Journal of Virology

| Editor's Pick | Computational Biology | Gem

Pathogen genomic surveillance and the AI revolution

Spyros Lytras,[1,2] Kieran D. Lamb,[2,3] Jumpei Ito,[1,4] Joe Grove,[2] Ke Yuan,[3,5,6] Kei Sato,[1,2,4,7,8,9,10] Joseph Hughes,[2] David L. Robertson[2]

AUTHOR AFFILIATIONS See affiliation list on p. 5.

ABSTRACT The unprecedented sequencing efforts during the COVID-19 pandemic paved the way for genomic surveillance to become a powerful tool for monitoring the evolution of circulating viruses. Herein, we discuss how a state-of-the-art artificial intelligence approach called protein language models (pLMs) can be used for effectively analyzing pathogen genomic data. We highlight examples of pLMs applied to predicting viral properties and evolution and lay out a framework for integrating pLMs into genomic surveillance pipelines.

KEYWORDS pLM, genomic surveillance, AI, language models, virus surveillance, public health

Genomic pathogen sequencing during the COVID-19 pandemic highlighted the critical role of surveillance in understanding the persistence of SARS-CoV-2 in the human population despite high levels of protective immunity and the effect of the changing circulating variants on public health, advising policymakers and informing vaccine design. With nearly 20 million SARS-CoV-2 genomes currently available in the GISAID database, the sequencing effort since 2020 has surpassed those of other viruses, followed by less than half a million influenza virus genomes (GISAID) and less than 50,000 HIV-1 genomes (GenBank) all collected over several decades. The almost real-time resolution of circulating SARS-CoV-2 variants allowed scientists to readily detect and characterize genome changes corresponding to the virus's functional and antigenic evolution in humans and anthroponosis events to other animals. This phenotypic change was apparent from the outset of the pandemic, with genomic surveillance revealing the first "functional" substitution, D614G (1, 2), and first significant antigenic substitution, N439K (3), both changes in the virus's Spike entry glycoprotein. In addition to such point mutations creating variants with altered phenotypic properties, the global sequencing effort uncovered several mechanisms this new human virus used to maintain its circulation in the population: recurring indel mutations generating new variants, e.g., deletion H69/V70 (4); independent occurrence of the same amino acid substitutions at the same site in different variants (convergence) (5, 6); the emergence of variants of concern associated with multiple beneficial mutations through saltation-like evolution linked to chronic infections (7, 8); and bringing novel sets of mutations together in one variant (recombination) (9, 10), which can have a combinatory beneficial effect for the virus (positive epistasis) (11).

The scientific and public health communities' response to the COVID-19 pandemic has been an unprecedented effort to track a novel virus's evolution in near real time. However, SARS-CoV-2 surveillance was fragmentary with global disparities in coverage (12), and as the COVID-19 pandemic is now deemed over, sequencing has been downscaled. Given the general failure to learn the importance of sufficient levels of widespread sequencing to understand novel viruses, it is thus optimistic to expect that the same amount of effort will be put into understanding the next virus that spills over into humans. In fact, many viruses that have a high potential for impacting global health—but primarily affect resource-limited countries, such as the six priority virus

Editor Suchetana Mukhopadhyay, Indiana University Bloomington, Bloomington, Indiana, USA

Address correspondence to Spyros Lytras, spyros@g.ecc.u-tokyo.ac.jp.

S.L. received consulting fees from the EcoHealth Alliance. J.I. received consulting fees and honoraria for lectures from Takeda. K.S. received consulting fees from Moderna and Takeda, as well as honoraria for lectures from Gilead Sciences, Moderna, and Shionogi & Co.

See the funding table on p. 6.

