Corrected 14 June 2021. See full text.

VIROLOGY

Learning the language of viral evolution and escape

Brian Hie[1,2], Ellen D. Zhong[1,3], Bonnie Berger[1,4]*, Bryan Bryson[2,5]*

The ability for viruses to mutate and evade the human immune system and cause infection, called viral escape, remains an obstacle to antiviral and vaccine development. Understanding the complex rules that govern escape could inform therapeutic design. We modeled viral escape with machine learning algorithms originally developed for human natural language. We identified escape mutations as those that preserve viral infectivity but cause a virus to look different to the immune system, akin to word changes that preserve a sentence's grammaticality but change its meaning. With this approach, language models of influenza hemagglutinin, HIV-1 envelope glycoprotein (HIV Env), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Spike viral proteins can accurately predict structural escape patterns using sequence data alone. Our study represents a promising conceptual bridge between natural language and viral evolution.

Viral mutations that allow an infection to escape from recognition by neutralizing antibodies have prevented the development of a universal antibody-based vaccine for influenza (*1*, *2*) or HIV (*3*) and are a concern in the development of therapies for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection (*4*, *5*). Escape has motivated high-throughput experimental techniques that perform causal escape profiling of all single-residue mutations to a viral protein (*1*–*4*). Such techniques, however, require substantial effort to profile even a single viral strain, and testing the escape potential of many (combinatorial) mutations in many viral strains remains infeasible.

Instead, we sought to train an algorithm that learns to model escape from viral sequence data alone. This approach is not unlike learning properties of natural language from large text corpuses (*6*, *7*) because languages such as English and Japanese use sequences of words to encode complex meanings and have complex rules (for example, grammar). To escape, a mutant virus must preserve infectivity and evolutionary fitness—it must obey a "grammar" of biological rules—and the mutant must no longer be recognized by the immune system, which is analogous to a change in the "meaning" or the "semantics" of the virus.

Currently, computational models of protein evolution focus either on fitness (*8*) or on functional or semantic similarity (*9*–*11*), but we want to understand both (Fig. 1A). Rather than developing two separate models of fitness and function, we developed a single model that simultaneously achieves these tasks. We leveraged state-of-the-art machine learning algorithms called language models (*6*, *7*), which learn the probability of a token (such as an English word) given its sequence context (such as a sentence) (Fig. 1B). Internally, the language model constructs a semantic representation, or an "embedding," for a given sequence (*6*), and the output of a language model encodes how well a particular token fits within the rules of the language, which we call "grammaticality" and can also be thought of as "syntactic fitness" (supplementary text, note S2). The same principles used to train a language model on a sequence of English words can train a language model on a sequence of amino acids. Although immune selection occurs on phenotypes (such as protein structures), evolution dictates that selection is reflected within genotypes (such as protein sequences),

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA. [3]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*Corresponding author. Email: bab@mit.edu (B.Be.); bryand@mit.edu (B.Br.)

