Signal Peptides Generated by Attention-Based Neural Networks

Zachary Wu[†,1], Kevin K. Yang[†,2], Michael J. Liszka[†,3], Alycia Lee[4], Alina Batzilla[5], David Wernick[3], David P. Weiner[3], Frances H. Arnold[*,1]

[1] Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA

[2] Generate Biomedicines, Cambridge, MA, USA; work performed while at California Institute of Technology

[3] BASF Enzymes, San Diego, CA, USA

[4] Department of Computational and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

[5] Department of Molecular Biotechnology, Heidelberg University, Heidelberg, Germany; work performed while at BASF Enzymes

KEYWORDS: machine learning, signal peptides, protein design, bacillus subtilis, secretion

ABSTRACT: Short (15-30 residue) chains of amino acids at the amino termini of expressed proteins known as signal peptides (SPs) specify secretion in living cells. We trained an attention-based neural network, the Transformer model, on data from all available organisms in Swiss-Prot to generate SP sequences. Experimental testing demonstrates that the model-generated SPs are functional: when appended to enzymes expressed in an industrial *Bacillus subtilis* strain, the SPs lead to secreted activity that is competitive with industrially used SPs. Additionally, the model-generated SPs are diverse in sequence, sharing as little as 58% sequence identity to the closest known native signal peptide and 73% ± 9% on average.

For cells to function, proteins must be targeted to their proper locations. Over one-third of the bacterial proteome that is synthesized in the cytoplasm is exported outside of it, and as a core requirement, the pathways that control localization are highly conserved across all domains of life.[1] To direct a protein through secretion pathways, organisms encode instructions in a leading short peptide sequence (typically 15-30 amino acids) called a signal peptide (SP).[2] SPs direct peptide chains to various export pathways, including the well-characterized Sec-[3–5] and Tat-mediated pathways.[6,7]

SPs have been engineered for a variety of industrial and therapeutic purposes, including increased export for recombinant protein production[2,8] and increasing the therapeutic levels of proteins secreted from industrial production hosts.[9] Secretion facilitates protein production by removing stress caused by protein accumulation in the cytoplasm, as well as by placing the burden of separation on the cells, which simplifies downstream processing.[10]

Due to the utility and ubiquity of protein secretion pathways, a significant amount of work has been invested in *identifying* SPs in natural protein sequences. Much of this work was pioneered by the groups behind the SignalP web server (http://www.cbs.dtu.dk/services/SignalP/), which first used artificial neural networks[11] and hidden Markov models[12] and now leverages modern deep learning architectures to model SPs.[13] An additional tool from the SignalP team, TargetP, is capable of identifying SP sequences and classifying them by the pathway used and the targeted intracellular or extracellular location.[14]

While this is a significant step toward modeling SP sequences from proteomic data, the challenging task of *generating* a SP sequence has yet to be validated *in vivo*. Indeed, the task of generating protein sequences of any kind is just beginning to be tackled.[15–19] Given a desired protein to target for secretion, there is no universally-optimal directing SP[20,21] and there is no reliable method for generating a SP with measurable activity. Instead, libraries of naturally-occurring SP sequences from the host organism or phylogenetically-related organisms are tested for each new protein secretion target.[21,22] That these libraries give functional SPs for new proteins is due to inherent "transferability" of SP sequences among multiple targets: empirically, roughly 50%[21] to 68%[20] of natural SPs paired to a protein secretion target produce measurable activity.

Although at one time the space of functional SP sequences was hypothesized to be quite large under the helical hairpin hypothesis,[23] subsequent research found that nature has designed SP sequences to interact with the necessary translocons in various pathways.[24] While researchers have attempted to generalize our understanding of SP–protein pairs by developing general SP design guidelines, those guidelines are heuristics at best and are limited to modifying existing SPs, not designing new ones.[2,25,26]

Here we present a machine translation model for generating SP sequences that have a high probability of being functional. Specifically, we trained a Transformer model[27] to predict SPs given the mature protein sequences of proteins annotated with SPs in Swiss-Prot[28] from all available organisms. These generated sequences are predicted by SignalP to have high probability of functioning as SPs. Upon *in vivo* validation in a gram-positive production organism, we find that 48% of constructs with generated SPs lead to secreted enzyme activity comparable to SPs used industrially. The functional generated sequences share as little as 58% sequence identity to the closest natural SP and 73% ± 9% on average.

Results

Model Description. We cast the SP generation problem as a translation problem by using the mature protein with the SP sequence removed as the source and the corresponding SP sequence as the output sequence. We employ the Transformer encoder-decoder architecture as first described by Vaswani *et al.* (ref. [27]) that leverages an attention mechanism,[29] which weights different positions over the entire sequence in order to determine a representation of that sequence, and remains a state-of-the-art architecture for machine translation between human languages.[30,31] Recent work has also applied the Transformer model to extract information from protein sequences for use in downstream protein function prediction and engineering tasks.[32,33]

Training Objective. We apply the Transformer architecture to SP prediction by treating each of the amino acids as a token (*cf.* machine translation, where words, characters, or subwords are tokens). The Transformer encoder maps an input sequence of tokens (the protein amino acids) to a sequence of continuous representations. Given these representations, the decoder then generates an output sequence (the SP amino acids) one token at a time. Each step in this generation depends on the generated sequence elements preceding the current step and continues until a special <END OF SP> token is generated. Figure 1 illustrates the modeling scheme. During training, we pass the decoder the true target SP. Training details can be found in Methods.

