Language models can identify enzymatic active sites in protein sequences

Yves Gaetan Nana Teukam[1,*], Loïc Kwate Dassi[1], Matteo Manica[1], Daniel Probst[1,2], Philippe Schwaller[1,2], and Teodoro Laino[1,2]

[1]IBM Research Europe, Saümerstrasse 4, 8803 Rüschlikon, Switzerland

[2]National Center for Competence in Research-Catalysis (NCCR-Catalysis), Switzerland

[*]yna@zurich.ibm.com

January 12, 2023

Abstract

Recent advances in language modeling have tremendously impacted how we handle sequential data in science. Language architectures have emerged as a hotbed of innovation and creativity in natural language processing over the last decade, and have since gained prominence in modeling proteins and chemical processes, elucidating structural relationships from textual/sequential data. Surprisingly, some of these relationships refer to three-dimensional structural features, raising important questions on the dimensionality of the information contained in sequential data. We demonstrate that the unsupervised use of a language model architecture to a language representation of bio-catalyzed chemical reactions can capture the signal at the base of the substrate-active site atomic interactions, identifying the three-dimensional active site position in unknown protein sequences. The language representation comprises a reaction-simplified molecular-input line-entry system (SMILES) for substrate and products, and amino acid sequence information for the enzyme. This approach can recover, with no supervision, 52.12% of the active site when considering co-crystallized substrate-enzyme structures as ground truth, vastly outperforming other attention-based models.

1 Introduction

Language Models (LMs) (e.g. BERT (*1*), GPT (*2*), and ELMo (*3*)) made the headlines being worldwide relevant for tasks such as information retrieval (*4*), text generation (*5–7*), and speech recognition (*8*). The ability of LMs to learn a probability distribution over sequences of words in relation to a domain-specific language is the primary factor contributing to their level of success. In essence, these architectures encode distinct vector representations (embeddings) of a word based on its context, uncovering linguistic relationships between the different words of the domain-specific language.

Large Language Models (LLMs) (*9, 10*) trained on massive and diverse corpora are demonstrating critical new abilities, from writing innovative content (*11*) to resolving simple math problems (*12*). These models achieved relatively high performance on new tasks for which they were not explicitly trained (also known as zero-shot learning tasks) (*13, 14*), most likely because the ability of language architecture to generalize on new tasks is the result of an unintentional multitask learning process. (*14*).

LMs, especially transformers and their derivatives (e.g. BERT (*1*), ALBERT (*15*), RoBERTa (*16*), etc.), had also a relevant impact on chemistry and biology, reaching state-of-the-art performance when fine-tuned on specific tasks (*17–21*). In chemistry, reagents, substrates, and products are usually depicted using a text representation such as SMILES (simplified molecular-input line-entry system) (*22, 23*). Using this domain-specific representation, scientists showed that LMs can learn to accurately map atoms between precursors and products with an unsupervised masked language modeling (MLM) (*24*), or predicting molecular properties using a BERT-like model trained in a semi-supervised way (*25*). With the extension of string-based representations to proteins, LMs can be used for uncovering hidden relationships in biological tasks. Unsupervised language models have been used for predicting mutational effect and secondary structure (*26*), improving long-range contact prediction (*27*), targeting binding sites (*28*) or capturing important biophysical properties governing protein shape (*28, 29*).

The identification of active site residues and the characterization of the corresponding protein function (*30–32*) is the next big scientific challenge, especially after the pioneering work on predicting proteins structure (*20, 33*). The activity of a protein is directly related to the structure of the active site (*34*), a spatial region with a joint/disjoint amino acid (AA) sequence evolved to interact with specific molecules under selective pressure. The fact that the amino acids (AA) in the active site have been conserved more than the entire sequence during evolution reflects their importance in providing unique structural features for enzyme function (*35, 36*). If the signal (or pattern) describing the 3D interaction of amino acids with the corresponding target molecules in the active site could be learned from the AA sequence and the molecular representation, protein function could be predicted using only sequential data with no use of explicit 3D structural features derived from co-homology strategies (*31, 37–42*), or from protein-protein interaction networks (*43*). Currently, only a few efforts, such as Pfam (*44*) and PSI-BLAST (*45*), took up the challenge of identifying active sites solely based on sequence similarity information.

Inspired by the work of (*24*), here we show that LMs can learn the signal characterizing the active sites AA using a linguistic representation for proteins and their molecular substrates (see Figure 1). We use a publicly available collection of enzymatic reactions (*46*), in which substrate molecules are represented with SMILES and the proteins with their AA linear sequence. The unsupervised training can recover 52.12% of the active sites when considering co-crystallized substrate-enzyme structures as ground truth without supervision.

This work confirms the versatility of LMs in extracting complex structural information from a sequence-based representation and demonstrates the effectiveness of using LMs for identifying protein functions.

