ACS Synthetic Biology

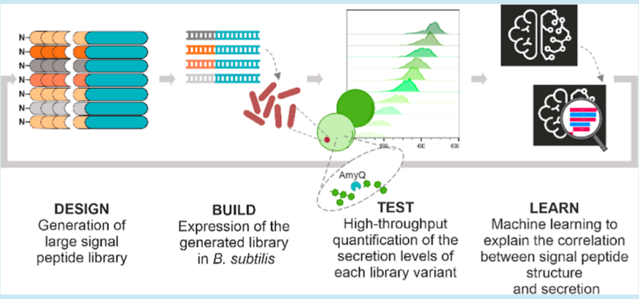
pubs.acs.org/synthbio Research Article

Signal Peptide Efficiency: From High-Throughput Data to Prediction and Explanation

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*

Cite This: *ACS Synth. Biol.* 2023, 12, 390−404 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information


N-
DESIGN
Generation of large signal peptide library
BUILD
Expression of the generated library in *B. subtilis*
AmyQ
TEST
High-throughput quantification of the secretion levels of each library variant
LEARN
Machine learning to explain the correlation between signal peptide structure and secretion

ABSTRACT: The passage of proteins across biological membranes via the general secretory (Sec) pathway is a universally conserved process with critical functions in cell physiology and important industrial applications. Proteins are directed into the Sec pathway by a signal peptide at their N-terminus. Estimating the impact of physicochemical signal peptide features on protein secretion levels has not been achieved so far, partially due to the extreme sequence variability of signal peptides. To elucidate relevant features of the signal peptide sequence that influence secretion efficiency, an evaluation of ∼12,000 different designed signal peptides was performed using a novel miniaturized high-throughput assay. The results were used to train a machine learning model, and a post-hoc explanation of the model is provided. By describing each signal peptide with a selection of 156 physicochemical features, it is now possible to both quantify feature importance and predict the protein secretion levels directed by each signal peptide. Our analyses allow the detection and explanation of the relevant signal peptide features influencing the efficiency of protein secretion, generating a versatile tool for the de novo design and in silico evaluation of signal peptides.

KEYWORDS: *signal peptide, protein secretion, amylase, Bacillus subtilis, nanoliter reactors, secretion efficiency*

■ INTRODUCTION

The general protein secretion (Sec) machinery is responsible for the translocation of bacterial, archaeal, and eukaryotic proteins across the cytoplasmic or endoplasmic reticular membranes.[1−5] Because of its high capacity for protein export from the cytoplasm, the Sec pathway of various microorganisms was engineered to generate cell factories for the production of commercially relevant secreted proteins, including enzymes and biopharmaceuticals.[2,6,7] Monoderm Gram-positive bacteria, like *Bacillus subtilis*, are preferred for this purpose, as products only need to pass a single membrane, which eases the secretion and subsequent recovery of bulk amounts of protein from the fermentation broth.[8−10]

N-terminal signal peptides (SPs) are responsible for guiding secretory proteins into the Sec pathway by interactions with chaperones, the membrane, and the membrane-embedded protein-conducting SecYEG channel and by maintaining an unfolded translocation-competent state of the translocated protein.[11] These ubiquitous targeting signals have been investigated for many years and, consequently, their structural features are well known. Essentially, SPs are composed of a positively charged N-region, a hydrophobic α-helical H-region, and a more polar C-region that frequently starts with a helix-breaking residue and that comprises the so-called signal peptidase (SPase) recognition site.[12−18] The functions of these three regions and the limits in their sequence variation were uncovered by extensive site-directed mutagenesis studies, involving the deletion or substitution of particular amino acids within the SPs of a range of different exported proteins.[19] Thus, positively charged residues in the N-region were shown to promote the initiation of protein translocation, explaining why introduction of negative charge in this region interferes with protein translocation.[20−22] The H-region promotes loop-like insertion of SPs into the membrane, explaining why deletion of the hydrophobic residues or insertion of charged residues interferes with productive protein export,[23,24] and why a turn-inducing residue like Gly is often present in the center of the H-region.[12−14,25,26] Furthermore, the presence of positively charged residues at the C-terminal end of the H-region interferes with effective protein translocation via Sec, as shown through studies on SPs that target proteins to the alternative twin-arginine protein translocation (Tat) pathway.[27,28] Amino acid substitutions in the polar C-region

Received: June 19, 2022
Published: January 17, 2023

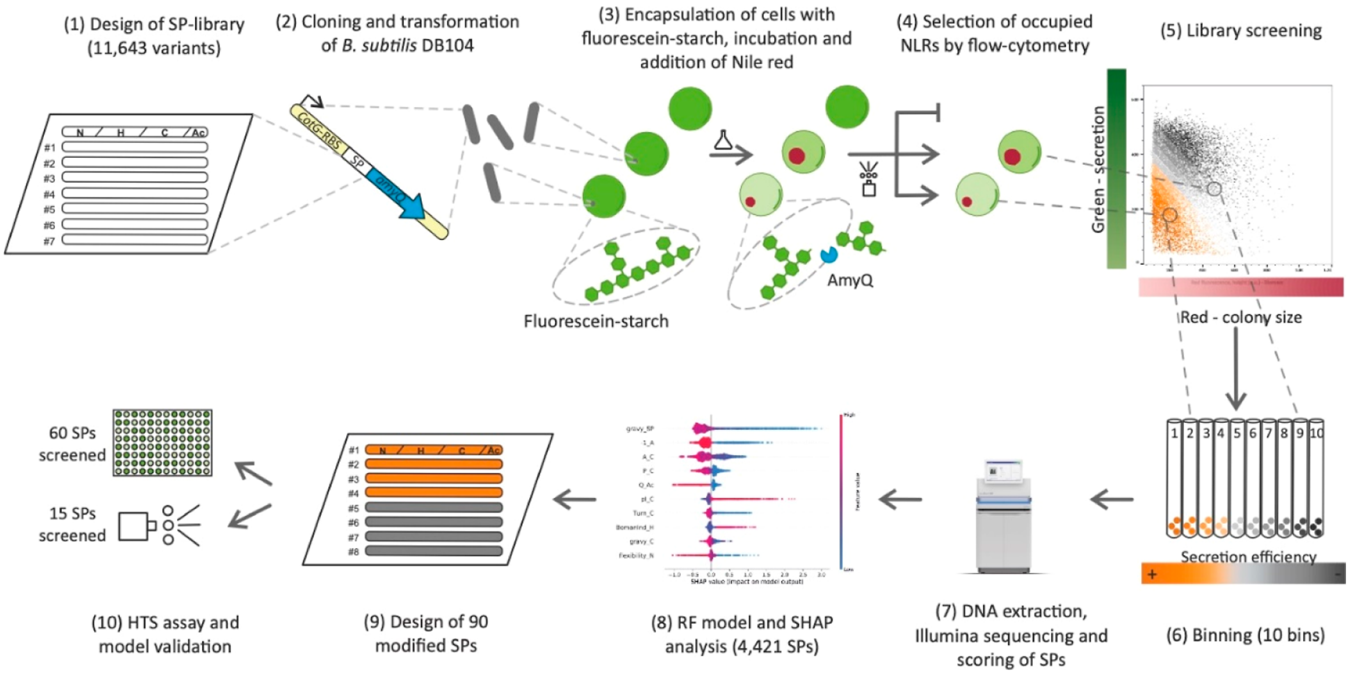

(1) Design of SP-library (11,643 variants)
(2) Cloning and transformation of B. subtilis DB104
(3) Encapsulation of cells with fluorescein-starch, incubation and addition of Nile red
(4) Selection of occupied NLRs by flow-cytometry
(5) Library screening
N / H / C / Ac
#1
#2
#3
#4
#5
#6
#7
CotG-RBS
SP
amyQ
Fluorescein-starch
AmyQ
Green - secretion
Red - colony size
1 2 3 4 5 6 7 8 9 10
Secretion efficiency
+
-
60 SPs screened
15 SPs screened
#8
SHAP value (impact on model output)
Feature value
High
Low
(10) HTS assay and model validation
(9) Design of 90 modified SPs
(8) RF model and SHAP analysis (4,421 SPs)
(7) DNA extraction, Illumina sequencing and scoring of SPs
(6) Binning (10 bins)

Figure 1. High-throughput characterization of the SP library. Experimental workflow: (1) A library of approximately 12,000 SP variants was designed by modifying key features (e.g., charge, length, and hydrophobicity); (2) the corresponding pool of oligonucleotides was cloned in frame with the sequence coding for mature AmyQ, and integrated into the *amyE* locus of *B. subtilis* DB104. (3) Clones were embedded in hydrogel beads, referred to as NLRs, containing fluorescein−starch (mean diameter of 500 μm; average occupation of 0.3 bacterial cells per NLR). During incubation in culture medium, single cells grew into microcolonies and secreted AmyQ, which degrades the fluorescein−starch into (still fluorescent) low molecular weight fragments that are lost from the NLR by diffusion. After incubation, biomass in the NLRs was labeled by adding nile red, a membrane-specific red fluorescent dye, and the NLRs were evaluated in two steps using a large particle flow cytometer. (4) First, all empty NLRs were identified and discarded; (5,6) second, occupied NLRs were sequentially sorted into 10 bins, based on their green to red signal ratio. The green fluorescence signal is inversely proportional to the amount of secreted amylase (AmyQ) in the NLR; the red signal is instead directly proportional to the colony size. Therefore, clones with a high secretion efficiency are located in the lower left corner of the dot plot (5) and have a low bin number. (7) DNA from the NLRs of each bin was recovered and SP occurrence in any given bin was determined by NGS, leading to the construction of a frequency table of SPs across bins, used to calculate the secretion efficiency of each SP variant as a WA. (8) WA values were subsequently combined with the features describing each SP to train a RF regressor model. The RF model was then studied using SHAP for explanation and quantification of the impact of each feature on the model output (i.e., WA). (9) Information obtained by combining the RF model with the SHAP analysis was used to generate new SP variants with defined secretion levels to validate the model. (10) Designed validation sequences were processed following the same HT screen, yet individually and not as a library. The secretion of amylase was quantified both with a MTP assay (60 SPs) and by the NLR-based screening protocol (15 SPs), and the results were compared to the predictions.

revealed the universally conserved SPase recognition site Ala^{-3}−X−Ala^{-1}, where X can be any amino acid, and −3 and −1 define the residue positions relative to the SPase cleavage site.[29] These studies also showed that a turn-inducing Pro residue in the C-region is important for SP cleavage by SPase,[29] which defines the N-terminus of the exported protein and is essential for its subsequent release from the membrane into the extracellular environment.[12−14,30] Despite these conserved features, species-specific variations have been observed, most notably differences in SP length and the SPase cleavage site.[15] Thus, SPs of Gram-positive bacteria, such as *B. subtilis*, are among the longest known SPs.[12−14,30]

All gathered knowledge about the SP structure allowed the development of algorithms that reliably detect the presence of a SP within a protein sequence.[31] However, there are no tools to predict the efficiency with which a given SP directs the secretion of a target protein of interest (POI).[32] In fact, finding an efficient SP to secrete a POI is currently still based on trial-and-error. Previous studies tested limited numbers of natural SP variants (i.e., up to 10^2) and analyzed the relationships between secretion efficiency and some SP features.[33−35] Such studies showed that the SP−POI match plays a crucial role in determining secretion efficiency,[14] but they did not unveil the underlying fundamental parameters.