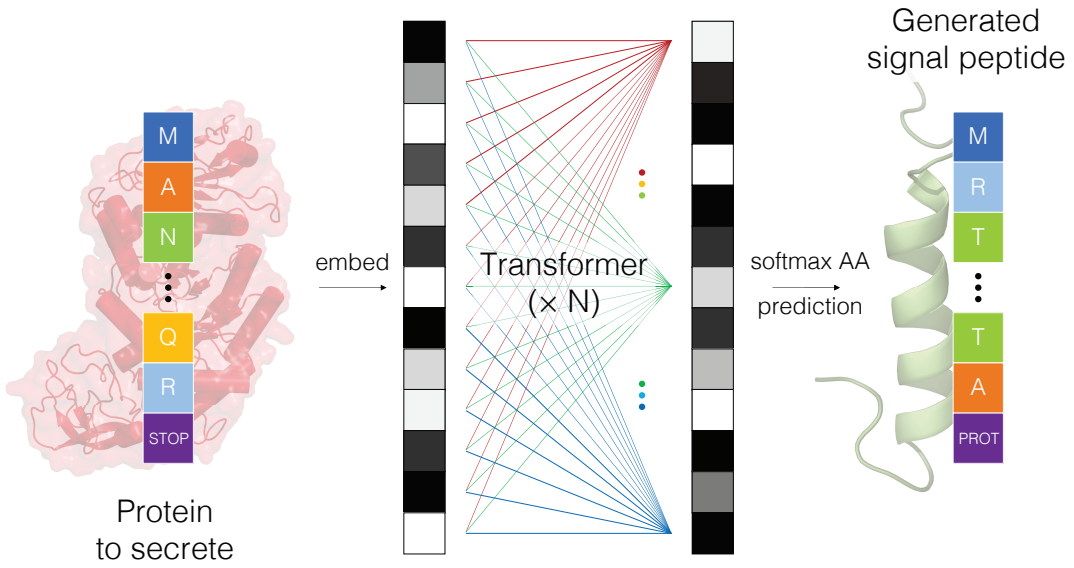


Figure 1. Sequence-to-sequence modeling for signal peptide (SP) amino acid sequences. During training, the first ~100 amino acids of the protein are tokenized and embedded as input for the Transformer Encoder-Decoder architecture with N=5 layers. The output is the SP amino acid sequence.

Training Data. From Swiss-Prot[28] we were able to extract over 40,000 SP–protein pairs from all domains of life. Protein secretory pathways are highly conserved, and others have found that incorporating data from all available organisms boosts accuracy in secretion prediction.[13] Additionally, we elect to train with SP-protein sequences over SP sequences alone, as experimental evidence suggests strong dependence on the protein sequence.[2,13,20] We selected sequence length maximum cutoffs of 70 amino acids and 105 amino acids for the SP and protein, respectively, to capture important motifs while keeping sequences short for more efficient training.

Model inputs are one-hot encodings of amino acid sequences of proteins to be secreted without their corresponding signal peptide. Mature proteins were padded or truncated to 105 residues, by observing that the loss during training did not decrease with longer input protein sequences, and the memory and computation required by the Transformer architecture scales quadratically with the sequence length. After truncation and removing duplicates, 25,000 SP–protein pairs remained, which were split randomly into training (80%) and validation (20%) sets. While we chose to restrict our search to the reviewed portion of UniProt (Swiss-Prot), most of the SPs returned were identified by computational annotation, and a future alternative is to incorporate sequences identified in TrEMBL, allowing a larger training set for better model prediction. Model outputs are one-hot encodings of signal peptide amino acid sequences, which are padded or truncated to 70 residues.

In addition to training on the full dataset, we also trained on filtered subsets of the full dataset for which we removed sequences with ≥75%, ≥90%, ≥95%, or ≥99% sequence identity to 28 enzymes from 4 families we selected for experimental validation (further described below) in order to test the model's ability to generalize to distant protein sequences. The Transformer model was then trained on each of these filtered datasets and used to generate sequences.

Machine Sequence Generation. Given a trained model that predicts sequence probabilities, there are many methods by which protein sequences can be generated.[17,19] One such method is beam search,[34] which generates a sequence by taking the most probable amino acid additions from the N-terminus. In traditional beam search, the size of the beam refers to the number of unique hypotheses with highest predicted probability for a specific input that are tracked at each generation step. For example, a beam size of 5 generates hypotheses from the N to C terminus, keeping the 5 most probable sequences as the sequence grows. In this study, we attempt to generate "generalist" SPs, which have higher probability of functioning across multiple input protein sequences. To this end, we employed an alternate form

of beam search, which we call "mixed input beam search" with a beam size of 5 over the decoder in identifying SPs. Our mixed input beam search generates SP hypotheses for *multiple* protein inputs, keeping the SP sequences with highest predicted probabilities. This generation process reflects the natural SPs' transferability between proteins to secrete, as 50-68% of natural SPs from related species exhibit measurable function when tested against specific enzymes.[20,21] By providing the Transformer model with multiple enzymes, the model has an opportunity to generate a sequence with high likelihood given multiple inputs, rather than being forced to generate a SP for an input it is unsure about.

For this study, we aimed to identify novel SPs (new amino acid sequences) and test them for secretion of ten enzymes across four families (amylases, lipases, proteases, and xylanases) in an industrial gram-positive bacterial (*Bacillus subtilis*) host. In addition to the ten enzymes tested, we also provided 31 other enzymes as inputs to generate SPs, in an effort to increase the transferability of generated SPs to multiple enzymes. The enzymes and the SPs generated for them can be found in Supplementary File 1. Predictions were made based on models trained on the four cutoffs for sequence identity described above.

The generated SP sequences from each cutoff (4 SPs for each target enzyme) were appended to different protein target sequences to test with SignalP. The generated SPs also showed high probability of functioning as predicted by SignalP 5.0 (average probability 90.4% ± 17.1%, Supplemental Section 1) and also contain many of the motifs common to SPs (positively charged N-terminus, hydrophobic core, and terminal AXA motif). While these heuristics could also be used to generate SPs, we find that the machine-learning approach generates SPs with significantly higher predicted probability of functioning than those generated by heuristics (p-value = 5×10^{-28}; Supplemental Section 1), which agrees with reported experimental difficulty in applying heuristics to designing SPs.[24] We also provide comparisons to sequences generated by HMMER[35] and a variational autoencoder[36] in Supplemental Section 1).

