Electronic Supplementary Information: Substrate Prediction for RiPP Biosynthetic Enzymes via Masked Language Modeling and Transfer Learning

Joseph D. Clark,[a] Xuenan Mi,[b] Douglas A. Mitchell,[c] and Diwakar Shukla *[bde]

[a]School of Molecular and Cellular Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801
[b]Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801
[c]Department of Chemistry, University of Illinois Urbana-Champaign, Urbana, IL, 61801.
[d]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801
[e]Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801.
Email: diwakar@illinois.edu

Table S1: Hyperparameter Grid for Downstream Substrate Prediction Model Optimization

Downstream model type	Hyperparameter	Values
LR	C	0.01, 0.1, 1, 5
RF	n_estimators	5, 25, 50, 100
AB	n_estimators	5, 25, 50, 100
SVC	C	0.01, 0.1, 1, 5
MLP	hidden_layer_sizes	50, 100, 200, 500

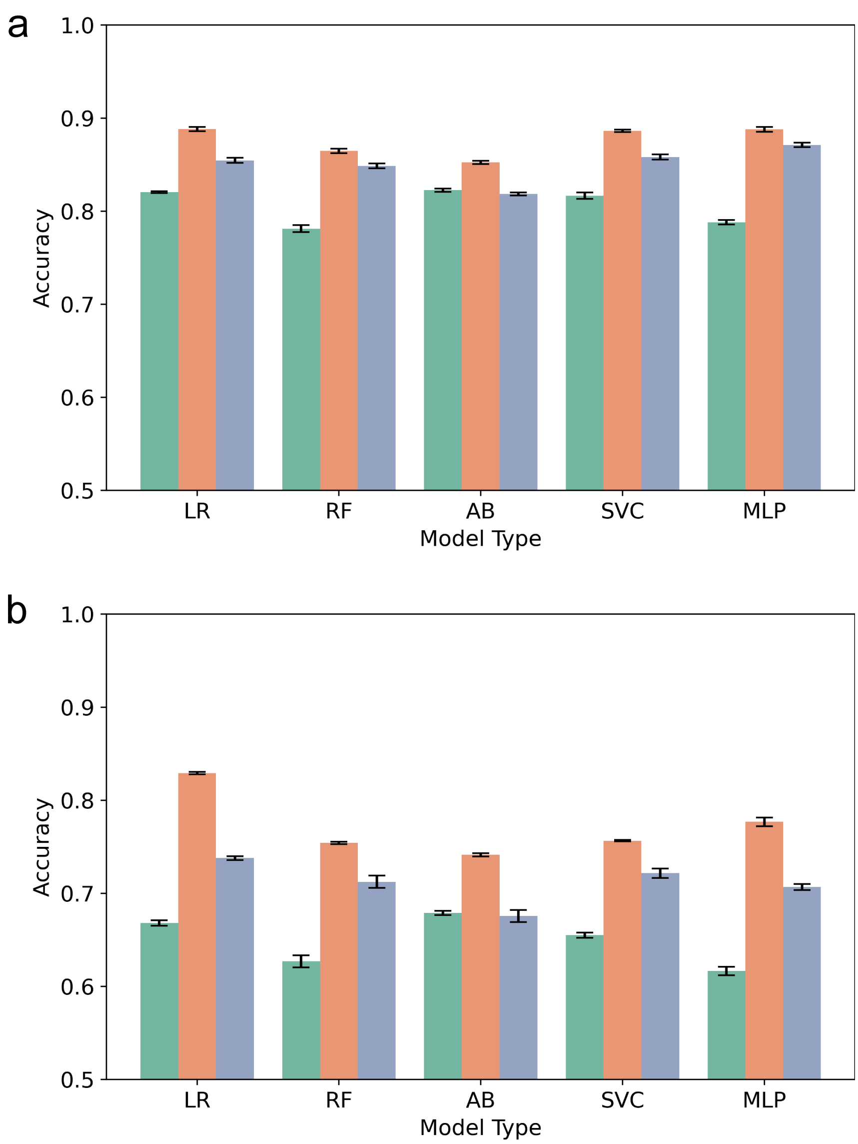

Figure S1: Accuracy of logistic regression (LR), random forest (RF), AdaBoost (AB), support vector classifier (SVC), and multi-layer perceptron (MLP) models trained using extended connectivity fingerprint (ECFP) representations of peptide sequences (green), embeddings from the protein language model ESM-2 (orange), and embeddings from the protein language model ProtBert (blue) for the a) LazBF substrate prediction task (n = 1,000) and b) the LazDEF substrate prediction task (n = 1,000). ESM-2 embeddings consistently outperform ECFP encodings and ProtBert embeddings.

