Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00055j

Reaction center prediction by analyzing attention of a chemical language model

Xiaoliang Xiong,† Ruizhen Jia,† Yang Tian,† Jingke Chen and Boxue Tian *

Pretrained chemical language models are widely used to predict molecular properties and chemical reactions, yet interpreting their internal attention mechanisms remains difficult. Here, we analyze attention matrices from a chemical language model to extract information relevant to reaction center prediction. We introduce Peak-Activated Binary Attention (PABA), which binarizes attention matrices by retaining only peak values based on a parameter alpha. Using PABA, we identify a top-ranked attention head that effectively predicts reaction centers. We further develop a supervised extension, Supervised PABA (SPABA), which achieves a Matthews correlation coefficient (MCC) of 0.73 and outperforms existing supervised methods for reaction center prediction. SPABA reduces dependence on explicit reaction templates while preserving high accuracy and generalizability, providing a robust framework for reaction center prediction.

Received 31st January 2026
Accepted 6th May 2026

DOI: 10.1039/d6dd00055j

rsc.li/digitaldiscovery

Introduction

Pretrained chemical language models (CLMs) achieve high accuracy in molecular property prediction and reaction prediction tasks.[1–7] These models rely on attention mechanisms and large-scale pretraining on unlabeled SMILES data. For example, Mol2Vec embeds similar molecular substructures into nearby vector representations, reducing sparsity and bit collision associated with traditional fingerprints.[8] SMILES-BERT and Mol-BERT apply self-supervised pretraining to improve predictive performance with limited labeled data.[9,10] ChemBERTa, ChemBERTa-2, and MolRoPE-BERT further enhance SMILES tokenization and positional encoding, leading to improved molecular feature extraction.[11–13] Collectively, CLMs strengthen molecular representations and improve performance on downstream tasks. However, the chemical information encoded within CLMs, particularly within their attention matrices, remains largely unexplored.

In reaction prediction and retrosynthesis planning, reaction center prediction is a critical subtask.[14–17] Reaction centers consist of the atoms and bonds directly involved in chemical transformations.[18] CLMs can predict reaction centers when provided with reactant information. For instance, Wang *et al.* developed Parrot, an attention-based model designed to improve reaction center identification and reaction condition prediction.[19] Lee *et al.* introduced HierRetro, which integrates atom- and bond-level representations to identify reaction centers for accurate retrosynthesis prediction.[20] Wang *et al.* proposed RetroPrime, which treats all atoms as potential reaction centers, predicts synthons, and subsequently converts them into final reactants, enabling effective handling of complex molecular transformations.[21]

Despite the strong predictive performance of CLMs, the chemical information encoded within these models remains largely opaque. Attention mechanisms capture interactions between atomic tokens and therefore offer a potential pathway for interpreting CLMs. In natural language processing, researchers commonly analyze attention by extracting maximum attention weights, averaging attention scores, or constructing maximum-weight spanning trees.[22–24] However, directly applying these approaches to chemical models presents challenges. SMILES representations include non-atomic symbols (*e.g.*, '#', '(·)'), which complicate the association between attention weights and specific atoms. In addition, no consensus exists on how to filter attention weights to retain chemically meaningful information while removing noise. Identifying attention heads that consistently focus on reaction centers, and demonstrating that attention distributions are non-random, are essential steps toward clarifying the functional role of attention mechanisms in CLMs.

In this study, we introduce Peak-Activated Binary Attention (PABA), an unsupervised approach that extracts chemically relevant information from CLMs for reaction center analysis, and Supervised Peak-Activated Binary Attention (SPABA), a supervised model for reaction center prediction. By analyzing attention matrices from representative molecules, we observe that high-frequency molecular fragments in specific attention heads closely correspond to known functional groups. We then assess whether individual attention heads also capture reaction

MOE Key Laboratory of Bioinformatics, State Key Laboratory of Molecular Oncology, Beijing Frontier Research Center for Biological Structure reactive center, School of Pharmaceutical Sciences, Tsinghua University, Beijing, 100084, China. E-mail: boxuetian@mail.tsinghua.edu.cn

† These authors contribute equally to this work.

centers. Using the USPTO benchmark reaction dataset, we find that the 7_7 attention head consistently achieves higher reaction center prediction accuracy than other heads.[25] Guided by this result, we train supervised models with multiple network architectures using the 7_7 attention matrices from each reaction SMILES as input. The best-performing model, SPABA_7_7, reaches a MCC of 0.73, exceeding the performance of LocalTransform (MCC = 0.60).[26] Furthermore, when evaluated on previously unseen reactions published in 2025, SPABA_7_7 accurately predicts reaction centers, demonstrating strong generalizability. Overall, SPABA_7_7 delivers superior accuracy and robustness compared with existing methods, establishing an effective and generalizable framework for reaction center prediction.

Results

Attention analysis and reaction center prediction workflow

We used a CLM developed by Chen *et al.*, pretrained on 700 million unlabeled molecules from the ChEMBL, PubChem, and ZINC datasets, to extract molecular features.[27–30] For each input SMILES string, the CLM produces 64 attention matrices (8 layers × 8 heads), with each matrix having dimensions $L \times L$, where L denotes the number of symbols in the SMILES sequence (Fig. 1a). Because atoms within a functional group interact more strongly with one another than with atoms outside the group, we hypothesized that CLM attention matrices encode functional group information. To test this hypothesis, we constructed MolDataset, which includes 10 000 randomly selected molecules from the ChEMBL, PubChem, and ZINC datasets. We then analyzed the attention matrices by binarizing all 64 attention heads and grouping the resulting molecular fragments according to the number of atoms in each fragment across all molecules in MolDataset. This analysis shows that frequently occurring fragments correspond to known functional groups (Table S1), demonstrating that CLM attention matrices capture functional group information.

Because many functional groups also serve as reaction centers, we further investigated whether attention matrices can predict reaction centers. We randomly selected 1000 reactions from the USPTO_50k dataset to construct a subset for reaction center prediction.[30] We developed the PABA approach, which applies a binarization algorithm to extract chemically relevant information from attention heads. Using PABA, we evaluated the unsupervised reaction center prediction accuracy of all 64 attention matrices for each reaction and ranked the attention heads accordingly. The results show that the 7_7 attention head ranks highest and outperforms all other heads (Fig. 1b). To identify the optimal binarization threshold for this head, we systematically evaluated thresholds ranging from 0.970 to 0.990 and analyzed the corresponding unsupervised prediction