Secreted Enzyme Activity Validation. We then tested the predicted SPs by expressing SP-protein pairs in a *Bacillus subtilis* host strain used for secretion of industrial enzymes. We expressed ten enzymes: 5 amylases, 1 lipase, 2 proteases, and 2 xylanases. Functional secretion was determined by testing fermentation supernatants with the corresponding enzyme activity assay, as described in Supplemental Table 2.

For the ten enzymes, we tested 1) SPs generated by the Transformer model 2) industrial SPs native to *Bacillus subtilis* (positive controls) and 3) SPs generated using random source amino acid sequences. The sequences for 1) and 2) can be found in Supplementary File 1. For the positive controls, we used six SPs represented in previous studies for industrial levels of protein secretion (AprE, LipB, YbdG, YcnJ, YkvV, and YvcE).[20,21,37] For 3), output SPs were generated by the Transformer model for input protein sequences, which were made by drawing randomly from random amino acid distributions following a) the *Bacillus* amino acid distribution, b) the bacterial amino acid distribution, and c) a uniform amino acid distribution. The sequences can be found in Supplemental Table 4, and the functional classification results are summarized in Table 1. The measured enzyme activities for each construct, as well as details for their functional classification, can be found in Supplemental Section 2. A total of 163 unique constructs were tested.

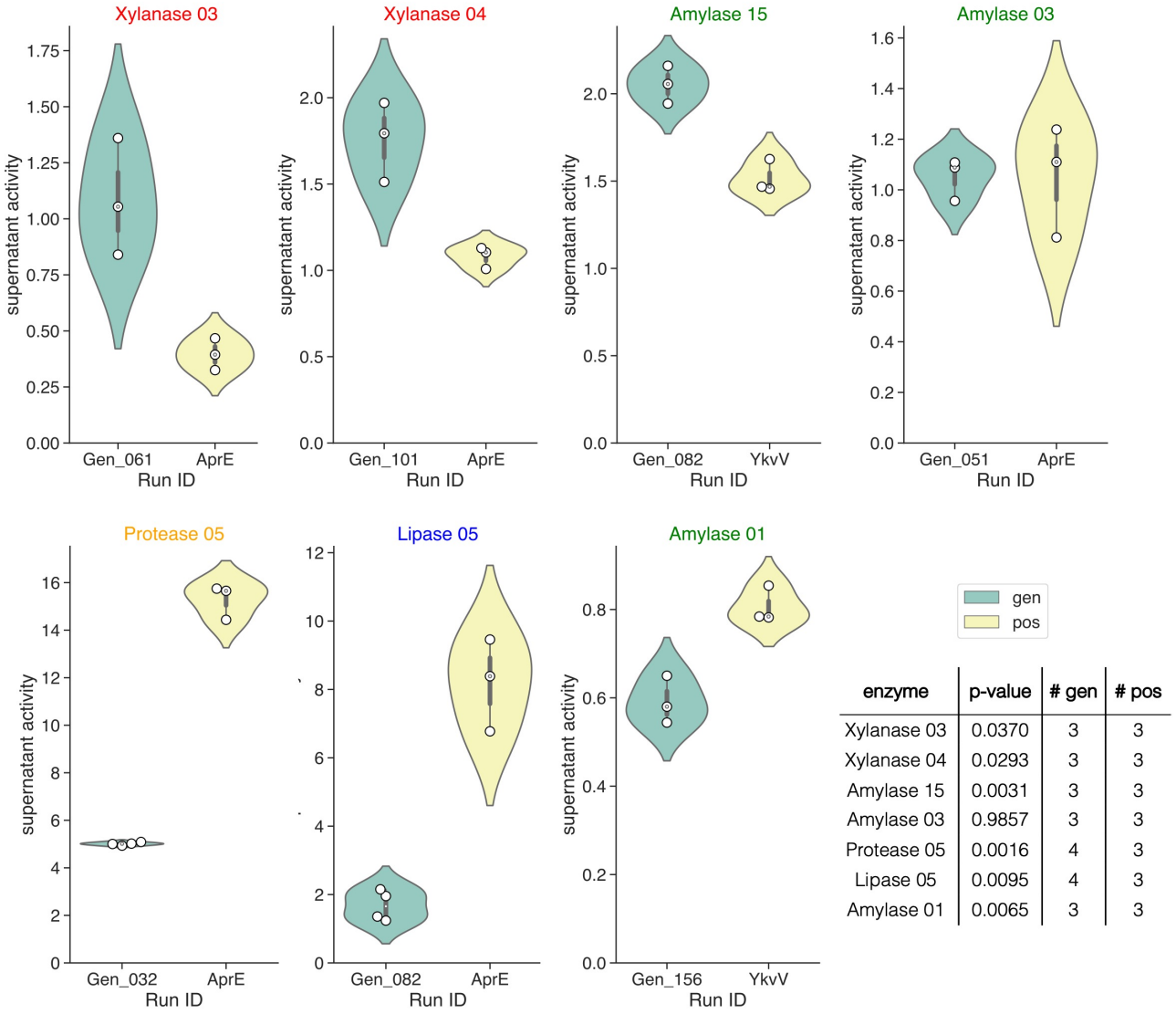
Table 1: Summary of protein-SP constructs that are functional.

	Num Functional	Num Tested	Percent Functional
SPs Generated for Random Inputs	1	18	6%
Natural SPs	27	34	79%
Generated SPs	53	111	48%

Functional classification is summarized in Supplemental Section 2, where enzyme activity in the supernatant is plotted for visual comparison. SPs Generated for Random Inputs were generated by the Transformer model given randomized amino acid sequences for the target protein, as detailed in Methods. Positive controls are naturally occurring *Bacillus* signal peptides. Generated SPs were generated for 41 proteins through mixed input beam search.

Using native SPs, 79% of constructs with 6 commonly used SPs resulted in secreted activity. We were pleased to find that a substantial fraction (48%) of the constructs containing a generated SP also resulted in significant secreted activity. SPs generated for random protein inputs were much less likely to lead to secreted activity. Only one construct containing an SP generated given random amino acid sequences gave some supernatant activity (Protease 05, Supplemental Section 2), which indicates that, in general, a real protein sequence is required for generating a sensible SP and the model is not relying on other artefacts for generation. Additionally, for the 21 generated functional SPs that were tested with multiple proteins, all 21 were functional for all proteins with which they were tested.

