Learning Mutational Semantics

Brian Hie
MIT CSAIL
Cambridge, MA 02139
brianhie@mit.edu

Ellen Zhong
MIT CSB and CSAIL
Cambridge, MA 02139
zhonge@mit.edu

Bryan Bryson
MIT Biological Engineering
Cambridge, MA 02139
bryand@mit.edu

Bonnie Berger
MIT Mathematics and CSAIL
Cambridge, MA 02139
bab@mit.edu

Abstract

In many natural domains, changing a small part of an entity can transform its semantics; for example, a single word change can alter the meaning of a sentence, or a single amino acid change can mutate a viral protein to escape antiviral treatment or immunity. Although identifying such mutations can be desirable (for example, therapeutic design that anticipates avenues of viral escape), the rules governing semantic change are often hard to quantify. Here, we introduce the problem of identifying mutations with a large effect on semantics, but where valid mutations are under complex constraints (for example, English grammar or biological viability), which we refer to as constrained semantic change search (CSCS). We propose an unsupervised solution based on language models that simultaneously learn continuous latent representations. We report good empirical performance on CSCS of single-word mutations to news headlines, map a continuous semantic space of viral variation, and, notably, show unprecedented zero-shot prediction of single-residue escape mutations to key influenza and HIV proteins, suggesting a productive link between modeling natural language and pathogenic evolution.[1]

1 Introduction

Much of the effort devoted to learning machine-intelligible representations of natural language semantics has been built on the "distributional hypothesis," in which the context and co-occurrence of words is assumed to provide insight into the meaning of words [25, 22, 35, 38, 41, 43]. While distributional semantics was developed to model human intuitive notions of "meaning," similar reasoning may be useful for domains beyond human intuition.

For example, like linguistic semantics, biological function is encoded by a sequence of tokens (the bases of nucleic acids or the amino acid residues of proteins) that is determined by a complex distributional structure. Promisingly, recent analyses of biological sequence inspired by tools for modeling natural language have been shown to improve prediction of biological function [9, 45, 5].

A pressing and still poorly understood biological problem is understanding how rapidly mutating viral proteins can evade recognition by "escaping" the immune system's antibodies. Viral escape, which can be caused by even a single-residue change, has prevented the development of a universal antibody-based vaccine for influenza [30, 33] or human immunodeficiency virus (HIV) [6]. However, the rules governing viral fitness are complex and a biological experiment that empirically tests the

[1] Code at https://github.com/brianhie/mutational-semantics-neurips2020.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

escape potential of all mutations to all viral strains would be prohibitively expensive. A key concept underlying this study is that, in order to escape the immune system, a mutation must not only preserve viral infectivity (i.e., it must be "grammatical") but it must also be functionally altered so that it is no longer recognized by the immune system's antibodies (i.e., it must have substantial "semantic change").

Here, we introduce the problem of searching for sequence mutations based on both high semantic change and grammatical validity, which we call constrained semantic change search (CSCS). This is in contrast to settings concerned with semantic similarity search, rather than change. To gain intuition, we apply CSCS to natural language and, to demonstrate broader impact, we apply CSCS to predict viral escape. Our key contributions are **(1)** we introduce the CSCS problem formulation and show how learned language models offer a compelling solution with strong empirical results on both natural language and biological applications, suggesting that the distributional hypothesis from linguistics is also useful for modeling pathogenic evolution; **(2)** we develop an unsupervised neural language model for viral proteins and show that it learns semantically meaningful embeddings; and **(3)** we use CSCS for zero-shot prediction of escape mutations for influenza and for HIV with quantitative results much higher than baseline methods. To our knowledge, we present the first computational model that effectively predicts viral escape, potentially enabling vaccine or therapeutic design that anticipates escape before it occurs.

2 Methods

2.1 Problem Formulation

Intuitively, our goal is to identify mutations that induce high semantic change (e.g., a large impact on biological function) while being grammatically acceptable (e.g, biologically viable). More precisely, we are given a sequence of tokens defined as $\mathbf{x} \triangleq (x_1, ..., x_N)$ such that $x_i \in \mathcal{X}, i \in [N]$, where $\mathcal{X}$ is a finite alphabet (e.g., characters or words for natural language, or amino acids for protein sequence). Let $\tilde{x}_i$ denote a mutation at position i and the mutated sequence as $\mathbf{x}[\tilde{x}_i] \triangleq (..., x_{i-1}, \tilde{x}_i, x_{i+1}, ...)$.

We first require a semantic embedding $\mathbf{z} \triangleq f_s(\mathbf{x})$, where $f_s : \mathcal{X}^N \to \mathbb{R}^K$ embeds discrete-alphabet sequences into a continuous space, where, ideally, closeness in embedding space would correspond to semantic similarity. We denote semantic change as the distance in embedding space, i.e.,

$$\Delta\mathbf{z}[\tilde{x}_i] \triangleq \|\mathbf{z} - \mathbf{z}[\tilde{x}_i]\| = \|f_s(\mathbf{x}) - f_s(\mathbf{x}[\tilde{x}_i])\| \tag{1}$$

where $\|\cdot\|$ denotes a vector norm. The grammaticality of a mutation is described by

$$p(\tilde{x}_i|\mathbf{x}), \tag{2}$$

which takes values close to zero if $\mathbf{x}[\tilde{x}_i]$ is not grammatical and close to one if it is grammatical.

Our objective combines semantic change and grammaticality as a linear combination

$$a(\tilde{x}_i; \mathbf{x}) \triangleq \Delta\mathbf{z}[\tilde{x}_i] + \beta p(\tilde{x}_i|\mathbf{x})$$

for each possible mutation $\tilde{x}_i$ and a user-specified parameter $\beta \in [0, \infty)$. Mutations $\tilde{x}_i$ are prioritized based on $a(\tilde{x}_i; \mathbf{x})$. We refer to ranking mutations based on semantic change and grammaticality as CSCS.

2.2 Algorithms

2.2.1 Language Modeling

Algorithms for CSCS could potentially take many forms; for example, separate algorithms could be used to compute $\Delta\mathbf{z}[\tilde{x}_i]$ and $p(\tilde{x}_i|\mathbf{x})$ independently, or a two-step approach might be possible that computes one of the terms based on the value of the other.

