Language Models Predict Drug Resistance from Complex Sequence Variation

by

Andy Tso

S.B. Computer Science and Engineering, Massachusetts Institute of Technology, 2020

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

© Massachusetts Institute of Technology 2021. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2021

Certified by .....
Bonnie Berger
Simons Professor of Mathematics
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Master of Engineering Thesis Committee

Language Models Predict Drug Resistance from Complex Sequence Variation

by

Andy Tso

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Mutation in viruses and bacteria presents a major barrier to the development of vaccines, antiviral drugs, and antibiotics. Recently, neural language models trained on viral protein sequence evolution have shown promise in their ability to predict viral escape mutations, potentially enabling more intelligent therapeutic design [6]. Hie et al.'s work puts forth the key conceptual advance that viral escape from human immunity occurs in the event of a mutation which simultaneously generates meaningful antigenic change while also preserving viral fitness. These ideas are analogous to the semantics and grammar of a language.

Theoretically, mutations that confer high semantic change while preserving high grammaticality may also be predictive of resistance to other types of evolutionary pressure as well. In this thesis, we show that language modeling of protein evolution can also predict mutations that confer drug resistance. We validate our language model predictions using known drug resistance mutations in HIV-1 protease and reverse transcriptase proteins and *Escherichia coli* beta-lactamase protein. Our results suggest a way to identify and potentially anticipate drug resistance mutations that generalizes across viruses and bacteria.

Thesis Supervisor: Bonnie Berger
Title: Simons Professor of Mathematics

Acknowledgments

Without the guidance and inspiration of several individuals, this work would not have been possible. I would first and foremost like to thank my thesis advisor, Dr. Bonnie Berger, for her continued guidance and unwavering support throughout my entire time with the Berger lab.

Special thanks to Brian Hie for his constant guidance and thoughtful commentary throughout the entire research process, which has helped to elevate the project to its current form. Brian's steadfast positivity and dedicated mentorship enabled me to innovatively think and grow as a researcher. I would also like to thank Ellen Zhong for her ongoing mentorship and support ever since I first joined the Berger lab. It was her strong mentorship that gave me the inspiration to continue to pursue research in computational biology.

Finally, I would like to thank my family for their endless love, support, and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mutation in pathogenic viruses and bacteria present a major barrier against the development of reliable vaccines and therapies. For example, viral mutation that escapes immunity presents a challenge to vaccine development and bacterial resistance to antibiotics threatens us with potentially untreatable infections. Unable to reliably depend on our current suite of antiviral and antibiotic drugs, we must seek to better understand drug resistance and potentially predict escape mutations within disease-causing viruses and bacteria before they become major problems (e.g., pandemics involving untreatable and fatal infectious diseases). Being able to predict the mutations which would allow viruses to escape human immunity and bacteria to develop drug resistance could serve as a valuable tool to intelligent therapeutic design.

Attempting to understand the mechanisms for drug resistance through wetlab experiments is often time-consuming because of the high complexity of biological systems. Furthermore, methods developed to understand one strain of a virus might not generalize to other viral strains or even other viral species or other infectious agents like bacteria. High throughput techniques such as deep mutational scans (DMS) have been developed to profile the resulting phenotypic changes from single-residue mutations, ignoring the prohibitively large set of combinatorial mutations possible. However, even a simple DMS which neglects combinatorial mutations requires a large amount of time and effort.

To this end, we seek more efficient means of studying escape through computa-

tional models. We draw inspiration from recent work with neural language models which shows that it is possible to extract functional information from sequences alone, thus, demonstrating that computation can be leveraged as a promising way to efficiently study viral escape [6].

As in all natural languages, there exists a set of rules (e.g., the grammar) which a linear sequence of tokens must obey. The encoding of complex ideas is the semantics of the language. Hie et al. posit that in the language of viral amino-acid sequences, grammaticality corresponds to and quantifies the virus's viability and infectivity, while semantic change corresponds to a different protein capable of escaping current drugs and human immunity [6]. The key idea is that in order to arrive at viral escape, a mutation must generate an amino acid sequence with both high grammaticality and semantic change. The task of searching for such a mutation is called a "constrained semantic change search" (CSCS) [6]. Previously, CSCS was used to predict escape mutations in influenza hemagglutinin, HIV Env, and SARS-CoV-2 Spike.

1.1 Related Work

The goal of this thesis is to understand the genrality of the language models to different evolutionary settings by showing how CSCS can model not only viral escape from vaccines but also resistance to artificial drug selection. To this end, we are interested in conducting empirical studies, applying the CSCS to predict drug resistance mutations in reverse transcriptase and protease within HIV and beta-lactamase protein in *Escherichia coli* (E. coli) bacteria.

The idea that co-occurrence patterns encode semantics, the distributional hypothesis, has been the basis of the practical success of language models first in natural language [15, 10] and more recently in biology. In biology in particular, we have seen recent success in using recurrent architectures for protein-sequence representation learning [1, 2, 16]. The main inspiration for this thesis is the recent work by Hie et al., which shows a novel use of neural language models to learn both semantic change and grammaticality to enable the prediction of viral escape mutations [6]. The key

idea contributed is that grammaticality and semantic change are analogous to and measure viral fitness and antigenic change, respectively.

Previous non-language model attempts have been made to predict virus mutations through statistical learning as well as neural networks [3, 19, 7]. Methods have also been developed to model and predict antibiotic resistance of nontyphoidal Salmonella [11]. We specifically compare the CSCS algorithm with other attempts to predict mutational effects on protein fitness using sequence variation alone. These methods include EVCouplings (independent and epistatic models) and a naive empirical mutation frequency-based approach following sequence alignment via MAFFT to understand which approach is most successful in predicting escape mutations [7, 8].

Chapter 2

Methods

2.1 CSCS: Problem Specification

The problem formulation is as in Hie et al., 2020 [6].

Let $\mathcal{X}$ be a finite alphabet which represents the set of all amino acids. Given

$$\mathbf{x} \triangleq (x_1, \cdots, x_N) \in \mathcal{X}^N \tag{2.1}$$

the wild-type protein sequence of N amino acids for a particular viral or bacterial protein, we are interested in finding the set of mutations which are most likely to result in escape.

Restricting our attention to single-token mutations, let $m_{i,x'} : \mathcal{X}^N \to \mathcal{X}^N$ be the mutation function which applies a substitution mutation of $x' \in \mathcal{X}$ at position $i \in [N]$ to its input protein sequence. As such, $m_{i,x'}(\mathbf{x})$ is a sequence produced as a result of a particular single-token substitution mutation based on $\mathbf{x}$:

$$m_{i,x'} = (x_1, \cdots, x_{i-1}, x', x_{i+1}, \cdots, x_N) \tag{2.2}$$

To quantify the semantics of a particular mutation, we first introduce embedding function $f_s : \mathcal{X}^N \to \mathbb{R}^K$ which embeds from the discrete space of token sequences to a K-dimensional continuous space, capturing the semantics of sequences. In particular,

the embedding function f_s is designed such that semantically similar sequences are close in a geometric sense with respect to a vector norm $\|\cdot\|$. Thus, we can interpret

$$\Delta \mathbf{z}_{i,x'}(\mathbf{x}) \triangleq \|f_s(\mathbf{x}) - f_s(m_{i,x'}(\mathbf{x}))\| \tag{2.3}$$

as the semantic change associated with applying mutation $m_{i,x'}(\cdot)$ to original protein sequence $\mathbf{x}$.

