Supporting Information

Reaction Center Prediction by Analyzing Attention of a Chemical Language Model

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

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.

* These authors contribute equally to this work.

To whom correspondence should be addressed:

Boxue Tian: boxuetian@mail.tsinghua.edu.cn

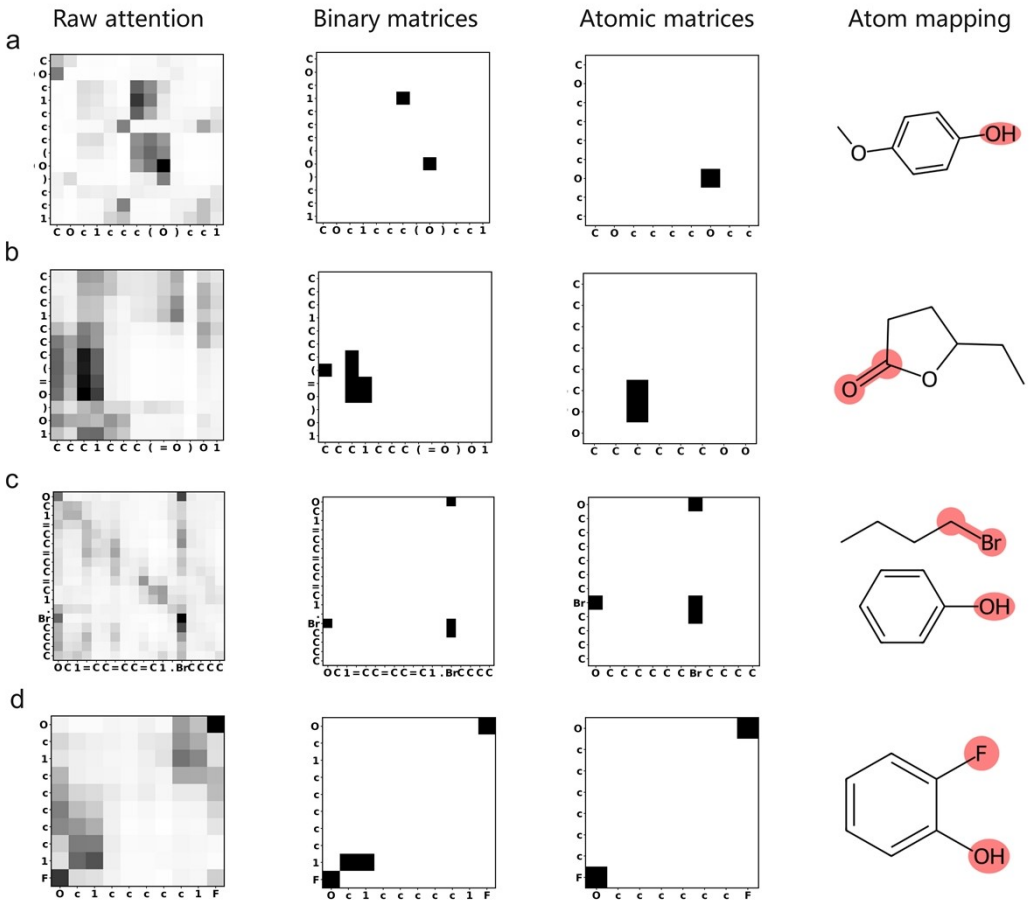

Fig. S1 Non-atomic symbol removal does not affect source data. (a) “COc1ccc(O)cc1”: After removing non-atomic symbols such as the ring index “1”, the model’s attention remains focused on the oxygen atom (O) and correctly corresponds to the phenolic hydroxyl (–OH) group, without introducing errors.(b) “CCC1CCC(=O)O1”: The attention is concentrated on the “C(=O)” fragment. After removing symbols such as “(” and “=”, it can still be accurately mapped to the carbonyl group (C=O), with no loss of key chemical information.(c) “OC1=CC=CC=C1.BrCCCC”: After removing non-atomic symbols such as “1” and “=”, the attention still correctly localizes on the R–OH and R–Br functional groups, indicating that the model focuses on atoms and their local structural environments rather than the symbols themselves.(d) “Oc1ccccc1F”: After removing the ring index “1”, the attention still accurately identifies the R–F and R–OH functional groups.