Table S2: Downstream LazBF Substrate Prediction Model Hyperparameters

PLM, N	LR	RF	AB	SVC	MLP
Vanilla, High-N	C=0.1	n_estimators=50	n_estimators=100	C=5	hidden_layer_sizes=500
Peptide, High-N	C=1	n_estimators=50	n_estimators=50	C=1	hidden_layer_sizes=200
LazBF, High-N	C=0.1	n_estimators=25	n_estimators=100	C=0.1	hidden_layer_sizes=50
LazDEF, High-N	C=0.1	n_estimators=100	n_estimators=100	C=5	hidden_layer_sizes=100
LazBCDEF, High-N	C=0.1	n_estimators=25	n_estimators=100	C=5	hidden_layer_sizes=50
Vanilla, Med-N	C=0.1	n_estimators=50	n_estimators=100	C=1	hidden_layer_sizes=50
Peptide, Med-N	C=0.01	n_estimators=50	n_estimators=100	C=1	hidden_layer_sizes=50
LazBF, Med-N	C=1	n_estimators=50	n_estimators=100	C=1	hidden_layer_sizes=50
LazDEF, Med-N	C=0.1	n_estimators=100	n_estimators=100	C=5	hidden_layer_sizes=200
LazBCDEF, Med-N	C=0.01	n_estimators=100	n_estimators=50	C=5	hidden_layer_sizes=500
Vanilla, Low-N	C=0.1	n_estimators=100	n_estimators=100	C=1	hidden_layer_sizes=100
Peptide, Low-N	C=0.01	n_estimators=100	n_estimators=25	C=1	hidden_layer_sizes=500
LazBF, Low-N	C=0.01	n_estimators=50	n_estimators=50	C=1	hidden_layer_sizes=100
LazDEF, Low-N	C=0.01	n_estimators=50	n_estimators=50	C=0.01	hidden_layer_sizes=500
LazBCDEF, Low-N	C=5	n_estimators=100	n_estimators=25	C=5	hidden_layer_sizes=100

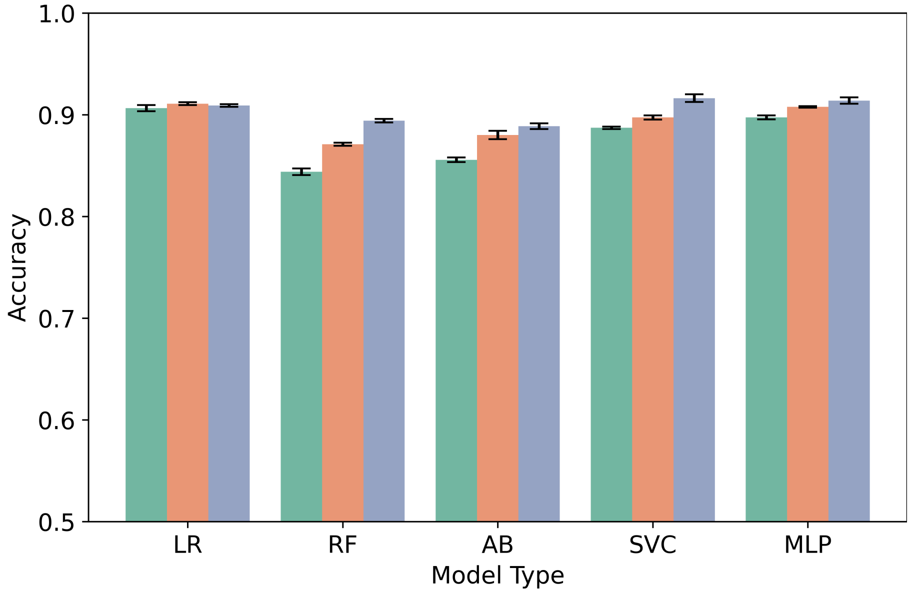

Figure S2: Accuracy of logistic regression (LR), random forest (RF), AdaBoost (AB), support vector classifier (SVC), and multi-layer perceptron (MLP) models trained using embeddings from a peptide language model trained on LazDEF substrates and non-substrates, with learning rates of 3e-4 (green), 3e-5 (orange), and 3e-6 (blue) for the LazBF substrate prediction task (n = 1,000). Embeddings from the model trained with a learning rate of 3e-6 outperform embeddings from models trained with higher learning rates.