Next, we define grammaticality which serves as a proxy for biological viability. With $\mathbf{x}_{-i} \triangleq (x_1, \cdots, x_{i-1}, -, x_{i+1}, \cdots, x_N)$ denoting the context surrounding position i in $\mathbf{x}$, we say that the grammaticality of a mutation is the probability associated with assigning token $x' \in \mathcal{X}$ to $\mathbf{x}_i$ conditional on context $\mathbf{x}_{-i}$

$$p_{\mathbf{x}_i|\mathbf{x}_{-i}}(x'|\mathbf{x}_{-i}) \tag{2.4}$$

Our objective function jointly optimizes an additive function of rank-normalized semantic change (2.3) and grammaticality (2.4) by assigning the score

$$a_{\mathbf{x}}(i, x') \triangleq \text{rank}(\Delta \mathbf{z}_{i,x'}(\mathbf{x})) + \beta \cdot \text{rank}(p_{\mathbf{x}_i|\mathbf{x}_{-i}}(x'|\mathbf{x}_{-i})) \tag{2.5}$$

to mutation $m_{i,x'}(\cdot)$ applied to original sequence $\mathbf{x}$ which induces semantic change $\Delta \mathbf{z}_{i,x'}(\mathbf{x})$ and maintains grammaticality $p_{\mathbf{x}_i|\mathbf{x}_{-i}}(x'|\mathbf{x}_{-i})$, for some parameter $\beta \in [0, \infty)$.

2.2 CSCS: Algorithm Overview

As proposed by Hie et al., we consider learning a language model which emits probability distributions over missing tokens given the surrounding context over all contexts [6, 10, 15, 4, 5, 13].

This is accomplished through first learning an embedding $\hat{f}_s(\cdot)$ to generate a latent variable $\hat{\mathbf{z}}_{-i} = \hat{f}_s(\mathbf{x}_{-i})$ such that the latent variable captures all the relevant context information. More precisely, we have Markov structure $\mathbf{x}_i \leftrightarrow \hat{\mathbf{z}}_{-i} \leftrightarrow \mathbf{x}_{-i}$ expressing conditional independence of the missing token's distribution from the context given the latent variable. Therefore, we can write

$$\hat{p}_{\mathbf{x}_i|\mathbf{x}_{-i},\hat{\mathbf{z}}_{-i}}(\cdot|\mathbf{x}_{-i},\hat{\mathbf{z}}_{-i}) = \hat{p}_{\mathbf{x}_i|\hat{\mathbf{z}}_{-i}}(\cdot|\hat{\mathbf{z}}_{-i}) \tag{2.6}$$

Next, we compute the semantic change of applying mutation $m_{i,x'}(\cdot)$ on sequence $\mathbf{x} \in \mathcal{X}^N$ as

$$\Delta\hat{\mathbf{z}}_{i,x'}(\mathbf{x}) \triangleq \|\hat{\mathbf{z}}(\mathbf{x}) - \hat{\mathbf{z}}(m_{i,x'}(\mathbf{x}))\|_1 \tag{2.7}$$

where

$$\hat{\mathbf{z}}(\mathbf{x}) \triangleq \frac{1}{N}\sum_{i=1}^{N} \hat{f}_s(\mathbf{x}_{-i}) \tag{2.8}$$

is the average embedding of contexts of $\mathbf{x}$ across all positions and $\|\cdot\|_1$ denotes L_1 norm so that embeddings spatially encode semantic similarity and differences.

Grammaticality is analogous to (2.4) where we learn distribution $\hat{p}$ so that

$$p_{\mathbf{x}_i|\mathbf{x}_{-i}}(\cdot|\mathbf{x}_{-i}) = \hat{p}_{\mathbf{x}_i|\hat{\mathbf{z}}_{-i}}(\cdot|\hat{f}_s(\mathbf{x}_{-i})) \tag{2.9}$$

encodes grammaticality.

We then utilize a bidirectional-LSTM as the encoder model for our semantic embedding of context $\mathbf{x}_{-i}$ [6].

$$\hat{\mathbf{z}}_{-i} = \hat{f}_s(\mathbf{x}_{-i}) = \begin{bmatrix} \text{LSTM}_f(g_f(x_1, \cdots, x_{i-1})) \\ \vdots \\ \text{LSTM}_r(g_r(x_{i+1}, \cdots, x_N)) \end{bmatrix} \in \mathbb{R}^K \tag{2.10}$$

with g_f and g_r being the outputs of the previous feed-forward and subsequent reverse-directed layers respectively; while LSTM_f and LSTM_r denote the final forward directed and final reverse-directed components.

Finally, the grammaticality distribution is obtained via

$$p_{\mathbf{x}_i|\mathbf{x}_{-i}}(\cdot|\mathbf{x}_{-i}) = \text{softmax}(\mathbf{W}\hat{\mathbf{z}}_{-i} + \mathbf{b}) \tag{2.11}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{X}|\times K}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{X}|}$ are learned model parameters.

With these components, our objective function analogous to (2.5) seeks to minimize

$$a_{\mathbf{x}}(i, x') \triangleq \text{rank}(\Delta\hat{\mathbf{z}}_{i,x'}(\mathbf{x})) + \beta \cdot \text{rank}(\hat{p}_{\mathbf{x}_i|\mathbf{x}_{-i}}(x'|\mathbf{x}_{-i})) \tag{2.12}$$

Alternatively, we can interpret score as a statistic which we can threshold for our binary hypothesis testing. That is for some threshold $\gamma \in [0, \infty)$ we decide whether or not a particular mutation is an escape via its score

$$\hat{H}(\mathbf{x}, (i, x')) = \begin{cases} \text{escape} & a_{\mathbf{x}}(i, x') \leq \gamma \\ \text{non-escape} & \text{otherwise} \end{cases}$$

which traces out the ROC curve for our decision rule.

2.3 Architecture Details

As in Hie et al., 2020, we use dense embedding to map each token in the alphabet $\mathcal{X}$ to a 20-dimensional point. There are two BiLSTM layers each with 512 units. Finally, we minimize categorical loss via Adam with the following parameters:

- learning rate 0.001 for all models except for the model associated with UniProt's beta-lactamase which utilized learning rate 0.0001, which we lowered due to observed instability during model training
- $\beta_1 = 0.9$
- $\beta_2 = 0.999$

Chapter 3

Results

3.1 Biologically meaningful semantic landscapes

We interpret the semantic embeddings learned by the language model by visualizing the embeddings in two dimensions Uniform Manifold Approximation and Projection (UMAP), which approximates the high-dimensional nearest-neighbor relationships in a low-dimensional space [9]. Overall, we find that the language model semantic embedding meaningfully captures various biologically relevant aspects of sequence variation.