Published 29 January 2025

pathogens causing human disease (MERS, Lassa, Nipah, Rift Valley Fever, Chikungunya, and Ebola) defined by the Coalition for Epidemic Preparedness Innovations—are largely under-surveilled. In addition to sequencing efforts, insights on pathogen evolution are further powered by time- and resource-consuming computational analyses, as well as validation with experimental approaches, such as *in vitro* deep mutational scanning (DMS) (13, 14). Advocating for more funding for pathogen genomic surveillance is undeniably the best approach forward; however, other priorities and resource limitations have made this approach disappointingly underwhelming so far (15). This leaves us with the need to develop enhanced computational approaches for answering the following questions: how do we monitor SARS-CoV-2 evolution and that of other endemic viruses when surveillance is diminished? How do we characterize the evolutionary landscape of neglected pathogens without knowing their underlying diversity? And can we predict which viruses have the highest zoonotic potential and prepare ourselves for an outbreak of a truly novel so-called "disease X" pathogen?

The field of artificial intelligence (AI) has recently experienced a boom in popularity and commercial applications, particularly in the form of large language models (LLMs). LLMs are state-of-the-art deep learning models trained on a large text corpus (set of words in the context of sentences) that identify fundamental properties of grammar and meaning (16). As such, the models can predict how changing a word can alter the meaning of a sentence given its understanding of these words in the specified context. In the same way, LLMs can be trained in any other corpus of contextualized characters, for example, amino acids in protein sequences, i.e., a protein language model (pLM) (17). Hie et al. (18) proposed that language models trained on protein sequences, instead of words, can be used to represent biological and biochemical properties of viruses. By introducing or "embedding" a new sequence into the model's latent representation space, the potential effect of amino acid substitutions in a given peptide sequence can be predicted (19). Applied in the context of virus entry glycoproteins, the authors suggest that the changes in the "meaning" of a protein correspond to antigenic change, while changes in the "grammar" of the protein reflect changes in viral fitness. A number of pLMs have been developed recently (20), a popular example being Evolutionary Scale Modeling (ESM), trained, in the case of ESM-2, on over 60 million unique protein sequences covering all known biological diversity (21, 22). The key merit is that these models have a generalized and transferable "understanding" of proteins and, therefore, can be used to analyze completely novel sequences without the need for a multiple sequence alignment as required by, e.g., AlphaFold models (23). So far, the ESM family of models has proven revolutionary for quick and efficient predictions of proteins' tertiary structure and protein design (22, 24, 25). But can the AI techniques be used for effective pathogen genomic surveillance?

Two of our recent studies have combined ESM-2 with the available SARS-CoV-2 genomic surveillance data to make inferences and predictions about the virus's biological properties and evolution. First, on the premises of the protein "meaning" and "grammar" scores proposed by Hie et al. (18), Lamb and colleagues (26) use an *in silico* deep mutational scanning approach where each site in the Spike glycoprotein is iteratively changed to each possible amino acid residue. The model can then predict which sites are likely to accommodate substitutions and which changes are likely to be beneficial or deleterious for the virus, providing a high-level understanding of the effect substitutions may have on the protein function and structure. This information is engraved in protein evolution but may not be captured by high-throughput *in vitro* DMS methods, such as combinatorial antigenic changes and epistatic substitutions. Second, Ito and colleagues (27) adapt the ESM-2 architecture on three layers of information: (i) the reproductive number associated with each SARS-CoV-2 variant that has circulated in the past, (ii) antibody evasion *in vitro* DMS data, and (iii) Spike sequences representing the wider diversity of the *Coronaviridae*. This process allows the model to learn about the properties of variants that encode distinct Spike proteins and the effect that substitutions on these proteins have on evading immunity. Combining this knowledge

on SARS-CoV-2 evolution with ESM-2's understanding of protein language, the authors create a model that can accurately predict the reproductive number of past and future SARS-CoV-2 variants, simply based on the virus's Spike protein sequence. Hence, models like ESM-2, which need little to no input on the circulating sequence diversity or the mechanisms behind virus evolution, can provide rapid valuable insights into pathogen biology.

