

DposFinder: an interpretable transformer model for predicting phage-derived polysaccharide depolymerases and their host capsular serotypes

Supplementary data

Table S1. Functional analyses of depolymerases included in the web platform of DposFinder.....	2
Table S2. Bacterial strains, phages and plasmids used in this study.....	3
Table S3. Oligonucleotides used in this study.....	5
Table S4. Performance comparison between DposFinder and model architectures of SpikeHunter and DePP.....	6
Table S5. Ablation study of DposFinder.....	7
Table S6. Prediction and validation of the specific capsular serotypes of the depolymerases derived from <i>K. pneumoniae</i> phages.....	8
Table S7. Candidate phages screened from 38 sequenced phages by using DposFinder and PhageHostLearn to infect 18 <i>K. pneumoniae</i> strains with different capsular serotypes.....	9
Fig. S1. Outline of the model performance of DposFinder, experimental validation, potential application and phage genome mining.	10
Fig. S2. Outline of collecting data, training models, and conducting evaluations of DposFinder.....	11
Fig. S3. Data characteristics of the curated depolymerase dataset.....	12
Fig. S4. Analysis of DposFinder's attention towards depolymerases (Dpos).....	13
Fig. S5. Benchmark evaluation reveals that DposFinder outperforms other models.....	14
Fig. S6. Subsequence similarity heatmap of depolymerase from <i>A. baumannii</i> phages.....	15
Fig. S7. Cartoon representations and attention visualization of validated depolymerases.....	16

Table S1. Functional analyses of depolymerases included in the web platform of DposFinder.

Analysis	Tool	Link
Protein comparison	NCBI BLASTp	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Protein properties	Biopython	https://biopython.org
Genomic context	gggenes	https://github.com/wilkox/gggenes ,
Sequence attention	Logomaker	https://github.com/jbkinney/logomaker
Protein secondary structure	S4pred	https://github.com/psipred/s4pred
Protein domain	InterproScan	https://www.ebi.ac.uk/interpro
Protein disorder region	IUPred3	https://iupred3.elte.hu/

Table S2. Bacterial strains, phages and plasmids used in this study.

<i>K. pneumoniae</i> strain	Description	Source or reference
2410	KL64 capsular serotype	Li, <i>et al.</i> , Antibiotics (Basel), 2021 (PMID: 33535705)
2226	KL47 capsular serotype	Wu, <i>et al.</i> , Front Microbiol, 2019 (PMID: 31849905)
293	KL2 capsular serotype	Wang, <i>et al.</i> , Virulence, 2018 (PMID: 29338592)
6418	KL1 capsular serotype	Zhang, <i>et al.</i> , Epidemiol Infect, 2020 (PMID: 32782064)
7246	KL25 capsular serotype	Zhang, <i>et al.</i> , Epidemiol Infect, 2020 (PMID: 32782064)
6415	KL23 capsular serotype	Zhang, <i>et al.</i> , Epidemiol Infect, 2020 (PMID: 32782064)
8734	KL36 capsular serotype	Zhang, <i>et al.</i> , Epidemiol Infect, 2020 (PMID: 32782064)
20-768	KL10 capsular serotype	Hu, <i>et al.</i> , Nat Microbiol, 2024 (PMID: 38424289)
19-618	KL51 capsular serotype	Hu, <i>et al.</i> , Nat Microbiol, 2024 (PMID: 38424289)
<i>K. pneumoniae</i> phage	Description	Source or reference
P2226	KL47 capsular target phage	Wu, <i>et al.</i> , Front Microbiol, 2019 (PMID: 31849905)
P0957	KL64 capsular target phage	This study, accession number: PQ133594
P293	KL2 capsular target phage	This study, accession number: PQ133605
P6418	KL1 capsular target phage	This study, accession number: PQ133615
P7246	KL25 capsular target phage	This study, accession number: PQ133620
PA45-9	KL23 capsular target phage	This study, accession number: PQ133626
P8734	KL36 capsular target phage	This study, accession number: PQ133625
P20-768	KL10 capsular target phage	This study, accession number: PQ133602
P19-618	KL51 capsular target phage	This study, accession number: PQ133600
P13a	KL110 capsular target phage	This study, accession number: PQ133595
P16	KL136 capsular target phage	This study, accession number: PQ133596
P17	KL110 capsular target phage	This study, accession number: PQ133597
P18-100	KL15 capsular target phage	This study, accession number: PQ133598
P19-411	KL102 capsular target phage	This study, accession number: PQ133599
P19-966	KL18 capsular target phage	This study, accession number: PQ133601
P2234	KL64 capsular target phage	This study, accession number: PQ133603
P2328	KL47 capsular target phage	This study, accession number: PQ133604
P29	KL54 capsular target phage	This study, accession number: PQ133606
P3	KL139 capsular target phage	This study, accession number: PQ133607
P5010	KL21 capsular target phage	This study, accession number: PQ133608
P5026	KL63 capsular target phage	This study, accession number: PQ133609
P5054	KL57 capsular target phage	This study, accession number: PQ133610
P6169	KL39 capsular target phage	This study, accession number: PQ133611
P6257	KL24 capsular target phage	This study, accession number: PQ133612
P6315	KL5 capsular target phage	This study, accession number: PQ133613
P6371	KL20 capsular target phage	This study, accession number: PQ133614
P6760	KL60 capsular target phage	This study, accession number: PQ133616
P6a	KL136 capsular target phage	This study, accession number: PQ133617
P700603	KL53 capsular target phage	This study, accession number: PQ133618
P7124	KL19 capsular target phage	This study, accession number: PQ133619
P7285	KL27 capsular target phage	This study, accession number: PQ133621
P7501	KL57 capsular target phage	This study, accession number: PQ133622
P7641	KL5 capsular target phage	This study, accession number: PQ133623

P7	KL10 capsular target phage	This study, accession number: PQ133624
PQ5104	KL54 capsular target phage	This study, accession number: PQ133627
PQ7254	KL17 capsular target phage	This study, accession number: PQ133628
Plasmid	Description	Source or reference
pSUMO3	plasmid of <i>Escherichia coli</i> , Amp ^R	LifeSensors, United States
pSK	plasmid of <i>Escherichia coli</i> , Apr ^R	Mayer MP, Gene, 1995 (PMID: 7557476)

Table S3. Oligonucleotides used in this study.