Generated SP-Enzyme Constructs Exhibit Activity Comparable to Natural Constructs. The model-generated SP-enzyme constructs are not only functional; they also exhibit activity similar to that of constructs with natural SPs. This is shown in Figure 2, which illustrates the highest performing natural and generated SP for each enzyme tested with both. Activities for all generated constructs can be found for comparison in Supplemental Section 3. Of the tested enzymes, approximately half exhibited higher or comparable secreted activity with machine-generated SPs. Thus, the generated SPs offer comparable and sometimes significantly higher activity compared to natural SPs, even with a generative model that was not specifically trained to optimize secretion levels.



enzyme	p-value	# gen	# pos
Xylanase 03	0.0370	3	3
Xylanase 04	0.0293	3	3
Amylase 15	0.0031	3	3
Amylase 03	0.9857	3	3
Protease 05	0.0016	4	3
Lipase 05	0.0095	4	3
Amylase 01	0.0065	3	3

Figure 2. Generated signal peptides enable secreted enzyme activities that are comparable to natural SPs. The highest-performing natural (labeled "pos") and machine-generated (labeled "gen") SPs are shown for the 7 enzymes where both were tested. Of these 7 enzymes, 4 exhibited the same or higher supernatant activity with generated SPs (top row), and 3 exhibited higher supernatant activity with native *Bacillus* SPs (bottom row). P-values are provided for reference for comparing the biological replicates of the best generated and natural SPs by a two-sided t-test with unequal variance for two independent samples of scores, where the null hypothesis is that the samples have identical expected values.

Generated Constructs Are Diverse in Sequence. The generated SPs occupy regions of sequence space that are not known to have been explored by naturally occurring SPs. The input *protein* sequences were removed at various sequence identity cutoffs from the training set to ensure that predictions were made for enzyme sequences that the trained model had never seen before. However, we did not specifically select for SP sequences that met a specified diversity threshold, as can be done to ensure sequence diversity.[38]

Interestingly, functional generated SPs share on average 73% ± 9% and as little as 58% sequence identity to the closest SP in Swiss-Prot (Figure 3A). Multiple sequence alignments (MSAs) for each of the best generated SPs identified for each enzyme can be found in Supplemental Section 5. We show the MSA for the most distant sequence in Figure 3B. In general, the generated SP retains characteristics of other natural SPs, such as a positively charged N-terminus, hydrophobic core, and AXA motif, while sharing low sequence identity (as low as 58%).

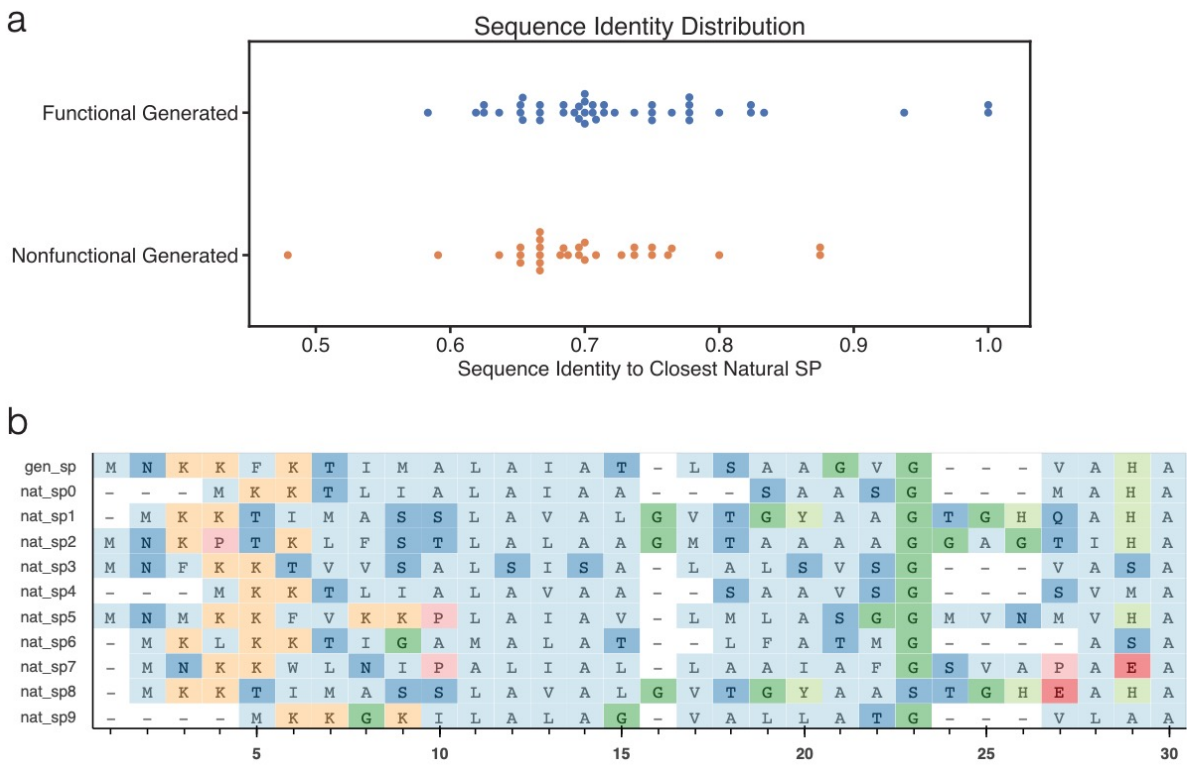


Figure 3. a) Percent sequence identity of various SPs to the closest matching natural SPs in Swiss-Prot, including 1) Functional Generated SPs (73% ± 9%) and 2) Nonfunctional Generated SPs (70% ± 8%) b) Multiple sequence alignment of the most diverse functional generated SP (58% identity to closest natural SP) with native SPs. Color groups follow those in ClustalW.[39]