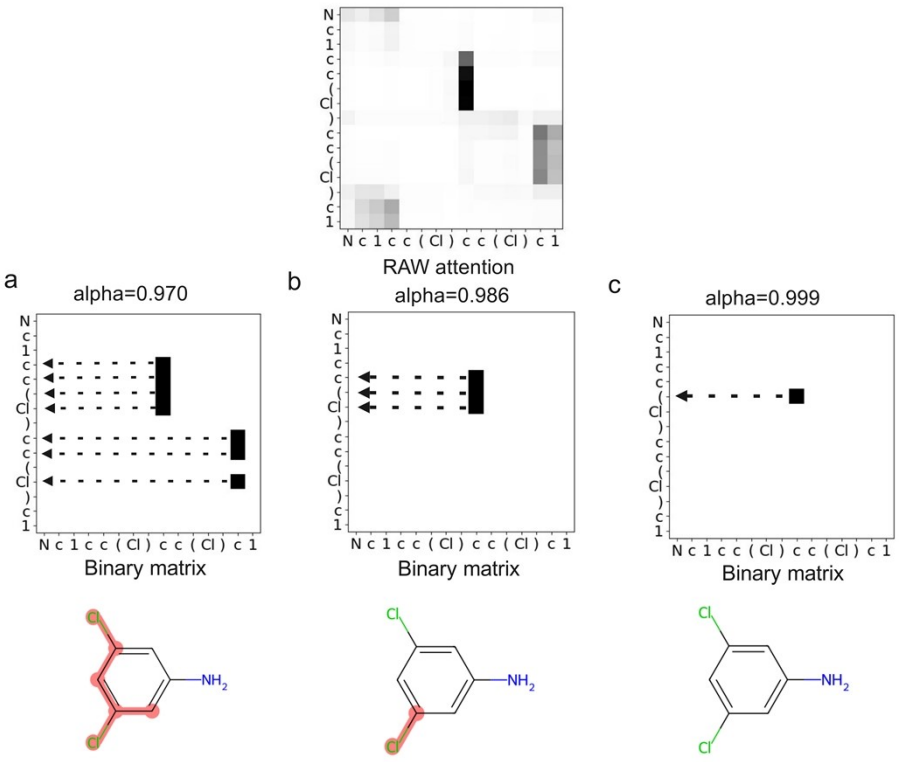

Fig. S2 Impact of the parameter α on model performance. In the raw attention, the model mainly focuses on the central black block. With α= 0.999, the binarized image loses most of this information, while α= 0.970 introduces secondary attention from the lower-right corner. Experimental results show that α= 0.986 best captures the central black block, extracting key features effectively.