For example, we will see below that the embedding preserves high level taxonomic information (e.g, taxonomic phylum) as well as protein class labels among various beta-lactamase sequences. Consistent with Hie et al.'s work, we also confirm that the embedding landscape also captures viral subtype information, with evident clustering of sequences with respect to the virus subtype in the HIV *pol* gene as well as individually, the subsequences of *pol* encoding the protease and reverse transcriptase proteins (Figure 3-1) [6]. Finally, with specific relevance to mutations which enable escape from antiviral drugs, we find that the semantic embedding landscape reliably capture's the drug treatment history of the patient from which the HIV sequence was retrieved (Figure 3-4). All of these point to evidence that it is indeed possible to learn semantically meaningful information from sequence data alone in this unsupervised learning setting [6].

3.1.1 Subtypes of HIV

In UMAP embedding space, we observe clustering of sequences by the subtype of the HIV virus from which the sequence was obtained, suggesting that the model had successfully learned subtype structure (Figure 3-1).

Most notably, subtypes B, C, and AE are consistently the largest main clusterings in all embeddings which are respectively the most common subtypes in

- Europe, the Americas, Japan, and Australia (subtype B)
- Southern Africa, Eastern Africa, India, Nepal, and parts of China (subtype C)
- AE represents a circulating recombinant form (CRF) prevalent in India

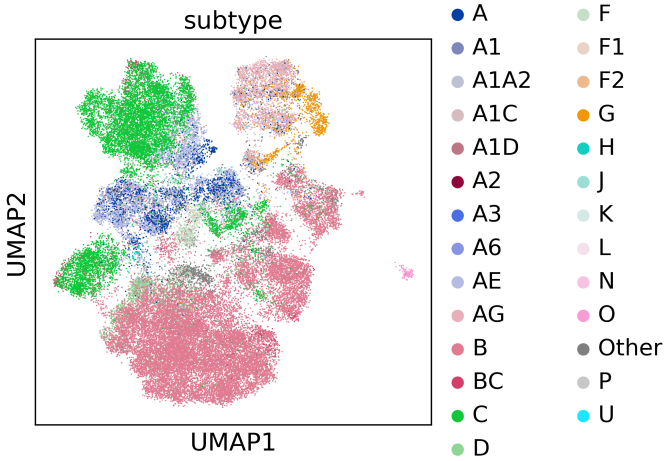
An interesting hypothesis is whether the landscape can be slightly modified to recapitulate to the geographic world map where the respective strains are most prevalent [12]. However, it is unclear from the given data as we clearly only have three distinctly large clusters.

3.1.2 Phylogeny of Bacteria

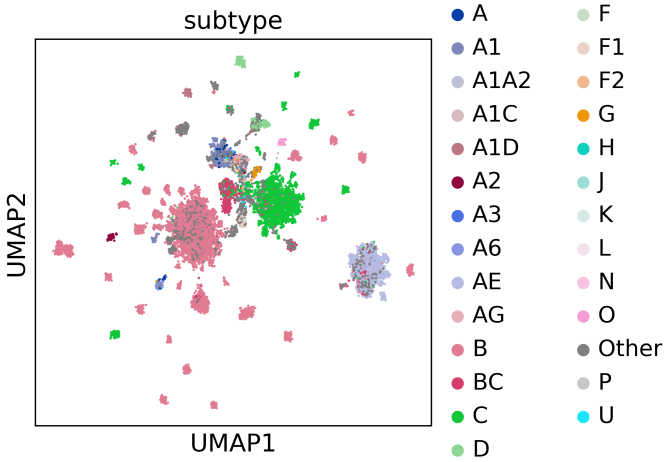
At a high level, via beta-lactamase protein sequences from many bacterial species, our model learns a semantic embedding which encapsulates phylogenetic structure. Specifically, clusters are seen in the spatial arrangement of sequences when visualized in the first two UMAP components that correspond to the phylum of the bacteria from which the sequence came from (Figure 3-2). We include the highest level snapshot we could obtain by choosing phylum, but we can also verify that at finer levels of phylogenetic classification, the semantic embedding landscape emits clusters which are yet further subdivided as one would expect.

3.1.3 Classes of Beta-Lactamase

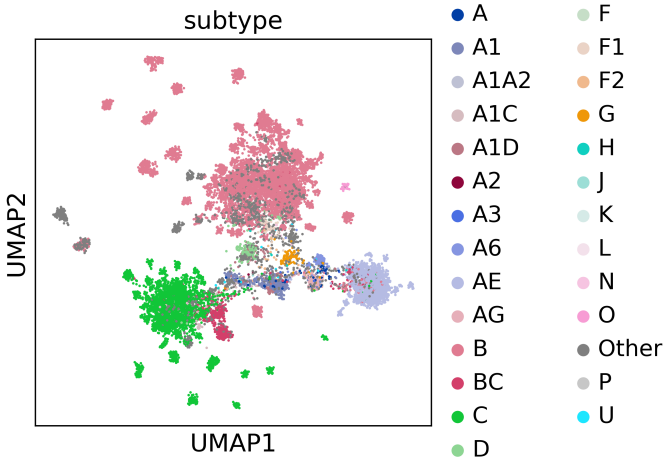
Mutations in bacterial beta-lactamase are a major threat to our use of beta-lactam based antibiotics to treat bacterially-caused infectious diseases. In fact, our heavy use



