Supporting Information for

Predicting 3D Structures of Lasso Peptides

Xingyu Ouyang,[1,2] Xinchun Ran,[1] Han Xu,[3] Runeem Al-Abssi,[1] Yi-Lei Zhao,[2]* A. James Link,[4]* and Zhongyue J. Yang[1,5-8]*

[1]*Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States* [2]*State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China* [3]*Neo Financial, 200 8 Ave SW #400, Calgary, AB T2P 1B5, Canada* [4]*Departments of Chemical and Biological Engineering, Chemistry, , and Molecular Biology, Princeton University, Princeton, New Jersey 08544, United States* [5]*Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [6]*Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [7]*Data Science Institute, Vanderbilt University, Nashville, Tennessee 37235, United States* [8]*Department of Chemical and Biomolecular Engineering, Vanderbilt University, Nashville, Tennessee 37235, United States*

Contents

Figure S1. Structure of microcin J25 4
Figure S2. Amino acid distribution for upper (left) and lower (right) plugs 5
Figure S3. Comparison of structure predicted by AlphaFold2, AlphaFold3 and ESMFold..... 6
Figure S4. Scaffold library construction of LassoPred's constructor 7
Figure S5. A 30 ns molecular dynamics trajectory for the linear core-peptide of microcin J25 8
Figure S6. Structure of the lasso peptide containing six cysteine residues 9
Text S1. Reconstruction of the isopeptide and plug position from predicted k-mer fragments in the plug classifier 10
Text S2. Comparison of fine-tuned ESM2 model 10
Text S3. Additional details for MD simulation clustering..... 10
Text S4. Estimation of the baseline probability of randomly guessing the correct plug position in an average-length lasso peptide 11
Table S1. Evaluation of AlphaFold3 prediction results for lasso peptides..... 12
Table S2. Evaluation of AlphaFold3 prediction accuracy for lasso peptides with a known annotation but undeposited structure in PDB 13
Table S3. Randomly sampled RODEO-mined lasso peptide sequences and their structural prediction results from AlphaFold3 14
Table S4. Sequences sampled by stratified sampling according to sequence length in RODEO genome mining sequences 15
Table S5 Dataset of 52 lasso peptides curated from PDB in March, 2024..... 16
Table S6. Models and hyperparameters for grid search 18
Table S7. Benchmark of k-mer fragmentation strategies for their performance in plug position prediction through repeated holdout validation 19
Table S8. Scoring formula for reconstruction of the plug position from predicted k-mer fragments in the plug classifier 20
Table S9. Performance benchmark for various reconstruction scores..... 21
Table S10. Comparison of plug prediction performance using various sequence featurization methods through repeated holdout validation..... 22
Table S11. Dataset splitting for training and testing on the selected dataset..... 23
Table S12. Predictive performance of machine learning models on the selected split..... 24
Table S13. Performance of isopeptide classifier and plug classifier on the selected dataset .. 25
Table S14. Data splitting test for plug prediction accuracy..... 26
Table S15. Model performance comparison of the original and clean dataset..... 27
Table S16. LassoPred's annotator prediction results on the hold-out test set of the selected dataset 28
Table S17. LassoPred's prediction results on the hold-out test set of the selected dataset 29
Table S18. Performance evaluation of LassoPred's constructor on the hold-out test set of the selected dataset..... 30

Table S19. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying the plug position 31

Table S20. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying plug locations on the selected dataset 32

Table S21. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying plug locations on the blind test 33

Table S22. Classification of data set (47 PDBs), based on number of disulfide bonds 34

References 35

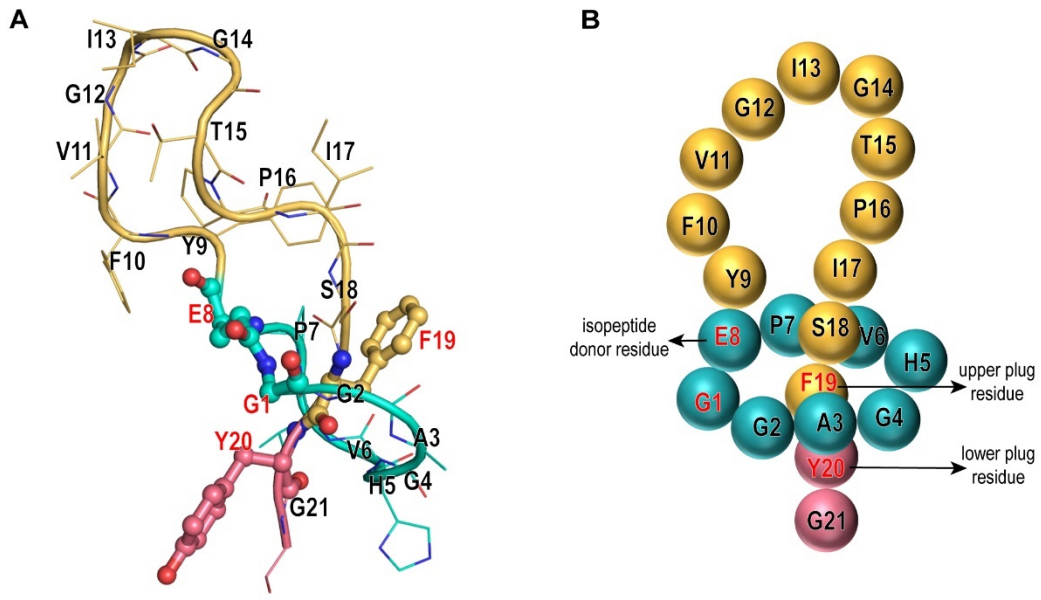

Figure S1. Structure of microcin J25. (A) PDB structure of microcin J25, PDB ID: 1PP5. The residues were displayed with lines, the key residues including isopeptide bond, upper plug and lower plug, were shown in sticks. (B) Cartoon representation of microcin J25, where residues are represented as balls, and key residues are labeled in red. We manually annotated the isopeptide residue, upper plug, lower plug, ring length, loop length, and tail length on each lasso peptide structure. For example, the residue that forms the isopeptide bond with the N-terminal residue was labeled as the "isopeptide residue" (E8). The aa distance between the N-terminal and the isopeptide residue was defined to be the "ring length" (8 aa). The residues positioned directly above and below the ring, were labeled as the "upper plug" (F19) and "lower plug" (Y20), respectively, where "above" refers to side chains located over the center or on the upper side of the macrolactam ring, and 'below" refers to one at the next position moving toward the C-terminal (side chains located lower than the macrolactam ring). The aa distance between the upper plug and the residue one position upper to the isopeptide residue was defined to be the "loop length" (11 aa). The "tail length" was determined by subtracting both the ring length and loop length from the total peptide length (2 aa).

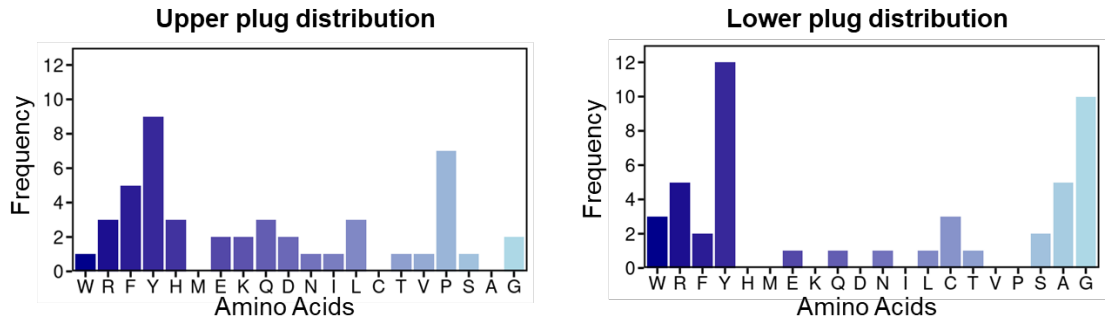

Figure S2. Amino acid distribution for upper (left) and lower (right) plugs. Lasso peptides data were collected from 47 known lasso peptide structures, with annotated plug regions.

