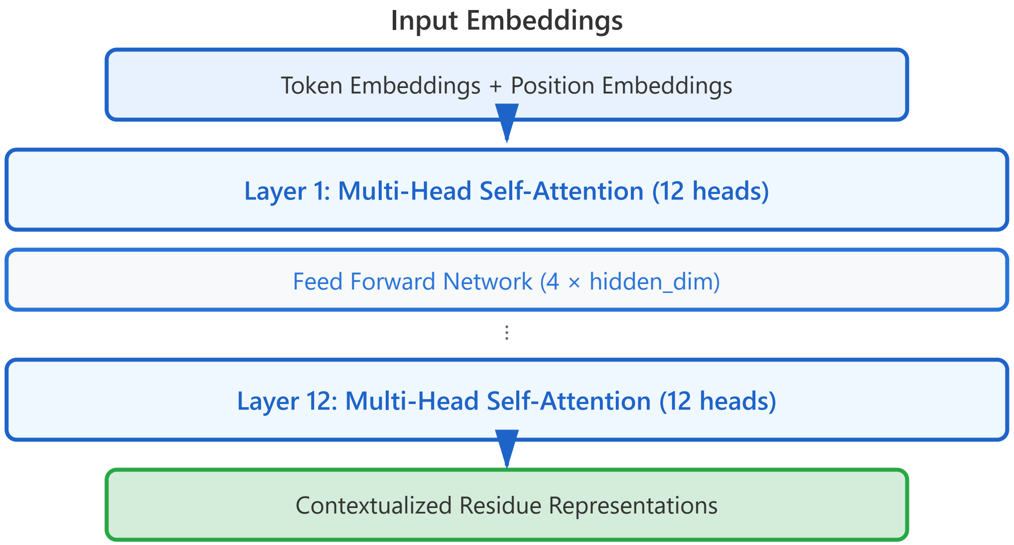
Function prediction

GO terms, EC numbers

ProtBERT Architecture

Architecture Overview

ProtBERT is based on the BERT architecture adapted for protein sequences. It employs a 12-layer bidirectional Transformer encoder that processes protein sequences to generate contextual embeddings for each amino acid residue.



Input Embeddings

Token Embeddings + Position Embeddings

Layer 1: Multi-Head Self-Attention (12 heads)

Feed Forward Network (4 × hidden_dim)

⋮

Layer 12: Multi-Head Self-Attention (12 heads)

Contextualized Residue Representations

Masked Language Modeling

Training Objective

ProtBERT is trained using Masked Language Modeling (MLM), where 15% of amino acids in the input sequence are randomly masked, and the model learns to predict the original amino acids based on bidirectional context.

Original:	M	K	L	F	V	A	I	G
Masked:	M	[M]	L	F	[M]	A	I	[M]
Predicted:	M	**K**	L	F	**V**	A	I	**G**

Bold predictions indicate correctly predicted masked positions

Masking Strategy (15% of tokens)

- **80%:** Replace with [MASK] token
- **10%:** Replace with random amino acid

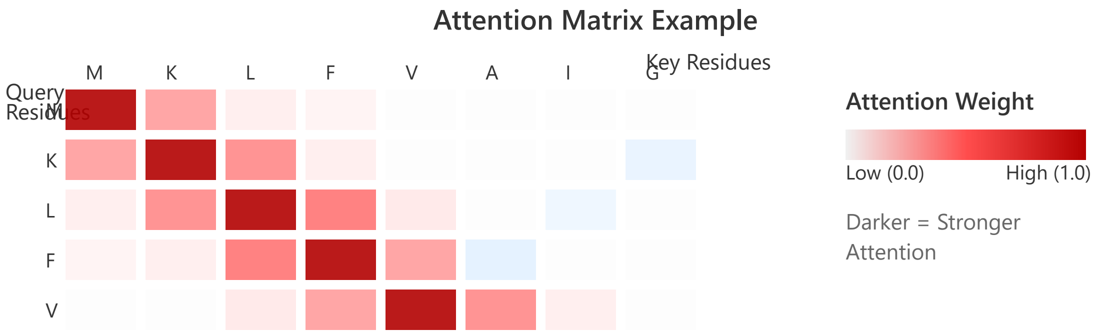
Training Data

- **UniRef100:** 217 million sequences
- **BFD:** 2.5 billion sequences

Attention Patterns

Learning Residue Interactions

The multi-head self-attention mechanism in ProtBERT learns to capture complex dependencies between amino acids. Different attention heads specialize in different types of interactions, from local sequential patterns to long-range contacts.



