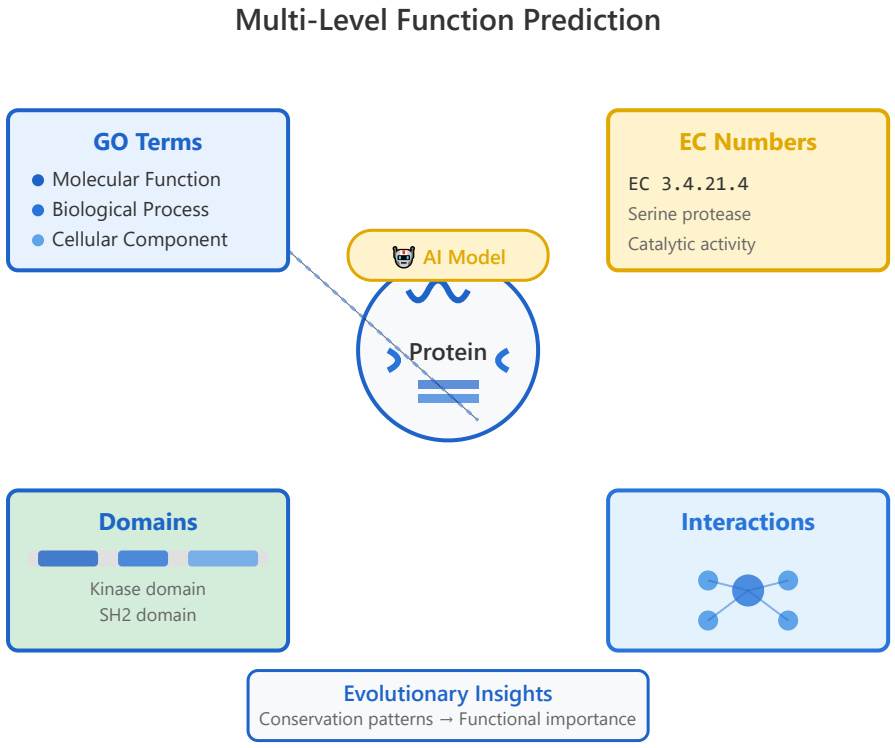
Protein Function Prediction



GO term prediction
Molecular/biological/cellular

EC number classification
Enzyme commission numbers

Domain annotation
Functional regions identification

Interaction prediction
Protein-protein networks

Evolutionary insights
Conservation-function mapping

1 GO Term Prediction

What are GO Terms?

Gene Ontology (GO) terms provide a standardized vocabulary to describe protein functions across three main domains. They enable consistent annotation of protein characteristics across different organisms and databases.

Three GO Categories:

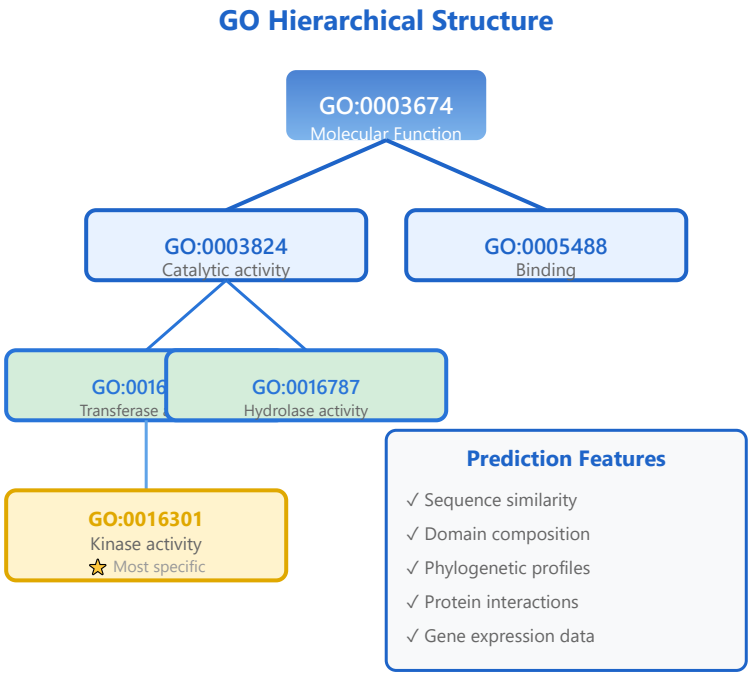
- **Molecular Function (MF):** Activities at the molecular level (e.g., catalytic activity, binding)
- **Biological Process (BP):** Larger processes accomplished by multiple molecular activities (e.g., signal transduction, metabolism)
- **Cellular Component (CC):** Location where a gene product is active (e.g., nucleus, mitochondrion)

Example: Protein Kinase Annotation

GO:0004672 - protein kinase activity

GO:0006468 - protein phosphorylation

GO:0005737 - cytoplasm



2 EC Number Classification

Enzyme Commission (EC) Numbers

EC numbers provide a hierarchical classification system for enzymes based on the chemical reactions they catalyze. Each EC number consists of four digits that progressively specify the enzyme's function.

EC Number Structure (EC a.b.c.d):

- **First digit (a):** Main enzyme class (1-7)
- **Second digit (b):** Subclass (substrate type)
- **Third digit (c):** Sub-subclass (specific substrate)
- **Fourth digit (d):** Serial number (specific enzyme)

Example: Trypsin Classification

```
EC 3.4.21.4 | | | └─ Trypsin (specific enzyme) | | └──── Serine endopeptidase | └────── Acting on peptide bonds └──────── Hydrolase
```

EC Classification System

Seven Main Enzyme Classes

- EC 1 - Oxidoreductases
- EC 2 - Transferases
- EC 3 - Hydrolases
- EC 4 - Lyases
- EC 5 - Isomerases
- EC 6 - Ligases
- EC 7 - Translocases

Detailed Example: Protein Kinase

EC 2.7.11.1 - Non-specific serine/threonine protein kinase

Reaction catalyzed:

`ATP + protein → ADP + phosphoprotein`

Substrate specificity:

Serine or threonine residues on target proteins

Cofactor requirement: Mg^{2+} or Mn^{2+}

Prediction Approaches

Modern EC number prediction combines sequence-based methods with structural information and substrate specificity patterns. Machine learning models trained on annotated enzyme databases can predict EC numbers with high accuracy.

3 Domain Annotation

What are Protein Domains?

Protein domains are distinct structural and functional units within a protein sequence. They are evolutionarily conserved regions that can fold independently and often retain function even when separated from the rest of the protein.

Key Characteristics:

- **Modularity:** Can be mixed and matched in different proteins
- **Conservation:** Similar domains found across different species
- **Function:** Each domain typically has a specific function
- **Independence:** Can fold and function independently

Common Domain Databases

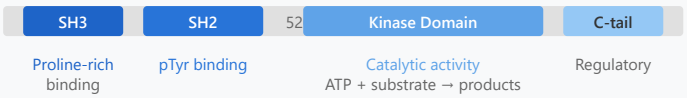
Pfam: Protein families database with HMM profiles

InterPro: Integrated resource of protein families

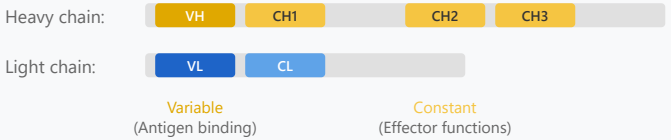
SMART: Simple Modular Architecture Research Tool

Multi-Domain Protein Architecture

Example 1: Src Kinase (c-Src)



Example 2: Immunoglobulin (Antibody)



Domain Prediction Methods

- **Sequence-based:** HMM profiles, pattern matching
- **Structure-based:** Fold recognition, structural alignment
- **Machine learning:** Deep learning models (e.g., AlphaFold)
- **Integrative:** Combining multiple evidence sources
- **Context-aware:** Domain co-occurrence patterns

4 Protein Interaction Prediction

Protein-Protein Interactions (PPIs)

Proteins rarely act alone in cells. Understanding how proteins interact with each other is crucial for deciphering cellular mechanisms, signaling pathways, and disease processes. Computational prediction of PPIs helps identify potential interaction partners.

Types of Protein Interactions:

- **Stable complexes:** Long-lasting, often structural interactions
- **Transient interactions:** Brief contacts for signaling or catalysis
- **Direct binding:** Physical contact between protein surfaces
- **Indirect associations:** Mediated through other molecules

Prediction Features

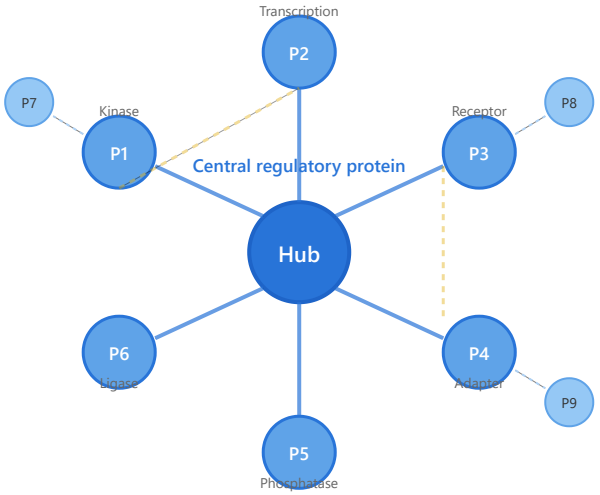
Genomic: Gene fusion, gene neighbors

Proteomic: Experimental PPI data

Evolutionary: Co-evolution signals

Structural: Interface complementarity

Protein Interaction Network



Interaction Types & Confidence

- Direct physical interaction (high confidence)
- Experimentally validated
- Predicted/indirect (medium confidence)
- Cross-pathway communication

5 Evolutionary Conservation & Function

Conservation-Function Relationship

Evolutionary conservation analysis provides powerful insights into protein function. Highly conserved regions across species typically indicate functional importance, as mutations in these areas are often detrimental and eliminated by natural selection.

Key Principles:

- **Sequence conservation:** Similar amino acids across species
- **Structural conservation:** Preserved 3D structure despite sequence variation
- **Functional residues:** Highly conserved catalytic and binding sites
- **Co-evolution:** Correlated mutations reveal interaction partners

Conservation Score Interpretation

Low → High conservation

High conservation (blue): Functionally critical
Low conservation (red): Tolerates variation

Evolutionary Conservation Analysis