Instead, we reasoned that a single approach could compute both terms simultaneously, based on learned language models that learn the probability distribution of a word given its context [38, 15, 43, 16, 44]. The language model we use throughout our experiments considers the full sequence context of a word and learns a latent variable probability distribution $\hat{p}$ and function $\hat{f}_s$, where, for all $i \in [N]$,

$$\hat{p}(x_i|\mathbf{x}_{[N]\setminus\{i\}}, \hat{\mathbf{z}}_i) = \hat{p}(x_i|\hat{\mathbf{z}}_i) \quad \text{and} \quad \hat{\mathbf{z}}_i = \hat{f}_s(\mathbf{x}_{[N]\setminus\{i\}}),$$

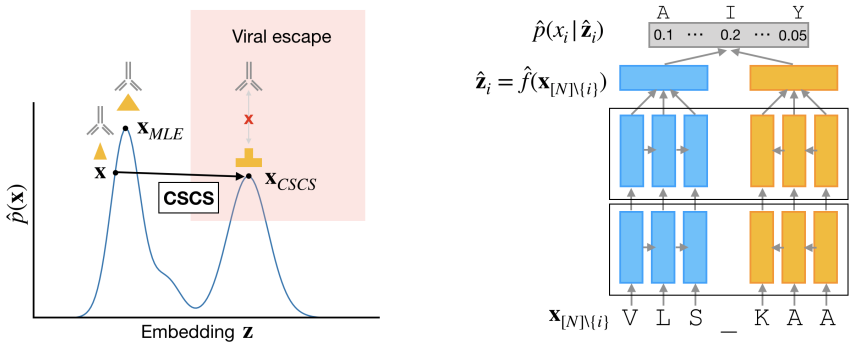

Figure 1: Constrained semantic change search (CSCS) for viral escape prediction. Left: Given an input sequence $\mathbf{x}$ and its semantics encoded by $\mathbf{z}$, CSCS aims to find a mutation to $\mathbf{x}_{\text{CSCS}}$ that causes the largest semantic change (high $\Delta\mathbf{z}$), while remaining grammatical (high $\hat{p}(\mathbf{x})$). Right: Language model architecture with two stacked BiLSTM layers instantiating the semantic embedding function $\hat{f}$, with the final language model output used as grammaticality.

i.e., latent variable $\hat{\mathbf{z}}_i$ encodes the context $\mathbf{x}_{[N]\setminus\{i\}} \triangleq (..., x_{i-1}, x_{i+1}, ...)$ such that x_i is conditionally independent of its context given the value of $\hat{\mathbf{z}}_i$.

We use different aspects of the language model to describe semantic change and grammaticality by setting terms (1) and (2) as

$$\Delta\mathbf{z}[\tilde{x}_i] \triangleq \|\hat{\mathbf{z}} - \hat{\mathbf{z}}[\tilde{x}_i]\|_1 \quad \text{and} \quad p(\tilde{x}_i|\mathbf{x}) \triangleq \hat{p}(\tilde{x}_i|\hat{\mathbf{z}}_i),$$

where $\hat{\mathbf{z}} \triangleq \begin{bmatrix}\hat{\mathbf{z}}_1^{\mathrm{T}} & \cdots & \hat{\mathbf{z}}_N^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ is the concatenation of embeddings for each token, $\hat{\mathbf{z}}[\tilde{x}_i]$ is defined similarly but for the mutated sequence, and $\|\cdot\|_1$ is the ℓ_1 norm, chosen because of more favorable properties compared to other standard distance metrics, though other metrics could be empirically quantified in future work [2].

Effectively, distances in embedding space are used to approximate semantic change and the emitted probability approximates grammaticality. We note that these modeling assumptions are not guaranteed to be perfectly specified, since, in the natural language setting for example, antonyms may also be close in embedding space and the language model output can also encode linguistic pragmatics in addition to grammaticality. However, we still find these modeling assumptions to have good empirical support.

Training or parameterizing the language model is separate from CSCS, and the novelty of CSCS is in leveraging these models in a new way. An advantage of this approach is that it does not require any bespoke modifications to the general language modeling framework, other than requiring a continuous latent variable. CSCS can therefore leverage the noted multitask generality of language models [44].

Importantly, this approach to CSCS is completely unsupervised. Rather than assume access to labels explicitly encoding semantics or grammaticality, the model instead extracts this information from a large unlabeled corpus. This is critical in domains, like viral genomics, in which large sequence corpuses are available but functional profiling is limited. These corpuses implicitly contain information related to grammaticality or infectivity (e.g., all sequences are grammatically acceptable or come from infectious virus), but the algorithm must learn these rules from data.

2.2.2 Architecture

Based on the success of recurrent architectures for protein-sequence representation learning [9, 45, 5], we use similar encoder models for viral protein sequences (**Figure 1**). Our model passes the full context sequence into bidirectional long-short-term-memory (BiLSTM) hidden layers. We used the concatenated output of the final LSTM layers as the semantic embedding, i.e.,

$$\hat{\mathbf{z}}_i \triangleq \left[\text{LSTM}_f(g_f(x_1, ..., x_{i-1}))^{\mathrm{T}} \quad \text{LSTM}_r(g_r(x_{i+1}, ..., x_N))^{\mathrm{T}}\right]^{\mathrm{T}}$$

where g_f is the output of the preceding forward-directed layer, LSTM_f is the final forward-directed LSTM layer, and g_r and LSTM_r are the corresponding reverse-directed components. The final output

probability is a softmax-transformed linear transformation of $\hat{\mathbf{z}}_i$, i.e.,

$$\hat{p}(x_i|\mathbf{x}_{[N]\setminus\{i\}}) \triangleq \text{softmax}(\mathbf{W}\hat{\mathbf{z}}_i + \mathbf{b})$$

for some learned model parameters $\mathbf{W}$ and $\mathbf{b}$. In our experiments, we used a 20-dimensional dense embedding for each element in the alphabet $\mathcal{X}$, two BiLSTM layers with 512 units, and categorical cross entropy loss optimized by Adam with a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Additional details on hyperparameter selection are given in Appendix 6.3.1.

2.2.3 Rank-Based Acquisition

Rather than acquiring mutations based on raw semantic change and grammaticality values, which may be on very different scales, we find that selecting β is much easier in practice when first rank-transforming the semantic change and grammaticality terms, i.e., acquiring based on

$$a'(\tilde{x}_i; \mathbf{x}) \triangleq \text{rank}(\Delta\mathbf{z}[\tilde{x}_i]) + \beta\,\text{rank}(p(\tilde{x}_i|\mathbf{x})).$$

All possible mutations $\tilde{x}_i$ are then given priority based on the corresponding values of $a'(\tilde{x}_i; \mathbf{x})$, from highest to lowest. Our empirical results have consistently good performance by simply setting $\beta = 1$ (equally weighting both terms), which we used in all experiments below unless otherwise noted. In this study, we deal with the unsupervised setting where β is a parameter but note that adding some supervision could learn β (or other, non-rank, transformations) from data.

2.2.4 Connection to Viral Escape

A language model is essentially a probability distribution over sequences learned from data. For any sequence $\mathbf{x}$, the model will output a predicted probability $p(\mathbf{x})$ of observing that sequence in the training data distribution. We call $p(\mathbf{x})$ "grammaticality" because in natural language tasks, $p(\mathbf{x})$ tends to be high for grammatically correct sentences. In the case of viral sequences, the training distribution consists of viral proteins that have evolved for high fitness/virality, so we hypothesize that high grammaticality corresponds to high viral fitness

However, high fitness alone does not indicate an escape mutation. A viral protein with a neutral mutation will have equally high fitness but will not look different to the immune system, i.e., it will have no "antigenic" change. To identify mutations that do lead to large antigenic changes, we use language models to internally compute a representation of the sequence, known as an "embedding", that they use to predict $p(\mathbf{x})$. If two sequences have similar embeddings, then they have similar distributions over sequence continuations given the input tokens. As a natural-language example, "the men advance", "the soldiers advance", and "the three advance" have a similar set of possible word continuations and would have similar embeddings, while "the cash advance" has a nearly disjoint set of continuations and thus a different embedding. We hypothesize that neutral mutations should not affect the distribution over amino acids at other positions, while mutations that affect antigenicity do affect the distribution over other positions. Thus, combining high sequence probability (high fitness) and a large change in embedding (antigenic change) indicates an escape mutation.