In recent years, machine learning (ML) and deep learning methods have been successfully leveraged to build tools to identify and classify peptides with different types of functions, such as antimicrobial activity[36] and antigen presentation by the major histocompatibility complex.[37] Here, the common theme is that a problem is often tackled with different approaches that are then benchmarked against each other. Some of the most common and successful ML models used in peptide prediction are the support vector machines[38−40] and random forests (RFs).[40−42] Recently, deep learning-based approaches have become popular,[43−46] also for SP prediction.[47,48] A relevant aspect in predictive models is the description of a peptide or protein used to train them, which can be based on information about its physicochemical properties,[38,41,46] its sequence,[44] and/or its structure.[38] Such information can be further encoded and fed to the algorithm in different ways. Common encoding methods include (pseudo-)amino acid composition[39,40,42−45] and positional matrices.[38,39] However, different models and feature-encoding methods may need to be applied, as each problem requires different and tailored combinations of tools.

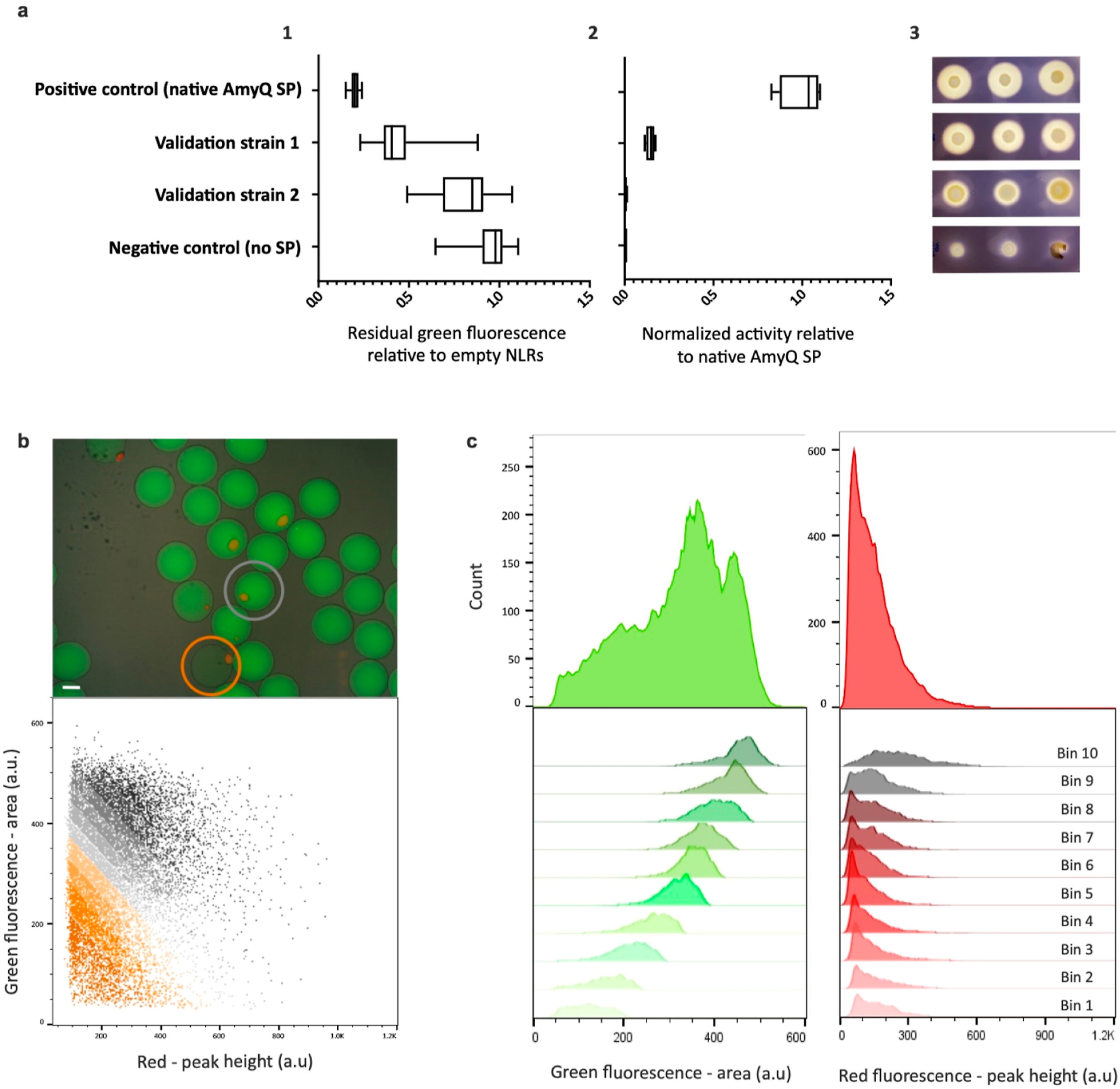


Figure 2. HT SP library screening in NLRs. (a) For initial validation of the NLR-based α-amylase assay, four *B. subtilis* strains secreting AmyQ to different levels were used: three strains with known SP amino acid sequences at the N-terminus (one of them with the native SP of AmyQ; positive control) and a *B. subtilis* strain synthesizing the amylase without an N-terminal SP (negative control). Amylase secretion of each strain was assessed using the (1) NLR-based assay, (2) MTP colorimetric assay, and (3) starch hydrolysis test. (1) For the NLR-based assay, the values represent the residual fluorescein-labeled starch still present in the occupied NLRs after cell growth, relative to the green fluorescence of the empty NLRs in the same population (set as 1). The recorded events were positive control, 24 occupied and 850 empty NLRs; validation strain 1, 99 and 4329; validation strain 2, 50 and 932; negative control, 124 and 859. (2) For the MTP assay, the values are calculated relative to the amylase activity produced by the positive control (having a value of 1) and four biological replicates were performed. (3) Starch hydrolysis tests based on the starch–iodine reaction.[23] (b) Top: overlay of bright-field and fluorescence microscopy images of NLRs after incubation in medium. Empty NLRs (no red dot) show a homogenous green fluorescence profile (no starch degradation), while NLRs harboring a colony (red dot) show different degrees of fluorescein-labeled starch degradation (orange circle: high secreter; gray circle: low secreter). Scale bar: 200 μm. Bottom: dot plot representing all occupied NLRs from one experiment (approximately 20,000 NLRs). The gating applied during the second sorting step is depicted in orange-gray color codes, which defines bins with distinct AmyQ secretion levels. (c) Green and red fluorescence profiles of all sorted events from the same experiment, both as a whole population (i.e., occupied NLRs; top panel) and divided into 10 equally sized bins (lower panel).

While ML and deep learning approaches can be used to predict a wide variety of biological functions of peptides, including SPs, it is important to bear in mind that there is also a need to understand their mechanisms of action and the contextual interplaying factors. Therefore, our present study was not only aimed at predicting the efficiency of various SPs fused to a particular POI, but also to elucidate the relevant physicochemical features of SPs that determine the secretion

