



Research article

Mutation prediction in the SARS-CoV-2 genome using attention-based neural machine translation

Darrak Moin Quddusi*, Sandesh Athni Hiremath and Naim Bajcinca

Chair of Mechatronics in the Faculty of Mechanical and Process Engineering, Rheinland-Pfalz Technical University of Kaiserslautern-Landau, Kaiserslautern 67663, Germany

* **Correspondence:** Email: darrak.quddusi@mv.uni-kl.de; Tel: +49631/205-3398;
Fax: +49631/205-4201.

Supplementary

Table S1. Dataset: tt_seq (total training sequences), t_seq (training sequences with duplicates removed), te_seq (total evaluation sequences), e_seq (evaluation sequences with duplicates removed).

Protein	tt_seq	t_seq	te_seq	e_seq
NSP1	130,903	864	330,212	2024
NSP3	102,071	13,347	170,659	23,999
NSP5	129,798	1002	315,672	2195
NSP8	132,472	577	349,002	852
NSP9	133,994	383	348,639	623
NSP13	127,019	2674	330,215	5921
NSP15	120,297	1232	319,678	2758

Table S2. Computational complexity of stacked biGRUs with attention mechanism for all seven targeted proteins.

Protein	n_{enc}	n_{dec}	big O	Complexity
NSP1	180	180	$O((2(180)(256^2)) + ((180)(180)(256)))$	55,480,320
NSP3	1945	1945	$O((2(1945)(256^2)) + ((1945)(1945)(256)))$	1,478,324,480
NSP5	306	306	$O((2(306)(256^2)) + ((306)(306)(256)))$	104,186,880
NSP8	198	198	$O((2(198)(256^2)) + ((198)(198)(256)))$	61,940,736
NSP9	113	113	$O((2(113)(256^2)) + ((113)(113)(256)))$	32,891,136
NSP13	601	601	$O((2(601)(256^2)) + ((601)(601)(256)))$	250,016,000
NSP15	346	346	$O((2(346)(256^2)) + ((346)(346)(256)))$	121,349,120