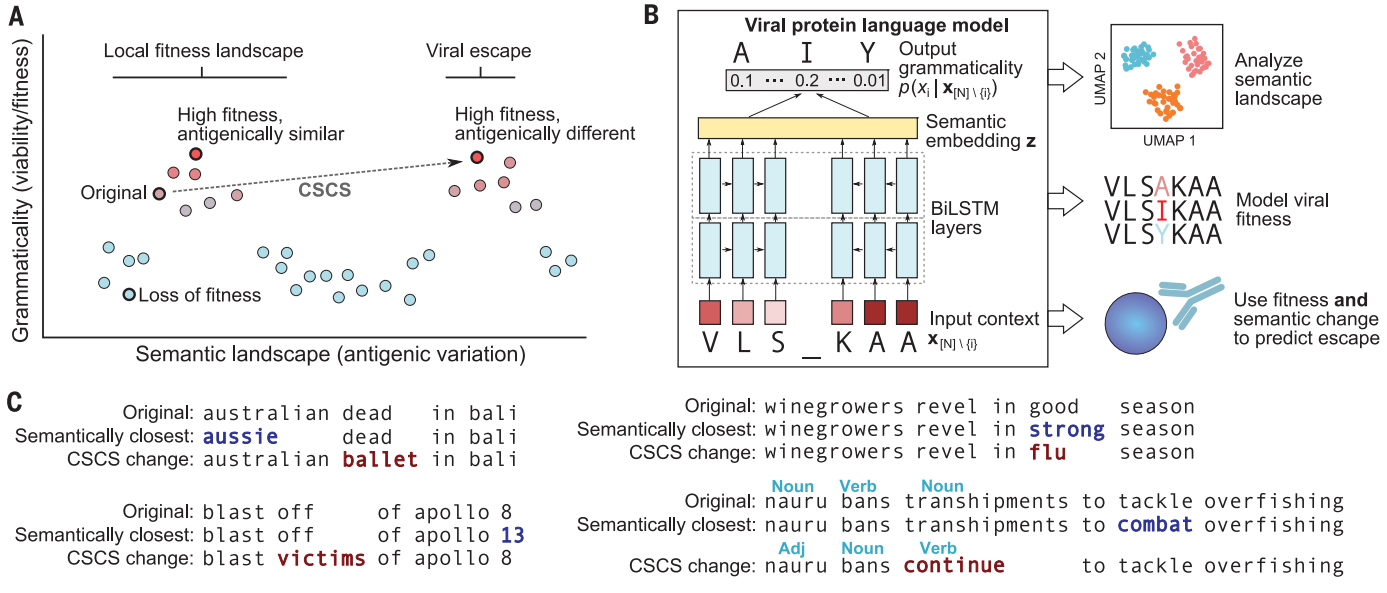

Fig. 1. Modeling viral escape requires characterizing semantic change and grammaticality. (**A**) Constrained semantic change search (CSCS) for viral escape prediction is designed to search for mutations to a viral sequence that preserve fitness while being antigenically different. This corresponds to a mutant sequence that is grammatical (conforms to the structure and rules of a language) but has high semantic change with respect to the original (for example, wild type) sequence. (**B**) A neural language model with a bidirectional long short-term memory (BiLSTM) architecture was used to learn both semantics (as a hidden layer output) and grammaticality (as the language model output). CSCS combines semantic change and grammaticality to predict escape (*12*). (**C**) CSCS-proposed changes to a news headline (implemented by using a neural language model trained on English news headlines) makes large changes to the overall semantic meaning of a sentence or to the part-of-speech structure. The semantically closest mutated sentence according to the same model is largely synonymous with the original headline.

Corrected 14 June 2021. See full text.

which language models can leverage to learn functional properties from sequence variation.

We hypothesize that (i) language model–encoded semantic change corresponds to antigenic change, (ii) language model grammaticality captures viral fitness, and (iii) both high semantic change and grammaticality help predict viral escape. Searching for mutations with both high grammaticality and high semantic change is a task that we call constrained semantic change search (CSCS) (Fig. 1C) (*12*). Our language model implementation of CSCS uses sequence data alone (which is easier to obtain than structure) and requires no explicit escape information (is completely unsupervised), does not rely on multiple sequence alignment (MSA) preprocessing (is "alignment-free"), and captures global relationships across an entire sequence (for example, because word choice at the beginning of a sentence can influence word choice at the end) (supplementary text, notes S2 and S3).

We assessed the generality of our approach across viruses by analyzing three proteins: influenza A hemagglutinin (HA), HIV-1 envelope glycoprotein (Env), and SARS-CoV-2 spike glycoprotein (Spike). All three are found on the viral surface, are responsible for binding host cells, are targeted by antibodies, and are drug targets (*1*–*5*). We trained a separate language model for each protein using a corpus of virus-specific amino acid sequences (*12*).

We initially sought to understand the semantic patterns learned by our viral language models. We therefore visualized the semantic embeddings of each sequence in the influenza, HIV, and coronavirus corpuses using Uniform Manifold Approximation and Projection (UMAP) (*13*). The resulting two-dimensional semantic landscapes show clustering patterns that correspond to subtype, host species, or both (Fig. 2), suggesting that the model was able to learn functionally meaningful patterns from raw sequence.

We quantified these clustering patterns, which are visually enriched for particular subtypes or hosts, with Louvain clustering (*14*) to group sequences on the basis of their semantic embeddings (fig. S1, A to C). We then measured the clustering purity on the basis of the percent composition of the most represented metadata category (sequence subtype or host species) within each cluster (*12*). Average cluster purities for HA subtype, HA host species, and Env subtype are 99, 96, and 95%, respectively, which are comparable with or higher than the clustering purities obtained with MSA-based phylogenetic reconstruction (Fig. 2, D and F, and fig. S1D) (*12*).