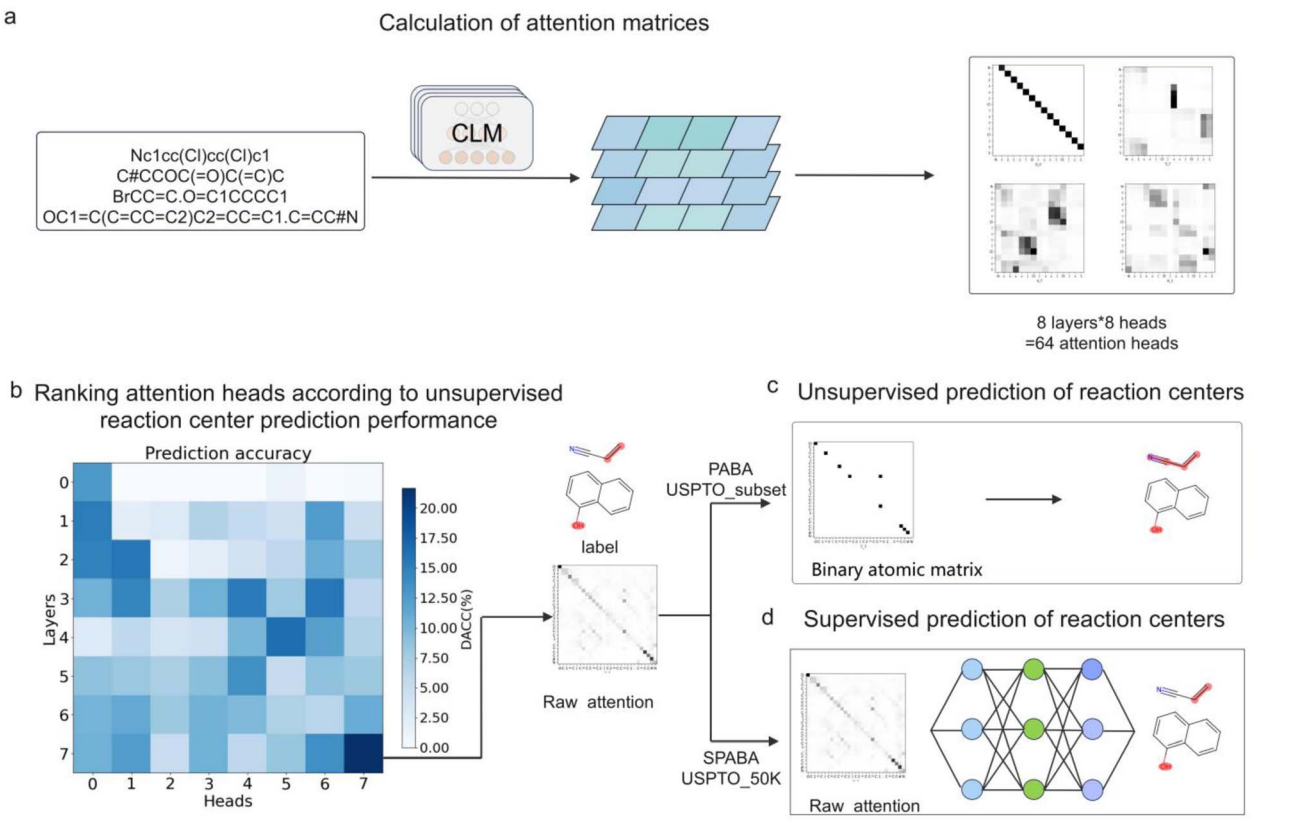

Fig. 1 Workflow for attention analysis. (a) Attention calculation. Molecular SMILES strings are fed into the CLM to obtain raw attention matrices. (b) Ranking of attention heads using PABA. (c) Unsupervised reaction center prediction using PABA. (d) Supervised reaction center prediction using SPABA.

performance (Fig. 1c). To further improve prediction accuracy, we developed SPABA, a supervised framework that uses the raw attention matrix from the 7_7 attention head as input (Fig. 1d).

Unsupervised reaction center prediction with PABA

The values in an attention matrix represent pairwise interactions between symbols. Because not all SMILES symbols correspond to atoms, we first removed all non-atomic symbols (Fig. 2a). We then binarized the raw attention matrices to generate peak-activated binary attention matrices by applying an alpha threshold that retains only the maximum attention values. After isolating matrices containing only atomic information, we projected regions with a value of 1 onto the corresponding atoms and highlighted these atoms using RDKit. Because functional groups strongly influence molecular properties, including chemical reactivity, we hypothesize that frequently occurring functional groups captured by attention matrices may highlight structurally preferred reactive regions and thus provide useful prior information for reaction center prediction. To evaluate the feasibility of reaction center prediction using PABA, we constructed a USPTO subset (USPTO_1K) consisting of randomly selected organic reactions from the USPTO dataset, each containing fewer than 40 atoms. We used the RxnMapper tool to generate atom mappings for reactions and defined reaction centers by comparing changes in atomic environments before and after the reaction. Because non-reaction-center atoms dominate SMILES representations, we used the MCC as the primary evaluation metric to account for label imbalance. In addition, because true reaction centers may vary during byproduct formation, we incorporated neighboring atoms into the evaluation. Specifically, we excluded predictions beyond the $X + 4$ criterion, where X denotes the true reaction center set, and defined this metric as "additionally defined ACCuracy (DACC)" (see Methods).[25]

We systematically evaluated the DACC values of all 64 attention heads for each reaction in the USPTO subset dataset by comparing their predictions with the ground-truth reaction centers and ranking the heads accordingly. This analysis identifies the 7_7 attention head as achieving the highest DACC among all heads (Fig. 2b). To further improve the unsupervised performance of this head, we exhaustively scanned the binarization threshold α across the range from 0.970 to 0.999. When α is set to 0.986, the 7_7 attention head attains its maximum DACC of 28.14% (309 of 1098 reactions) (Fig. 2c). Based on this result, we selected $\alpha = 0.986$ as the optimal binarization threshold for PABA_7_7.

Next, we compared PABA_7_7 with two representative unsupervised baseline methods: the Gasteiger charge method and the Functional Group method. The Gasteiger method identifies reaction centers by estimating atomic charges, whereas the Functional Group method directly assigns functional groups as reaction centers, which can introduce errors in molecules that contain multiple non-reactive functional groups. PABA_7_7 achieves a MCC of 0.29, outperforming both the Gasteiger method (0.23) and the Functional Group method (0.27). Relative to these approaches, PABA_7_7 improves the MCC by 0.06 and 0.02, respectively (Fig. 2d). For the DACC metric, PABA_7_7 reaches 28.14%, while the Gasteiger and