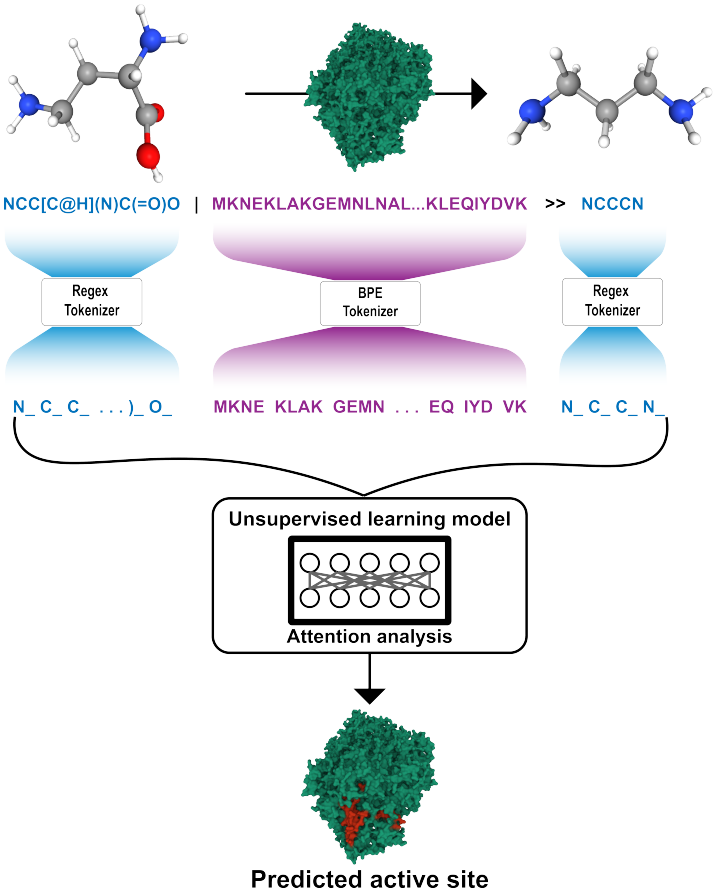


Figure 1: **RXNAAMAPPER pipeline**. A BERT model (*1*) is trained on a combination of organic and enzymatic reaction SMILES using MTL (*47*), leveraging atom-level tokenization and MLM (*48*) for the SMILES components, while Byte Pair Encoding (BPE) tokenization and n-gram MLM for the amino acid sequence part. 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 of the enzyme bio-catalyzing the reaction with no supervision or structural information. The active sites are represented in our plot as red regions in the molecule.

2 Results

Sequence compression and representation

Before language modeling, amino acid sequences must be numerically encoded using an encoding scheme that gives each amino acid sequence a unique vector representation. This encoding method acts as a map from the input amino acids to a point in the protein representation space. The embedding should be able to capture key features of each element (also called a token) that is encoded and should be able to preserve the relationship between the encoded elements (typically expressed geometrically through a vectorial representation of the encoded tokens).

Here we consider a Transformer model based on BERT that in a standard setup handles a maximum of 512 input tokens to generate the encoding vector. For architectures like ours, 512 input tokens size is a pretty strict limitation due to the large memory footprint coming from self-attention layers, whose complexity scales quadratically with the input sequence length (*49*). In protein modeling, the model must see the entire amino acid sequence in order to learn crucial structural information. Given that amino acid sequences can be prohibitively long, finding a good compression and representation scheme becomes a fundamental task before the model training.

We trained different Byte-Pair Encoding (BPE) tokenizers with various settings (as described in the method section) to find the set of parameters that maximizes the compression of the amino acid sequences in terms of vocabulary size and sequence length. The compression power of the tokenizers trained has been tested on a dataset of random sequences from Uniprot $(n = 600K)$. Figure 2B shows a negative correlation between the vocabulary size and the median number of tokens for the same dataset. This result confirms that by increasing the vocabulary size we are implicitly increasing the length of BPE tokens in our vocabulary, as we merge the most frequently occurring fragment of sequences into single subwords or fragments. The BPE tokenizer giving the best compression rate ($c = 66,8\%$) has been trained on the dataset made of amino acid sequences of lengths between 600 and 700, and setting the vocabulary size to 75K tokens. The comparison of the best performing BPE with a simple character-Level tokenizer (ByChar), which splits sequences into single amino acids, shows that the median number of tokens per sequence in our test dataset drastically drops to 152 (see Figure 2A and Table S1). Therefore by using this tokenization scheme, we overcome the architectural limitations and train our model on broader corpora.

Active site prediction

Enzyme binding sites are areas on an enzyme's surface specifically intended to interact with other molecules. Enzymes can have many types of binding sites that perform distinct tasks and engage different molecules. The most significant is the active site, which includes catalytic residues to carry out the enzymatic reaction on a substrate. We trained the RXNAAMapper (a BERT-base model combined with BPE tokenization on the amino acid sequences

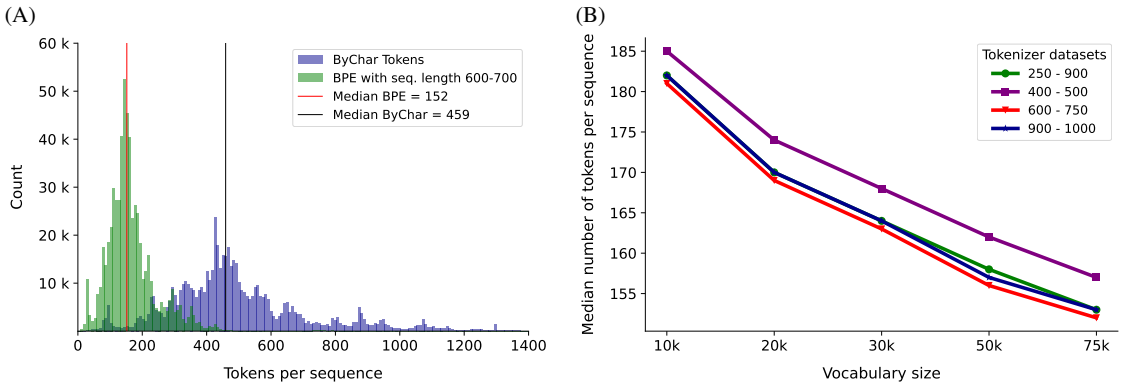


Figure 2: **BPE analysis**. On the left is the density distribution plot of the number of tokens using the BPE tokenizer chosen vs the ByChar tokenizer. As compared to the ByChar tokenization, the BPE splits infrequent fragments into two or more fragments and also merges the most frequent ones into longer fragments. Its application results in shortening the sequences with respect to the ByChar tokenization scheme. On the right is the median of the BPE for each configuration of our grid search.

developed in this work), BERT-base, BERT-Large, ProtAlbert, and ProtBert (details in methods) and compared them on the task of active site predictions using 777 amino acid sequences from PLIP (with related active sites) as ground truth. The predictions are based on the self-attention analysis of the models. Self-attention modules, a key component in Transformers-based models, is an attention mechanism that connects distinct points in a single sequence to calculate a representation of the same sequence. To match the required output dimension, the separate attention 'heads' are commonly concatenated and multiplied by a linear layer (*50*) forming a multi-head attention system. With this architecture, the specific heads in the model may selectively focus on useful sections of the input sequence and so capture the relationship between them. Throughout the analysis of the attention mechanism in our model, we found out that the specific combination of attention heads, the layer at which the attention scores are extracted, and the number of amino acid tokens to bind with each reactant's atom (a.k.a top_k), led to different performances in the prediction of active sites. Given our architectural design, the combination giving us the highest overlap score when predicting the active sites is $head = 10$, $layer = 5$, and $top_k = 6$.