SignalP Does Not Discriminate Between Functional and Non-Functional Constructs with Generated Signal Peptides. Interestingly, the functional classification accuracy of the best server, SignalP 5.0,[13] on the generated SPs is quite low. Figure 4 shows a receiver-operating curve (ROC) that displays true positive rate versus false positive rate for secretion probabilities generated by SignalP. As a reminder, random guessing gives an Area Under the Curve (AUC) of 0.50. SignalP performs quite poorly, with an AUC of only 0.59. However, there are a few differences in our modeling and validation approaches worth noting. First, our model is based on the Transformer architecture, whereas SignalP relies on bidirectional long short-term memory (LSTM) cells for longer range sequence interactions. Empirically, attention-based models currently have generally higher accuracy than LSTMs for protein tasks.[33] Additionally, our specific validation task of secreting functional enzyme in *Bacillus subtilis* differs from that of SignalP, which aims to assign a probability for sequences functioning as SPs from genomic data across many domains of life. Therefore, although SignalP may have the ability to discern natural SPs from other sequences, it does not appear to classify machine-generated SPs in *Bacillus* well, as previously shown by Brockmeier and coworkers.[20] This low accuracy may result from an inability to predict expression in the desired host, which SignalP is not trained for. In the future, SignalP may be adapted for specific production organisms in a feedback loop with our model, which is capable of generating functional sequences to test. We attempted to identify general protein properties from Biopython[40] that differed between the functional and nonfunctional SPs, but were unable to identify any statistically significant differences (Supplemental Section 4).

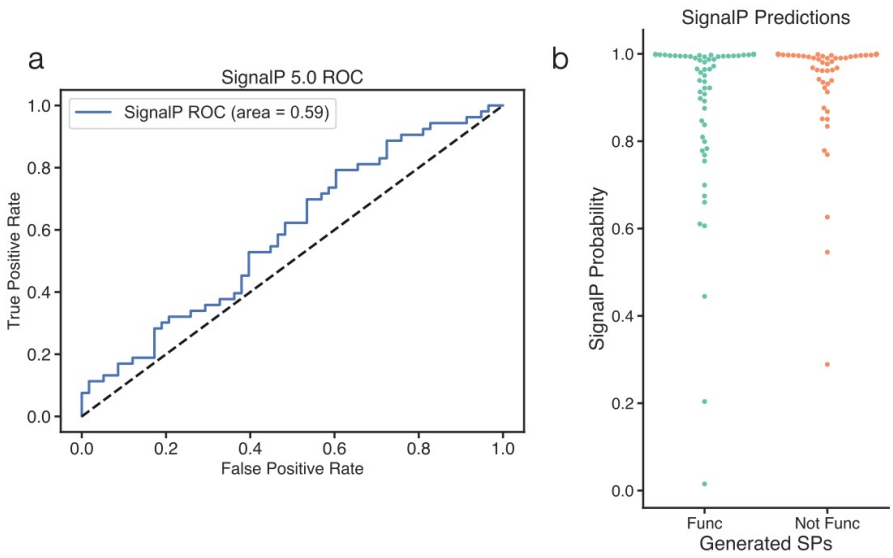


Figure 4. a) Receiver Operating Characteristic (ROC) curve for the prediction of functional constructs with machine-generated SPs. The SignalP 5.0 web server, an exemplary tool for natural SP annotation, performs poorly on this task, with AUC=0.59 (compared to 0.50 for random guess). b) Probability predictions for functional and nonfunctional generated SP constructs. Most constructs are predicted to be functional with high probability.

Discussion

We describe the application of a sequence-to-sequence model to generate functional peptide sequences that have not been identified in nature. These sequences accomplish the same function of directing enzyme secretion to the *B. subtilis* supernatant, yet they share as little as 58% sequence identity, and on average 73% ± 9%, to the closest-aligned recorded SP and thus explore new regions of sequence space. Enzymes with machine-generated SPs are expressed with activity levels comparable to those of natural SPs used in industrial enzyme production, although they were not explicitly designed to maximize secretion levels.

This work builds upon existing efforts in protein sequence generation with deep learning by providing *in vivo* validation of predictions. In one other case in which predictions were validated experimentally, 24% of the malate dehydrogenase enzymes from a generative adversarial network (GAN) by in-sample generation were functional.[15] Our approach uses a sequence-to-sequence approach trained with SP-protein pairs. We were pleased to find that a high fraction of generated constructs (98%) were predicted to be functional by SignalP (Supplemental Section 1), and a significant fraction (48%) of constructs were in fact functional *in vivo*. While SignalP is optimistic in its predictions, the lower fraction that is functional *in vivo* and the volume of heuristics developed for modifying existing SPs[2,25,26] suggest this remains a challenging engineering task.

Interestingly, the leading existing model trained for identifying SPs is not able to accurately distinguish functional machine-generated SP sequences from those that are not functional. Because our model generates sequences that an advanced critic (SignalP) is not able to discriminate among, coordinating these two systems in an adversarial approach could increase accuracy for both sequence generation and discrimination.

Important for both natural and synthetic SPs is whether they are transferable between secretion targets and host organisms. In this study, limited to a single round of experimentation, we used a generation strategy with multiple protein inputs with the goal of maximizing the probability of the SP functioning for any protein sequence. Of the functional generated SPs tested with multiple proteins, all 21 were classified as functional when prepended to all tested proteins. With knowledge that the Transformer model can generate functional protein sequences, probing the accuracy with which this translation strategy is able to generate SPs specific to desired secretion proteins and whether these specific SPs are transferable are potential future directions. Additionally, augmenting the generation process by conditioning on desired metadata, such as the host species, as outlined recently by Madani and coworkers,[19] may allow tuning SP sequences to different production organisms, as we have observed that the length distribution of generated SPs (20.3 ± 4.1) is lower than that of both Uniprot's *Bacillus* SPs (25.5 ± 6.3) and Brockmeier's set of *Bacillus* SPs (28.5 ± 5.2).

