SecretoGen: towards prediction of signal peptides for efficient protein secretion

Felix Teufel[12*] **Carsten Stahlhut**[2] **Jan Christian Refsgaard**[2] **Henrik Nielsen**[3]
Ole Winther[13] **Dennis Madsen**[2*]
[1]University of Copenhagen [2]Novo Nordisk A/S [3]Technical University of Denmark
felix.teufel@bio.ku.dk, dnnm@novonordisk.com

Abstract

Signal peptides (SPs) are short sequences at the N terminus of proteins that control their secretion in all living organisms. Secretion is of great importance in biotechnology, as industrial production of proteins in host organisms often requires the proteins to be secreted. SPs have varying secretion efficiency that is dependent both on the host organism and the protein they are combined with. Therefore, to optimize production yields, an SP with good efficiency needs to be identified for each protein. While SPs can be predicted accurately by machine learning models, such models have so far shown limited utility for predicting secretion efficiency. We introduce **SecretoGen**, a generative transformer trained on millions of naturally occuring SPs from diverse organisms. Evaluation on a range of secretion efficiency datasets show that SecretoGen's perplexity has promising performance for selecting efficient SPs, without requiring training on experimental efficiency data.

1 Introduction

Signal peptides (SPs) are short N-terminal sequences that target proteins to the secretory pathway and are found in all living organisms. They are of great importance to protein biology and engineering as they form the main mode of exporting proteins from the cell, so that e.g. enzymes, signalling molecules or antibodies can fulfill their function in the cell's environment. SP-mediated protein secretion is a multi-step process that depends on many different binding partners that together form the secretory pathway. The success of the process relies on the SP having a biochemical structure that facilitates recognition by all binding partners [1].

Given the evolutionary divergence of organisms, these structural requirements differ with increasing evolutionary distance. It has been known for a long time that there are systematic differences between groups of organisms [2, 3]. Also within a single organism SP structural requirements are not universal, as the SP needs to be *compatible* with the sequence of the mature protein [4]. When industrially producing a protein of interest in a host organism, an SP needs to be found that results in efficient expression (Figure 1), so that manufacturing is both ecologically and commercially viable.

The underlying biology of this compatibility remains elusive, and available experimental data that combines SPs with various proteins has so far proven too limited to learn to predict compatibility in a generalizable way. However, positive data – in the sense of SP-protein pairs that result in secretion – can be observed in nature, as any naturally occurring combination can be assumed to have evolved to sufficient efficiency for expressing the protein at the needed level to be viable. We therefore investigate using generative neural networks to learn the conditional distribution of SPs from positive data only, and explore their applicability for the efficiency prediction problem.

NeurIPS 2023 Generative AI and Biology Workshop.

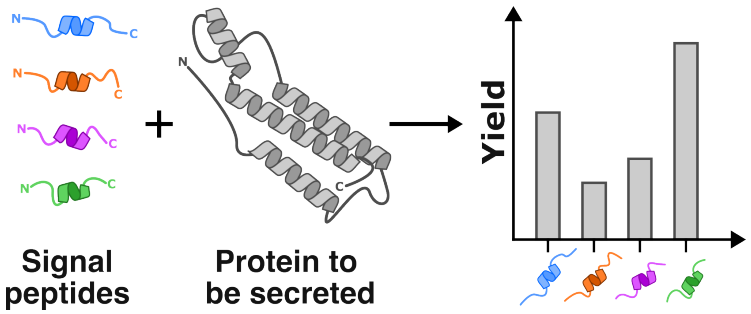

Figure 1: The signal peptide engineering problem. Given a protein to be produced, a signal peptide that results in high secretory expression yields needs to be found.

In summary, we introduce

- **SP efficiency prediction as a ranking problem**, using four experimental datasets for evaluation.
- **SecretoGen**, a 270 million parameter conditional generative transformer trained on 4.9 million predicted SPs.