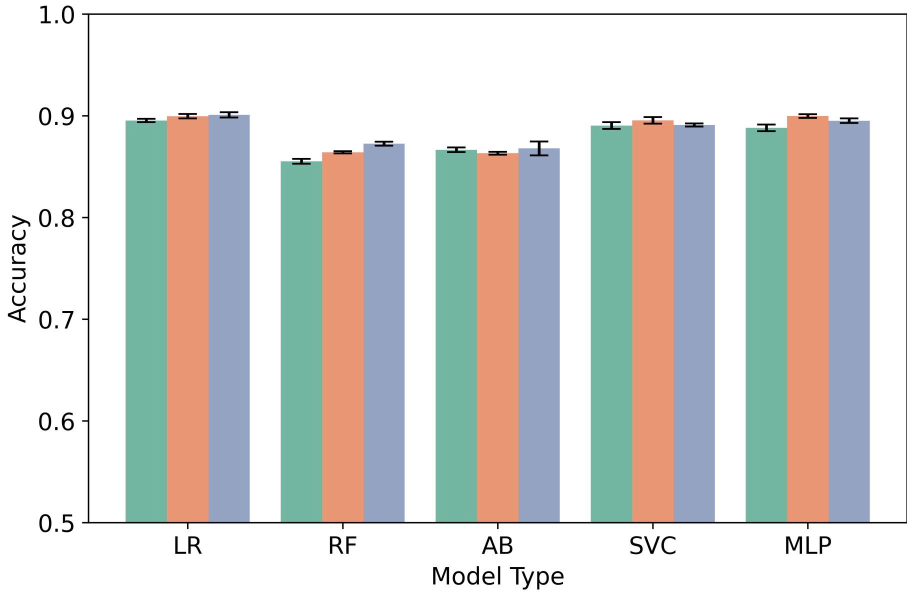

Figure S3: Accuracy of logistic regression (LR), random forest (RF), AdaBoost (AB), support vector classifier (SVC), and multi-layer perceptron (MLP) models trained using embeddings from a peptide language model trained on LazDEF substrates and non-substrates, with a batch size of 64 (green), 128 (orange), and 256 (blue) for the LazBF substrate prediction task (n = 1,000). Embeddings from the model trained with a batch size of 256 perform similar or better embeddings from models trained with lower batch sizes.