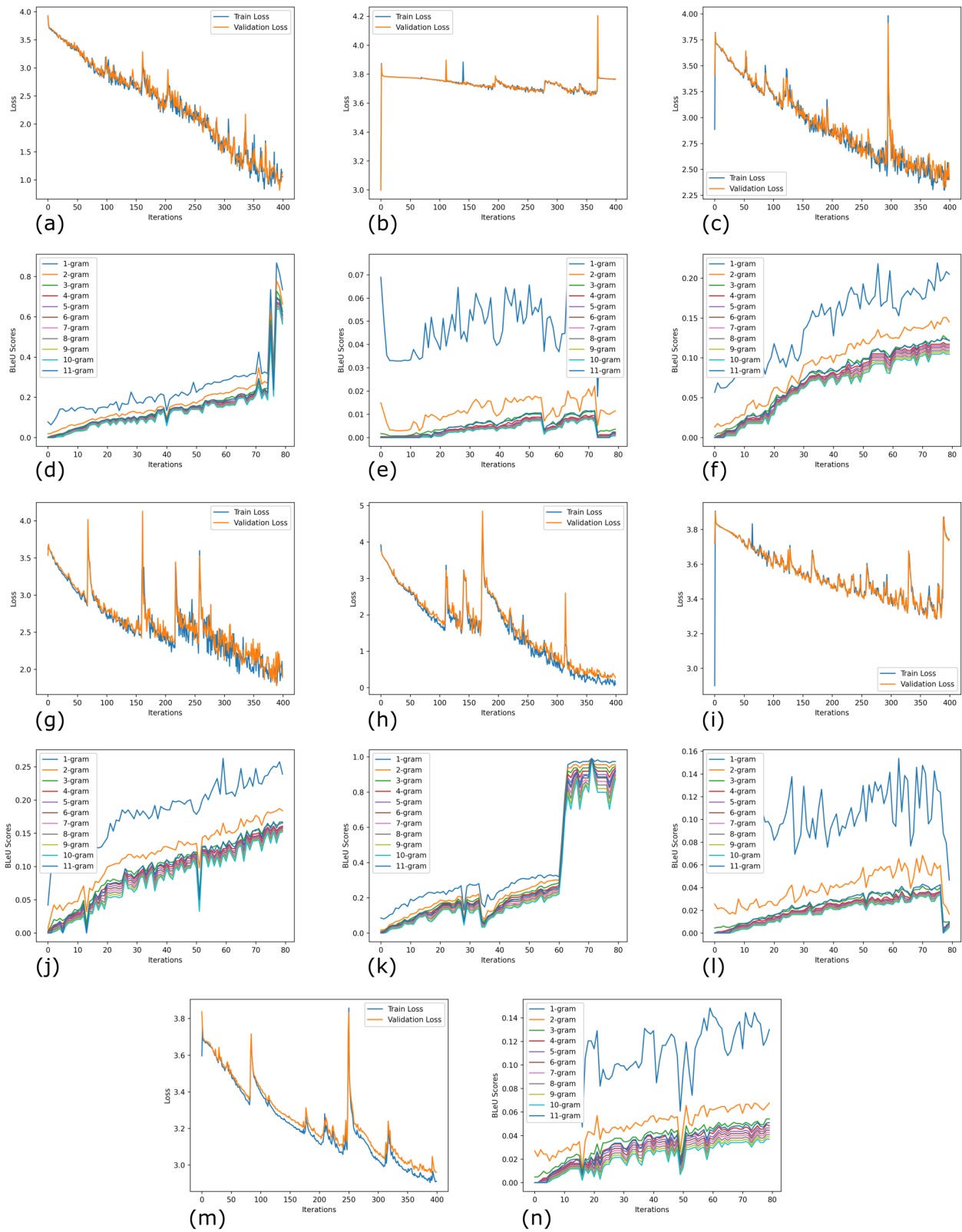


Figure S1. Vanilla RNNs. Training loss for (a) NSP1, (b) NSP3, (c) NSP5, (g) NSP8, (h) NSP9, (i) NSP13, and (m) NSP15. BLEU scores for (d) NSP1, (e) NSP3, (f) NSP5, (j) NSP8, (k) NSP9, (l) NSP13, and (n) NSP15.

