Supplementary Information for InstructPLM: Aligning Protein Language Models to Follow Protein Structure Instructions

Jiezhong Qiu[1,2*†], Junde Xu[1,4†], Jie Hu[1†], Hanqun Cao[4†], Liya Hou[1], Zijun Gao[1], Xinyi Zhou[4], Anni Li[3], Xiujuan Li[3], Bin Cui[1], Fei Yang[1], Shuang Peng[1], Ning Sun[1], Fangyu Wang[1], Aimin Pan[1], Jie Tang[5], Jieping Ye[6], Junyang Lin[6], Jin Tang[1], Xingxu Huang[1], Pheng Ann Heng[1,4], Guangyong Chen[1*]

[1*]Zhejiang Lab.
[2]Zhejiang University.
[3]Nanjing Normal University.
[4]The Chinese University of Hong Kong.
[5]Tsinghua University.
[6]Alibaba.

*Corresponding author(s). E-mail(s): jiezhongqiu@zhejianglab.com; gychen@zhejianglab.com;
†These authors contributed equally to this work.

The PDF file includes:

Supplementary Table 1 to Supplementary Table 7 and Supplementary Figure 1 to Supplementary Figure 4.

Supplementary Results

Implementation Details

To ensure the reproducibility of our results and facilitate the understanding of our model's architecture, we provide a detailed description of the hyperparameter and training settings employed in InstructPLM.

Initialization

- **Linear Layers**: The weight parameters of the linear layers are initialized using a truncated normal distribution with a mean of 0 and a standard deviation of 0.02. The biases are initialized to zero.
- **Layer Normalization**: The weights and biases of the layer normalization modules are uniformly initialized to 1.0 and 0, respectively. This initialization scheme ensures a standardized starting point for the training process.

Hyperparameters

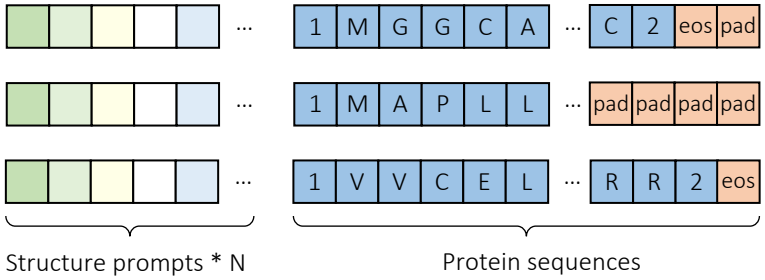
We show our model hyperparameters and training configuration in Supplementary Table 1.

Supplementary Table 1 Summary of model hyperparameters.

Hyperparameter	Value	Description
Model		
Max length	512	Max number of amino acids of input structures.
Learnable Queires	256	Number of learnable queries.
Number of Heads	16	Number of heads in multi-head cross attention.
Generation		
Temperature	0.15/0.8	Control the diversity of generated sequences.
Top-p	0.9	Control the diversity of generated sequences.
Training		
Learning Rate	0.0001	Learning rate for the optimizer.
Batch Size	16	Number of samples per step.
Data Parallel	4	Number of batches per step.
Epochs	200	Number of times to iterate over the entire dataset.
Warmup Steps	5000	Number of steps to gradually increase the learning rate at the beginning of training.
Weight Decay	0.1	Penalizing large weights.
Precision Training Mode	bfloat16	Utilizes bfloat16 for mixed precision training.
Optimizer	Adam	Optimization method for updating weights.
Loss Function	Cross-Entropy	Loss function for training the model.

Tokenizer

We use the same tokenizer as the original ProGen2. The vocabulary of the tokenizer comprises 30 tokens, including 20 amino acids, 5 special tokens, and 5 unused tokens. Among the 5 special tokens, we use '1' and '2' to represent the N-terminal and C-terminal sides of the protein sequence respectively, representing the end of the sequence by the end-of-sequence token ($\langle eos \rangle$). To handle sequences with different lengths, the padding token ($\langle pad \rangle$) will be added at the end of the sequences. Since the structure prompt is added at the beginning of the protein sequence to indicate the start of the sequence, thus we do not use the $\langle bos \rangle$ token. Fig. 1 shows an example of a constructed data batch.



Supplementary Fig. 1 Illustration of batch construction. We concat the structure prompt with protein sequence to train the InstructPLM, and special tokens '1', '2', ⟨*eos*⟩ and ⟨*pad*⟩.

Supplementary Table 2 Perplexity with/without special tokens.

