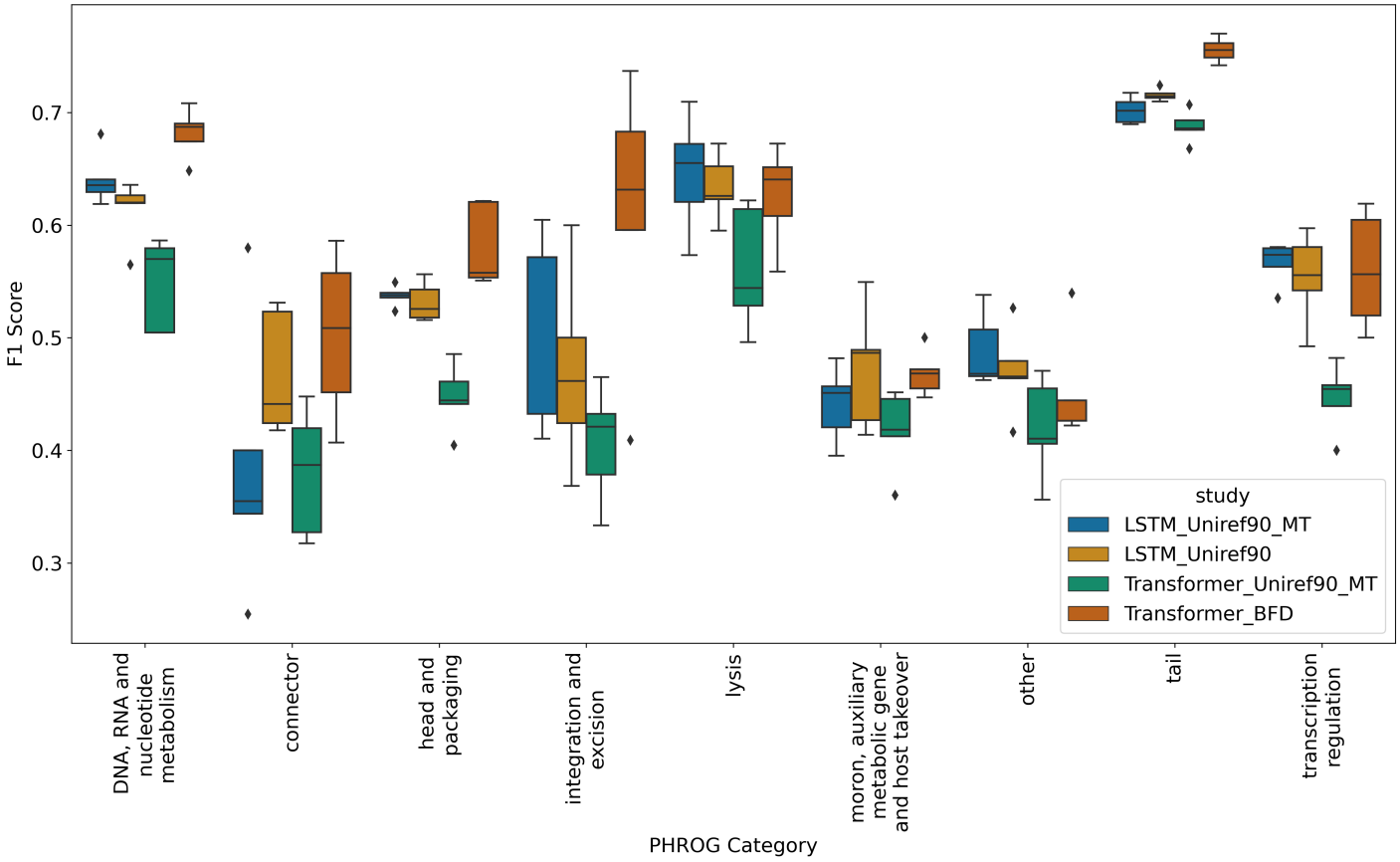
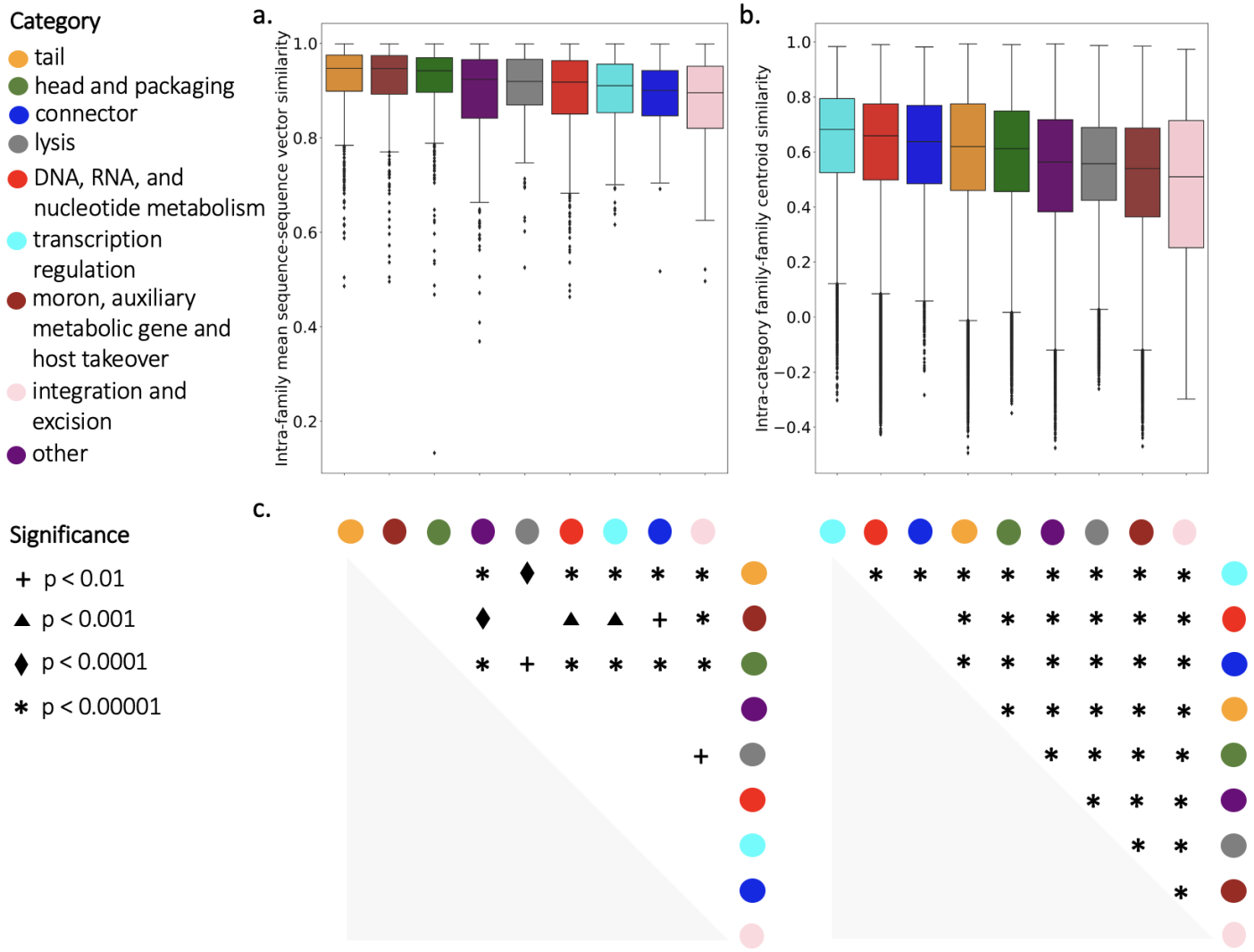