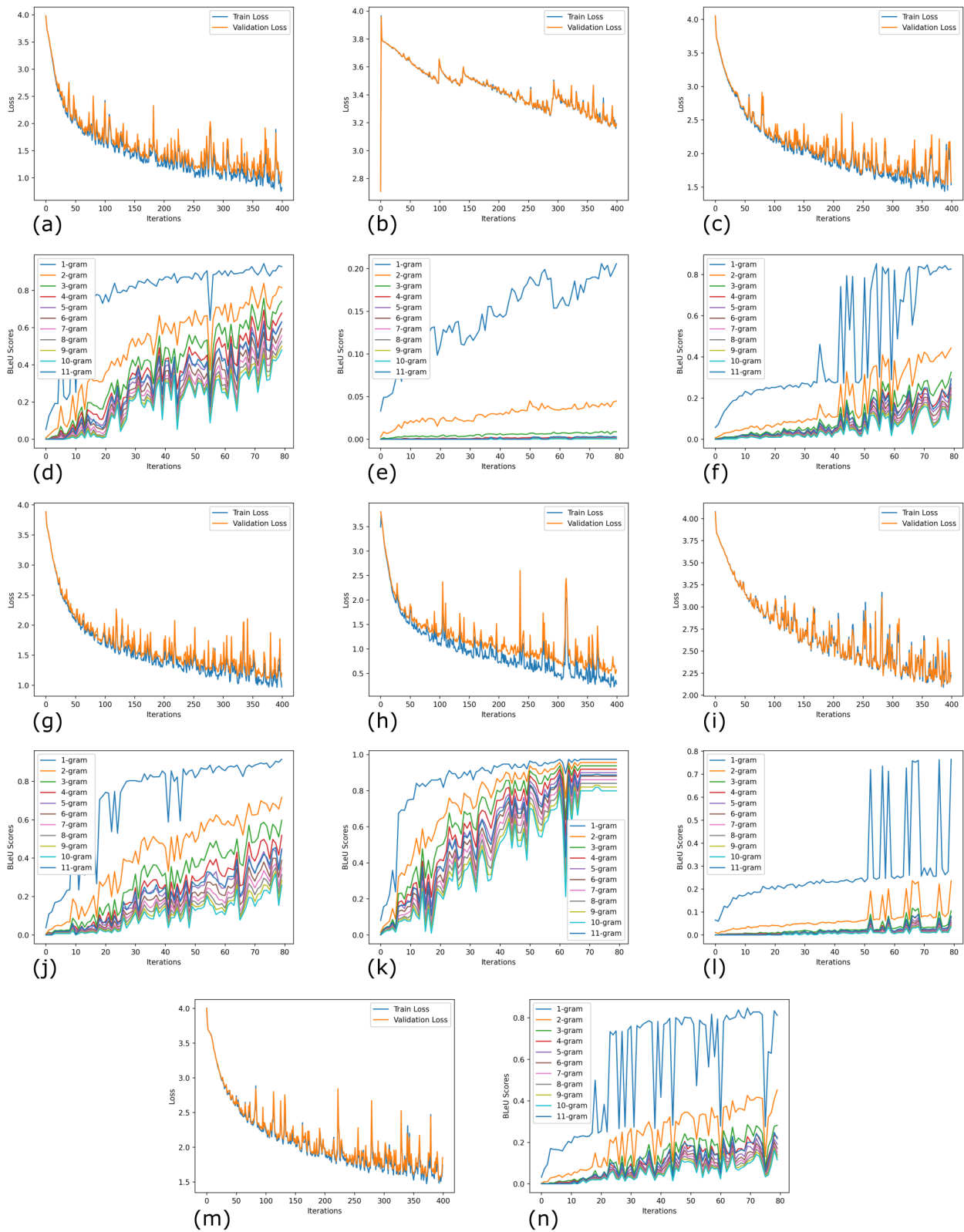


Figure S2. Simple LSTMs. Training loss for (a) NSP1, (b) NSP3, (c) NSP5, (g) NSP8, (h) NSP9, (i) NSP13, and (m) NSP15. BLEU scores for (d) NSP1, (e) NSP3, (f) NSP5, (j) NSP8, (k) NSP9, (l) NSP13, and (n) NSP15.

