Computational and Structural Biotechnology Journal 23 (2024) 1929–1937

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Language models can identify enzymatic binding sites in protein sequences

Yves Gaetan Nana Teukam [a,*], Loïc Kwate Dassi [a], Matteo Manica [a], Daniel Probst [a,b], Philippe Schwaller [a,b], Teodoro Laino [a,b]

[a] IBM Research Europe, Saümerstrasse 4, 8803 Rüschlikon, Switzerland
[b] National Center for Competence in Research-Catalysis (NCCR-Catalysis), Switzerland

ARTICLE INFO

Dataset link: https://github.com/rxn4chemistry/biocatalysis-model

Dataset link: https://github.com/rxn4chemistry/rxnaamapper

Dataset link: https://doi.org/10.5281/zenodo.7530180

ABSTRACT

Recent advances in language modeling have had a tremendous impact on 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 encoded within sequential data. Here, 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-binding site atomic interactions. This allows us to identify the three-dimensional binding 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, 52.13% of the binding 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 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.), have also had 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 have shown 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, such as predicting mutational effect, secondary structure [26], long-range contact prediction [27], binding site targeting [28], capturing important biophysical properties governing protein shape [28,66], modeling bacterial activity [29], or predicting peptide binding [30].

The identification of binding and active site residues and the characterization of the corresponding protein function [31–33] pose a significant scientific challenge, particularly following groundbreaking research on protein structure prediction [20,34]. The activity of a protein is directly correlates with the structure of its binding site [35], a spatial region hosting a contiguous or non-contiguous amino acid (AA)

[*] Corresponding author.
E-mail address: yna@zurich.ibm.com (Y.G. Nana Teukam).

https://doi.org/10.1016/j.csbj.2024.04.012
Received 15 February 2024; Received in revised form 5 April 2024; Accepted 5 April 2024
Available online 30 April 2024

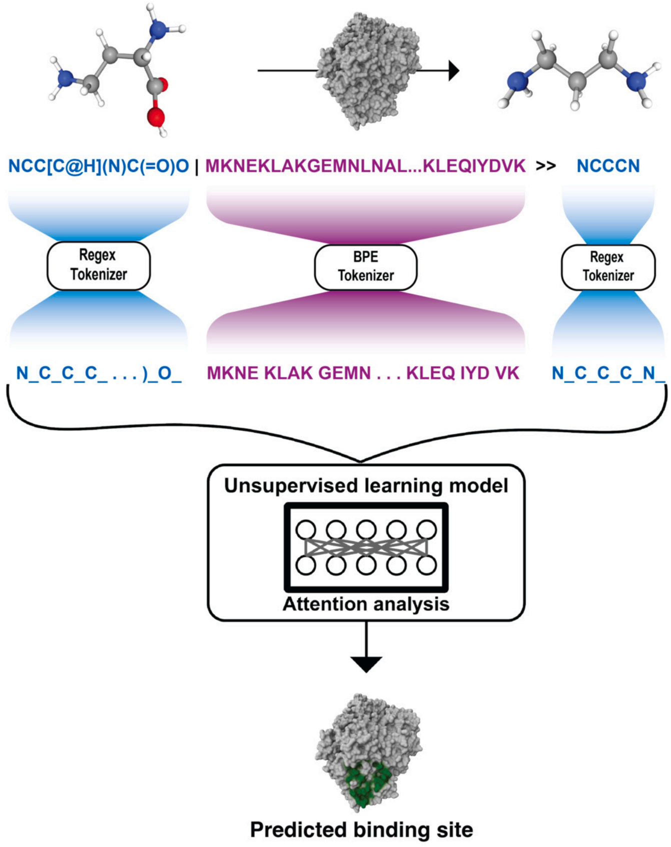
NCC[C@H](N)C(=O)O | MKNEKLAKGEMNLNAL...KLEQIYDVK >> NCCCN
Regex Tokenizer
BPE Tokenizer
Regex Tokenizer
N_C_C_C_ . . .)_O_
MKNE KLAK GEMN . . . KLEQ IYD VK
N_C_C_C_N_
Unsupervised learning model
Attention analysis
Predicted binding site

Fig. 1. RXNAAMAPPER pipeline. A BERT model [1] is trained on a combination of organic and enzymatic reaction SMILES using MTL [52], leveraging atom-level tokenization and MLM [53] 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 binding site of the enzyme bio-catalyzing the reaction with no supervision or structural information. The binding sites are represented in our plot as red regions in the molecule.

sequence. This region has evolved to facilitate selective interactions with specific molecules. The conservation of amino acids in the binding site over the entire sequence during evolution underscores their crucial role in providing unique structural features for enzyme function [36,37]. Learning the signal describing the 3D interaction of amino acids with the target molecules in the binding site from the AA sequence and the molecular representation could enable the prediction of protein function derived from co-homology strategies [32,38–43], or protein-protein interaction networks [44]. Tools like Pfam [45] and PSI-BLAST [46] can help infer active site location based on sequence similarity or preserved domains information. Other approaches, such as those proposed by Zhang et al. [47] and Pande, Raheja, and Livesay [48], leverage machine learning methods like support vector machines (SVM) [49] and multilayer perceptron (MLPs) [50] for catalytic site prediction.