As the protein modeling field moves toward machine generation of protein sequences, our understanding of protein similarity must evolve as well. Similarities have historically been measured by a weighted alignment of linear sequences. However, we are finding that machine learning is capable of interpolating in modeled latent space to reach regions of sequence space that nature has yet to explore, as nature has significant physical limitations on its engineering strategies (and our databases, although large, are woefully incomplete). By challenging and supplementing nature's generation strategy with machine-generated sequences to more fully sample sequence space, we can unlock sequences with new properties and functions.

Materials and Methods

Model training. We trained a Transformer Encoder-Decoder with 5 layers and a hidden dimension of 550. Each layer had 6 attention heads. The model was trained for 100 epochs with a dropout rate of 0.1 in each attention head and after each position-wise feed-forward layer. Following the original

Transformer paper,[27] we used periodic positional encodings and the Adam optimizer. We increased the learning linearly for the first 12500 batches from 0 to 1e-4 and then decayed by *n_steps ^ -0.03* after the linear warmup. Models were trained on 1 NVIDIA V100 GPU through a generous grant from the Caltech Amazon Web Services Compute program.

Data augmentation. We used varying sub-sequences of the mature protein sequences as source sequences in order to augment our training dataset, to diminish the effect of choosing one specific length cutoff, and to make the model more robust. For mature proteins of length $L < 105$, the model receives the first $L - 10$, $L - 5$, and L residues as training inputs. For mature proteins of $L >= 105$, the model receives the first *95, 100, and 105* amino residues as training inputs. Data for signal peptides were collected from UniProt.

Bacterial strains, DNA design, and library construction. The expression vector was constructed from the *Bacillus subtilis* shuttle vector pHT01 by removal of the BsaI restriction sites and replacing the inducible Pgrac promotor with the constitutive promotor Pveg. However, IPTG was included during expression to ensure no residual or off-site inhibition from the LacI fragment still included on the pHT vector. Signal peptide sequences predicted from the model were reverse translated into DNA sequences for synthesis using JCat[41] for codon optimization with *Bacillus subtilis* (strain 168). Each gene of interest was modeled at four homology cut-offs resulting in 4 predicted signal peptides. These 4 signal peptides were synthesized as a single DNA fragment with spacers including the BsaI restriction sites. 8 individual colonies were picked from each group of 4 predicted signal peptides. Protein sequences were selected from literature reports of enzymes expressed in *Bacillus* host systems. Supplemental Excel File 1 lists the enzymes used in this work and their reported amino acid sequence. Signal peptide and protein DNA sequences were ordered from Twist Biosciences and cloned into their *E. coli* cloning vector. *Bacillus subtilis* PY97 was the base strain used for the expression of enzymes. Native enzymes that could interfere with measurement were knocked out as indicated in Supplemental Table 3.

The expression vector backbone, gene of interest, and SP fragments were amplified via PCR with primers including BsaI sites and assembled through Golden Gate Assembly, with a linker GGGGCT sequence (encoding Glycine and Alanine) between the generated SP and the target protein. Primers used to amplify each fragment are listed in Supplemental Table 1. Each linear DNA fragment was agarose gel purified for use in Golden Gate assembly reactions. The Golden Gate reactions were performed with 700ng vector PCR product, 100ng signal peptide group PCR product, and 300 ng gene of interest PCR product in 20µl reactions (2µl 10x T4 Ligase Buffer, 2µl 10x BSA, 0.8µl BsaI-HFv2, 1µl T4 Ligase). The reactions were cycled 35 times (10min, 37°C; 5 min, 16°C) then heat inactivated (5 min, 50°C; 5min, 80°C) before being stored at 4°C for use directly.

Enzyme expression and functional characterization. All *Bacillus* strains were transformed by natural competency as previously described.[42] Transformations were plated on LB agar (10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl, 15g/l agar) supplemented with 5µg/ml chloramphenicol and grown overnight at 37°C. Single colonies were picked and grown overnight in 96-well plates (Whatman #7701-5200) with LB containing 17 µg/ml chloramphenicol then stored as glycerol stocks. For enzyme expression, cultures were seeded from glycerol stocks into 100 µl LB media and grown overnight at 37°C. A 10 µl aliquot of the overnight culture was transferred into 500 µl of 2xYT media (16 g/l Tryptone, 10 g/l yeast extract, 5 g/l NaCl) containing 1mM IPTG and incubated for 48 hrs at either 30°C or 37°C with shaking (900 rpm, 3 mm throw). Culture supernatants were clarified by centrifugation (4000 rpm, 10 min) and used directly in enzyme activity assays. Strains were grown and expressed in at least three biological replicates from each original picked colony.

Enzyme expression quantification was attempted via SDS-PAGE (BioRad Criterion 10-20 % Tris-HCl) but the observed expression level was below a quantifiable limit. Enzyme expression was too low to reliably quantify with SDS-PAGE, so the relative expression of each enzyme was approximated by activity measurements. Enzyme activity was measured in the linear response range for each substrate and reaction condition as listed in Supplemental Table 2. Intracellular enzyme expression was assessed by washing the cell pellet after the supernatant was removed, and then resuspending in 500 µl of 50 mM HEPES buffer with 2 mg/ml Lysozyme and incubated for 30 minutes at 37 °C. The resuspended material was centrifuged again and used directly in enzyme activity assays.

SPs Generated for Random Inputs. SPs were generated by the trained Transformer model with 99% sequence identity cutoff for randomized protein inputs following a) the *Bacillus* amino acid distribution, b) the bacterial amino acid distribution, and c) a uniform amino acid distribution. The same mixed input beam search generation approach was used as detailed in *Machine Sequence Generation*. These sequences can be found below in Supplemental Table 4.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

- Supplementary Tables detailing 1) primers used to generate linear DNA fragments, 2) reaction conditions, 3) strains used, 4) control sequences generated, 5) distribution of protein and SP lengths as obtained from UniProt *and* Supplementary Sections detailing 1) comparison of alternate generation approaches, 2) functionality classification from experimental validation, 3) activity assays at higher dilution, 4) sequence characteristics of functional vs nonfunctional generated SPs, and 5) all MSAs for functional SPs, (PDF)
- Supplementary File 1: amino acid sequences of proteins and signal peptides (Excel)

AUTHOR INFORMATION

Corresponding Author

Frances H. Arnold – Department of Chemistry and Chemical Engineering, California Institute of Technology; orcid.org/0000-0002-4027-364X; Email: frances@cheme.caltech.edu

Author Contributions

Z.W., F.H.A., and K.K.Y. conceived and directed this study. K.K.Y., A.L., and Z.W. obtained training data and trained the models. Z.W., M.J.L., and D. Wernick planned the *in vivo* experimental validation. M.J.L., and A.B. performed the experimental validation. Z.W. analyzed the experimental results. D. Weiner advised the study. Z.W., F.H.A., K.K.Y., and M.J.L. wrote the paper. All authors edited and approved the manuscript. †These authors contributed equally.