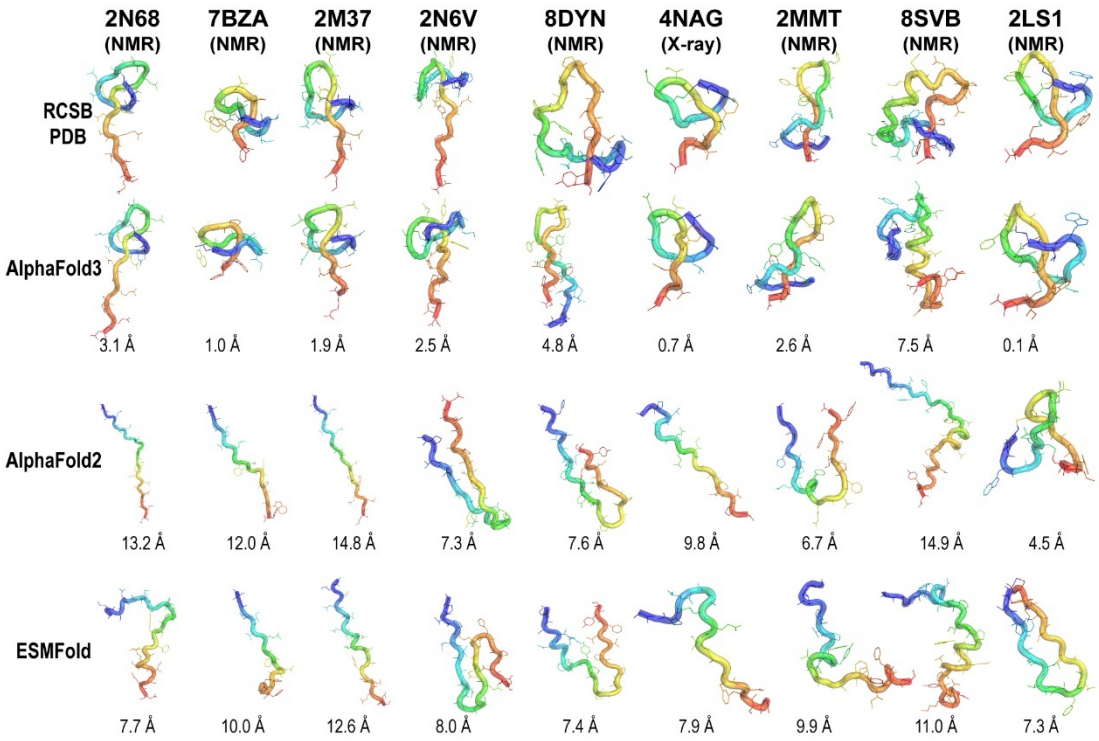

Figure S3. Comparison of structure predicted by AlphaFold2, AlphaFold3 and ESMFold. RMSD were calculated based on Cα Atoms. The structure ribbon is colored from blue at the N-terminus to red at the C-terminus.

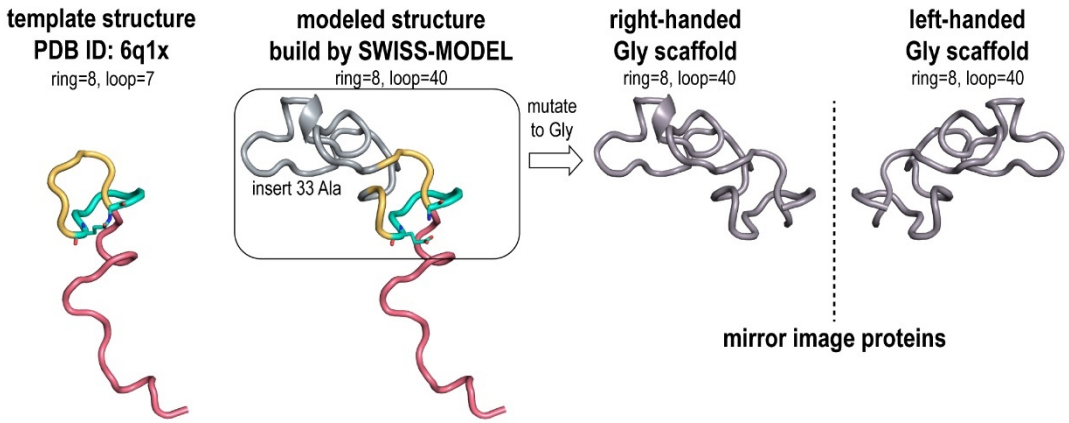

Figure S4. Scaffold library construction of LassoPred's constructor. To build a lasso peptide scaffold with a ring length of 8 aa and a loop length of 40 aa, the constructor starts from a known PDB structure template (PDB ID: 6q1x), and extends the loop length from 7 aa to 40 aa by inserting Ala residues using SWISS-MODEL[1]. These residues are then mutated to Gly. Finally, the left-handed scaffold is constructed by creating a mirror image of the right-handed Gly scaffold before adding the L-configuration amino acid.

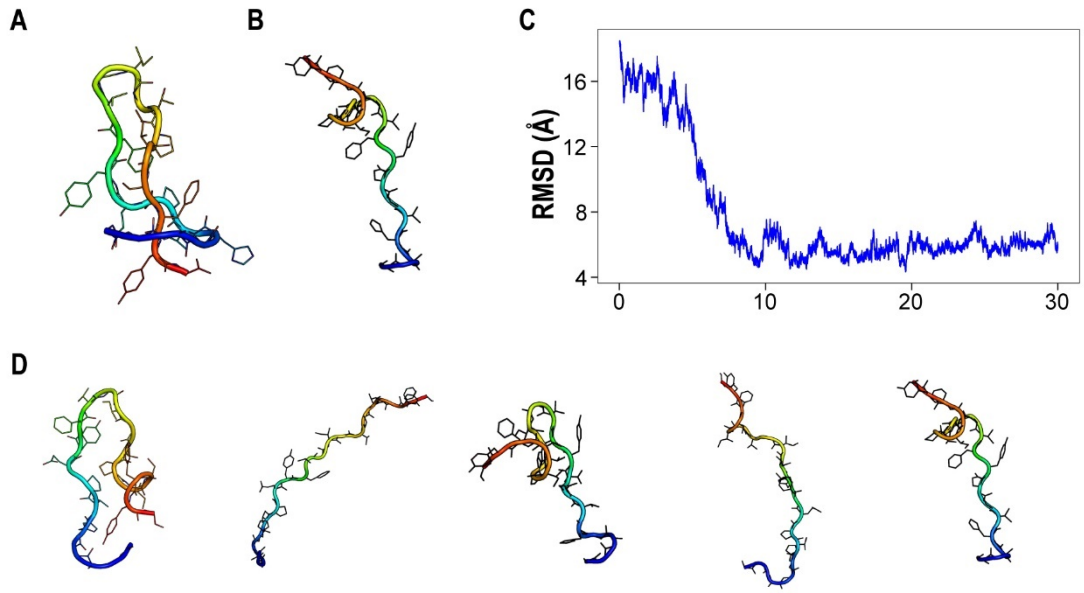

Figure S5. A 30 ns molecular dynamics trajectory for the linear core-peptide of microcin J25. (A) PDB structure of microcin J25 (PDB ID: 1Q71). (B) Initial structure of the linear core-peptide of microcin J25 built using the tleap module in Amber20[2]. (C) RMSD of Cα atoms over a 30 ns of classical MD simulation for the linear core-peptide of microcin J25, referencing to the first snapshot of the NMR structure (PDB ID: 1Q71). During the MD trajectory, none of the sampled snapshots fall within 4 Å RMSD compared to the PDB reference. (D) Representative snapshots from the 30 ns MD simulation, displayed from left to right based on clustered population rankings.