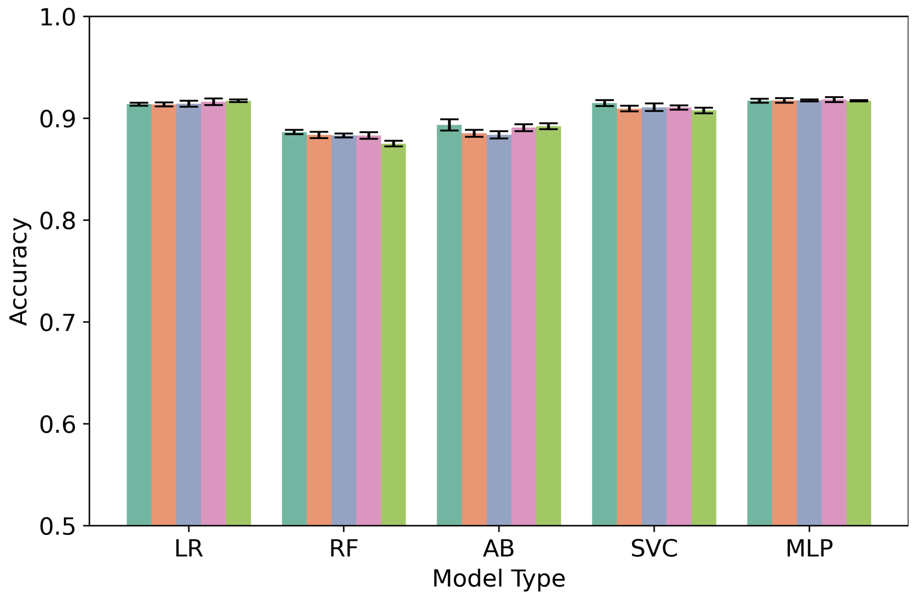

Figure S4: Accuracy of logistic regression (LR), random forest (RF), AdaBoost (AB), support vector classifier (SVC), and multi-layer perceptron (MLP) models trained using embeddings from a peptide language model trained on LazDEF substrates and non-substrates for one epoch (green), two epochs (orange), three epochs (blue), four epochs (pink), and five epochs (lime) for the LazBF substrate prediction task (n = 1,000). Embeddings from the models trained for more than one epoch did not increase the performance of LazBF substrate classifiers.

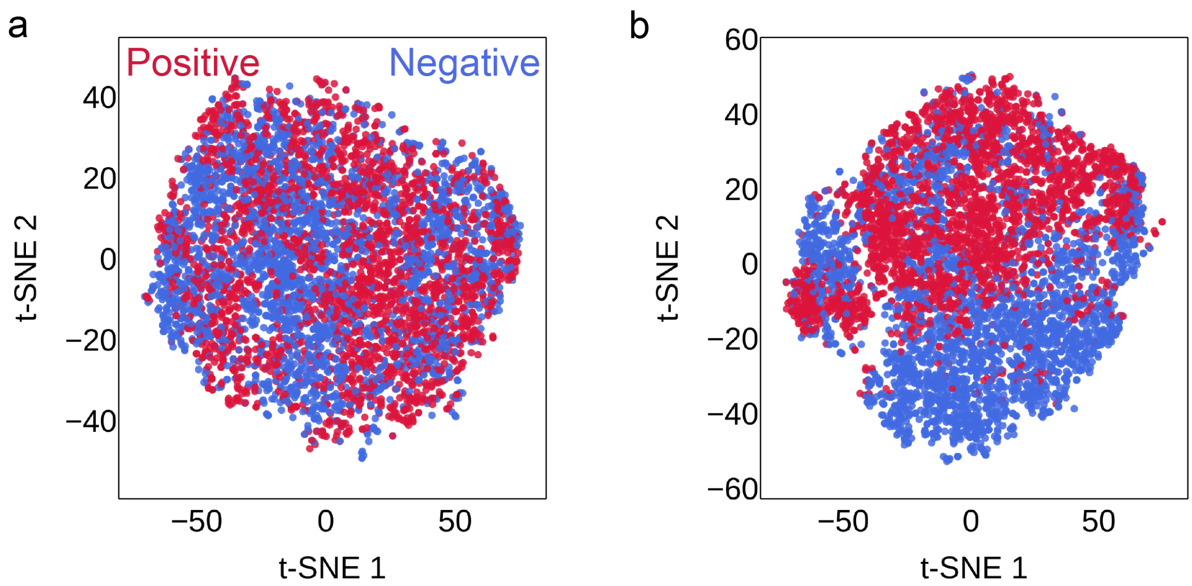

Figure S5: t-SNE visualization of the embedding space of ESM trained on non-LazA peptides a) LazDEF substrates/non-substrates, and b) LazBF substrates/non-substrates. Substrates are red and non-substrates samples are blue.