2 Related works

Barash et al. [5] pioneered the field of generating secretion-optimized SPs by training a hidden Markov model capable of predicting SP sequences on 416 human samples. Some generated SPs improved secretion of a model protein compared to natural SPs. Wu et al. [6] proposed to frame SP generation as a sequence-to-sequence task, conditioning on the mature protein sequence and generating a corresponding SP autoregressively. Their SPGen model was trained on 25,000 reviewed sequences from UniProt. 48% of the model-generated SPs were found to be functional in *B. subtilis*.

3 Methods

3.1 Data

Training data UniProt [7] contains 3,827 experimentally verified SPs. As this number is rather limited for training generative models, we augment the data by predicting UniProt reference proteomes using SignalP 6.0 [8]. In total, we obtain 4,940,012 protein sequences paired with SP sequences from 11,097 organisms. Detailed data processing is laid out in A.1.

Efficiency data We curate data from four experiments [6, 9, 10, 11] that experimentally measured the secretion efficiency of SPs for selected proteins in suitable host organisms (Table 1). We focus on studies that have at least 100 measurements available within a single experiment. While there are more studies available that measured secretory yields of different SPs, many are not reusable for modeling purposes due to the used protein sequences not being reported [12, 13, 14]. For brevity, studies are referred to by their first author name throughout this work.

Table 1: Benchmark sets for SP efficiency prediction.

Name	Count	Label	Organism	Protein	Origin of SP sequences
Grasso et al. [9]	4,421	Enrichment	*B. subtilis*	alpha-Amylase from *B. amyloliquefaciens*	Natural and mutants
Wu et al. [6]	162	Binary	*B. subtilis*	various enzymes	Natural and model generated
Xue et al. [10]	322	Extracellular Yield	*S. cerevisiae*	alpha-Amylase from *A. oryzae*	Natural
Zhang et al. [11]	114	Extracellular Yield	*B. subtilis*	Xylanase from *B. pumilus*	Natural

3.2 Zero-shot efficiency prediction

Ultimately, generative SP models should be capable of designing novel SPs with good protein yields. However, as this requires experimental validation, it is impractical during model development and can become a limiting factor. We instead propose to focus on the problem of predicting efficiency, as this can be evaluated on existing datasets. We frame this as a ranking problem: given a list of candidate SPs for a protein, predict scores to order the SPs by efficiency. Performance is evaluated as the Spearman rank correlation coefficient (ρ). In cases where efficiency labels are binary (secretion successful/failed), we compute the area under the receiver operating characteristic curve (AUC).

3.3 Baseline methods

While this work is the first to propose SP ranking as a task, there are available methods that are amenable to the problem. Firstly, there are SP predictors, algorithms that detect SPs in natural protein sequences. While they are not trained to predict efficiency, their SP probability can still be used as a baseline, as it is capable of prioritizing true SPs over other, non-SP sequences. SP predictors have proven useful for some engineering applications [15]. We choose SignalP 5.0 [16] and 6.0 [8] as representative methods. Autoregressive protein LMs (pLMs) can score the perplexity of input amino acid (AA) sequences. In such models, the SP-protein compatibility problem is therefore phrased as a generic "N terminus - mature protein" likelihood. We benchmark ProGen2 [17], the largest available autoregessive pLM (A.3). Lastly, SPGen can be used to compute perplexities of SPs given a protein.