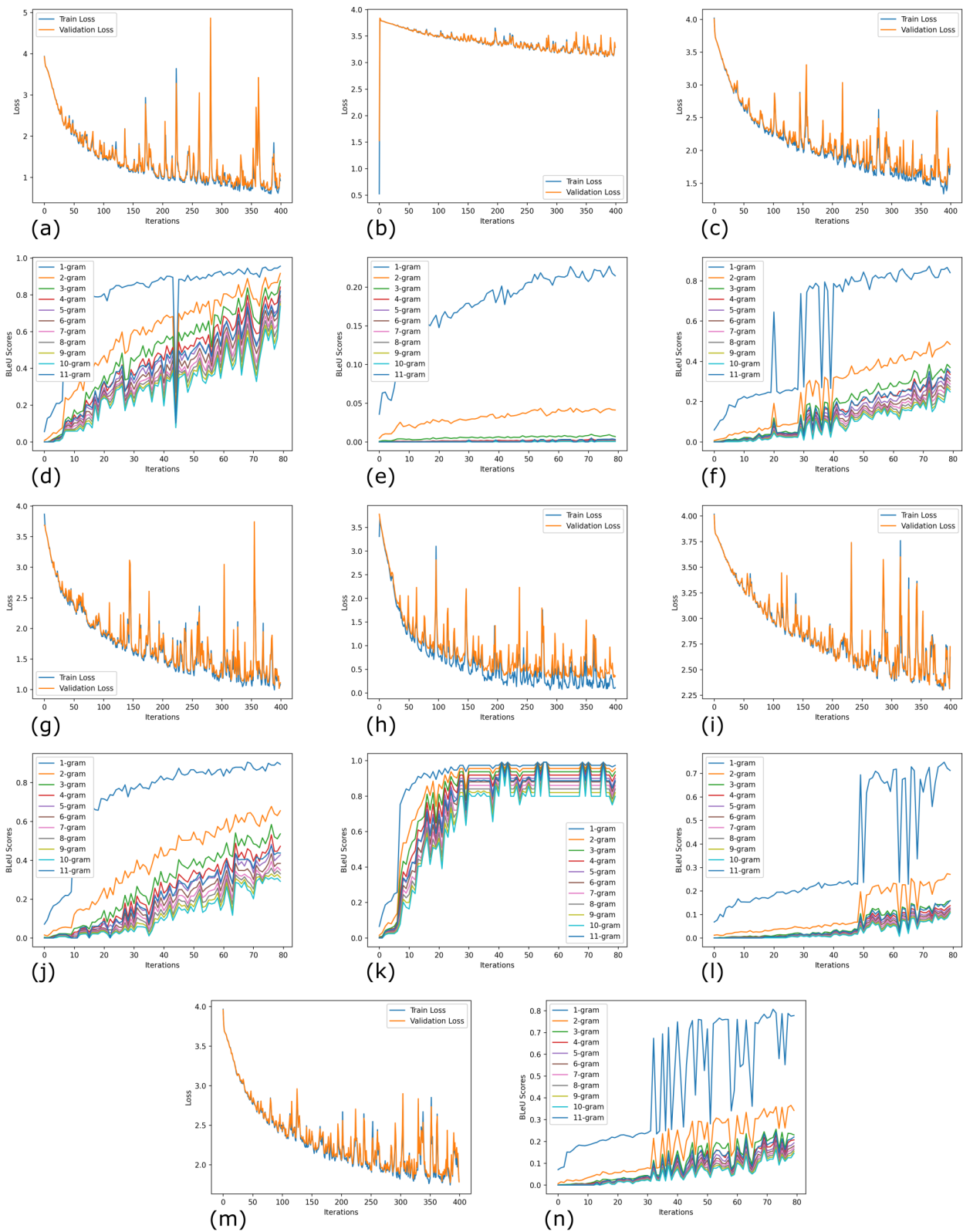


Figure S3. Simple GRUs. Training loss for (a) NSP1, (b) NSP3, (c) NSP5, (g) NSP8, (h) NSP9, (i) NSP13, and (m) NSP15. BLEU scores for (d) NSP1, (e) NSP3, (f) NSP5, (j) NSP8, (k) NSP9, (l) NSP13, and (n) NSP15.