Building on Schwaller's work [24], we demonstrate that Language Models (LMs) can capture the signal characterizing AA binding sites using linguistic representations for proteins and their molecular substrates. We leverage a publicly available dataset of enzymatic reactions, where substrate molecules are represented with SMILES notation and proteins with their AA linear sequences. Unsupervised training can identify up to 52.13% of binding sites when considering co-crystallized substrate-enzyme structures as ground truth.

Inspired by the work of Schwaller et al. [24], here we show that LMs can capture the signal characterizing AA binding sites using a linguistic representation for proteins and their molecular substrates (see Fig. 1). We leverage a publicly available dataset of enzymatic reactions [51], where substrate molecules are represented with SMILES notation and proteins with their AA linear sequences. Unsupervised training can identify 52.13% of binding sites when considering co-crystallized substrate-enzyme structures as ground truth.

Table 1
ECREACT dataset division divided into groups based on the level of information.

Groups	N° 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

2. Methods

2.1. Dataset

We consider a dataset of 1 million organic reactions from USPTO [54], combined with a dataset of bio-catalyzed reactions called ECREACT [51]. The USPTO covers a wide range of chemical reactions, represented as reaction SMILES. This dataset is commonly employed in various applications within the field of chemistry, particularly in the development and training of machine learning models for reaction prediction and analysis due to its accessibility and extensive coverage of organic chemical reactions [24,55–57]. On the other hand, ECREACT contains 62,222 reactions with a unique reaction-EC combination. The entries from this dataset can be classified based on their Enzyme Commission numbers (EC numbers) and grouped as shown in Table 1. ECREACT is the result of the combination of bio-catalyzed reactions from four databases: Brenda [58], Rhea [59], PathBank [60], and MethaNetX [61]. It predominantly features transferase-catalyzed reactions (EC 2.x.x.x), making up around 53.5% of entries due to the inclusion of non-primary lipid pathways from PathBank. Oxidoreductases and hydrolases account for 24.5% and 10.7% respectively, with lyases, isomerases, ligases, and translocases making up the rest. To align with the original ECREACT dataset, we maintained consistent data splits based on products to prevent bias towards specific reaction products during model training. By stratifying data using unique SMILES strings for products, we mitigate biases effectively. While removing duplicate reactions, we retained duplicated sequences or SMILES to expose the model to diverse variations and enhance learning robustness. Prior to data augmentation, our dataset contained 70.8% unique substrate SMILES and 57.9% unique product SMILES.

2.2. Data processing

Using EC numbers as the single filter, we mapped the EC numbers to their corresponding AA sequences from Uniprot [62]. To enrich the training data with additional protein context, we constructed "augmented" reactions for each successful mapping. In these augmented reactions, we kept the original reactants and products from the ECREACT entry, but replaced the EC number with the corresponding retrieved protein sequence. Essentially, this transformed the data representation from a functional classification (EC number) to a more detailed protein sequence-based representation. A filtering step is performed to reduce the overrepresentation of certain EC numbers by limiting the maximum number of sequences per EC number to 10K. In cases where the number of sequences exceeded the limit for a particular EC number, we randomly selected 10K sequences from the set. While the data augmentation approach successfully enriched the training data with additional protein context, our analysis revealed that 96% of the protein sequences in the dataset are unique. Notably, the uniqueness of protein sequences stands out, influenced by the prevalence of certain reactions in datasets like USPTO, emphasizing popularity and organic chemistry biases.

2.3. Tokenization

Language models operate on numerical data, requiring text transformation into a numerical representation. An important step in the conversion is the tokenization step, which involves 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 [63] (character-level tokenization). The selection of the tokenizer for amino acids sequences is more complex. The complexity derives from: (a) the length of the sequences and (b) the limited number of input tokens supported by LMs. To overcome these limitations, we compressed 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 Byte-Pair Encoding (BPE) 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 created subsets contained sequences falling within specific length intervals, including 250-900, 400-500, 600-750, and 900-1K amino acids. Each dataset consists of 400K sequences. By adopting this approach, we aimed to maintain the diversity of our search space while accommodating the unique characteristics of sequences across various length ranges. Additionally, to ensure comprehensive coverage and representation within our training data, we ensured to sample sequences from all EC classes present in the dataset.

2.4. 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 [53] (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 [51] and USPTO [54], Multi-task Transfer Learning (MTL) [52] 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 Fig. 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 [64] 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 [65]. Inspired by this seminal work, here we also decided to include in our study the following models: ProtAlbert [66] and ProtBert [66] (pre-trained on protein data and fine-tune on biocatalyzed data), BERT-base (pre-trained on various data modalities), BERT-base [1] and BERT-Large [1] (trained from scratch on biocatalyzed data).

2.5. Binding site prediction

The prediction of the binding 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 [67]. 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 binding 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 binding region. Hereinafter, the method combined with BERT-base and the BPE will be referred to as RXNAAMapper.

Algorithm 1 Binding site prediction.