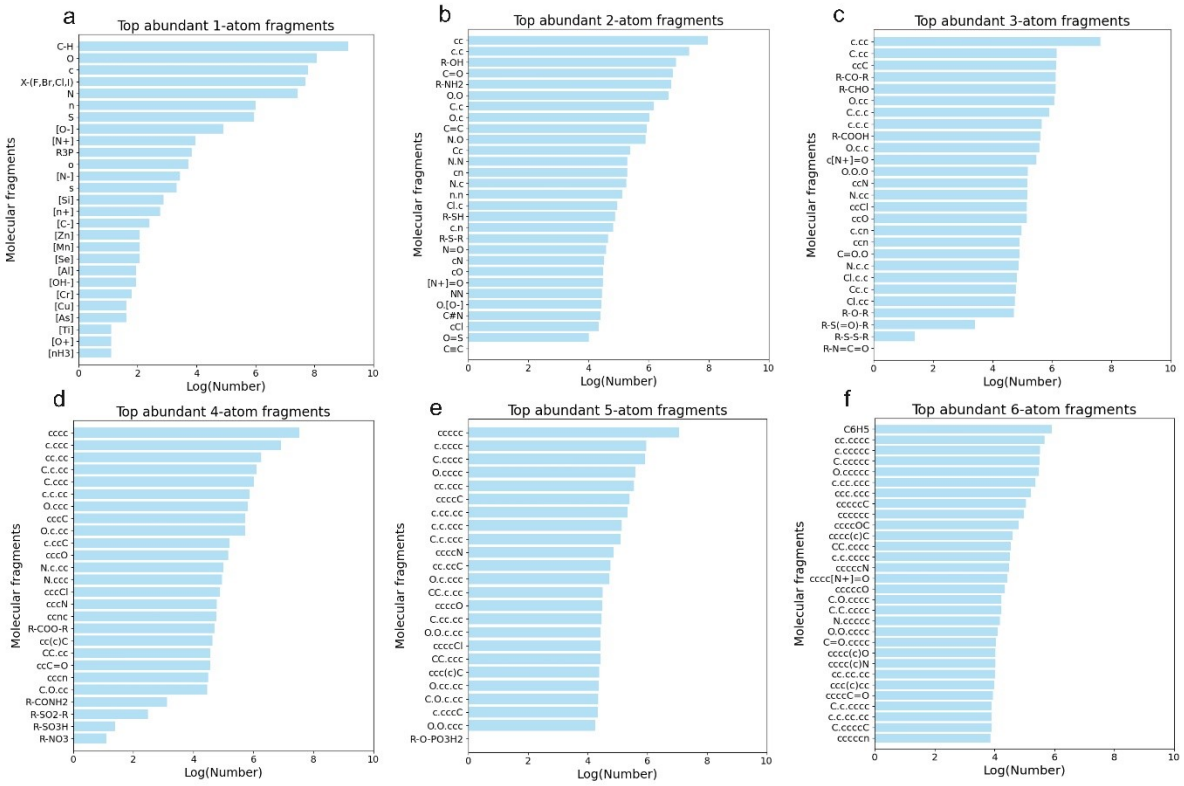

Fig. S3 Molecular fragments frequently attend to functional groups. (a) 1-atom fragments mainly contain halogen elements, such as X- (e.g., F, Br, Cl, I). (b) 2-atom molecular fragments primarily include common functional groups, e.g., R-OH, R-NH2, C=C. (c) 3-atom fragments mainly include R-CO-R, R-CHO, R-COOH, etc. (d) 4-atom molecular fragments mainly comprise R-COO-R, R-CONH2, etc. (e) Only one functional group was identified among 5-atom fragments, R-O-PO3H2. (f) Among 6-atom fragments, benzene ring groups had the highest abundance.