Another innovative new method for viral escape forecasting has shown how using the known diversity of related viruses can largely improve the predictive power of such models (28, 29). Along the same lines, pLMs can be fine-tuned on specific groups of peptides to improve the model's understanding of that group's unique biological properties (30). We can imagine a starting model such as ESM-2, pretrained on all known proteins, that is subsequently fine-tuned on the proteome of a virus group with members affecting global health, or even all known virus receptor-binding proteins. Embedding a single protein sequence of a new virus genome into such a model would produce actionable predictions on viral fitness and antigenicity, as well as identify substitutions that may drive these functional changes. This process does not require an alignment and is essentially agnostic to the unsampled diversity of the circulating pathogen. Even though genomic surveillance is still necessary, only a fraction of the circulating diversity needs to be sampled for these models to provide predictive insights into the pathogen's function and potential evolutionary trajectories. The current AI models' usefulness does not extend to epidemiological insights, but sequencing is not required for this, with multiple cheaper and more widely available diagnostic tools existing for most viral pathogens.

It is clear that AI and protein language models hold great and mostly untapped potential for making genomic pathogen surveillance more effective, responsive, and less costly. The COVID-19 pandemic has ignited the development of extensive genomic surveillance infrastructure globally, which will likely be applied in future public health emergencies. Now, as part of ongoing preparedness efforts, is the right time to urge scientists developing the computational side of such infrastructure to consider implementing pLM methods in their pipelines (Fig. 1). Of course, improvements can be made to the methods themselves, for example, expanding and balancing the data sets these models are trained on (31). A better understanding of overall protein diversity—and specifically virus diversity—can largely aid the predictive ability of these models; thus, we note the paramount importance of animal virus discovery in pandemic preparedness. Not only for finding where viruses with zoonotic potential reside but also for improving the predictive models' understanding of virus biology and for rational design of drug and vaccine interventions.

The inherent caveat of machine learning approaches, being restricted by the diversity of the data sets they are trained on, means that predictiveness can reach an "out-of-distribution" problem. That is, genuine inferences about a protein sequence can only be made if it falls within the distribution of sequence contexts the model was trained on. Metagenomic virus discovery and high-throughput reanalysis of existing sequencing read data sets can largely aid in diversifying training data beyond the formally annotated viral protein diversity and expanding the models' context distribution (32–34). In addition to diversity in the sequence training data set, the same caveat applies to functional data used for model training. Hence, aside from continuous virus surveillance and discovery, experimental molecular virology and phenotypic virus characterization—potentially driven by initial AI model predictions—should be regularly performed, and the resulting data used for training models (Fig. 1). The need for updating training data sets should naturally come hand-in-hand with frequently updating the models themselves. Even though training a pLM from scratch can be very computationally intensive, updating existing pLMs with specialized tasks on new data is substantially easier, making this technology extremely fitting for dynamically updating data sets (27).

In the presence of comprehensive epidemiological, sequencing, and phenotypic data for a virus, combinatorial predictive models can be developed that perform well for