For each trained model, we determined their performance by selecting the combination giving the highest overlap between the predictions and ground truth from PLIP (details in the methods section). We find that RXNAAMapper performs consistently better than the other unsupervised sequence-based methods, even though the product information has been completely omitted, given the nature of the dataset. Among the active sites predicted by RXNAAMapper, up to 52.12% overlap with the ground truth whereas ProtAlbert, ProtBert, BERT-base, BERT-Large, and the random model, reached 18.27%, 20.01%, 15.26%, 45.88%, and 14.07%, respectively. As a reference, we report the overlap score obtained by homology-based on Pfam annotations (67.31%).

Active site prediction based on homology models, like the one obtained using Pfam annotations, can help recover active sites. However, these approaches use heuristics-based methods, giving rise to high frequencies of false positive rates (Pfam-based = 61.68%). The high false positive rate suggests that the area predicted as active sites are too large concerning the actual size of active sites. While active sites usually account for just 10-20% of the volume of an enzyme (*51*), Pfam's-based active sites on average account for 61.68% of the size of the input sequences. Our model predicted shorter stretches of the input sequences as active sites (on average 41.5%) while maintaining a lower false positive rate compared to the homology-based predictions (47.89%).

	Overlap Score	False Positive Rate
Random Model	14.07%	13.80%
BERT-base	15.26%	14.92%
BERT-Large + BPE	45.88%	42.71%
ProtAlbert	18.27%	27.33%
ProtBert	20.01%	16.42%
RXNAAMapper (ours)	52.12%	47.89%
Pfam-based	67.31%	61.68%

Table 1: **Performance on sequence-based active site prediction**. Reported in the table are the overlap score and the false positive rates for the active site prediction using PLIP as ground truth for the seven methods considered: a random model, Pfam alignment-based model, a pre-trained BERT-base model, a pre-trained BERT-Large model coupled with a BPE tokenizer, ProtAlbert, ProtBert, and RXNAAMapper. Among these models, Pfam-based predictions are based on homology present within Pfam families. The others are attention based models extracted from unsupervised language models.

We inspect the performance of our model and the Pfam-based across different enzyme classes (see Figure 4A) and reaction classes (see Figure 4B). Certain types of enzymes (e.g. Transferases) and reaction classes (e.g. functional group interconversion (FGI)) have a better compromise of overlap score and false positive rate with respect to other classes. In some cases, like with Lysases, our model have a very different behavior compared to the homology-based predictions. Our model is more conservative by generating shorter active sites, which results in a lower false positive rate and overlap score. Pfam-based instead generates long active sites. The high false positive rate of Pfam-based prediction correlates with the inefficacy of sequence alignments to produce accurate results when the sequence identity goes below a certain threshold (*52*). Despite alignment-based methods, our methodology (an alignment-free approach) captures evolutionary events without the assumption that homologous sequences are the consequence of a succession of linearly organized and more or less conserved sequence regions.

We further compared our prediction and those from homology-based by looking at the distance between the barycenter of the grid boxes centered on the predicted active sites and the ground truths. Although our model has lower overlap scores compared to the Pfam-based, our ability to control the false positive rate is reflected on the

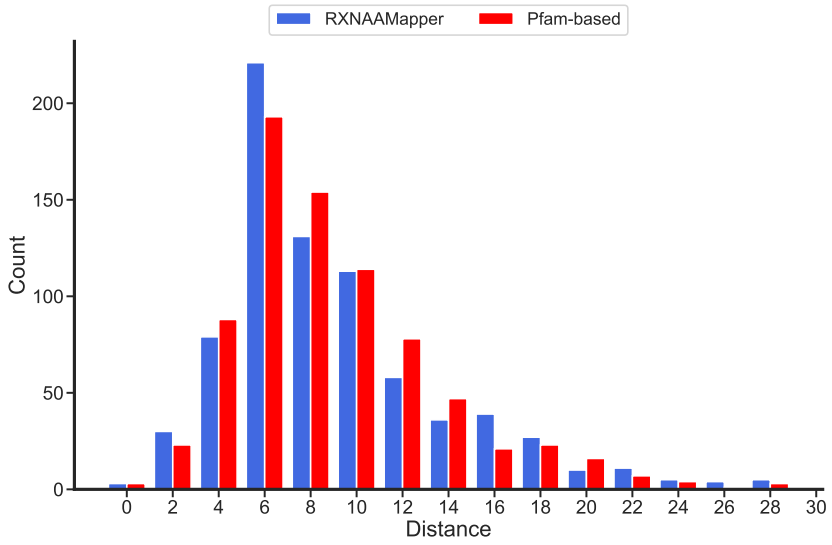


Figure 3: **Active sites distance from ground truth**. Distribution plot depicting the distance of the predicted active site from the PLIP annotations. For both predictions, the distribution is right skewed reflecting the correctness of the predictions. RXNAAMapper exhibits a distribution peak at lower values, confirming the superior accuracy of its predictions in comparison to Pfam annotations.

barycenter of our predictions to be spatially closer to the ground truth (see Figure 3 and Figure S1).

Figure 5 shows a typical comparison of Pfam-based and RXNAAMapper predictions overlapped with the PLIP ground truth. Notably, RXNAAMapper exhibits better false positive rates than Pfam alignments, while matching the active site reported in PLIP.