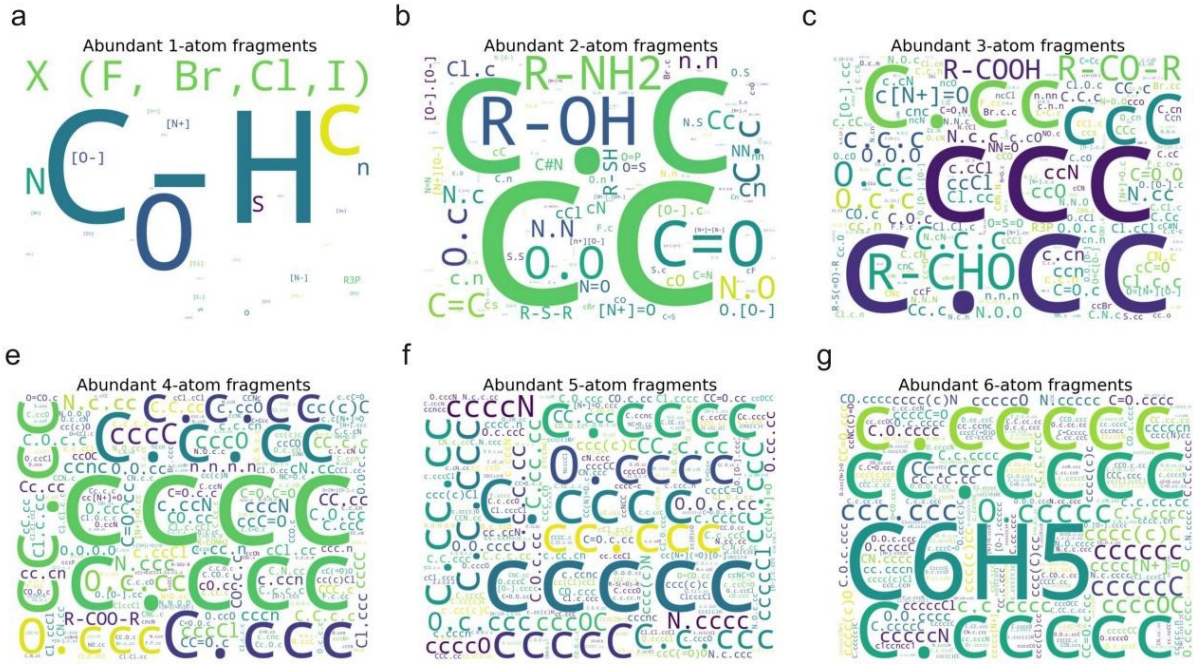

Fig. S4 Word cloud visualization of six fragment categories. (a) One-atom fragments rank with halogen groups (X-: F, Br, Cl, I) highest, while [Sn][Fe-2] ranks lower. (b) Two-atom fragments show R-OH and R-NH2 ranking high, while rare elements like '[nH3].n', '[Se].c' rank low. (c) Three-atom fragments place R-CO-R and R-CHO at the top, while 'S.c.s', 'C.cc', and 'CC[n+]' rank lower. (d) The R-COO-R group ranks high, while 'cCC#N', 'CC.cs', and 'Br.c.c.c' rank lower. (e) In five-atom samples, fragments are larger, and functional groups are smaller. (f) In six-atom samples, C6H5 ranks first, followed by 'cc.CCcc' and 'c.ccccc', while 'c.c(=O)S.n' and 'CC=O.c.co' rank lower.

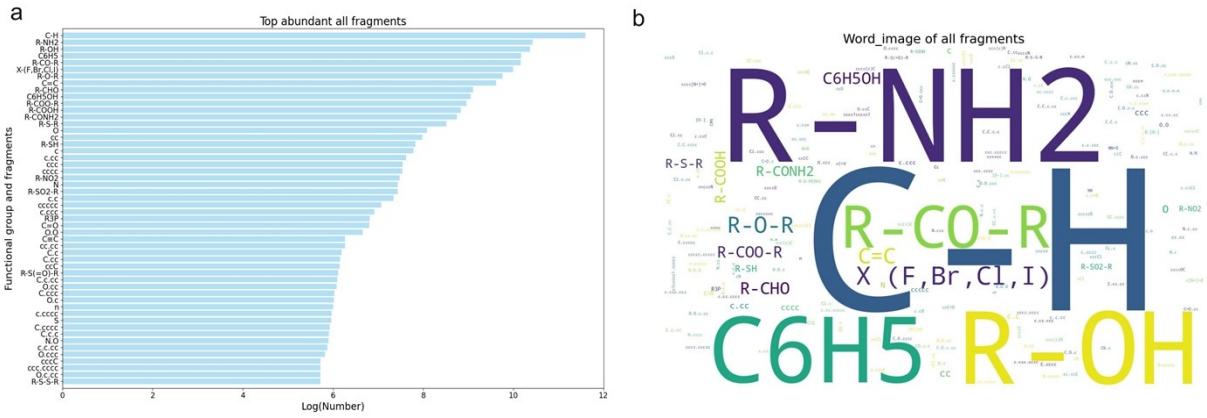

Fig. S5 Functional group ranking across atomic fragments. (a) C-H, R-NH2, and R-OH are the top three, respectively. Because most of the molecules contain C-H bonds, C-H can be ranked first. (b) The ranking of all fragments was visualized using a word cloud diagram.