Within the HA landscape, clustering patterns suggest interspecies transmissibility. The sequence for 1918 H1N1 pandemic influenza belongs to the main avian H1 cluster, which contains sequences from the avian reservoir for 2009 H1N1 pandemic influenza (Fig. 2C and fig. S1, A to C). Antigenic similarity between H1 HA from 1918 and 2009, although nearly a century apart, is well supported (*15*). Within the landscape of SARS-CoV-2 Spike and homologous proteins, clustering proximity is consistent with the suggested zoonotic origin of several human coronaviruses (Fig. 2G), including bat and civet for SARS-CoV-1, camel for Middle East respiratory syndrome-related coronavirus (MERS-CoV), and bat and

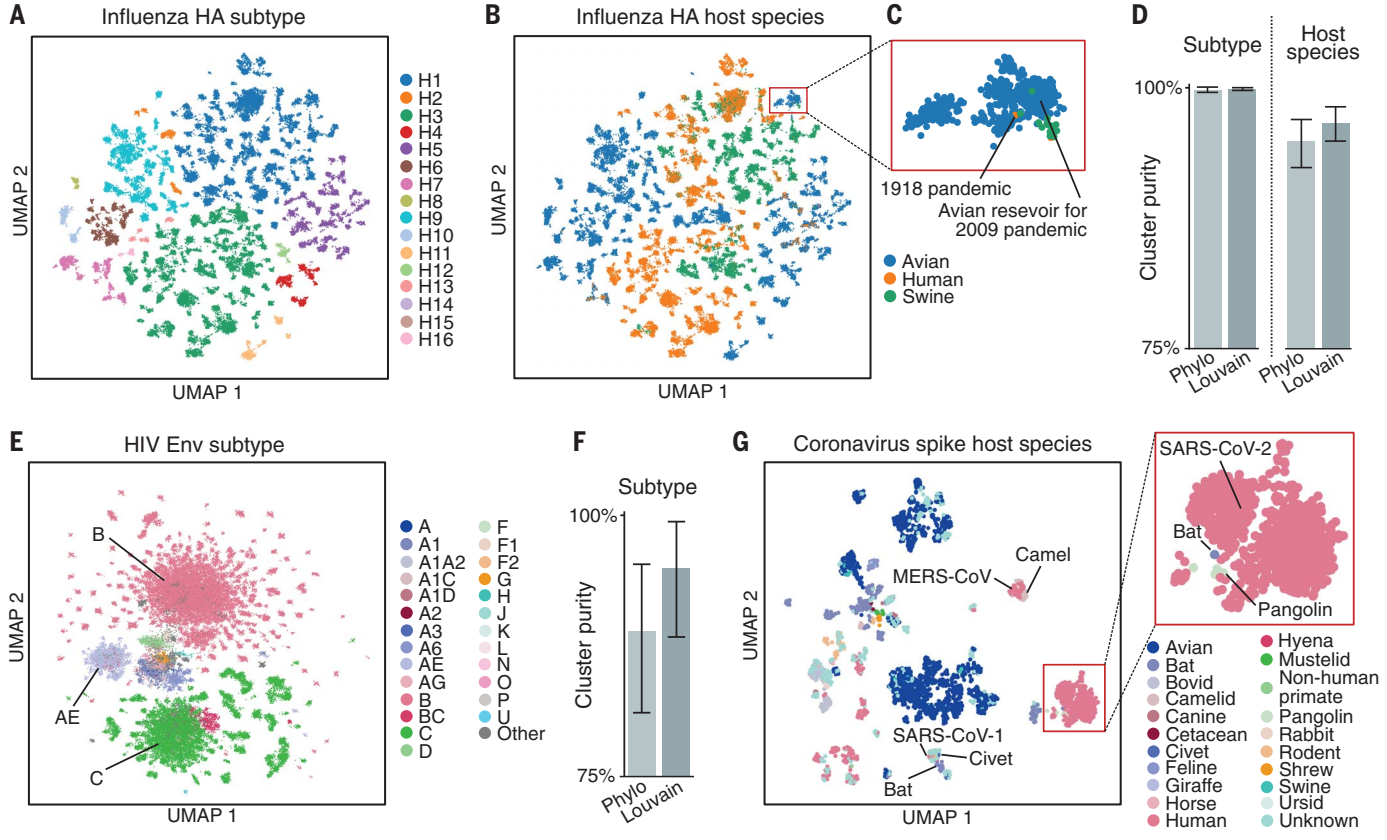

Fig. 2. Semantic embedding landscape is antigenically meaningful. (**A** and **B**) UMAP visualization of the high-dimensional semantic embedding landscape of influenza HA. (**C**) A cluster consisting of avian sequences from the 2009 flu season onward also contains the 1918 pandemic flu sequence, which is consistent with their antigenic similarity (*15*). (**D**) Louvain clusters of the HA semantic embeddings have similar purity with respect to subtype or host species compared with phylogenetic sequence clustering (Phylo). Bar height, mean; error bars, 95% confidence. (**E** and **F**) The HIV Env semantic landscape shows subtype-related distributional structure and high Louvain clustering purity. Bar height, mean; error bars, 95% confidence. (**G**) Sequence proximity in the semantic landscape of coronavirus spike proteins is consistent with the possible zoonotic origin of SARS-CoV-1, MERS-CoV, and SARS-CoV-2.