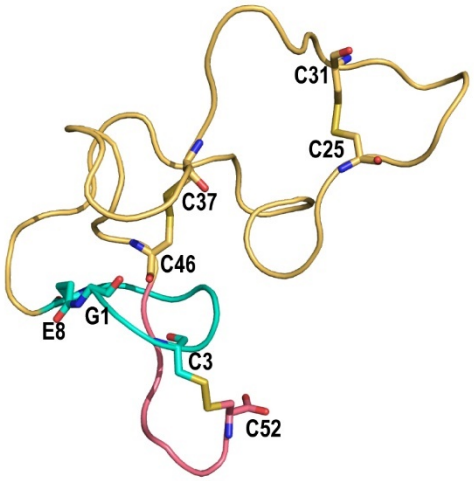

Figure S6. Structure of the lasso peptide containing six cysteine residues. The database ID for this structure is LP_OBZ16328, with the sequence GSCGWGAENIYLDRTDYYEYLQKNCVSGNNCNYETVCQIKDPPDTCYSNANC. The disulfide bonds were optimized using MD equilibration and are displayed as sticks.

Text S1. Reconstruction of the isopeptide and plug position from predicted k-mer fragments in the plug classifier.

The reconstruction strategy leverages fragment predictions from the isopeptide or plug classifier to determine the position of the isopeptide or plug residue. The k-mer fragmentation approach consistently generates k-1 fragments labeled as '0', which are fragments that reside across two sequence regions, corresponding to the isopeptide donor for the isopeptide classifier and the plug position for the plug classifier. When k > 2, algorithms need to be developed to identify which residue within these '0' fragments correspond to the isopeptide-donor or plug position. A common approach to locate the key residue is by calculating the joint probability of k-1 consecutive fragments being labeled '0'. For instance, with tripeptide fragmentation, the label sequence might appear as '1-0-0-2'. The occurrence of '00' at residues N_x and N_{x+1} suggests a potential isopeptide/plug residue at N_{x+1}. By ranking the joint probability, we can identify residues that most likely serve as the isopeptide/plug for the LaP.

To determine the fragmentation strategy, we benchmarked different fragment lengths for their prediction performance, especially the accuracy of identifying the correct isopeptide/plug residues from the top 1-3 predictions, such as dipeptide (2-mer), tripeptide (3-mer), and tetrapeptide (4-mer) (Table S7). In the case of dipeptide fragmentation, the joint probability refers to the likelihood of fragments being labeled '0'. Since '0' labels are often accompanied by '1' and '2' labels, we also explored different schemes for calculating joint probability (Table S8) and assessed their performance (Table S9).

Text S2. Comparison of fine-tuned ESM2 model.

We have fine-tuned the ESM2 language model by incorporating an additional two-layer MLP (multi-layer perceptron). This fine-tuning process was conducted using the same training and test datasets as the selected dataset (Table S11), consisting of 37 lasso sequences for training and 10 for testing. The MLP has an input dimension of 1280, matching the sequence length. We developed the binary classification models to locate the plug locations. The model was optimized using Binary Cross Entropy Loss (BCELoss)[3].

Text S3. Additional details for MD simulation clustering.

To perform clustering analysis on a MD simulation trajectory, we strip water molecules, sodium ions, and chloride ions from the trajectory to focus the analysis on the protein. We apply the autoimage command to ensure the molecules are imaged correctly within the periodic boundary conditions. The clustering analysis is then performed on the alpha carbon atoms (CA) of the protein using the cluster command, specifying 5 clusters and sampling every 5th frame to reduce computational load while maintaining representative conformations. Representative structures for each cluster are output in PDB format.

Text S4. Estimation of the baseline probability of randomly guessing the correct plug position in an average-length lasso peptide.

Based on 4,749 RODEO-predicted lasso peptides, the average length of the core sequence is 22 amino acids. Therefore, we took microcin J25 as an example, the lasso peptide whose core sequence has 21 amino acids, to compute the baseline probability of identifying the plug position. The sequence of microcin J25 is GGAGHVPEYFVGIGTPISFYG. A large portion of this sequence, specifically VGIGTPISF, could potentially host the plugs. This is determined by excluding the first 8 amino acids as the ring sequence, also by the prior knowledge that the minimum loop length is 3 and the minimum tail length is 2 amino acids. In this case, the probability of randomly identifying the correct upper plug is only 11.1% (1/9).

Table S1. Evaluation of AlphaFold3 prediction results for lasso peptides. Assessments were based on the formation of a lariat knot-like structure, the formation of the isopeptide bond, and the correct location of the upper plug. Although AlphaFold3 cannot create a covalent C–N bond in the isopeptide motif, we define a structure to contain a reasonable lariat knot-like structure when the distance between the carboxyl C of the isopeptide donor (Glu or Asp) and the amine N of the N-terminus residue is less than 4.0 Å. Among the 47 PDB structures of lasso peptides, AlphaFold3 achieves a correct prediction in 37 structures (79%).

PDB ID	Lariat knot-like structure	Broken isopeptide	Wrong upper plug	Wrong folding
1Q71	1			
1RPB	1			
2LS1		1 (d(C–N) = 4.3 Å)		
2LTI			1	
2LX6				1
2M37	1			
2MFV	1			
2MLJ	1			
2MMT	1			
2MMW	1			
2MW3	1			
2N5C	1			
2N68	1			
2N6U	1			
2N6V	1			
3NJW	1			
4NAG	1			
5D9E	1			
5GVO	1			
5JPL	1			
5JQF	1			
5OQZ	1			
5TJ1	1			
5UI6	1			
5UI7	1			
5XM4	1			
5ZCN				1
6AK0	1			
6B5W	1			
6M19	1			
6MW6	1			
6POR	1			
6Q1X		1 (d(C–N) = 4.5 Å)		
6XTH	1			
6XTI	1			
7BW5	1			
7BZ7	1			
7BZ8	1			
7BZ9		1 (d(C–N) = 4.2 Å)		
7BZA	1			
7CU6	1			
7EES				1
7JS6	1			
7LCW	1			
7ZWJ				1
8DYN				1
8SVB				1

Table S2. Evaluation of AlphaFold3 prediction accuracy for lasso peptides with a known annotation but undeposited structure in PDB[4-11]. Assessments were based on the formation of a lariat knot-like structure, the formation of the isopeptide bond, and the correct location of the upper plug. We define a AF3-predicted structure to contain a reasonable lariat knot-like structure when the distance between the carboxyl C of the isopeptide donor (Glu or Asp) and the amine N of the N-terminus residue is less than 4.0 Å. The pTM (predicted TM-score) is a confidence metric used in AlphaFold3 to assess the accuracy of predicted protein structures, which is introduced to estimate how well the predicted structure aligns with the true structure by comparing the topology or global fold of the protein.

