ISGC 2022

N°872 / OC
TOPIC(s) : Artificial intelligence / Mechanism investigations

Identification of Enzymatic Active Sites with Unsupervised Language Modeling

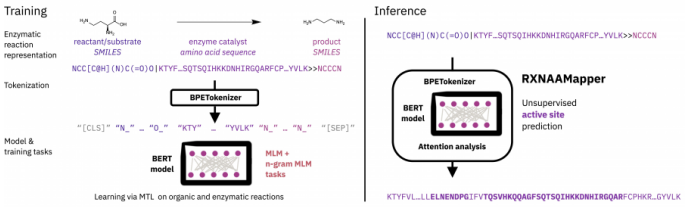
AUTHORS

Matteo MANICA / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON
Loïc KWATE DASSI / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON
Daniel PROBST / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON
Philippe SCHWALLER / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON
Yves Gaetan GAETAN TEUKAM / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON
Teodoro LAINO / IBM RESEARCH EUROPE, SÄUMERSTRASSE 4, RÜSCHLIKON

PURPOSE OF THE ABSTRACT

The first decade of genome sequencing saw a surge in the characterization of proteins with unknown functionality. Even still, more than 20% of proteins in well-studied model animals have yet to be identified, making the discovery of their active site one of biology's greatest puzzle. Herein, we apply a Transformer architecture to a language representation of bio-catalyzed chemical reactions to learn the signal at the base of the substrate-active site atomic interactions. The language representation comprises a reaction simplified molecular-input line-entry system (SMILES) for substrate and products, complemented with amino acid (AA) sequence information for the enzyme. We demonstrate that by creating a custom tokenizer and a score based on attention values, we can capture the substrate-active site interaction signal and utilize it to determine the active site position in unknown protein sequences, unraveling complicated 3D interactions using just 1D representations. This approach exhibits remarkable results and can recover, with no supervision, 31.51% of the active site when considering co-crystallized substrate-enzyme structures as a ground-truth, vastly outperforming approaches based on sequence similarities only. Our findings are further corroborated by docking simulations on the 3D structure of few enzymes. This work confirms the unprecedented impact of natural language processing and more specifically of the Transformer architecture on domain-specific languages, paving the way to effective solutions for protein functional characterization and bio-catalysis engineering.

FIGURES



	Overlap Score	False Positive Rate
Random Model	4.98%	84.20%
Pfam	24.01%	78.01%
BERT-base	28.98%	75.56%
RXNAAMapper (ours)	**31.51%**	**66.63%**

FIGURE 1

RXNAAMAPPER pipeline

A BERT model is trained on a combination of organic and enzymatic reaction SMILES using MTL, leveraging atom-level tokenization and MLM for the SMILES components, while BPE tokenization and n-gram MLM for the AA sequence part (left).
The trained model is

FIGURE 2

Performance on sequence-based active site prediction

Reported in the table the overlap
score and the false positive rates for the active site prediction using PLIP as a ground-truth for the four methods considered: a random model, Pfam alignment-based model, a pretrained BERT model and RXNAAMapper.

KEYWORDS

Enzymatic Reactions | Protein Language Modeling | Chemical Language Modeling | Active Sites

BIBLIOGRAPHY