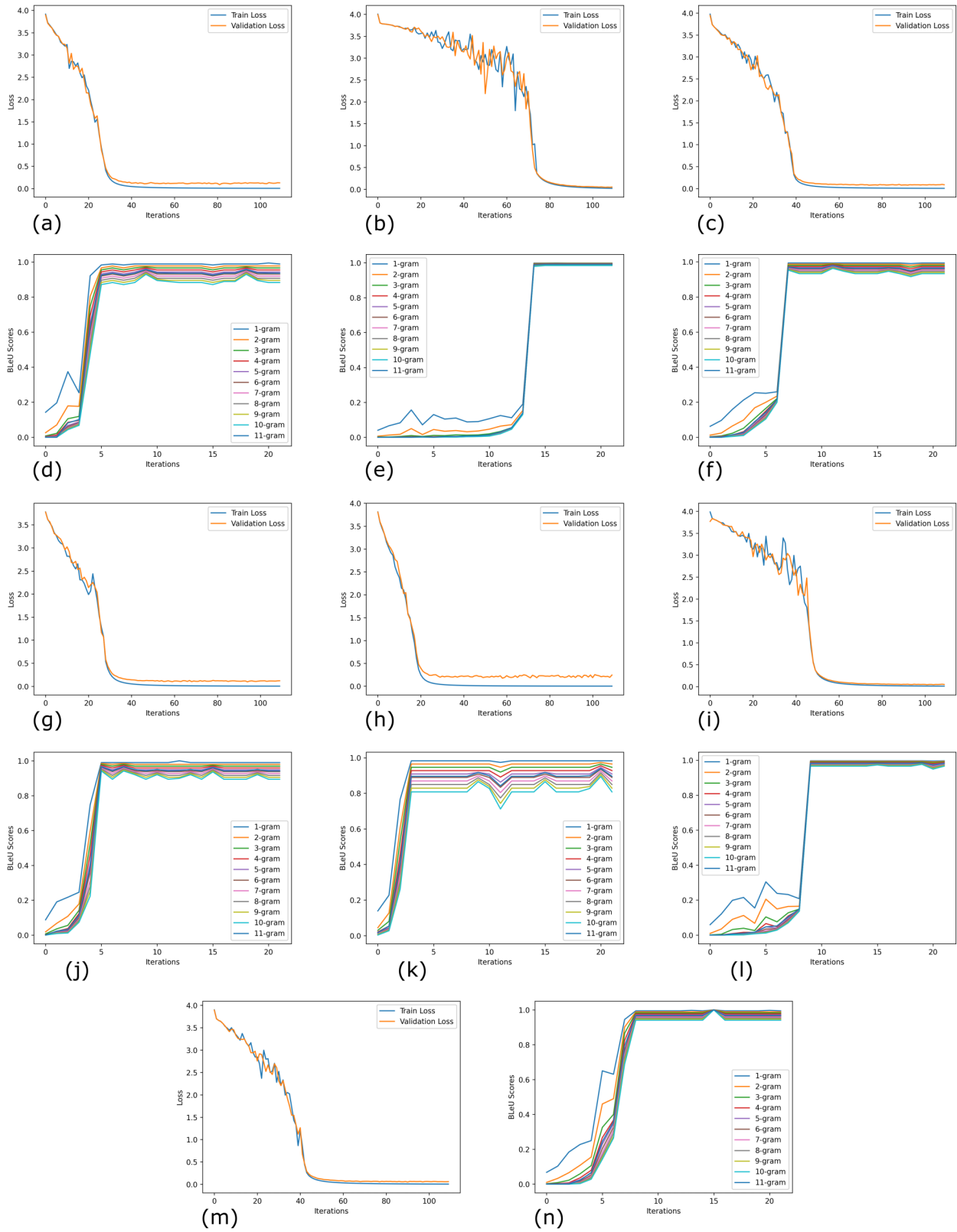


Figure S4. Bi-directional GRUs with attention. Training loss for (a) NSP1, (b) NSP3, (c) NSP5, (g) NSP8, (h) NSP9, (i) NSP13, and (m) NSP15. BLEU scores for (d) NSP1, (e) NSP3, (f) NSP5, (j) NSP8, (k) NSP9, (l) NSP13, and (n) NSP15.

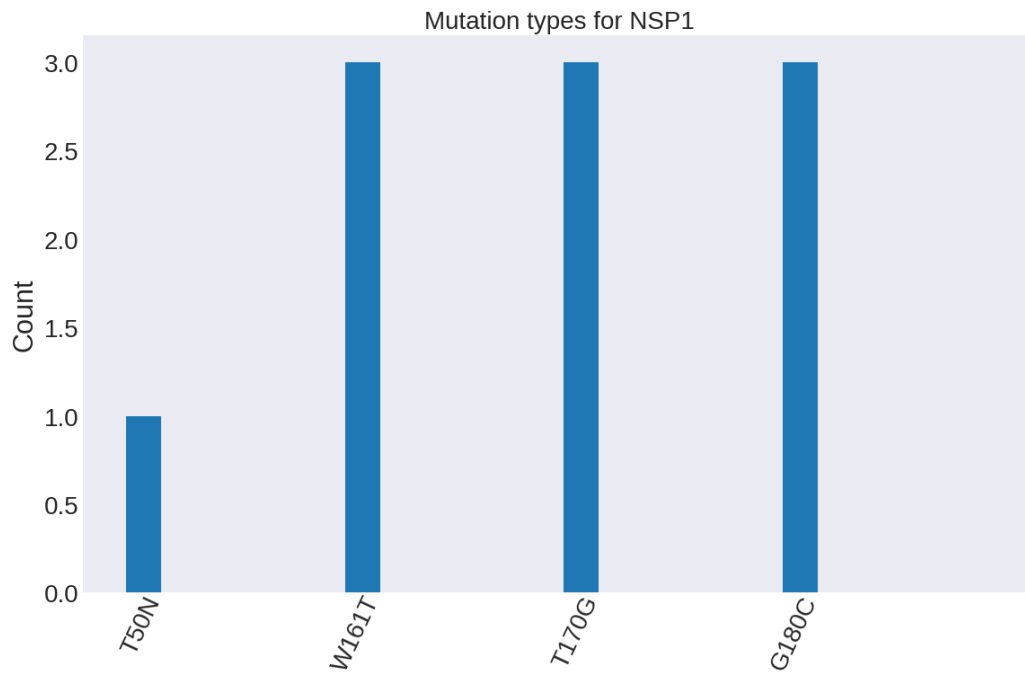


Figure S5. Predicted frequently occurring mutations in NSP1.