Name	Sequence (5'–3')	Description
K2-P293-gp43-F	CAGCAGCAGACGGGAGGATCCATGATAAAGAACGATTTTAACCAGC	Amplification of P293-gp43
K2-P293-gp43-R	CTCGAGTGCGGCCGCAAGCTTTTACAAGAAGTTTACGATAACG	Amplification of P293-gp43
K1-P6418-gp1-F	CAGCAGCAGACGGGAGGATCCATGGCATTAAATTAGATTAGT	Amplification of P6418-gp1
K1-P6418-gp1-R	CTCGAGTGCGGCCGCAAGCTTTTAGAGATACACCTCCCAGG	Amplification of P6418-gp1
K25-P7246-gp2-F	CAGCAGCAGACGGGAGGATCCATGGACCAAGACATTAACACAG	Amplification of P7246-gp2
K25-P7246-gp2-R	CTCGAGTGCGGCCGCAAGCTTTTAGTTAGGCAGAAGCGTTG	Amplification of P7246-gp2
K25-P7246-gp3-F	CAGCAGCAGACGGGAGGATCCATGTAAAGTCAGATTTTAACCA	Amplification of P7246-gp3
K25-P7246-gp3-R	CTCGAGTGCGGCCGCAAGCTTTTATGAGTGAACCTCGACCA	Amplification of P7246-gp3
K23-PA45-9-gp23-F	CAGCAGCAGACGGGAGGATCCATGGCTTTAACTAAAGCTACTAT	Amplification of PA45-9-gp23
K23-PA45-9-gp23-R	CTCGAGTGCGGCCGCAAGCTTTTAAACGCCAGTCAGGCGA	Amplification of PA45-9-gp23
K36-P8734-gp30-F	CAGCAGCAGACGGGAGGATCCATGCTACAGAGTGTTAAAATTTTC	Amplification of P8734-gp30
K36-P8734-gp30-R	CTCGAGTGCGGCCGCAAGCTTTTAATCAGCCACAAACAAGTC	Amplification of P8734-gp30
K10-P20-768-gp42-F	CAGCAGCAGACGGGAGGATCCATGTTACCAAATTTAACCAACC	Amplification of P20-768-gp42
K10-P20-768-gp42-R	CTCGAGTGCGGCCGCAAGCTTTTAGTATCGCTCAATCATCCA	Amplification of P20-768-gp42
K51-P19-618-gp16-F	CAGCAGCAGACGGGAGGATCCATGAGTCATTATAATGGAGGTG	Amplification of P19-618-gp16
K51-P19-618-gp16-R	CTCGAGTGCGGCCGCAAGCTTTTATGCTGAAGCTATAGACC	Amplification of P19-618-gp16

Table S4. Performance comparison between DposFinder and model architectures of SpikeHunter and DePP. performance is evaluated through 5-fold cross-validation on the training dataset, The results show average values with 95% confidence intervals. FC, fully connected network; RF, random forest.

Model	Accuracy	Precision	Recall	F1
ESM2+conv1d+transformer (DposFinder)	0.979±0.013	0.973±0.019	0.935±0.068	0.952±0.031
ESM2+FC (SpikeHunter)	0.971±0.013	0.961±0.025	0.913±0.066	0.935±0.031
Amino acid derived feature+RF (DePP)	0.884±0.029	0.960±0.023	0.787±0.053	0.863±0.024

Table S5. Ablation study of DposFinder. performance is evaluated through 5-fold cross-validation on the training dataset, The results show average values with 95% confidence intervals. conv1d, 1-dimensional convolution layer.

Model	Accuracy	Precision	Recall	F1
ESM2+conv1d+transformer (DposFinder)	0.979±0.013	0.973±0.019	0.935±0.068	0.952±0.031
ESM2+conv1d	0.955±0.017	0.915±0.053	0.891±0.069	0.900±0.040
ESM2+transformer	0.976±0.014	0.960±0.013	0.935±0.059	0.947±0.033
ESM2+finetune	0.970±0.014	0.934±0.013	0.935±0.059	0.934±0.033
TAPEBert+conv1d+transformer	0.962±0.006	0.940±0.037	0.897±0.056	0.916±0.015

Table S6. Prediction and validation of the specific capsular serotypes of the depolymerases derived from *K. pneumoniae* phages.

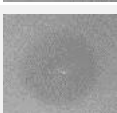
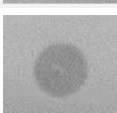
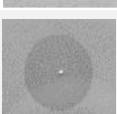
Phage	Predicted depolymerase	Predicted capsular serotype of depolymerase	Validated capsular serotype	Spot test
P0957	Dpo15	KL64	KL64	
P2226	Dpo43	KL47	KL47	
P293	Dpo44	KL2/KL3/KL13	KL2	
P6418	Dpo1	KL1	KL1	
P19-618	Dpo16	KL61/KL134	KL51	
P20-768	Dpo42	KL10	KL10	
P8734	Dpo30	KL36/KL43	KL36	
PA45-9	Dpo23	KL23/KL54/KL58	KL23	
P7246	Dpo2	KL25	KL25	
P7246	Dpo3	KL10	KL10	

Table S7. Candidate phages screened from 38 sequenced phages by using DposFinder and PhageHostLearn to infect 18 *K. pneumoniae* strains with different capsular serotypes.