Corrected 14 June 2021. See full text.

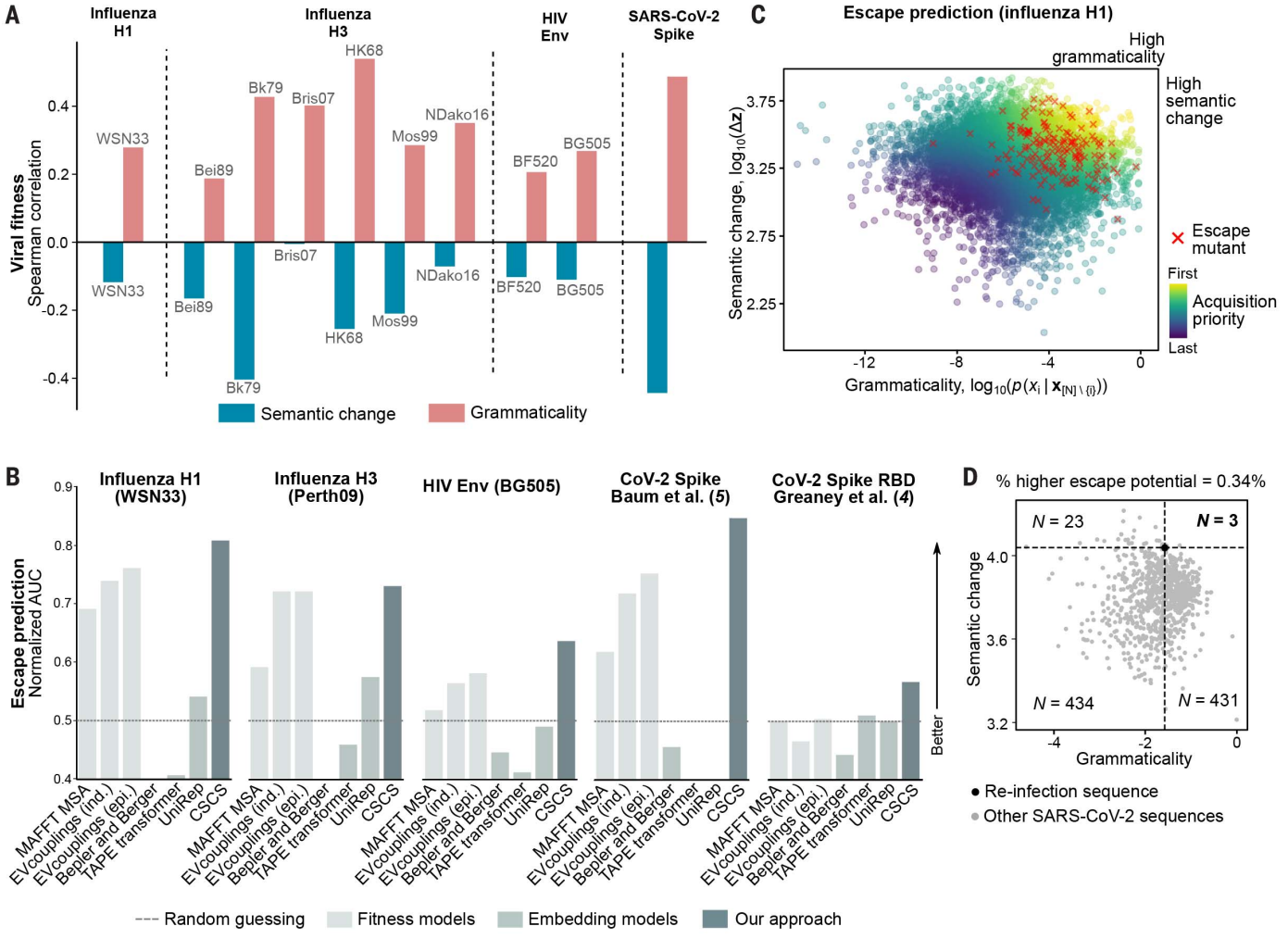

Fig. 3. Biological interpretation of language model semantics and grammaticality enables escape prediction. (**A**) Whereas grammaticality is positively correlated with fitness, semantic change has negative correlation, suggesting that most semantically altered proteins lose fitness. (**B** and **C**) However, a mutation with both high semantic change and high grammaticality is more likely to induce escape. Considering both semantic change and grammaticality enables identification of escape mutants that is consistently higher than that of previous fitness models or generic functional embedding models. (**D**) Across 891 surveilled SARS-CoV-2 Spike sequences, only three have both higher semantic change and grammaticality than a Spike sequence with four mutations that is associated with a potential reinfection case.

pangolin for SARS-CoV-2 (*16*). Analysis of these semantic landscapes strengthens our hypothesis that our viral sequence embeddings encode functional and antigenic variation.

We then assessed the relationship between viral fitness and language model grammaticality using high-throughput deep mutational scan (DMS) characterization of hundreds or thousands of mutations to a given viral protein. We obtained datasets measuring replication fitness of all single-residue mutations to A/WSN/1933 (WSN33) HA H1 (*17*), combinatorial mutations to antigenic site B in six HA H3 strains (*18*), or all single-residue mutations to BG505 and BF520 HIV Env (*19*), as well as a dataset measuring the dissociation constant (K_d) between combinatorial mutations to yeast-displayed SARS-CoV-2 Spike receptor-binding domain (RBD) and human ACE2 (*20*), which we used to approximate the fitness of Spike.

