Identification of Enzymatic Active Sites with Unsupervised Language Modelling

Anonymous Author(s)
Affiliation
Address
email

Abstract

The first decade of genome sequencing saw a surge in the characterisation 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 is able to recover, with no supervision, 31.51% of the active site when considering co-crystallized substrate-enzyme structures as a ground truth, largely outperforming sequence alignment-based approaches. 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 characterisation and bio-catalysis engineering.

1 Introduction

The number of anticipated genes has exploded as high-throughput sequencing technologies have improved dramatically. However, a huge percentage of these newly found genes codifies proteins with no known function, and the identification of active site is often the only viable way to deduce their activity. The most well-established and reliable approaches for identifying active sites are methods like PFAM [Mistry et al., 2020] and PSI-BLAST [Altschul et al., 1997], which are based on sequence similarity.

Active sites are enzyme regions that have evolved specifically to interact with other molecules. Because of their importance in enzyme function, the amino acids (AA) in the active site are more conserved during evolution than the entire sequence. As a result, the signal (or pattern) describing the 3D interaction of amino acids and molecules in the active site may be a valuable source of data for function prediction when represented with a specific domain language.

In the last two years, BERT [Devlin et al., 2018], and more in general models pretrained via masked language modelling, have become the de-facto standard for different language understanding problems [Vig, 2019, Hoover et al., 2019, Rogers et al., 2020]. Self-supervised training techniques enabled the use of unlabelled and unstructured data sources to learn patterns in large corpora without

Submitted to NeurIPS 2021 AI for Science Workshop.

requiring complex annotations. For example, Rao et al. [2019] evaluated how language models pretrained on amino acid (AA) sequences can be used to predict protein properties, such as, function and secondary structure. Schwaller et al. [2021] showed how language models trained via masked language modelling (MLM) on organic chemical reaction data, represented using SMILES [Weininger, 1988], can leverage attention to efficiently and accurately map atoms between precursors and products with no supervision.

Here, we propose to capture the signal between the active site amino acids and the atoms of the interacting molecules to identify protein active sites. We extend the work of Schwaller et al. [2021] by developing a new method that employs a publicly available collection of enzymatic reactions. We use SMILES for modelling each interacting molecule combined with the AA sequence for the enzyme's fundamental structure. This approach is able to recover, with no supervision, 31.51% of the active sites when considering co-crystallized substrate-enzyme structures as a ground truth, largely outperforming current state-of-the-art methods.

2 Methods

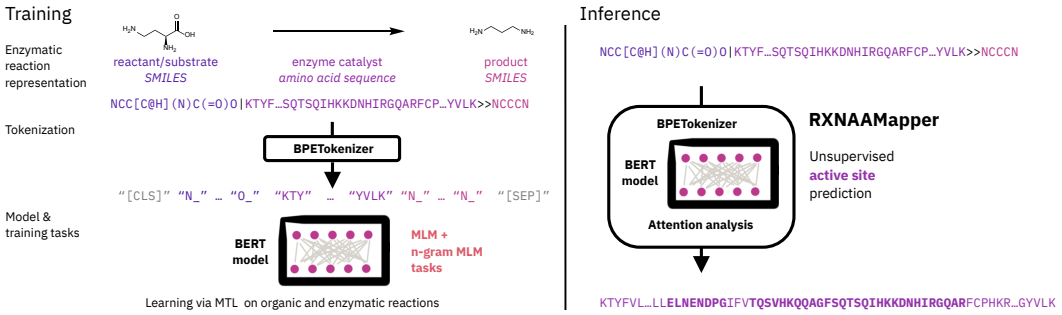

Figure 1: **RXNAAMAPPER pipeline**. A BERT model is trained on a combination of organic and enzymatic reaction SMILES using MTL, leveraging atom-level tokenization and MLM for the SMILES components, while BPE tokenization and n-gram MLM for the AA sequence part (left). The trained model is used in inference to define a score, based on the attention values computed on the reaction SMILES provided as input, which allows the prediction of the active site (bold-face AAs) of the enzyme biocatalyzing the reaction with no supervision or structural information (right).

We repurposed BERT [Devlin et al., 2018] by jointly training it with Masked Language Modelling (MLM) and a n-gram Masked Language Modelling [Xiao et al., 2020] (n-gram MLM, with $n = 3$), leveraging Multi-task Transfer Learning (MTL) [Pesciullesi et al., 2020] on a combination of reaction SMILES representing organic reactions [Lowe, 2012] (weight assigned 0.1) and bio-catalyzed reactions [Probst et al., 2021] (weight assigned 0.9). For enzymatic reactions, each example consists of a reaction SMILES complemented with the AA sequence representation of the enzyme of interest (see Figure 1 for a depiction). As we train the model via MLM and n-gram MLM, we sparsely mask the reactants and the products, and densely mask the enzyme sequence. To reduce the length of the resulting reactions [Filipavicius et al., 2020], we built a sub-word vocabulary for AA sequences using Byte Pair Encoding (BPE) [Gage, 1994] learned on UniProt [Consortium, 2020].