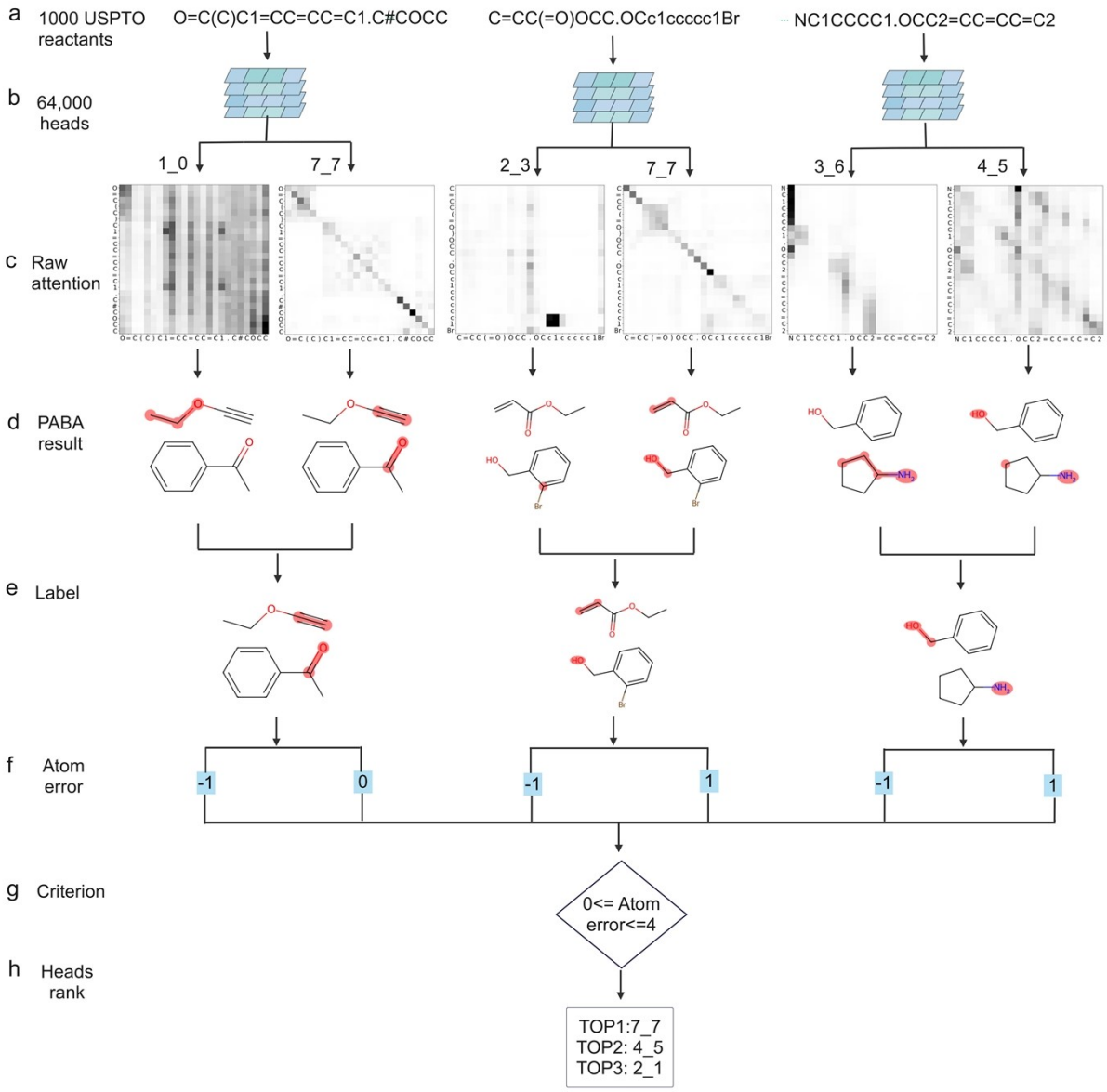

Fig. S6 Approach for discovering key attention heads.(a) Reactant molecules sampled from the USPTO dataset are input into a pretrained chemical language model (CLM).(b) The model produces attention matrices from all 64 heads (8 layers × 8 heads).(c) Raw self-attention maps from selected heads are visualized.(d) The PABA method binarizes attention using a peak-ratio threshold to extract high-attention atoms, highlighting candidate reaction centers.(e) Ground-truth reaction center labels are obtained from atom-mapped reactions.(f) Atom error is computed by comparing predicted atoms with labeled reaction centers.