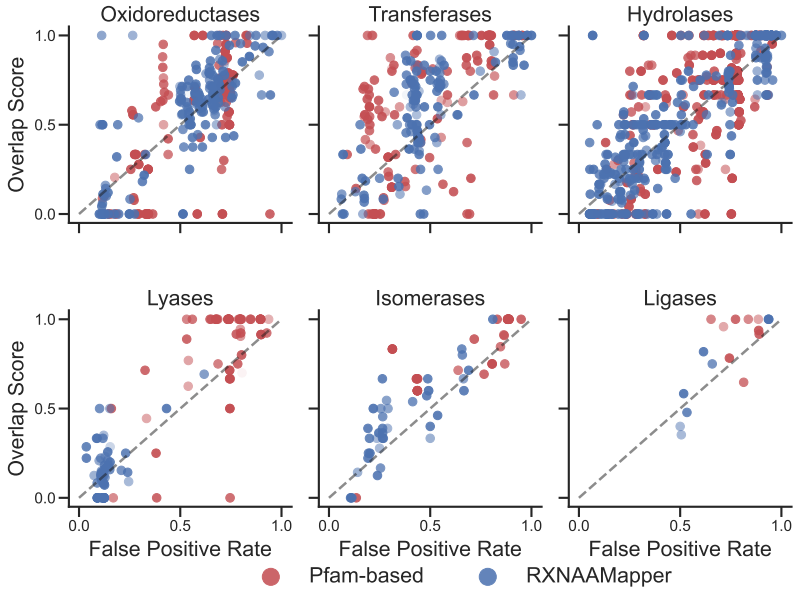
Our approach demonstrates the potential of language models to retrieve important structural information from a text representation. This potential reflects in our model's capacity to control false positives while generating predictions that are spatially closer to the actual active sites.

Structural validation

For a chemical reaction to take place, the substrates must bind to the enzyme at its active site. This area is divided into two parts: the binding site and the catalytic site. Some of the residues in the binding site aid in the substrate (reactants) binding to the enzyme. A chemical reaction is aided by the catalytic site. In addition to specificity, the substrate binding site supplies binding energy to keep the substrate engaged on the active site throughout the catalytic process. The interaction between the ligand and the enzyme does not occur at random locations but around the active sites to produce specifically high binding energies toward the ligand.

We further evaluated the active site predictions from RXNAAMapper by using them to compute the binding

(A) Comparison at EC class level



(B) Comparison at reaction class level

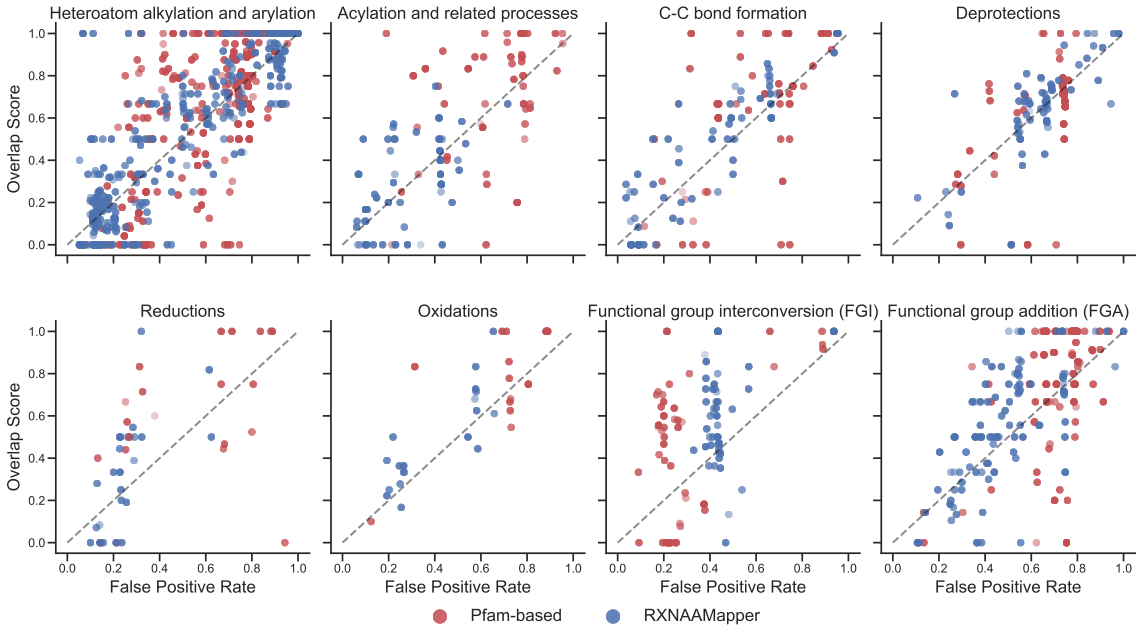


Figure 4: **Pfam-based and RXNAAMapper performances with respect to the EC classes and reaction classes**. The models exhibit different performances for different types of reactions and enzymes, underlying the modeling complexity of certain types of reactions and enzymes. For almost every enzyme class and reaction class, our model's predictions are on the bottom left of the figures (lower overlap score and false positive rate), while Pfam-based predictions are on the symmetrically opposite side of the figure (higher overlap and false positive rate). This highlights the ability of our model to predict active sites while keeping the false positive rate within descent frequencies. The transparency in the figures correlates with the distance between the barycenter of the grid boxes centered on the predicted active sites and the one centered on the ground truth. The closer the prediction is to the ground truth, the more opaque the point.

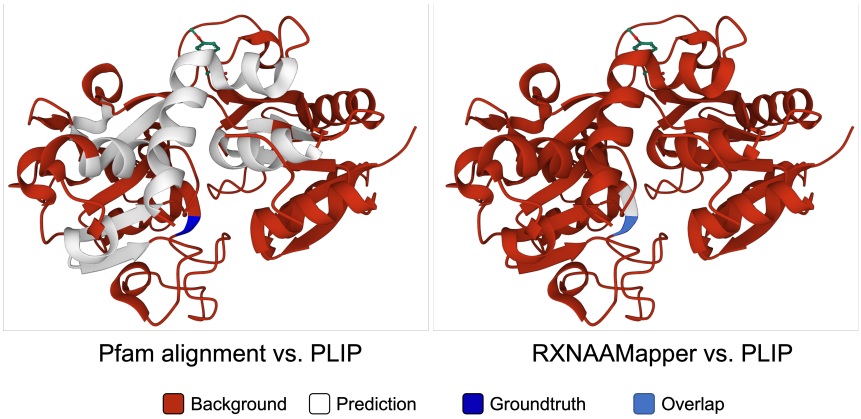


Figure 5: **Experiment results**. Comparison of the prediction from Pfam alignments (left) and RXNAAMapper (right) using PLIP as a ground-truth for a hydrolase (PDB id: 4FJP) interacting with zinc (SMILES: `[Zn]`). The area in white represents the predicted active site region, while the blue area represents the ground truth of active sites. The red area depicts the backbone of the protein.

energy on a subset of 2213 enzyme-ligand pairs representing the maximum number of pairs extractable from the intersection of our dataset and PLIP database (for more details in the Methods section). We used Autodock Vina (*53*), a molecular docking tool, to estimate the binding energies. We compare the predicted binding energy generated from RXNAAMapper's predicted active sites with those from the corresponding experimental ones provided by PLIP. The binding energy calculations performed on the active site predicted by RXNAAMapper accurately match the one using PLIP experimental information, with a difference of 0.37 kcal/mol between the two on average.