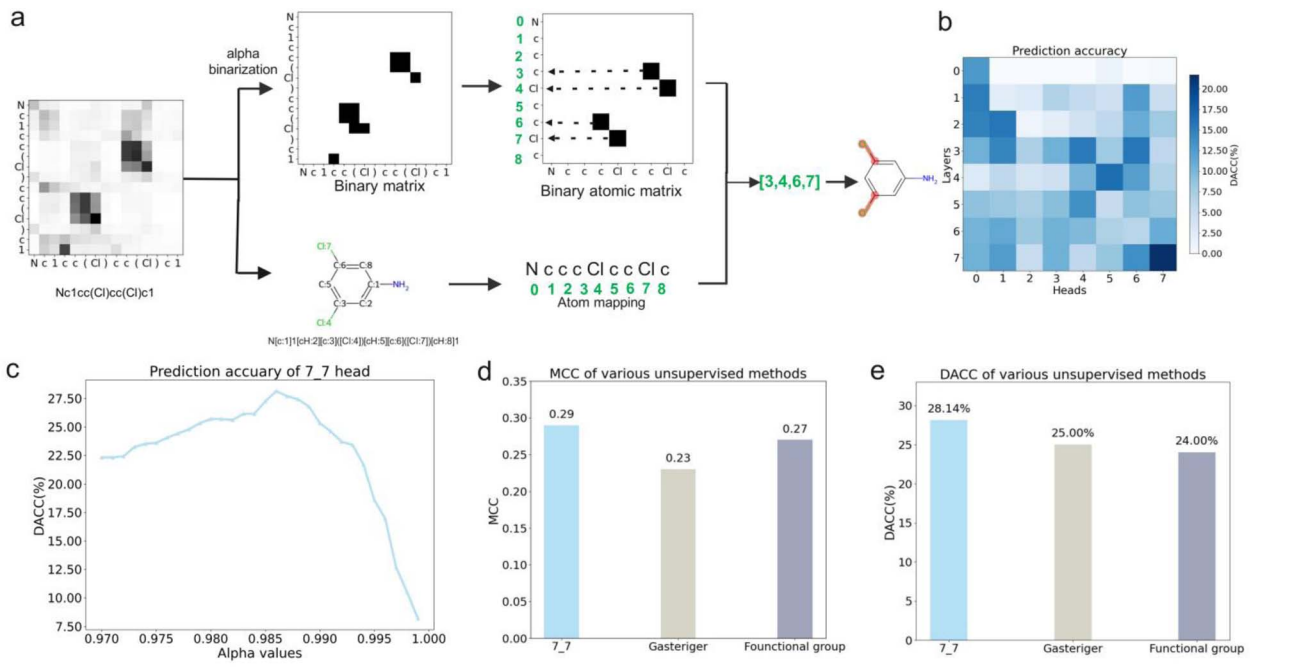

Fig. 2 Unsupervised reaction center prediction with PABA. (a) Binarization of attention matrices. Raw attention matrices are binarized, and non-atomic symbols are removed. (b) Ranking of attention heads by the performance of the unsupervised reaction center predictions. DACC values for all 64 attention heads are computed and displayed in a heatmap. (c) Threshold optimization for the top-ranked head (7_7). A sweep of the binarization threshold α was performed from 0.970 to 0.999. (d and e) Performance comparison between the PABA_7_7 model and other unsupervised methods in terms of MCC (d) and DACC (e).

Functional Group methods achieve 25.00% and 24.00%, respectively (Fig. 2e). Accordingly, PABA_7_7 demonstrates DACC gains of 3.14% over the Gasteiger method and 4.14% over the Functional Group method. Together, these results show that PABA_7_7 delivers higher accuracy than existing unsupervised approaches and indicate that attention head 7_7 encodes chemically meaningful information relevant to reaction center prediction.

Supervised reaction center prediction with SPABA

Because the unsupervised PABA_7_7 model achieves an MCC of only 0.29, its predictive capability remains limited. We therefore sought to improve performance through supervised training using the 7_7 attention head (Fig. 3a). Specifically, we input molecules from the USPTO_50K dataset into a pretrained chemical language model and extracted the corresponding 7_7 attention matrices. To identify the optimal network architecture, we evaluated three model types—fully connected (FC), convolutional neural network (CNN), and Transformer—yielding five architectural variants (Fig. 3b). Based on MCC, the 2FC, 2CNN+2FC, 2CNN+3FC, 3CNN+2FC, and Transformer architecture achieved values of 0.64, 0.66, 0.65, 0.66, and 0.73, respectively. Their corresponding DACC values were 52.34%, 58.23%, 57.29%, 58.21%, and 74.56% (Fig. 3c). The Transformer-based model delivers the highest performance and is therefore designated SPABA_7_7. In addition, we compared SPABA_7_7 with models trained using embedding attentions (SPABA_embedding) and averaged attentions (SPABA_average). SPABA_7_7 improves performance by 0.04 and 0.16 over these