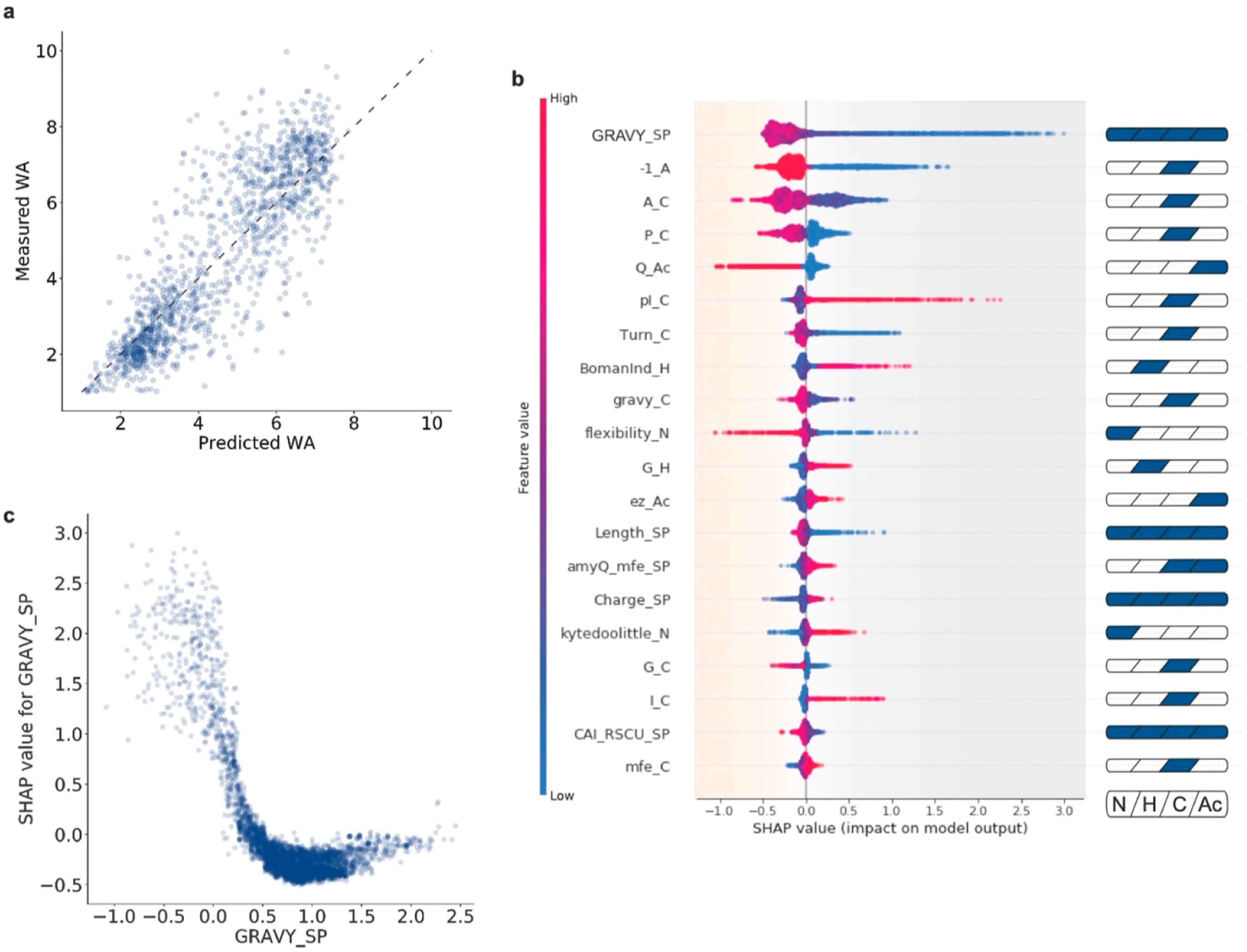


Figure 3. Model explanation. (a) Predictions of the trained RF regressor model on the test set. On the ordinate, WA values for SPs belonging to the test set and measured with the NLR-based amylase assay are reported; on the abscissa, their WA values predicted with the trained RF model are displayed. Note that measurements and prediction show a high degree of agreement with a calculated MSE of 1.22 WA, thus indicating a good performance of the generated RF model. The dashed line represents the ideal situation where all predicted and measured values would align. (b) SHAP summary plot of the 20 most impactful features: **GRAVY_SP**, overall SP hydrophobicity; **−1_A**, Ala at −1 of the AxA cleavage site; **A_C**, frequency of Ala in the C-region; **P_C**, frequency of helix-breaking Pro in the C-region; **Q_Ac**, frequency of glutamine in the Ac-region; **pI_C**, pI of the C-region; **Turn_C** indicates a helix-breaking residue at the end of the H-region; **Bomanlnd**, protein−protein interaction in the SP; **Gravy_C**, hydrophobicity of the C-region; **Flexibility_N**, measure for flexibility and charge in the N-region; **G_H**, frequency of Gly in the H-region; **Ez_Ac**, potential for Ac-region insertion in lipid membranes; **Length_SP**, overall length of the SP; **amyQ_mfe_SP**, minimum folding energy of the RNA secondary structure encoded by the *sp-amyQ* gene fusion; **Charge_SP**, charge of the SP; **Kytedoolittle_N**, hydrophobicity of amino acids in the N-region; **G_C**, frequency of Gly in the C-region; **I_C**, frequency of Ile in the C-region; **CAI_RSCU_SP**, codon adaptation index of the SP; **Mfe_C**, minimum folding energy of the RNA secondary structure in the C-region-encoding sequence (for full descriptions, see Supporting Information Table S2). A high dispersion of SHAP values on the abscissa indicates a broad effect of the respective feature on the model. Each data point represents a specific SP, the color of the data point indicates the value of that feature in the feature-specific scale, and the position on the abscissa indicates the SHAP value for that particular feature. SHAP values for the whole data set sum up to the base value of the model (4.45 WA, average model output calculated over the 4421 selected SPs). Positive SHAP values indicate a negative impact on the model outcome and vice versa. Cartoons on the right highlight the corresponding SP parts of each particular feature. (c) SHAP-dependence plot for "GRAVY_SP", which is a two-dimensional representation of the information summarized by the first line of panel b. The GRAVY index is represented on the abscissa: negative values indicate low hydrophobicity, and positive ones indicate high hydrophobicity. On the ordinate, SHAP values are displayed: negative values indicate a beneficial effect on protein secretion, and positive ones indicate a detrimental effect. Vertical dispersion of SHAP values for similar GRAVY indexes can be explained through the interaction effect between features (described by SHAP interaction values, summarized in Supporting Information Figure S7 and Supporting Information Figure S8). To exemplify, the high variability visible in the negative range of the GRAVY index is to be attributed mainly to the feature "−1_A" (Supporting Information Figure S6). The data imply that a very low hydrophobicity will have a strong negative impact on protein secretion, while a GRAVY index value of around 1.0 will be most favorable for protein secretion.

efficiency of the POI. To achieve our aim, we devised a workflow (Figure 1), based on the Design−Build−Test−Learn (DBTL) cycle approach. A SP library was designed using 134 known wild-type SPs from *B. subtilis* as template, but we expanded the SP diversity by including targeted modifications in the SP sequences that altered physicochemical features in either the entire SPs or the different SP regions, with minimal effect on other features. In particular, the modified features included the amino acid composition and physicochemical properties of the SPs and addressed the SPs both at the amino acid and nucleotide sequence levels. This SP library was fused to the secreted α-amylase AmyQ from *Bacillus amyloliquefaciens*, and high-throughput (HT) quantification of the AmyQ secretion efficiency[49] was then used to generate a training data set for a ML model. Finally, the impact of physicochemical features on the secretion efficiency was estimated using TreeSHAP[50−53] (hereafter SHAP).

RESULTS AND DISCUSSION

High-Throughput SP Library Screening in Nanoliter Reactors. Starting from the selection of 134 known wild-type SPs from *B. subtilis*, a library of 11,643 unique SPs (Supporting Information Table S1) was rationally designed to expand the sampling space and the variance of naturally occurring SP sequences. We individually modified 7 specific physicochemical features on 94 designated levels (Supporting Information Table S2), while concomitantly minimizing their influence over related ones (e.g., editing the charge while avoiding a significant hydrophobicity change). Furthermore, each SP was treated both as a single sequence, and as four juxtaposed segments that included the N-, H- and C-regions plus a short stretch of three residues after the SPase cleavage site (i.e., the Ac-region). The designed SP-library was then introduced into *B. subtilis* strain DB104 using a genome-integrating vector. A total of 160,000 clones was harvested, achieving a 10X coverage of the SP-library.

The secretion efficiency associated with each SP variant was determined by measuring amylolytic degradation of fluorescein-labeled starch upon encapsulation of the library strains in so-called nanoliter reactors (NLRs).[54,55] In this assay, green fluorescence of each NLR is rapidly measured via flow cytometry, allowing the assessment of secretion levels of active AmyQ (Figures 2 and S1, Supporting Information). For initial validation, the HT methodology was compared to two alternative assays: a microtiter plate (MTP) format using a synthetic substrate (Figure 2a, 2) and a starch hydrolysis test using agar plates[56] (Figure 2a, 3). The results of the different assays are comparable and show the highest sensitivity and dynamic range for the NLR-based assay.

Next, we performed the HT screening with the SP library measuring simultaneously amylase activity and bacterial biomass. The latter was achieved by incubating the NLRs with nile red, a hydrophobic red dye that interacts with cell membranes[57] and fluorescently labels the microcolonies. Occupied NLRs were separated into 10 equally populated bins, based on enzymatic activity per biomass unit (i.e., the ratio between green and red signals). Variant collection was followed by DNA sequencing to determine the abundance of each SP variant in each bin and, as a control, in the original library after transformation and before sorting the NLRs. Occurrence values were used to generate a weighted average (WA), assuming equidistance between bins, and this WA was used as an efficiency score for each SP.

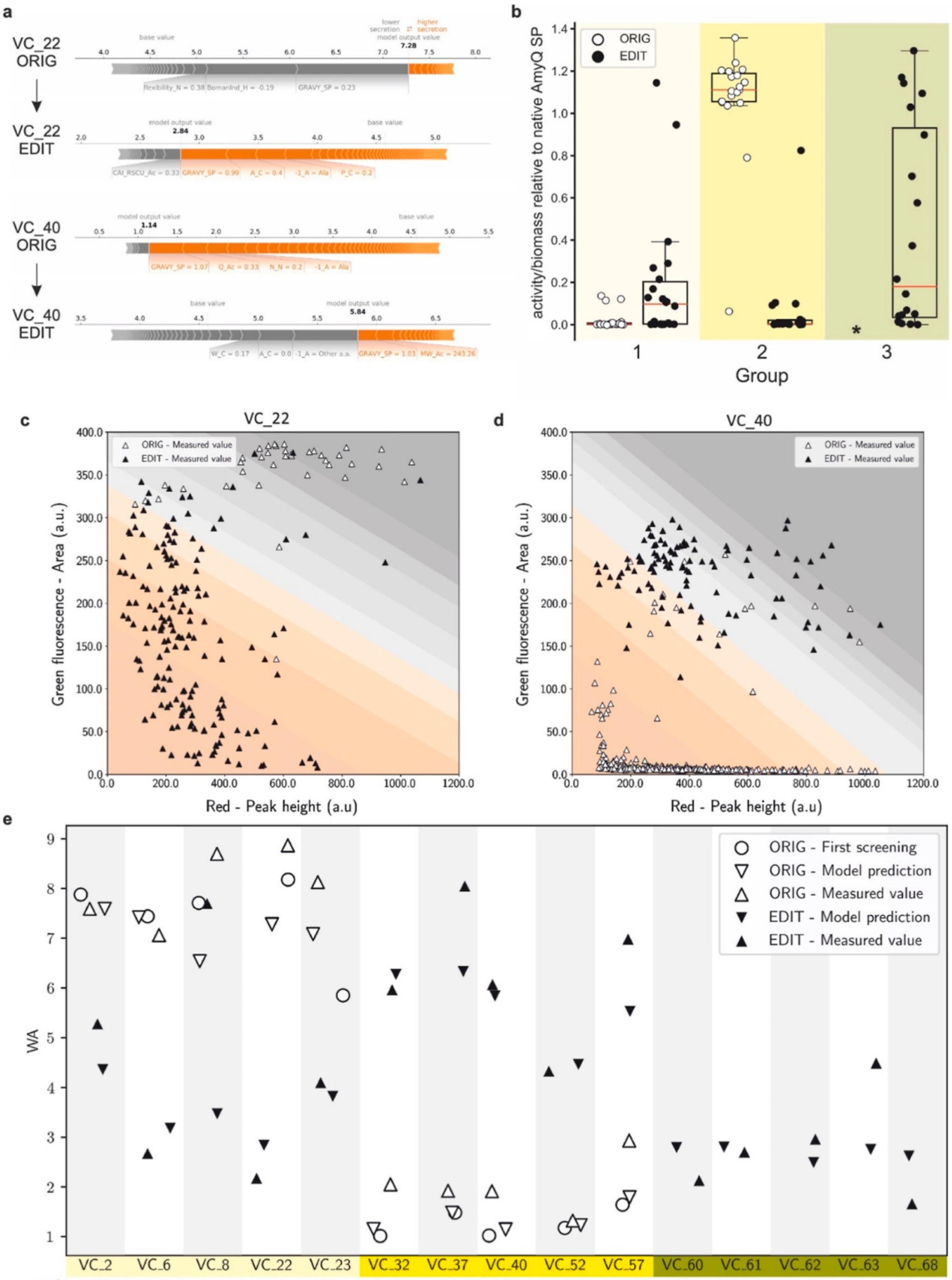
As shown by sequencing, 92% of the 11,643 unique rationally designed SPs were successfully introduced into *B. subtilis*, while 83% were retrieved after screening (Supporting Information Table S3). Such reduction may relate to SP-dependent impaired growth and the resulting high background-to-noise ratios for small colonies.