Model	Perplexity on CATH 4.2↓		
	Short	Single-chain	All
InstructPLM-*full vocab.*	3.28	3.20	2.71
InstructPLM	3.22	3.17	2.68

Training and Evaluation

During the training phase, following [1–4] we first filter the CATH 4.2 dataset with sequence lengths less than 500. We exclude the loss calculation of the tokens where the corresponding structure is missing, as same as other special tokens: '1', ⟨*eos*⟩, and ⟨*pad*⟩. We keep the loss calculation of special token '2' to teach the model where to stop. We evaluate InstructPLM on the CATH 4.2 validation set at the end of each epoch, and perform early stopping based on validation LM-loss, which can drastically reduce the training cost. As the vocabulary of InstructPLM is larger than other baselines, for fair comparison, we only calculated the perplexity on 20 valid amino acids. The perplexity on the whole vocab is shown in Supplementary Table 2. Our methods are still optimal even with a larger vocabulary. To evaluate designed sequences, we truncate each sequence between '1' and '2' from the original output of InstructPLM resulting in a sequence with pure amino acid tokens.

More Results on TS50/TS500

We further test InstructPLM on the TS50 and TS500 datasets. As Supplementary Table 3 shows, our methods are better than the best baselines by large margins.

Designed Sequences for PETase and L-MDH

As for PETase, Supplementary Table 4 and Supplementary Table 5 list the 15 designed protein sequences. Supplementary Fig. 2 shows the sequence alignment results by ENDscript [9] of PETase generated by InstructPLM, Fast-PETase [10] and wild-type PETase. As for L-MDH, Supplementary Table 6 and Supplementary Table 7 list the 15 designed protein sequences. Supplementary Fig. 4 shows the sequence

Supplementary Table 3 Sequence design performance on TS50 and TS500. The metrics include perplexity (exponentiated categorical cross-entropy loss per residue) and recovery rate (percentage of correctly predicted residues). The best performance is **bold**, while the best baseline is indicated with an underline.

Model	Perplexity ↓		Recovery ↑	
	TS50	TS500	TS50	TS500
StructGNN [5]	5.40	4.98	43.89	45.69
GraphTrans [5]	5.60	5.16	42.20	44.66
GVP [6]	4.71	4.20	44.14	49.14
GCA [7]	5.09	4.72	47.02	47.74
AlphaDesign [8]	5.25	4.93	48.36	49.23
ProteinMPNN [3]	3.93	3.53	54.43	58.08
PiFold [1]	3.86	3.44	58.72	60.42
InstructPLM (Relative Gain)	**2.29** (-32.9%)	**2.42** (-26.45%)	**67.99** (+15.79%)	**64.22** (+6.28%)

alignment results by ENDscript [9] of active L-MDH generated by InstructPLM and wild-type L-MDH. An example of L-MDH designed by InstructPLM is illustrated in Supplementary Fig. 3.

References

[1] Gao, Z., Tan, C. & Li, S. Z. Pifold: Toward effective and efficient protein inverse folding (2022).

[2] Zheng, Z. et al. Structure-informed language models are protein designers. bioRxiv 2023–02 (2023).

[3] Dauparas, J. et al. Robust deep learning–based protein sequence design using proteinmpnn. Science **378**, 49–56 (2022).

[4] Hsu, C. et al. Learning inverse folding from millions of predicted structures. International Conference on Machine Learning 8946–8970 (2022).

[5] Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. Advances in neural information processing systems **32** (2019).

[6] Jing, B., Eismann, S., Suriana, P., Townshend, R. J. & Dror, R. Learning from protein structure with geometric vector perceptrons. arXiv preprint arXiv:2009.01411 (2020).

[7] Tan, C., Gao, Z., Xia, J., Hu, B. & Li, S. Z. Global-context aware generative protein design. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASS 1–5 (2023).

Supplementary Table 4 Designed PETase protein sequences (1-8), ✓indicates sequences where activity are detected; ○denotes sequences exhibiting activity superior to the wild-type.