Lasso Peptide Name	Lariat knot-like structure	Broken isopeptide	Wrong upper plug	Wrong folding	pTM
RES-701-1				1	0.01
RES-701-3				1	0.02
Capistruin	1				0.02
Lariatin A				1	0.02
Fuscanodin				1	0.03
Cellulonodin-1				1	0.02
Cellulonodin-2				1	0.03
Burhizin				1	0.24
Mycetohabin-15			1		0.02
Mycetohabin-16		1 (d(C–N) = 4.5 Å)			0.02
Caulonodin IV				1	
Caulonodin VI				1	0.03

Table S3. Randomly sampled RODEO-mined lasso peptide sequences and their structural prediction results from AlphaFold3 (see the dataset from Data.csv).

LP_ID	Core Sequence	Lariat knot-like structure (0: No; 1: Yes)
LP_14726	KVGSNVDGMGGFL	0
LP_NBU09133	GGTAGAVGDGTTST	0
LP_WP_093597056	GSASGSSDASGQSFN	1
LP_MBS1363547	STGKDNDDGYGYDKCA	1
LP_MCO5187178	SPSPGTFESGQGAGFKS	0
LP_WP_201866942	RGGEPIWEEVVVPWDYWV	0
LP_HBY96999	STTELPDLTDSTMREGERR	0
LP_15288	GQFNKRYLDIFFMTTPYTWG	0
LP_WP_184005307	GSGVLPGDEFIGLNPAGLSDE	1
LP_ALS29698	YGKGVTEVDSVTINDKDIYDPS	0
LP_WP_127536561	AGTGFRQIDWISEHDADLYDPTS	0
LP_WP_160661078	STSSGNRTDAAYANQAPLITGALS	0
LP_ESZ75755	SGLSIAGSELLLHRNSRMAAFMSQA	0
LP_1338	GIKQLGEPDFEDGDHYGLVGAISTDD	0
LP_MCB2018628	GGSAGVIETKSANTGGAQCGGNMTKRC	0
LP_WP_161355608	GSSAGAYIDGGTVPWIYSHFRPGGGSWE	0
LP_17779	TINPSNIDDVNEETGDEGQGSFSVVKEEG	0
LP_WP_191758872	GWRGWRRDHRHHRRRRRHGGGHGGHGGHGG	0
LP_MCA0197304	TGSLNGQEDLVENGVCTMIDINNINNMYNMC	0
LP_WP_126051746	NGPGNANADCFDVSDGKTETHPGRSNASCLGS	0
LP_WP_078432603	WNLKGTHHDGAWTEHISNPHDNGDGTTTESQMS	0
LP_WP_098257454	GCGGWGCETGLNNYSYHQEGDYCRDTIPGESGQC	0
LP_ADP78328	GRWGWGRDWLWRSFPRYYGGGGGAIIIVGGGGRRY	0
LP_MBW4652682	AVGDVSFSDTVFLSASVNPVAIGDSAGSRDAIVVPL	0
LP_WP_000714527	GKSGWGAEGFTLDKVGSRMTNKFQSTHIEPGVLKTTK	0
LP_MBK1699556	GSRLAISDSWFGTDGNDGLLGPKCDPDSDFLACTADGS	0
LP_WP_214430368	LLGKSDKDDFLFFPGSSSPASNPATNSPITGDGSVDLHQ	0
LP_RBP88292	NGPGNANPDCFDVSEGRHETHEGQSNASCLPGDIGGSFES	0
LP_WP_215607803	GSGAVNLGDSVIFTNLPRGANDLEIESVGSVDVTIDFEDLF	0
LP_WP_021169006	GDGDSDNVDGSGYYTSSQCNKDCQQYPCPNPHNCPYGDAQPL	0
LP_GAX44210	VSGRNAIRDTLFFNGQPTNVQSDDSLDLLTTSNGSPTTRPAGQ	0
LP_WP_098681776	GCGGWGSEAWTFDDSRSQFRWIQLKGQGGDRSVCVCTSISSEEC	0
LP_PSO63441	MLGEASAQDALVWAGDPIARSSGAKPAAGGSFDLESKSDLDWDAW	0
LP_WP_190649936	IGDFILPENFVHYFSKEQISSELKEAGFRLEFYSELNYGHAVGIAE	0
LP_RAM51547	YSGEIGFQDLFFGNPDTAPTKGNGTLNACIQVNGQCLDPNAKGKGID	0
LP_WP_011938274	RGVFWEGKELYPNAGFREKNHIQICIRNLNCIKGYFHPRKPLDSYPTP	0
LP_16988	YSFDSKDGDVFYGQTNATNQNGTGSLNACITYKTGLKKGDCVNPSAKGV	0
LP_WP_029518785	GECGWGTENATFDKTGAYKTTKRRMELVQMFPQEVRVCKVVEACSSDSNEC	0
LP_OBZ16328	GSCGWGAENIYLDRTDYYEYLQKNCVSGNNCNYETVCQIKDPPDTCYSNANC	0
LP_ANC22478	GDCSWGKENFALDETGAYYTKKRKYVHAGWLPGYKEVFKCAVVDGCSSDSNQC	0

Table S4. Sequences sampled by stratified sampling according to sequence length in RODEO genome mining sequences.

Name	Sequence
^{LP}Test1	EIVVAGDETSGT
^{LP}Test2	GWYGQHWDGPTGQRD
^{LP}Test3	GILQGNEPMGGEPVPGISEE
^{LP}Test4	SGMGTQNEGVWFMFGMVTICNFNGTRYPCM
^{LP}Test5	GDNMANFDKPLTTKDILGKLAVIKRKNHLIEVDSTVYEKLGKLIVKIYS LNYYATYFILNLYNFIKV

Table S5 Dataset of 52 lasso peptides curated from PDB in March, 2024. The duplicate data that were removed from the dataset (PDB ID: 1PP5, 6N60, 4CU4, 1RPC, 2M8F). In total, 47 unique LaP PDBs were used for building machine learning models.