We subsequently characterized the sensitivity of the NLR-based secretion assay by comparing values from 95 randomly picked library clones in NLRs versus MTPs (Supporting Information Figure S2). As 73 of the selected 95 variants could not be measured using the standard MTP format, we applied also a classical starch hydrolysis test on agar plates to verify the low-secreting variants (Supporting Information Figure S3). These experiments highlighted the superior sensitivity of the NLR-based secretion assay.

Machine Learning Model to Predict SP Efficiency and Model Explanation. To test and train our ML model, we evaluated the number of physicochemical features in the SP data set. Starting from an initial set of 267 features, 156 informative features were retained to describe each SP (Supporting Information Table S2). This step removed features either presenting no variability or exhibiting a high correlation with another feature in the training data set. A further reduction of dimensionality proved to be unnecessary, as the PCA analysis showed that each of the principal components contributed to the explained variation (Supporting Information Figure S4). Additionally, the same number of components was necessary to describe the whole variance of the 11,643 unique rationally designed SPs and the 4421 informative SPs, indicating that, despite the loss in the total number of data points, there was no loss in the variation of the data set. In contrast, the PCA showed that, to explain the same variation, more principal components are needed within the designed library than for the wild-type set of SPs, underpinning the improvement of the assayed space gained with our design. The array of 156 features is thus to be considered as the independent variable, and the single value of secretion efficiency (WA), as the dependent one.

Three-quarters of our data set were used to train a RF regression algorithm, resulting in a mean squared error (MSE) of 1.75 WA, while the remaining quarter was used as a test set, resulting in an MSE of 1.22 WA (Figure 3a). After this first validation, we proceeded to provide explanations for the RF model predictions. Due to the complexity in explaining such a developed RF model,[58,59] SHAP[50−52] was used to extract information about the importance of the features and their interaction effects (Supporting Information Figure S5).

Due to the large number of features fed into the model and the notable amount of information provided by SHAP, only the most relevant and representative findings are discussed. To fully explore the model, a Jupyter notebook and an interactive tool (File S1) are available as Supporting Information data. The 20 most impactful features in our model are shown in Figure 3b. Some of these features were already documented in literature,[32,60,61] for instance, the overall SP hydrophobicity ("GRAVY_SP"), the helix-breaking residue at the end of the H-region ("P_C" and "Turn_C"), or the cleavage consensus sequence (e.g., "−1_A" and "A_C"). Notably, even for such known features, the impact on secretion could so far only be qualitatively estimated based on their distributions in wild-type SPs. With our approach, a more precise quantification is now achieved, establishing favorable, neutral or detrimental values and their impact on the predicted secretion efficiency. Additionally, it is now possible to determine relationships between features and secretion efficiency (e.g., linear, sigmoidal, and monotonic), as illustrated with the dependency plot for a simple feature, such as "GRAVY_SP" (Figure 3c), whose wild-type distribution is known but only includes positive values.[61,62] Our model analysis shows that functional positive GRAVY values are favorable, while negative GRAVY values will be detrimental. Moreover, thanks to the applied



	VC_2	VC_6	VC_8	VC_22	VC_23	VC_32	VC_37	VC_40	VC_52	VC_57	VC_60	VC_61	VC_62	VC_63	VC_68
○	7.88	7.44	7.71	8.18	5.85	1.01	1.48	1.02	1.17	1.64	N.A.	N.A.	N.A.	N.A.	N.A.
▽	7.59	7.42	6.54	7.28	7.08	1.15	1.48	1.14	1.23	1.79	N.A.	N.A.	N.A.	N.A.	N.A.
△	7.59	7.06	8.70	8.87	8.13	2.05	1.92	1.91	1.32	2.93	N.A.	N.A.	N.A.	N.A.	N.A.
▼	4.36	3.18	3.47	2.84	3.82	6.27	6.33	5.84	4.46	5.53	2.79	2.80	2.49	2.75	2.62
▲	5.28	2.67	7.70	2.17	4.10	5.96	8.05	6.07	4.33	6.98	2.13	2.70	2.96	4.49	1.66

Figure 4. Assay and model validation. (a) SHAP force plots for SP variants VC_22 and VC_40 before ("ORIG") and after ("EDIT") editing. The impact of each of the most relevant features on secretion efficiency is quantitatively assessed. Each segment is sized proportionally to its impact on the model; their summation is equal to the difference between the base value (4.45 WA for all SPs) and the output value (i.e., the predicted WA value of each SP). Features colored in gray have a negative impact on the secretion efficiency of the specified SP, while features colored in orange have a positive impact. (b) Box plot showing amylase activities of the three groups of SPs used for validation as measured by MTP assays: group 1, 30 originally (ORIG) poorly secreting SPs edited into 30 improved SPs (EDIT) (light yellow); group 2, 30 originally (ORIG) highly efficient SPs edited into 30 poorly secreting SPs (EDIT) (yellow); and group 3, 21 pseudo-randomly designed SPs (dark yellow). The circles represent the average amylase activity (measured with the MTP assay, relative to the efficiency of the native SP of AmyQ defined as 1) of individual SPs before (white circles, "ORIG") and after (black circles, "EDIT") editing (*, ORIG versions of pseudo-randomly designed SPs do not exist, since they were not present in the original SP library). (c,d) Dot plots showing the individual recorded events of NLR-based analyses; panel c shows results obtained with SP VC_22 ORIG (white triangles) and VC_22 EDIT (black triangles), whereas panel d shows results obtained with SP VC_40 ORIG (white triangles) and VC_40 EDIT (black triangles). The numbers of recorded events were 43 for the VC_22 ORIG; 174 for VC_22 EDIT; 411 for the VC_40 ORIG; and 103 for the VC_40 EDIT). White triangles indicate NLRs harboring strains secreting AmyQ with the original

Figure 4. continued

variant from the SP-library (ORIG), while black triangles indicate NLRs with strains secreting AmyQ with the edited SP (EDIT). In the background, the 10 different bins are indicated using the same color code as in Figure 2b. (e) Summary plot of the model validation. For each of the 15 selected SPs, 5 data points are shown: open symbols indicate the original variant from the SP-library, while black symbols designate the engineered SP derivative. The open circles mark the WA measured during the initial screening, triangles pointing downward denote WA values predicted by the model, whereas triangles pointing upward denote WA values measured during model validation. Groups of SPs are highlighted in different shades of yellow as in (a). Below the graph, values are listed to allow a more detailed comparison.

segmentation approach (i.e., four juxtaposed regions for each SP), it is possible to visualize how particular features can have different relevance, depending on whether we consider the whole SP or only a single region. This is clearly exemplified by the feature "Charge_SP", for which an overall value lower than +2 increases secretion efficiency, and a slightly negative charge is even more favorable. In contrast, inspection of the charge of the N-region, represented by "Flexibility_N" (Supporting Information Table S2), shows that values close to +2 or higher favor secretion. Analogously, different features describing the same region can influence each other's impact. For instance, the feature "BomanInd_H", which positively correlates with charge and negatively with hydrophobicity (Supporting Information Tables S1 and S2), shows that a high level of hydrophobicity (low "BomanInd_H") in the H-region can favor secretion. At the same time, judged by the feature "G_H" (Gly content in the H-region), it appears that high hydrophobicity due to a high Gly content is not favorable, most likely because Gly reduces the stability of α-helices. Because of the applied feature selection process, one feature (e.g., "BomanInd_H") may be representative of similar ones (e.g., "pI_H" and "Charge_H"), which sets a limit to our immediate understanding of the influence of some properties. Nonetheless, with the present approach, we can retain, explain, and trace back to their correlating counterparts, physicochemical properties of SPs, rather than less biologically significant indicators (e.g., principal components of a PCA or the "D-score" from SignalP[33,63]).

With SHAP, it is possible to analyze and quantify pairwise interactions between features, which explains why equal values of the same feature can influence the model to different extents. For instance, the vertical dispersion of "GRAVY_SP", which is especially pronounced for negative hydrophobicity values (Figure 3c), is to be attributed to the feature "−1_A" (Supporting Information Figure S6). Furthermore, overall interactions seem to play minor roles in our model, as the most impactful interaction ("Q_Ac"−"A_C") has limited impact on the overall output (Supporting Information Figures S7 and S8). One possible explanation is that interactions occur at orders higher than the second and as such would not be represented in the model.

Assay and Model Validation Through Rationally Edited and Pseudo-Randomly Designed SPs. To validate our model, we decided (i) to rationally tune the secretion efficiency of screened SPs and (ii) to in silico-screen a library of pseudo-randomly designed SPs for high secretion levels. From the previously screened SPs, we selected 30 sequences that poorly (group 1) and 30 that highly (group 2) directed AmyQ secretion, and we manually modified their nucleotide and amino acidic sequences to invert their efficiency (Figure 4a and Supporting Information Table S4). An interactive exploration of original and edited SPs is possible through File S2 (original) and File S3 (edited). Additionally, we generated 4,903 pseudo-randomly designed SPs, predicted their secretion efficiency, and picked 32 among the potentially best-performing SPs (average WA of selected SPs is 2.64) to be tested (group 3). Out of the 92 SPs selected, 39 of the manually modified (groups 1 and 2) and 21 of the newly designed SPs (group 3) were successfully cloned and tested for amylase activity in the MTP assay, showing substantial difference compared to their original counterparts (groups 1 and 2) and very effective secretion (group 3), respectively (Figure 4b). Remarkably, out of the 21 in silico pseudo-randomly designed SPs (group 3), 5 showed a secretion efficiency higher than AmyQ with its native SP.

To further validate the quality of the model in predicting SP efficiency for AmyQ secretion (i.e., WA value), we selected five SPs from each of the three groups and analyzed their behavior in the NLR-based assay. Dot plots of two variants with the respective original ("ORIG") and manually edited ("EDIT") versions are shown in Figure 4c,d, respectively, and highlight the clear shift in secretion efficiency for the two versions depicted in Figure 4a,b. Figure 4e summarizes the WAs of the 15 selected SPs and compares them with the WAs obtained at each step of the workflow (i.e., library screening, model predictions, and validation). Remarkably, 11 of the tested SPs fell within one unit of difference (i.e., ±1 WA) from the predicted value, implying that the proposed workflow is indeed a powerful tool to quantify the efficiency of engineered SPs.