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

2.6. Evaluation

We use a set of 777 co-crystallized ligand-protein pairs from the Protein-Ligand Interaction Profiler (PLIP) [68] 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 [45]; (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 binding site as a set of non-overlapping segments in a sequence. If S with $|S| = n$ is a sequence of amino acid residues, the binding region B_s of S is defined as $B_s = \{(a_i, b_i)\}_i^m$, where a_i and b_i are the index boundaries of the segment i. The overlap score $(OS(B, B_s))$ between the predicted binding region $B = \{(a_{pi}, b_{pi})\}_i^n$ and the ground-truth $B_s = \{(a_{si}, b_{si})\}_i^m$ is defined as:

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

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 binding site predictions, we evaluated the binding energy computed with Autodock Vina [69,70] considering predicted binding sites and the

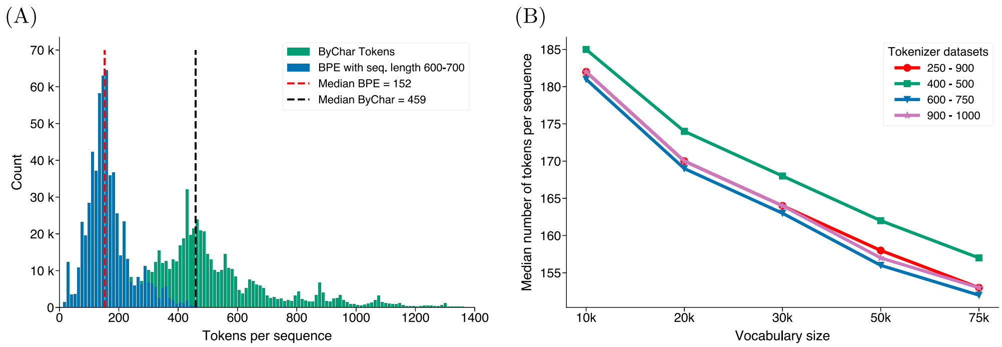

Fig. 2. BPE analysis. The left plot displays the density distribution of token counts obtained using the BPE tokenizer, contrasting it with the ByChar tokenizer used in our model training. Unlike ByChar, BPE splits infrequent fragments into two or more tokens while merging the most frequent ones into longer tokens. This application leads to a sequence shortening effect compared to ByChar tokenization. On the right, the median of BPE token counts is shown for each configuration in our grid search. The results indicate that larger vocabularies lead to more significant data compression.

ground truth from PLIP. We chose these enzyme-ligand pairs by first matching PDBs and amino acid sequences with annotated binding 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 binding sites. We computed the Cartesian coordinates of the ligand and receptor molecules, which are generally retrieved from the PDB [71] or PDBQT [72] 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 binding site where the ligand is to be docked. The box has been found by averaging the 3D coordinates of the atoms of the binding site and setting the box side length to 50 Å.

3. Results

3.1. 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 [73]. This quadratical growth in complexity translates to a dramatic increase in processing time, contrasting the model's ability to analyze the input text in a timely manner. 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 (see Figure S2, 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 ($n = 600K$). Fig. 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. When analyzing the impact of vocabulary size and sequence length, we observed that sequences in the training set ranging from 600 to 700 amino acids, paired with a vocabulary size of 75K, achieved optimal compression with a median token count of 152 (see Fig. 2A and Table S1). This represents a compression rate of 66.8% compared to the baseline character-level tokenizer (ByChar) which tokenizes sequences into individual amino acids and had a median token count of 459. Sequences within the 600-750 amino acid range strike a balance between capturing sequence information effectively and keeping input lengths manageable for language model processing. This range enables efficient compression by the tokenizer while retaining essential details. The use of a 75k vocabulary size allows the tokenizer to cover a wide range of amino acid combinations and rare subword structures, enhancing its ability to represent diverse sequences accurately and contribute to effective compression. Therefore by using this tokenization scheme, we overcome the architectural limitations and train our model on broader corpora.

We rely on two distinct tokenization methods for the AA sequences and the molecular substrates (SMILES). We applied BPE only to AA sequences to compress the resulting token sequence length. On the other hand, the SMILES strings representing substrates naturally allow for tokenization at atoms and bonds level resulting in relatively short token sequences, making extensive compression unnecessary. Therefore we opted for a regular expression-based tokenization approach [55,63].

3.2. Binding 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 developed in this work), BERT-base, BERT-Large, ProtAlbert, and ProtBert (details in methods) and compared them on the task of binding site predictions using 777 amino acid sequences from PLIP (with related binding 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 [67] forming a multi-head attention system. During our analysis we explored how three key factors influenced our model's performance in predicting binding sites. First, different attention heads prioritize different information within the sequence. We found that head 10 achieved the best results, suggesting

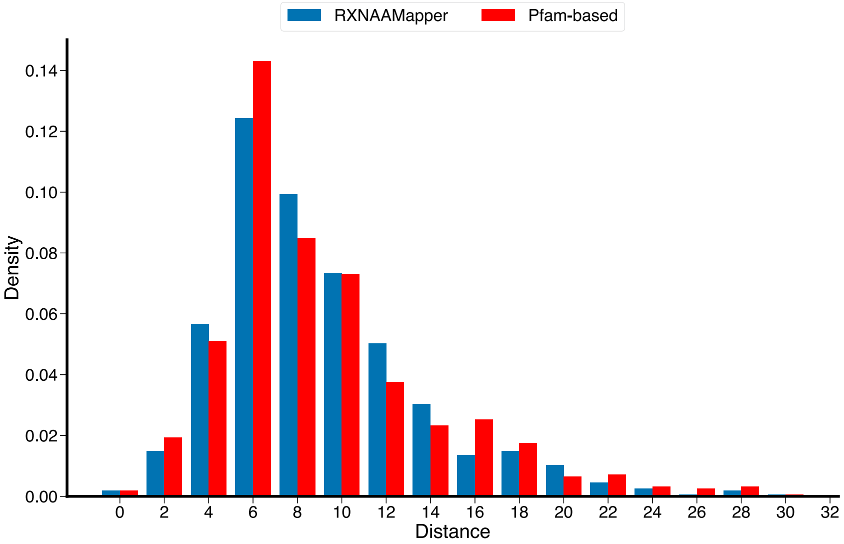

Fig. 3. Binding sites distance from ground truth. Distribution plot depicting the distance of the predicted binding sites from the PLIP annotations. For both predictions, the distribution is right skewed reflecting the correctness of the predictions. RXNAAMapper exhibits a distribution peak at lower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