3 Discussion

The prediction of protein active sites, which are critical and conserved functional areas of proteins, is crucial to improving our understanding of protein function. Moreover, the ability to detect these regions in an unsupervised fashion, solely relying on AA sequence information, allows an initial characterization of novel proteins.

Herein, we tackled the problem by introducing RXNAAMapper, a technology that uses pre-trained language models based on textual representations of biochemical reactions to identify active sites in long amino acid sequences. When tested on PLIP protein-ligand interactions, our approach outperforms other sequence-based methods by identifying more than 52% of proteins' active regions with a lower false-positive rate. The combination of a model like a BERT-base and a BPE tokenization system leverage an as-of-now unexplored potential.

One of the main limitations in applying language models to enzymatic reactions is the computational burden introduced by handling long sequences. Compressing the representations using efficient tokenization strategies mitigates the problem, but it has also the detrimental effect of discarding data points that may contain useful

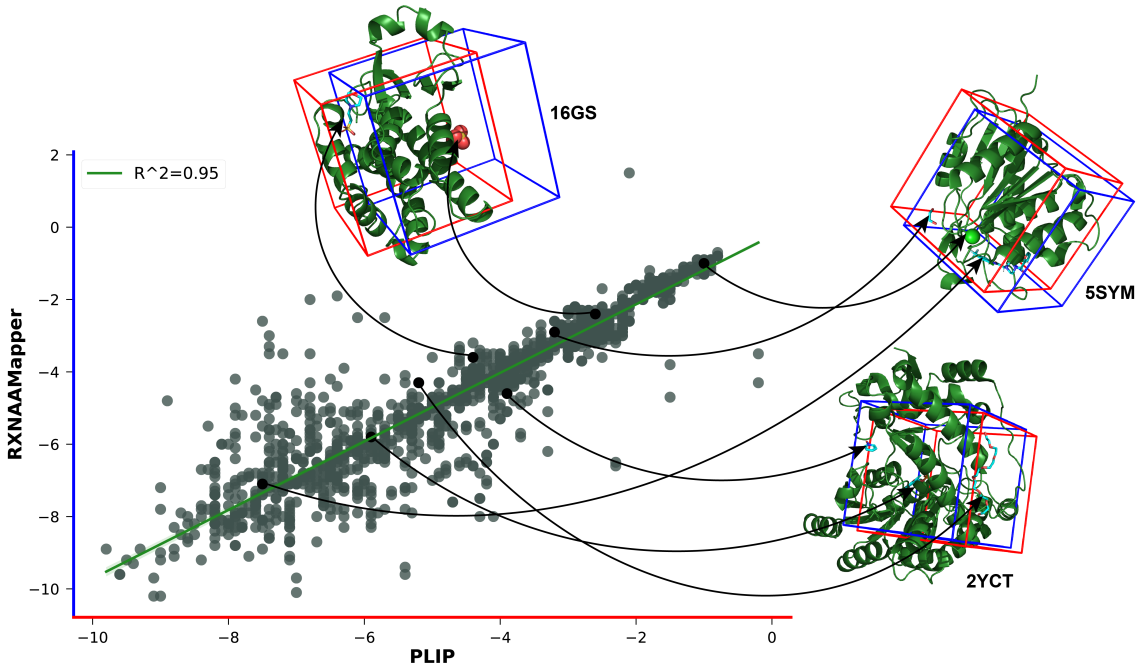


Figure 6: **Negative binding energy of 2213 enzyme-ligand pairs**. The figure shows how the energy scores deriving from the RXNAAMapper active site predictions are in the same range with respect to those predicted from PLIP (R^2=0.95).

information. The use of a BPE tokenizer allowed us to train our model on entire amino acid sequences by compressing sequences in a lossless fashion. Leveraging the full sequence is a key component of our model as amino acids that in a protein sequence are far away may come together in the 3D representation. Unlike other models evaluated in this paper, RXNAAMapper demonstrated the ability to capture the syntax of bio-catalyzed reactions and grasp relevant features of the AA sequences by detecting regions of importance via reaction language modeling.

To further validate the active sites predicted with RXNAAMapper, we compared the binding energy of a set of enzyme-ligand pairs using the active sites from PLIP as ground truth. The binding energy computed from our predicted active sites matched those predicted from the ground truth. This is particularly important because the identification of the active sites helps to predict the binding energy and therefore model the binding sites. Residues present in binding sites, particularly catalytic residues in active sites, are among the most critical residues in an enzyme structure. Thus, identifying them is critical for improving the design of biocatalytic processes.

This work set a stepping stone towards novel in-silico approaches for protein function identification, and it is further evidence of the amazing ability of language models to retrieve 3D structure information from 1D sequential

representation. We are confident that future language models with yet-to-be-unveiled capabilities will continue to offer innovative solutions to complex tasks using domain-specific languages without supervision.

4 Methods

Dataset

We consider a dataset of 1 million organic reactions from USPTO (*54*), split as reported in schwaller2019molecular, combined with a dataset of bio-catalyzed reactions called ECREACT (*46*). ECREACT contains 62,222 reactions with a unique reaction-EC combination. The entries from this dataset can be classified based on their EC numbers and grouped as in Table 2. ECREACT is the result of the combination of bio-catalyzed reactions coming from 4 databases: Brenda (*55*), Rhea (*56*), PathBank (*57*), and MethaNetX (*58*). For our analysis, we consider the entire dataset regardless of the level of information contained in the EC level.