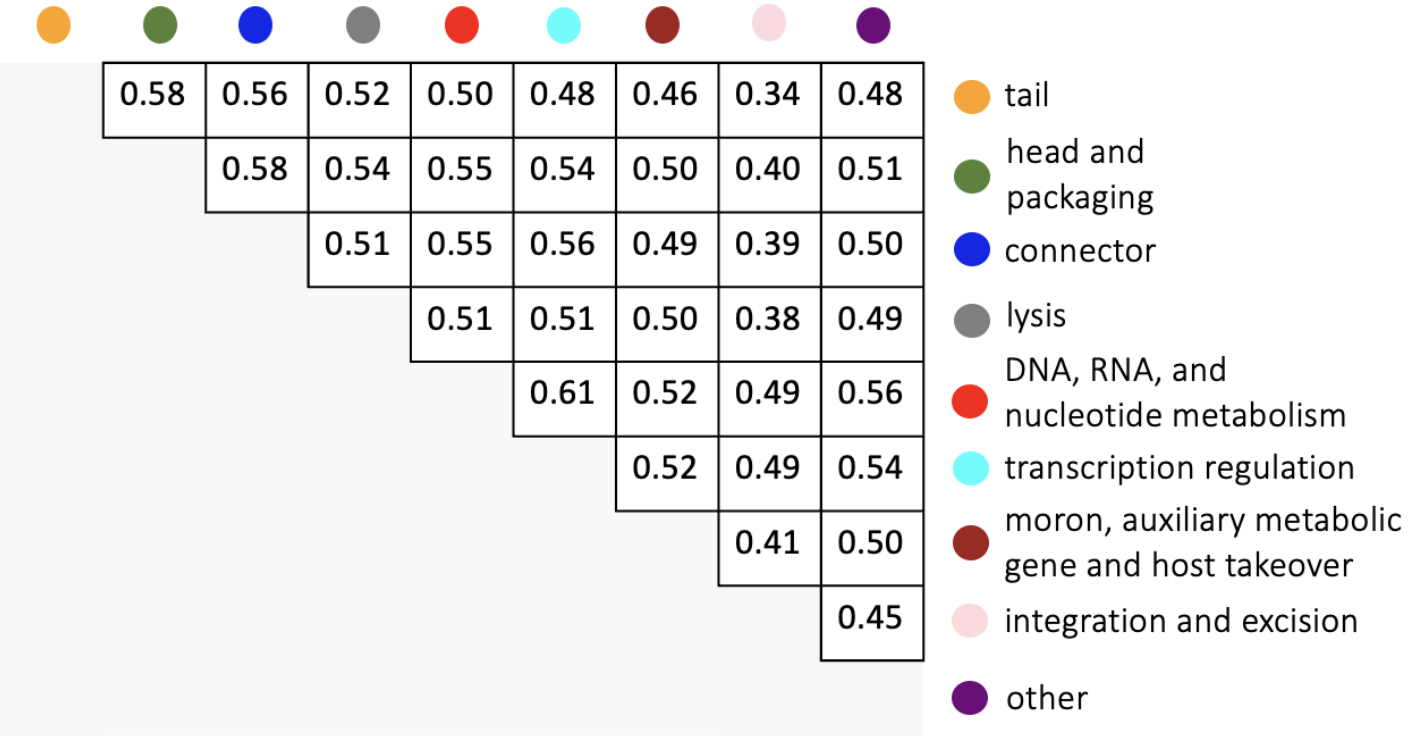