PDB ID	Sequence	Total Len. (aa)	Iso.	Ring Len. (aa)	UP.	Loop Len. (aa)	Tail Len. (aa)
1PP5	GGAGHVPEYFVGIGTPISFYG	21	8	8	19	11	2
1Q71	GGAGHVPEYFVGIGTPISFYG	21	8	8	19	11	2
6N60	GGAGHVPEYFVGIGTPISFYG	21	8	8	19	11	2
4CU4	GGAGHVPEYFVGIGTPISFYG	21	8	8	19	11	2
1RPB	CLGIGSCNDFAGCGYAVVCFW	21	9	9	15	6	6
1RPC	CLGIGSCNDFAGCGYAVVCFW	21	9	9	15	6	6
2LS1	CVWGGDCTDFLGCGTAWICV	20	9	9	15	6	5
2LTI	GLSQGVEPDIGQTYFEESRINQD	23	9	9	17	8	6
2LX6	GAFVGQPEAVNPLGREIQG	19	8	8	15	7	4
2M37	GLSQGVEPDIGQTYFEESR	19	9	9	14	5	5
2M8F	GPTPMVGLDSVSGQYWDQHAPLAD	24	9	9	15	6	9
2MFV	GGPLAGEEMGGITT	14	7	7	10	3	4
2MLJ	SIGDSGLRESMSSQTYWP	18	9	9	16	7	2
2MMT	GGAGHVPEYFVRGDFPISFYG	21	8	8	19	11	2
2MMW	GGAGHVPEYFVRGDTPISFYG	21	8	8	19	11	2
2MW3	SLGSSPYNDILGYPALIVIYP	21	9	9	14	5	7
2N5C	GFGSKPLDSFGLNFF	15	8	8	12	4	3
2N68	GLSQGVEPDIGQTYFEESRINQD	23	9	9	14	5	9
2N6U	GLTQIQALDSVSGQFRDQLG	20	9	9	15	6	5
2N6V	GPTPMVGLDSVSGQYWDQHAPLAD	24	9	9	15	6	9
3NJW	GLPWGCPSDIPGWNTPWAC	19	9	9	14	5	5
4NAG	GGPLAGEEIGGFNVPG	16	7	7	10	3	6
5D9E	GTLTPGLPEDFLPGHYMPG	19	9	9	15	6	4
5GVO	GLPIGWWIERPSGWYFPI	18	9	9	14	5	4
5JPL	GRPNWGFENDWSCVRVC	17	8	8	12	4	5
5JQF	GIEPLGPVDEDQGEHYLFAGG	21	9	9	15	6	6
5OQZ	GAPSLINSEDNPAFPQRV	18	9	9	16	7	2
5TJ1	GVGFGRPDSILTQEQAKPM	19	8	8	14	6	5
5UI6	GGKGPIFETWVTEGNYYG	18	8	8	16	8	2
5UI7	GSDGPIIEFFNPNGVMHYG	19	8	8	17	9	2
5XM4	GPPGDRIEFGVLAQLPG	17	8	8	12	4	5
5ZCN	DGMGEEFIEGLVRDSLYPPAG	21	9	9	14	5	7
6AK0	CLGVGSCVDFAGCGYAVVCFW	21	9	9	15	6	6
6B5W	GVGFGRPDSILTQEQAKPM	19	8	8	15	7	4
6M19	LVVIVQADWNAPGFF	15	8	8	12	4	3
6MW6	GGVGKIIEYFIGGGVGRYG	19	8	8	17	9	2
6POR	GGDGSIAEYFNRPMHIHDWQIMDSGYYG	28	8	8	26	18	2
6Q1X	GVLGNDAEGITLLPLCFKPICIPTLPPLTGGHA	33	8	8	15	7	18
6XTH	GSRGWGFEPGVRCLIWCD	18	8	8	11	3	7
6XTI	GGGGRGYEYNKQCLIFC	17	8	8	12	4	5
7BW5	GPKGDFPDVGDGRILAG	17	8	8	11	3	6
7BZ7	LVVIVQADWNAPGWY	15	8	8	12	4	3
7BZ8	LVAIVQADWNAPGWF	15	8	8	12	4	3
7BZ9	LVVAVQADWNAPGWF	15	8	8	12	4	3
7BZA	LVVIVQADWNAPGWF	15	8	8	12	4	3
7CU6	LCVIVQADWNCPGWF	15	8	8	12	4	3

7EES	GTIDPQNSEEHPVLSRRLEN	20	9	9	16	7	4
7JS6	LLGRSGNDRLILSKN	15	8	8	11	3	4
7LCW	GSKYSDTADESSYRW	15	9	9	13	4	2
7ZWJ	SKKSKPGDGIRGKGVRG	17	8	8	13	5	4
8DYN	GHSVDRIPEYFGPPGLPGPVLFYS	24	9	9	22	13	2
8SVB	GGGGPTPEYFLMPIDPAWLQANLP NTGKYN	30	8	8	28	20	2

Table S6. Models and hyperparameters for grid search.

Model	Grid
KNeighborsClassifier	n_neighbors: [5, 10] weights: [uniform distance] algorithm: [ball_tree kd_tree brute] leaf_size: [15 30] p: [1 2]
SVC	C: [0.1 1.0 10.0] kernel: [linear poly rbf sigmoid] degree: [3 4 5] gamma: [scale]
RandomForestClassifier	n_estimators: [50 100 150] criterion: [log_loss] max_depth: [None 10 20] min_samples_split: [2 4 5 6] min_samples_leaf: [1 2 3 4]
GradientBoostingClassifier	n_estimators: [20 50 100] criterion: [friedman_mse squared_error] min_samples_split: [2 5] min_samples_leaf: [1 2]

Table S7. Benchmark of k-mer fragmentation strategies for their performance in plug position prediction through repeated holdout validation. Using sequence embedding derived from ESM_L33_650M, each of the 100 splits was tested with four different machine learning models: Random Forest Classifier (RFC), K-Neighbors Classifier (KNC), Gradient Boosting Classifier (GBC), and Support Vector Classifier (SVC). The parameters for each model were optimized through grid search, and the model that achieves the highest accuracy is used in the performance statistics. A k-mer refers to a piece of k amino acids extracted from the lasso peptide sequence. The ring-truncated k-mer refers to a strategy in which the sequence piece corresponding to the ring region is removed in building the plug classifier. As the minimum loop length for experimentally characterized LaPs is 3 amino acids, the truncation strategy was not performed on 4-mers and 5-mers. The AUC ROC value is weighted by one-versus-rest test. Top 1, 2, 3 accuracy refers to the accuracy of finding the correct plug position from the top 1, 2, 3-ranked residues, respectively. In the main text, 2-mer, 3-mer, and 4-mer correspond to dipeptide, tripeptide, and tetrapeptide, respectively. Values are presented as mean ± standard deviation.

	ROC AUC (ovr-weighted)	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
k=2 (ring-truncated)	0.91±0.04	0.60±0.15	0.75±0.14	0.85±0.10	0.74±0.12	0.85±0.09	0.91±0.06
k=3 (ring-truncated)	0.91±0.03	0.52±0.15	0.70±0.14	0.82±0.12	0.67±0.13	0.81±0.10	0.90±0.07
k=2	0.94±0.02	0.57±0.15	0.72±0.12	0.80±0.10	0.72±0.13	0.83±0.09	0.89±0.07
k=3	0.95±0.02	0.53±0.14	0.68±0.13	0.82±0.11	0.68±0.13	0.80±0.09	0.90±0.07
k=4	0.96±0.02	0.47±0.15	0.64±0.14	0.79±0.12	0.63±0.15	0.78±0.11	0.88±0.08

Table S8. Scoring formula for reconstruction of the plug position from predicted k-mer fragments in the plug classifier. Here, i represents the index of a k-mer, and the probabilities of the k-mer being classified into categories 0, 1, and 2 are denoted as $P(0)$, $P(1)$, and $P(2)$, respectively. Z scores are used to standardize the data for each category, represented as $Z(P(0))$, $Z(P(1))$ and $Z(P(2))$. The boundary k-mer is expected to be in category 0, with the preceding k-mer in category 1 and the following k-mer in category 2. The scoring function is used to evaluate the likelihood of the i-th k-mer becoming the boundary k-mer.

	Scoring Method
2-mer scoring1	$P_i(0)$
2-mer scoring2	$P_{i-1}(1) * P_i(0) * P_{i+1}(2)$
2-mer scoring3	$P_{i-1}(1) * P_i(0)$
2-mer scoring4	$P_i(0) * P_{i+1}(2)$
2-mer scoring5	$Z(P_{i-1}(1)) + Z(P_i(0)) + Z(P_{i+1}(2))$
3-mer scoring1	$P_i(0) * P_{i+1}(0)$
3-mer scoring2	$P_i(0) * P_{i+1}(0) + P_{i-1}(1) * P_{i+2}(2)$
3-mer scoring3	$P_{i-1}(1) * P_i(0) * P_{i+1}(0) * P_{i+2}(2)$
3-mer scoring4	$P_{i-1}(1) + P_i(0) + P_{i+1}(0) + P_{i+2}(2)$
3-mer scoring5	$P_{i-1}(1) * P_i(0) * P_{i+1}(0)$
3-mer scoring6	$P_i(0) * P_{i+1}(0) * P_{i+2}(2)$
3-mer scoring7	$Z(P_{i-1}(1)) + Z(P_i(0)) + Z(P_{i+1}(0)) + Z(P_{i+2}(2))$
3-mer scoring8	$0.5 * Z(P_{i-1}(1)) + Z(P_i(0)) + Z(P_{i+1}(0)) + 0.5 * Z(P_{i+2}(2))$

Table S9. Performance benchmark for various reconstruction scores. This benchmark evaluates the performance of various scoring functions in reconstructing the plug position based on outcomes of k-mer classification. The assessment was conducted using the full sequences and the ESM2 L33 featurization method. Different scoring methods are summarized in Table S8. Values are presented as mean ± standard deviation.