Language model grammaticality was significantly correlated (table S1, *t*-distribution *P* values) with viral fitness across all viral strains and across studies that examined single or combinatorial mutations (Fig. 3A), even though our language models were not given any explicit fitness-related information nor trained on the DMS mutants. When we compared viral fitness with the magnitude of mutant semantic change (rather than grammaticality), we observed significant negative correlation (table S1, *t*-distribution *P* values) in 8 out of 10 strains tested (Fig. 3A). This makes sense biologically because a mutation with a large effect on function is on average more likely to be deleterious and result in a loss of fitness. These results suggest that "grammaticality" of a given mutation captures fitness information and add an additional dimension to our understanding of how semantic change encodes perturbed protein function.

We then tested whether combining semantic change and grammaticality enables us to predict mutations that lead to viral escape. Our experimental setup involved making, in silico, all possible single-residue mutations to a given viral protein sequence; next, each mutant was ranked according to the CSCS objective that combines semantic change and grammaticality. We validated this ranking on the basis of enrichment of experimentally verified mutants that causally induce escape from neutralizing antibodies. Three of these causal escape datasets used a DMS with antibody selection to identify escape mutations to WSN33 HA H1 (*1*), A/Perth/16/2009 (Perth09) HA H3

Corrected 14 June 2021. See full text.