(a) HIV Protease



(b) HIV Reverse Transcriptase



(c) HIV *pol* gene

Figure 3-1: Subtype Landscapes of HIV

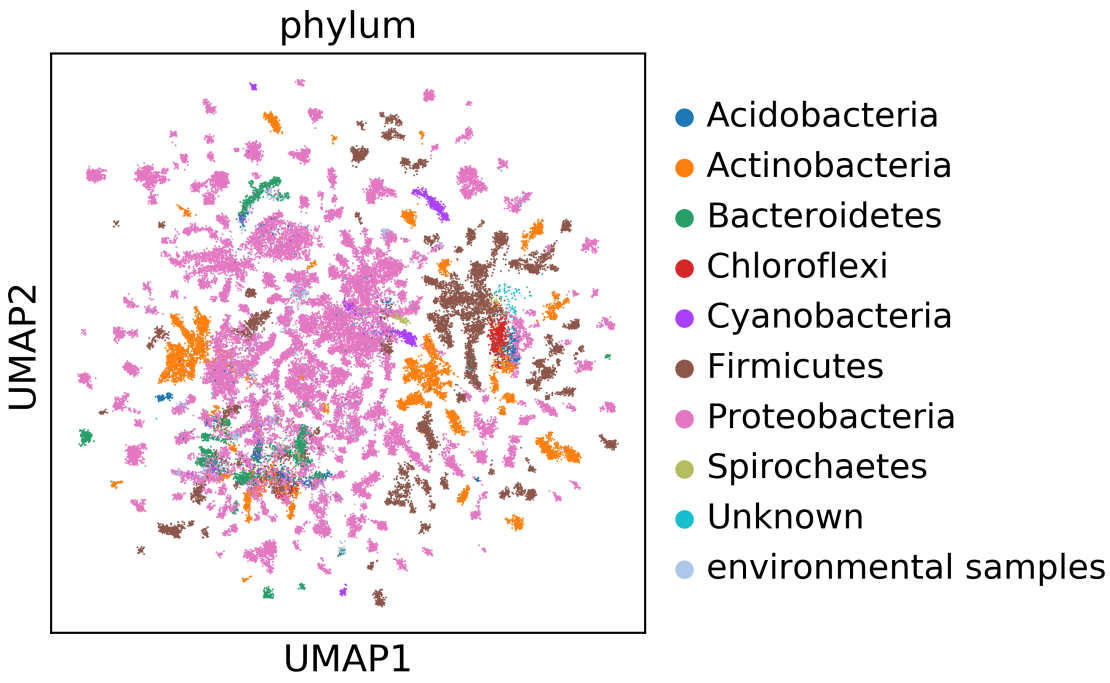


Figure 3-2: Phylums of UniProt database beta-lactamase sequences

of beta-lactam based antibiotics such as Penicillin has resulted in the rapid mutation of the beta-lactamase gene which confers resistance against these drugs, developing multiple classes of the protein across bacterial species [14].

In our experiments, the model successfully learns to embed protein sequences in a way that captures the specific subclass of the sequence (Figure 3-3).

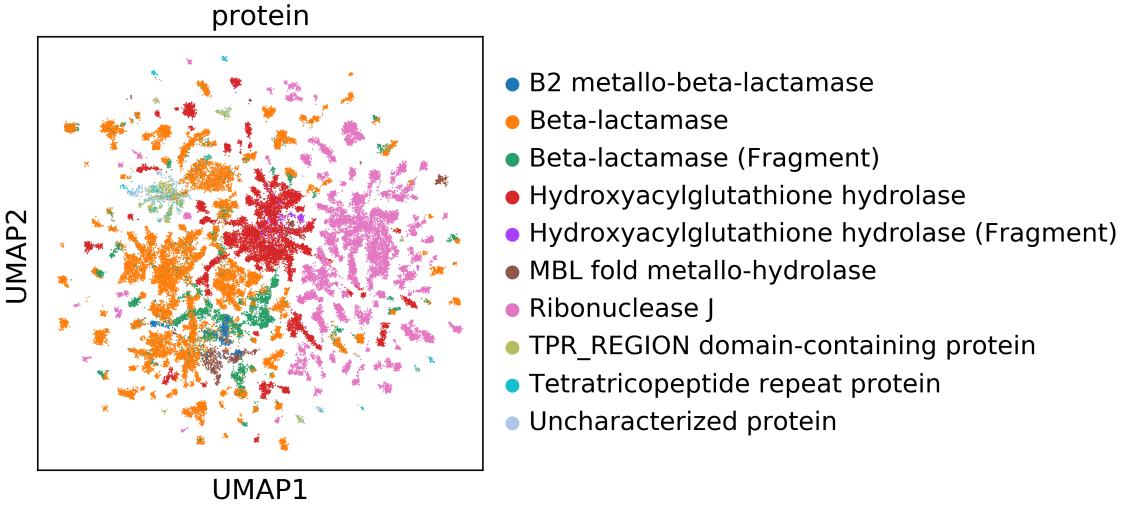


Figure 3-3: Protein class of UniProt database beta-lactamase sequences

3.1.4 Drug Treatment History

In the LANL HIV datasets, viral sequences come from patients undergoing treatment. Each of these sequences comes tagged with metadata detailing the set of drug treatments the patient has undergone prior to viral sequencing. With the assumption that patient revisits are in part due to remaining symptoms after treatment, it is likely that the drug resistance of a particular sequence is correlated with the drug treatment history appearing in its metadata tag. There is a slight spatial structure with clustering seen among the drugs Lopinavir and protease-inhibitor, with insufficient data from other classes of drugs to make strong conclusions.

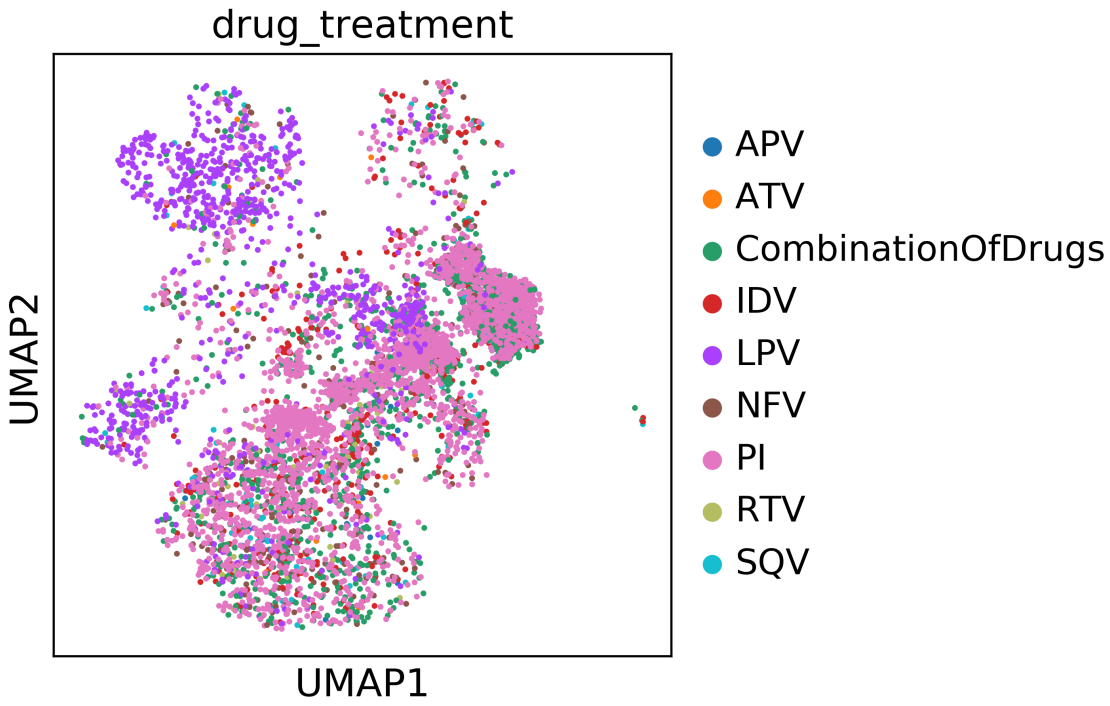


Figure 3-4: Protein class of UniProt database beta-lactamase sequences

3.2 HIV Protease and Reverse Transcriptase

We first train our two BiLSTM models on unaligned sequences of protease and reverse transcriptase (73,110 and 6,164 unique sequences respectively). Following sequence alignment, we perform single-residue escape mutation prediction via CSCS and find