Algorithm 1 Active Site Prediction

procedure RXNAAMAPPER($P \in \mathbb{R}^{r \times m}, k$)
  active_site = `set()`
  for `i in 1..r` **do**
    line = $P[i]$
    for `j in argmax(line, k)` **do**
      active_site.add(j)
  return active_site

The prediction of the active regions of proteins is unsupervised and entirely based on the analysis of the attention values computed by the pretrained language model after encoding a reaction. If we label $S \in \mathbb{R}^{l \times d}$ ($l = r + m + p$) as the embedding of a given reaction and r, m, and p refer to

the length of the reactants, the enzyme, and the products, respectively, a forward pass of S through the model yields a sequence S' with the same dimension as S. Each encoder block computes the attention matrix $A \in \mathbb{R}^{l \times l}$ of the sequence S provided as input[Vaswani et al., 2017]. We construct a matrix $P \in \mathbb{R}^{r \times m}$ by summing two sub-matrices of A, describing the link between reactants and enzymes: $P = A[1:r, 1:m] + A[r+1:r+1+m, 1:r]^T$. We use the matrix P as shown in the Algorithm 1 to predict the active regions via a consensus scheme where each reactant's atom has k votes to choose its best-bound enzyme's token. The selected enzyme's tokens are uniquely gathered in a set and are considered as the active region of the protein. Hereinafter, the method will be referred to as RXNAAMapper.

3 Evaluation

We use a set of 5K co-crystallised ligand-protein pairs from the Protein–Ligand Interaction Profiler (PLIP) [Salentin et al., 2015] as ground-truth, to perform a two-fold evaluation: (1) a sequence-based assessment benchmarking RXNAAMapper against a statistical baseline (Random Model), a pretrained BERT model on natural language (BERT-base) and the alignments retrieved from Pfam [Mistry et al., 2020]; and (2) a structural validation with protein-ligand binding energies computed with docking.

For the sequence-based evaluation, we use an overlap score between the prediction and the ground-truth, as well as the false positive rate. The overlap score is defined considering the active site as a set of non-overlapping segments in a sequence. If S with $|S| = n$ is a sequence of amino acid residues, the active region A_s of S is defined as $A_s = \{(a_i, b_i)\}_i^m$, where a_i and b_i are the index boundaries of the segment i. The overlap score ($OS(A, A_s)$) between the predicted active region $A = \{(a_{pi}, b_{pi})\}_i^n$ and the ground-truth $A_s = \{(a_{si}, b_{si})\}_i^m$ is defined as:

$$OS(A, A_s) = \frac{\sum_i^n \sum_j^m \max(0, \min(b_{pi}, b_{sj}) - \max(a_{pi}, a_{sj}))}{\sum_i^m (b_{si} - a_{si})}$$

Besides the overlap score, the false positive rate (FPR) of the predictions is defined as:

$$FPR = \frac{\sum_i^n (b_{pi} - a_{pi}) \mathbb{1}_{\bigwedge_{j=1}^{m} [a_{pi}, b_{pi}] \cap [a_{sj}, b_{sj}] = \emptyset}}{\sum_i^n (b_{pi} - a_{pi})}$$

For the structural assessment, on a set of protein-ligand active sites predictions, we evaluated the binding energy computed with Autodock Vina [Eberhardt et al., 2021, Trott and Olson, 2010] considering predicted active sites and the ground-truth from PLIP.

4 Experiments and Results

In our experiments, we train models using a dataset of organic reactions from USPTO [Lowe, 2012] ($\sim$1M, splits as reported in Schwaller et al. [2019]) combined with a dataset of bio-catalyzed reactions from ECREACT [Probst et al., 2021]. ECREACT records were further processed by using UniProt to replace and augment Enzyme Commission (EC) numbers with protein sequences from the corresponding EC classes. The resulting dataset has been further split into a training set ($\sim$7M reactions) and a validation set ($\sim$4K reactions). BERT has been trained with 4,094 as the batch size (leveraging gradient accumulation) and LAMB [You et al., 2020] as the optimization algorithm with 20,928 optimization steps.

Table 1 reports the performance of the different approaches on the PLIP groundtruth. Notably, RXNAAMapper performs consistently better than other methods, even though the product information has been completely omitted given the nature of the dataset. Figure 2a shows an exemplar comparison of Pfam-based and RXNAAMapper predictions overlapped with the PLIP ground-truth for a ligand-protein pair. Notably, RXNAAMapper controls the false positives better than Pfam alignments and matches the active site as reported in PLIP.