3.4 The SecretoGen model

We aim to model the conditional likelihood of SPs given a protein of interest and a host organism. To this end, we train a generative sequence-to-sequence transformer that conditions on the AA sequence m of the protein of interest and a token o denoting the organism. For an SP AA sequence x of length N, the conditional likelihood $p(x|m, o)$ is factorized autoregressively over the SP's sequence x, conditioning the probability of the next AA x_{i+1} on all its preceding AAs $x_{\leq i}$.

$$p(x|m, o) = \frac{1}{N} \sum_{i=1}^{N} \log p(x_{i+1}|x_{\leq i}, m, o) \tag{1}$$

We achieve this by introducing an organism embedding layer (Emb) that maps the organism of a sequence to a learned vector representation. This vector then serves as the first input to the decoder (Dec) which generates the SP sequence. The protein sequence is processed by the encoder (Enc). The likelihood $p(x_{i+1}|x_{\leq i}, m, o)$ of the AA at position $i + 1$ is thus computed as

$$p(x_{i+1}|x_{\leq i}, m, o) = \text{Dec}(x_{\leq i}, \text{Emb}(o), \text{Enc}(m)) \tag{2}$$

We use a standard transformer [18] with 12 encoder and 12 decoder layers and a hidden dimension of 1024 for all our experiments. We exploit the fact that organisms are related by evolution, rather than just constituting categorical labels, and embed organisms in a hierarchical manner for more efficient sharing of information between different organisms. We learn embeddings for 8 levels L of the taxonomy tree (*Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, Species*) and compute the organism embeddings that are used for conditioning as their sum (A.2). All taxonomy level embeddings $\mathbf{e}_l$ are randomly initialized and learned during training.

$$\text{Emb}(o) = \sum_{l \in L}^{L} \mathbf{e}_l \tag{3}$$

SecretoGen's parameters θ are optimized using the next token prediction objective on the train set of naturally evolved $\{x, m, o\}$ triplets predicted from UniProt.

$$Loss\,(x) = -\sum_{i=1}^{N} \log p_\theta(x_{i+1}|x_{\leq i}, m, o) \tag{4}$$

Table 2: SecretoGen ablation models. x = SP, m = mature protein, o = organism.

Model	Mature protein	Organism ID	Test perplexity ↓
$p(x\|m, o)$	Yes	Yes (8 levels)	**9.40**
$p(x\|m, o_{simple})$	Yes	Yes (1 level)	9.49
$p(x\|o)$	No	Yes	9.67
$p(x\|m)$	Yes	No	9.74
$p(x)$	No	No	10.12

4 Results

4.1 Including the organism of origin improves SP language modeling performance

We first investigate whether the learned distribution of SPs x is indeed conditional on a) the protein sequence m and b) the organism o. We train ablations of the SecretoGen model and evaluate the perplexity of the held out test sequences. The model that conditions on both m and o has the lowest perplexity, indicating that these two sources of information are useful for capturing the distribution of SPs (Table 2). Encoding organisms as multiple levels of the taxonomy tree outperforms using the species ID (=one level) only.

4.2 Zero-shot secretion efficiency prediction

We evaluate the SecretoGen model with full conditioning information and all baseline models on the SP ranking benchmark datasets. SecretoGen's perplexity shows the highest performance for ranking SPs by efficiency on three out of four datasets (Table 3). When evaluating ProGen2 and SPGen, we find that different model checkpoints show inconclusive performance, with no checkpoint clearly outperforming another. Surprisingly, except on the Wu dataset, all evaluated models perform better than random, without having been trained on efficiency data.

Performance varies by dataset, with SecretoGen performing strongly on Grasso, but only weakly on Zhang, even though both datasets are in *B. subtilis*. Closer inspection of the quantitative datasets reveals that both Grasso and Xue exhibit a bimodal distribution of efficiency values, with the lower mode representing SPs that result in poor secretion (Figure A3). In Zhang, which is the smallest dataset, this bimodality is less pronounced. SecretoGen ranking performance is to a large part driven by assigning high perplexity to sequences in the low-efficiency mode, thereby partly acting as a filter that discards low-efficiency SPs. If less low-efficiency SPs are present in a dataset, performance is expected to be of the order observed on Zhang.