#	Reference strain	Phage validated by single-spot assay	Top 5 candidate phages predicted by PhageHostLearn					Candidate Phages predicted by DposFinder				
			1	2	3	4	5	1	2	3	4	5
1	KL110	P6418	P6418 (0.892)	P2328 (0.022)	P2226 (0.02)	P5054 (0.017)	P7246 (0.004)	P6418 (0.988)	P2328 (0.498)			
2	KL10	P20-768, P7	P6371 (0.03)	P6760 (0.008)	P19-966 (0.006)	P293 (0.005)	P7246 (0.003)	P7124 (0.970)	P19-966 (0.932)	P20-768 (0.930)	P6371 (0.919)	P6760 (0.768)
3	KL15	P18-100	P293 (0.001)	P6a (0.001)	P19-411 (0.0)	PQ5104 (0.0)	P7501 (0.0)	N/A				
4	KL17	PQ7254	P2492 (0.003)	P5010 (0.001)	P6760 (0.001)	P16 (0.001)	P19-618 (0.0)	P2492 (0.460)	PQ7254 (0.261)			
5	KL18	P19-966	P18-100 (0.039)	P8734 (0.001)	P19-618 (0.001)	P6760 (0.001)	P6a (0.0)	N/A				
6	KL19	P6570, P7124	P7 (0.027)	P5010 (0.001)	P13 (0.001)	P293 (0.001)	P6760 (0.001)	P6570 (1.0)	P7124 (1.0)			
7	KL2	P293	P2328 (0.001)	P7501 (0.001)	P293 (0.001)	PQ7254 (0.001)	PA45-9 (0.001)	P2328 (0.498)	P293 (0.183)	P6169 (0.180)		
8	KL20	P6371	PQ5104 (0.002)	P29 (0.001)	P5054 (0.001)	PA45-9 (0.001)	P7501 (0.0)	N/A				
9	KL21	P5010	P6257 (0.0)	P7246 (0.0)	P6371 (0.0)	P293 (0.0)	P18-100 (0.0)	N/A				
10	KL23	PA45-9	P5010 (0.021)	P2492 (0.003)	P16 (0.002)	P2226 (0.001)	P2328 (0.001)	N/A				
11	KL24	P6257	P6760 (0.0)	P19-966 (0.0)	P18-100 (0.0)	P6257 (0.0)	P7501 (0.0)	P6257 (0.274)	P7641 (0.033)			
12	KL25	P7246	P5010 (0.002)	P7501 (0.001)	P6371 (0.001)	PQ5104 (0.001)	P293 (0.001)	P7124 (0.764)				
13	KL27	P7285	P5010 (0.001)	P6257 (0.0)	P19-411 (0.0)	P7501 (0.0)	P6371 (0.0)	P7285 (0.957)				
14	KL36	P8734	P5010 (0.977)	P19-966 (0.077)	P2226 (0.052)	P6760 (0.05)	P6371 (0.034)	P6760 (0.468)	P19-966 (0.446)	P8734 (0.046)		
15	KL39	P6169	P7 (0.008)	P5026 (0.003)	P18-100 (0.002)	P293 (0.002)	P6a (0.001)	N/A				
16	KL47	P2226, P2328	P5010 (0.707)	P2226 (0.0)	P6760 (0.0)	P6257 (0.0)	P19-618 (0.0)	P2328 (1.0)	P2226 (0.990)			
17	KL5	P6315, P7641	P19-966 (0.0)	P2226 (0.0)	P6760 (0.0)	P293 (0.0)	P2328 (0.0)	P6315 (0.401)				
18	KL51	P19-618	P6760 (0.003)	P19-966 (0.003)	P293 (0.001)	P7246 (0.001)	PQ5104 (0.0)	N/A				
19	KL53	P700603	P5010 (0.001)	P2492 (0.0)	P6315 (0.0)	P6760 (0.0)	P19-966 (0.0)	P700603 (0.990)				
20	KL54	P29, PQ5104	P6257 (0.014)	P8734 (0.002)	P2492 (0.002)	P29 (0.002)	P19-966 (0.002)	P29 (0.843)	PQ5104 (0.842)	P6257 (0.274)	P7641 (0.033)	
21	KL57	P7501, P5054	P7501 (0.624)	P2328 (0.084)	P2226 (0.02)	P6760 (0.014)	P19-966 (0.008)	P7501 (0.493)	P5054 (0.491)			
22	KL62	P2492	P19-618 (0.026)	P7 (0.015)	P2492 (0.006)	P6257 (0.001)	P2226 (0.001)	P2492 (0.460)				
23	KL63	P5026	P2226 (0.024)	P7246 (0.013)	P6257 (0.012)	P19-966 (0.011)	P2328 (0.009)	N/A				
24	KL64	P0957, P2234, P2410	P5010 (0.153)	P5054 (0.0)	PQ7254 (0.0)	P6257 (0.0)	P19-618 (0.0)	P2410 (1.0)	P0957 (0.995)	P2234 (0.989)		
25	KL102	P19-411	PQ5104 (0.003)	P5010 (0.001)	P7285 (0.001)	P6a (0.001)	P19-411 (0.001)	P19-411 (0.329)				
26	KL110	P13a	P7 (0.0)	P5026 (0.0)	P5010 (0.0)	PQ5104 (0.0)	P29 (0.0)	N/A				
27	KL136	P6a, P16	P5010 (0.002)	P6a (0.0)	P6760 (0.0)	P2226 (0.0)	P16 (0.0)	N/A				
28	KL139	P3	P18-100 (0.002)	P5026 (0.001)	P5010 (0.001)	P16 (0.0)	P2492 (0.0)	P18-100 (0.199)	P5026 (0.167)			

True positive prediction (the predicted phage matches the phage validated by single-spot assay)