that its focus on specific relationships was particularly relevant for this task Figure S5. Second, within this model paradigm, information flows through the layers sequentially, refining the model's abstraction. We observed that layer 5 provided the optimal level of context for binding site prediction. Finally, the model computed interactions between individual amino acids and reactant atoms. By focusing on the top 6 atoms, we achieved the most accurate predictions, indicating that capturing local interactions was key. The set of parameters giving the highest overlap between our predictions and ground truth is $head = 10$, $layer = 5$, and $top_k = 5$.

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. Among the binding sites predicted by RXNAAMapper, up to 52.13% overlap with the ground truth whereas ProtAlbert, ProtBert, BERT-base, BERT-Large, and the random model, reached 18.27%, 20.03%, 15.27%, 45.91%, and 14.07%, respectively predictions (Table 2). As a reference, we report the overlap score obtained by homology-based on Pfam annotations (67.37%).

When comparing the performance of the sequence-based models, RXNAAMapper and BERT-Large + BPE, the Wilcoxon-rank test [74] revealed a statistically significant difference in the "Overlap Score" metric ($p < 0.001$), where RXNAAMapper showed a higher overlap compared to BERT-Large + BPE.

Predictions based on homology models, like the one obtained using Pfam annotations, can help recover binding 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 binding sites are too large. This might be attributed to the inclusion in the preserved domains of regions containing the active site, housing both the binding and the catalytic sites. However, active sites usually account for just 10-20% of the volume of an enzyme [75], while Pfam's-based approach predicted [45] on average 61.99% of the sequence as sites of interest. Our model predicted shorter stretches of the input sequences as binding sites (on average 48.06%) while maintaining a lower false positive rate (47.89%) compared to the homology-based.

We inspect the performance of our model and the Pfam-based across different enzyme classes (see Fig. 4A) and reaction classes (see Fig. 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. Our model predictions differ from the homology-based ones, particularly in the case of Lysases. It demonstrates a more conservative approach by generating shorter predictions (on average 44.5 amino acids), leading to a lower false positive rate and overlap score. In contrast, Pfam-based predictions tend to be longer (on average 254 amino acids), contributing to a higher false positive rate. This discrepancy is linked to the limitations of sequence alignments in accurately predicting results when sequence identity falls below a specific threshold [76]. Conversely to 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.

Table 2
Performance on sequence-based binding site prediction. Reported in the table are the overlap score and the false positive rates for the binding 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.

	Overlap Score	False Positive Rate
Random Model	14.07%	13.80%
BERT-base	15.27%	14.92%
BERT-Large + BPE	45.91%	42.71%
ProtAlbert	18.27%	16.23%
ProtBert	20.03%	16.42%
RXNAAMapper (ours)	52.13%	47.89%
Pfam-based	67.37%	61.68%

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

Fig. 5 shows an example of Pfam-based and RXNAAMapper predictions overlapped with the PLIP ground truth. Despite RXNAAMapper