3 Related Work

The CSCS problem is related to work focused on identifying the best interventions to structured data to produce a desired outcome [40, 42]. Such work often assumes a dataset that includes both the observed features and corresponding outcomes, which allows for supervised learning. In contrast, we assume no explicit labels of semantic change and must resort to unsupervised learning to extract this information. This is because in domains like viral mutation, data that directly measures viral fitness is very limited, while unlabeled sequence data is abundant.

Importantly, our CSCS task is distinct from representation learning tasks that construct semantically meaningful embeddings, but CSCS does stand to benefit from innovation in representation learning. Using hidden states in a language model to represent natural language semantics has been an influential and productive idea [43]. Rather than acquiring mutations based on greatest semantic change as in CSCS, acquisition based instead only on lowest $\Delta\mathbf{z}[\tilde{x}_i]$ essentially performs semantic similarity search among all sequences that differ by a single token.

In biological applications, neural language models have been developed to learn unsupervised or weakly supervised protein sequence embeddings that encode generic protein similarity [9, 45, 5]. To

Original: `australian dead in bali` **Original**: `winegrowers revel in good season`
CSCS: `australian ballet in bali` **CSCS**: `winegrowers revel in flu season`
Original: `nauru bans transhipments to tackle overfishing`
CSCS: `nauru bans continue to tackle overfishing`

Figure 2: Example CSCS-proposed mutations to news headlines show large changes to the headline meaning or to the syntactic part-of-speech structure.

our knowledge, however, no previous work has considered how mutations affect these embeddings, nor have such methods been applied to evolutionary change. Furthermore, while many variants of recurrent or transformer-based architectures have been proposed for protein sequence modelling tasks, we note any such current or future language model architecture could be used in CSCS.

Some work in computational biology has focused on identifying deleterious mutations in human or mammalian genomes with clinical relevance [51, 46]. However, these approaches are based on direct supervision under the assumption that rare or poorly conserved mutations are deleterious. Such an assumption, however, does not apply to escape mutations, which could be both frequent or infrequent in a population. Viral genomes are also more highly variable than mammalian genomes (e.g., "Drake's rule"), so aligning mutations across viral strains is more difficult [20, 14, 48].

Most computational analyses specific to viral mutation require rich metadata beyond raw sequence or make virus-specific assumptions [8, 54] (for example, vaccine-related temporal patterns in influenza, which are absent for HIV). Most similar to our approach, models exist for learning viral fitness from a large sequence corpus [27, 26]. These approaches, however, requires time-consuming and error-prone multiple sequence alignment (MSA) preprocessing [29] and only consider pairwise information couplings among residues, which, as demonstrated below, limit performance when predicting escape. To our knowledge, our work is the first to effectively model viral escape that generalizes to any relevant genomic sequence from diverse viruses, without the need for sequence alignment, complex metadata, or special assumptions on mutational processes.

4 Results

To demonstrate how CSCS can alter semantics while preserving grammaticality, we gain intuition by first applying CSCS in a natural language setting before demonstrating broader impact by applying CSCS to biological sequence mutation in viruses. We find that CSCS-mutated headlines are semantically altered (quantified via changes in part-of-speech (POS) structure and distance in WordNet hierarchy) while remaining grammatical. Using a language model trained on a large corpus of influenza sequences, we find that CSCS-mutated viral sequences are predictive of escape mutations (i.e., "grammatical" mutations that preserve biological viability and infectivity but that also alter the protein's "semantics" thereby enabling escape from vaccines or treatments) that were identified by independent biological experiments. To assess generality, we perform this zero-shot escape prediction in two different influenza subtypes and in HIV.

4.1 News Headlines

Setup and Training Data. We sought to confirm our intuitions of "semantic change" and "grammaticality" by applying CSCS to single-word changes in news headlines. Our training corpus consisted of 1,186,018 headlines from the Australian Broadcasting Corporation from 2003 through 2019 (Appendix 6.1.1) [34].

Language Model Selection. We selected our model architecture by holding out a test set of headlines from 2016 onward (179,887 headlines, about 15%) and evaluating cross entropy loss for the language modeling task. We used a cross-validation strategy within the training set to grid search hyperparameters (Appendix 6.3.1). Our BiLSTM model with access to the full context (described above) obtained a training and test loss of 2.2 and 6.0, respectively. Performance decreased when replacing the LSTM hidden layers with densely-connected layers (train loss = 2.3, test loss = 7.2) or when removing access to the right context, i.e., a language model task $p(x_i|\mathbf{x}_{[i-1]})$ (train loss = 4.2, test loss = 6.5).

Table 1: Headline Semantic Change Results.

	Median % POS Change		Median WordNet Similarity	
Setting	NLTK	FLAIR	Pathwise	Wu-Palmer
Semantically closest (smallest $\Delta\mathbf{z}[\tilde{x}_i]$)	0.00%	0.00%	0.143	0.546
CSCS-proposed (highest $a'(\tilde{x}_i; \mathbf{x})$)	16.7%	14.3%	0.0833	0.235
two-sided t-test P	$<10^{-308}$	$<10^{-308}$	$<10^{-308}$	$<10^{-308}$

Table 2: Grammatical Acceptability Results

	Number Acceptable (Out of 300)			
Setting	Human 1	Human 2	Human Consensus	CSCS/Original Binomial P
CSCS-proposed ($\beta = 0.25$)	130	158	104	9.1×10^{-8}
CSCS-proposed ($\beta = 1$)	200	192	174	0.25
Original headline	223	233	197	N/A

Significant Semantic Change. For each headline, we considered all possible single-word mutations and picked the top according to the CSCS objective. Proposed mutations resulted in sentences that are qualitatively and quantitatively different than the original (**Figure 2**). CSCS often proposed word mutations that substantially change the part-of-speech (POS) structure. We quantified this observation by looking at the percentage of words in the mutated headline that had a different POS from the original headline. Using the NLTK POS tagger [10], the CSCS-proposed headline changed the POS of 16.7% of the words; using the FLAIR POS tagger [3], the median change was 14.3% of the words in the headline (**Table 1**). In contrast, the median POS change for the semantically-closest mutated headline (i.e., closest $\Delta\mathbf{z}[\tilde{x}_i]$) was 0% for both POS taggers (**Table 1**). Even when POS was not changed, CSCS proposed strikingly different word mutations, which we quantified using semantic similarity scores based on distance in the WordNet hierarchy [39, 28]. Specifically, for noun-to-noun and verb-to-verb changes, we selected the first WordNet synset corresponding to the depluralized or deconjugated version of the word. Across all these changes, the semantically-closest mutation had a median pathwise similarity of 0.14 and a median Wu-Palmer similarity [53] of 0.55 (both measures are between 0 and 1, inclusive, where 1 indicates high similarity, i.e., the same synset). In contrast, the CSCS-proposed mutation had a median pathwise similarity of 0.08 and median Wu-Palmer similarity of 0.24 (**Table 1**). Mean and standard deviation statistics, with similar trends, are also provided in **Table S1**. For both POS change and WordNet similarity, the difference between the CSCS-proposed and the semantically closest mutation are highly significant (two-sided independent t-test $P < 10^{-308}$). These results, supported by a qualitative examination of the changes (e.g., **Figure 2**), show that CSCS-mutated headlines are quite semantically different.