Table 3: Zero-shot SP efficiency prediction performances. We use perplexities or classifier probabilities for ranking, and evaluate correlation of the ranked list with ground truth experimental efficiency. ρ = Spearman's rank correlation coefficient, AUC = Area under the receiver operator characteristic curve. Perplexities are inverted for computing ρ and AUC. Multiple available checkpoints are evaluated for ProGen2 and SPGen.

	Model	Quantitative datasets (ρ ↑)			Qualitative datasets (AUC ↑)
		Grasso	**Xue**	**Zhang**	**Wu**
Classifiers	SignalP 6.0	0.34	0.21	0.21	0.37
	SignalP 5.0	0.57	0.27	0.14	**0.70**
Generative pLMs	ProGen2 (base-fwd)	0.37	0.35	0.17	0.54
	ProGen2 (large-fwd)	0.51	0.29	0.15	0.53
	ProGen2 (xlarge-fwd)	0.56	0.37	0.17	0.53
	ProGen2 (base-rev)	0.46	0.39	0.05	0.48
	ProGen2 (large-rev)	0.43	0.44	0.04	0.47
	ProGen2 (xlarge-rev)	0.46	0.43	0.01	0.51
Generative SP LMs	SPGen (75)	0.29	0.30	0.10	0.51
	SPGen (90)	0.29	0.16	**0.28**	0.54
	SPGen (95)	0.16	0.24	0.18	0.57
	SPGen (99)	0.23	0.24	0.18	0.48
	SecretoGen	**0.61**	**0.49**	0.24	**0.70**

5 Discussion

We introduce SecretoGen, a conditional generative model of naturally occurring SPs. Our results confirm that when modeling SP sequences, it is beneficial to take both the mature protein sequence and the organism into account. We show that training on predicted data is a viable strategy, and that the distribution of naturally occurring SP-protein pairs encodes information about efficiency. While SecretoGen performs better than previous approaches at zero-shot efficiency prediction, we note that overall, the achieved performance level is still modest. Presumably, further improvement of performance is needed so that SP selection efforts for protein production applications can be guided by SecretoGen.

A remaining limitation of the secretion efficiency benchmark collection is that it only covers two host organisms. While *S. cerevisiae* and *B. subtilis* are two of the most popular microbial secretory expression hosts, other relevant systems such as CHO for mammalian proteins are missing as we could not identify any study with publicly available data of the required scale. Moreover, for two of the benchmark datasets, the recombinant problem could be considered comparably mild, as the target proteins are from the same *Bacillus* genus as the expression host.

Acknowledgements

We acknowledge EuroHPC Joint Undertaking for awarding us access to MeluXina at LuxProvide, Luxembourg.

Code and Data availability

Code and data are available at `https://github.com/fteufel/SecretoGen`.

References

[1] Gunnar von Heijne. Life and death of a signal peptide. *Nature*, 396(6707):111–113, November 1998. Number: 6707 Publisher: Nature Publishing Group.

[2] Gunnar von Heijne and Lars Abrahmsèn. Species-specific variation in signal peptide design Implications for protein secretion in foreign hosts. *FEBS Letters*, 244(2):439–446, 1989.

[3] Ning Zheng and Lila M. Gierasch. Signal Sequences: The Same Yet Different. *Cell*, 86(6):849–852, September 1996. Publisher: Elsevier.

[4] Chong Peng, Chaoshuo Shi, Xue Cao, Yu Li, Fufeng Liu, and Fuping Lu. Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Frontiers in Bioengineering and Biotechnology*, 7, 2019.

[5] Steve Barash, Wei Wang, and Yanggu Shi. Human secretory signal peptide description by hidden Markov model and generation of a strong artificial signal peptide for secreted protein expression. *Biochemical and Biophysical Research Communications*, 294(4):835–842, June 2002.

[6] Zachary Wu, Kevin K. Yang, Michael J. Liszka, Alycia Lee, Alina Batzilla, David Wernick, David P. Weiner, and Frances H. Arnold. Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synthetic Biology*, 9(8):2154–2161, August 2020. Publisher: American Chemical Society.