N/A

No candidate phage was predicted

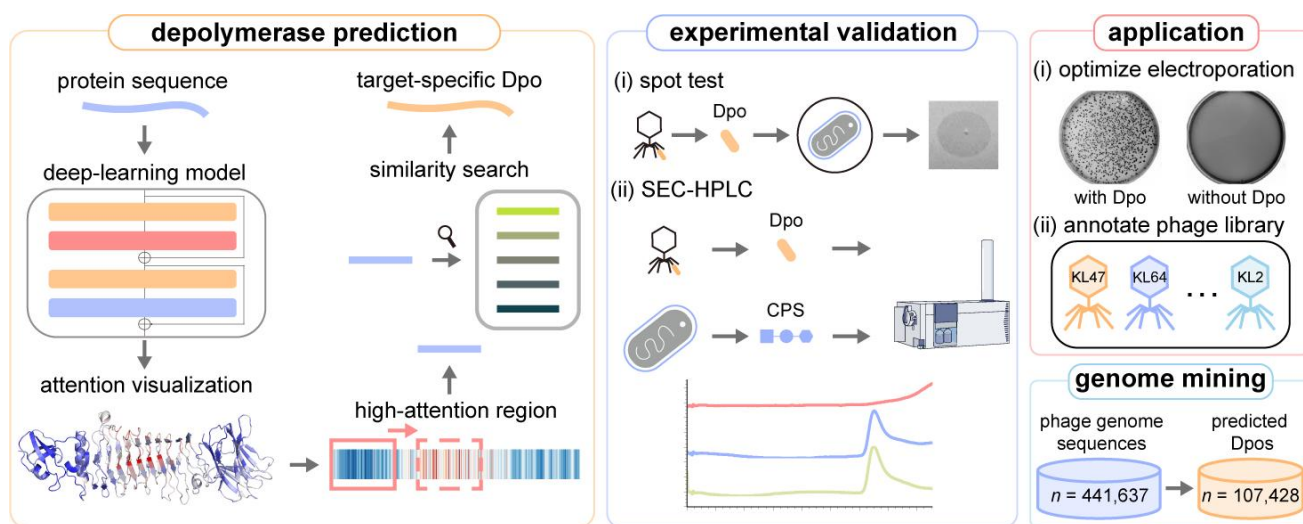


Fig. S1. Outline of model performance of DposFinder, experimental validation, potential application and phage genome mining. DposFinder is an interpretable transformer model for predicting phage-derived polysaccharide depolymerases (Dpo) and their corresponding capsular serotypes of host bacteria.

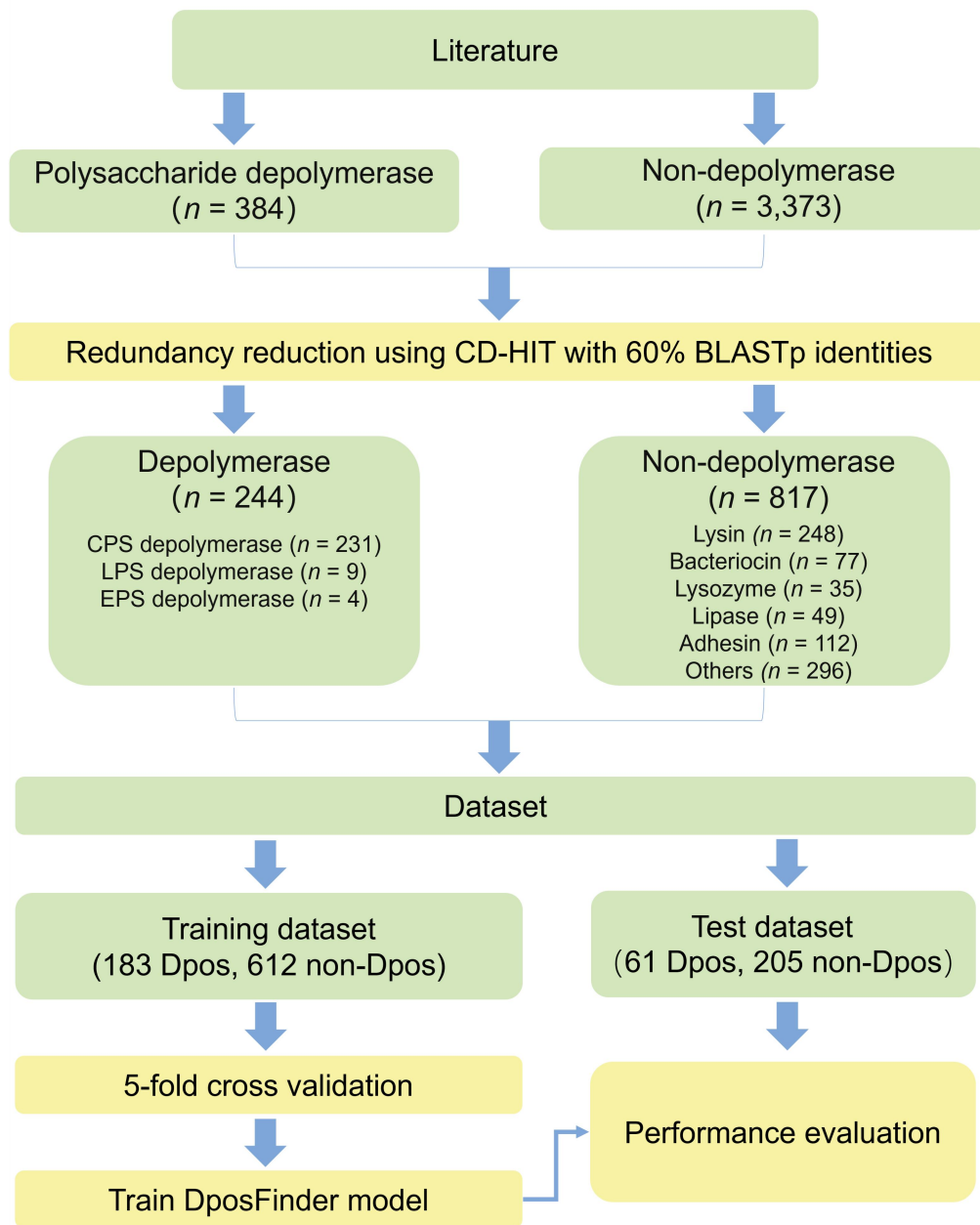


Fig. S2. Outline of collecting data, training models, and conducting evaluations of DposFinder.