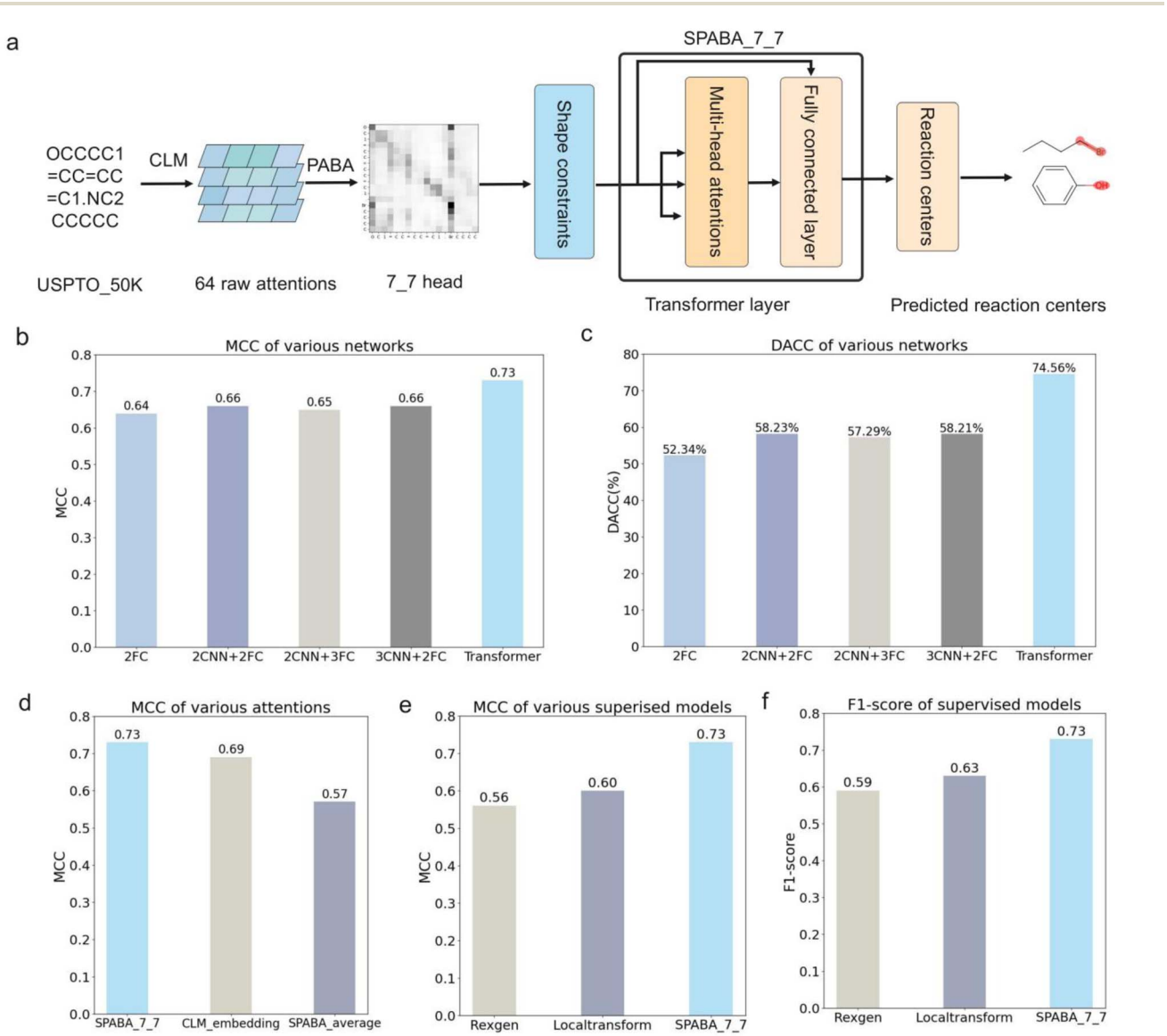

Fig. 3 Supervised reaction center prediction with specific attention heads. (a) Supervised learning workflow with using a transformer architecture with an optimal attention head. (b and c) Performance of different network architectures as measured by MCC (b) and DACC (c). (d) Performance of different attention processing methods as measured by MCC. (e and f) Performance of different supervised models as measured by MCC (e) and F1-score (f).

two approaches, respectively, demonstrating that direct use of the 7_7 attention matrix yields the most accurate reaction center predictions (Fig. 3d).

To assess the predictive performance of SPABA_7_7, we compared it with two representative supervised methods, Rexgen and LocalTransform, using identical training and testing datasets.[25,31] Rexgen formulates reaction prediction as a graph-editing task by identifying which chemical bonds are broken or formed. LocalTransform first extracts reaction templates from the dataset, identifies reaction centers, and then trains a graph neural network to predict these centers. The results show that Rexgen achieves an MCC of 0.56 and an F1-score of 0.59, while LocalTransform reaches an MCC of 0.60 and an F1-score of 0.63 (Fig. 3e). In contrast, SPABA_7_7 attains an MCC of 0.73 and an F1-score of 0.73 (Fig. 3f). Relative to Rexgen and LocalTransform, SPABA_7_7 improves MCC by 0.17 and 0.13, respectively, and increases the F1-score by 0.14 and 0.10. These results demonstrate that SPABA_7_7 delivers substantially higher accuracy for reaction center prediction than existing supervised approaches.

Case study of reaction center Prediction

To evaluate the generalizability of the SPABA model, we selected four types of newly reported organic synthesis reactions published in 2025 that include both major and minor product information as test cases. We then performed a comparative analysis of SPABA, Rexgen, and LocalTransform to assess their ability to identify the core reaction centers of the major products. Fig. 4a shows a representative example of a nickel-catalyzed Suzuki–Miyaura cross-coupling reaction between aliphatic methanesulfonates and arylboronic acids. This reaction overcomes the traditional limitation of Suzuki–Miyaura couplings that require halogenated electrophiles and, for the first time, employs aliphatic alcohol derivatives (methanesulfonates) as coupling partners to form C(sp3)–C(sp2) bonds.[32] The core reaction centers for the major product formation include the saturated carbon atom in compound 1 bonded to the methanesulfonyloxy group (–OMs) and the aromatic carbon atom in compound 2 attached to the boron atom. Although both Rexgen and LocalTransform identify these core reaction center atoms, only SPABA produces predictions that are fully consistent with the ground-truth annotations. Fig. 4b presents an example of a one-step synthesis of *m*-phenylenediamine derivatives through a tandem reaction of cyclohexenones with secondary amines, proceeding *via* a sequence of 1,2-addition, 1,4-addition, and dehydrogenation steps. This strategy eliminates the need for pre-installed substituents or complex directing groups.[33] The core reaction centers for the major product formation include the carbon–carbon double bond and carbonyl group in compound 3, as well as the amino nitrogen atom in compound 4. In this case, Rexgen and LocalTransform each identify only a subset of the true reaction centers, whereas SPABA accurately predicts all core reaction center atoms.

Fig. 4c presents a reaction in which an originally minor phenol oxidation pathway becomes the dominant pathway through precise control of reaction conditions, enabling the efficient synthesis of novel pyrroloquinone monoketals.[34] The core reaction centers for formation of the major product include the hydroxyl group and its directly bonded carbon atom in compound 5, as well as the two adjacent aromatic carbon atoms connected to the phenolic hydroxyl group. In this case, the Rexgen model incorrectly identifies the nitrogen atom and one carbon atom in the Boc group as reaction centers, while the LocalTransform model recognizes only the hydroxyl group and its attached carbon atom. In contrast, the SPABA model accurately identifies all core reaction center atoms. Fig. 4d shows an example of ruthenium-catalyzed direct C–H phosphorylation of indoles at the C4 position.[35] This strategy eliminates the need to pre-protect the highly reactive C2 and C3 positions and overcomes the limitations of traditional approaches that depend on directing groups or fail to achieve selective C4 functionalization. The core reaction centers for formation of the major product are the C4 carbon atom of the indole ring in compound 6 and the phosphorus atom in compound 7. The results indicate that the Rexgen model produces incorrect predictions for the reactive atoms in both molecules, and the LocalTransform model identifies only the phosphorus and oxygen atoms in compound 7. By contrast, the SPABA model accurately predicts all core reaction center atoms. Across all cases examined, only the SPABA model consistently and correctly identifies the complete set of reaction center atoms responsible for major product formation.

Discussion

The attention mechanism, a central component of large language models, encodes relationships between tokens by quantifying their pairwise interactions. In the PABA framework, unsupervised learning based on the binary attention matrix of the 7_7 attention head enables prediction of bimolecular and trimolecular reaction centers. Because CLMs are trained exclusively on single-molecule data, the ability of the 7_7 attention head to capture multi-molecular reaction centers likely arises from atomic interaction patterns learned from single-molecule SMILES representations.[26,36,37] By exploiting these learned atomic interactions encoded in the 7_7 attention head, the SPABA approach achieves high accuracy in reaction center prediction. Beyond reaction center identification, our findings further indicate that attention matrices may be broadly applicable to other tasks, including molecular property prediction and single-step retrosynthesis.[10,11,38–41]

Because SPABA_7_7 achieves higher accuracy than supervised models that rely on CLM embeddings, our strategy of using a specific attention head for a defined task can be readily extended to other language models and application domains. For example, this approach could support solubility prediction using protein language models (PLMs). We have recently applied related strategies to identify protein words and to use these protein words to predict protein-small molecule interactions.[42] More broadly, attention mechanisms can enable systematic investigation of whether, and how, specific molecular features influence properties such as solubility, stability, and reactivity, thereby supporting the rational design of