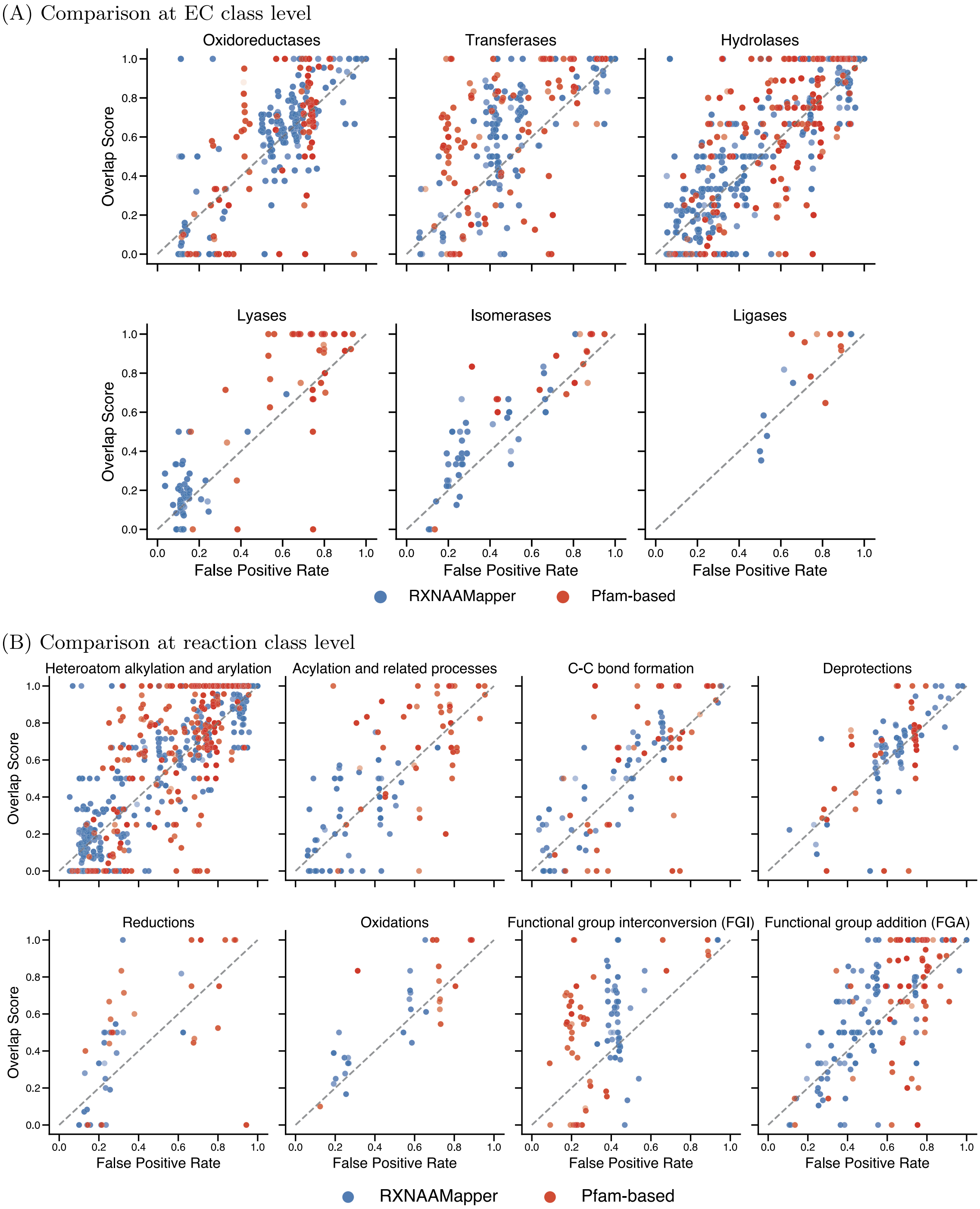

Fig. 4. Pfam-based and RXNAAMapper performances with respect to the EC classes and reaction classes. The models exhibit different performances for different types of reactions and enzymes, underlying the modeling complexity of certain types of reactions and enzymes. For almost every enzyme class and 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 binding 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 binding sites and the one centered on the ground truth. The closer the prediction is to the ground truth, the more opaque the point.

is covering fewer amino acids (35.9%) than the Pfam-based approach (66.3%), it demonstrated a recall rate comparable to Pfam-based predictions. Furthermore, the lower false positive rate observed in RXNAAMapper predictions (33.5% compared to 65.5% in Pfam-based predictions) suggests a more focused prediction of relevant amino acids, minimizing the inclusion of irrelevant ones. To complement our analysis, in the supplementary section we report a comparison of RXNAAMapper with a supervised approach showing competitive results (see Figure S3).

While RXNAAMapper did not surpass Pfam in terms of ground truth overlap, it offers distinct advantages in specific contexts, such as for Transferases and Lyases. Unlike homology-based methods, RXNAAMapper performs better even in scenarios where homologous sequences or functional annotations are scarce or absent, as observed for Lyase en-

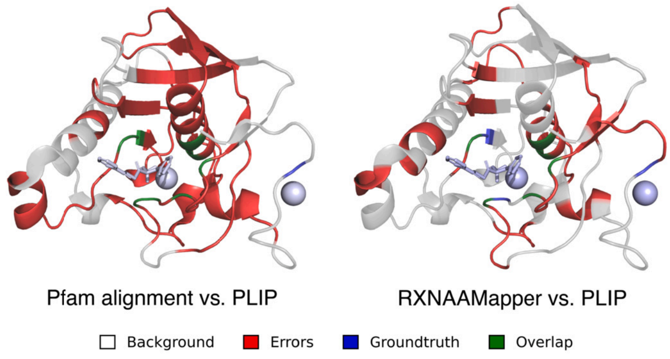

Fig. 5. Experiment results. Comparison of the prediction from Pfam alignments (left) and RXNAAMapper (right) using PLIP as a ground-truth (PDB id: 6JFR) interacting with S-(2-oxo-2-phenylethyl) (2R)-2-benzyl-4,4,4-trifluorobutanethioate (K3U) and Nickel (II) ion). The area in red represents the predicted binding site region, while the blue area represents the ground truth of bindingsites. The white area depicts the backbone of the protein.

zymes. This capability to generalize across different enzyme classes enables RXNAAMapper to navigate diverse biological contexts, contributing to its versatility and applicability beyond the limitations of homology-based approaches.

3.3. Structural validation

A successful chemical reaction hinges on the precise interaction between substrate and enzyme. The active site, delineated as a specific region on the enzyme comprising two crucial subdomains: the substrate-binding site and the catalytic site, fulfills distinct functional roles within the enzymatic mechanism. The binding site secures the substrate in position, while the catalytic site orchestrates the actual chemical transformation. This cooperation not only ensures specificity, but also drives the process by supplying energy to sustain substrate engagement.