Table S3: Downstream LazDEF Substrate Prediction Model Hyperparameters

PLM, N	LR	RF	AB	SVC	MLP
Vanilla, High-N	C=0.1	n_estimators=100	n_estimators=100	C=1	hidden_layer_sizes=100
Peptide, High-N	C=5	n_estimators=50	n_estimators=100	C=5	hidden_layer_sizes=200
LazBF, High-N	C=1	n_estimators=50	n_estimators=100	C=5	hidden_layer_sizes=100
LazDEF, High-N	C=0.01	n_estimators100=	n_estimators=25	C=1	hidden_layer_sizes=50
LazBCDEF, High-N	C=0.1	n_estimators=100	n_estimators=100	C=5	hidden_layer_sizes=50
Vanilla, Med-N	C=0.1	n_estimators=100	n_estimators=50	C=1	hidden_layer_sizes=500
Peptide, Med-N	C=0.1	n_estimators=50	n_estimators=100	C=1	hidden_layer_sizes=50
LazBF, Med-N	C=0.01	n_estimators=100	n_estimators=100	C=1	hidden_layer_sizes=100
LazDEF, Med-N	C=0.01	n_estimators=100	n_estimators=100	C=0.1	hidden_layer_sizes=50
LazBCDEF, Med-N	C=0.1	n_estimators=100	n_estimators=100	C=5	hidden_layer_sizes=50
Vanilla, Low-N	C=0.01	n_estimators=100	n_estimators=100	C=1	hidden_layer_sizes=500
Peptide, Low-N	C=0.1	n_estimators=100	n_estimators=25	C=5	hidden_layer_sizes=500
LazBF, Low-N	C=0.01	n_estimators=100	n_estimators=25	C=1	hidden_layer_sizes=100
LazDEF, Low-N	C=0.01	n_estimators=25	n_estimators=25	C=0.1	hidden_layer_sizes=50
LazBCDEF, Low-N	C=0.1	n_estimators=25	n_estimators=50	C=5	hidden_layer_sizes=100

Table S4: Hyperparameters for Masked Language Modeling

Hyperparameter	Peptide-ESM	LazBF-ESM	LazDEF-ESM	LazBCDEF-ESM
Learning rate	3×10^{-6}	3×10^{-6}	3×10^{-6}	3×10^{-6}
Learning rate scheduler	Linear	Linear	Linear	Linear
Precision	fp16	fp16	fp16	fp16
Batch size	256	256	256	256
Weight decay	0.01	0.01	0.01	0.01
Training epochs	2	1	1	1
Adam Beta 1	0.9	0.9	0.9	0.9
Adam Beta 2	0.999	0.999	0.999	0.999
Adam Epsilon	1×10^{-8}	1×10^{-8}	1×10^{-8}	1×10^{-8}

Table S5: Hyperparameters for Fine-Tuned Models with 35M and 650M Parameters

Hyperparameter	650M parameters	35M parameters
Learning rate	2×10^{-4}	2×10^{-4}
Learning rate scheduler	Linear	Linear
Precision	fp16	fp16
Batch size	256	128
Weight decay	0.01	0.01
Dropout probability	0.1	0.1
Gradient accumulation steps	2	2
Training epochs	1	1
Adam Beta 1	0.9	0.9
Adam Beta 2	0.999	0.999
Adam Epsilon	1×10^{-8}	1×10^{-8}

Table S6: Classification Accuracy of Fine-Tuned Models with 35M and 650M Parameters

	LazBF test set	LazDEF test set	LazBCDEF test set
650M parameters	99.4%	99.2%	95.8%
35M parameters	99.4%	99.2%	95.8%

Table S7: Classification accuracy of fine-tuned LazBF substrate prediction models with different dropout probabilities

Dropout probability	LazBF test set	LazDEF test set	LazBCDEF test set
0.1	99.34%	50.93%	52.42%
0.2	99.37%	50.86%	52.37%
0.3	99.36%	51.00%	52.39%
0.4	99.35%	50.94%	52.34%
0.5	99378%	50.98%	52.48%

Table S8: Zero-shot Classification Accuracy of Downstream Models

	LazBF test set	LazDEF test set
LazBF SVC (Low-N)	-	54.2%
LazBF SVC (Med-N)	-	58.3%
LazBF SVC (High-N)	-	54.7%
LazDEF SVC (Low-N)	70.2%	-
LazDEF SVC (Med-N)	72.1%	-
LazDEF SVC (High-N)	70.5%	-

Table S9: Classification accuracy of fine-tuned LazBF substrate prediction model across 10 epochs of training

Epoch	LazBF	LazDEF	LazBCDEF
1	99.2%	50.9%	52.5%
2	99.3%	51.2%	52.4%
3	99.4%	51.0%	52.3%
4	99.4%	50.9%	52.4%
5	99.4%	51.0%	52.6%
6	99.4%	51.0%	52.5%
7	99.4%	51.2%	52.6%
8	99.4%	51.1%	52.5%
9	99.4%	51.1%	52.6%
10	99.4%	51.0%	52.3%

Table S10: Classification accuracy of fine-tuned LazDEF substrate prediction model across 10 epochs of training