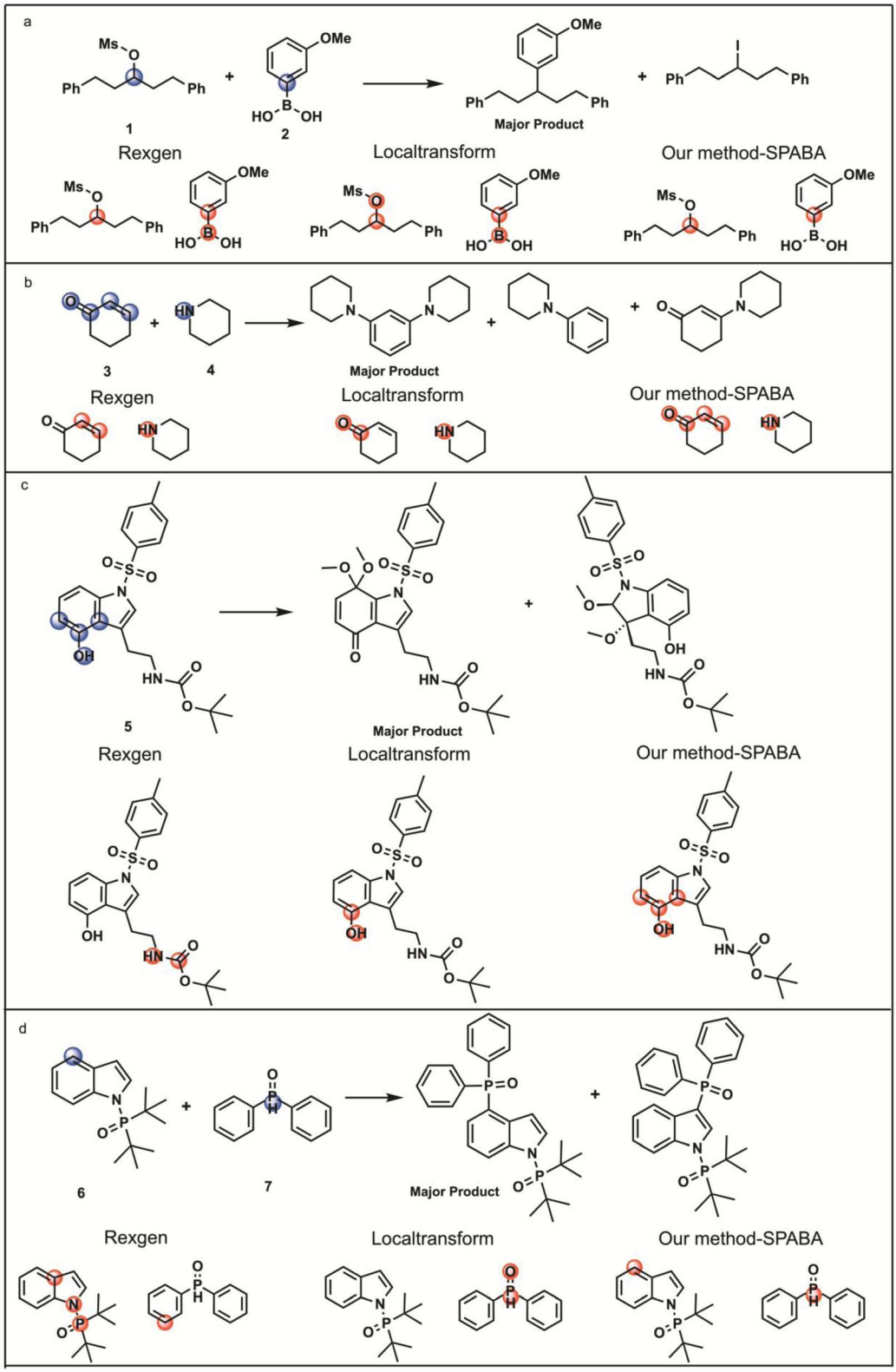

Fig. 4 | Case study of reaction center Prediction. (a) Example of the nickel-catalyzed Suzuki–Miyaura cross-coupling reaction between aliphatic methanesulfonates and arylboronic acids. (b) Example of the one-step synthesis of *m*-phenylenediamine derivatives *via* the tandem reaction of cyclohexenones with secondary amines. (c) Example of the synthesis of novel pyrroloquinone monoketals by converting originally minor phenol oxidation pathways into major pathways. (d) Example of the ruthenium-catalyzed direct C–H phosphorylation of indoles at the C4 position.

molecules with desirable characteristics. In addition, attention matrices can help identify key binding sites involved in complex protein-molecule interactions, improving mechanistic understanding of molecular recognition. This capability is particularly valuable in drug discovery, where accurate prediction of binding affinity between drug candidates and their targets remains a central challenge. Extending the PABA framework to these applications could substantially improve the efficiency and effectiveness of molecular and drug design.

Conclusion

In conclusion, we propose the PABA approach, which reveals chemically meaningful information encoded in pretrained CLMs through systematic analysis of attention matrices. By examining attention matrices across reactions in the USPTO_1K dataset, we identify the 7_7 attention head as exhibiting substantially higher relevance to reaction centers than all other heads. Further analysis based on binarization of attention weights using the parameter alpha shows that the 7_7 attention head achieves the strongest predictive performance in an unsupervised setting, outperforming traditional unsupervised methods such as the Gasteiger charge and functional group approaches. Building on this insight, the SPABA model attains an MCC of 0.73 and delivers higher prediction accuracy and overall performance than both Rexgen and LocalTransform. Moreover, evaluation on previously unseen four representative reactions demonstrates that SPABA accurately identifies reaction centers in novel reactions, highlighting its strong generalizability. Together, these results establish a new methodological framework for reaction center prediction that offers clear advantages in both accuracy and generalizability over existing approaches. Given that USPTO data is patent-derived, it is naturally biased toward successful or common reaction types, which is a limitation of the current study. Notably, the exact reaction center may depend on the reaction conditions, which are not reflected in the training data containing only molecules. Future work incorporating reaction conditions into the prediction framework is needed.

Methods

Samples and attention data collection

Three datasets were constructed for attention data collection. A molecular dataset containing 10k molecules was used for molecular fragment extraction. The USPTO-1k dataset was employed to identify and determine attention heads associated with reaction centers. The USPTO-50k dataset was used for supervised learning, including model training, validation, and testing. To ensure structural alignment, all SMILES notations were converted into hydrogen-depleted forms using RDKit. Versions of the software libraries were Python:3.12.1, Numpy 1.26.4, Pandas:2.2.1, Rdkit: 2023.09.6, Pytorch:2.2.2, Seaborn:0.13.2, Scipy:1.1.3.2. Additionally, the CLM used in this study (https://github.com/WeilabMSU/Pretrain-Models/blob/main/README.md) does not provide an attention interface. Attention matrices for all layers and heads were computed using eqn (1). We extracted the 8 $Q \times K^T$ heads matrices from each layer of the model's functional function, sequentially saved the data from 8 layers, and ultimately uploaded it to the main function (Fig. 1a). A total of 64 heads' attention matrices were preserved for each sample.