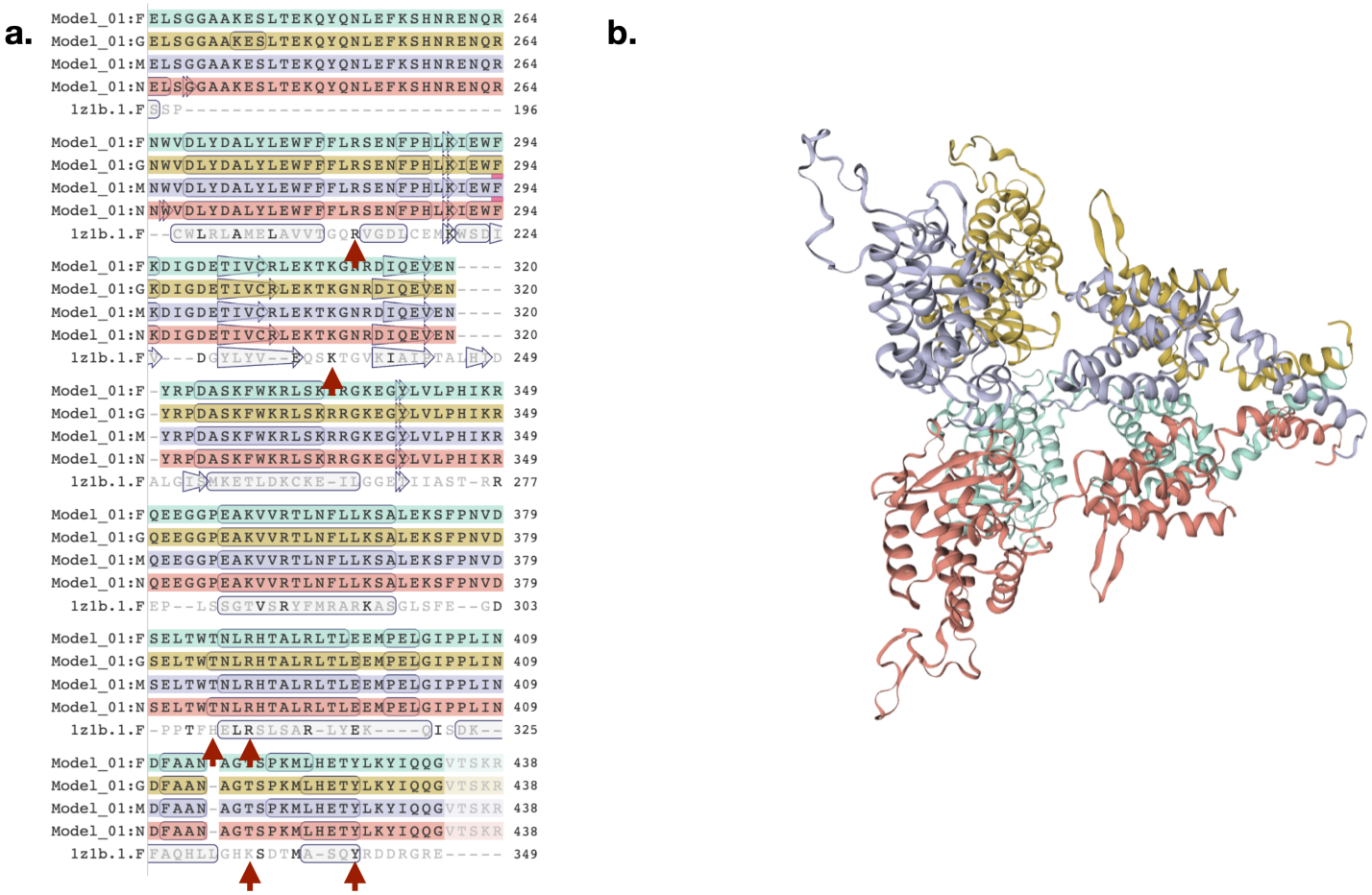
Supplemental Figure 1: Performance of four different PLM-based representations for viral VPF functional classification. Embedded proteins were used to train and evaluate PHROGs functional annotation classification. Performance is across five-fold training-testing splits of PHROGs VPFs. Study is described by the model architecture, protein source, and whether the PLM is trained with a multi-task training objective (MT). Performance is measured as F1-score.