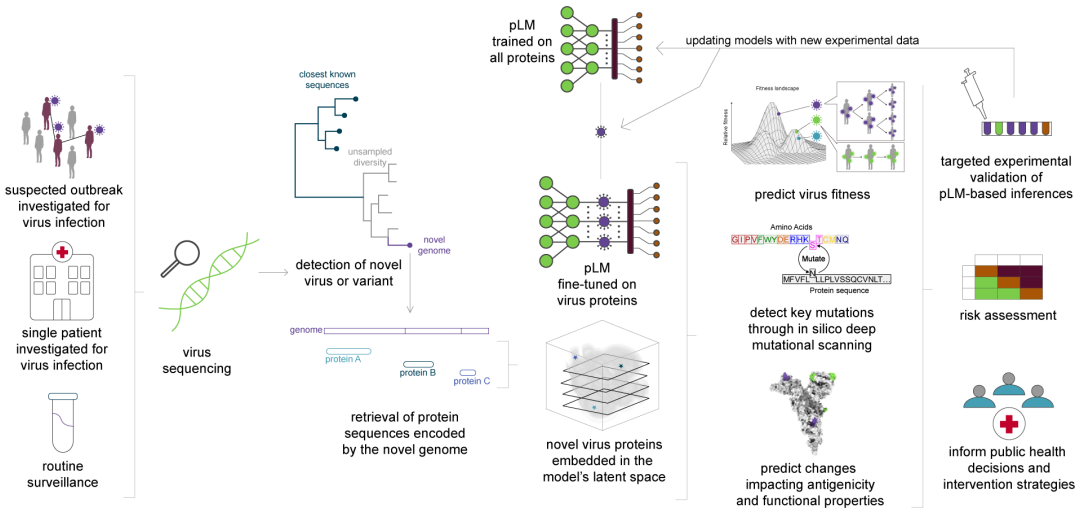

FIG 1 Overview of integrating pLMs to pathogen genomic surveillance pipelines to more effectively inform public health decisions on potential outbreaks.

variant prediction (35). However, this breadth of data is rarely available. Despite the "out-of-distribution" problem described above, AI models do have the extrapolative capacity to make accurate predictions in the absence of comprehensive data. The future of pathogen preparedness should focus on efficiency and effectiveness in translating data to successful public health policies and decisions. Integrating AI-driven methods into our routine preparedness pipelines will pave the way for making fast and informative risk assessments about novel pathogens. This approach reduces the need for comprehensive sequencing or immediate experimental validation, which can become targeted and frequently feedback to the models (Fig. 1).

When discussing the need for data to train AI models, we should also highlight the importance of effectively acknowledging data contributors. Acknowledging sequence submitters has been an important discussion throughout the COVID-19 pandemic where global availability of SARS-CoV-2 sequences was critical for public health decisions (36). For effective and equitable use of the AI-based technologies discussed here, we believe that all published models should have full transparency of their training data sets, including acknowledging all parties that led to the data collection. Notably, much of the sampling and sequencing takes place in resource-limited areas. Acknowledging these efforts is essential, but it must be accompanied by guarantees that data submitters can also benefit from the models trained on their contributions. Simply making model weights and code available is insufficient. In many regions, local computing resources are prohibitively expensive, while technical expertise is lacking, and necessary infrastructure may be unavailable. To address this, investment in AI development should include technical training for data contributors and establishing models available through cloud computing where only internet access is required for using the models instead of costly local infrastructure. This should apply to academia, but also—perhaps even more importantly—to industry since most "foundational" models to date have been developed by tech companies that hold the necessary resources and capital to support such projects. In these settings, ethics committees should be in place to enforce equitable AI model development. Ensuring that data providers are acknowledged and have full access to the resulting models is the only way to sustain continuous updating and, subsequently, the predictive performance of these models.

So far, genomic surveillance has complemented public policy after a viral outbreak has already occurred. The generalizable nature of pLMs has the potential to help us create transformative tools where a single genome sequence allows us to predict and respond to spillovers, understand variant effects (26), predict functional interactions (37), and what interventions are required before widespread transmission takes off in

the human population, enabling truly actionable preparedness strategies. The compute capacity required to build such models will require access to intensive GPU resources and expertise not usually available to academic biologists. Interdisciplinary partnerships, know-how, and training will need to be in place in preparation for the next virus outbreak. Without a doubt, viruses will continue to spill over and cause outbreaks that we can prevent if we are equipped with the right tools; the time for action is now.