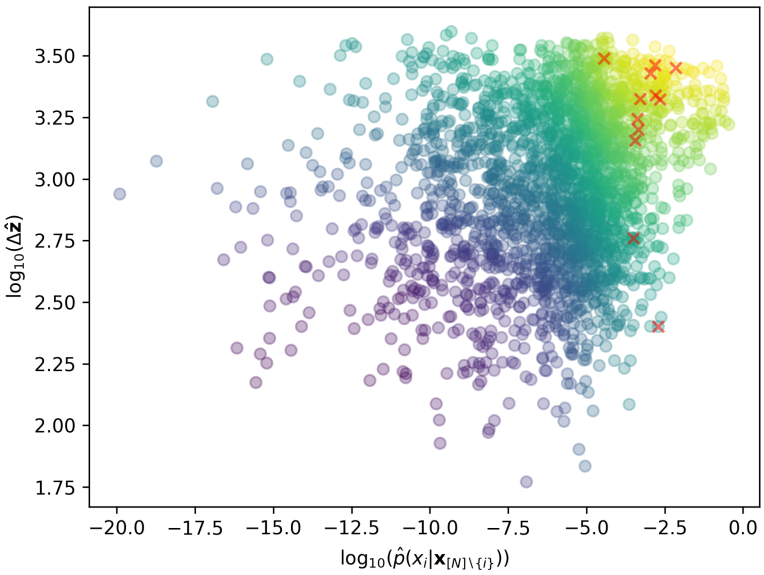
that regions marked with a high priority for acquisition based on high rank-normalized grammaticality and semantic change scores contain the known escape mutations (Figure 3-5). From sequence data alone, CSCS learns in a zero-shot learning setting to predict with high accuracy escape mutations which it had not previously seen. These escape mutations are obtained from the Stanford University HIV drug resistance database, and we classify phenotypes indicated in the highest bucket of drug fold resistance as escape mutations [17].

In particular, we confirm that both grammaticality and semantic change contain relevant information for the prediction of viral escape, where grammaticality tracks biological viability and semantic change indicates a meaningful antigenic change. With regards to protease and reverse transcriptase which are both encoded on the *pol* gene, a prediction scheme made purely based on ranked grammaticality seems to perform better than one solely based on semantic change, and combining the two in CSCS improves the AUC for the semantic change-based predictions. In both settings, CSCS clearly outperforms both the independent and epistatic models of EVcouplings by a sizeable margin and has similar performance to the Mutation Frequency method (Figure 3-6). With the comparatively similar performance between CSCS and the naive Mutation Frequency method, it is likely that the patterns of the viral HIV PR and RT "languages" are not particularly difficult to learn. We suspect that the relatively poor performance of the EVcouplings methods in comparison to CSCS and Mutation Frequency suggests that perhaps some of the input preprocessing performed by EVcouplings was detrimental to its performance to which CSCS is not prone.

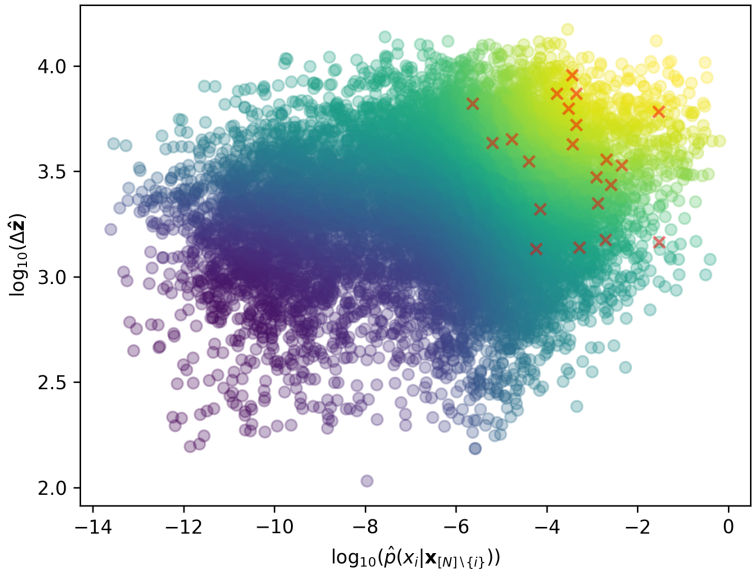
3.3 E. coli beta-lactamase

As demonstrated in Hie et al.'s work and confirmed in our HIV experiments with protease and reverse-transcriptase, through sequence data alone language models are able to

- learn biologically meaningful semantics embedding landscapes



(a) CSCS: HIV Protease



(b) CSCS: HIV Reverse Transcriptase

Figure 3-5: CSCS acquisition of HIV protease and reverse-transcriptase single-residue escape mutations. Each circle represents a single-residue mutation with red Xs drawn over escape mutations. Light yellow points are acquired first, dark blue points are acquired last.