Epoch	LazBF	LazDEF	LazBCDEF
1	69.8%	99.0%	63.2%
2	71.4%	99.0%	62.7%
3	70.4%	99.1%	63.6%
4	70.2%	99.0%	61.2%
5	71.1%	99.0%	62.1%
6	68.6%	99.1%	62.3%
7	69.7%	99.0%	61.8%
8	69.3%	99.0%	61.5%
9	69.5%	99.1%	61.2%
10	68.7%	99.1%	61.8%

Table S11: Classification accuracy of fine-tuned LazBCDEF substrate prediction model across 10 epochs of training

Epoch	LazBF	LazDEF	LazBCDEF
1	64.7%	59.1%	95.2%
2	64.7%	58.9%	95.5%
3	65.4%	59.1%	95.6%
4	62.5%	59.5%	95.5%
5	61.3%	58.0%	95.3%
6	62.7%	58.3%	94.9%
7	62.3%	58.0%	95.1%
8	62.2%	58.9%	94.8%
9	62.9%	58.4%	94.8%
10	62.8%	58.6%	94.9%

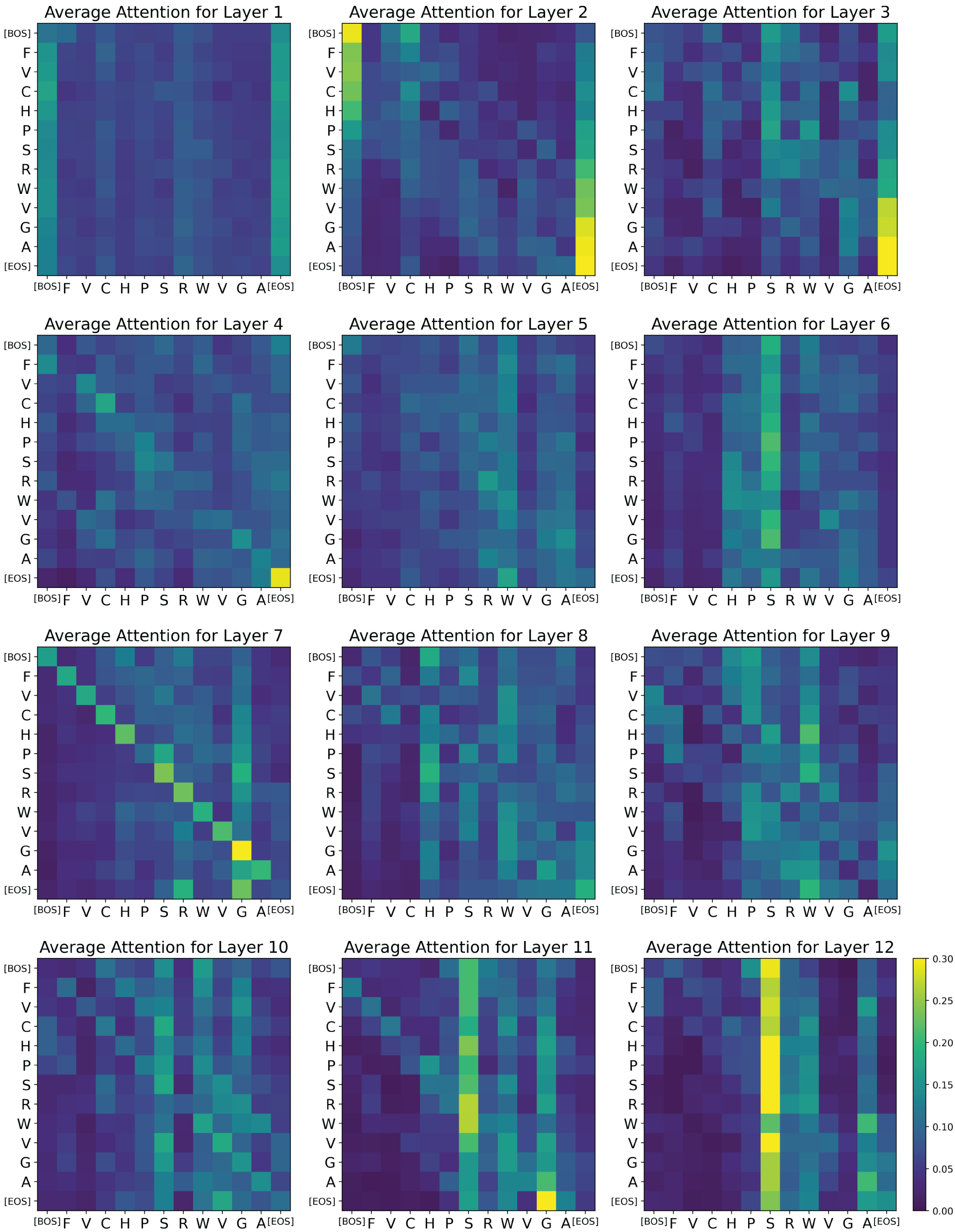

Figure S6: The average attention for all 12 layers of the fine-tuned LazBF-ESM for the LazBF substrate FVCHPSR-WVGA.

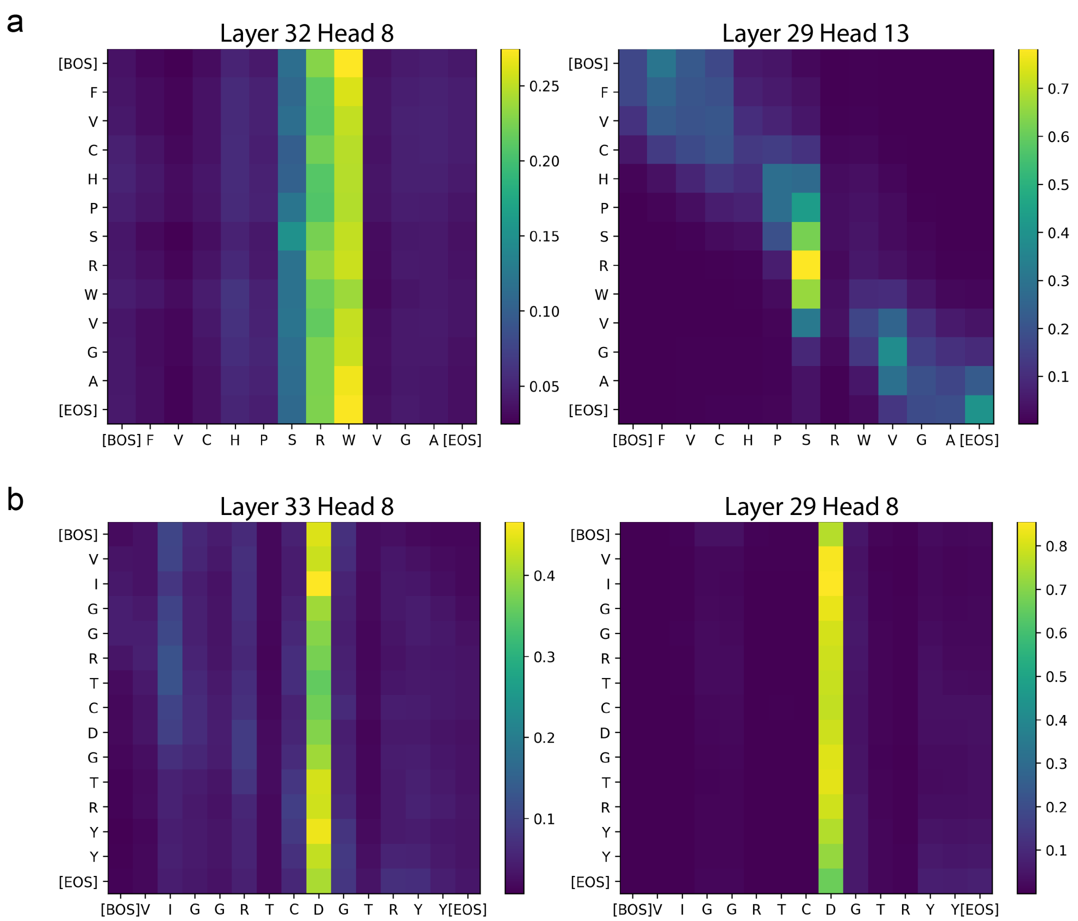

Figure S7: Attention maps from the fine-tuned LazBF-ESM with 650M parameters. [BOS] and [EOS] tokens mark the "beginning of sequence" and "end of sequence" respectively. a) Attention heads from the later layers highlight a motif with high pairwise epi-scores in a LazBF substrate. c) Attention heads from the later layers highlight a residue important for substrate fitness in a LazDEF substrate.