Groups	**N^o of reactions**	**Level of information**
No EC number	55,115	No information about the enzyme
EC-level 1 (EC1)	55,707	Enzyme class
EC-level 1-2 (EC2)	56,222	EC1 + Subclass
EC-level 1-3 (EC3)	56,579	EC2 + sub-subclass
EC-level 1-4 (EC4)	62,222	EC3 + serial number in the subclass

Table 2: ECREACT dataset division divided into groups based on the level of information

Data processing

Using EC numbers as the single filter, we mapped the EC numbers to their corresponding AA sequences from Uniprot (*59*). A filtering step is performed to reduce the overrepresentation of certain EC numbers by limiting to 10K the maximum number of sequences per EC number. In case of exceeding for certain EC number, we randomly selected 10K AA sequences from its set. Then finally paired the substrate-product with the AA sequences.

Tokenization

Language models operate on numerical data requiring text transformation into a numerical representation. An important step in the conversion of text to vectors is the tokenization step. Tokenization is the task of dividing a text into its constituent parts. Two different tokenization approaches are used to deal with the dual representation of the molecular entries in our dataset (SMILES and amino acids sequences). For SMILES, we use a Regex-based transformation (*60*) (character-level tokenization). While for amino acid sequences, the selection of the tokenizer

is a bit more complex. The complexity derives from: (a) the length of the sequences; (b) the limited number of input tokens supported by LMs. To overcome these limitations, compress our input sequences. We trained several tokenizers with various settings to maximize the sequence compression using a grid search over sequence length and vocabulary size.

All the tokenizers trained are BPE (Byte-Pair Encoding) tokenizers. To investigate the effect of the vocabulary size on the tokenizers' compression abilities, we set this variable to 10k, 20k, 30k, 50k, and 75k. We create four small datasets of randomly selected AA sequences in the following ranges: 250-900, 400-500, 600-750, and 900-1000. Each dataset consists of 400K sequences.

Models and training procedure

Herein, we consider different transformer architectures, i.e., BERT (*1*) and Albert (*15*), exploring various approaches: training from scratch, fine-tuning, and pre-trained models.

In training, to better control the token masking and handle the different lengths of the enzymatic reaction components, the model variants have been jointly optimized with Masked Language Modeling (MLM) and an n-gram Masked Language Modeling (*48*) (n-gram MLM, by randomly masking out 15% of the input tokens). MLM and n-gram MLM have been applied to substrates/products and enzymes respectively. As the models are trained on different datasets, i.e. ECREACT (*46*) and USPTO (*54*), Multi-task Transfer Learning (MTL) (*47*) has been adopted on a combination of reaction SMILES representing organic reactions (weight assigned 0.1) and bio-catalyzed reactions (weight assigned 0.9) to create a task-specific language model able to understand bio-catalyzed reactions.

The use of USPTO in a transfer learning process is to ease the understanding of generic chemistry and SMILES syntax. 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.

Six million (6M) reactions subset of the preprocessed ECRACT was chosen at random and used as the training set for all language models. Another 2.5M ECREACT subset was chosen as the validation set to compute the validation loss. For all the language models we used default hyper-parameters from the HuggingFace implementation. As an optimizer we adopted ADAM (*61*) for 50,000 training steps.

It has been recently shown that large pre-trained models on natural language can be fine-tuned on different data modalities to attain comparative performance with respect to models trained on downstream tasks (*62*). Inspired by this seminal work, here we also decided to include in our study the following models: ProtAlbert (*63*) and ProtBert (*63*) (pre-trained on protein data and fine-tune on bioctalyzed data), BERT-base (pre-trained on various data modalities), BERT-base (*1*) and BERT-Large (*1*)(trained from scratch on bioctalyzed data).

Active site prediction

The prediction of the active regions of proteins is unsupervised and entirely based on the analysis of the attention values computed by the pre-trained 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 (*50*). 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 the protein's active region. Hereinafter, the method combined with BERT-base and the BPE will be referred to as RXNAAMapper.

Algorithm 1 Active Site Prediction

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

Evaluation

We use a set of 5K co-crystallized ligand-protein pairs from the Protein-Ligand Interaction Profiler (PLIP) (*64*) as ground-truth, to perform a two-fold evaluation: (1) a sequence-based assessment benchmarking RXNAAMapper against two fine-tuned protein language models (ProtAlbert and ProtBert), a statistical baseline (Random Model), two pre-trained BERT models on natural language (BERT-base and BERT-Large) and the alignments retrieved from Pfam (*44*); (2) a structural validation with protein-ligand binding energies computed with docking. We used Pfam annotations for a fair assessment with existing methods using sequence information only. 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 2213 protein-ligand active site predictions, we evaluated the binding energy computed with Autodock Vina (*53, 65*) considering predicted active sites and the ground truth from PLIP. We chose these enzyme-ligand pairs by first matching PDBs and amino acid sequences with annotated active sites from PLIP. Then filtering reactions catalyzed by enzymes not present in our training set. We then selected the reactions having unique combinations of PDB, EC number, ligands, and predicted active sites. We computed the Cartesian coordinates of the ligand and receptor molecules, which are generally retrieved from the PDB (*66*) or PDBQT (*67*) for the protein, and PDB, PDBQT, or Mol2 for the ligand. To calculate the binding free energy of a ligand to an enzyme, we first computed a grid box centered on the active site where the ligand is to be docked. The box has been found by averaging the 3D coordinates of the atoms of the active site and setting the box side length to 50 Å.

Data and code availability

The ECREACT data set is publicly available at the URL https://github.com/rxn4chemistry/biocatalysis-model. The code is available at the URL https://github.com/rxn4chemistry/rxnaamapper. Structures of docked proteins and results are available at the URL https://doi.org/10.5281/zenodo.7530180.

Acknowledgments

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References

1. J. Devlin, M. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Association for Computational Linguistics, 2019), vol. 1, pp. 4171–4186.

2. S.-Y. Su, Y.-S. Chuang, Y.-N. Chen, *Dual Inference for Improving Language Understanding and Generation* (arXiv, 2020).