	ROC AUC (ovr-weighted)	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
2-mer scoring1	0.94±0.02	0.57±0.15	0.72±0.12	0.80±0.10	0.72±0.13	0.83±0.09	0.89±0.07
2-mer scoring2	0.94±0.02	0.48±0.15	0.64±0.14	0.75±0.13	0.64±0.14	0.77±0.11	0.85±0.09
2-mer scoring3	0.94±0.02	0.49±0.15	0.67±0.14	0.79±0.14	0.65±0.14	0.79±0.11	0.87±0.09
2-mer scoring4	0.94±0.02	0.52±0.15	0.66±0.14	0.74±0.12	0.67±0.14	0.79±0.10	0.84±0.08
2-mer scoring5	0.94±0.02	0.53±0.14	0.67±0.14	0.76±0.13	0.68±0.13	0.79±0.10	0.86±0.08
3-mer scoring1	0.95±0.02	0.53±0.14	0.68±0.13	0.82±0.11	0.68±0.13	0.80±0.09	0.90±0.07
3-mer scoring2	0.95±0.02	0.46±0.14	0.60±0.13	0.73±0.13	0.62±0.13	0.74±0.11	0.83±0.09
3-mer scoring3	0.95±0.02	0.48±0.14	0.63±0.14	0.80±0.12	0.63±0.13	0.77±0.11	0.88±0.08
3-mer scoring4	0.95±0.02	0.48±0.14	0.64±0.13	0.79±0.11	0.64±0.13	0.78±0.10	0.88±0.07
3-mer scoring5	0.95±0.02	0.48±0.13	0.65±0.13	0.81±0.12	0.64±0.12	0.78±0.10	0.89±0.07
3-mer scoring6	0.95±0.02	0.49±0.15	0.64±0.14	0.75±0.13	0.64±0.14	0.77±0.11	0.85±0.09
3-mer scoring7	0.95±0.02	0.51±0.14	0.66±0.13	0.81±0.12	0.67±0.12	0.79±0.10	0.89±0.07
3-mer scoring8	0.95±0.02	0.52±0.14	0.68±0.13	0.82±0.11	0.68±0.12	0.80±0.09	0.90±0.07

Table S10. Comparison of plug prediction performance using various sequence featurization methods through repeated holdout validation. The assessment was conducted using ring-truncated sequences and the dipeptide fragmentation. “ESM2 L33” refers to the output from the 33rd layer of the “esm2_t33_650M_UR50D” model[12]. “ESM2 L32 Seq” is the output from the 32nd layer of the same model with mean representation of the peptide sequence, and “ESM2 L33 Seq” refers to the output from the 33rd layer with mean sequence representation. “SaProt 35M AF2” features were adopted from the SaProt 35M AF2 models [13]. “AF2 Distogram” is a predicted pairwise residue distances generated by the AlphaFold2 model, “AF2 Structure Module” is structural predictions generated by the AlphaFold2 model’s structure module. “AF2 Masked MSA” derived from the masked multiple sequence alignment used by the AlphaFold2 model. “AF2 Predicted LDDT” is the predicted local distance difference test (LDDT) scores, which indicate the confidence of the AlphaFold2 model in its structural predictions. Machine learning models and relevant hyperparameters were selected using a grid search across four algorithms: Random Forest Classifier (RFC), K-Neighbors Classifier (KNC), Gradient Boosting Classifier (GBC), and Support Vector Classifier (SVC). The machine learning model for each feature, based on the average performance dataset, is labeled accordingly. Values are presented as mean ± standard deviation.

	ROC AUC (ovr-weighted)	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
ESM2 L33 (SVC)	0.91±0.04	0.60±0.15	0.75±0.14	0.85±0.10	0.74±0.12	0.85±0.09	0.91±0.06
ESM2 L32 Seq (SVC)	0.90±0.04	0.58±0.15	0.72±0.13	0.83±0.11	0.72±0.13	0.83±0.09	0.90±0.07
ESM2 L33 Seq (SVC)	0.90±0.04	0.58±0.15	0.72±0.13	0.83±0.11	0.72±0.13	0.83±0.09	0.90±0.07
SaProt 35M AF2 (SVC)	0.77±0.06	0.44±0.13	0.62±0.13	0.73±0.14	0.60±0.13	0.76±0.10	0.84±0.09
AF2 Distogram (SVC)	0.67±0.05	0.20±0.12	0.39±0.14	0.53±0.12	0.32±0.17	0.55±0.15	0.68±0.11
AF2 Structure Module (RFC)	0.80±0.05	0.39±0.13	0.59±0.13	0.74±0.12	0.54±0.14	0.73±0.11	0.84±0.08
AF2 Masked MSA (KNC)	0.57±0.05	0.37±0.15	0.57±0.15	0.69±0.13	0.53±0.16	0.71±0.13	0.81±0.10
AF2 Predicted LDDT (SVC)	0.84±0.05	0.21±0.12	0.42±0.13	0.58±0.14	0.33±0.16	0.58±0.13	0.72±0.12

Table S11. Dataset splitting for training and testing on the selected dataset. The selected dataset was determined to be the one with an average performance metrics across the 100 splits used in the repeated holdout validation (with ESM2 L33 feature). The performance metrics include Accuracy (the accuracy to classify fragments to Class 0, 1, 2), F1 Score (the F1 score to classify fragments to Class 0, 1, 2), Overall ROC AUC (ovr-micro), ROC AUC Class 0, ROC AUC Class 1, ROC AUC Class 2, Top 1 Accuracy, Top 2 Accuracy, and Top 3 Accuracy.

Dataset	PDB IDs
Training Set	5ZCN, 7BW5, 2LTI, 1Q71, 2MMW, 8DYN, 2N6U, 2MFV, 5UI7, 5D9E, 7JS6, 8SVB, 7ZWJ, 6Q1X, 3NJW, 7CU6, 7BZ8, 6M19, 7BZA, 7BZ7, 2MW3, 5OQZ, 6XTI, 6B5W, 2LX6, 7EES, 5JPL, 2LS1, 6XTH, 1RPB, 6AK0, 6POR, 5UI6, 2M37, 2N68, 7LCW, 2MLJ
Test Set	2MMT, 2N5C, 2N6V, 4NAG, 5GVO, 5JQF, 5TJ1, 5XM4, 6MW6, 7BZ9

Table S12. Predictive performance of machine learning models on the selected split (Table S11). The hyperparameters were determined through grid search, and the values represent the average performance on the validation set using 5-fold cross-validation.

		RandomForest Classifier	SupportVector Classifier	GradientBoostingClassifier	KNeighborsClassifier
Isopeptide classifier	Fragment classification accuracy	0.72	0.90	0.82	0.82
	Fragment classification F1 Score (weighted)	0.67	0.90	0.81	0.82
	Overall ROC AUC (ovr-weighted)	0.91	0.96	0.90	0.91
Plug classifier	Fragment classification accuracy	0.71	0.69	0.69	0.63
	Fragment classification F1 Score (weighted)	0.67	0.69	0.67	0.62
	Overall ROC AUC (ovr-weighted)	0.81	0.85	0.79	0.73

Table S13. Performance of isopeptide classifier and plug classifier on the selected dataset (Table S11).