Protein ID	Length	Activity	Sequence
PETase-1	261	✓	PETNPYAKGPTPTAASLEASAGQFTVQSFRVSKPIGFGAGTVYYPTDAGGRVGAIAIVPGYTQTQAA IRWWGPRLASHGFVVMTIDTNSTFDQPDSRASQLMAALDQVATLNSTSSSPIYGKVDTTRQGVMGHS MGGGGSLISAMNNPSLKAAIPLAPWHSSTNFSGVQVPTMIIGAQSDTVASVSSHSIPFYNSIPSTVPKAY LELLGGGHFCPNSPNTTIAKYAISWMKRFVDGDTRYDQFLCPAPSGTAISDYRTTCP
PETase-2	262	✓	SETNPYSRGPDPTRASLEASAGPFTVSSFSVSRPLGFGAGTVFYPTDAGGRVPAIAIAPGFTQTQSSVK WWGPRLASHGFVVVTIDTNSTFDQPDSRSAQLMAALDQVSRLNSQSSSPIYGKVDTTRQGVMGHSMG GGGSLISAQNNPDLKAAIPLAPWHTSTNFSSVQVPTLIIGAENDSIAPVSTHSIPFYTSLPSTLDKAYLEL AGASHFAPNSSNTTIAKYSISWLKRFIDDDTRYEQFLCPPPPSSAISDFRDTCPH
PETase-3	263	○	METNPFHRGPNPTAASLQASAGPFTVSSFRVSRPRGFGAGTVYYPTDAGGRVPAIAIVPGFTQRQSSI RWWGPMLASHGFVVMTIDTNSTFDQPDSRSDQLMAALNQLATLNSTASSPIYNKVDTTRQGVMGHS MGGGGSLISAQDNPNLKAALPMAPWNSSTNFSGVTVPTLIFACQADTIAPVSSHAQPFYNSLPSSTPKA YLELVGASHFAPNSSNTAIAKYAISWLKRFVDNDTRYEQFLCPPPPNSNLISEYRDTCPH
PETase-4	262	○	GETNPYTRGPDPTVASLEASAGPYTVQSFTVSRPRGFGAGEVFYPTDAGGQVPAIAIVPGFTQRQSSV MWWGPRLASHGFVVMTIDTNSTLDQPDSRSAQLMAALDQVASLNSTRSSPIYNKVDTTRQGVMGHS MGGGGSLIAAQDNPNLKASAPMAPWHISTNFSAVTVPTLIVGCEGDTIAPVASHSEPFYSSLSSGPEKA YLELNNASHFCPNSPNSTNGKYAISWLKRFVDNDTRYEQFLCPPPPNSAISDYRNSCPN
PETase-5	262	✓	AETNPYTRGPDPTTASLEASAGPFTVSSFSVSRPVGFGAGTVFYPTDAGGKVGAIAIVPGFTARQSSIK WWGPRLASHGFVVMTIDTNSVFDQPDSRSNQLMAALNQLATLNSTSSSPIYGKVDTTRQGVMGHSM GGGGALIAAQDNPSLKAAAPLTPWHSTVNFSNVQVPTFIIGAENDTVAPVSSHAIPFYTSIPSSIPKAYM ELKGASHFAPNSSNTTIAKYSIAWLKRFIDNDTRYDQFLCPPPSNSAISDYRNTCPY
PETase-6	262	○	RETNPYSRGPDPTTASLEASAGPFTVQSFSVSRPAGFGAGTVYYPTDAGGRVGAIAVVPGYTQSQSS VKWWGPRLASHGFVVMTIDTISTYDQPDSRANQLLAALDQVSKLNSTSSSPIYGKVDTTRQAVMGHS MGGGGSLIAAMDNPSLKAAIPMAPWHTSTDFSGVTVPTLIFGAQNDTVAPVSSHSQPFYNSIPSSTPK AYLELLGGSHFAPNSSNTAIAKYSISWLKRFVDNDTRYDQFLCPPPNGTAISDYRSTCPH
PETase-7	269	○	AETNPYSRGPDPTRASLEASAGPFTVSSFSVSRPLGFGAGQVYYPTDAGGKVGAVAIVPGYTQTQSA VKWWGPRLASHGFVVMTIDTQSTYDQPDSRATQLMAALNQIVTLNSTSSSPIYNKVDTTRQGVMGHS MGGGGALISAMNNPNLKAAIPMAPWHTSTNFSNVTVPTFIIACESDSVAPVSSHAEPFYNSLPASTPKA YLELVGGGHFCPNSPNTGNDSLGIMGRMTISWLKRFLDGDTRYSQFLCGAPANSDTISDYRSTCPF
PETase-8	269	○	METNPYVKGPDPTTASLEASAGPFTVQSFQVSNPRGFGAGQVYYPTDAGGRVPAIAIVPGFTQTQAS IQWWGPRLASHGFVVMTIDTNSPFDQPDSRSTQLMAALNQVANLNSTSSSPIYNKVDTSRQGVMGHS MGGGGSLIAAANNPSLKAALPLAPWHTSTNFSSVQVPTFIMGAQNDTIAPVSSHSIPFYNSLPASLPKA YMEIVGGGHLCPNSPNSGGRCTAGSIARYAISWLKRFLDGDTRYDQFLCPAPSGGATSDFRDTCTY

Supplementary Table 5 Designed PETase protein sequences (9-15), ✓indicates sequences where activity are detected; ○denotes sequences exhibiting activity superior to the wild-type.