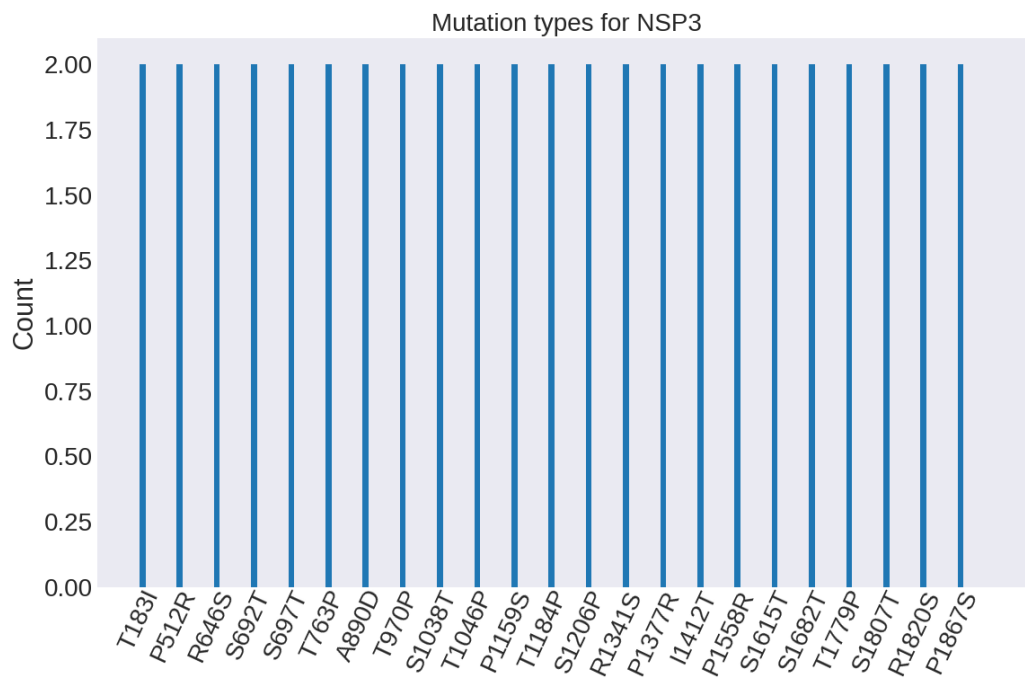


Figure S6. Predicted frequently occurring mutations in NSP3.

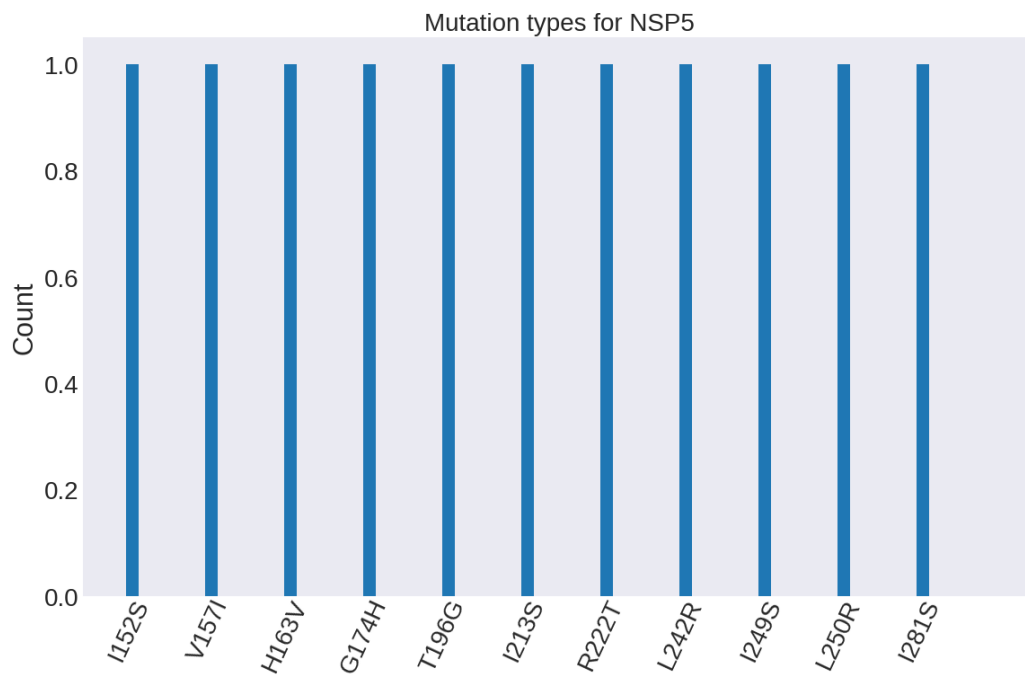


Figure S7. Predicted frequently occurring mutations in NSP5.