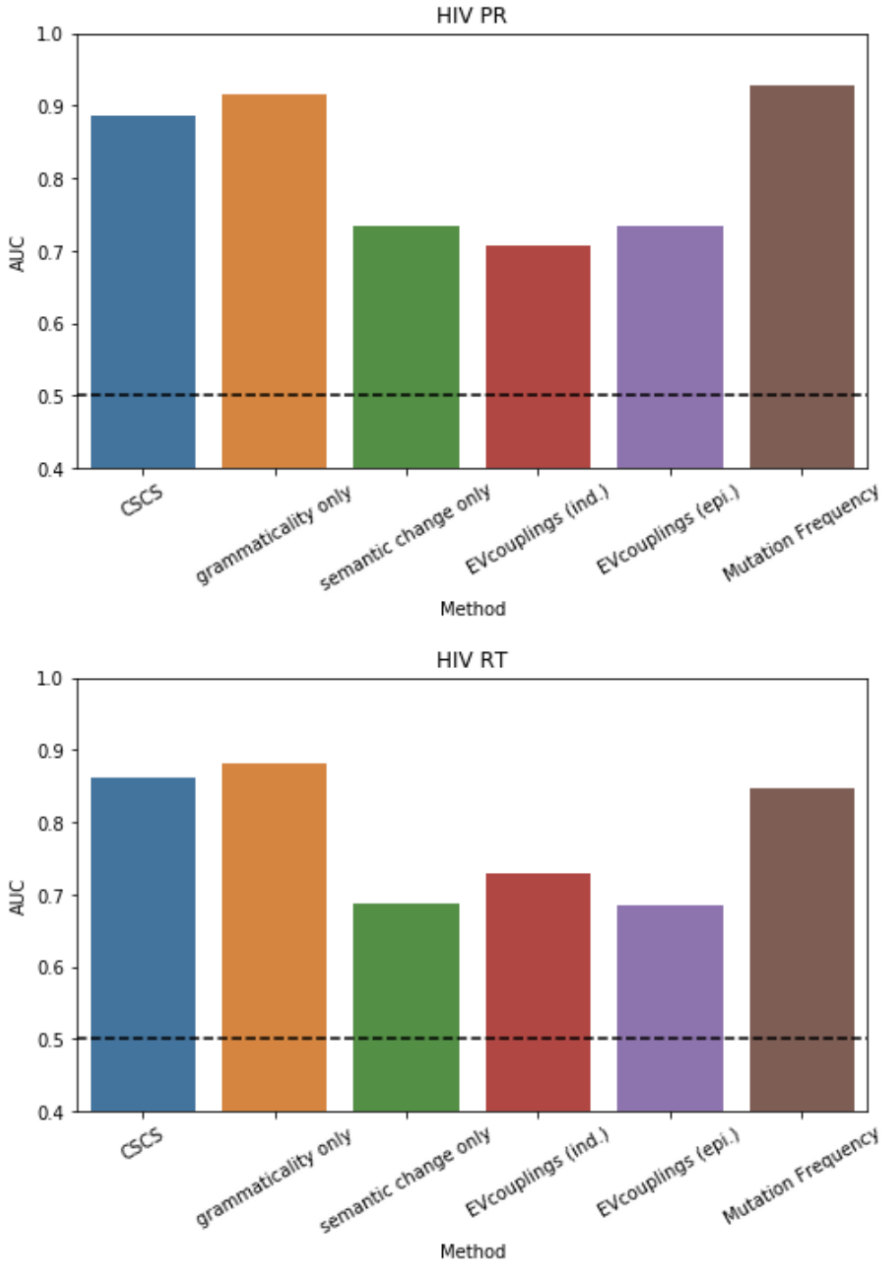


Figure 3-6: CSCS: HIV PR and RT Benchmarked Performance. Higher AUCs are better.

- quantify biological fitness via grammaticality
- combine these ideas of semantic change and grammaticality to predict escape via CSCS [6]

However simultaneously, we must have sufficiently rich sequence data for training such language models successfully for the purpose of escape prediction inference.

In the following beta-lactamase escape prediction setting, we consider training a BiLSTM language model based on NCBI's Anti-Microbial Resistance (AMR) dataset, which consists of 2,893 unique sequences. Then, based on a recent study, which screens deep mutants of the beta-lactamase protein among *E. coli* for resistance to beta-lactam antibiotics, we consider two different sets of screened deep mutants as "escape" in our validation tests [18]

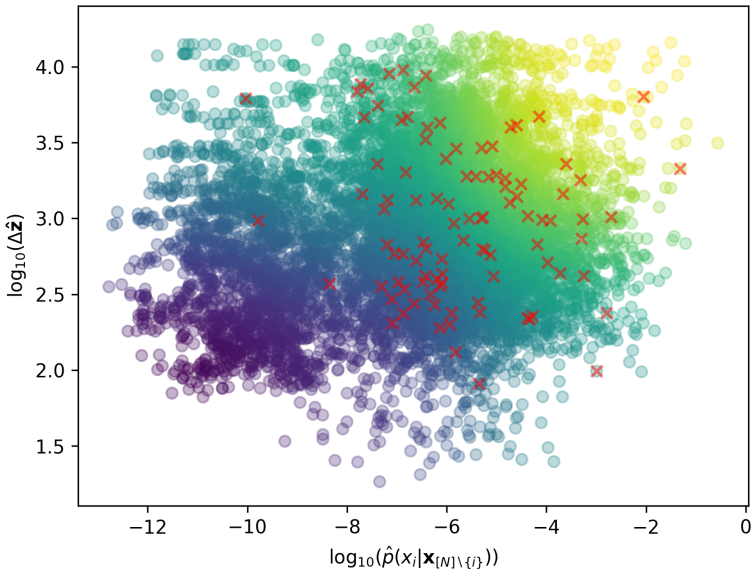
- **Antibiotic Resistant Mutants**: Mutants having an antibiotic IC50 concentration greater than that of the wild-type for any of the beta-lactam based antibiotic drugs (PIP, ATM, FEP)
- **Combination Resistant Mutants**: Mutants having an antibiotic IC50 concentration greater than that of the wild-type for any of the beta-lactam based antibiotic drugs (PIP, ATM, FEP) in the presence of avibactam (AVI), an inhibitor of beta-lactamase

With these two sets of escape mutations to validate against, we perform CSCS and also consider the performance of models using only semantic change and grammaticality alone. Benchmarking the performance of CSCS, we find statistically significant AUCs compared to a null hypothesis of random guessing in both settings, with p-values of 0.0069 and 0.0001 for combination resistance and antibiotic resistance respectively.

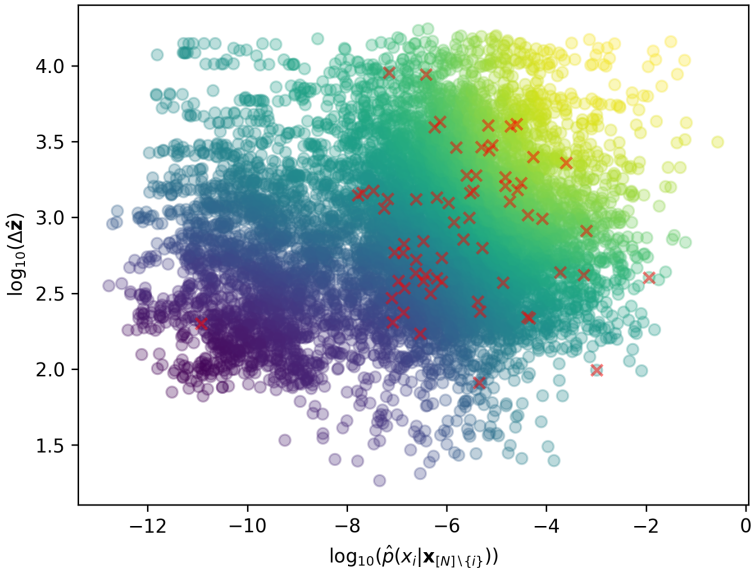
Although the raw AUC scores are smaller than in the HIV PR and RT results, CSCS clearly outperforms both EVcouplings methods in combination resistance prediction and does slightly better for antibiotic resistance prediction. Furthermore, it is not clear whether either the EVcouplings and Mutation Frequency methods perform

better than random guessing, as the independent EVcouplings method achieves an AUC which is less than 0.5 in the combination resistance prediction setting, while CSCS clearly achieves statistically significant results.

Along with the few number of training sequences available (fewer than half as many sequences available for RT and more than a whole order of magnitude less than the number of sequences available for PR), these benchmarks suggest that indeed, the training sequences were not sufficiently rich to enable the language model to sufficiently learn the semantics and grammar of the language. Thus, CSCS performs statistically significantly better than random guessing but it is likely its AUC performance could be further improved with access to a richer training sequence dataset.