Phage-encoded polysaccharide depolymerases (Dpos) and non-depolymerases are curated manually from publicly available databases and relevant literature. Depolymerases include capsular polysaccharide (CPS) depolymerases, lipopolysaccharide (LPS) depolymerases, and exopolysaccharide (EPS) depolymerases. Non-depolymerases comprise lysins, bacteriocins, lysozymes, lipases, adhesins, and other phage-encoded proteins.

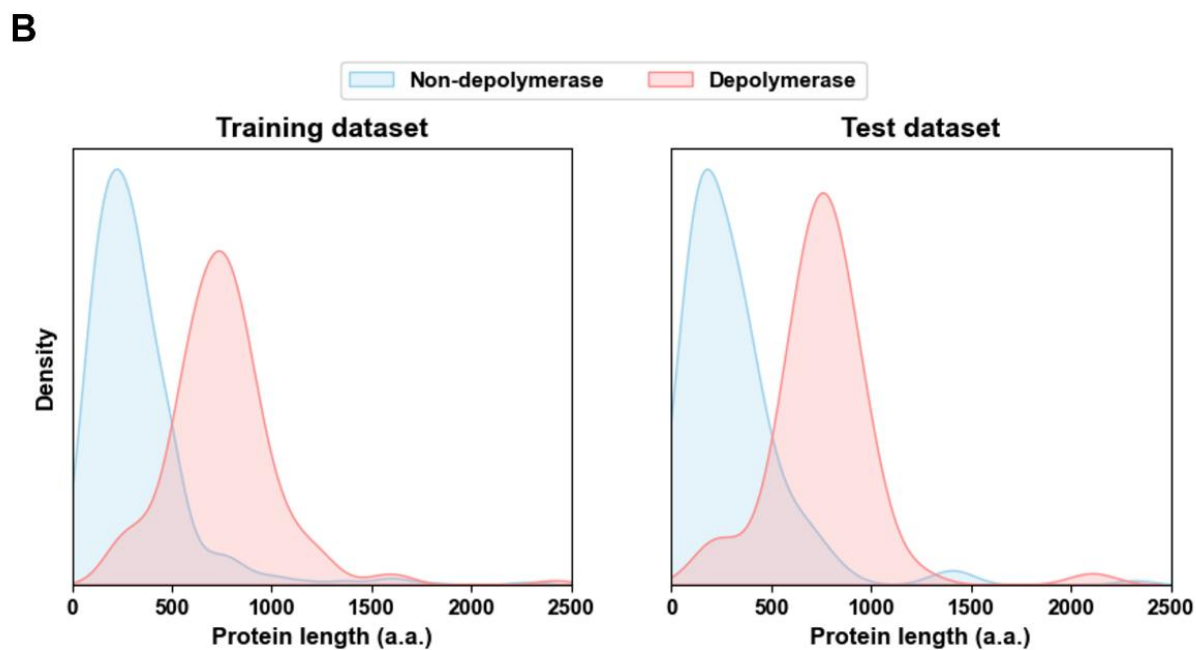
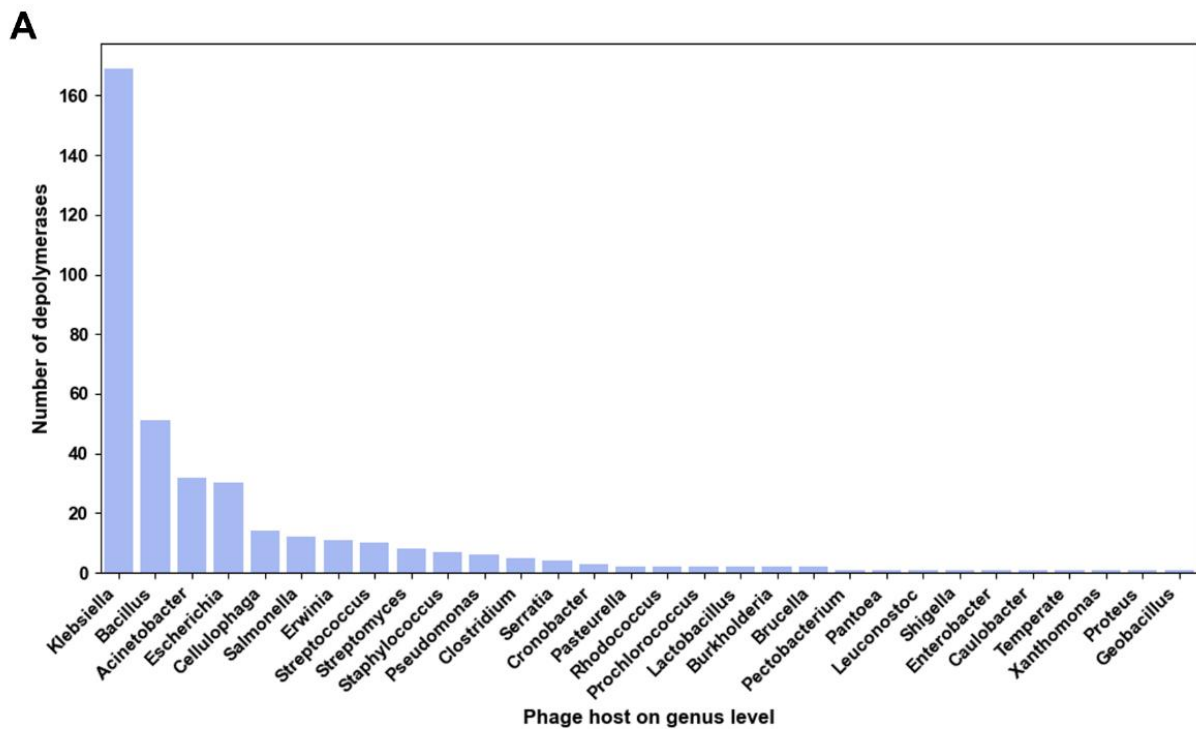


Fig. S3. Data characteristics of the curated depolymerase dataset. (A) The distribution of phage host genera for the phage-derived depolymerases. (B) The length distribution of depolymerases and non-depolymerases in the training dataset (left) and test dataset (right).