The amylase quantification experiments in MTP cultivations show that, although the model was trained only with NLR-based data, its predictions mostly retain their validity at larger scale. To the best of our knowledge, we have thus achieved an unprecedented accuracy in the prediction of SP efficiency and present the first example of successful model-driven de novo design of highly effective SPs. In fact, in silico SP design based on our trained model already proved very effective in the present proof-of-principle study, since the best predicted SPs turned out to direct high-level secretion.

CONCLUSIONS

Altogether, we conclude that our approach can detect and explain the relevant SP features that influence the efficiency of protein secretion. It sets the stage for in silico tuning and de novo design of SPs. Although we limited our present study to one protein, the workflow can be extended to other industrially or biomedically relevant POIs by applying different enzymatic assays[64] and novel HT analytical systems.[49,65,66] For the future, we advocate an iteration of our approach to obtain further insights into the general features that influence protein secretion. This may be achieved either by using a fractional factorial design[67,68] to ameliorate the design space (e.g., combining regions with differently modified features rather than editing one at a time) or through the screening of different POIs (e.g., fused to the same set of SPs). Data sets thus obtained will further improve the generalizability and reliability on prediction and design of SPs directing high secretion levels. As a result, far smaller numbers of SPs will be screened, or SP sequences will directly be designed in silico, for instance with our pseudo-random approach, or by exploiting a novel ML-based tool for SP generation.[69] We are therefore confident that, with less experimental testing, our approach will deliver a deeper understanding of SP function and more accurate, better tunable, and highly productive protein secretion systems.

■ MATERIALS AND METHODS

Library Design. To identify the most relevant physicochemical features influencing the secretion efficiency directed by SPs, we designed a SP mutant library starting from 134 sequences (Supporting Information Table S1) of known or highly probable *B. subtilis* wild-type SPs. These SPs were initially selected based on literature[33] and via predictions by various computational tools (SignalP4.1,[70] SignalP3.0,[63] Phobius[71]). Next, the selected SPs were manually curated to remove false positives, knowing the final localization of their cognate native protein. As a point of novelty, we considered the SP sequences both as a single sequence (i.e., the whole SP) and as the juxtaposition of four separate parts, namely, the canonical N-, H-, and C-regions, and a region referred to as "after cleavage" (Ac-region), which consists of the first three amino acid residues (AAs) after the expected SPase cleavage site. The Phobius tool for SP predictions was used to determine the boundaries of the four regions constituting each SP, still with partial manual curation based on evidence from literature. After defining the four regions for each SP, physicochemical properties were calculated for each region independently as well as for the complete SP. The 227 calculated properties are listed in Supporting Information Table S2, while the respective methods of calculation and further explanations are reported in Supporting Information Table S1.

From each of these 134 SPs, 94 mutant sub-libraries of 134 elements each were created. In each sub-library, only one feature at a time was edited, while modifications to other interdependent features (e.g., the charge of an AA sequence affects also its isoelectric point and hydrophobicity) were minimized. Edited features at the AA level were hydrophobicity, charge, and length; edited features at the nucleotide level were codon usage and RNA secondary structures. The full list of varied features is presented in Supporting Information Table S2. For each selected feature, multiple target levels (usually four or five) were chosen. The rationale for selecting target levels was to allow for some expansion of the investigated design space without diverging too much from the biologically meaningful space of the wild-type SPs. The resulting SP-library was composed of the 94 sub-libraries and included a total of 11,643 unique sequences, which are presented in Supporting Information Table S1.

pSG01 Plasmid Construction. The plasmid pSG01 (Supporting Information Figure S9, and see Supporting Information Table S5 for the full plasmids list) was developed within this study in order to be used as a chromosomal integration vector for expression of the SP-library. To this end, the previously constructed genome-integrating vector pCS75[72] (Supporting Information Table S5) was cleaved with PmeI and EagI (NEB); the resulting fragments were separated on a 0.8% agarose gel, and the 7.8 kbp band, delimited by two regions homologous to the *B. subtilis amyE* gene, was excised and purified with the QIAquick gel extraction kit (Qiagen). The DNA sequence encoding the AmyQ mature protein (P00692) (i.e., without its SP) was ordered as a single gBlock G1 (Integrated DNA Technologies, Inc.) (see Supporting Information Table S5 for the full nucleotide sequence), amplified with primers P1 and P2 (see Supporting Information Table S5 for a full primer list), digested with the same restriction enzymes as the vector and purified with the DNA clean & concentrator-25 kit (ZymoSearch). The two DNA fragments were ligated, and the ligation mix was directly used to transform 10-beta competent *Escherichia coli* cells (NEB), to amplify pSG01. The resulting plasmid was verified and used to transform dam$^-$/dcm$^-$ competent *E. coli* cells (NEB), from which demethylated pSG01 was obtained for all downstream applications to increase the efficiency of *B. subtilis* transformation.[73] Notably, 5′ to the SP-less *amyQ* gene, plasmid pSG01 contains two BsmBI (a type IIS restriction enzyme) restriction sites at 11 nt distance, which are oppositely oriented so that cleavage occurs upstream of each restriction enzyme recognition sequences, thus allowing for scar-less insertion of properly oriented DNA fragments. Moreover, this feature allows for the insertion of multiple DNA fragments in one step. After transformation of *B. subtilis*, pSG01 will integrate into the *amyE* gene, thereby disrupting the main source of amylase activity in *B. subtilis*.

Expression Strains and Cloning of the Library. *B. subtilis* strain DB104,[74] which lacks two major extracellular proteases, was selected to produce the library of designed SPs fused to AmyQ.

To obtain the final SP-library, pSG01 was endowed with two DNA fragments, using the two BsmBI restriction sites upstream of the SP-less *amyQ* gene: one fragment contained the P_{veg} promoter,[75] the native mRNA stabilizer of *cotG*,[76] and a strong RBS from the *pre*(*mob*) gene of pUB110,[77] obtained as a single gBlock G2 (Integrated DNA Technologies, Inc.; Supporting Information Table S5); the other fragment coded for one of the 11,643 designed SPs (ordered as an oligo pool from Twist Bioscience). Both fragments were designed to be amplified with P1 and P2 primers (Supporting Information Table S5) and to present two terminal BsmBI cleavage sites generating complementary sticky ends to the vector for sequential assembly. Cloning was carried out using the StarGate[78] methodology, and the resulting construct constitutively expressed the gene coding for the mature AmyQ fused at the N-terminus with one of the 11,643 designed SPs. A total of 3 mL of StarGate reaction was mixed with 63 mL of competent *B. subtilis* DB104 that has been prepared using a modified Spizizen protocol.[72] After 1 h of recovery at 37 °C and 250 rpm, cells were plated on 62 Q-trays (Nunc Square BioAssay Dishes product n. 240835, Thermo Fisher) each containing 200 mL of 2xPY medium (16 g/L peptone, 10 g/L yeast extract, 5 g/L NaCl) supplemented with 15 g/L agar and 300 μg/mL spectinomycin. After cell plating, the Q-trays were incubated at 30 °C for 20 h.

The total number of grown colonies was estimated using a QPix 450 (Molecular Devices) automated microbial screening system. Two rounds of transformation were performed in order to obtain approximately 160,000 colonies, corresponding to a 10× coverage of the SP-library, and estimating 10% of clones containing pSG01 without inserts (data not shown). Plates were scraped to collect all colonies and rinsed with 2xPY. The collected cells were then transferred to several 50 mL Falcon tubes, mixed, and concentrated by centrifugation at 3000*g* for 5 min. The pellets were resuspended in 2xPY, the cell suspensions were pooled, thoroughly mixed, and supplemented with glycerol to a final concentration of 10% (v/v). The glycerol stock was aliquoted, snap frozen, and stored at −80 °C. The cell concentration in the glycerol stocks, as determined by the optical density at 600 nm, was approximately 5.8 × 10^9 cells/mL.

Substrate Preparation for NLR-Based Amylase Assay. Dry corn starch (Sigma-Aldrich, S9679) was re-suspended in 90/10 DMSO/water (v/v) to a final concentration of 2% (w/v), boiled for 30 min, and allowed to cool to room temperature. An aliquot of 100 mL of the prepared solution was basified with 1 M NaOH until it reached a pH ≥ 9, then mixed with 1 mL of the reactive dye 5-([4,6-dichlorotriazin-2-yl]amino)fluorescein hydrochloride (DTAF) (Sigma-Aldrich), previously dissolved in DMSO (20 mg/mL). After 1 h incubation at room temperature, the solution was neutralized with glacial acetic acid to stop the reaction, and the fluorescein-starch was precipitated with ethanol to remove the remaining free dye. The precipitated starch was resuspended in DMSO and subsequently ground with glass beads at 30 Hz for 20 min (Retsch). The resulting fluorescein-starch preparation was stored at 4 °C and used as the substrate employed to monitor amylase activity within the NLR-based assay described below.[54,55]

Cultivation of Strains in NLRs. NLRs were synthesized starting from a mix of bacterial glycerol stocks, fluorescein−starch and sodium alginate, which was processed through a laminar jet break-up encapsulator (Nisco Engineering) to generate a monodisperse bead population. To prepare the mix, 200 μL of fluorescein−starch (4% w/v in DMSO) was diluted in 2 mL of resuspension medium (4 g/L yeast extract, 1 g/L tryptone, 20 mM TRIS pH 7) and added to 16 mL of sodium alginate 2.5% (w/v) aqueous solution. The number of bacterial cells to be included was defined to achieve an average occupation of 0.3 cells per NLR. To this end, the corresponding volume of the bacterial glycerol stock was added to the resuspension medium to reach a final volume of 2 mL, which was then mixed with the fluorescein−starch alginate preparation.

For NLR formation, the encapsulator was equipped with a 150 μm nozzle and operated with a flow rate of 3.3 mL/min and a frequency of 650 Hz.[79] This delivered NLRs with an average diameter of 500 μm (corresponding to a volume of approximately 65 nL). NLRs were allowed to harden for 15 min in 100 mM aqueous $CaCl_2$, then isolated using a cell strainer (100 μm mesh size, Falcon, Becton Dickinson), and washed once with 10 mM aqueous $CaCl_2$. NLRs were transferred into growth medium (4 g/L yeast extract, 1 g/L tryptone, 20 mM TRIS pH 7, 4 mM $CaCl_2$, and 300 μg/mL spectinomycin) with 0.5% (v/v) amylopectin to a final concentration of 100 g wet NLRs/L in Erlenmeyer flasks. The reactors were incubated in a shaker (150 rpm, room temperature) for approximately 13 h to allow cells to grow into microcolonies. NLRs were then recovered and washed twice with screening buffer (10 mM $CaCl_2$, 10 mM TRIS pH 8). During each wash, the beads were allowed to sediment in a 50 mL Falcon tube, the supernatant was discarded, and buffer was added to achieve a concentration of 12.5 g of wet NLRs/L. Prior to screening, 40 μL of nile red (Chemodex) (1 g/L in 90/10 DMSO/water, v/v) was added for every gram of wet NLRs to fluorescently stain the cells. The NLRs were incubated for 20 min under gentle shaking, washed once more with the screening buffer to remove surplus dye, and then subjected to flow cytometry and microscopic analysis. Bright-field and fluorescence microscopy images were recorded using an Axio Observer II with an AxioCam MR3 camera (Carl Zeiss Microscopy) to control for proper NLR synthesis and cell growth. For a detailed description of the flow cytometry analysis, see the section below.