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q,K, V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

The PABA framework

During the training of CLM, special tokens '<*s*>' were added to both ends of the SMILES strings. To align the number of atoms with the coordinates, we removed the two outermost rows and columns of the attention matrix. Additionally, to prevent inconsistencies between the size of the attention tokens matrix and the number of atoms in the SMILES strings, we constrained elements in the input SMILES strings that contained multiple atoms (*e.g.*, R–Cl and R–Br) to be treated as single elements. Subsequently we generated grayscale images with their corresponding atomic coordinates (Fig. 2a). By applying a binarization threshold alpha to the raw attention, we obtained binarized matrices containing the critical information.[43–47] Since RDKit only numbers atoms in SMILES strings, non-atomic symbols (*e.g.*, numbers, parentheses, and "@") were removed to prevent interference when matching binary matrix atoms with SMILES atoms. Non-atomic symbols such as '1' and ' = ' in SMILES are used to represent topological relationships or bond types but do not correspond to actual atoms. In our SMILES processing pipeline, these symbols are removed by precisely identifying their positions using a Python script. After attention binarization, we retain only tokens corresponding to atoms, thereby enabling alignment from token-level attention to atom-level attention. Examples are shown in Fig. S1. Non-atomic symbols is deleted during the atomic mapping step in the PABA, which does not affect the original source data. After the above operations, the algorithm identified positions with a value of 1 in the binary atomic matrix and projected them onto the vertical axis to locate attention atoms. Using RDKit's Draw function, the reactive center atoms were highlighted based on SMILES numbering and attention coordinates (Fig. 2a).

To extract the most critical information from the attention mechanism, we applied an alpha threshold to binarize the raw attention matrix: elements greater than the threshold were set to 1, while those below the threshold were set to 0. Specifically, the two-dimensional attention matrix was first flattened into a one-dimensional vector, and each element was ranked and assigned a rank value (eqn (2)). The algorithm used $\alpha|A|$ as the cutoff, converting values to 1 if above cutoff and 0 if below in a same-length one-dimensional matrix. Finally, this one-dimensional matrix was reshaped into a binarized matrix with the same dimensions as the original attention matrix.

$$B_{ij} = \begin{cases} 1, & M_{ij} \leq \alpha|A| \\ 0, & M_{ij} > \alpha\,|A| \end{cases} \quad (2)$$

where $|A|$ denotes the total number of elements in the attention matrix, M_{ij} denotes the rank index of $|A|$ among all attention elements, and α is the binarization threshold ratio.

The choice of the binarization threshold alpha impacted the results: while a higher alpha value filtered out more minor information, an excessively high alpha led to the loss of critical information. In the current sample, when alpha was set to 0.99, the binarized matrix accurately captured reactive center the most critical information in the raw attention, specifically the block structures that were significantly darker than other regions (Fig. S2b). However, further increasing the alpha value (*e.g.*, 0.999) caused the dark regions in the binarized matrix to shrink and lost the most critical information, even resulting in the disappearance of functional groups (*e.g.*, R–Cl) (Fig. S2c). Conversely, when alpha was reduced to 0.97, minor information (*e.g.*, atoms on the benzene ring) began to appear (Fig. S2a). Therefore, the setting of the alpha value must be optimized based on the requirements of the specific task to ensure the retention of critical information while avoiding the introduction of excessive noise.

Extracting molecular fragments

After processing the attention matrices of MolDataset through binarization, projection, and vertical coordinate processing, we obtained the molecular fragments of interest and their coordinates. To identify functional groups in fragments, we built a dictionary of 25 functional groups containing their names, SMARTS patterns, simplified formulas, and atom counts (Table S1). Using the 'HasSubstructMatch (RDKit)' method, we matched the molecular fragments of interest with the SMARTS in the functional group dictionary and output functional groups. For ease of statistical analysis, we compiled the SMILES strings, the SMILES fragments generated from the coordinates, and the matched functional groups into the dictionary.

To improve the identification of functional groups in molecular fragments, we filtered attention samples by comparing fragment the number of atoms with functional group sizes. For example, though fragments "cccccc" and "ccccccC" both match benzene rings, the extra "C" atom in the latter introduced identification errors. To address this, we restricted benzene detection to 6-atom fragments for precise identification. Following the above approach, we classified samples into 6 categories by fragment atom count and matched them to functional group sizes. Subsequently, we sorted the number of matched functional groups in each class for quantitative analysis and visualization (Fig. S3). To avoid imbalance between functional groups and fragments, we log-transformed sample counts and showed only the top 30 groups. Meanwhile, we aggregated the functional groups of all fragments and arranged them in a specific order (Fig. S4 and S5).

Finding the heads position for predicting reaction center

To accurately identify the head positions for reaction center prediction, we constructed the USPTO-1k and annotated reaction atom criterion based on actual chemical reaction to establish the ground truth. Using the PABA algorithm with a fixed alpha of 0.986, we binarized the attention matrix, and projected it onto the vertical axis to obtain the positions of reactive atoms. Subsequently, we saved the SMILES, positions of reactive atoms (predicted atoms), and the positions of the heads. To identify the most accurate attention heads among 64 heads across 1k samples, we compared their predictions with the ground truth. Since predictions with excess reactive atoms were meaningful for some reactions, we defined AtomError as the number of predicted atoms that were below or above the ground truth value (eqn (3)). As the true reaction center might shift during byproduct formation, adjacent atoms were also considered in the prediction evaluation. We defined DACC as the proportion of samples with an atom error satisfying $0 \leq$ AtomError ≤ 4, and excluded all predictions that violated the $X + 4$ criterion (eqn (4)).

$$\text{AtomError} = \begin{cases} |P| - |G|, & if\ G \subseteq P \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

where P denotes the predicted atom set, G denotes the ground-truth reaction-center atom set, and $|\cdot|$ represents the number of atoms in the set.

$$\text{DACC} = \frac{|\{i|0 \leq \text{AtomError}_i \leq 4|\}|}{N} \tag{4}$$

where i represents the index of each sample and N represents the total number of test samples.

We established two variables, flag and head location, to store the AtomError of a specific attention mechanism and the corresponding head position, respectively (Fig. S6). If a head had zero error compared to the ground truth (AtomError = 0), flag_0 was set to 1, and the position of the head was saved. To rank the heads, we aggregated all heads with atom error of 0–3 and identified the top two heads. In the USPTO-1k dataset, statistical analysis showed that the top-ranked 7_7 attention head appeared 309 times, whereas the second-ranked attention head appeared 258 times (Fig. S7c and S7d). Therefore, the 7_7 attention head exhibited significantly higher saliency at reaction centers than other attention heads and was thus selected as the focus of subsequent studies.

Finding the alpha values for predicting reaction center

To precisely determine the optimal value of the binarization threshold alpha, we conducted a traversal of alpha values ranging from 0.970 to 0.999 for 7_7 heads. Based on the flag mentioned earlier, we counted the occurrences of 7_7 head under $X + 4$ condition (Fig. S7a). The results show that as alpha increased from 0.97 to 0.986, the number of samples with 0–2 atom errors consistently increased, while those with 3–4 atom errors exhibited only a slight decrease. When α was further increased beyond 0.986, a pronounced decline was observed in samples with 3–4 atom errors. Therefore, $\alpha = 0.986$ corresponded to the maximum total number of samples within $X + 4$ condition. For the 7_7 head, at $\alpha = 0.986$, the sample counts for atom errors values of 0–4 were 17, 54, 76, 82, and 80, respectively, with a total of 309 correctly predicted samples (Fig. S7b). Since the total number of correctly predicted samples 7_7

reached a peak of 309 at $\alpha = 0.986$, we ultimately determined 0.986 to be the optimal value for alpha (Fig. 2c).