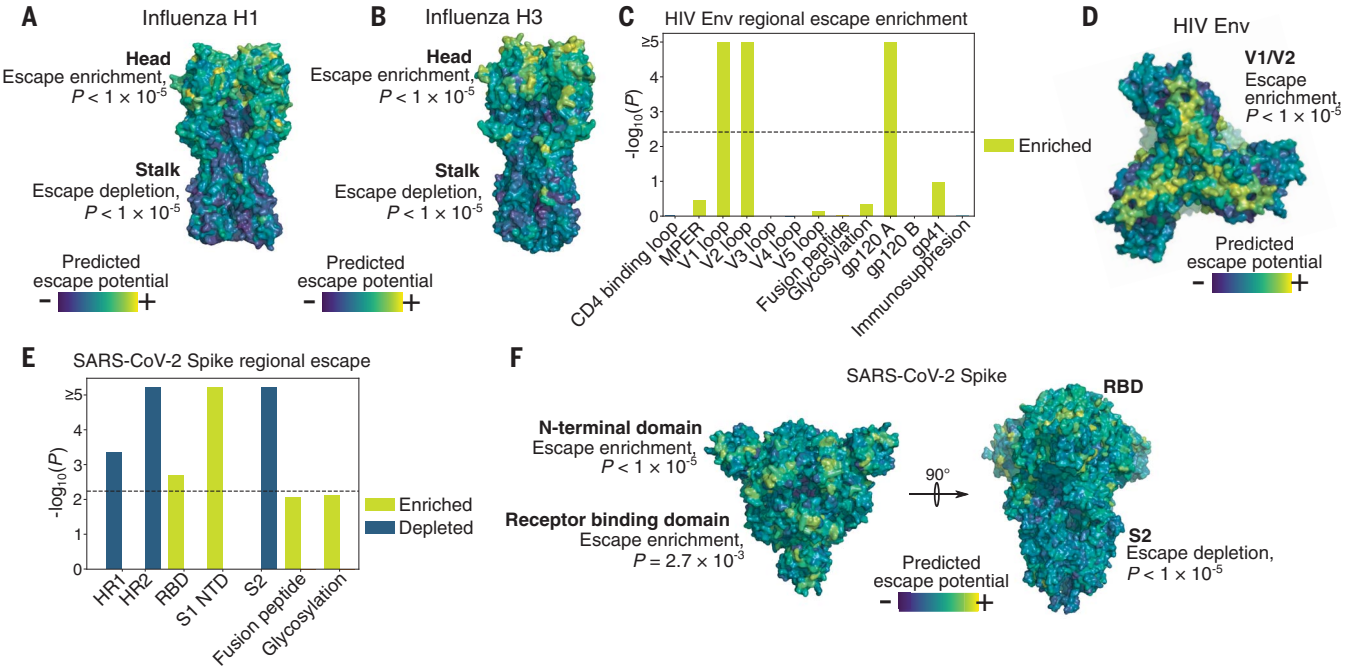

Fig. 4. Structural localization of predicted escape potential. (**A** and **B**) HA trimer colored by escape potential. (**C**) Escape potential *P* values for HIV Env. The gray dashed line indicates the statistical significance threshold. (**D**) The Env trimer colored by escape potential, oriented to show the V1/V2 regions. (**E** and **F**) Potential for escape in SARS-CoV-2 Spike is significantly enriched at the N-terminal domain and receptor binding domain (RBD) and significantly depleted at multiple regions in the S2 subunit. The gray dashed line indicates the statistical significance threshold.

(*2*), and BG505 Env (*3*). The fourth identified escape mutations to SARS-CoV-2 Spike by using natural replication error after in vitro passages under antibody selection (*5*), whereas the fifth performed a DMS to identify mutants that affect antibody binding to yeast-displayed Spike RBD (*4*).

We computed the area under the curve (AUC) of acquired escape mutations versus the total acquired mutations (*12*). In all five cases, escape prediction with CSCS resulted in both statistically significant and strong AUCs of 0.81, 0.73, 0.64, 0.85, and 0.57 for H1 WSN33, H3 Perth09, Env BG505, Spike, and Spike RBD, respectively (one-sided permutation-based $P < 1 \times 10^{-5}$ for H1, H3, Env, and Spike; $P = 2 \times 10^{-4}$ for Spike RBD) (Fig. 3B, and table S2). We did not provide the model with any information on escape, a setup in machine learning referred to as "zero-shot prediction" (*7*). The AUC decreased when ignoring either grammaticality or semantic change, evidence that both are useful in predicting escape (Fig. 3C, fig. S2A, and table S2). Although semantic change is negatively correlated with fitness, it is positively predictive (along with grammaticality) of escape (table S2), indicating that functional mutations are often deleterious, but when fitness is preserved, they are associated with antigenic change and subsequent escape from immunity.

We also tested how well alternative models of fitness (each requiring MSA preprocessing) (*8*, *21*) or of semantic change (pretrained on generic protein sequence) (*9*–*11*) predict escape, although these models are not explicitly designed for escape prediction. Fitness models associate more frequently observed patterns with higher fitness and greater escape potential, whereas semantic models associate larger functional changes with escape (*12*). CSCS with our viral language models was more predictive of escape across all five datasets (Fig. 3B and fig. S2A). Moreover, the individual grammaticality or semantic change components of our language models often outperformed benchmark models (table S2).

Language modeling can also characterize sequence changes beyond single-residue mutations, such as from accumulated replication error or recombination (*22*), although our approach is agnostic to how a sequence acquires its mutations. We therefore estimated the antigenic change and fitness of a set of four mutations to the SARS-CoV-2 Spike associated with a reported reinfection event (*23*). Among 891 other distinct, surveilled Spike sequences, we found that only three (0.34%) represent both higher semantic change and grammaticality (Fig. 3D). We estimate significant escape potential of these four mutations (random mutant null distribution $P < 1 \times 10^{-8}$) (*12*), and we observed similar patterns for known antigenically dissimilar sequences (fig. S2B) (*12*). Our analysis suggests a way to quantify the escape potential of interesting combinatorial sequence changes, such as those from possible reinfection (*23*), and calls for more information that relates combinatorial mutations to reinfection and escape.