(a) CSCS: E. coli beta-lactamase antibiotic resistance prediction



(b) CSCS: E. coli beta-lactamase antibiotic + AVI combination resistance prediction

Figure 3-7: CSCS acquisition of E. coli single-residue escape mutations. Each circle represents a single-residue mutation with red Xs drawn over escape mutations. Light yellow points are acquired first, dark blue points are acquired last.

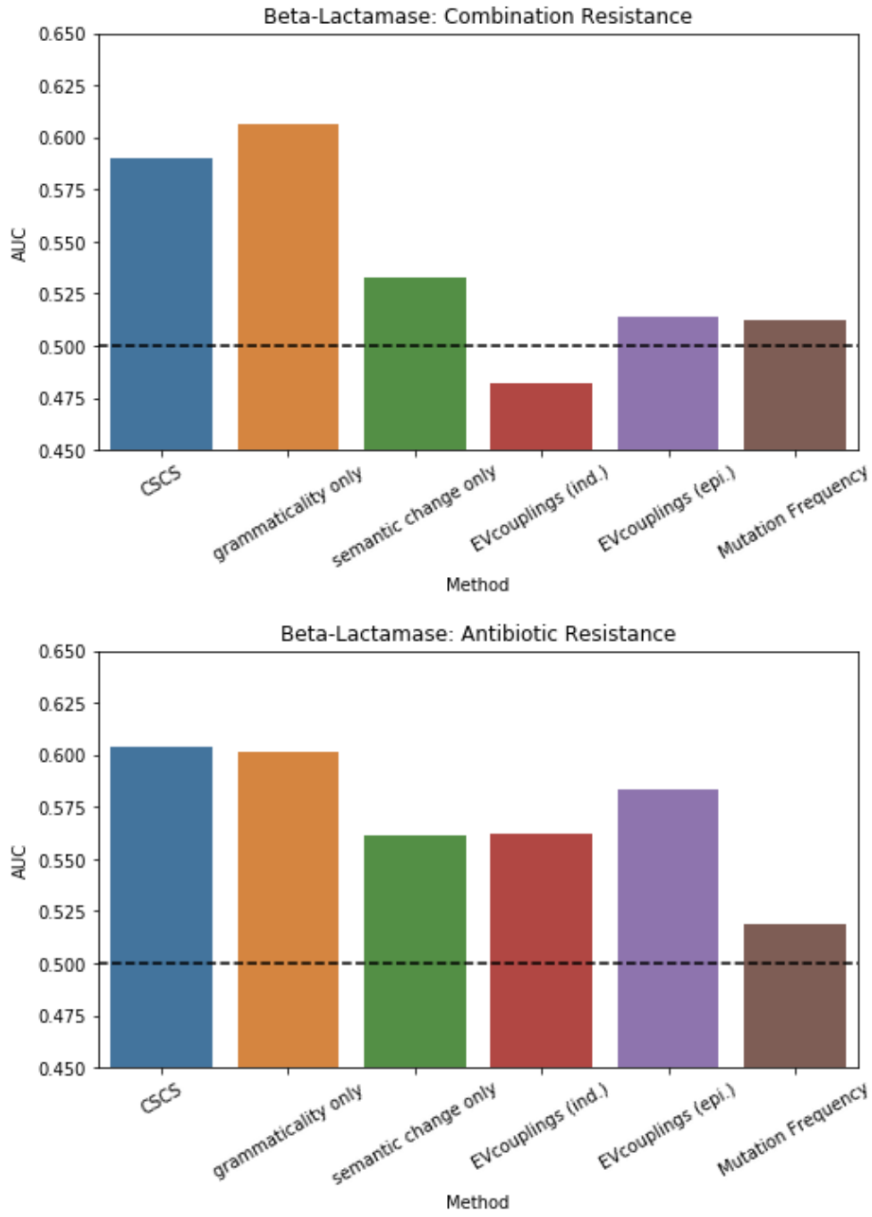


Figure 3-8: CSCS: beta-lactamase benchmarked performance. Higher AUCs are better.

Chapter 4

Conclusion and Future Directions

Mutation in viruses and bacteria presents a major barrier to the development of vaccines, antiviral drugs, and antibiotics. Being able to predict mutations that escape our current drugs or human immunity could lead to more intelligent therapeutic design. This work draws on recent work showing the promise of neural language models in predicting these escape mutations and applies it in understanding HIV-1 protease and reverse transcriptase. It also extends the application of the CSCS framework to understanding bacterial proteins by considering beta-lactamase in *E. coli*.

Indeed, we confirm that when given a sufficiently rich sequence dataset, one can train a neural language model to not only succeed in prediction of escape mutations but also construct a biologically-meaningful semantic landscape. From learning HIV subtype structure to capturing phylogenetic structure in bacteria, it is evident that neural language models are doing much more than pattern recognition for the sake of prediction. Instead, neural language models are capable of picking apart the biologically meaningful information embedded in protein sequences, which is a key contribution by Hie et al.

Coupling semantics with the notion of grammaticality, CSCS makes successful predictions when trained on a sufficiently rich sequence dataset. In this research, we extend the success of CSCS in the viral escape prediction setting and show the method generalizes for bacterial escape prediction as well. Nonetheless, when we have

insufficiently rich datasets, the learning is limited and the success of CSCS prediction falls slightly accordingly. To this end, we can reasonably expect the quality of the inference performed by CSCS generally improves with more data.

There are several potential areas for future research. The first is to consider training a model for a bacterial protein for which there exists a sufficiently rich sequence dataset (e.g. nuclear proteins of influenza, rpoB) in order to further verify the generality of CSCS. Additionally, the research could also have been extended to account for combinatorial mutations in a way that preserves computational tractability. Furthermore, improving the methodology to be able to understand not only substitution mutations but also deletions and insertions is an additional interesting direction to explore.

Though mutations constantly pose a challenge to researchers hoping to develop vaccines, antiviral drugs, and antibiotics; there is a recently expanding suite of methodologies to help. We hope that our work provides useful results and methodologies to researchers hoping to more intelligently design therapeutics.