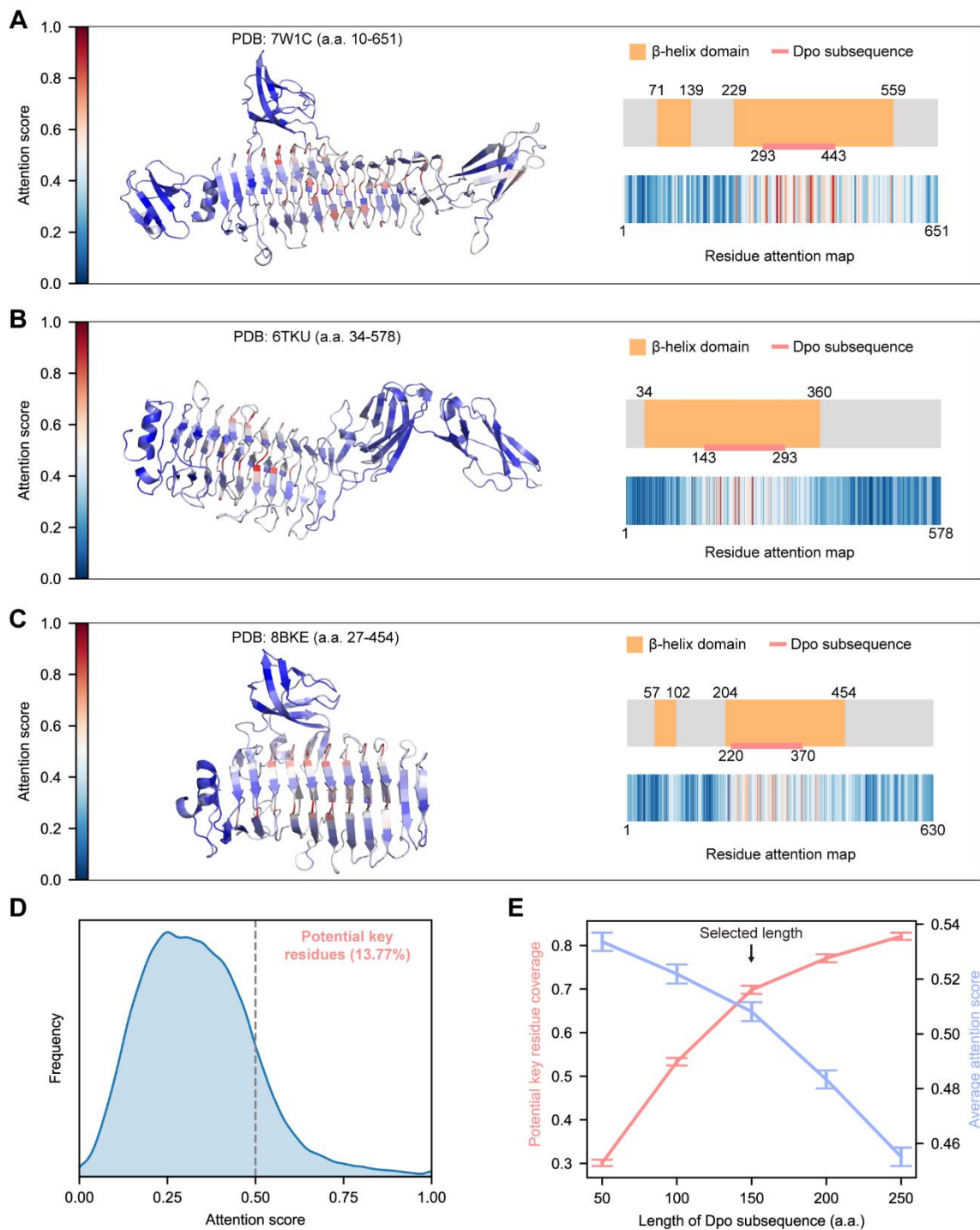


Fig. S4. Analysis of DposFinder's attention towards depolymerases (Dpos). (A–C) Cartoon representations of KL1-specific depolymerase (PDB ID: 7W1C) (A), KL21-specific depolymerase (6TKU) (B) and KL63-specific depolymerase (8BKE) (C). The color of the cartoon indicates the attention scores generated by DposFinder. The regions with higher scores are shown in red, while those with lower scores are depicted in blue. The entire attention map, the boundaries of the β -helix domain and subsequences extracted by DposFinder are shown on the right. (D) The distribution of attention scores for each residue within 384 reported depolymerases. Residues with attention scores greater than 0.5 are defined as potential key residues. (E) Coverage rate of potential key residues and average attention score for different lengths of depolymerase subsequences. The length of 150 amino acid residues is chosen as the optimal length for the subsequence.