To further assess whether our model could learn structurally relevant patterns from sequence alone, we scored each residue on the basis of the CSCS objective, visualized escape potential across the protein structure, and quantified enrichment or depletion of escape (*12*). Escape potential is significantly enriched in the HA head (permutation-based $P < 1 \times 10^{-5}$) and significantly depleted in the HA stalk (permutation-based $P < 1 \times 10^{-5}$) (Fig. 4, A and B; fig. S3; and table S3), which is consistent with literature on HA mutation rates and supported by the successful development of antistalk broadly neutralizing antibodies (*24*). We also detected, consistent with existing knowledge, a significant enrichment (permutation-based $P < 1 \times 10^{-5}$) of escape mutations in the V1/V2 hypervariable regions of the HIV Env (Fig. 4, C and D; fig. S3; and table S3) (*3*). Our model only learns escape patterns linked to mutations, rather than post-translational changes such as glycosylation that contribute to HIV escape (*3*), which may explain the lack of escape potential specifically assigned to Env glycosylation sites (Fig. 4C and table S3).

The escape potential within the SARS-CoV-2 Spike is significantly enriched in both the RBD (permutation-based $P = 2.7 \times 10^{-3}$) and N-terminal domain (permutation-based

$P < 1 \times 10^{-5}$), whereas escape potential is significantly depleted in the S2 subunit (permutation-based $P < 1 \times 10^{-5}$) (Fig. 4, E and F; fig. S3; and table S3). These results are supported by the greater evolutionary conservation at S2 antigenic sites (*25*). Our model of Spike escape therefore suggests that immunodominant antigenic sites in S2 (*5*, *25*) may be more stable target antibody epitopes and underscores the need for more exhaustive causal escape profiling of Spike in regions beyond the RBD.

Our study leverages the principle that evolutionary selection is reflected in sequence variation. This principle may allow CSCS to generalize beyond viral escape to different kinds of natural selection (such as T cell selection) or drug selection. CSCS and its components could be used to select elements of a multivalent or mosaic vaccine. Our techniques also lay the foundation for more complex modeling of sequence dynamics. We anticipate that the "distributional hypothesis" from linguistics (*26*), in which co-occurrence patterns can model complex concepts and on which language models are based, can further inform viral evolution.

REFERENCES AND NOTES

1. M. B. Doud, J. M. Lee, J. D. Bloom, *Nat. Commun.* **9**, 1386 (2018).
2. J. M. Lee *et al.*, *eLife* **8**, e49324 (2019).
3. A. S. Dingens, D. Arenz, H. Weight, J. Overbaugh, J. D. Bloom, *Immunity* **50**, 520–532.e3 (2019).
4. A. J. Greaney *et al.*, *Cell Host Microbe* 10.1016/j.chom.2020.11.007 (2020).
5. A. Baum *et al.*, *Science* **369**, 1014–1018 (2020).
6. M. Peters *et al.*, Deep contextualized word representations. *Proc. NAACL-HLT*, 2227–2237 (2018).
7. A. Radford *et al.*, *OpenAI Blog* **1**, 9 (2019).
8. T. A. Hopf *et al.*, *Bioinformatics* **35**, 1582–1584 (2019).
9. T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure. arXiv:1902.08661 [cs.LG] (2019).
10. R. Rao *et al.*, Evaluating protein transfer learning with TAPE. *Proc. Adv. Neural Inf. Process. Syst.*, 9686–9698 (2019).
11. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat. Methods* **16**, 1315–1322 (2019).
12. Materials and methods are available as supplementary materials.
13. L. McInnes, J. Healy, UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv:1802.03426 [stat.ML] (2018).
14. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
15. R. Xu *et al.*, *Science* **328**, 357–360 (2010).
16. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, *Nat. Med.* **26**, 450–452 (2020).
17. M. B. Doud, J. D. Bloom, *Viruses* **8**, 155 (2016).
18. N. C. Wu *et al.*, *Nat. Commun.* **11**, 1233 (2020).
19. H. K. Haddox, A. S. Dingens, S. K. Hilton, J. Overbaugh, J. D. Bloom, *eLife* **7**, e34420 (2018).
20. T. N. Starr *et al.*, *Cell* **182**, 1295–1310.e20 (2020).
21. K. Katoh, D. M. Standley, *Mol. Biol. Evol.* **30**, 772–780 (2013).
22. Y. Xiao *et al.*, *Cell Host Microbe* **19**, 493–503 (2016).
23. K. K.-W. To *et al.*, *Clin. Infect. Dis.* ciaa1275 (2020).
24. E. Kirkpatrick, X. Qiu, P. C. Wilson, J. Bahl, F. Krammer, *Sci. Rep.* **8**, 10432 (2018).
25. S. Ravichandran *et al.*, *Sci. Transl. Med.* **12**, eabc3539 (2020).
26. Z. S. Harris, *Word* **10**, 146–162 (1954).
27. B. Hie, brianhie/viral-mutation: viral-mutation release 0.3. Zenodo (2020).
28. B. Hie, Data for "Learning the language of viral evolution and escape". Zenodo (2020).