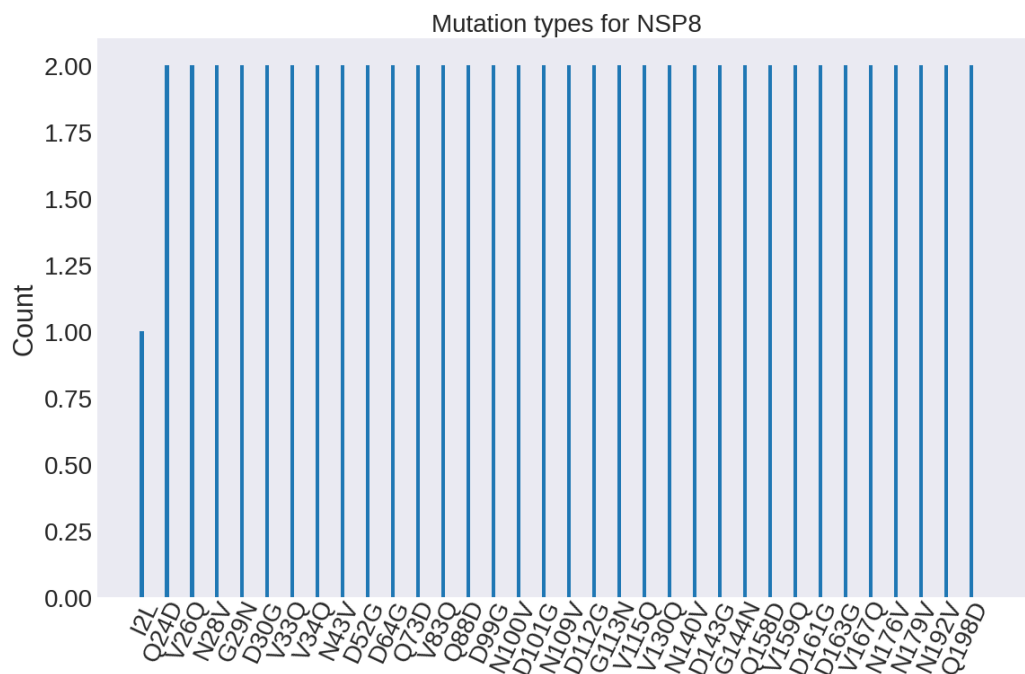


Figure S8. Predicted frequently occurring mutations in NSP8.

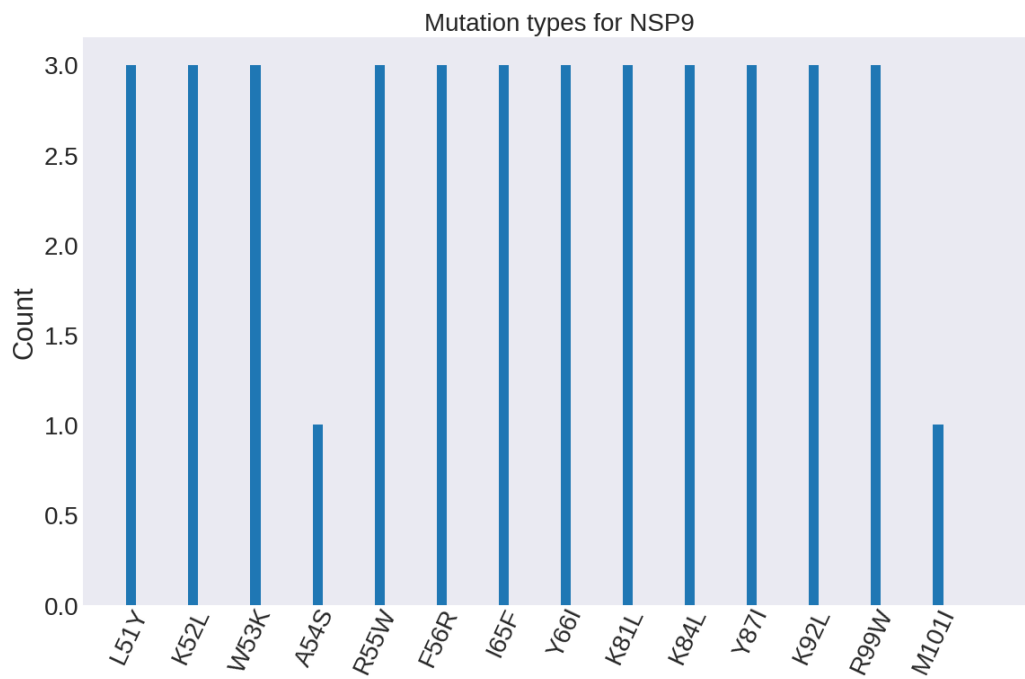


Figure S9. Predicted frequently occurring mutations in NSP9.