Grammaticality Preservation. We quantified grammaticality by asking human volunteers (12 in total) to provide grammatical acceptability labels. All humans were native English speakers with college degrees. Two humans were assigned to the same 150-headline text, blinded to the mutational status, and were asked to only evaluate grammaticality and not the content of the phrase, giving a binary "yes" or "no" label. Out of 300 original headlines, two humans provided a consensus "yes" grammatical label for 197 headlines (**Table 2**). The 300 corresponding CSCS-mutated headlines had 174 headlines with a consensus "yes" grammaticality; though lower, the number is within statistical error (two-sided binomial P-value of 0.25 compared with original). When we lowered β from 1 to 0.25, thereby reducing the influence of $\hat{p}(x_i|\hat{\mathbf{z}}_i)$, consensus grammaticality of the 300 CSCS-mutated headlines dropped significantly to 104 (binomial $P = 9.1 \times 10^{-8}$; **Table 2**). These results suggest that by considering $\hat{p}(x_i|\hat{\mathbf{z}}_i)$, CSCS can preserve grammaticality. In general, CSCS of natural language produces intuitively satisfactory results and may be relevant to work in computational humor [52].

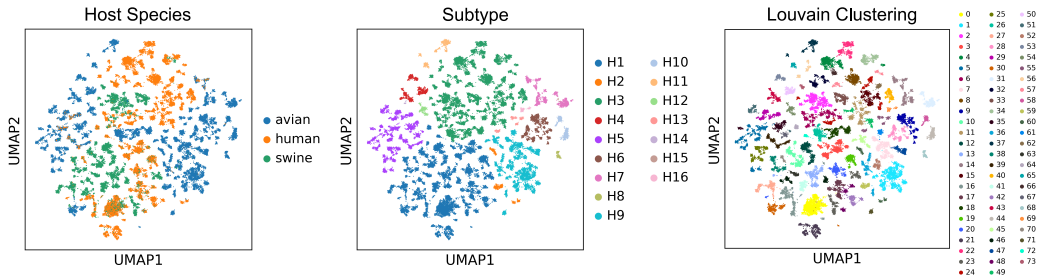

Figure 3: Semantic embedding space of influenza HA visualized in two-dimensions via UMAP [37] and colored by host species, subtype, or cluster labels from Louvain clustering [11].

4.2 Influenza

4.2.1 Language Model Training

Training Data and Model Selection. Our training data consists of 44,999 unique influenza A hemagglutinin (HA) amino acid sequences (around 550 residues in length) observed in animal hosts from 1908 through 2019. HA is a highly variable protein on the surface of influenza responsible for binding to host cells [24]. Since immunity to influenza is acquired by developing antibodies that bind and thereby neutralize HA, mutations to HA can lead to loss of immunity by reducing antibody binding affinity (i.e., immunological "escape") [30, 33]. Data was obtained from the NIAID Influenza Research Database (IRD) [55] through the web site at `http://www.fludb.org` (Appendix 6.1.2). These sequences were all obtained from animal hosts and thus, at least implicitly, encode viral viability and infectivity. We evaluated language model performance with a test set of held-out HA sequences where the first recorded date was before 1990 or after 2017, yielding a test set of 7,497 out of 44,999 sequences (about 17%). We again observed that a model with both an LSTM architecture and access to the full sequence context had the best train and test loss (Section 2.2.2).

Semantically Meaningful Embedding Structure. To improve our confidence that the embeddings are functionally meaningful, we leverage tools for unsupervised exploration of high-dimensional data. We trained our language model on the full IRD HA corpus, averaged $\hat{\mathbf{z}}_i$ across all residues in each sequence (to enable comparison across variable length sequences), and visualized the resulting embedding in two dimensions with Uniform Manifold Approximation and Projection (UMAP) [37, 17, 7].This results in clear structure corresponding to influenza subtype and host species (**Figure 3**), which we quantify via unsupervised Louvain clustering [11]. Within each cluser, on average, 99.8% of sequences come from a single influenza subtype and 96.2% come from a single host species, indicating high correspondence between semantic structure and biologically important metadata.

4.2.2 Zero-Shot Escape Prediction with CSCS

H3N2 Causal Escape Dataset. We validate the ability for CSCS to prioritize escape mutations using an interventional dataset by Lee et al., who made all possible single-residue mutations to HA from the `A/Perth/16/2009` (H3N2) strain and assessed which mutants preserve viral infectivity and induce escape [36]. To quantify escape, Lee et al. measured the overrepresentation of infectious viral sequences after immune selection by neutralizing human antibodies. These mutants therefore preserve infectivity and causally induce escape from neutralizing antibodies.

CSCS Enrichment of Acquired Escapes. Based on the language model trained over the full IRD HA corpus (Section 4.2.1), we computed $a'(\tilde{x}_i; \mathbf{x})$ for all possible single-residue mutations to the `A/Perth/16/2009` HA sequence. We emphasize that *none of these mutants were present in the training corpus*. The mutants identified by CSCS are substantially enriched for experimentally-verified escapes from Lee et al. [36], e.g., 4 out of the top 5 hits were confirmed escapes (**Figure 4**). We quantified enrichment by computing the area under the curve (AUC) obtained by plotting acquired escape mutations versus total acquired mutations based on $a'(\tilde{x}_i; \mathbf{x})$, normalized by the maximum area to produce a score between 0 and 1, inclusive, where 0.5 indicates the expected value of random guessing. The AUC obtained by the full CSCS objective is 0.771, compared to 0.709 when acquiring

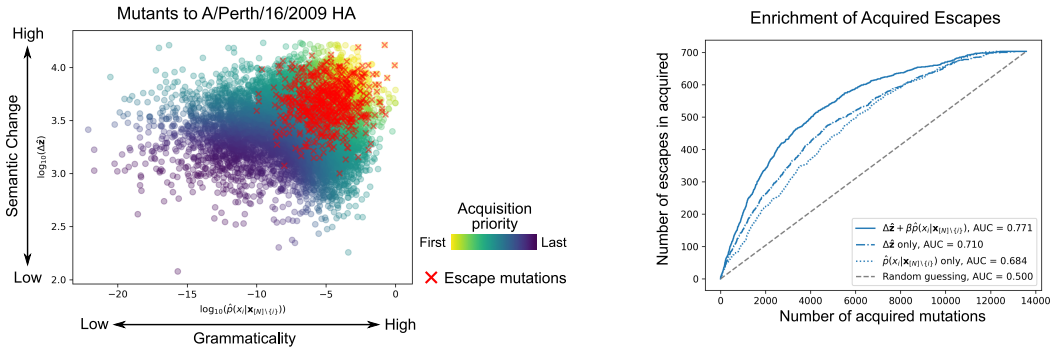

Figure 4: Left: Escape mutants (red Xs) to `A/Perth/16/2009` from Lee et al. [36] have high semantic change and grammaticality. Right: Escape mutants are substantially enriched in top CSCS-acquired mutants; see **Table 3**.