Protein ID	Length	Activity	Sequence
PETase-9	262	○	METNPYHRGPDPTCASLEASAGPFTVRSFRVSSPLGFGAGTVYYPTDAGGRVPAIAIAPGFTQTQSSV KWWGPRLASHGFVVMTIDTNSTFDQPDSRSSQLMAALNQVSTLNSSSSSPIYGKVDTSRQGVMGHSM GGGGSLIAAQNNPSLKAAIPLAPWHSSTNFSAVTVPTLIFGAQNDTIAPVSSHSIPFYNSLPSSTEKAYLE LKGASHFAPNSPNSVIAKYSISWMKRFIDNDTRYEQFLCPSPSSSSISDYRNTCPN
PETase-10	263	✓	SETNPFSRGPDPTSASLEASAGPFTVSSFSVSRPAGFGAGTVYYPTDAGGRVGAIAIVPGYTGTQSSVK WWGPRLASHGFVVMTIDTESTFDQPDSRASQLMAALDQLAVLNSQRSSPIYNKVDTSRQGVMGHSM GGGGALISAMDNPSLKAAIPLAPWHASTNFSGVTVPTLIFGCESDSVAPVSSHASPFYNSLPASTDKAY LELTNGGHFCPNSPNTTIAKYGIAWMKRFVDNDTRYDQFLCPPPAGSPAISQYRDTCPN
PETase-11	262	○	GETNPFSRGPNPTSASLEASAGPFTVSSFSVSRPLGFGAGTVFYPTDAGGRVGAIAIVPGYTATQSSIR WWGPRLASHGFVVMTIDTISTFDQPDSRSSQLMAALDQLANLNSTRSSPIYGKVDTTRQGVMGHSMG GGGSLISAQDNPNLKAAIPLTPWHASTNFSNVQVPTLIIGAENDTIAPVASHAEPFYNSLPSSLDKAYLE LNNASHFAPNSPNTTIAKYSISWLKRFIDNDTRYDQFLCPAPTSTAISDYRNTCPH
PETase-12	263	○	SETNPYHRGPNPTAASLEASAGAFTVSSFSVSRPAGFGAGQVYYPTDAGGRVPAVAIVPGFTQTQSSV KWWGPRLASHGFVVITIDTNSTFDQPDSRSSQLMAALNQVATLNSTASSPIYGKVDTSRQGVMGHSM GGGGSLISAMNNPGLKAAAPMAPWHVSTNFSSVTVPTFVIGAQNDTIAPVSSHSIPFYNSLPSSLPKAYI EIAGGSHFCPNSPNTTIAKYAISWLKRFVDNDTRYSQFLCPSPSDSAISSYRSTCPNS
PETase-13	262	○	METNPFHRGPNPTAASLEASAGPFTVQSFRVSRPLGFGAGTVYYPTDAGGRVPAIAIVPGYTQNQSS VKWWGPRLASHGFVVMTIDTSSTFDQPDSRSAQLMAALNQIATLNSSASSPIYNKVDTSRQGVMGHS MGGGGSLISAMNNPNLKAAIPMAPWHISTNFSSVTVPTLIIGAQADTVAPVSSHAIPFYNSIPASTPKAY LELAGASHFFPNTTNTPTFAQMIAIWFKRFLDADQRYSQFACPPPNASTLSSFRNSCP
PETase-14	264	○	AETNPYHRGPNPTSASLEASAGPFTVSSFSVSRPVGFGAGQVYYPTDAGGRVPAVAIVPGFTQNQGS VRWWGPMLASHGFVVMTIDTNSTFDQPDSRSAQLMAALDQVSTLNSSTSSPIYGKVDTARQGVMGH SQGGGGALISAQNNPGLKAAAPLAPWHVSTNFSGVTVPTFVIACQNDTIAPVSTHALPFYNSLPASTP KAYLELTGGGHYCPNSPNTGLERYGIAWMKRFLDNDTRYDQFLCPPNSGSSSDISRWSNRCPL
PETase-15	263	○	GETNPYTRGPNPTTASLEASAGPFTVQSFSVSRPLGFGAGTVYYPTDAGGRVGAIAIVPGYTQTQSSI RWWGPRLASHGFVVMTIDTISTFDQPDSRSNQLMAALDQLANLNSQRSSPIYNKVDTTRQGVMGHSM GGGGSLIAAANNPGLKAAIPLAPWHSSTNFSSVTVPTLIFGASNDTTAPVSTHSIPFYNSLPSSLPKAYM ELKGASHFAPNSTNSVIAKYSISWLKRFVDNDTRYDQFLCPPPSGTTLISDYRDNCPY

Supplementary Table 6 Designed L-MDH protein sequences (1-8), ✓indicates sequences where activity are detected; ○denotes sequences exhibiting activity superior to the wild-type.