If alginate beads with known SP variants needed to be incubated together in the same vessel (to guarantee identical incubation conditions) and differentiated later in the flow cytometry analysis, the NLRs were synthesized with different concentrations of pacific-blue (Ex 410 nm, Em 455 nm) labeled amino dextran (AD). Two concentrations, corresponding to 12 and 2.4 μL of the pacific-blue AD stock solution (20 mg/mL in 0.2 M sodium bicarbonate, pH 8.3) per mL of fluorescein−starch alginate, were added. This polymer is not a substrate for AmyQ (data not shown) and does not interfere with the recording of fluorescein-based fluorescence (Ex 492 nm, Em 516 nm). Instead, the pacific-blue content can be read out in the violet spectrum. Conjugation of the dye to the polymer was achieved by adding 5 mg of the amine-reactive pacific blue succinimidyl ester (Thermo Fisher) to a solution of 20 mg AD (Fina Biosolutions) per mL of 0.2 M sodium bicarbonate (pH 8.3). The reaction was incubated for 6 h at room temperature. Then, TRIS pH 7 was added to a final concentration of 50 mM to stop the reaction, and the solution was aliquoted and frozen.

Throughout the study, different *B. subtilis* strains, all generated in the same fashion and with the same vector, were analyzed using the NLR-based amylase assay. These included (1) *B. subtilis* producing AmyQ with its native SP (positive control, PC), (2) *B. subtilis* carrying the empty vector, without an inserted SP (negative control, NC), (3) *B. subtilis* transformed with the SP-library, fused to AmyQ, or (4) *B. subtilis* producing AmyQ with SP variants with defined modification. The PC (1) and the NC (2) strains and two variants producing AmyQ with known SPs were used to estimate the dynamic range and sensitivity of the NLR-based amylase activity assay (Figure 2a and Supporting Information Figure S1). Fifteen strains producing AmyQ with SP variants with defined modification (4) were encapsulated and used to validate both the NLR-based screening assay and the model.

NLR-Based Screening. The NLR-based screening of the clones carrying the SP-library was performed with a large particle flow cytometer, which allowed to read out the amount of starch, of cells, and, if applicable, of amino dextran in each NLR, based on different fluorescence signals. Specifically, we used a BioSorter (Union Biometrica) to record for each NLR green (excitation laser 488 nm, beam splitter DM 562, emission filter BP 510/23 nm), red (excitation laser 561 nm, TR mirror, emission filter BP 615/24 nm) and violet fluorescence (excitation laser 405 nm, beam splitter DM 495, emission filter BP 445/40 nm).

Each screening round was performed in two sequential steps. During the first step, all events were analyzed in bulk mode, at a maximum of 90 Hz, and NLRs with a positive red fluorescence (peak height, i.e., presence of colonies stained with nile red) were sorted into a 50 mL Falcon tube, containing 5 mL of screening buffer. The isolated population represented approximately 20% of all the NLRs, in agreement with the occupation estimated from the cell concentration in the glycerol stocks. Prior to the second step, the values of green and red fluorescence of each sorted NLR were graphically visualized using the FlowPilot software provided by the BioSorter manufacturer. The graph was then used to divide all events in 10 bins based on the ratio between green fluorescence (area, representative of amylase activity, and secretion levels) and red fluorescence (peak height, representative of total biomass). The bin width was thus adjusted to have 10% of the events sorted in step 1, in each bin. For the second step, sorted NLRs were run through the Biosorter ten consecutive times, every time isolating in bulk mode the events falling in one bin. In particular, the sorting started from the bin with the lowest green to red ratio (i.e., highest secretion/biomass ratio), bin 1, and then moved progressively to bins with higher green/red ratios. The screening analysis was repeated until the number of occupied NLRs (i.e., positive red fluorescence) reached 160,000, to ensure a 10× coverage of the SP-library. Additionally, after cell encapsulation and growth in the NLRs, 53,588 occupied NLRs were sorted in three rounds, without performing any further binning and treated separately. This sample, named hereafter "occupation control", was processed and sequenced with the 10 bins and later used to gather information about the library coverage and the *B. subtilis* population at this step of the workflow.

To recover the NLR-embedded cells, binned samples were incubated for 10 min under gentle shaking with 2xTY medium (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl) supplemented with potassium phosphate buffer (pH 7) to a final concentration of 0.2 M, at which point, full dissolution of the cross-linked calcium alginate had been achieved. Bacterial cells were pelleted by centrifugation (4000*g*, 30 min), the supernatant was discarded, and the pellet was stored at −80 °C.

Genomic DNA Extraction and NGS Library Preparation. Samples from the 10 bins, the occupation control, and the initial glycerol stock were thawed on ice, centrifuged for 1 min at 16,000*g*, and the supernatant was discarded. Afterward, cells were resuspended in 0.85% (w/v) aqueous NaCl, supplemented with 250 μg/mL of RNase A (Macherey-Nagel), and 0.5 mg/mL of lysing enzymes from *Trichoderma harzianum* (Sigma-Aldrich, L4142). After incubating for 10 min at 37 °C, EDTA and SDS were added to final concentrations of 15 mM and 1.2%, respectively. Samples were vortexed thoroughly, ammonium acetate was added to a concentration of 2.5 M, and then, samples were vortexed again. Precipitated proteins were pelleted by centrifugation at 22,000*g* for 15 min at 4 °C. The supernatant was transferred to a fresh reaction cup, supplemented with an equal volume of 2-propanol and gently mixed. DNA was then pelleted by centrifugation at 22,000*g* for 40 min at 4 °C. The supernatant was discarded, and the pellet was washed twice with ice-cold 70% ethanol, dried, and resuspended in 10 mM Tris−HCl pH 7.5.

Each sample was then amplified by PCR, using phusion polymerase (NEB), with primers P3−P15 (Supporting Information Table S5) that anneal immediately up- and downstream of the inserted SP sequence in pSG01, adding barcodes to identify the sample (primers P3−P15 in Supporting Information Table S5, containing Illumina Nextera tagmentation adapters and, in each forward primer, a specific barcode). PCR products were then purified and recovered in milliQ water. Amplicons were analyzed with a bioanalyzer (LabChip GXII, Caliper Life Sciences) using a 5K HT DNA chip, to check size and concentration of the fragment. The 12 PCRs products, corresponding to the 10 bins and the two controls, were pooled and sequenced as a Nextera library (Illumina) by the company BaseClear B.V. (Leiden, NL) on a NovaSeq machine (Illumina) in paired ends, for a total of 26,175,197 2 × 150 bp reads. For both forward and reverse raw reads, the Phred scores had an average of 36 and a median of 37.

Reads Pre-Processing and Mapping. The software FastQC version 0.11.8[80] was used for quality inspection of the sequencing data. First, possible adapters were removed from the 3′ end of the reads (read-trough adapters), since they could confound the merging process when the read length and insert size are comparable. To this end, the software NGmerge[81] version 0.2dev was used in "adapter removal" mode, with 0 mismatches allowed. Sequences were thus merged into longer pseudoreads using PEAR[82] version 0.9.11, with a minimum overlap of 5nt and a *p*-value of 0.001. This yielded 26,105,901 pairs of reads (99.735% of the total reads) to be merged, with the remaining reads unassembled and no read discarded. Pseudoreads were then sorted in the 10 bins and the two controls, based on the respective barcodes, using the "fastx_barcode_splitter.pl" script from FASTX-Toolkit[83] looking only at the 5′ ("*--bol*" option) and allowing only one mismatch. This resulted in 25,980,025 (99.254% of the total reads) demultiplexed pseudoreads. Any remaining adapter (including the barcode) at both 5′ and 3′ of the assembled reads was removed using cutadapt[84] version 2.3, without any read loss. The obtained pseudoreads were then mapped to the reference sequences (i.e., the designed SPs) using BBMap[85] version 37.93, with "*perfectmode*" activated; and behavior for ambiguously mapped reads was set to "best alignment" (Supporting Information Table S3). Occurrences for each bin and both controls were counted for each of the designed sequences, and the resulting frequency table was later used for model construction (Supporting Information Table S1).

Data Preprocessing, Feature Extraction, Model Construction, and Interpretation. To identify the possible influence of investigated features on protein secretion, we decided to train a simple ML model. This procedure, combined with an interpretation of the model, would allow us to obtain a predictive model that could yield important mechanistic insights into the features determining the secretion efficiency of different SPs.

First of all, sequences with low abundance, corresponding to less than 255 reads in the most populated bin, were discarded. This resulted in 4421 informative SPs, which were used to train and test the model. As a different number of NLRs was collected for each bin, the occurrence of reads was normalized across bins so that they contained the same number of NLRs. To score SPs, we assumed that bins were equidistant and each bin had an average value corresponding to its number. A WA, that is, the summation of bin values weighted on the relative frequencies of reads, was calculated for each SP and used as a secretion score. The WA values of selected SPs could thus range from 1 (i.e., the best secreting SPs with all occurrences detected in bin 1) to 10 (i.e. the worst secreting SPs with all occurrences detected in bin 10).

From the 227 calculated features, 22 were discarded because they showed no variation either in the designed SP-library or in the informative SPs data set, which was a subset of the designed library (Supporting Information Table S2). Furthermore, it was decided to minimize the number of features with a correlation coefficient higher than 0.7 to avoid a spread of importance, as attributed by the model, among them. Thus, out of the initial 227 features, 96 were retained, while 110 features, with correlation coefficients higher than 0.7, were selected for clustering. Clustering was carried out through affinity propagation[86] using the scikit-learn[87] python package with standard parameters. Notably, affinity propagation was selected as the clustering algorithm, because of its intrinsic capability of inferring the total number of clusters. This resulted in 14 clusters of correlating features, out of which 22 features were selected and added back to the feature set. Specifically, for 12 of the clusters, the centroid was selected, for 1, cluster the centroid and an additional feature were selected, and for the last cluster with lowly correlating features, all of the 7 features were included. Altogether, this procedure resulted in a total of 116 features, to which 40 Boolean dummy variables were added that describe the AAs in positions −3 and −1, respective to the SPase cleavage site. This resulted in the final selection of 156 features describing the selected SPs (Supporting Information Table S2). In order to verify that the selected features were relevant (i.e., provided variance) within the data sets, and thus needed to train the model, a principal component analysis was performed on (i) the 134 wild-type known SPs, (ii) the set of 11,643 unique SP (i.e. the SP-library), and (iii) the set of 4421 informative SPs (Supporting Information Figure S4).