Attention Matrix Example

Darker = Stronger Attention

Types of Patterns Learned

- **Local patterns:** Adjacent residue correlations (α-helices, β-sheets)

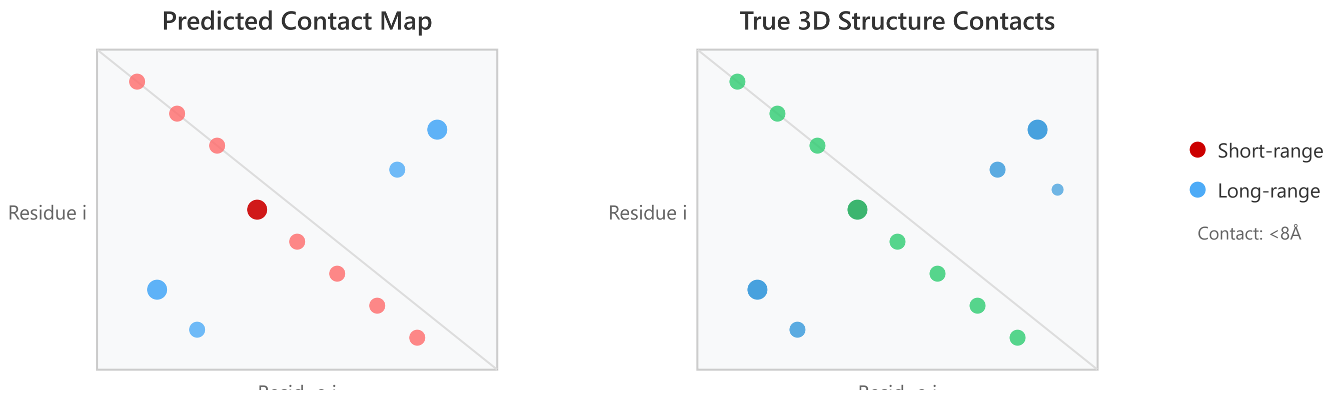
Head Specialization

- **Heads 1-4:** Focus on local sequence context

Structural Insights

Capturing 3D Contact Maps

ProtBERT learns implicit structural information from sequence data alone. The attention weights in deeper layers show strong correlation with actual 3D contacts in protein structures, enabling structure prediction tasks without explicit structural training.



Structural Features Learned

- **Secondary structure:** α-helices, β-sheets, loops with 85%+ accuracy

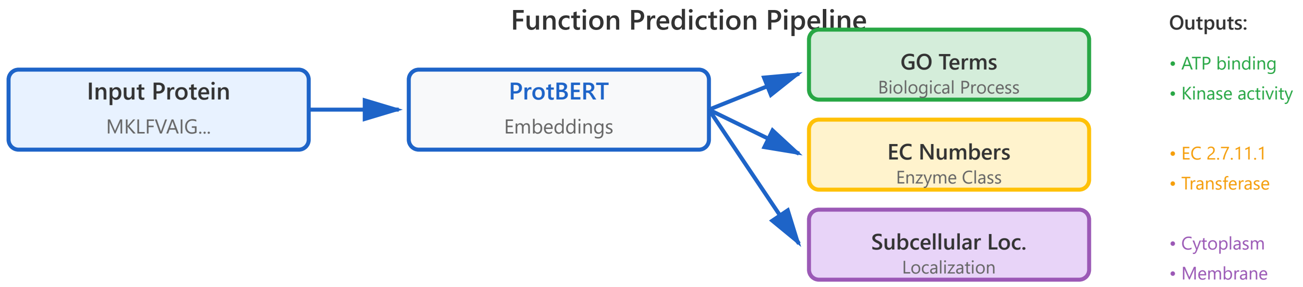
Applications

- **Structure prediction:** Input features for AlphaFold-like models
- **Protein design:** Guide mutations to maintain structure

Function Prediction

Predicting Protein Function

ProtBERT representations can be fine-tuned or used as features for various protein function prediction tasks, including Gene Ontology (GO) term annotation and Enzyme Commission (EC) number classification.



Function Prediction Pipeline

Input Protein
MKLFVAIG...

ProtBERT
Embeddings

GO Terms
Biological Process

EC Numbers
Enzyme Class

Subcellular Loc.
Localization

Outputs:

- ATP binding
- Kinase activity

- EC 2.7.11.1
- Transferase

- Cytoplasm
- Membrane

State-of-the-Art Performance

GO Term Prediction: Fmax = 0.51 | EC Prediction: Accuracy = 89.2% | Localization: Accuracy = 87.5%

Gene Ontology (GO) Terms

GO annotations describe protein functions in three categories:

Enzyme Commission (EC) Numbers

EC numbers classify enzyme functions hierarchically: