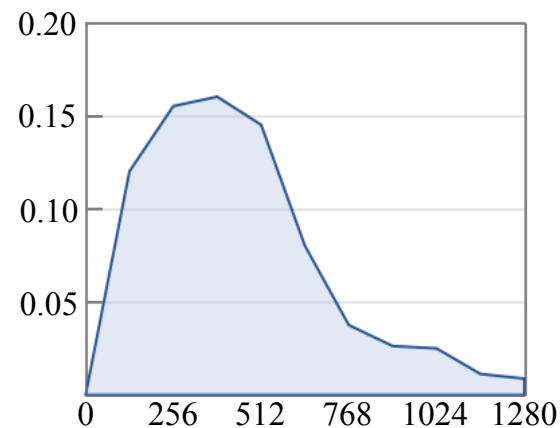
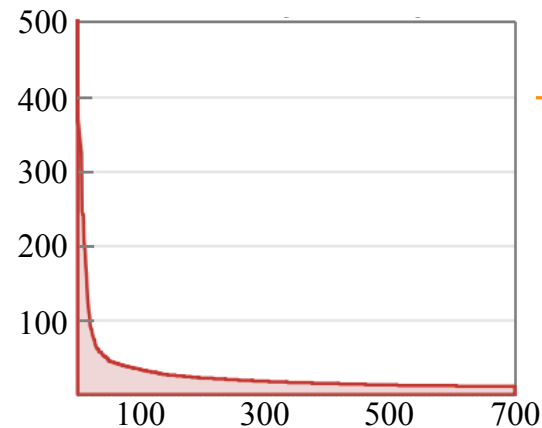


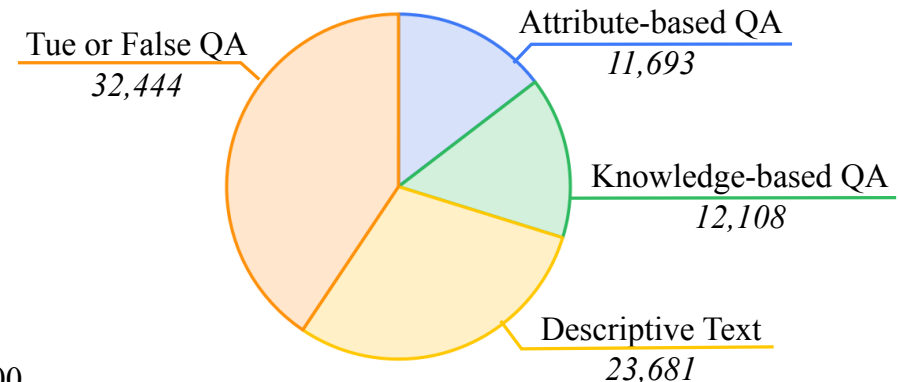
(a) Taxonomy



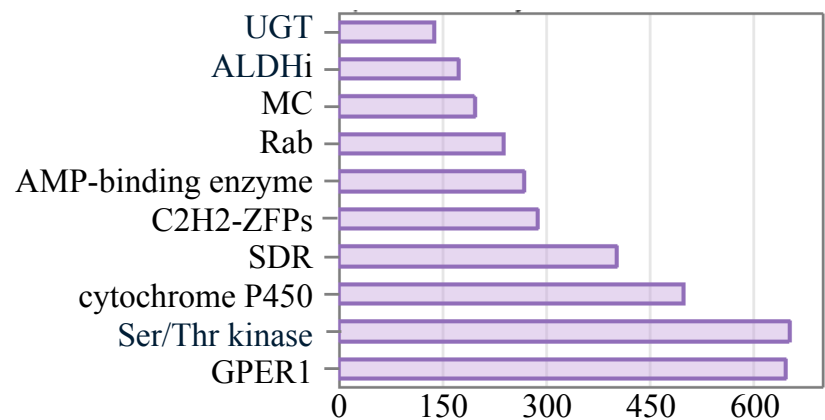
(c) Sequence length



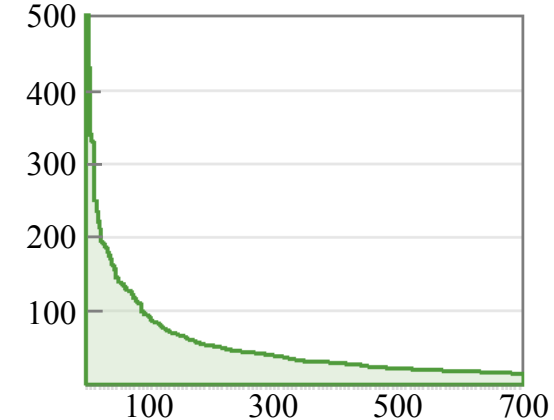
(e) Catalytic activity



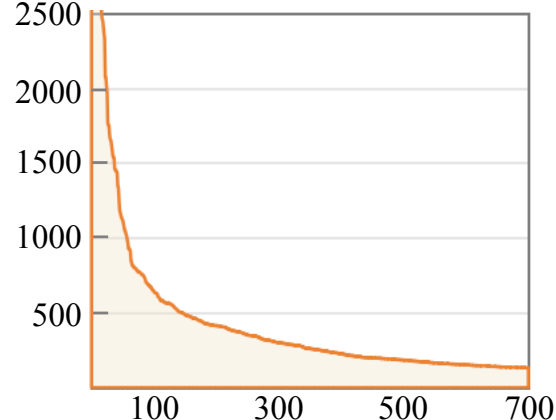
(g) Sample counts across the four QA task types



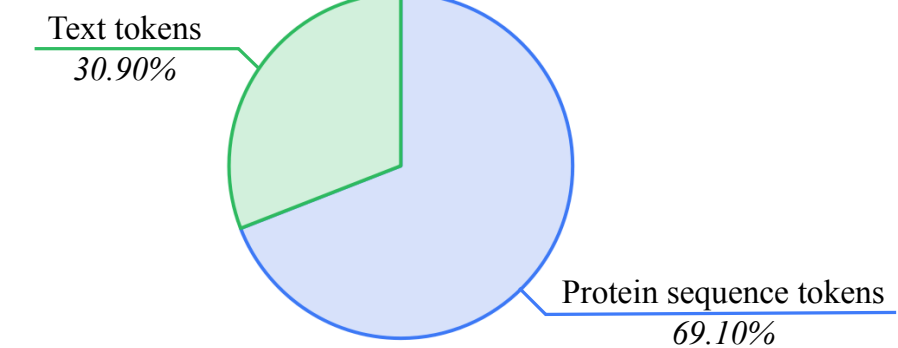
(b) Protein family



(d) Domain



(f) Gene Ontology



(h) Proportion of text tokens versus amino-acid sequence tokens