[7] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, January 2021.

[8] Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D. Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, pages 1–3, January 2022. Publisher: Nature Publishing Group.

[9] Stefano Grasso, Valentina Dabene, Margriet M. W. B. Hendriks, Priscilla Zwartjens, René Pellaux, Martin Held, Sven Panke, Jan Maarten van Dijl, Andreas Meyer, and Tjeerd van Rij. Signal Peptide Efficiency: From High-Throughput Data to Prediction and Explanation. *ACS Synthetic Biology*, 12(2):390–404, February 2023. Publisher: American Chemical Society.

[10] Songlyu Xue, Xiufang Liu, Yuyang Pan, Chufan Xiao, Yunzi Feng, Lin Zheng, Mouming Zhao, and Mingtao Huang. Comprehensive Analysis of Signal Peptides in Saccharomyces cerevisiae Reveals Features for Efficient Secretion. *Advanced Science*, 10(2):2203433, 2023.

[11] Weiwei Zhang, Mingming Yang, Yuedong Yang, Jian Zhan, Yaoqi Zhou, and Xin Zhao. Optimal secretion of alkali-tolerant xylanase in Bacillus subtilis by signal peptide screening. *Applied Microbiology and Biotechnology*, 100(20):8745–8756, October 2016.

[12] Ulf Brockmeier, Michael Caspers, Roland Freudl, Alexander Jockwer, Thomas Noll, and Thorsten Eggert. Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *Journal of Molecular Biology*, 362(3):393–402, September 2006.

[13] Keiro Watanabe, Yoshiki Tsuchida, Naoko Okibe, Haruhiko Teramoto, Nobuaki Suzuki, Masayuki Inui, and HideakiYR 2009 Yukawa. Scanning the Corynebacterium glutamicum R genome for high-efficiency secretion signal sequences. *Microbiology*, 155(3):741–750. Publisher: Microbiology Society,.

[14] Pamela O'Neill, Rajesh K. Mistry, Adam J. Brown, and David C. James. Protein-Specific Signal Peptides for Mammalian Vector Engineering. *ACS Synthetic Biology*, 12(8):2339–2352, August 2023. Publisher: American Chemical Society.

[15] Xin Yu, Merlinda Conyne, Marc R. Lake, Karl A. Walter, and Jing Min. In silico high throughput mutagenesis and screening of signal peptides to mitigate N-terminal heterogeneity of recombinant monoclonal antibodies. *mAbs*, 14(1):2044977, December 2022. Publisher: Taylor & Francis.

[16] José Juan Almagro Armenteros, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, 37(4):420–423, April 2019. Number: 4 Publisher: Nature Publishing Group.

[17] Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. arXiv:2206.13517 [cs, q-bio].

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.

[20] Chong Peng, Yixue Guo, Shaodong Ren, Cen Li, Fufeng Liu, and Fuping Lu. SPSED: A Signal Peptide Secretion Efficiency Database. *Frontiers in Bioengineering and Biotechnology*, 9, 2022.

[21] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20, pages 1–16, Atlanta, Georgia, November 2020. IEEE Press.

[22] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

A Appendix

A.1 Data

A.1.1 Training data

UniProt [7] contains 3,829 experimentally verified signal peptides as of release 2022_04. While this is sufficient for training discriminative classifiers with good performance, it makes a rather limited dataset for generative modeling, even more so when the aim is to model the differences between varying species. To overcome the scarcity of experimental data and enable coverage of many species, we instead augment the data by predicting UniProt reference proteomes using SignalP 6.0 [8]. From the reported performance, we expect a false positive rate of 5-15% for the main SP class Sec/SPI, depending on the organism group, which we consider acceptable for training generative models. Throughout the paper, the term SP refers to the class Sec/SPI. We predict signal peptides in the archaeal, bacterial and eukaryotic reference proteomes from UniProt release 2022_05. In total, we obtain 4,940,012 sequences with signal peptides for 11,097 organisms. We use MMseqs2 [19] to cluster the data at 30% identity and split it into 20 partitions, balancing for the organisms. We train on 18 partitions and validate and test on the remaining two, corresponding to a 90/5/5 split. While this is not a homology split in the strict sense, as there can be identities higher than 30% between members of two different partitions, it still minimizes train-test homology overlap compared to random approaches. We opt for MMseqs2 clustering because exact partitioning approaches scale poorly to a dataset of this size and overestimation of performance is less of a concern for LM perplexity metrics, compared to supervised models on smaller downstream task datasets.