For a subset of PLIP's protein-ligand pairs, we used Autodock Vina to compute the protein-ligand binding energy for the RXNAAMapper's predicted active sites and for the corresponding experimental ones provided by PLIP. We computed the binding energies by averaging the 3D coordinates of the active sites atoms and setting the box side length to 50 Å. In Figure 2b we show the binding energies

Table 1: **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.

	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%**

for the two active sites: RXNAAMapper and PLIP. We can appreciate how the binding energy calculations performed on the active site predicted by RXNAAMapper are accurately matching the one using the PLIP experimental information.

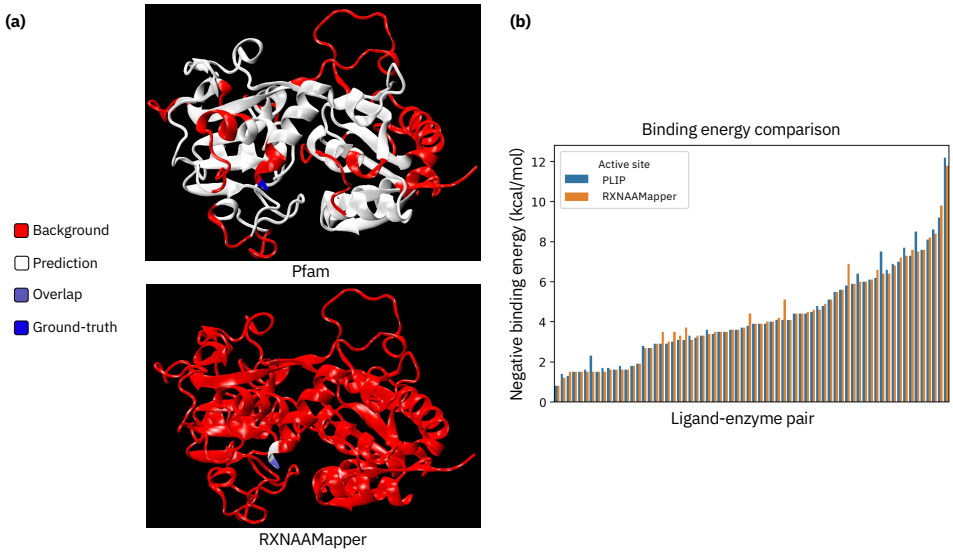

Figure 2: **Experiment results**. **(a)**: comparison of the prediction from Pfam alignments (top) and RXNAAMapper (bottom) using PLIP as a ground-truth for an hydrolase (PDB id: 4FJP) interacting with zinc (SMILES: `[Zn]`). **(b)**: reported in the barplot the negative binding energy computed for a subset of protein-ligand pairs from PLIP using two configurations for the active site: experimental from PLIP (blue) and predicted via RXNAAMapper (orange).

5 Discussion

The unsupervised detection of active sites using only protein sequence information is an important step to identify the function of uncharacterized proteins. We presented RXNAAMapper, a technology that uses pretrained language models on text-based molecule representations to identify active sites in long amino acid sequences. When compared with protein-ligand interactions from PLIP, our approach outperforms other sequence-based baselines reconstructing >30% of proteins' active regions while better controlling false positives. Furthermore, using docking studies, we have been able to validate the RXNAAMapper active sites predictions by comparing the binding energy configurations with the experimental references. These results show how large language models can inherently capture 3D structural information and reaction mechanisms from 1D representations, thus, opening up novel avenues for their consistent application in enzyme engineering.

References

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 10 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa913. URL `https://doi.org/10.1093/nar/gkaa913`.

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 09 1997. ISSN 0305-1048. doi: 10.1093/nar/25.17.3389. URL `https://doi.org/10.1093/nar/25.17.3389`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, 2019.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276*, 2019.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32:9689, 2019.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. *arXiv preprint arXiv:2010.12148*, 2020.

Giorgio Pesciullesi, Philippe Schwaller, Teodoro Laino, and Jean-Louis Reymond. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature communications*, 11(1):1–8, 2020.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Daniel Probst, Matteo Manica, Yves Gaëtan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino. Molecular transformer-aided biocatalysed synthesis planning. *ChemRxiv*, 2021. doi: 10.26434/chemrxiv.14639007. URL `https://app.dimensions.ai/details/publication/pub.1138314574andhttps://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/60c75919842e6599a7db4990/original/molecular-transformer-aided-biocatalysed-synthesis-planning.pdf`.

Modestas Filipavicius, Matteo Manica, Joris Cadow, and Maria Rodriguez Martinez. Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks. *arXiv preprint arXiv:2012.03084*, 2020.

Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL `https://doi.org/10.1093/nar/gkaa1100`.

Ashish Vaswani, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, N. Gomez Aidan, Kaiser Lukasz, and Polosukhin Illia. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Sebastian Salentin, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021. doi: 10.1021/acs.jcim.1c00203. URL https://doi.org/10.1021/acs.jcim.1c00203. PMID: 34278794.

Oleg Trott and A. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31, 2010.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2020.