Protein ID	Length	Activity	Sequence
L-MDH-1	327		MVKACVLGASGGIGVPLSLLLKQSPLVDDLALYDVVNTPGVAADLSHIDSVANVAGHLPDDDGLKKA LAGADVVVIPAGVPRKPGMTRDDLFNINASIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAA GVFDPKRLFGVTTLDVVRAQTFTQELTGKKDPTAVVIPVVGGHSGETIVPLFSLVPASFPADKLDALT KRVQYGGDEVVKAKDGAGSATLSMAYAGYRFAEKVIRALKGESGVVEPTFVYLPGVAGGEEIAKAT GVEYFSVPVELGPNGAEKAINPLEGVTAEEKKLLEVAIPGLKKNIEKGVAFVQNPPQ
L-MDH-2	325	✓	MVKACVLGASGGIGVPLSLLLKQSPLVDELSLYDIANTPGVAADLSHIDSVANVAGHLPKDDGLKKAL AGADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAAG VFDPKRLFGVTTLDVVRAQTFTAEFTGQKDPSKTVIPVIGGHSGETIVPLFSQAKPAVNIPADKLDAL VNRVQFGGDEVVKAKDGAGSATLSMAYAGFRFAESVLKALKGEKGIIEPTFVYLPGVEGGEAIKKAT GCDFFSVPIELGPEGAEKAINPLSNVTEQEKKLLEACVEGLKGNIAKGVDFVK
L-MDH-3	326		MVKACVLGASGGIGQPLALLLKNSPLVDDLSLYDVVNTPGVAADLSHISSVANVAGYLPKDDGLKKAL AGADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEFCPKAFILIISNPVNSTVPIAAEVLKAAG VFDPKRLFGVTTLDVVRAQTFTAEFTGQKDPAELVIPVVGGHSGETIVPLFSLAKPAVDIPKDKYDAL VNRVQFGGDEVVKAKDGAGSATLSMAYAGYRFAEKVIKAAKGESGIVEPTYVYLPGIPGGDAIAKET GCDFFSVPIELGPDGAKKAINVIEGATAEEKKLLEACIAGLKKNIEKGVAFVAK
L-MDH-4	329		AVKAAVLGAAGGIGQPLSLLLKNSPLVDELSLYDIVNTPGVAADLSHISSVANVKGHLPEDDGLKKALT GADVVVIPAGVPRKPGMTRDDLFNINASIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAQTFTAEFTGQKDPAKLKIPVVGGHSGETIVPLFSLAAPSVKIPADKLDALT NRVQFGGDEVVKAKDGAGSATLSMAFAGFRFAEAVLKAIKGETGIVEPTFVYLPGVAGGDAIAKAT GVDYFSVPIELGPNGAEKAINVVADVTEKEKALLAKAIEGLKGNIAKGVAFAHNPPQ
L-MDH-5	328		MVKACVLGASGGIGVPLSYLLKSSPLVDDLALYDIVNTPGVAADLSHISSVANVAGHLPADDGLAKAL AGADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAG VFDPKRLFGVTTLDVVRAQTFTQEFTGQKDPAKLKIPVVGGHSGETIVPLFSLGKPAVDIPADKLDAL TNRVQFGGDEVVKAKDGAGSATLSMAYAGFRFAEKVIKAVKGEAGVVEPSFVYLPGVAGGEAIAKA TGVDFFSVPVELGPNGAEKAINPLAEVTDKEKALLAKAIEGLKGNIAKGVEFANAPK
L-MDH-6	329		MVKVAVLGAAGGIGQPLSLLLKQSPLVDELALYDVVNTPGVAADLSHISTVAKVSGHLPEDDGLKKAL TGADVVIIPAGVPRKPGMTRDDLFKINAGIVRDLATGIAKYCPKAFVLVISNPVNSTVPIVAETLKKAG VFDPKRVFGVTTLDVVRASTFVSELTGDKQPAKLTVPVIGGHSGETIVPLISQAKPSVAIPADKLDALV NRVQFGGDEVVKAKDGAGSATLSMAYAGFRFAEKVIRAAKGESGVVEPTFVYLPGVEGGAEIQKAT GAEYFSVPVELGPNGAEKALNIVEGATEAEKKLLDAAIAGLKGNIAKGVEFAKNPPA
L-MDH-7	326		MVKACVLGASGGIGVPLSLLLKESPLIDELALYDIVNTPGVAADLSHIDSVAKVAGHLPADDGLKKALA GADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAQTFTQEFTGQKDPSKVTIPVVGGHSGETIVPLFSLATPAVTIPKEKYDALV NRVQFGGDEVVKAKDGAGSATLSMAYAGFRFAESVLRAIKGEKGIVEPTYVYLPGVPGGDAIAKETG CDFFSVPVELGPNGAEKAINPLAGATEAEKKLLEACIPDLKKNIEKGVAFAHA
L-MDH-8	329		MVKACVLGASGGIGVPLSYLLKMSPLVDDLALYDIVNTPGVAADLSHISSIAKIDGHLPKDDGLKKALA GADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAQTFTQEFTGQKDPSVITIPVVGGHSGETIVPLFSKVEGVSFSDEDIEKLTHR VQFGGDEVVKAKDGAGSATLSMAFAGYRFAEKVLKALKGEKGVVEPTFVYLPGVEGGEEIKKATGV DYFSVPIELGPNGAEKAVNVLSGVNEYEKKLLEACYAGLKGNIEKGVEFIKNPPPK