Table 3: Escape Prediction Results

	Normalized AUC		
Model	Influenza H1	Influenza H3	HIV Env
MAFFT MSA	0.697	0.598	0.523
EVcouplings (independent)	0.706	0.691	0.536
EVcouplings (epistatic)	0.726	0.687	0.552
$\Delta\mathbf{z}[\tilde{x}_i]$ alone	0.664	0.709	0.622
$p(\tilde{x}_i\|\mathbf{x})$ alone	0.820	0.684	0.667
CSCS ($\Delta\mathbf{z}[\tilde{x}_i]$ and $p(\tilde{x}_i\|\mathbf{x})$)	**0.834**	**0.771**	**0.692**

solely based on semantic change and 0.684 when acquiring solely based on grammaticality (**Figure 4**; **Table 3**), indicating that both are informative for determining escape. We obtained these results without direct supervision or explicit escape training data.

Benchmark of Existing Approaches. Though to our knowledge no previous method has been explicitly designed for escape prediction, we compare with standard viral fitness model strategies that are the closest to our unsupervised problem setting. The first strategy performs MSA of the viral sequence corpus and acquired escapes are simply those with the highest observed mutational frequency [12, 31, 4, 21]; our two benchmark methods that leverage this strategy are MAFFT MSA [29] and EVcouplings independent [26] (see Appendix 6.2.2 for more information). The second strategy also requires MSA followed by parameter estimation in a Potts model [27, 26], which incorporates pairwise residue information; we use the EVcouplings epistatic model that implements this approach, which is described in greater detail in Appendix 6.2.2. For influenza, we observed consistently higher AUCs obtained by CSCS over all benchmark methods (**Table 3**), noting that these methods were not specifically designed for viral escape prediction. We also tested pretrained protein sequence embedding models [9, 45, 5], not trained on viral corpuses, to see if their representations automatically transferred to viral escape prediction (Appendix 6.2.3), but this was not the case (**Table S2**), indicating that specific viral training data greatly improves escape prediction.

CSCS of H1N1 Viral Mutations. We evaluated CSCS on HA from another flu strain, `A/WSN/1933`, from a different HA subtype (H1 instead of H3) for which causal escape mutations were also determined by the same experimental procedure above, albeit with a more limited set of neutralizing antibodies [19]. Using the same language model trained on the IRD corpus, we ranked all possible single-residue mutations of `A/WSN/1933` HA, $\tilde{x}_i$, based on $a'(\tilde{x}_i; \mathbf{x})$. We again found substantial enrichment of escapes (observed in [19]) in the top mutations; the normalized AUC of acquired escape mutations versus total acquired mutations was 0.834 (**Table 3**). We note that none of these mutated sequences were present in the training data. In contrast, other approaches had lower enrichment of acquire escape mutants (normalized AUC $\leq$ 0.726; **Tables 3** and **S2**). Though similar causal escape

data is not available for other influenza strains, this additional validation increases our confidence that escape prediction with CSCS generalizes across strains.

4.3 HIV

Setup and Training Data. To assess generality to other viral proteins, we analyzed the HIV-1 Envelope (Env) protein, which, like influenza HA, is responsible for binding and entering host cells and is also targeted by antibodies [6]. Env is larger than influenza HA (about 850 residues compared to around 550) and more readily escapes immune selection due to viral mutation, even within the same host [47]. We train our language model on 60,857 unique Env sequences from the Los Alamos National Laboratory (LANL) HIV database (Appendix 6.1.3) [23]. We used the same language model architecture as in the influenza HA experiments. We again observed functionally-meaningful patterns when visualizing the semantic embeddings of Env sequences (**Figure S1**).

Zero-Shot Escape Prediction with CSCS. We applied CSCS to a dataset quantifying the infectivity and escape potential of all single-residue mutations to Env from the `BG505.T332N` strain of HIV, using a similar experimental procedure as that for HA from the two influenza strains described above [18]. We ranked all single-residue mutations $\tilde{x}_i$ of `BG505.T332N` by the CSCS objective $a'(\tilde{x}_i; \mathbf{x})$. We again observed enrichment of escape mutations when acquiring based on both semantic change and grammaticality, though with a weaker enrichment than observed for influenza HA (normalized AUC = 0.692; **Table 3**), suggesting that the semantic complexity of HIV Env might be more difficult to model with existing training data. However, CSCS escape prediction performance still exceeds that of other models (normalized AUC $\leq$ 0.574; **Tables 3** and **S2**).

5 Discussion

Here we show that a learning-based, distributional approach to modeling viral sequence achieves unprecedented insight into evolution and escape, suggesting a timely and important direction for the machine learning community. Excitingly, we demonstrate that the distributional hypothesis is a productive assumption for analysis of viral variation. This is not obvious, since it may be possible for non-causal mutations to widely co-occur with causal escape mutations [32], but our results suggest that many of the mutations that alter distributional structure are also causal escape mutations (perhaps due to pressure on viral sequences to maintain both diversity and economy, thereby diminishing the importance of non-causal mutants).

The CSCS problem in general is useful for any domain in which substantial functional change is desirable but the feature changes are limited or constrained. For example, in exploring differences in human-versus-machine perception, it may be desirable to generate entities that are perceived as similar by humans but as vastly different by algorithms, or vice versa. Though we focus on zero-shot, unsupervised escape prediction, some supervision could be useful in improving performance (e.g., learning β from a handful of examples).

A broader problem is in modeling other changes aside from mutations, like insertions and deletions, or more complex sequence changes. CSCS that accommodates insertions and deletions (about four times rarer than mutations in viruses [48]) could likewise model semantic change as a shift in the embedding space and grammaticality as some function of an emitted language model probability. While single-token changes allow for interpretability and efficiency, CSCS could be extended to multi-token changes (e.g., by combining the individual mutational probabilities to approximate the joint probability), though the search problem then becomes combinatorial. It may also be possible to evolve a sequence over multiple timesteps, each with a new single-token change, to produce complex sequence designs.

Broader Impact

We hope that this work leads to broad positive impact by (1) encouraging those in the machine learning community to contribute to understanding and combatting viruses (and infectious disease more broadly) and by (2) providing state-of-the-art prediction of how viruses can mutate around neutralization, which could be useful as part of rational design of vaccines or therapies. *In silico* models of how mutation leads to pathogenesis might help reduce both the resources and risks

associated with experimentally characterizing viral mutants. A primary goal of infectious disease research in general is to mitigate and prevent pandemic disease events among the global human population, which lead to widespread mortality, suffering, and economic disruption.

In computationally predicting mutations that induce escape or improve viral fitness, misuse could potentially take the form of using such methods to increase the pathogenicity of an existing viral strain. Experimental biologists, policy makers, and ethicists have already devoted and continue to devote a substantial amount of consideration to the ethics of such "gain-of-function" research (GOFR) [1, 49, 50]. As computational biologists become part of the GOFR conversation, attention to ethics is paramount and the scientific community should continue to preserve and strengthen the existing combination of experimental and policy safeguards.