Funding Sources

This work was supported by BASF through the California Research Alliance (CARA), the National Science Foundation Division of Chemical, Bioengineering, Environmental and Transport Systems (CBET-1937902), a National Science Foundation Graduate Fellowship GRF2017227007 (to Z.W.), and through generous research credits provided by Amazon Web Services.

Competing Interests

Provisional patent applications have been filed based on the results presented here.

Notes

The trained Transformer model for generating signal peptides and the data used to train the model will be available at https://github.com/fhalab/SPGen.

ACKNOWLEDGMENTS

The authors would like to thank Yisong Yue, Taehwan Kim, and other instructors of the Spring 2017 CS159 course at Caltech for initial guidance, and Zheyuan (Steve) Guo and Lucas Schaus for helpful discussions. Additionally, the authors would like to thank the team members of BASF Enzymes for being gracious hosts over the course of this project and Twist Biosciences for providing DNA at educational rates.

ABBREVIATIONS

SP signal peptide; LSTM long short-term memory; ROC receiver operating characteristic; IPTG isopropyl β-d-1-thiogalactopyranoside

REFERENCES

(1) Tsirigotaki, A.; De Geyter, J.; Šoštarić, N.; Economou, A.; Karamanou, S. Protein Export through the Bacterial Sec Pathway. *Nat. Rev. Microbiol.* **2016**, *15* (1), 21–36. DOI: 10.1038/nrmicro.2016.161.

(2) Low, K. O.; Mahadi, N. M.; Illias, R. M. Optimisation of Signal Peptide for Recombinant Protein Secretion in Bacterial Hosts. *Appl. Microbiol. Biotechnol.* **2013**, *97* (9), 3811–3826. DOI: 10.1007/s00253-013-4831-z.

(3) Wickner, W. The Enzymology Of Protein Translocation Across The Escherichia Coli Plasma Membrane. *Annu. Rev. Biochem.* **1991**, *60* (1), 101–124. DOI: 10.1146/annurev.biochem.60.1.101.

(4) Driessen, A. J. M.; Manting, E. H.; van der Does, C. The Structural Basis of Protein Targeting and Translocation in Bacteria. *Nat. Struct. Biol.* **2001**, *8* (6), 492–498. DOI: 10.1038/88549.

(5) Osborne, A. R.; Rapoport, T. A.; van den Berg, B. Protein Translocation by the Sec61/SecY Channel. *Annu. Rev. Cell Dev. Biol.* **2005**, *21*, 529–550. DOI: 10.1146/annurev.cellbio.21.012704.133214.

(6) Berks, B. C.; Palmer, T.; Sargent, F. Protein Targeting by the Bacterial Twin-Arginine Translocation (Tat) Pathway. *Curr. Opin. Microbiol.* **2005**, *8* (2), 174–181. DOI: 10.1016/j.mib.2005.02.010.

(7) Natale, P.; Brüser, T.; Driessen, A. J. M. Sec- and Tat-Mediated Protein Secretion across the Bacterial Cytoplasmic Membrane-Distinct Translocases and Mechanisms. *Biochim. Biophys. Acta - Biomembr.* **2008**, *1778* (9), 1735–1756. DOI: 10.1016/j.bbamem.2007.07.015.

(8) Mori, A.; Hara, S.; Sugahara, T.; Kojima, T.; Iwasaki, Y.; Kawarasaki, Y.; Sahara, T.; Ohgiya, S.; Nakano, H. Signal Peptide Optimization Tool for the Secretion of Recombinant Protein from Saccharomyces Cerevisiae. *J. Biosci. Bioeng.* **2015**, *120* (5), 518–525. DOI: 10.1016/j.jbiosc.2015.03.003.

(9) Zhang, L.; Leng, Q.; Mixson, A. J. Alteration in the IL-2 Signal Peptide Affects Secretion of Proteins in Vitro and in Vivo. *J. Gene Med.* **2005**, *7* (3), 354–365. DOI: 10.1002/jgm.677.

(10) Mergulhão, F. J. M.; Summers, D. K.; Monteiro, G. A. Recombinant Protein Secretion in Escherichia Coli. *Biotechnol. Adv.* **2005**, *23* (3), 177–202. DOI: 10.1016/j.biotechadv.2004.11.003.

(11) Nielsen, H.; Engelbrecht, J.; Brunak, S.; Heijne, G. Von. A Neural Network Method for Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites. *Int. J. Neural Syst.* **1997**, *8* (05n06), 581–599. DOI: 10.1142/S0129065797000537.

(12) Nielsen, H.; Krogh, A. Prediction of Signal Peptides and Signal Anchors by a Hidden Markov Model. *Intell. Syst. Mol. Biol.* **1998**, *6*, 122–130.

(13) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37* (4), 420–423. DOI: 10.1038/s41587-019-0036-z.

(14) Armenteros, J. J. A.; Salvatore, M.; Emanuelsson, O.; Winther, O.; Heijne, G. von; Elofsson, A.; Nielsen, H. Detecting Novel Sequence Signals in Targeting Peptides Using Deep Learning. *Life Sci. Alliance* **2019**, *2* (5). DOI: 10.1101/639203.