3. M. E. Peters, *et al.*, *Deep contextualized word representations* (arXiv, 2018).

4. S. Zhuang, H. Li, G. Zuccon, *Deep Query Likelihood Model for Information Retrieval* (2021), pp. 463–470.

5. J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, J.-R. Wen, *Pretrained Language Models for Text Generation: A Survey* (arXiv, 2022).

6. A. Radford, *et al.*, *Language Models are Unsupervised Multitask Learners* (2019).

7. T. B. Brown, *et al.*, *Language Models are Few-Shot Learners* (arXiv, 2020).

8. T. Hori, J. Cho, S. Watanabe, End-to-end speech recognition with word-based rnn language models. *2018 IEEE Spoken Language Technology Workshop (SLT)* pp. 389–396 (2018).

9. F. Xu, U. Alon, G. Neubig, V. Hellendoorn, *A systematic evaluation of large language models of code* (2022), pp. 1–10.

10. J. Wei, *et al.*, *Emergent Abilities of Large Language Models* (2022).

11. T. B. Brown, *et al.*, Language models are few-shot learners. *CoRR* **abs/2005.14165** (2020).

12. K. Noorbakhsh, M. Sulaiman, M. Sharifi, K. Roy, P. Jamshidi, *Pretrained Language Models are Symbolic Mathematics Solvers too!* (arXiv, 2021).

13. T. Kojima, S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, *Large Language Models are Zero-Shot Reasoners* (2022).

14. V. Sanh, *et al.*, Multitask prompted training enables zero-shot task generalization. *CoRR* **abs/2110.08207** (2021).

15. L. Zhenzhong, *et al.*, Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

16. Y. Liu, *et al.*, *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (2019).

17. P. Schwaller, *et al.*, Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **11** (2020).

18. A. C. Vaucher, *et al.*, Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **12**, 2573 (2021).

19. R. Rao, *et al.*, MSA Transformer. *bioRxiv* (2021).

20. J. Jumper, *et al.*, Highly accurate protein structure prediction with alphafold. *Nature* **596**, 1-11 (2021).

21. A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, T. Laino, Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence* **3**, 485–494 (2021).

22. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31–36 (1988). Publisher: American Chemical Society.

23. D. Weininger, A. Weininger, J. L. Weininger, SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences* **29**, 97–101 (1989). Number: 2 Publisher: American Chemical Society.

24. P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **7**, eabe4166 (2021).

25. S. Wang, Y. Guo, Y. Wang, H. Sun, J. Huang, *SMILES-BERT: large scale unsupervised pre-training for molecular property prediction* (2019), pp. 429–436.

26. A. Rives, *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

27. R. Rao, *et al.*, Evaluating protein transfer learning with tape. *Advances in neural information processing systems* **32**, 9689 (2019).

28. J. Vig, *et al.*, *BERTology Meets Biology: Interpreting Attention in Protein Language Models* (2020).

29. A. Elnaggar, *et al.*, Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* (2020).

30. A. Chatterjee, Protein active site structure prediction strategy and algorithm. *International Journal of Current Engineering and Technology* **7**, 1092-1096 (2017).

31. A. Yousaf, T. Shehzadi, A. Farooq, K. Ilyas, Protein active site prediction for early drug discovery and designing. *International Review of Applied Sciences and Engineering* **13**, 98 - 105 (2021).

32. T.-D. Nguyen-Trinh, K. Lee, R. Kusuma, Y. Y. Ou, Prediction of atp-binding sites in membrane proteins using a two-dimensional convolutional neural network. *Journal of Molecular Graphics and Modelling* **92** (2019).

33. M. Baek, *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, eabj8754 (2021).

34. Z.-P. Liu, L.-Y. Wu, Y. Wang, X. Zhang, L. Chen, Bridging protein local structures and protein functions. *Amino acids* **35**, 627-50 (2008).

35. A. Sharir-Ivry, Y. Xia, Quantifying evolutionary importance of protein sites: A tale of two measures. *PLOS Genetics* **17**, e1009476 (2021).

36. G. Bartlett, C. Porter, N. Borkakoti, J. Thornton, Analysis of catalytic residues in enzyme active sites. *Journal of molecular biology* **324**, 105-21 (2002).

37. S. Sankararaman, F. Sha, J. F. Kirsch, M. I. Jordan, K. Sjölander, Active site prediction using evolutionary and structural information. *Bioinformatics* **26**, 617–624 (2010).

38. J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose, G. De Fabritiis, DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

39. I. Kozlovskii, P. Popov, Protein–Peptide Binding Site Detection Using 3D Convolutional Neural Networks. *Journal of chemical information and modeling* **61**, 3814–3823 (2021).

40. J. Yang, A. Roy, Y. Zhang, Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **29**, 2588–2595 (2013).

41. I. Kozlovskii, P. Popov, Spatiotemporal identification of druggable binding sites using deep learning. *Communications Biology* **3**, 618 (2020).

42. M. N. Wass, L. A. Kelley, M. J. E. Sternberg, 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research* **38**, W469–W473 (2010).

43. C. Zhang, P. L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research* **45**, W291-W299 (2017).

44. J. Mistry, *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412-D419 (2020).

45. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).

46. D. Probst, *et al.*, Biocatalysed synthesis planning using data-driven learning. *Nat Commun* **13, 964** (2022).

47. G. Pesciullesi, P. Schwaller, T. Laino, J.-L. Reymond, Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature communications* **11**, 1–8 (2020).

48. D. Xiao, *et al.*, Ernie-gram: ransraining with explicitly n-gram masked language modeling for natural language understanding. *arXiv preprint arXiv:2010.12148* (2020).

49. S. Sun, K. Krishna, A. Mattarella-Micke, M. Iyyer, *Do Long-Range Language Models Actually Use Long-Range Context?* (2021), pp. 807–822.

50. A. Vaswani, *et al.*, Attention is all you need. *CoRR* **abs/1706.03762** (2017).

51. *Enzymes Are Wonderful Catalysts* (John Wiley Sons, Ltd, 2012), chap. 3, pp. 26–49.

52. A. Chattopadhyay, N. Diar, D. Flower, A statistical physics perspective on alignment-independent protein sequence comparison. *Bioinformatics (Oxford, England)* **31** (2015).

53. E. Jérôme, D. Santos-Martins, A. Tillack, S. Forli, Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling* (2021).