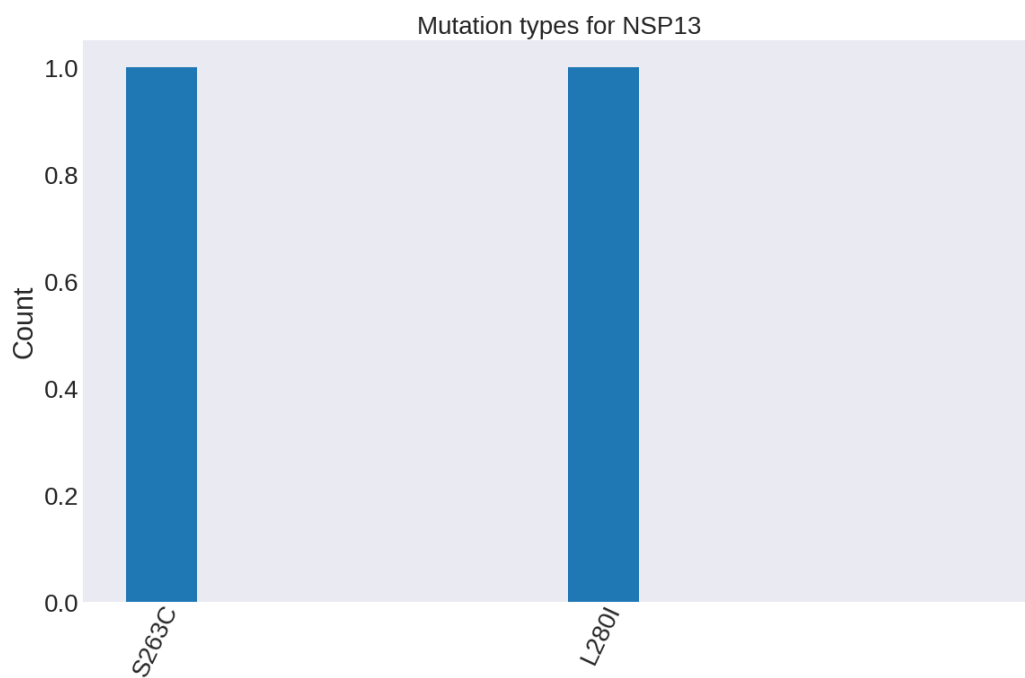


Figure S10. Predicted frequently occurring mutations in NSP13.

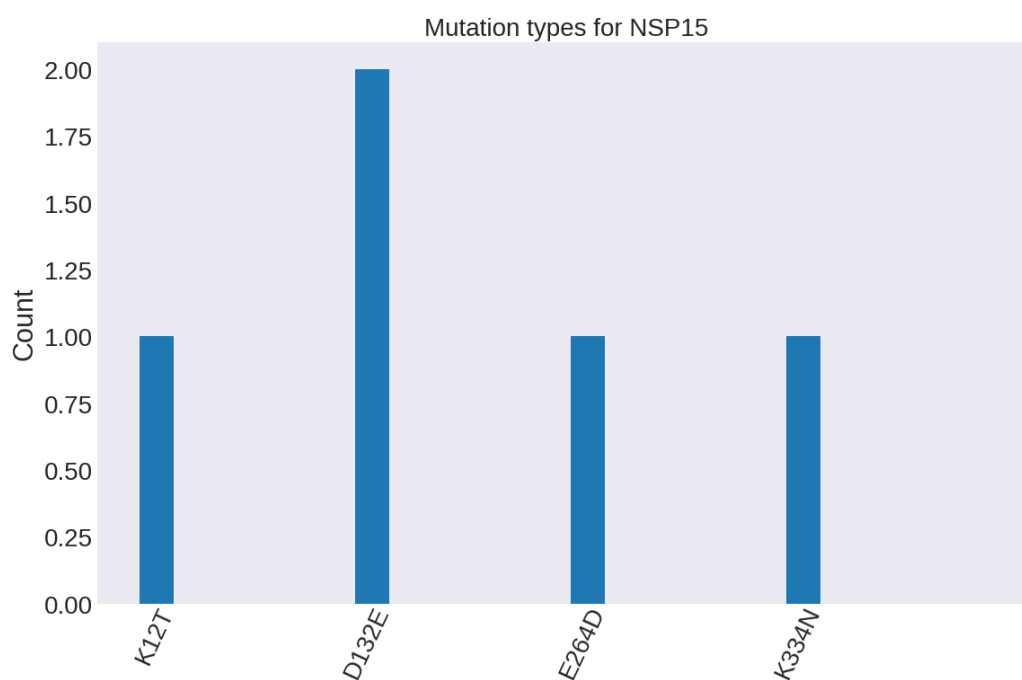


Figure S11. Predicted frequently occurring mutations in NSP15.



AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)