Work in this area should continue to rely on direct experimental validation of computational prediction so that any system failures can be identified and corrected. Global viral surveillance already benefits from international cooperation through entities like the World Health Organization and collaborations like the Global Virome Project [13], and both the IRD and LANL HIV databases already have substantial global coverage across six continents [55, 23]. Preventing datasets from bias toward certain geographies or human populations underscores the already high priority given to viral monitoring at a global scale.

Acknowledgments and Disclosure of Funding

We thank Alejandro Balazs, Owen Leddy, Adam Lerer, Allen Lin, Adam Nitido, Uma Roy, and Aaron Schmidt for helpful discussions. We thank Steven Chun, Benjamin DeMeo, Ashwin Narayan, An Nguyen, Sarah Nyquist, and Alexander Wu for assistance with the manuscript. B.H. and E.Z. are partially funded by NIH grant R01 GM081871 (to B. Berger). B.H. is partially funded by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG). E.Z. is partially funded by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP).

References

[1] Doing Diligence to Assess the Risks and Benefits of Life Sciences Gain-of-Function Research. *The White House, President Barack Obama*, 2014.

[2] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *International conference on database theory*, volume 1973, pages 420–434. 2001.

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 54–59, 2019.

[4] T. M. Allen, M. Altfeld, S. C. Geer, E. T. Kalife, C. Moore, K. M. O'Sullivan, I. DeSouza, M. E. Feeney, R. L. Eldridge, E. L. Maier, D. E. Kaufmann, M. P. Lahaie, L. Reyor, G. Tanzi, M. N. Johnston, C. Brander, R. Draenert, J. K. Rockstroh, H. Jessen, E. S. Rosenberg, S. A. Mallal, and B. D. Walker. Selective Escape from CD8+ T-Cell Responses Represents a Major Driving Force of Human Immunodeficiency Virus Type 1 (HIV-1) Sequence Diversity and Reveals Constraints on HIV-1 Evolution. *Journal of Virology*, 79(21):13239–13249, 2005.

[5] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, 2019.

[6] Kathryn Twigg Arrildt, Sarah Beth Joseph, and Ronald Swanstrom. The HIV-1 Env protein: A coat of many colors. *Current HIV/AIDS Reports*, 9(1):52–63, 2012.

[7] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019.

[8] Trevor Bedford, Steven Riley, Ian G. Barr, Shobha Broor, Mandeep Chadha, Nancy J. Cox, Rodney S. Daniels, C. Palani Gunasekaran, Aeron C. Hurt, Anne Kelso, Alexander Klimov, Nicola S. Lewis, Xiyan Li, John W. McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner,

Derek J. Smith, Marc A. Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A. Russell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, 2015.

[9] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *7th International Conference on Learning Representations*, cs.LG:1902.08661, 2019.

[10] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[11] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[12] John A. Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, 2007.

[13] Dennis Carroll, Peter Daszak, Nathan D. Wolfe, George F. Gao, Carlos M. Morel, Subhash Morzaria, Ariel Pablos-Méndez, Oyewale Tomori, and Jonna A.K. Mazet. The Global Virome Project. *Science*, 359(6378):872–874, 2018.

[14] José M. Cuevas, Pilar Domingo-Calap, Marianoel Pereira-Gómez, and Rafael Sanjuán. Experimental Evolution and Population Genetics of RNA Viruses. *The Open Evolution Journal*, 3(1):9–16, 2009.

[15] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. *Advances in Neural Information Processing Systems*, pages 3079–3087, 2015.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, cs.CL(1810.04805), 2018.

[17] Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, 15(11):e1008432, 2019.

[18] Adam S. Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D. Bloom. An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes. *Immunity*, 50(2):520–532.e3, 2019.

[19] Michael B. Doud, Juhye M. Lee, and Jesse D. Bloom. How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Nature Communications*, 9(1):1386, 2018.

[20] J. W. Drake. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences of the United States of America*, 88(16):7160–7164, 1991.

[21] Guido Ferrari, Bette Korber, Nilu Goonetilleke, Michael K.P. Liu, Emma L. Turnbull, Jesus F. Salazar-Gonzalez, Natalie Hawkins, Steve Self, Sydeaka Watson, Michael R. Betts, Cynthia Gay, Kara McGhee, Pierre Pellegrino, Ian Williams, Georgia D. Tomaras, Barton F. Haynes, Clive M. Gray, Persephone Borrow, Mario Roederer, Andrew J. McMichael, and Kent J. Weinhold. Relationship between functional profile of HIV-1 specific CD8 T cells and epitope variability with the selection of escape mutants in acute HIV-1 infection. *PLoS Pathogens*, 7(2):e1001273, 2011.

[22] John Rupert Firth. *A Synopsis of Linguistic Theory, 1930-1955*. 1957.

[23] Brian Foley, Cristian Apetrei, Ilene Mizrachi, Andrew Rambaut, Bette Korber, Thomas Leitner, Beatrice Hahn, James Mullins, and Steven Wolinsky. HIV Sequence Compendium 2018. *HIV Sequence Compendium*, LA-UR 18-2, 2018.

[24] Steven J. Gamblin and John J. Skehel. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of Biological Chemistry*, 285(37):28403–28409, 2010.

[25] Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, 1954.

[26] Thomas A. Hopf, Anna G. Green, Benjamin Schubert, Sophia Mersmann, Charlotta P.I. Schärfe, John B. Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J. Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J. Draizen, Christian Dallago, Chris Sander, and Debora S. Marks. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, 2019.

[27] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017.

[28] Daniel Jurafsky and James Martin. *Speech and Language Processing*. Pearson Education, Inc., 2014.

[29] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[30] Hyunsuh Kim, Robert G. Webster, and Richard J. Webby. Influenza Virus: Dealing with a Drifting and Shifting Pathogen. *Viral Immunology*, 31(2):174–183, 2018.

[31] Björn F. Koel, David F. Burke, Theo M. Bestebroer, Stefan Van Der Vliet, Gerben C.M. Zondag, Gaby Vervaet, Eugene Skepner, Nicola S. Lewis, Monique I.J. Spronken, Colin A. Russell, Mikhail Y. Eropkin, Aeron C. Hurt, Ian G. Barr, Jan C. De Jong, Guus F. Rimmelzwaan, Albert D.M.E. Osterhaus, Ron A.M. Fouchier, and Derek J. Smith. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, 2013.

[32] Katia Koelle, Sarah Cobey, Bryan Grenfell, and Mercedes Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza a (H3N2) in humans. *Science*, 314(5807):1898–1903, 2006.

[33] Adam J. Kucharski, Justin Lessler, Jonathan M. Read, Huachen Zhu, Chao Qiang Jiang, Yi Guan, Derek A.T. Cummings, and Steven Riley. Estimating the Life Course of Influenza A(H3N2) Antibody Responses from Cross-Sectional Data. *PLoS Biology*, 13(3):e1002082, 2015.

[34] Rohit Kulkarni. A Million News Headlines. *kaggle*, therohk/mi, 2020.

[35] Thomas K. Landauer and Susan T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211, 1997.