(g) A criterion is applied, where predictions with 0≤Atom error≤4 are considered acceptable.(h) Based on this criterion, attention heads are ranked according to their performance, with top-performing heads (e.g., 7_7, 4_5, 2_1) identified for downstream analysis.

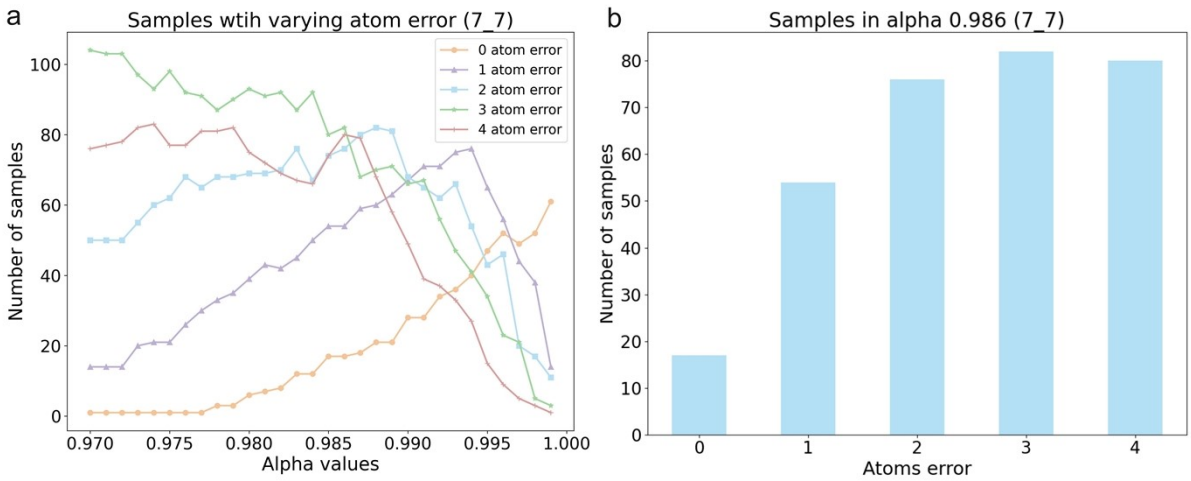

Fig. S7 Identification of the optimal α. (a) Effect of different binarization thresholds α on the atom error distribution (7_7 attention head), (b) At this threshold, the numbers of samples with atom errors of 0-4 were 17, 54, 76,82 and 80, respectively, giving a total of 309 samples.