	Isopeptide Classifier	Plug Classifier
Overall ROC AUC (ovr-weighted)	0.98	0.90
Fragment Classification F1 Score (weighted)	0.90	0.81
Fragment Classification Accuracy	0.90	0.81
Boundary Accuracy	1.00 (ring-loop boundary)	0.6 (loop-tail boundary)
Class 1 Accuracy	0.86 (ring label)	0.94 (loop label)
Class 2 Accuracy	1.00 (loop label)	0.71 (tail label)
Top 1 Accuracy	1.00	0.60
Top 2 Accuracy	1.00	0.80
Top 3 Accuracy	1.00	0.90

Table S14. Data splitting test for plug prediction accuracy using ring-truncated dipeptide fragmentation, ESM2_L33 embedding, and the SVC model with optimized hyperparameters on the selected holdout set. The 47-sequence dataset was split under different test set sizes. For each splitting ratio, performance metrics were assessed via repeated holdout validation of 100 splits. Values are presented as mean ± standard deviation.

Test set size (seqs)	ROC AUC (ovr-weighted)	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
8	0.91±0.04	0.58±0.15	0.76±0.14	0.86±0.11	0.72±0.13	0.85±0.09	0.92±0.07
10	0.92±0.04	0.63±0.13	0.75±0.13	0.85±0.10	0.76±0.10	0.85±0.09	0.91±0.06
12	0.91±0.04	0.61±0.10	0.75±0.10	0.85±0.09	0.75±0.08	0.86±0.07	0.92±0.05
16	0.90±0.03	0.59±0.09	0.73±0.09	0.84±0.08	0.74±0.07	0.84±0.06	0.91±0.05
24	0.88±0.02	0.57±0.09	0.72±0.08	0.83±0.08	0.72±0.08	0.84±0.06	0.91±0.05

Table S15. Model performance comparison of the original and clean dataset using ring-truncated dipeptide fragmentation, ESM2_L33 embedding, and the SVC model with optimized hyperparameters on the selected holdout set. The original dataset includes 47 sequences (37 in the training set and 10 in the test set). The clean dataset with <80% similarity sequences include 36 sequences (29 in the training set and 7 in the test set). The removed PDB IDs are: 2MMT, 2MMW, 6AK0, 2M37, 2N68, 7BZ8, 7BZ9, 6M19, 7BZA, 7BZ7, 6B5W. For both datasets, all performance metrics were evaluated using repeated holdout validation across 100 splits, applying a 4:1 training-to-test ratio and stratified sampling. Values are presented as mean ± standard deviation.

	ROC AUC (ovr-weighted)	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
Original Dataset	0.92±0.04	0.63±0.13	0.75±0.13	0.85±0.10	0.76±0.10	0.85±0.09	0.91±0.06
Clean Dataset	0.88±0.05	0.49±0.18	0.69±0.19	0.86±0.13	0.64±0.17	0.80±0.14	0.92±0.08

Table S16. LassoPred's annotator prediction results on the hold-out test set of the selected dataset (Table S11). "Iso." refers to the isopeptide residue, and "UP" refers to the upper plug residue. Correctly predicted results are highlighted in red.

	PDB ID	True Iso.	Predicted Iso.	True UP	Predicted UP1	Predicted UP2	Predicted UP3
^{test}V1	2MMT	8	8	19	19	17	14
^{test}V2	2N5C	8	8	12	12	13	11
^{test}V3	2N6V	9	9	15	19	15	14
^{test}V4	4NAG	7	7	10	12	13	10
^{test}V5	5GVO	9	9	14	15	14	16
^{test}V6	5JQF	9	9	15	15	18	16
^{test}V7	5TJ1	8	8	14	15	13	17
^{test}V8	5XM4	8	8	12	12	14	11
^{test}V9	6MW6	8	8	17	17	15	13
^{test}V10	7BZ9	8	8	12	12	13	11

Table S17. LassoPred's prediction results on the hold-out test set of the selected dataset (Table S11). RMSD1-RMSD3 correspond to the top-1 to top-3 results, with the correct annotations highlighted in red. RMSD values were calculated with reference to the experimental structures. "all Cα" refers to the calculations using the Cα atoms of the full lasso peptide. "key Cα" refers to the calculations using local interlocked structural moiety that consists of the isopeptide-donating residue, plug residues, along with their adjacent residues (a total of 4 Cα atoms).

PDB ID	RMSD1 (all Cα)	RMSD2 (all Cα)	RMSD3 (all Cα)	RMSD1 (key Cα)	RMSD2 (key Cα)	RMSD3 (key Cα)
2MMT	2.6	4.1	7.4	1.6	1.8	6.6
2N5C	3.3	2.1	2.5	1.1	1.6	1.0
2N6V	6.2	4.8	5.9	3.1	1.1	2.3
4NAG	4.9	5.6	4.4	1.3	1.1	0.4
5GVO	2.0	1.2	4.2	1.1	0.0	2.6
5JQF	3.5	4.8	3.2	0.6	2.8	2.0
5TJ1	3.0	4.2	3.9	2.0	1.5	2.2
5XM4	3.1	3.2	3.0	1.0	1.1	1.1
6MW6	3.2	4.4	7.0	2.0	1.6	5.1
7BZ9	3.1	2.0	2.7	1.5	1.7	1.4

Table S18. Performance evaluation of LassoPred's constructor on the hold-out test set of the selected dataset (10 samples, Table S11). The predicted structures were selected based on the correct annotations or the top 1 prediction result (PDB ID 5TJ1). RMSD were calculated with all measurements using the experimental structures as reference. Average RMSD and standard deviation are given in Ångströms.

	LassoPred Optimized Results					
Test ID	**PDB ID**	**RMSD**	**RMSD(ring)**	**RMSD(loop)**	**RMSD(tail)**	**RMSD(ring+loop)**
^{test}V1	2MMT	2.6	0.8	2.9	0.0	2.5
^{test}V2	2N5C	3.3	1.6	1.8	0.2	2.9
^{test}V3	2N6V	4.8	2.0	1.0	2.4	2.0
^{test}V4	4NAG	4.4	0.5	0.2	2.4	0.5
^{test}V5	5GVO	1.2	0.3	0.0	1.4	0.2
^{test}V6	5JQF	3.5	1.5	1.1	2.1	1.6
^{test}V7	5TJ1	3.0	1.0	1.6	2.2	2.0
^{test}V8	5XM4	3.1	1.5	1.0	1.2	2.3
^{test}V9	6MW6	3.2	0.9	3.4	0.0	3.4
^{test}V10	7BZ9	3.1	1.5	1.1	0.8	3.0
	AVG±SD	3.2±0.9	1.2±0.5	1.4±1.0	1.3±0.9	2.0±1.0

Table S19. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying the plug position. LassoPred annotator is a machine learning model based on Support Vector Classifier and the ESM2-based approach is a deep learning method that uses the fine-tuned ESM2 model with a two-layer MLP, see details in Text S2. All performance metrics were evaluated using repeated holdout validation across 100 splits, applying a 4:1 training-to-test ratio and stratified sampling. Values are presented as mean ± standard deviation.

	ROC AUC	Top 1 Accuracy	Top 2 Accuracy	Top 3 Accuracy	Top 1 F1 Score	Top 2 F1 Score	Top 3 F1 Score
LassoPred Annotator	0.91±0.04 (ovr-weighted)	0.60±0.15	0.75±0.14	0.85±0.10	0.74±0.12	0.85±0.09	0.91±0.06
ESM2 Annotator	0.89±0.03	0.64±0.11	0.69±0.10	0.81±0.07	0.76±0.15	0.80±0.12	0.90±0.04

Table S20. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying plug locations on the selected dataset (Table S11). The ESM2-based model is the same as the one used in Table S19.