The SPABA framework

The USPTO_50K dataset was first processed using a pretrained model to obtain attention-based representations. The data were then randomly split into training, validation, and testing sets with a ratio of 8 : 1 : 1 for network training and evaluation. Considering that both the labels (z) (eqn (5)) and prediction targets (y) (eqn (6)) are binary, we employed BCEWithLogitsLoss (eqn (7) and (8)), which integrates the sigmoid activation with binary cross-entropy in a numerically stable manner, effectively mitigating overflow and underflow issues caused by extreme logit values. Subsequently, we conducted a systematic hyperparameter search, including optimization of the network architecture and learning rate.

Let the model output logits be:

$$z = (z_1, z_2, \ldots, z_N) \tag{5}$$

and the corresponding ground-truth labels be:

$$y = (y_1, y_2, \ldots y_N),\ y_i \in \{0, 1\} \tag{6}$$

where N denotes the number of atoms or prediction dimensions.

The BCEWithLogitsLoss is defined as:

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log\delta(z_i) + (1 - y_i)\log(1 - \sigma(z_i))] \tag{7}$$

where the sigmoid function $\sigma(\cdot)$is given by:

$$\delta(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

To identify the optimal network architecture, we systematically designed and assessed multiple neural network architectures, including FC, CNN, and Transformer-based models (Fig. S9). We began by constructing a baseline model consisting of 2FC. On this basis, two convolutional layers were introduced to form the 2CNN+2FC architecture, enabling evaluation of the impact of convolutional operations on model performance. Subsequently, the network complexity was further increased by adding one additional fully connected layer (2CNN + 3FC) or one additional convolutional layer (3CNN + 2FC) to the 2CNN + 2FC structure, allowing for a systematic analysis of how increased architectural complexity affects prediction performance.

In parallel, a Transformer architecture was introduced as a comparative model to assess the effectiveness of self-attention-based designs for this task. The input length is first constrained to 512 × 512 and subsequently encoded by a two-layer Transformer architecture. With residual connections and nonlinear activation applied, the network produces a 512 dimensional output vector. Detailed formulations are described below.

Input representation

Each molecule is represented as an embedded SMILES sequence forming an input tensor (eqn (9)):

$$X \in R^{B\times L\times d} \tag{9}$$

where B denotes the batch size, $L = 512$ is the number of tokens, and $d = 512$ is the embedding dimension. We denote the hidden representation at the l-th layer, with $H^{(0)} = X$.

Transformer encoder

The embedded SMILES representations are processed by a multi-layer Transformer encoder. The model consists of $N = 2$ Transformer encoder layers.

Multi-Head Self-Attention (MHSA) (eqn (10)–(14)). Given the input $H^{(l-1)}$, the h-th attention head is computed as:

$$Q_h = H^{(l-1)}W_h{}^{Q} \tag{10}$$

$$K_h = H^{(l-1)}W_h{}^{K} \tag{11}$$

$$V_h = H^{(l-1)}W_h{}^{V} \tag{12}$$

where W_h^Q, W_h^K, $W_h{}^V \in R^{d\times d}$ and $d_{kk} = d/H$ with $H = 8$ attenton heads.

The scaled dot-product attention is defined as:

$$\mathrm{Attention}(Q_h, K_h, V_h) = \mathrm{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)V_h \tag{13}$$

The outputs of all heads are concatenated and linearly projected:

$$\mathrm{MHSA}(H^{(l-1)}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H)W^O \tag{14}$$

Residual connection and layer normalization(eqn (15))

$$Z^{(l)} = \mathrm{LayerNorm}(H^{(l-1)} + \mathrm{MHSA}(H^{(l-1)})) \tag{15}$$

Feed-forward network(eqn (16)–(18))

$$\mathrm{FFN}(Z^{(l)}) = \max(0, Z^{(l)}W_1 + b_1)W_2 + b_2 \tag{16}$$

The output of the l-th encoder layer is:

$$H^{(l)} = \mathrm{LayerNorm}(Z^{(l)} + \mathrm{FFN}(Z^{(l)})) \tag{17}$$

After $N = 2$ layers, the encoder produces:

$$H \in R^{B\times 512\times 512} \tag{18}$$

Flattening operation

To enable atom-level prediction, the encoder output is flattened into a one-dimensional vector (eqn (19)):

$$H_{\mathrm{flat}} = \mathrm{Flatten}(H) \in R^{B\times(512\cdot 512)} \tag{19}$$

Fully connected prediction. The flattened representation is passed through two fully connected layers (eqn (20)–(22)).

First, a hidden representation is computed using a ReLU activation:

$$Z = \mathrm{ReLU}(H_{\mathrm{flat}} W_{\mathrm{f}}^{(1)} + b_{\mathrm{f}}^{(1)}),\ W_{\mathrm{f}}^{(1)} \in R^{(512^2)\times 1024} \tag{20}$$

Dropout regularization is then applied:

$$Z' = \mathrm{Dropout}(Z) \tag{21}$$

The final output of Transformer network is obtained as:

$$Y = Z' W_{\mathrm{f}}^{(2)} + b_{\mathrm{f}}^{(2)},\ Y \in R^{B\times 512} \tag{22}$$

Atom-level reaction center predictions

Transformer outputs are projected onto SMILES to identify reaction-center atoms.

The sigmoid function constrains the output to the range (0,1) (eqn (23))

$$\hat{Y} = \delta(Y) = [\hat{y}_{i,j}] \in (0,1)^{B\times 512} \tag{23}$$

Constrain the output to the length of the SMILES (eqn (24)). The valid SMILES length of the i-th sample is g_i.Then the set of predicted reaction center atoms for the i -th sample is represented as:

$$\hat{y}^{(i)} = [\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,gi}] \in \{0, 1\}^{gi} \tag{24}$$

Reaction centers are identified by thresholding (eqn (25)):

$$\tilde{y}_{i,j} = \begin{cases} 1, \hat{y}_{i,j} \geq 0.5 \\ 0, \hat{y}_{i,j} < 0.5 \end{cases} \tag{25}$$

eg., input reaction "BrCC=C.O=C1CCCC1" output is "[1,0,0,0,0,0,1,1,1,0,0,0,0,0,0]", reaction center is "Br", "C=O".

Author contributions

Xiaoliang Xiong: formal analysis, methodology, software, visualization, writing – original draft; Ruizhen Jia: investigation, validation, writing – review & editing; Yang Tian: data curation; Jingke Chen: software; Boxue Tian: conceptualization, methodology, supervision, writing – review & editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no competing interests.