Supplementary Table 7 Designed L-MDH protein sequences (9-15), ✓indicates sequences where activity are detected; ○denotes sequences exhibiting activity superior to the wild-type.

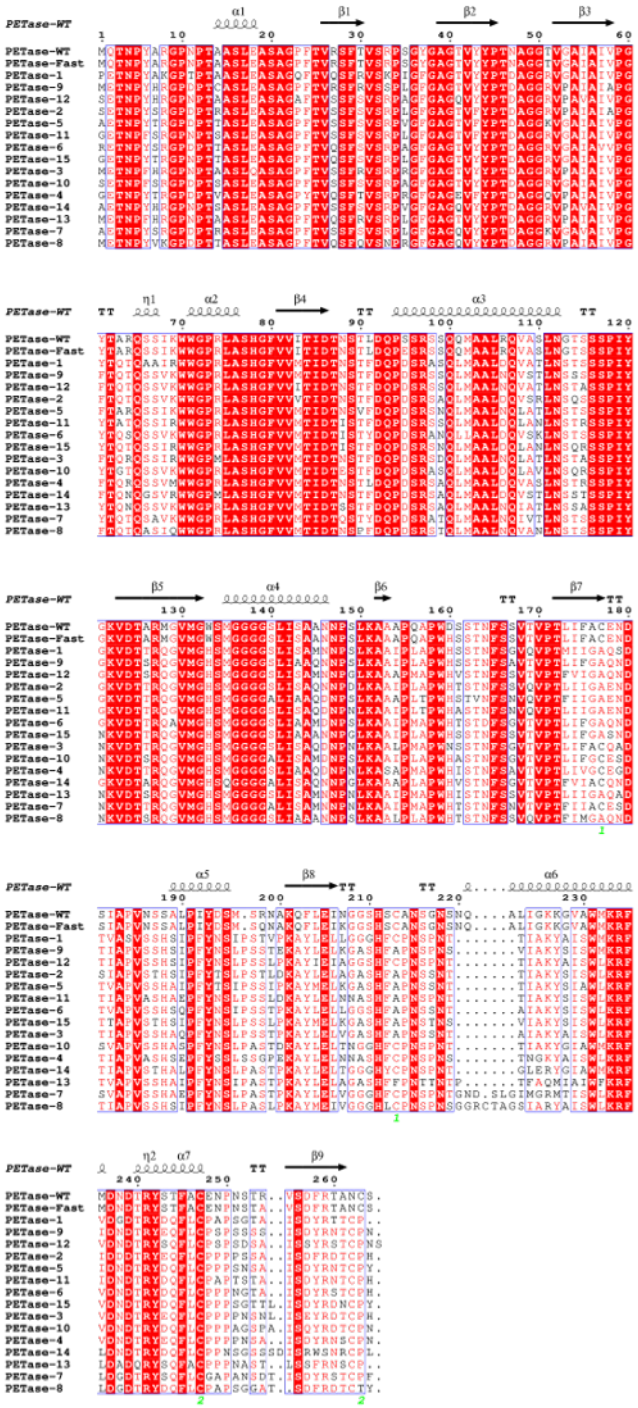
Protein ID	Length	Activity	Sequence
L-MDH-9	330		MVKACVLGASGGIGVPLSWLLKQSPLVDDLALYDVVNTPGVAADLSHINSVAKVAGHLPADDGLKKA LTGADVVLIPAGVPRKPGMTRDDLFNVNASIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAA GVFDPKRLFGVTTLDVVRAQTFTQEFTGQKDPSVVVIPVIGGHSGETIVPLFSKATAGVKLDPEKLDA LVNRVQFGGDEVVKAKDGAGSATLSMAYAGYRFAEKVIKAAKGESGIVEPSYVYLPGVEGGEAIKKA TGVDFFSVPIELGPEGAQKAINVVGNVTEAEKKLLAAAIEGLKGNIAKGVEFVKNPPPK
L-MDH-10	330		MVKVAVLGAAGGIGQPLSLLLKTSPLVDELSLYDVVNTPGVAADLSHISTVAKVSGHLPQDDGLKKAL TGADVVVIPAGVPRKPGMTRDDLFKINAGIVRDLAKGIAQHCPKAFVLVISNPVNSTVPIVAEVFKKA GVFDPKRLFGVTTLDVVRASTFVSEVTGEKDPAKFKIPVVGGHSGVTILPLLSQATPAVKFPKEKLDA LTNRIQFGGDEVVKAKDGAGSATLSMAYAGARFANSLLRAAAGEKGIVTPTFVYLPGVEGGEAISKA TGTDFFSTPVEFGPNGAEKAINVVNGITPEEKKLLEACYAGLKGNIEKGVEFVKNPPPA
L-MDH-11	329		MVKACVLGASGGIGQPLSLLLKQSPLVDELSLYDVVNTPGVAADLSHISSVANVAGHLPADDGLKKAL AGADVVVIPAGIPRKPGMTRDDLFNINASIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAQTFTAEITGQKDPASIVIPVIGGHSGETIVPLFSQATPAVNLPADKLDALTK RVQFGGDEVVKAKDGAGSATLSMAYAGFRFAEKVIKAAKGESGVVEPSFVYLPGVAGGEKIAAETG VDFFSVPIELGPNGAEKAINVVEGVNEKEKALLAAAVAGLKGNIAKGVAFVQNPPK