We trim proteins to their first 120 N-terminal amino acids for training and testing SecretoGen.

A.1.2 Efficiency datasets

Grasso et al. The Grasso [9] dataset contains high-throughput measurements of 4,421 SPs with *B. amyloliquefaciens* α-Amylase expressed in *B. subtilis*. Efficiency was scored as a unitless enrichment metric based on high throughput sequencing of fluorescence-sorted constructs.

We obtained the train and test data of the Grasso et al. random forest efficiency predictor from the supplementary materials and extracted the mature protein sequence with correct N terminus from the provided G1 sequence (AmyQ, P00692). G1 excludes the 3 native mature n-terminal AAs valine-asparagine-glycine, as these positions form part of the library and are varied in each SP-protein construct. Insertion of SPs with BsmBI directly fuses them to the N terminus of the protein sequence. In this dataset, a low *WA* (weighted average of fluorescence bins) value indicates a high secretion efficiency. In order to match other efficiency metrics, where a higher value indicates higher efficiency, we inverted *WA* as $WA' = 10 - WA$.

Wu et al. The Wu [6] dataset contains 162 measurements of generated and natural SPs in combination with various enzymes, expressed in *B. subtilis*. Efficiency was determined as a binary value based on fluorescence assays on culture supernatants, indicating whether an SP resulted in secretion or not.

Synthetic SP and mature protein sequences were obtained from the provided supplementary tables. AprE, LipB, YbdG, YkvV and YvcE SP sequences were sourced from Zhang et al. [11], as referred to in the paper. For AprE, we follow Zhang et al. and Brockmeyer et al. [12] and keep a valine at position -3 (UniProt lists an alanine for -3). The YcnJ SP sequence was not found in Zhang et al. and sourced from UniProt. A glycine-alanine linker was added to the N terminus of mature protein sequences. Efficiency data was obtained from the swarmplots provided in the supplement.

Xue et al. The Xue [10] dataset contains 322 measurements of natural SPs with *A. oryzae* α-Amylase expressed in *S. cerevisiae*. Efficiency was determined as units/mL using a fluorescence assay of the supernatant.

SP sequences, the α-Amylase sequence and efficiency data were obtained through communication with the authors.

Table A1: Performance of perplexity and length bias corrected perplexity.

Metric	Grasso	Xue	Zhang	Wu
Perplexity	0.61	**0.49**	0.24	**0.70**
Z-scored perplexity	**0.63**	**0.49**	**0.25**	0.58

Zhang et al. The Zhang [11] dataset contains 114 measurements of of natural SPs with *B. pumilus* α-Amylase expressed in *B. subtilis*. Efficiency was determined as units/mL using a fluorescence assay of the supernatant.

SP efficiency data was downloaded in curated format from SPSED [20]. We followed the cloning protocol to derive the mature protein sequence. Starting from Xylanase sequence KU301789, we aligned the Xyn-Up primer to remove the native SP. Xyn-Up adds an EcoRI restriction site upstream of the protein. Complementarily, SPs are amplified using primers that add an EcoRI site downstream of the SP. The full constructs therefore have a glycine-serine linker between the SP and the mature protein. The 114 SPs were measured twice, using promoters P43 and Pglvm. The correlation between the two experiments is 97%. We report results on the Pglvm dataset.