	LassoPred Annotator (for plug locations)	ESM2 Annotator (for plug locations)
Overall ROC AUC	0.90 (ovr-weighted)	0.71
Fragment Classification F1 Score	0.81 (weighted)	0.72
Fragment Classification Accuracy	0.81	0.70
Boundary Accuracy	0.6 (loop-tail boundary)	/
Class 1 Accuracy	0.94 (loop label)	/
Class 2 Accuracy	0.71 (tail label)	/
Top 1 Accuracy	0.60	0.70
Top 2 Accuracy	0.80	0.80
Top 3 Accuracy	0.90	0.90

Table S21. Performance comparison of the LassoPred annotator and fine-tuned ESM2-based classifiers for identifying plug locations on the blind test. The performance was assessed based on the blind test of 12 LaPs (Table S2) with a known annotation but undeposited structure in PDB.

	LassoPred Annotator (for plug locations)	ESM2 Annotator (for plug locations)
Top 1 Accuracy	0.58	0.58
Top 2 Accuracy	0.67	0.75
Top 3 Accuracy	0.92	0.75

Table S22. Classification of data set (47 PDBs), based on number of disulfide bonds. Class I lasso peptides contain two disulfide bonds. Class II peptides contain no disulfide bond. Class III and Class IV peptides each have one disulfide bond, with Class III bonds between the macrocyclic ring and the tail, and Class IV bonds within the tail itself. Notably, the structure of 7CU6 does not belong to any of the four classes currently used to classify lasso peptides[14].

PDB ID	sequence	Number of Cys	Class
2MFV	GGPLAGEEMGGITT	0	II
7JS6	LLGRSGNDRLILSKN	0	II
2N5C	GFGSKPLDSFGLNFF	0	II
7CU6	LCVIVQADWNCPGWF	2	NA
6M19	LVVIVQADWNAPGFF	0	II
7BZ8	LVAIVQADWNAPGWF	0	II
7BZ9	LVVAVQADWNAPGWF	0	II
7BZA	LVVIVQADWNAPGWF	0	II
7BZ7	LVVIVQADWNAPGWY	0	II
7LCW	GSKYSDTADESSYRW	0	II
4NAG	GGPLAGEEIGGFNVPG	0	II
7BW5	GPKGDFPDVGDGRILAG	0	II
7ZWJ	SKKSKPGDGIRGKGVRG	0	II
5XM4	GPPGDRIEFGVLAQLPG	0	II
6XTI	GGGGRGYEYNKQCLIFC	2	IV
5JPL	GRPNWGFENDWSCVRVC	2	IV
5OQZ	GAPSLINSEDNPAFPQRV	0	II
6XTH	GSRGWGFEPGVRCLIWCD	2	IV
5GVO	GLPIGWWIERPSGWYFPI	0	II
5UI6	GGKGPIFETWVTEGNYYG	0	II
2MLJ	SIGDSGLRESMSSQTYWP	0	II
5TJ1	GVGFGRPDSILTQEQAKPM	0	II
5UI7	GSDGPIIEFFNPNGVMHYG	0	II
5D9E	GTLTPGLPEDFLPGHYMPG	0	II
3NJW	GLPWGCPSDIPGWNTPWAC	2	III
6B5W	GVGFGRPDSILTQEQAKPM	0	II
2LX6	GAFVGQPEAVNPLGREIQG	0	II
6MW6	GGVGKIIEYFIGGGVGRYG	0	II
2M37	GLSQGVEPDIGQTYFEESR	0	II
2N6U	GLTQIQALDSVSGQFRDQLG	0	II
7EES	GTIDPQNSEEHPVLSRRLEN	0	II
2LS1	CVWGGDCTDFLGCGTAWICV	4	I
5ZCN	DGMGEEFIEGLVRDSLYPPAG	0	II
2MMT	GGAGHVPEYFVRGDFPISFYG	0	II
2MMW	GGAGHVPEYFVRGDTPISFYG	0	II
1Q71	GGAGHVPEYFVGIGTPISFYG	0	II
5JQF	GIEPLGPVDEDQGEHYLFAGG	0	II
2MW3	SLGSSPYNDILGYPALIVIYP	0	II
6AK0	CLGVGSCVDFAGCGYAVVCFW	4	I
1RPB	CLGIGSCNDFAGCGYAVVCFW	4	I
2LTI	GLSQGVEPDIGQTYFEESRINQD	0	II
2N68	GLSQGVEPDIGQTYFEESRINQD	0	II
8DYN	GHSVDRIPEYFGPPGLPGPVLFYS	0	II
2N6V	GPTPMVGLDSVSGQYWDQHAPLAD	0	II
6POR	GGDGSIAEYFNRPMHIHDWQIMDSGYYG	0	II
8SVB	GGGGPTPEYFLMPIDPAWLQANLPNTGKYN	0	II
6Q1X	GVLGNDAEGITLLPLCFKPICIPTLPPLTGGHA	2	IV

References

1. Schwede, T., et al., *SWISS-MODEL: an automated protein homology-modeling server.* Nucleic acids research, 2003. **31**(13): p. 3381-3385.
2. Case, D.A., et al., *Amber 2021*. 2021: University of California, San Francisco.
3. Ruby, U. and V. Yendapalli, *Binary cross entropy with deep learning technique for image classification.* Int. J. Adv. Trends Comput. Sci. Eng, 2020. **9**(10).
4. Katahira, R., et al., *Solution structure of endothelin B receptor selective antagonist RES-701-1 determined by 1H NMR spectroscopy.* Bioorganic & Medicinal Chemistry, 1995. **3**(9): p. 1273-1280.
5. Shihoya, W., et al., *Structure of a lasso peptide bound ETB receptor provides insights into the mechanism of GPCR inverse agonism.* bioRxiv, 2024: p. 2023.12.30.573741.
6. Knappe, T.A., et al., *Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of Burkholderia thailandensis E264.* Journal of the American Chemical Society, 2008. **130**(34): p. 11446-11454.
7. Iwatsuki, M., et al., *Lariatins, antimycobacterial peptides produced by Rhodococcus sp. K01− B0171, have a lasso structure.* Journal of the American Chemical Society, 2006. **128**(23): p. 7486-7491.
8. Koos, J.D. and A.J. Link, *Heterologous and in vitro reconstitution of fuscanodin, a lasso peptide from Thermobifida fusca.* Journal of the American Chemical Society, 2018. **141**(2): p. 928-935.
9. Cao, L., et al., *Cellulonodin-2 and Lihuanodin: Lasso peptides with an aspartimide post-translational modification.* Journal of the American Chemical Society, 2021. **143**(30): p. 11690-11702.
10. Bratovanov, E.V., et al., *Genome mining and heterologous expression reveal two distinct families of lasso peptides highly conserved in endofungal bacteria.* ACS Chemical Biology, 2019. **15**(5): p. 1169-1176.
11. Zimmermann, M., et al., *Characterization of caulonodin lasso peptides revealed unprecedented N-terminal residues and a precursor motif essential for peptide maturation.* Chemical Science, 2014. **5**(10): p. 4032-4043.
12. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
13. Su, J., et al., *Saprot: Protein language modeling with structure-aware vocabulary.* bioRxiv, 2023: p. 2023.10. 01.560349.
14. Liu, T., et al., *Rational generation of lasso peptides based on biosynthetic gene mutations and site-selective chemical modifications.* Chemical Science, 2021. **12**(37): p. 12353-12364.