L-MDH-12	329		MVKACVLGASGGIGVPLSLLLKQSPLIDDLSLYDVVNTPGVAADLSHIDSIANVAGHLPADDGLKKALA GADVVVIPAGVPRKPGMTRDDLFNINASIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAQTFTAEFTGQKDPASVNIPVVGGHSGETIVPLFSLAKPSVDIPADKYDALV NRVQFGGDEVVKAKDGAGSATLSMAFAGFRFAEKVLKAIKGESGVVEPTFVYLPGVAGGEAIAKAT GVDFFSVPVELGAEGAAKAVNVLAEANDYEKKLLAKAYEGLQGNIAKGVEFIQNPPK
L-MDH-13	329	✓	MVKACVLGASGGIGQPLALLLKQSPLVDELALYDVVNTPGVAADLSHISSIANVAGHLPADDGLKKAL AGADVVVIPAGIPRKPGMTRDDLFNTNASIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAAG VFDPKRLFGVTTLDVVRAQTFTQEFTGQKDPAKTVIPVVGGHSGETIVPLFSLGKPAVKLDADKYDA LVNRVQFGGDEVVKAKDGAGSATLSMAYAGFRFAEKVIKAVKGEKGLVEPTYVYLPGVEGGEEIKK ATGCDFFSVPVELGPNGAEKAVNVVSAATDKEKALLAKAIEGLKTNIAKGIEFVKNPPA
L-MDH-14	328	✓	MVKAAVLGASGGIGVPLSLLLKQSPLIDELSLYDIVNTPGVAADLSHISSVANVKGFLPKDDGLKHALA GADVVVIPAGVPRKPGMTRDDLFNTNASIVRDLVKGVAEYAPKAFILIISNPVNSTVPIAAEVLKAAGV FDPKRLFGVTTLDVVRAETFTQEFTGQKDPATVNIPVIGGHSGETIVPLFSKATPSVNIPADKLDALT NRVQFGGDEVVKAKDGAGSATLSMAFAGFRFAEKVLRAIKGEKGIVEPSYVYLPGVAGGDAIKKAT GCDFFSVPIELGPNGAEKAINVVSEVSDKEKALLEAAVAGLKGNIEKGVAFVTSNP
L-MDH-15	329		MVKACVLGASGGIGVPLSWLLKQSPLVDDLALYDIVNTPGVAADLSHISSVAKVSGHLPADDGLKKAL TGADVVVIPAGVPRKPGMTRDDLFNINAGIVRDLVKGVAEYCPKAFILIISNPVNSTVPIAAEVLKAAG VFNPQRLFGVTTLDVVRAQTFTAELTGQKDPAAVVIPVIGGHSGETIVPLFSAAKPAVSIPQDKYDAL VNRVQFGGDEVVKAKDGAGSATLSMAFAGYRFAEKVLKAAAGEKGLIEPSFVYLPGVPGGEAIAKE TGVDFFSVPIELGPNGAEKAINVLSDVTDEEKALLKAAIEGLKGNIAKGVEFAHNPPQ

[8] Gao, Z., Tan, C. & Li, S. Z. Alphadesign: A graph protein design method and benchmark on alphafolddb. arXiv preprint arXiv:2202.01079 (2022).