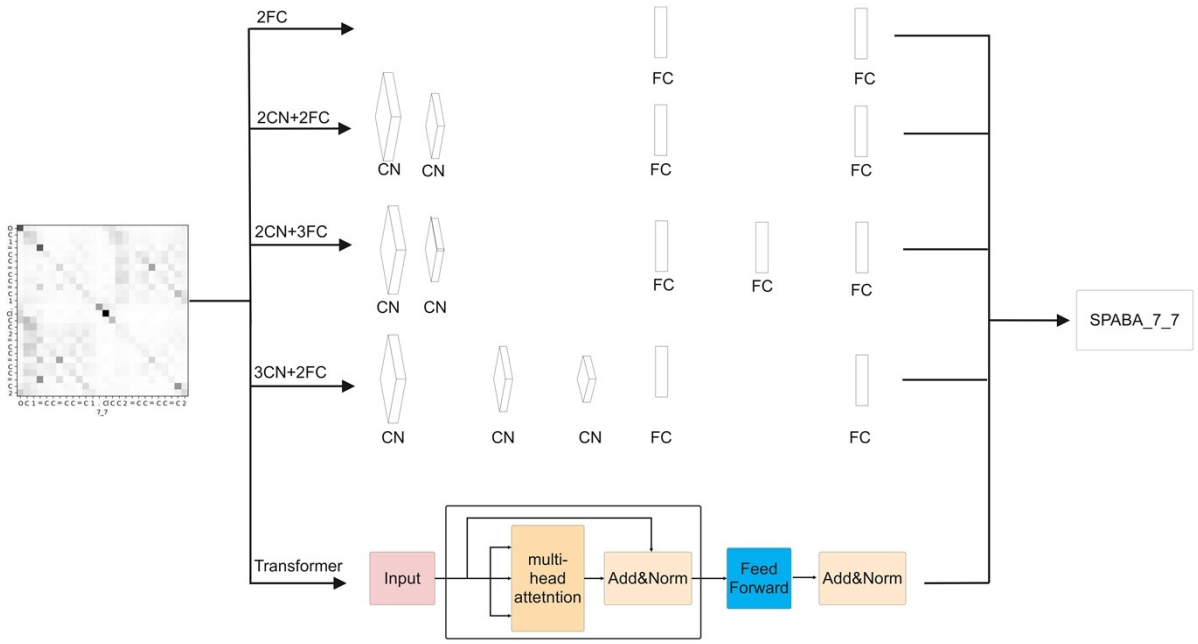

Fig. S8 Schematic of the network architecture. Five network architectures composed of FC, CN, and Transformer were constructed, including 2FC, 2CNN+2FC, 2CNN+3FC, 3CNN+2FC, to compare model performance under different structural complexities and identify the optimal architecture.

Table S1. Dictionary of functional groups

Functional group	SMARTS	Formula	Atoms
Alkene	[CX3]=[CX3]	C=C	2
Alkyne	[CX2]#[CX2]	C≡C	2
Aromatic	c1ccccc1	C6H5	6
Halide	[F,Cl,Br,I]	X (F, Cl, Br, I)	1
Alcohol	[OX2H]	R-OH	2
Phenol	c1ccc(O)cc1	C6H5OH	7
Ether	[OD2]([#6])[#6]	R-O-R	3
Aldehyde	[CX3H1](=O)[#6]	R-CHO	3
Ketone	[CX3](=O)[#6]	R-CO-R	3
Carboxylic Acid	[CX3](=O)[OX2H1]	R-COOH	3
Ester	[CX3](=O)[OX2][#6]	R-COO-R	4
Amide	[NX3][CX3](=[OX1])[#]	R-CONH2	4
Amine	[NX3][#6]	R-NH2	2
Nitrate	[NX3](=O)([OX1-])[OX1-]	R-NO3	4
Nitro	[NX3](=O)[OX1-]	R-NO2	3
Sulfonic Acid	S(=O)(=O)[O-]	R-SO3H	4
Thiol	[SX2H]	R-SH	2
Thioether	[SX2][#6]	R-S-R	2
Disulfide	[SX2][SX2]	R-S-S-R	3
Sulfoxide	[SX3](=O)[#6]	R-S(=O)-R	3
Sulfone	[SX4](=O)(=O)[#6]	R-SO2-R	4
Phosphine	[PX3]	R3P	1
Phosphate	P(=O)(O)(O)O	R-O-PO3H2	5
Isocyanate	N=C=O	R-N=C=O	3
Isothiocyanate	N=C=S	R-N=C=S	3
Cyano	[NX1]#[CX2]	R-C≡N	2