To construct the model, the matrix of 4421 informative SPs and 156 features was used as the independent variable, while the array of 4421 WA values was used as the dependent variable. These matrices were split with the Kennard-Stone algorithm[88,89] into a training set and a test set of 3095 and 1326 informative SPs, respectively. From available models, a RF[90,91] Regressor model from scikit-learn[87] was implemented. To identify the best hyperparameters for the RF model, a five-time cross-validation grid search was performed on the training set. From the 15,435 tested combinations of hyperparameters, the following set of hyperparameters, balancing predictive power and size of the model, was selected: max depth 25, max features 156, min samples leaf 0.0001 of the training set, min samples split 0.001 of the training set, and estimators 75. The model was subsequently evaluated on the test set and scored calculating the MSE between measured and predicted values (Figure 3a).

The RF model was analyzed to gain mechanistic insights and an explanation of the model itself. For this task, the TreeSHAP method from the SHAP (SHapley Additive exPlanation)[50−52] package was used since, being based on Shapley values, it is advantageous in terms of consistency, allows for a more reliable comparison of feature attribution values, and allows users to understand the model explanation. A further advantage of SHAP is that it includes both local and global explanations, thereby providing explainability for both the whole data set and the individual SPs. Nevertheless, it should be emphasized that SHAP only provides an explanation of the model based on the contribution of individual features to the final output. SHAP does not necessarily uncover the causal relationships between individual SP features and the actual protein secretion efficiency as displayed by a bacterial cell. Furthermore, it is noteworthy that SHAP provides the possibility to determine the type of relationship between each individual feature and the predicted output and to determine second-order interactions that occur between features.

Assay Validation. To show both the reliability of the NLR-based amylase activity assay and to assess the correctness of the model, we used two orthogonal procedures to measure the amylase activity from strains producing AmyQ with selected SPs: a commercial assay in 96-wells MTPs and a hydrolysis test on starch agar plates.[56] For the MTP assay, bacteria were precultured overnight into 2xPY supplemented with 300 μg/mL spectinomycin, subsequently diluted 100-fold in the same medium and grown for 7.5 h at 37 °C and 250 rpm. Throughout the validation of the NLR-based amylase assay and of the model, two cultivation vessels were applied: deep-well plates filled with 300 μL of medium and/or Erlenmeyer shake flasks (25 mL culture volume in 250 mL flask). Aliquots of the different cultures were collected, the cells were pelleted, and 9 μL of the supernatants were mixed with 9 μL of Ceralpha reagent (Megazymes). The reaction mixtures were incubated for 20 min (standard version) or 90 min (sensitive version) at room temperature on a shaker (1000 rpm), and then the reactions were stopped through the addition of 200 μL of 1% (w/v) 2-amino-2-(hydroxymethyl)-propane-1,3-diol (Tris-base, pH 9). The amylase activity was then measured by monitoring the absorbance at 405 nm with a Tecan Infinite M200 Pro. Similar to the NLR-based assay, the optical density at 600 nm of the cultures was measured and used to normalize all samples for the biomass. Eventually, the OD-normalized absorption values were expressed relative to the amylase activity obtained from cultures that secreted AmyQ with its native SP (PC), defined as 100%, and to the amylase activity in the growth medium of a strain containing pSG01 without an inserted SP sequence (NC), defined as 0%.

For the hydrolysis test on starch agar plates, glycerol stocks of the selected variants were diluted 100-fold in 300 μL 2xPY supplemented with 300 μg/mL spectinomycin and grown until they reached the mid-exponential phase (6 h, 37 °C, 250 rpm). An aliquot of 2 μL of the cell culture was spotted on 2xPY-agar plates supplemented with 300 μg/mL spectinomycin and 0.2% (w/v) potato starch (Sigma-Aldrich). After overnight incubation at 37 °C, the plates were flooded with Lugol's iodine (Carl Roth), which interacts with starch and generates a dark color. Where starch is degraded, a clear zone arises, for example, around a colony, and the area of this clear degradation zone is approximately proportional to the amount of amylase secreted.[56] The standard MTP assay was used for initial validation of the NLR-based assay (Figure 2a, 2) and the screening strategy. To further validate the assay, 95 SP-AmyQ fusions, randomly picked from the 4,421 variants used to train and test the model, were subjected to the MTP assay (see activity values in Supporting Information Table S6 and Figure S2). As 72 randomly picked variants showed no amylase activity, presumably due to too low secretion levels, the sensitive version of the MTP assay was applied, and AmyQ activities could be determined for 15 more clones.

The starch hydrolysis on plate was also applied for the partial validation of the NLR-based assay (Figure 2a, 3), and if no amylase activity could be detected with the sensitive MTP assay (i.e. $\text{Abs}_{405} < 0.1$; Supporting Information Figure S3).

Model Validation. To further validate our model, small sets of SPs were manually edited to tune their predicted secretion levels. 30 SPs directing high-level secretion of AmyQ (i.e., "good performers") were manually edited until the model predicted them to direct AmyQ secretion with poor efficiency (group 1). Similarly, another 30 SPs directing low-level secretion of AmyQ were edited in order to improve their efficiency (group 2) (see Supporting Information Table S4 for full list of SPs).

As an additional validation approach, we generated pseudo-random SP AA sequences, with a home-made script, described as follows. Based on 134 sequences of known and highly probable SPs, 7 dictionaries were calculated that map each AA to its relative frequency: one for the N-region (excluding the initial Met), one for the H-region, one for the C-region except the last 3 residues, one for each −3, −2 and −1 position relative to the SPase cleavage site, and the last one for the Ac-region (i.e., positions +1, +2 and +3 together). Using these values, 10,000 sequences for each region were generated as follows: for the N-region a Met was always placed in front of a stretch of 1 to 10 residues built on the frequency table; the H-region simply consisted of a stretch of 9 to 16 residues built on the frequency table; the C-region was built juxtaposing a stretch of 4 to 11 residues to the 3 single residues for positions −3,-2, and −1, each based on its frequency table; and for the N-terminus of the mature protein after SPase cleavage, a stretch of 3 residues based on its frequencies was built. To minimize the occurrence of SPs with features too far from the distribution of the known, or probably representing wild-type *B. subtilis* SPs, the Kolmogorov-Smirnov[92,93] statistic test was applied, which compared the distribution of various features (data not shown). In case the number of similar distributions (considered as all those features with a calculated *p*-value above 0.1) was below a certain threshold, i.e., at least 21, 16, 18, and 17 features, respectively, for the N-, H-, C-, and Ac-regions, the batch of 10,000 sequences was discarded and the process was repeated. In such a way, the four designated regions (i.e. the N-, H-, C-, and Ac-regions) were built independently from each other and their possible combinations or interactions were not considered. When 10,000 sequences for each region were generated, they were juxtaposed to form a full SP. Then, sequences equal or longer than 33 AAs (calculating the length up to and including the Ac-region) were discarded, thus resulting in 4903 valid SPs. To generate the relative coding sequences, the AA sequences were retro-translated with an unambiguous dictionary where only the most frequent codon for each AA was present. Subsequently, having both a nucleotide and an AA sequence for each SP, the 156 features of the final model were calculated, and the respective SPs' secretion efficiency (i.e., the WA value) was predicted using the RF model. A set of 32 SPs (group 3) was selected with the requirement to be predicted by our model as very good secretors (see Supporting Information Table S4 for full list of SPs). Eventually, the different features for each SP were also analyzed and explained through SHAP.

For manually edited as well as pseudo-randomly designed SPs, the respective SP-encoding sequences were ordered (Twist Biosciences), cloned in pSG01, and used to transform *B. subtilis* DB104, following the same procedure as applied for the generation of the SP-library. For the 61 successfully constructed SP-AmyQ fusions, the amylase activity was monitored in the MTP assay as described above. Additionally, 15 of these fusions (5 for each of the three groups) were further tested via the NLR-based amylase activity assay, to verify the model in the same experimental setup used to generate it (Figure 4).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.2c00328.

Captions corresponding to Supporting Information files (PDF)

Figure S1. Optimization of NLR culture conditions (PDF)

Figure S2. Comparison of the NLR- and MTP-based amylase assays (PDF)

Figure S3. Hydrolysis tests on starch agar plates (PDF)

Figure S4. Principal component analysis (PCA) over data sets with wild-type, designed, and reliably measured SPs (PDF)

Figure S5. Explanation of the applied SHAP methodology (PDF)

Figure S6. Second-order interactions between SP features (PDF)

Figure S7. SHAP interaction plot (PDF)

Figure S8. SHAP summary plot (PDF)

Figure S9. Map of plasmid pSG01 (PDF)

Table S1. SP sequences and feature descriptions (XLSX)

Table S2. Overview of feature editing, processing, and clustering (XLS)

Table S3. Summarized report of mapped reads (XLS)

Table S4. Data on SP validation (XLSX)

Table S5. Oligonucleotide sequences and plasmids (XLS)

Table S6. Summary of the MTP amylase assay of 95 clones (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Jan Maarten van Dijl − *Department of Medical Microbiology, University of Groningen, University Medical Center Groningen, Groningen 9700 RB, The Netherlands;* orcid.org/0000-0002-5688-8438;
Phone: +31503615187; Email: j.m.van.dijl01@umcg.nl

Tjeerd van Rij − *DSM Biotechnology Center, Delft 2613 AX, Netherlands;* orcid.org/0000-0002-0301-2598;
Phone: +31628441843; Email: tjeerd.rij-van@dsm.com

Authors

Stefano Grasso − *Department of Medical Microbiology, University of Groningen, University Medical Center Groningen, Groningen 9700 RB, The Netherlands; DSM Biotechnology Center, Delft 2613 AX, Netherlands;* Present Address: Lesaffre International, 101 Rue de Menin, 59700 Marcq-en-Barœul, France

Valentina Dabene − *Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland; FGen AG, Basel 4057, Switzerland*

Margriet M. W. B. Hendriks − *DSM Biotechnology Center, Delft 2613 AX, Netherlands*

Priscilla Zwartjens − *DSM Biotechnology Center, Delft 2613 AX, Netherlands*

René Pellaux − *FGen AG, Basel 4057, Switzerland*

Martin Held − *Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland*

Sven Panke − *Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland*

Andreas Meyer − *FGen AG, Basel 4057, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.2c00328

Author Contributions

S.G. and V.D., J.M.D., A.M., and T.R. contributed equally. J.M.D., M.H., S.P., A.M., and T.R. conceived the project. S.G.

and P.Z. designed the libraries. S.G. designed and performed cloning and transformation. V.D. and R.P. designed and optimized the NLR-based enzymatic assay. S.G. and V.D. performed the screening assay. S.G. prepared the sequencing library and analyzed the NGS data. S.G. and M.H. generated and explained the ML model. V.D. validated the assay and the model. S.G. and V.D. drafted the manuscript. S.P., J.M.D., A.M., and T.R. supervised the project and revised the manuscript.