Supplemental Figure 2: Evaluation of functional in embedding similarities of constituent families between functional categories. (a) Distribution of family average sequence-sequence similarity. (b) Distribution of family-family centroid similarity. (c) Significance of pairwise category distribution comparison using independent t-test with Bonferroni correction.



Supplemental Figure 3: Inter-category similarity for PHROGs functional categories. Pairwise family centroid similarities were calculated for every combination of families between the two categories. Score is the average over all comparisons.



Supplementary Figure 4: Comparative protein structure modelling of a novel integrase sequence supports annotation as a tyrosine recombinase. (a) Target/template alignment between the *Prochlorococcus* PAC1 sequence (indicated as Model_01), and the template sequence 1Z1B, the phage lambda integrase. Active site residues Arg 212, Lys 235, His 308, Arg 311, His 333, and Tyr 342, are indicated with red arrows. (b) Homology model of *Prochlorococcus* PAC1 sequence based on template 1Z1B. Colors indicate individual monomers of the homo-tetramer template protein structure in both a and b.

Category	Number of families	Total proteins
connector	133	34,054
DNA, RNA, and nucleotide metabolism	1,065	98,104
head and packaging	946	90,584
integration and excision	105	16,688
lysis	299	30,997
moron, auxiliary metabolic gene and host takeover	458	22,566
other	560	37,349
tail	1,219	102,955
transcription regulation	303	39,386
unknown function	33,792	395,697

Supplemental Table 1: Viral protein families (VPFs) and total number of proteins manually annotated to 10 function categories in the Prokaryotic virus Remote Homologous Groups (PHROGs) database.

Project	Model name	Model architecture	Protein source	Mutl-task training objective	Number of proteins	Training objective
ProSE[30]	MT-LSTM	LSTM	Unifref90	yes	76,215,872	Masked token + predicting contact between residues in protein structure + structural similarity of proteins by SCOP hierarchy
ProSE[30]	DLM-LSTM	LSTM	Unifref90	no	76,215,872	Masked token
ProteinBERT[29]	ProteinBERT	Transformer	Uniref90	yes	$\sim$106M	Masked token + GO term annotation
ProtTrans[28]	protbert_bfd	Transformer	BFD	no	2,122M	Masked token

Supplemental Table 2: PLMs evaluated for viral protein VPF functional classification. Four PLMs were used to embedd PHROGs VPF proteins for training and testing the viral function classifier. Model name is name from original study. PLMs were chosen to vary in the model architecture, protein source, and use of multiple training objectives.

Category	Support	Precision	Recall	F1-score
connector	2,145	0.61	0.62	0.62
DNA, RNA, and nucleotide metabolism	18,972	0.83	0.90	0.86
head and packaging	14,520	0.75	0.77	0.76
integration and excision	402	0.94	0.89	0.92
lysis	2,046	0.60	0.62	0.61
moron, auxiliary metabolic gene and host takeover	4,097	0.63	0.68	0.65
other	7,087	0.65	0.60	0.63
tail	11,060	0.61	0.90	0.72
transcriptional regulation	535	0.58	0.71	0.64
unknown function	20,078	0.71	0.45	0.55
weighted average	80,939	0.71	0.71	0.70

Supplemental Table 4: Functional classifier performance for EFAM VPFs labeled with PHROGs annotation. PHROGs functional annotation was assigned to EFAM VPFs using HMM matching. Support indicates how many EFAM VPFs matched PHROGs VPFs per category.