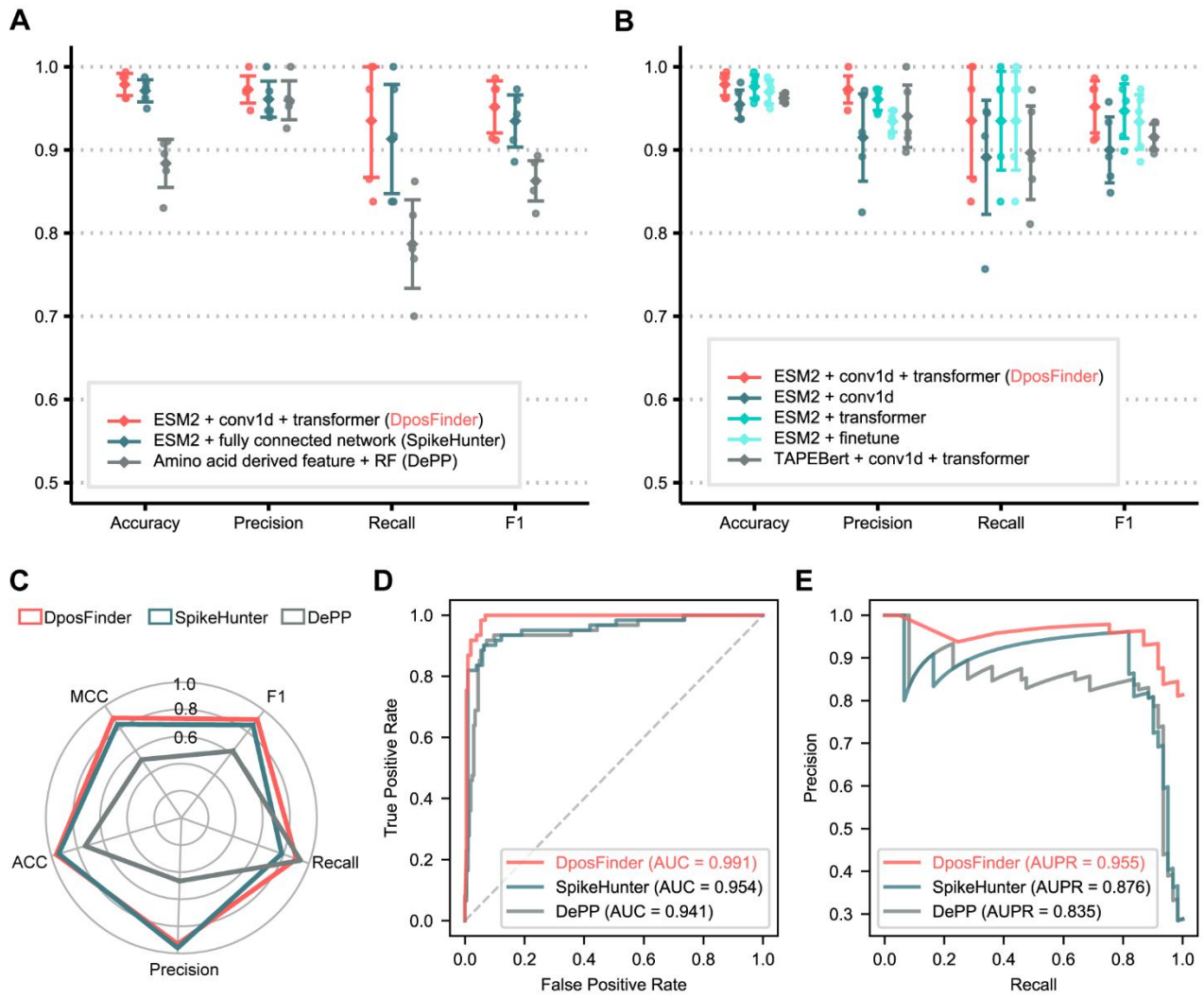


Fig. S5. Benchmark evaluation reveals that DposFinder outperforms other models. (A and B) Performance Comparison by five-fold cross-validation on the training dataset. (A) Comparison of DposFinder's performance with two other Dpo prediction tools, SpikeHunter and DePP. The results show average values with 95% confidence intervals. (B) Comparison of DposFinder against four designed models. The results show average values with 95% confidence intervals. (C–E) Benchmark evaluation of DposFinder, SpikeHunter and DePP on the independent test dataset. (C) Redar chart. (D) ROC curve. (E) Precision-Recall (PR) curve.