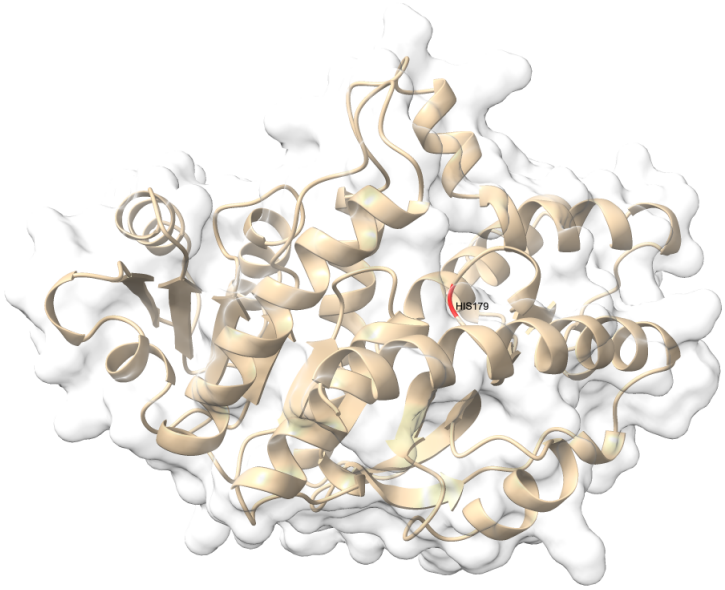
[9] Xavier, Robert, Patrice & Gouet. Deciphering key features in protein structures with the new endscript server. Nucleic acids research (2014).

[10] Lu, H. et al. Machine learning-aided engineering of hydrolases for pet depolymerization. Nature **604**, 662–667 (2022).

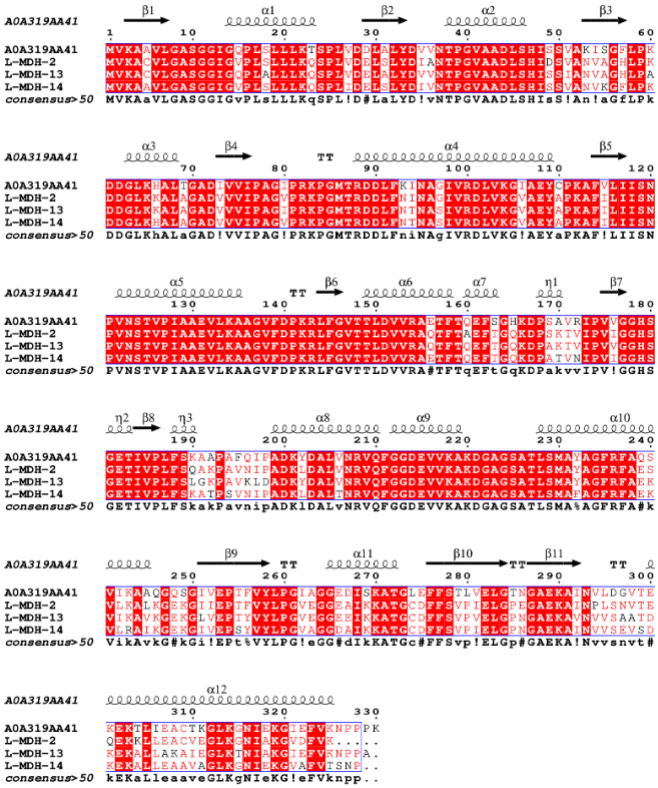
[11] Chelico, L., Pham, P., Petruska, J. & Goodman, M. F. Biochemical basis of immunological and retroviral responses to dna-targeted cytosine deamination by activation-induced cytidine deaminase and apobec3g. Journal of Biological Chemistry **284** (2009).



Supplementary Fig. 2 Alignment by ENDscript [9] of PETase generated by InstructPLM, Fast-PETase[10] and Wiled-type PETase. All of them (15/15) have the strictly conserved catalytic triad SER134-HIS212-ASP180. Not all sequences (4/15) have two intramolecular disulfide bridges (DS1 and DS2) and all of them (15/15) have DS2 connects the C-terminal helix and the last loop.



Supplementary Fig. 3 An example of L-MDH designed by InstructPLM, the protein structure is predicted by ESMFold. It has the same activate site HIS179 with Uniprot: A0A319AA41 which hit the UniRule PIRSR: PIRSR000102-1 [11].



Supplementary Fig. 4 Alignment by ENDscript [9] of L-MDH(Uniprot:A0A319AA41) and L-MDH which generate by InstructPLM and have detectable enzymatic activity. It has the same activate site HIS179 with Uniprot: A0A319AA41 which hit the UniRule PIRSR: PIRSR000102-1 [11].