[36] Juhye M. Lee, Rachel Eguia, Seth J. Zost, Saket Choudhary, Patrick C. Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C. Hurt, Seema S. Lakdawala, Scott E. Hensley, and Jesse D. Bloom. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *eLife*, 27(8):e49324, 2019.

[37] Leland McInnes and John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, stat.ML(1802.03426), 2018.

[38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[39] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[40] Jonas Mueller, David N. Reshef, George Du, and Tommi Jaakkola. Learning optimal interventions. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1039–1047, 2017.

[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing (EMNLP), Proceedings of the Conference*, 2014.

[42] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[43] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[45] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating Protein Transfer Learning with TAPE. *Advances in Neural Information Processing Systems*, pages 9686–9698, 2019.

[46] Philipp Rentzsch, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

[47] Douglas D. Richman, Terri Wrin, Susan J. Little, and Christos J. Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4144–4149, 2003.

[48] Rafael Sanjuán, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. Viral Mutation Rates. *Journal of Virology*, 84(19):9733–9748, 2010.

[49] S. Schultz-Cherry, R. J. Webby, R. G. Webster, A. Kelso, I. G. Barr, J. W. McCauley, R. S. Daniels, D. Wang, Y. Shu, E. Nobusawa, S. Itamura, M. Tashiro, Y. Harada, S. Watanabe, T. Odagiri, Z. Ye, G. Grohmann, R. Harvey, O. Engelhardt, D. Smith, K. Hamilton, F. Claes, and G. Dauphin. Influenza gain-of-function experiments: Their role in vaccine virus recommendation and pandemic preparedness. *mBio*, 5(6):e02430–14, 2014.

[50] Michael J. Selgelid. Gain-of-Function Research: Ethical Analysis. *Science and Engineering Ethics*, 22(4):923–964, 2016.

[51] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F. McRae, Yanjun Li, Jack A. Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, Jinbo Xu, Serafim Batzloglou, Xiaolin Li, and Kyle Kai How Farh. Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, 50(8):1161–1170, 2018.

[52] I. M. Suslov. How to realize "a sense of humour" in computers? *arXiv*, cs.CL(0711.3197), 2007.

[53] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, 1994.

[54] Rui Yin, Emil Luusua, Jan Dabrowski, Yu Zhang, and Chee Keong Kwoh. Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. *Bioinformatics*, page btaa050, 2020.

[55] Yun Zhang, Brian D. Aevermann, Tavis K. Anderson, David F. Burke, Gwenaelle Dauphin, Zhiping Gu, Sherry He, Sanjeev Kumar, Christopher N. Larsen, Alexandra J. Lee, Xiaomei Li, Catherine MacKen, Colin Mahaffey, Brett E. Pickett, Brian Reardon, Thomas Smith, Lucy Stewart, Christian Suloway, Guangyu Sun, Lei Tong, Amy L. Vincent, Bryan Walters, Sam Zaremba, Hongtao Zhao, Liwei Zhou, Christian Zmasek, Edward B. Klem, and Richard H. Scheuermann. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1):D466–D474, 2017.

6 Appendix

6.1 Additional dataset details

6.1.1 Headline dataset

Preprocessed headlines (stripped of punctuation, space-delimited, and lower-cased) from the Australian Broadcasting Corporation (early-2013 through the end of 2019) were obtained from `https://www.kaggle.com/therohk/million-headlines`.

6.1.2 Flu IRD dataset

Influenza HA amino acid sequences were downloaded from the "Protein Sequence Search" section of `https://www.fludb.org`. We only considered complete HA sequences from virus type A, but did not filter based on subtype, strain, date, host, geography, or country.

6.1.3 HIV LANL dataset

Sequences were downloaded from the "Sequence Search Interface" at `https://www.hiv.lanl.gov`. All complete HIV-1 Env sequences were downloaded, excluding sequences that the database had labeled as "problematic." To ensure that our sequences corresponded to complete viral haplotypes, we only considered sequences that had length between 800 and 900 amino acid residues, inclusive.

6.2 Additional baseline method details

We benchmark our escape prediction experiments against models that try to estimate the evolutionary fitness of a viral protein based on some assumptions. Notably, viral fitness models are not equivalent to escape prediction, since mutations that preserve fitness may be neutral with respect to escape (fitness models better correspond to the "grammaticality" term in CSCS). However, in the absence of existing unsupervised models that are directly built to perform unsupervised escape prediction, viral fitness models are the most related that attempt to solve a conceptually close problem.

6.2.1 Alignment-based frequency fitness model

This baseline model for viral fitness assumes that higher mutational frequencies in a corpus correspond to higher fitness and that residue-level fitness information is independent across the viral sequence; this fitness model is widely adopted due to its simplicity [12, 31, 4, 21].

We first perform MSA with the MAFFT software package (version 7.453) within the respective corpuses (influenza or HIV sequences). After sequence alignment was performed, we considered each position in the viral sequence of interest (influenza strains `A/Perth/16/2009` or `A/WSN/1933`, or HIV strain `BG505.T332N`). At a given position, we computed the frequency of other amino acids that were aligned to that position across all other sequences in the corpus. Sequences were acquired based on the highest observed frequencies across all possible single-residue mutations.

For influenza, we found that performance (in terms of normalized AUC) improved when restricting sequence alignment to the corresponding subtype (H1 sequences for `A/WSN/1933` and H3 sequences for `A/Perth/16/2009`) For HIV, we found that performance improved when only restricting alignments to the local neighborhood of `BG505.T332N`, defined by sequences that differ by a maximum of 15 residues. In general, we found that sequence alignment is dramatically affected by the sequences that are included in the corpus. For a best-case comparison, we report the highest performance over different sequence inclusion strategies.

We also used a conceptually similar implementation of this strategy provided by the EVcouplings pipeline [26] (`https://github.com/debbiemarkslab/EVcouplings`) using default parameters. We trained the EVcouplings independent model on the same corpus of viral sequences used to train our language models.

6.2.2 Alignment-based Potts model

A common critique of the above strategy for modelling viral fitness is that the independence assumption is limiting. Biologically, two residues can co-evolve, especially if they are physically and

biochemically related in the three-dimensional structure of the protein, a phenomenon referred to as "epistasis." A solution is to incorporate pairwise residue information by learning a probabilistic model in which each residue position corresponds to a random variable and pairwise potentials can encode epistatic relationships.

Hopf et al. learned such a model based on a Potts model formulation; we describe the general formulation here and leave implementation details to the original paper [27]. Given a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ where x_i comes from an alphabet $\mathcal{X}$ that is the set of all amino acids and a gap character, the model assigns an energy score to each sequence as

$$E(\mathbf{x}; \mathbf{h}, \mathbf{J}) \triangleq \sum_{i=1}^{N} h_i x_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} J_{ij} x_i x_j.$$

This term is scaled to be a valid probability distribution

$$p(\mathbf{x}; \mathbf{h}, \mathbf{J}) = \frac{1}{Z} \exp\{-E(\mathbf{x}; \mathbf{h}, \mathbf{J})\}$$

where $Z = \sum_{\mathbf{x}'} \exp\{-E(\mathbf{x}'; \mathbf{h}, \mathbf{J})\}$. The parameters are learned by a maximum likelihood procedure using a number of critical heuristics that Hopf et al. use to allow for efficient inference and parameter regularization [27, 26]. We use the pipeline provided by Hopf et al. at `https://github.com/debbiemarkslab/EVcouplings` with default parameters. We trained the EVcouplings epistatic model on the same corpus of viral sequences used to train our language models.