The relation between enzyme and substrate within the active site relies heavily on the binding site's ability to hold the substrate in place and provide stabilizing energy. This specificity and stability are crucial for efficient chemical reactions. Therefore, accurately predicting binding sites becomes a key step in understanding and optimizing these interactions.

To evaluate RXNAAMapper's ability to predict effective binding sites, we employed a docking approach. We tested our model on 2213 enzyme-ligand pairs (extracted from our dataset and the PLIP database - see Methods section) and compared the predicted binding energies (using Autodock Vina with RXNAAMapper's sites) with experimental data from PLIP. The results, with an average difference of only 0.37 kcal/mol, demonstrate RXNAAMapper's potential for accurate binding site prediction and its valuable contribution to understanding enzyme-ligand interactions (Fig. 6).

4. Discussion

The prediction of protein binding 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 binding sites in amino acid sequences. When tested on PLIP protein-ligand interactions, our approach outperforms other sequence-based methods by identifying more than 52% of proteins' binding 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 [73,77,78]. 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 information. The use of a BPE tokenizer allowed us to train our model

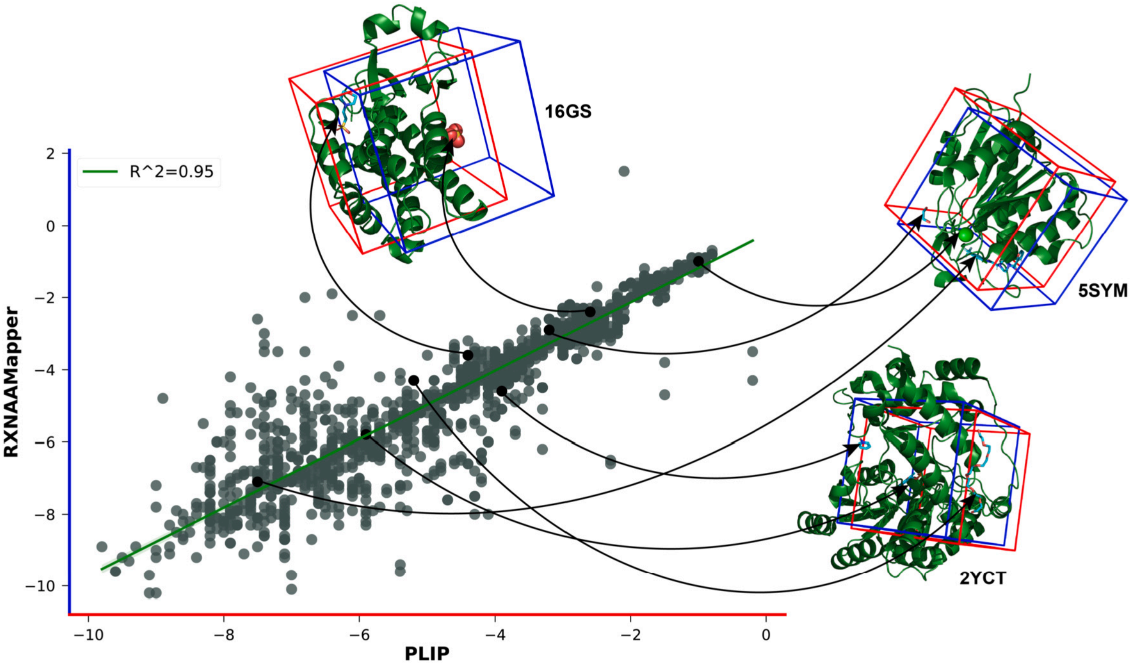

Fig. 6. Negative binding energy of 2213 enzyme-ligand pairs. The figure shows how the energy scores deriving from the RXNAAMapper binding site predictions are in the same range with respect to those predicted from PLIP ($R^2 = 0.95$).

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 predictions of RXNAAMapper, we evaluated the inferred binding sites of several enzyme-ligand pairs, achieving close agreement with experimentally determined sites (from the PLIP database). This ability to accurately identify binding sites is crucial, as it allows for modeling their interactions and predicting binding energies. Crucially, this analysis paves the way for predicting entire active sites. However, to fully evaluate the reconstruction of the entire active site, necessitates accurate annotations of catalytic sites due to their critical role in enzyme function.

Our method demonstrates superior control over false positives and is reflected on the barycenter of our predictions to be spatially closer to the ground truth. This represents a significant advancement as RXNAAMapper, unlike homology methods relying on evolutionary relationships, can predict binding sites for proteins with limited or no annotations. For future advancements in binding site prediction, exploring multi-modal deep learning model integration such as Mixtral 8x7B [79] and implementing fine-tuning strategies customized for biological sequences and interactions [80] are crucial steps to enhance predictive capabilities.

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.

CRediT authorship contribution statement

Yves Gaetan Nana Teukam: Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Loïc Kwate Dassi:** Data curation. **Matteo Manica:** Conceptualization, Supervision. **Daniel Probst:** Data curation. **Philippe Schwaller:** Data curation. **Teodoro Laino:** Conceptualization, Supervision.

Declaration of competing interest

The authors declare that there are no conflict of interest.

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.

Acknowledgements

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.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.04.012.