Data availability

The code supporting this study is openly available on GitHub at https://github.com/xiongxxl/PABA and https://github.com/xiongxxl/SPABA. A stable version of the code, including all software versions, trained models, and workflows, has been deposited in Zenodo with the DOI: https://doi.org/10.5281/zenodo.19447472. The pretrained chemical language model used in this study is accessible at https://github.com/WeilabMSU/PretrainModels. The molecular datasets (ChEMBL, PubChem, ZINC) and reaction datasets (USPTO_50k, USPTO_1K) are publicly available from their original sources as cited in the manuscript. All data generated or analyzed during this study are included in the published article and its supplementary information (SI) files.

Supplementary information: detailed α parameters, additional examples, supplementary figures and tables, and analysis results that support the findings of this work. See DOI: https://doi.org/10.1039/d6dd00055j.

Acknowledgements

This work was supported by Tsinghua University Initiative Scientific Research Program (No. 20231080030) and the Tsinghua-Peking University Center for Life Sciences (No. 20111770319).

Notes and references

1 Z. Li, M. J. Jiang, S. Wang and S. G. Zhang, *Drug Discovery Today*, 2022, **27**, 103373.
2 M. E. Mswahili, J. H. Hwang, J. C. Rajapakse, K. Jo and Y. S. Jeong, *J. Cheminf.*, 2025, **17**, 17.
3 J. Hu, *arXiv*, 2024, preprint arxiv:2406.06553, DOI: 10.48550/arXiv.2406.06553.
4 A. Sultan, M. Rausch-Dupont, S. Khan, O. Kalinina, D. Klakow and A. Volkamer, *arXiv*, 2025, preprint arXiv:2503.03360, DOI: 10.48550/arXiv.2503.03360.
5 X. Zhang, C. Qian, B. Yang, H. Jin, S. Wu, J. Xia, F. Yang and L. Zhang, *J. Pharm. Anal.*, 2025, **15**, 101465.
6 Y. Cao, T. Zhang, X. Zhao and H. Li, *J. Chem. Inf. Model.*, 2025, **65**, 1990–2002.
7 Z. Chen, Z. Fang, W. H. Tian, Z. G. Long, C. Z. Sun, Y. F. Chen, H. Yuan, H. L. Li and M. Lan, *Proc. AAAI Conf. Artif. Intell.*, 2025, **39**, 84–92.
8 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.
9 S. Wang, Y. Guo, Y. Wang, H. Sun and J. Huang, *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage*, 2019, pp. 429–436, DOI: 10.1145/3307339.3342186.
10 J. Li and X. Jiang, *Wirel. Commun. Mob. Comput.*, 2021, **2021**, 7181815.
11 S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020, preprint arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.09885.
12 W. Ahmad, E. Simon, S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2022, preprint arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885\.
13 Y. W. Liu, R. S. Zhang, T. F. Li, J. Jiang, J. Ma and P. Wang, *J. Mol. Graphics Modell.*, 2023, **118**, 108344.
14 M. Das, A. Ghosh and R. B. Sunoj, *J. Comput. Chem.*, 2024, **45**, 1160–1176.

15 A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.
16 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
17 I. V. Tetko, P. Karpov, R. V. Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 5575.
18 W. A. Warr, *Mol. Inf.*, 2014, **33**, 469–476.
19 X. Wang, C. Y. Hsieh, X. Yao, J. Wang and Y. Q. Lin, *Research*, 2023, **6**, 0231.
20 S. Yun and W. B. Lee, *arXiv*, 2024, preprint arXiv:2411.19503, doi: DOI: 10.48550/arXiv.2411.19503.
21 X. R. Wang, Y. Q. Li, J. Z. Qiu, G. Y. Chen, H. X. Liu, B. B. Liao, C. Y. Hsieh and X. J. Yao, *Chem. Eng. J.*, 2021, **420**, 129845.
22 P. M. Htut, J. Phang, S. Bordia and S. R. Bowman, *arXiv*, 2019, preprint arXiv:1911.12246, DOI: 10.48550/arXiv.1911.12246.
23 E. Voita, P. Serdyukov, R. Sennrich and I. Titov, *arXiv*, 2018, preprint arXiv:1805.10163, DOI: 10.48550/arXiv.1805.10163.
24 G. Tang and J. Nivre, *arXiv*, 2018, preprint arXiv:1810.07595, DOI: 10.48550/arXiv.1810.07595.
25 D. M. Lowe, *PHD thesis*, University of Cambridge, 2012.
26 S. Chen and Y. S. Jung, *Nat. Mach. Intell.*, 2022, **4**, 772–780.
27 D. Chen, J. Zheng, G. W. Wei and F. Pan, *J. Phys. Chem. Lett.*, 2021, **12**, 10793–10801.
28 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Uhlén, *et al.*, *Nucleic Acids Res.*, 2017, **45**, D945–D954.
29 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, *et al.*, *Nucleic Acids Res.*, 2016, **44**(1), D1202–D1213.
30 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
31 C. W. Coley, W. G. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.
32 C. D. Wong, L. C. Bradford, N. Hirbawi and E. R. Jarvo, *Angew. Chem., Int. Ed.*, 2025, **137**, e202509657.
33 H. Kimura, T. Yatabe and K. Yamaguchi, *J. Am. Chem. Soc.*, 2025, **147**, 27238–27250.
34 J. P. Tuccinardi and J. L. Wood, *J. Am. Chem. Soc.*, 2025, **147**, 5736–5742.
35 X. Y. Gou, Y. X. Zhi, C. T. Wang, Y. Lei, L. Zhao, B. S. Zhang and Y. M. Liang, *Org. Lett.*, 2025, **27**, 11415–11421.
36 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
37 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J. L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 144–152.
38 Y. Wan, C.-Y. Hsieh, B. Liao and S. Zhang, *Proc. Mach. Learn. Res.*, 2022, **162**, 22475–22490.
39 R. Irwin, S. Dimitriadis, J. Z. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.
40 Z. Zheng, F. Florit, B. Jin, H. Wu, S. C. Li, K. Y. Nandiwale, C. A. Salazar, J. G. Mustakis, W. H. Green and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2025, **137**, e202418074.
41 J. Choi, S. Kim and Y. Jung, *J. Am. Chem. Soc.*, 2025, **147**, 39113–39122.
42 H. Chen, J. Zhong, X. Zhang, J. Chen, L. Guo, X. Xiong, X. Zhang, X. Liu, B. Xiao and B. Tian, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.01.20.633699.
43 J. Yousefi, University of Guelph, Ontario, Canada, 2011, DOI:DOI: 10.13140/RG.2.1.4758.9284.
44 J. Sauvola and M. Pietikäinen, *Pattern Recogn.*, 2000, **33**, 225–236.
45 B. Gatos, I. Pratikakis and S. J. Perantonis, *Pattern Recogn.*, 2006, **39**, 317–327.
46 Y. Liu and Y. Wang, *arXiv*, 2025, preprin, arXiv:2503.08017, DOI: 10.48550/arXiv.2503.08017.
47 L. O'Gorman, *CVGIP Graph. Models Image Process.*, 1994, **56**, 494–506.