A.2 SecretoGen details

Hyperparameters The SecretoGen model is a 12+12 layer encoder-decoder transformer with 16 attention heads in each layer, a hidden size of 1024 and a feedforward size of 2048. Models were trained for 30 epochs with a batch size of 1024 and a learning rate of 5×10^{-5} with linear warmup for 20000 steps and linear decay for 180000 steps using Adam. Checkpoints were selected based on the validation next token prediction loss. Dropout was set to 0.1. Training was done using DeepSpeed [21] with ZeRO stage 1.

dropout, optimizer etc.

Organism embedding layer Organisms are related by evolution and can be classified in various taxonomy levels. While in principle it should be possible to learn organism embedding vectors that correctly recapitulate their relative orientation to another, it might be hard for a model to learn so if only a very limited amount of data is available for any given organism, as it is the case for signal peptides. We instead propose to learn these embeddings in a hierarchical manner for more efficient sharing of information between different organisms. We learn embeddings for 8 levels L of the taxonomy tree (*Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, Species*) and compute the organism embeddings that are used for conditioning as their sum (Equation A1). Preliminary results had indicated that learning organism embeddings in this hierarchical manner has better convergence than learning categorical organism embeddings without sharing of information (Figure A1).

$$\mathbf{embedding}_{org} = \sum_{l \in L}^{L} \mathbf{embedding}_{l} \tag{A1}$$

Length dependency of perplexity While perplexity in itself is adjusted for length (it is divided by the sequence length), we find that there is a clear length dependency of perplexity values assigned by SecretoGen, with an optimum at lengths 16-19 (Figure A2). As this dependency is possibly an artefact of the training data distribution that does not represent a biologically meaningful effect on the level of individual sequences, we attempted to correct for this effect by following the methodology laid out in Durbin et al. [22] for correcting length biases in profile HMMs using Z-scores. We z-scored the perplexity at each length, with the mean and standard deviation of the perplexity estimated on the training data. On two out of four benchmark datasets, this correction yields a slight improvement, while it resulted in a substantial reduction on the Wu dataset (Table A1).

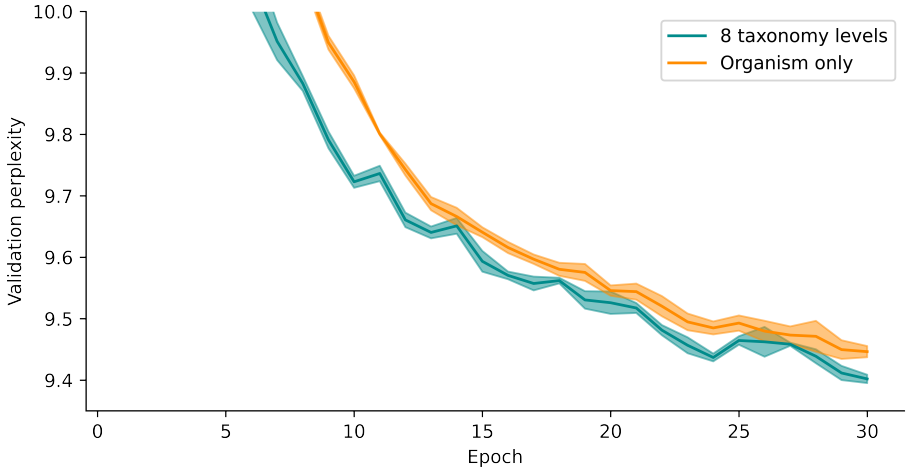

Figure A1: Learning organism embeddings as sums of taxonomy levels leads to better convergence. Experiments were performed using a smaller transformer model (6+6 layers, dimension 512) and a preliminary dataset of 4.2 million SPs. The means and standard deviations of three replicate runs with different random seeds are shown.