References

[1] Devlin J, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol. 1. Association for Computational Linguistics; 2019. p. 4171–86.
[2] Su S-Y, Chuang Y-S, Chen Y-N. Dual inference for improving language understanding and generation. https://doi.org/10.48550/ARXIV.2010.04246, 2020.
[3] Peters ME, et al. Deep contextualized word representations. https://doi.org/10.48550/ARXIV.1802.05365, 2018.
[4] Zhuang S, Li H, Zuccon G. Deep query likelihood model for information retrieval. ISBN 978-3-030-72239-5, Mar. 2021. p. 463–70.
[5] Li J, et al. Pretrained language models for text generation: a survey. https://doi.org/10.48550/ARXIV.2201.05273, 2022.
[6] Radford A, et al. Language models are unsupervised multitask learners; 2019.
[7] Brown TB, et al. Language models are few-shot learners. https://doi.org/10.48550/ARXIV.2005.14165, 2020.
[8] Hori T, Cho J, Watanabe S. End-to-end speech recognition with word-based Rnn language models. In: 2018 IEEE spoken language technology workshop (SLT); 2018. p. 389–96.
[9] Xu F, et al. A systematic evaluation of large language models of code; June 2022. p. 1–10.
[10] Wei J, et al. Emergent abilities of large language models. https://doi.org/10.48550/arXiv.2206.07682, June 2022.
[11] Yuan A, et al. Wordcraft: Story Writing With Large Language Models. https://doi.org/10.1145/3490099.3511105, 2022.
[12] Noorbakhsh K, et al. Pretrained language models are symbolic mathematics solvers too!. https://doi.org/10.48550/ARXIV.2110.03501, 2021.
[13] Kojima T, et al. Large language models are zero-shot reasoners. https://doi.org/10.48550/arXiv, May 2022.
[14] Sanh V, et al. Multitask prompted training enables zero-shot task generalization. CoRR, arXiv:2110.08207, 2021.
[15] Zhenzhong L, et al. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv:1909.11942, 2019.
[16] Liu Y, et al. RoBERTa: a robustly optimized BERT pretraining approach; July 2019.
[17] Schwaller P, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci Mar. 2020;11. https://doi.org/10.1039/C9SC05704H.
[18] Vaucher AC, et al. Inferring experimental procedures from text-based representations of chemical reactions. Nat Commun 2021;12:2573. https://doi.org/10.1038/s41467-021-22951-1.
[19] Rao R, et al. MSA transformer. bioRxiv 2021. https://doi.org/10.1101/2021.02.12.430858.
[20] Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature Aug. 2021;596:1–11. https://doi.org/10.1038/s41586-021-03819-2.
[21] Toniato A, et al. Unassisted noise reduction of chemical reaction datasets. Nat Mach Intell 2021;3(6):485–94. https://doi.org/10.1038/s42256-021-00319-w.
[22] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci Feb. 1988;28(1):31–6. https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society.
[23] Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci May 1989;29(2):97–101. https://doi.org/10.1021/ci00062a008. ISSN: 0022020600022. Number: 2 Publisher: American Chemical Society.
[24] Schwaller P, et al. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 2021;7(15):eabe4166.
[25] Wang S, et al. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics; 2019. p. 429–36.
[26] Rives A, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118:15.
[27] Rao R, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst 2019;32:9689.
[28] Vig J, et al. BERTology meets biology: interpreting attention in protein language models. In: International conference on learning representations; 2020.
[29] Hu Y, et al. T4SEpp: a pipeline integrating protein language models to predict bacterial type IV secreted effectors. Comput Struct Biotechnol J 2024.
[30] Yadav S, et al. TCR-ESM: employing protein language embeddings to predict TCR-peptide-MHC binding. Comput Struct Biotechnol J 2024;23:165–73.
[31] Chatterjee A. Protein active site structure prediction strategy and algorithm. Int J Curr Eng Technol June 2017;7:1092–6. https://doi.org/10.14741/Ijcet/22774106/7.3.2017.53.
[32] Yousaf A, et al. Protein active site prediction for early drug discovery and designing. Int Rev Appl Sci Eng 2021;13(1):98–105. https://doi.org/10.1556/1848.2021.00315.
[33] Nguyen-Trinh T-D, et al. Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network. J Mol Graph Model July 2019;92. https://doi.org/10.1016/j.jmgm.2019.07.003.
[34] Baek M, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science July 2021;373:eabj8754. https://doi.org/10.1126/science.abj8754.

[35] Liu Z-P, et al. Bridging protein local structures and protein functions. Amino Acids May 2008;35:627–50. https://doi.org/10.1007/s00726-008-0088-8.

[36] Sharir-Ivry A, Xia Y. Quantifying evolutionary importance of protein sites: a Tale of two measures. PLoS Genet Apr. 2021;17:e1009476. https://doi.org/10.1371/journal.pgen.1009476.

[37] Bartlett G, et al. Analysis of catalytic residues in enzyme active sites. J Mol Biol Dec. 2002;324:105–21. https://doi.org/10.1016/S0022-2836(02)01036-7.

[38] Sankararaman S, et al. Active site prediction using evolutionary and structural information. Bioinformatics Jan. 2010;26(5):617–24.

[39] Jiménez J, et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics May 2017;33(19):3036–42.

[40] Kozlovskii I, Popov P. Protein–peptide binding site detection using 3D convolutional neural networks. J Chem Inf Model Aug. 2021;61(8):3814–23.

[41] Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics Aug. 2013;29(20):2588–95.