ACKNOWLEDGMENTS

We thank A. Balazs, H. Fraser, O. Leddy, A. Lerer, A. Lin, A. Nitido, U. Roy, and A. Schmidt for helpful discussions. We thank S. Chun, B. DeMeo, A. Narayan, A. Nguyen, S. Nyquist, and A. Wu for assistance with the manuscript. **Funding:** B.H. is partially supported by the U.S. Department of Defense (DOD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG). E.Z. is partially supported by the National Science Foundation (NSF) Graduate Research Fellowship. **Author contributions:** All authors conceived the project and methodology. B.H. performed the computational experiments and wrote the software. All authors interpreted the results and wrote the manuscript. **Competing interests:** B.H., B.Be., and B.Br. have filed a provisional patent application (serial no. 53/049,676) related to this work. **Data and materials availability:** Code, scripts for plotting and visualizing, and pretrained models are deposited to Zenodo at doi:10.5281/zenodo.4034681 (*27*) and are also available at https://github.com/brianhie/viral-mutation. We used the following publicly available datasets for model training: Influenza A HA protein sequences from the NIAID Influenza Research Database (IRD) (www.fludb.org); HIV-1 Env protein sequences from the Los Alamos National Laboratory (LANL) HIV database (www.hiv.lanl.gov); *Coronavidae* spike protein sequences from the Virus Pathogen Resource (ViPR) database (www.viprbrc.org/brc/home.spg?decorator=corona); SARS-CoV-2 Spike protein sequences from NCBI Virus (www.ncbi.nlm.nih.gov/labs/virus/vssi); and SARS-CoV-2 Spike and other Betacoronavirus spike protein sequences from GISAID (www.gisaid.org). We used the following publicly available datasets for fitness and escape validation: Fitness single-residue DMS of HA H1 WSN33 from Doud and Bloom (2016) (*17*); Fitness combinatorial DMS of antigenic site B in six HA H3 strains from Wu *et al.* (*18*); Fitness single-residue DMS of Env BF520 and BG505 from Haddox *et al.* (*19*); ACE2 binding affinity combinatorial DMS of Spike RBD from Starr *et al.* (*20*); Escape single-residue DMS of HA H1 WSN33 from Doud *et al.* (2018) (*1*); Escape single-residue DMS of HA H3 Perth09 from Lee *et al.* (*2*); Escape single-residue DMS of Env BG505 from Dingens *et al.* (*3*); Escape mutations of Spike from Baum *et al.* (*5*); Escape single-residue DMS of Spike RBD from Greaney *et al.* (*4*); Training and validation datasets are deposited to Zenodo at doi:10.5281/zenodo.4029296 (*28*), and links to this data are also available at https://github.com/brianhie/viral-mutation.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6526/284/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S3
References (*29*–*49*)
MDAR Reproducibility Checklist

8 July 2020; accepted 9 November 2020
10.1126/science.abd7331

Science

Learning the language of viral evolution and escape

Brian HieEllen D. ZhongBonnie BergerBryan Bryson

Science, 371 (6526), • DOI: 10.1126/science.abd7331

Natural language predicts viral escape

Viral mutations that evade neutralizing antibodies, an occurrence known as viral escape, can occur and may impede the development of vaccines. To predict which mutations may lead to viral escape, Hie *et al.* used a machine learning technique for natural language processing with two components: grammar (or syntax) and meaning (or semantics) (see the Perspective by Kim and Przytycka). Three different unsupervised language models were constructed for influenza A hemagglutinin, HIV-1 envelope glycoprotein, and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike glycoprotein. Semantic landscapes for these viruses predicted viral escape mutations that produce sequences that are syntactically and/or grammatically correct but effectively different in semantics and thus able to evade the immune system.

Science, this issue p. 284; see also p. 233