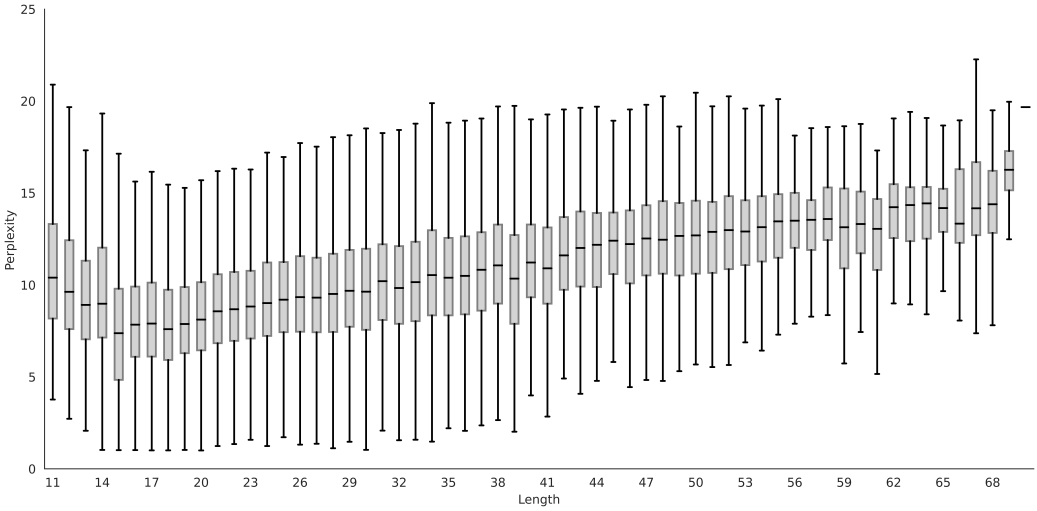

Figure A2: Perplexity is length dependent. The distribution of SecretoGen perplexity at each length is shown for the training set.

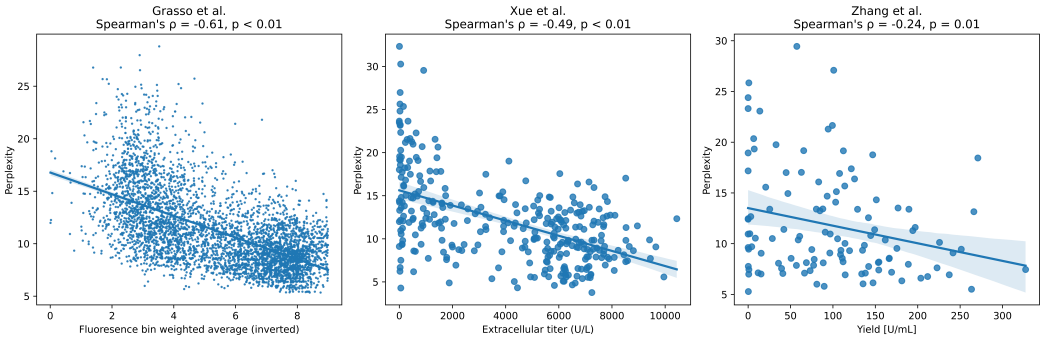

Figure A3: Length-adjusted SecretoGen perplexity vs. SP efficiency. Grasso and Xue show a bimodal distribution of efficiency values.

A.3 Baseline details

ProGen2 We evaluate multiple checkpoints of ProGen2. The checkpoints differ in their number of parameters (medium: 762M; large: 2.7B; xlarge: 6.4B). Additionally, ProGen2 models support computing the perplexity of the reversed sequence, as they were trained both on N-to-C and C-to-N formatted sequences. We report performance of both the N-to-C mode (`fwd`) and the C-to-N mode (`rev`).

SPGen SPGen provides multiple checkpoints with the same model size, differing in their training data. The checkpoints (`75,90,95,99`) refer to different percent sequence identity cutoffs that were used to remove sequences similar to the enzymes used in the study for evaluation.