54. D. M. Lowe, Extraction of chemical structures and reactions from the literature, Ph.D. thesis, University of Cambridge (2012).

55. A. Jäde, *et al.*, Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49** (2020).

56. P. Bansal, *et al.*, Rhea, the reaction knowledgebase in 2022. *Nucleic acids research* **50** (2021).

57. D. Wishart, *et al.*, Pathbank: a comprehensive pathway database for model organisms. *Nucleic acids research* **48** (2019).

58. M. Ganter, T. Bernard, S. Moretti, J. Stelling, M. Pagni, Metanetx.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics (Oxford, England)* **29** (2013).

59. T. U. Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480-D489 (2020).

60. P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, "found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* **9** (2017).

61. D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization* (arXiv, 2014).

62. K. Lu, A. Grover, P. Abbeel, I. Mordatch, *Pretrained Transformers as Universal Computation Engines* (2021).

63. A. Elnaggar, *et al.*, *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing* (arXiv, 2020).

64. S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, M. Schroeder, Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research* **43**, W443–W447 (2015).

65. O. Trott, 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).

66. H. Berman, *et al.*, The protein data bank. *Nucleic acids research* **28**, 235-42 (2000).

67. N. O'Boyle, *et al.*, Open babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 33 (2011).

Supplementary Materials for

Language models can identify enzymatic active sites in protein sequences

Yves Gaetan Nana Teukam *et al.*[1]

[*]yna@zurich.ibm.com

January 12, 2023

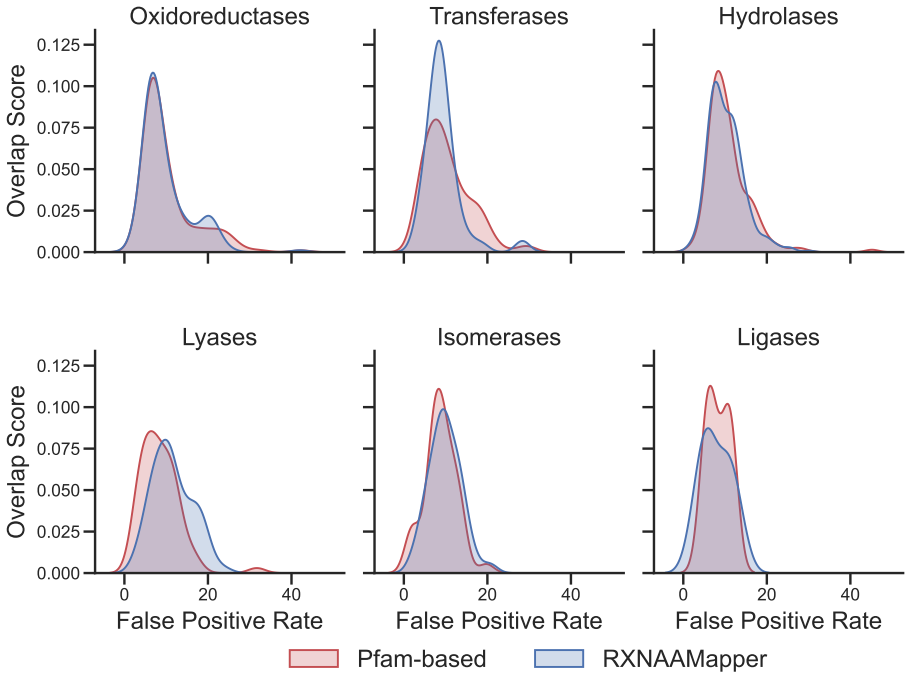
This PDF file includes:

Figure 1 and Table 1

	BPE datasets			
Vocabulary size	**200-900**	**400-500**	**600-750**	**900-1K**
10K	182	185	181	182
20K	170	174	169	170
30K	164	168	163	164
50K	158	162	156	157
75K	153	157	**152**	153

Table 1: Median number of tokens per sequence

(A) EC class level



(B) Reaction class level

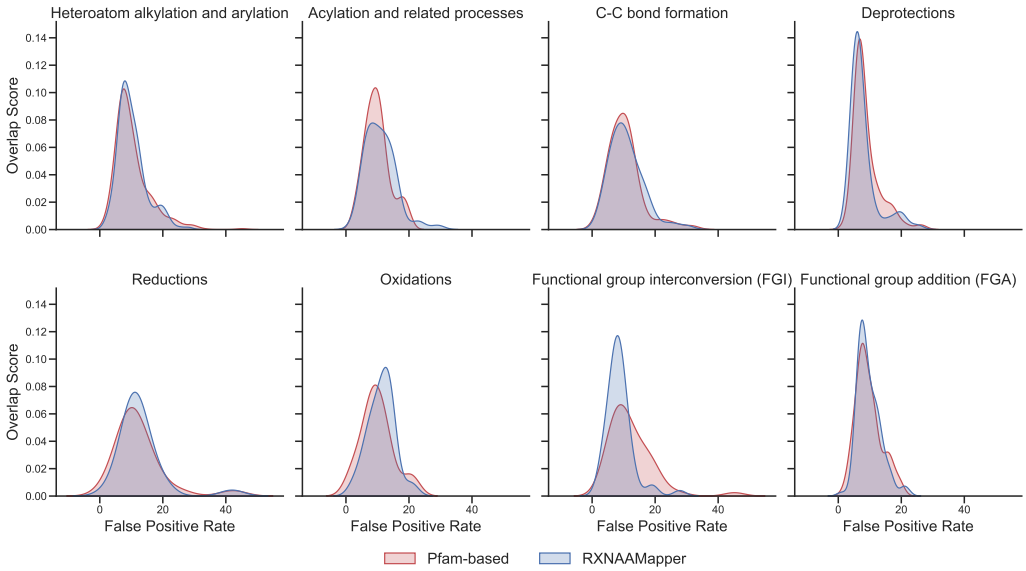


Figure 1: **Distance between the predicted active sites and ground truth**. The distance between the barycenter of the grid boxes centred on the predicted active sites and the ground facts was used to compare our prediction to those from the homology-based. Figure 1A has been computed by grouping the points in out set by EC classes, while Figure 1B on reaction classes.