Appendix A

Tables

Method	**AUC**
CSCS	0.886
grammaticality only	0.916
semantic change only	0.733
EVcouplings (ind.)	0.707
EVcouplings (epi.)	0.735
Mutation Frequency	0.928

Table A.1: HIV PR

Method	**AUC**
CSCS	0.862
grammaticality only	0.881
semantic change only	0.688
EVcouplings (ind.)	0.729
EVcouplings (epi.)	0.685
Mutation Frequency	0.846

Table A.2: HIV RT

Method	AUC
CSCS	0.604
grammaticality only	0.601
semantic change only	0.561
EVcouplings (ind.)	0.562
EVcouplings (epi.)	0.582
Mutation Frequency	0.519

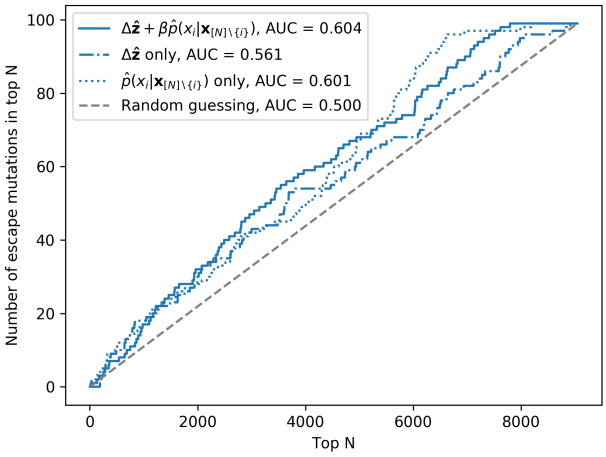
Table A.3: E. Coli beta-lactamase antibiotic resistance

Method	AUC
CSCS	0.590
grammaticality only	0.606
semantic change only	0.533
EVcouplings (ind.)	0.482
EVcouplings (epi.)	0.514
Mutation Frequency	0.512

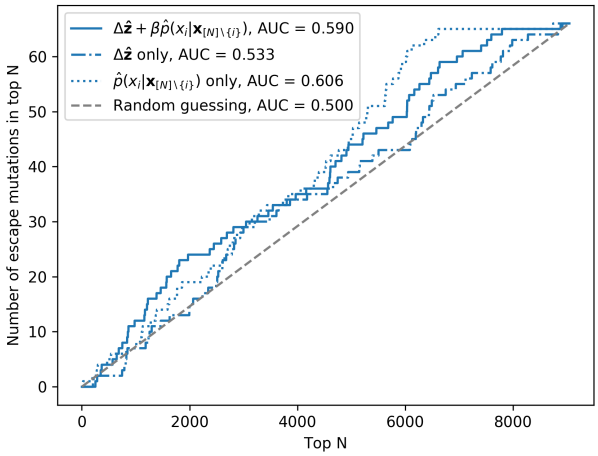
Table A.4: E. Coli beta-lactamase combination resistance

Appendix B

Figures

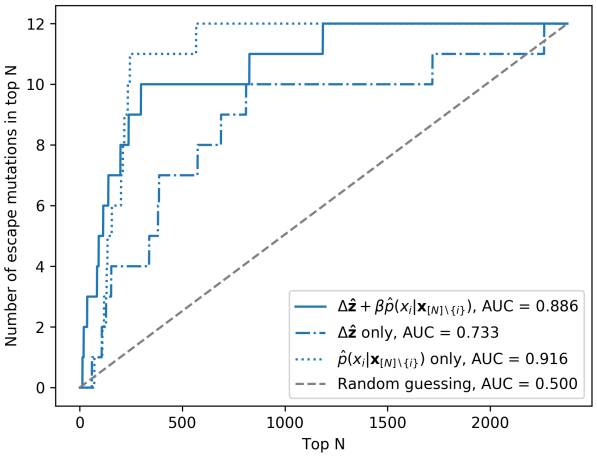


(a) Antibiotic Resistance

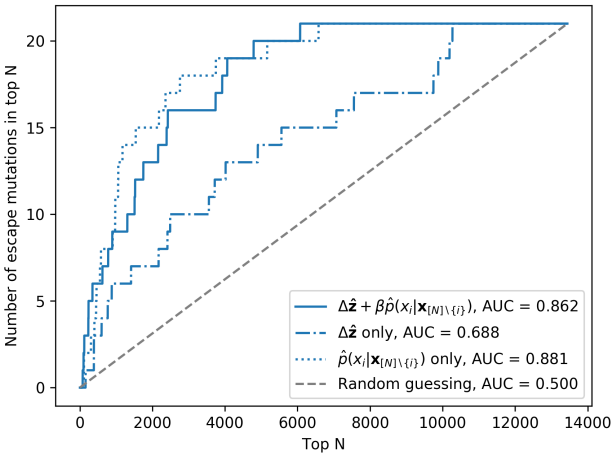


(b) Combination Resistance

Figure B-1: beta-lactamase ROC Curves



(a) HIV Protease



(b) HIV Reverse Transcriptase

Figure B-2: HIV ROC Curves

Bibliography

[1] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G.M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 2019.

[2] T. Bepler and B Berger. Learning protein sequence embeddings using information from structure. *7th International Conference on Learning Representations*, 2019.

[3] E. Cilia, S. Teso, S. Ammendola, and et al. Predicting virus mutations through statistical relational learning. *BMC Bioinformatics 15*, 2014.

[4] A.M. Dai and Q. V. Le. Semi-supervised sequence learning. *Adv. Neural Inf. Process. Syst.*, 2015.

[5] Chang M.-W. Lee K. Devlin, J. and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv cs.CL*, 2018.

[6] Brian Hie, Ellen Zhong, Bonnie Berger, and Bryson Bryan. Learning the language of viral evolution and escape. *Science*, 2020.

[7] T.A. Hopf, A.G. Green, B. Schubert, S. Mersmann, C.P.I. Schärfe, J.B. Ingraham, A. Toth-Petroczy, K. Brock, A.J. Riesselman, P. Palmedo, and et al. The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics 35*, 2019.

[8] K. Katoh and D.M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol*, 2013.

[9] L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv cs.CL*, 2013.

[11] Marcus Nguyen, Wesley Long, Patrick F. McDermott, Randall J. Olsen, Robert Olsen, Rick L. Stevens, Gregory H. Tyson, Shaohua Zhao, and James J. Davis. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *Journal of ClinicalMicrobiology*, 2019.

[12] J. Novembre, T. Johnson, K. Bryc, and et al. Genes mirror geography within europe. *Nature*, 2008.

[13] Neumann M.-Iyyer M. Gardner M. Clark C. Lee K. Peters, M. and L. Zettlemoyer. Deep contextualized word representations. *Proc. NAACL-HLT*, 2018.

[14] A. Philippon, Dény P. Slama, P., and R. Labia. A structure-based classification of class a -lactamases, a broadly diverse family of enzymes. *Clinical microbiology reviews*, 2016.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog 1, 9*, 2019.

[16] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song. Evaluating protein transfer learning with tape. *Adv. Neural Inf. Process. Syst.*, 2019.

[17] S. Y. Rhee, M. J. Gonzales, and et al. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res 31(1): 298-303*, 2003.

[18] D. Russ, F. Glaser, and E. et al. Shaer Tamar. Escape mutations circumvent a tradeoff between resistance to a beta-lactam and resistance to a beta-lactamase inhibitor. *Nature Communications*, 2020.

[19] M. A. Salama, A. E. Hassanien, , and A. Mostafa. The prediction of virus mutation using neural networks and rough set techniques. *EURASIP J. Bioinform. Syst. Biol*, 2016.