6.2.3 Pretrained sequence embedding models

We tested if the sequence embeddings produced by models trained on generic protein sequence corpuses [9, 45, 5] would be informative with respect to escape. We used the pretrained transformer model from Rao et al. [45] and the pretrained UniRep model from Alley et al. [5], both obtained through `https://github.com/songlab-cal/tape`. We used the pretrained model with full soft symmetric alignment and protein structure information from Bepler et al. [9], available through `https://github.com/tbepler/protein-sequence-embedding-iclr2019`. Rather than training exclusively on a large viral sequence corpus, as we did, these methods trained on corpuses containing generic protein sequences.

Each single-residue escape mutant was embedded using the pretrained model and mutant sequences were acquired based on the largest changes to the embedding based on the ℓ_1-distance. The results are provided in **Table S2**.

6.3 Additional experimental details

6.3.1 Language model hyperparameter selection

We performed a small-scale grid search using categorical cross entropy loss after 20 training epochs on the headline and influenza datasets to select the language model architecture and hyperparameters based on a random 80%/20% cross-validation split of the training set. Hyperparameter ranges were influenced by previous applications of recurrent architectures to protein sequence representation learning [9]. We tested hidden unit dimensions of 128, 256, and 512. We tested architectures with one or two hidden layers. We tested three hidden-layer architectures: a densely connected neural network with access to both left and right sequence contexts, an LSTM with access to only the left context, and a BiLSTM with access to both left and right sequence contexts. We tested two Adam learning rates (0.01 and 0.001). All other architecture details described in Section 2.2.2 were fixed to reasonable defaults. In total, we tested 36 conditions and ultimately used a BiLSTM architecture with two hidden layers of 512 hidden units each, with an Adam learning rate of 0.001. We used the same architecture for all experiments. In general, we noted that increasing model capacity only served to improve performance.

6.3.2 Headline semantic change quantification

POS tagging was done using the English `pos_tag()` function with default parameters from the `nltk` Python package (`https://www.nltk.org`) and separately using the default POS `SequenceTagger` from `flair` (`https://github.com/flairNLP/flair`).

CSCS-mutated words were compared to the original word based on WordNet synset similarities. We only considered words where the POS (labeled by `nltk`) was preserved, where the POS was a noun or a verb (i.e., `NN`, `NNS`, or `VB`), and where the depluralized or deconjugated word was present in `nltk`'s WordNet. We used the `pattern` Python package (`https://github.com/clips/pattern`) to depluralize words or to conjugate verbs into the infinitive form.

6.3.3 Computational resources

Training on the influenza HA dataset requires approximately a week of training and around three hours to evaluate all possible single escape sequences. On our largest dataset (HIV Env), our training implementation finished within 2.5 weeks and escape prediction inference requires eight hours. Models were trained with an Nvidia Tesla V100 PCIe 32GB GPU. Experiments were run with Python 3.7 on Ubuntu 18.04.

6.3.4 Code and data availability

Code and datasets used in this paper's experiments has been made available as supplementary data.

Table S1: Additional Headline Semantic Change Results

	Mean ± S.Dev. % POS Change		Mean ± S.Dev. WordNet Similarity	
Setting	NLTK	FLAIR	Pathwise	Wu-Palmer
Semantically closest (smallest $\Delta\mathbf{z}[\tilde{x}_i]$)	8.40% ± 13.3%	5.64% ± 10.5%	0.266 ± 0.280	0.536 ± 0.298
CSCS-proposed (highest $a'(\tilde{x}_i; \mathbf{x})$)	18.9% ± 15.3%	15.5% ± 14.3%	0.0833 ± 0.0756	0.235 ± 0.145

Table S2: Additional Escape Prediction Results (pretrained sequence embeddings)

	Normalized AUC		
Model	Influenza H1	Influenza H3	HIV Env
Alley et al. pretrained $\Delta\mathbf{z}$ [5]	0.482	0.452	0.534
Bepler et al. pretrained $\Delta\mathbf{z}$ [9]	0.660	0.644	0.561
Rao et al. pretrained $\Delta\mathbf{z}$ [45]	0.584	0.526	0.574
CSCS ($\Delta\mathbf{z}[\tilde{x}_i]$ and $p(\tilde{x}_i\|\mathbf{x})$)	**0.834**	**0.771**	**0.692**

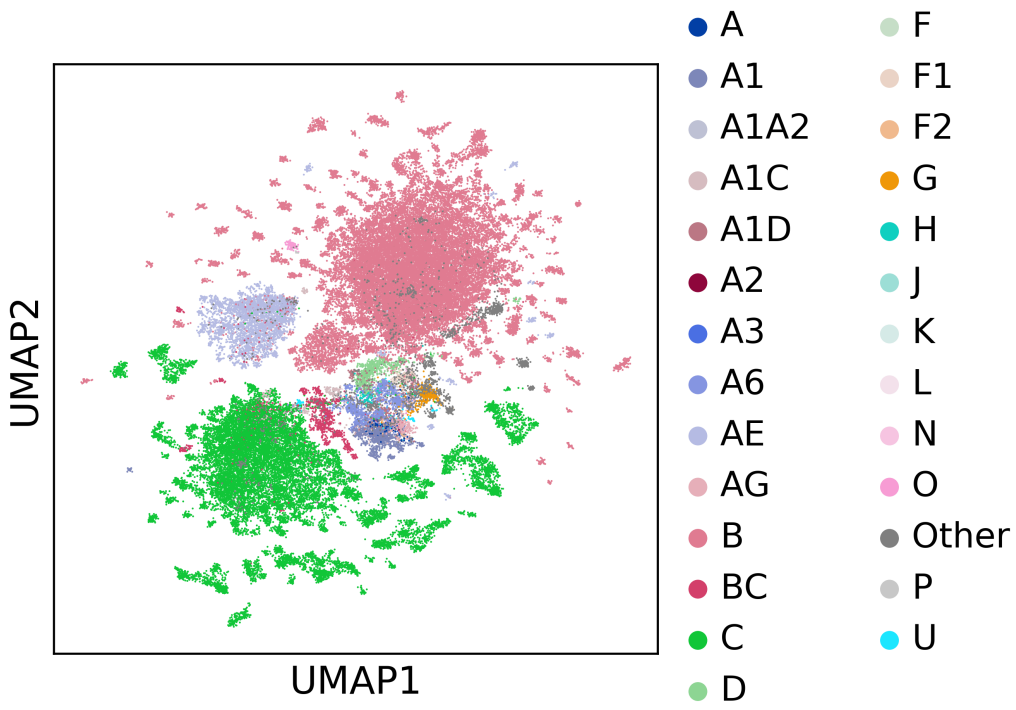

Figure S1: UMAP visualization of unique HIV Env sequences colored by subtype. Large, dominating clusters corresponding to B, C, and AE subtypes may be due to the lack of vaccine pressure on HIV, compared to influenza.