(15) Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Zrimec, J.; Poviloniene, S.; Rokaitis, I.; Laurynenas, A.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding Functional Protein Sequence Space Using Generative Adversarial Networks. *bioRxiv* **2019**, 789719. DOI: 10.1101/789719.

(16) Costello, Z.; Martin, H. G. How to Hallucinate Functional Proteins. *arXiv* **2019, 1903.00458**.

(17) Brookes, D. H.; Park, H.; Listgarten, J. Conditioning by Adaptive Sampling for Robust Design. *Proc. Mach. Learn. Res.* **2019**, *97*, 773–782.

(18) Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep Generative Models of Genetic Variation Capture the Effects of Mutations. *Nat. Methods* **2018**, *15* (10), 816–822. DOI: 10.1038/s41592-018-0138-4.

(19) Madani, A.; Mccann, B.; Naik, N.; Shirish, N.; Namrata, K.; Raphael, A.; Socher, P. H. R. ProGen : Language Modeling for Protein Generation. *bioRxiv* **2020 982272**. DOI: 10.1101/2020.03.07.982272

(20) Brockmeier, U.; Caspers, M.; Freudl, R.; Jockwer, A.; Noll, T.; Eggert, T. Systematic Screening of All Signal Peptides from Bacillus Subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-Positive Bacteria. *J. Mol. Biol.* **2006**, *362* (3), 393–402. DOI: 10.1016/j.jmb.2006.07.034.

(21) Degering, C.; Eggert, T.; Puls, M.; Bongaerts, J.; Evers, S.; Maurer, K. H.; Jaeger, K. E. Optimization of Protease Secretion in Bacillus Subtilis and Bacillus Licheniformis by Screening of Homologous and Heterologous Signal Peptides. *Appl. Environ. Microbiol.* **2010**, *76* (19), 6370–6376. DOI: 10.1128/AEM.01146-10.

(22) Hemmerich, J.; Rohe, P.; Kleine, B.; Jurischka, S.; Wiechert, W.; Freudl, R.; Oldiges, M. Use of a Sec Signal Peptide Library from Bacillus Subtilis for the Optimization of Cutinase Secretion in Corynebacterium Glutamicum. *Microb. Cell Fact.* **2016**, *15* (1), 208. DOI: 10.1186/s12934-016-0604-6.

(23) Engelman, D. M.; Steitz, T. A. Insertion of Proteins into and across Membranes : The Helical Hairpin Hypothesis. **1981**, *23* (February), 411–422. DOI: 10.1016/0092-8674(81)90136-7.

(24) Duffy, J.; Patham, B.; Mensa-Wilmot, K. Discovery of Functional Motifs in H-Regions of Trypanosome Signal Sequences. *Biochem. J.* **2010**, *426* (2), 135–145. DOI: 10.1042/BJ20091277.

(25) Owji, H.; Nezafat, N.; Negahdaripour, M.; Hajiebrahimi, A.; Ghasemi, Y. A Comprehensive Review of Signal Peptides: Structure, Roles, and Applications. *Eur. J. Cell Biol.* **2018**, *97* (6), 422–441. DOI: 10.1016/j.ejcb.2018.06.003.

(26) Freudl, R. Signal Peptides for Recombinant Protein Secretion in Bacterial Expression Systems. *Microb. Cell Fact.* **2018**, *17* (1), 52. DOI: 10.1186/s12934-018-0901-3.

(27) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, 5998–6008.

(28) Consortium, U. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506–D515. DOI: 10.1093/nar/gky1049.

(29) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* **2014**, *abs/1409.0*.

(30) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language

Understanding. *arXiv* **2018, 1810.04805**.

(31) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019, 1907.11692**.

(32) Rives, A.; Goyal, S.; Meier, J.; Guo, D.; Ott, M. C.; Zitnick, L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *bioRxiv* **2019**. DOI: 10.1101/622803.

(33) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *Adv. Neural Inf. Process. Syst.* **2019**, 9686–9698.

(34) Graves, A. Sequence Transduction with Recurrent Neural Networks. *ICML Work. Represent. Learn.* **2012**.

(35) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics*. 1998. DOI: 10.1093/bioinformatics/14.9.755.

(36) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*; 2014.

(37) Zhang, W.; Yang, M.; Yang, Y.; Zhan, J.; Zhou, Y.; Zhao, X. Optimal Secretion of Alkali-Tolerant Xylanase in Bacillus Subtilis by Signal Peptide Screening. *Appl. Microbiol. Biotechnol.* **2016**, *100* (20), 8745–8756. DOI: 10.1007/s00253-016-7615-4.

(38) Bedbrook, C. N.; Yang, K. K.; Robinson, J. E.; Gradinaru, V.; Arnold, F. H. Machine Learning-Guided Channelrhodopsin Engineering Enables Minimally-Invasive Optogenetics. *Nat. Methods* **2019**, *16*, 1176–1184. DOI: 10.1038/s41592-019-0583-8

(39) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22* (22), 4673–4680. DOI: 10.1093/nar/22.22.4673.

(40) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. DOI: 10.1093/bioinformatics/btp163.

(41) Grote, A.; Hiller, K.; Scheer, M.; Münch, R.; Nörtemann, B.; Hempel, D. C.; Jahn, D. JCat: A Novel Tool to Adapt Codon Usage of a Target Gene to Its Potential Expression Host. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W526–W531. DOI: 10.1093/nar/gki376.

(42) Koo, B. M.; Kritikos, G.; Farelli, J. D.; Todor, H.; Tong, K.; Kimsey, H.; Wapinski, I.; Galardini, M.; Cabal, A.; Peters, J. M.; Hachmann, A. B.; Rudner, D. Z.; Allen, K. N.; Typas, A.; Gross, C. A. Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus Subtilis. *Cell Syst.* **2017**, *4*, 291–305. DOI: 10.1016/j.cels.2016.12.013.

Authors are required to submit a graphic entry for the Table of Contents (TOC) that, in conjunction with the manuscript title, should give the reader a representative idea of one of the following: A key structure, reaction, equation, concept, or theorem, etc., that is discussed in the manuscript. Consult the journal's Instructions for Authors for TOC graphic specifications.