ACKNOWLEDGMENTS

We acknowledge funding from the UK Medical Research Council (MRC, MC_UU_12014/12, MC_UU_00034/5, MR/V01157X/1, and a Doctoral Training Programme in Precision Medicine studentship for KDL, MR/N013166/1) and the UK Research and Innovation (UKRI) to the G2P2 consortium (MR/Y004205/1). J.I. is supported by grants from Japan Science and Technology Agency PRESTO (JPMJPR22R1), JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (JP23K14526), and Mitsubishi UFJ Financial Group Vaccine Development Research Support. J.G. is supported by the Wellcome Trust and Royal Society through a Sir Henry Dale Fellowship (107653/Z/15/Z). K.S. is supported by the Agency for Medical Research and Development (AMED) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) Program (JP24jf0126002), AMED Strategic Center of Biomedical Advanced Vaccine Research and Development for Preparedness and Response (SCARDA) Japan Initiative for World-leading Vaccine Research and Development Centers "UTOPIA" (JP243fa627001), AMED SCARDA Program on R&D of new generation vaccine including new modality application (JP243fa727002), AMED Research Program on Emerging and Re-emerging Infectious Diseases (JP24fk0108690), Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (KAKENHI) (JP24H00607), and Mitsubishi UFJ Financial Group Vaccine Development Research Support.

AUTHOR AFFILIATIONS

[1]Division of Systems Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
[2]MRC-University of Glasgow Centre for Virus Research, Glasgow, Scotland, United Kingdom
[3]School of Computing Science, University of Glasgow, Glasgow, Scotland, United Kingdom
[4]International Research Center for Infectious Diseases, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
[5]School of Cancer Sciences, University of Glasgow, Glasgow, Scotland, United Kingdom
[6]Cancer Research UK Scotland Institute, Glasgow, Scotland, United Kingdom
[7]Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[8]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan
[9]International Vaccine Design Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
[10]Collaboration Unit for Infection, Joint Research Center for Human Retrovirus Infection, Kumamoto University, Kumamoto, Japan

AUTHOR ORCIDs

Spyros Lytras http://orcid.org/0000-0003-4202-6682
Kieran D. Lamb http://orcid.org/0000-0002-3011-5189
Jumpei Ito http://orcid.org/0000-0003-0440-8321
Joe Grove http://orcid.org/0000-0001-5390-7579
Ke Yuan http://orcid.org/0000-0002-2318-1460
Kei Sato http://orcid.org/0000-0003-4431-1380
Joseph Hughes http://orcid.org/0000-0003-2556-2563

David L. Robertson http://orcid.org/0000-0001-6338-0221

FUNDING

Funder	Grant(s)	Author(s)
UKRI | Medical Research Council (MRC)	MC_UU_12014/12, MC_UU_00034/5,MR/V01157X/1	Joseph Hughes David L. Robertson
UKRI | Medical Research Council (MRC)	MR/N013166/1	Kieran D. Lamb
UK Research and Innovation (UKRI)	MR/Y004205/1	David L. Robertson
MEXT | JST | Precursory Research for Embryonic Science and Technology (PRESTO)	JPMJPR22R1	Jumpei Ito
MEXT | Japan Society for the Promotion of Science (JSPS)	JP23K14526	Jumpei Ito
Wellcome Trust (WT)	107653/Z/15/Z	Joe Grove
Japan Agency for Medical Research and Development (AMED)	JP24jf0126002, JP243fa627001, JP243fa727002, JP24fk0108690	Kei Sato
MEXT | Japan Society for the Promotion of Science (JSPS)	JP24H00607	Kei Sato

AUTHOR CONTRIBUTIONS

Spyros Lytras, Conceptualization, Visualization, Writing – original draft, Writing – review and editing | Kieran D. Lamb, Conceptualization, Writing – review and editing | Jumpei Ito, Visualization, Writing – review and editing | Joe Grove, Writing – review and editing | Ke Yuan, Writing – review and editing | Kei Sato, Writing – review and editing | Joseph Hughes, Conceptualization, Writing – review and editing | David L. Robertson, Conceptualization, Writing – review and editing