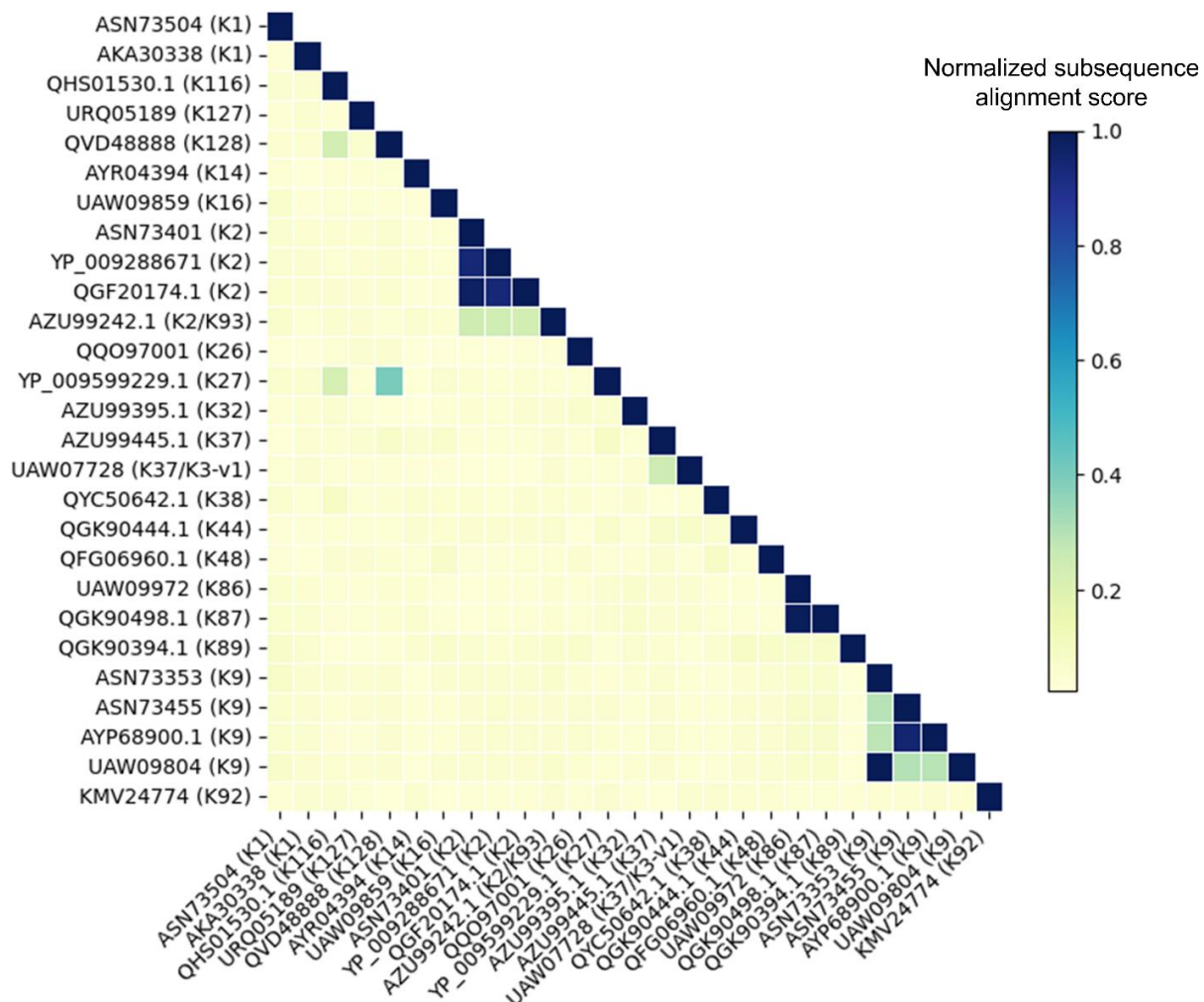


Fig. S6. Subsequence similarity heatmap of depolymerase from *A. baumannii* phages. To evaluate the effectiveness of our approach for predicting capsular serotype, we analyzed reported depolymerases from *A. baumannii* phages, for which capsular specificity (K specificity) has been experimentally determined. Subsequences from 27 *A. baumannii* phage depolymerases were extracted using DposFinder. Sequence alignment of these subsequences revealed that depolymerases targeting the same K-type exhibit similar sequence patterns, suggesting that subsequence alignment can be leveraged to predict the K-type targeted by these depolymerases.

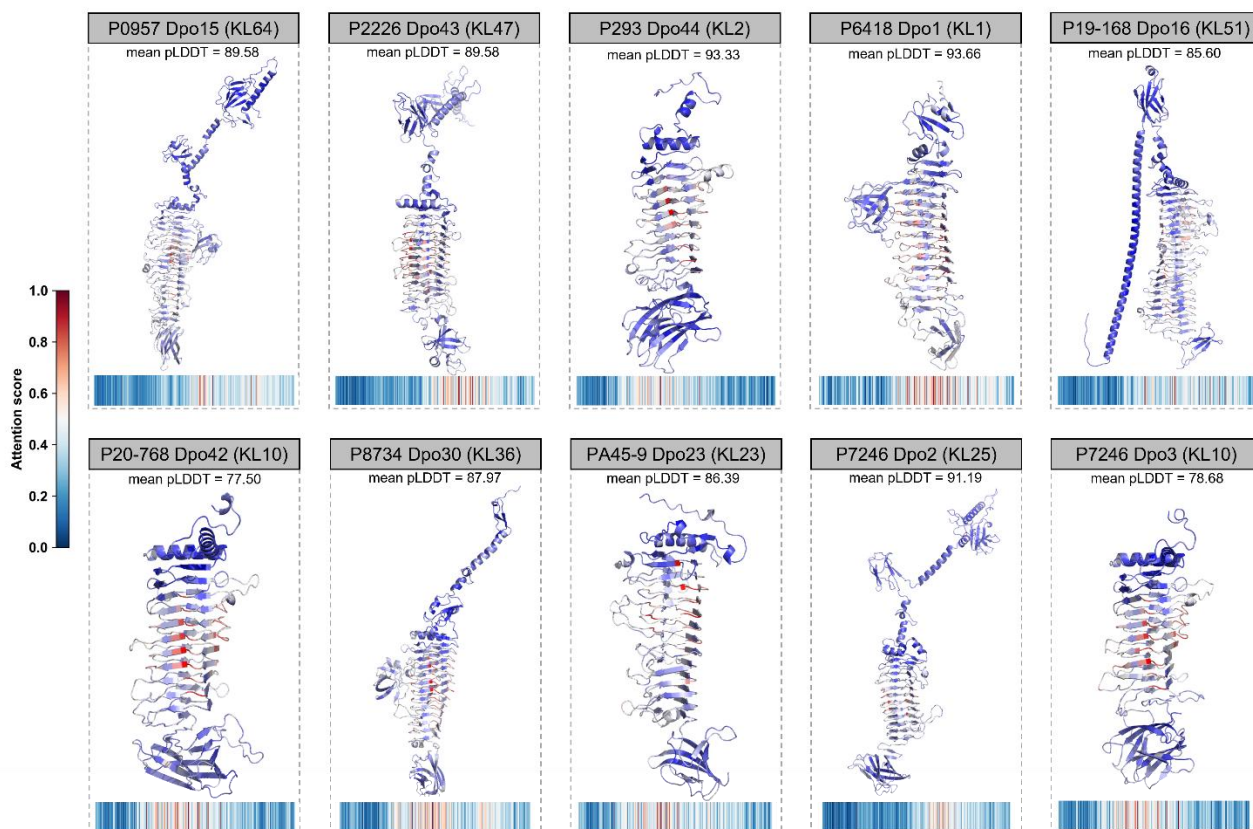


Fig. S7. Cartoon representations and attention visualization of validated depolymerases. These structures were modeled using AlphaFold2. The color of the cartoon corresponds to the attention scores generated by DposFinder, with regions exhibiting higher scores depicted in red and those with lower scores in blue.