[42] Kozlovskii I, Popov P. Spatiotemporal identification of druggable binding sites using deep learning. Commun Biol 2020;3(1):618.

[43] Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res May 2010;38(2):W469–73.

[44] Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res May 2017;45(W1):W291–9. https://doi.org/10.1093/nar/gkx366. eprint: https://academic.oup.com/nar/article-pdf/45/W1/W291/23741003/gkx366.pdf.

[45] Mistry J, et al. Pfam: the protein families database in 2021. Nucleic Acids Res Oct. 2020;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D412/35363969/gkaa913.pdf.

[46] Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res Sept. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389. eprint: https://academic.oup.com/nar/article-pdf/25/17/3389/3639509/25-17-3389.pdf.

[47] Zhang T, et al. Accurate sequence-based prediction of catalytic residues. Bioinformatics 2008;24(20):2329–38.

[48] Pande S, Raheja A, Livesay DR. Prediction of enzyme catalytic sites from sequence using neural networks. In: 2007 IEEE symposium on computational intelligence and bioinformatics and computational biology. IEEE; 2007. p. 247–53.

[49] Hearst M, et al. Support vector machines. IEEE Intell Syst Appl 1998;13(4):18–28. https://doi.org/10.1109/5254.708428.

[50] Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1994.

[51] Probst D, et al. Biocatalysed synthesis planning using data-driven learning. Nat Commun 2022;13:964. https://doi.org/10.1038/s41467-022-28536-w.

[52] Pesciullesi G, et al. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nat Commun 2020;11(1):1–8.

[53] Xiao D, et al. ERNIE-Gram: ransraining with explicitly N-gram masked language modeling for natural language understanding. arXiv:2010.12148, 2020.

[54] Lowe DM. Extraction of chemical structures and reactions from the literature. PhD thesis. University of Cambridge; 2012.

[55] Schwaller P, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(9):1572–83.

[56] Thakkar A, et al. Unbiasing retrosynthesis language models with disconnection prompts. ACS Cent Sci 2023;9(7):1488–98.

[57] Bai R, et al. Transfer learning: making retrosynthetic predictions based on a small chemical reaction dataset scale to a new level. Molecules 2020;25(10):2357.

[58] Jäde A, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res Nov. 2020;49. https://doi.org/10.1093/nar/gkaa1025.

[59] Bansal P, et al. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res Nov. 2021;50. https://doi.org/10.1093/nar/gkab1016.

[60] Wishart D, et al. PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res Oct. 2019;48. https://doi.org/10.1093/nar/gkz861.

[61] Ganter M, et al. MetaNETX.org: a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics (Oxford, England) Jan. 2013;29. https://doi.org/10.1093/bioinformatics/btt036.

[62] Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res Nov. 2020;49(D1):D480–9. https://doi.org/10.1093/nar/gkaa1100. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf.

[63] Schwaller P, et al. “Found in translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci Nov. 2017;9. https://doi.org/10.1039/C8SC02339E.

[64] Kingma DP, Ba J. Adam: a method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980, 2014.

[65] Lu K, et al. Pretrained transformers as universal computation engines; Mar. 2021.

[66] Elnaggar A, et al. ProtTrans: towards cracking the language of life’s code through self-supervised deep learning and high performance computing. https://doi.org/10.48550/ARXIV.2007.06225, 2020.

[67] Vaswani A, et al. Attention is all you need. CoRR, arXiv:1706.03762, 2017.

[68] Salentin S, et al. PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res 2015;43(W1):W443–7.

[69] Jérôme E, et al. AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. J Chem Inf Model July 2021. https://doi.org/10.1021/acs.jcim.1c00203.

[70] Trott O, Olson A. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31.

[71] Berman H, et al. The protein data bank. Nucleic Acids Res Feb. 2000;28:235–42.

[72] O’Boyle N, et al. Open babel: an open chemical toolbox. J Cheminform Oct. 2011;3:33. https://doi.org/10.1186/1758-2946-3-33.

[73] Sun S, et al. Do long-range language models actually use long-range context? Jan. 2021. p. 807–22.

[74] Woolson RF. Wilcoxon signed-rank test. Wiley encyclopedia of clinical trials. John Wiley & Sons, Ltd. ISBN 9780471462422, 2008. p. 1–3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780471462422.eoct979.

[75] Enzymes are wonderful catalysts. In: Introduction to enzyme and coenzyme chemistry. John Wiley & Sons, Ltd. ISBN 9781118348970, 2012. p. 26–49. Chap. 3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118348970.ch3.

[76] Chattopadhyay A, Diar N, Flower D. A statistical physics perspective on alignment-independent protein sequence comparison. Bioinformatics (Oxford, England) Mar. 2015;31. https://doi.org/10.1093/bioinformatics/btv167.

[77] Sanford C, Hsu DJ, Telgarsky M. Representational strengths and limitations of transformers. Adv Neural Inf Process Syst 2024;36.

[78] Alman J, Song Z. The fine-grained complexity of gradient computation for training large language models. arXiv:2402.04497, 2024.

[79] Jiang AQ, et al. Mixtral of experts. arXiv:2401.04088, 2024.

[80] Yuan Q, Tian C, Yang Y. Genome-scale annotation of protein binding sites via language model and geometric deep learning. bioRxiv 2023:2023–11.