REFERENCES

1. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Sheffield COVID-19 Genomics Group. 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827. https://doi.org/10.1016/j.cell.2020.06.043
2. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, Southgate J, Johnson R, Jackson B, Nascimento FF, et al. 2021. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 184:64–75. https://doi.org/10.1016/j.cell.2020.11.020
3. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva Filipe A, Wojcechowskyj JA, Davis C, Piccoli L, Pascall DJ, Dillen J, et al. 2021. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell 184:1171–1187. https://doi.org/10.1016/j.cell.2021.01.037
4. Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, Harvey WT, Lytras S, Mohamed A, Gallo G, et al. 2021. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep 35:109292. https://doi.org/10.1016/j.celrep.2021.109292
5. Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, Lucaci AG, Giandhari J, Naidoo S, Pillay Y, Singh L, Lessells RJ, NGS-SA, COVID-19 Genomics UK (COG-UK), Gupta RK, Wertheim JO, Nekturenko A, Murrell B, Harkins GW, Lemey P, MacLean OA, Robertson DL, de Oliveira T, Kosakovsky Pond SL. 2021. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell 184:5189–5200. https://doi.org/10.1016/j.cell.2021.09.003
6. Ito J, Suzuki R, Uriu K, Itakura Y, Zahradnik J, Kimura KT, Deguchi S, Wang L, Lytras S, Tamura T, et al. 2023. Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. Nat Commun 14:2671. https://doi.org/10.1038/s41467-023-38188-z
7. Harari S, Tahor M, Rutsinsky N, Meijer S, Miller D, Henig O, Halutz O, Levytskyi K, Ben-Ami R, Adler A, Paran Y, Stern A. 2022. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. Nat Med 28:1501–1508. https://doi.org/10.1038/s41591-022-01882-4
8. Hill V, Du Plessis L, Peacock TP, Aggarwal D, Colquhoun R, Carabelli AM, Ellaby N, Gallagher E, Groves N, Jackson B, et al. 2022. The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. Virus Evol 8:veac080. https://doi.org/10.1093/ve/veac080
9. Tamura T, Ito J, Uriu K, Zahradnik J, Kida I, Anraku Y, Nasser H, Shofa M, Oda Y, Lytras S, et al. 2023. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. Nat Commun 14:2800. https://doi.org/10.1038/s41467-023-38435-3
10. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, Haldenby S, Hill V, Lucaci A, McCrone JT, Nicholls SM, O'Toole Á, Pacchiarini N, Poplawski R, Scher E, Todd F, Webster HJ, Whitehead M, Wierzbicki C, COVID-19 Genomics UK (COG-UK) Consortium, Loman NJ, Connor TR, Robertson DL, Pybus OG, Rambaut A. 2021. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell 184:5179–5188. https://doi.org/10.1016/j.cell.2021.08.014
11. Martin DP, Lytras S, Lucaci AG, Maier W, Grüning B, Shank SD, Weaver S, MacLean OA, Orton RJ, Lemey P, et al. 2022. Selection analysis identifies unusual clustered mutational changes in Omicron lineage BA.1 that likely impact spike function. bioRxiv 39:2022.01.14.476382. https://doi.org/10.1101/2022.01.14.476382

12. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, Ho J, Tegally H, Githinji G, Agoti CN, et al. 2022. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun 13:7003. https://doi.org/10.1038/s41467-022-33713-y
13. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, King NP, Veesler D, Bloom JD. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182:1295–1310. https://doi.org/10.1016/j.cell.2020.08.012
14. Cao Y, Jian F, Wang J, Yu Y, Song W, Yisimayi A, Wang J, An R, Chen X, Zhang N, et al. 2023. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. Nature 614:521–529. https://doi.org/10.1038/s41586-022-05644-7
15. Hill V, Githinji G, Vogels CBF, Bento AI, Chaguza C, Carrington CVF, Grubaugh ND. 2023. Toward a global virus genomic surveillance network. Cell Host Microbe 31:861–873. https://doi.org/10.1016/j.chom.2023.03.003
16. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. 2018. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 1 (Long Papers). Association for Computational Linguistics, Stroudsburg, PA, USA. https://doi.org/10.18653/v1/N18-1202
17. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20:723. https://doi.org/10.1186/s12859-019-3220-8
18. Hie B, Zhong ED, Berger B, Bryson B. 2021. Learning the language of viral evolution and escape. Science 371:284–288. https://doi.org/10.1126/science.abd7331
19. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. 2023. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet 55:1512–1522. https://doi.org/10.1038/s41588-023-01465-0
20. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B. 2022. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44:7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381
21. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118. https://doi.org/10.1073/pnas.2016239118
22. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130. https://doi.org/10.1126/science.ade2574
23. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
24. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, Badkundri R, Shafkat I, Gong J, Derry A, Molina RS, Thomas N, Khan Y, Mishra C, Kim C, Bartie LJ, Nemeth M, Hsu PD, Sercu T, Candido S, Rives A. 2024. Simulating 500 million years of evolution with a language model. Synthetic Biology. https://doi.org/10.1101/2024.07.01.600583
25. ESM Team. 2024. ESM Cambrian: revealing the mysteries of proteins with unsupervised learning. evolutionary scale website. Available from: https://evolutionaryscale.ai/blog/esm-cambrian
26. Lamb KD, Hughes J, Lytras S, Young F, Koci O, Herzig J, Lovell SC, Grove J, Yuan K, Robertson DL. 2024. From a single sequence to evolutionary trajectories: protein language models capture the evolutionary potential of SARS-CoV-2 protein sequences. Bioinformatics. https://doi.org/10.1101/2024.07.05.602129
27. Ito J, Strange A, Liu W, Joas G, Lytras S, Sato K. 2024. A protein language model for exploring viral fitness landscapes. https://doi.org/10.1101/2024.03.15.584819
28. Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Ritter D, Sander C, Gal Y, Marks DS. 2023. Learning from prepandemic data to forecast viral escape. Nature 622:818–825. https://doi.org/10.1038/s41586-023-06617-0
29. Rochman ND, Koonin EV. 2023. Learn from the past to predict viral pandemics. Nature New Biol 622:700–702. https://doi.org/10.1038/d41586-023-02931-9
30. Yang W, Liu C, Li Z. 2023. Lightweight fine-tuning a pretrained protein language model for protein secondary structure prediction. Bioengineering. https://doi.org/10.1101/2023.03.22.530066
31. Ding F, Steinhardt J. 2024. Protein language models are biased by unequal sequence sampling across the tree of life. Bioinformatics. https://doi.org/10.1101/2024.03.07.584001
32. Chikhi R, Raffestin B, Korobeynikov A, Edgar R, Babaian A. 2024. Logan: planetary-scale genome assembly surveys life's diversity. Bioinformatics. https://doi.org/10.1101/2024.07.30.605881
33. Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, Banfield JF, de la Peña M, Korobeynikov A, Chikhi R, Babaian A. 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 602:142–147. https://doi.org/10.1038/s41586-021-04332-2
34. Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, Tian J-H, Zhang S, Zeng Z-Y, Gou Q-Y, Xin G-Y, Le S-J, Xia Y-Y, Zhou Y-L, Hui F-M, Pan Y-F, Eden J-S, Yang Z-H, Han C, Shu Y-L, Guo D, Li J, Holmes EC, Li Z-R, Shi M. 2024. Using artificial intelligence to document the hidden RNA virosphere. Cell 187:6929–6942. https://doi.org/10.1016/j.cell.2024.09.027
35. Maher MC, Bartha I, Weaver S, di Iulio J, Ferri E, Soriaga L, Lempp FA, Hie BL, Bryson B, Berger B, Robertson DL, Snell G, Corti D, Virgin HW, Kosakovsky Pond SL, Telenti A. 2022. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.06.21.21259286
36. Mallapaty S. 2024. New virus-genome website seeks to make sharing sequences easy and fair. Nature 633:501–502. https://doi.org/10.1038/d41586-024-02864-x
37. Liu D, Young F, Lamb KD, Claudio Quiros A, Pancheva A, Miller C, Macdonald C, Robertson DL, Yuan K. 2024 PLM-interact: extending protein language models to predict protein-protein interactions. Bioinformatics. https://doi.org/10.1101/2024.11.05.622169