Funding

This work was supported by the European Union's Horizon 2020 Program, Marie Skłodowska-Curie Actions (MSCA), under REA grant agreement no. 642836.

Notes

The authors declare no competing financial interest.

Data availability: HTML files with Supporting Information train and test data, the ML model, and the SHAP explainer have been deposited on Zenodo with the following doi:10.5281/zenodo.6448806 and are available at https://doi.org/10.5281/zenodo.6448806. In particular, these include an interactive Jupyter notebook to fully explore the model (File S1, interactive SHAP force plot for the whole train and test set), an interactive file to explore the original SPs (File S2, interactive SHAP force plot for the 59 SPs used to validate the model in their original form), and an interactive file to explore the edited SPs (File S3, interactive SHAP force plot for the 59 SPs used to validate the model in their engineered form). Raw Illumina reads have been deposited on NCBI SRA under BioProject accession number: PRJNA825264. Flow cytometry data have been deposited on FlowRepository under the experiment ID FR-FCM-Z57A.

■ ACKNOWLEDGMENTS

We would like to thank Nivitha Punniyamoorthy and Steven Schmitt for the technical support provided during the long hours spent at the Biosorter. We would also like to thank Irsan Kooi, Marcel Hillebrand, Rianne van der Hoek, and Ana Bulović for the fruitful discussions.

■ REFERENCES

(1) Desvaux, M.; Hébraud, M.; Talon, R.; Henderson, I. R. Outer Membrane Translocation: Numerical Protein Secretion Nomenclature in Question in Mycobacteria. *Trends Microbiol.* **2009**, *17*, 338−340.

(2) Anné, J.; Economou, A.; Bernaerts, K. Protein Secretion in Gram-Positive Bacteria: From Multiple Pathways to Biotechnology. *Protein and Sugar Export and Assembly in Gram-positive Bacteria*; Bagnoli, F., Rappuoli, R., Eds.; Springer: Cham (DE), 2016; Vol. *404*, pp 267−308.

(3) Park, E.; Rapoport, T. A. Mechanisms of Sec61/SecY-Mediated Protein Translocation Across Membranes. *Annu. Rev. Biophys.* **2012**, *41*, 21−40.

(4) Tsirigotaki, A.; De Geyter, J.; Šoštarić, N.; Economou, A.; Karamanou, S. Protein Export through the Bacterial Sec Pathway. *Nat. Rev. Microbiol.* **2017**, *15*, 21−36.

(5) Chen, Y.; Shanmugam, S. K.; Dalbey, R. E. The Principles of Protein Targeting and Transport Across Cell Membranes. *Protein J.* **2019**, *38*, 236−248.

(6) Demain, A. L.; Vaishnav, P. Production of Recombinant Proteins by Microbes and Higher Organisms. *Biotechnol. Adv.* **2009**, *27*, 297−306.

(7) Rettenbacher, L. A.; Arauzo-Aguilera, K.; Buscajoni, L.; Castillo-Corujo, A.; Ferrero-Bordera, B.; Kostopoulou, A.; Moran-Torres, R.; Núñez-Nepomuceno, D.; Öktem, A.; Palma, A.; Pisent, B.; Puricelli, M.; Schilling, T.; Tungekar, A. A.; Walgraeve, J.; Humphreys, D.; von der Haar, T.; Gasser, B.; Mattanovich, D.; Ruddock, L.; van Dijl, J. M. Microbial Protein Cell Factories Fight Back? *Trends Biotechnol.* **2022**, *40*, 576−590.

(8) Desvaux, M.; Hébraud, M.; Talon, R.; Henderson, I. R. Secretion and Subcellular Localizations of Bacterial Proteins: A Semantic Awareness Issue. *Trends Microbiol.* **2009**, *17*, 139−145.

(9) van Dijl, J. M.; Hecker, M. Bacillus Subtilis: From Soil Bacterium to Super-Secreting Cell Factory. *Microb. Cell Fact.* **2013**, *12*, 3.

(10) Harwood, C. R.; Cranenburgh, R. Bacillus Protein Secretion: An Unfolding Story. *Trends Microbiol.* **2008**, *16*, 73−79.

(11) Kaushik, S.; He, H.; Dalbey, R. E. Bacterial Signal Peptides-Navigating the Journey of Proteins. *Front. Physiol.* **2022**, *13*, 1398.

(12) Tjalsma, H.; Bolhuis, A.; Jongbloed, J. D.; Bron, S.; van Dijl, J. M. Signal Peptide-Dependent Protein Transport in Bacillus Subtilis: A Genome-Based Survey of the Secretome. *Microbiol. Mol. Biol. Rev.* **2000**, *64*, 515−547.

(13) Tjalsma, H.; van Dijl, J. M. Proteomics-Based Consensus Prediction of Protein Retention in a Bacterial Membrane. *Proteomics* **2005**, *5*, 4472−4482.

(14) Peng, C.; Shi, C.; Cao, X.; Li, Y.; Liu, F.; Lu, F. Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Front. Bioeng. Biotechnol.* **2019**, *7*, 139.

(15) von Heijne, G.; Abrahmsèn, L.; Abrahmsén, L. Species-Specific Variation in Signal Peptide Design. Implications for Protein Secretion in Foreign Hosts. *FEBS Lett.* **1989**, *244*, 439−446.

(16) Perlman, D.; Halvorson, H. O. A Putative Signal Peptidase Recognition Site and Sequence in Eukaryotic and Prokaryotic Signal Peptides. *J. Mol. Biol.* **1983**, *167*, 391−409.

(17) Gelis, I.; Bonvin, A. M. J. J.; Keramisanou, D.; Koukaki, M.; Gouridis, G.; Karamanou, S.; Economou, A.; Kalodimos, C. G. Structural Basis for Signal-Sequence Recognition by the Translocase Motor SecA as Determined by NMR. *Cell* **2007**, *131*, 756−769.

(18) Musial-Siwek, M.; Rusch, S. L.; Kendall, D. A. Selective Photoaffinity Labeling Identifies the Signal Peptide Binding Domain on SecA. *J. Mol. Biol.* **2007**, *365*, 637.

(19) Pugsley, A. P. *Protein Targeting*; Academic Press: San Diego, 1989.

(20) Inouye, S.; Soberon, X.; Franceschini, T.; Nakamura, K.; Itakura, K.; Inouye, M. Role of Positive Charge on the Amino-Terminal Region of the Signal Peptide in Protein Secretion across the Membrane. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 3438−3441.

(21) Iino, T.; Takahashi, M.; Sako, T. Role of Amino-Terminal Positive Charge on Signal Peptide in Staphylokinase Export across the Cytoplasmic Membrane of Escherichia Coli. *J. Biol. Chem.* **1987**, *262*, 7412−7417.

(22) Vlasuk, G. P.; Inouye, S.; Ito, H.; Itakura, K.; Inouye, M. Effects of the Complete Removal of Basic Amino Acid Residues from the Signal Peptide on Secretion of Lipoprotein in Escherichia Coli. *J. Biol. Chem.* **1983**, *258*, 7141−7148.

(23) Fikes, J. D.; Bankaitis, V. A.; Ryan, J. P.; Bassford, P. J. Mutational Alterations Affecting the Export Competence of a Truncated but Fully Functional Maltose-Binding Protein Signal Peptide. *J. Bacteriol.* **1987**, *169*, 2345−2351.

(24) Bedouelle, H.; Bassford, P. J.; Fowler, A. V.; Zabin, I.; Beckwith, J.; Hofnung, M. Mutations Which Alter the Function of the Signal Sequence of the Maltose Binding Protein of Escherichia Coli. *Nature* **1980**, *285*, 78−81.

(25) Emr, S. D.; Schwartz, M.; Silhavy, T. J. Mutations Altering the Cellular Localization of the Phage λ Receptor, an Escherichia Coli Outer Membrane Protein. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5802−5806.

(26) Michaelis, S.; Beckwith, J. Mechanism of Incorporation of Cell Envelope Proteins in Escherichia Coli. *Annu. Rev. Microbiol.* **1982**, *36*, 435−465.

(27) Bogsch, E.; Brink, S.; Robinson, C. Pathway Specificity for a ΔpH-Dependent Precursor Thylakoid Lumen Protein Is Governed by

a "Sec-Avoidance" Motif in the Transfer Peptide and a "Sec-Incompatible" Mature Protein. *EMBO J.* **1997**, *16*, 3851−3859.

(28) Tullman-Ercek, D.; DeLisa, M. P.; Kawarasaki, Y.; Iranpour, P.; Ribnicky, B.; Palmer, T.; Georgiou, G. Export Pathway Selectivity of Escherichia Coli Twin Arginine Translocation Signal Peptides. *J. Biol. Chem.* **2007**, *282*, 8309−8316.

(29) Kuhn, A.; Wickner, W. Conserved Residues of the Leader Peptide Are Essential for Cleavage by Leader Peptidase. *J. Biol. Chem.* **1985**, *260*, 15914−15918.

(30) Dalbey, R. E.; Wang, P.; van Dijl, J. M. Membrane Proteases in the Bacterial Protein Secretion and Quality Control Pathway. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 311−330.

(31) Nielsen, H.; Tsirigos, K. D.; Brunak, S.; von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* **2019**, *38*, 200−216.

(32) 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*, 422−441.

(33) 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*, 393−402.

(34) 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*, 6370−6376.

(35) Mathiesen, G.; Sveen, A.; Brurberg, M.; Fredriksen, L.; Axelsson, L.; Eijsink, V. G. Genome-Wide Analysis of Signal Peptide Functionality in Lactobacillus Plantarum WCFS1. *BMC Genomics* **2009**, *10*, 425.

(36) Aronica, P. G. A.; Reid, L. M.; Desai, N.; Li, J.; Fox, S. J.; Yadahalli, S.; Essex, J. W.; Verma, C. S. Computational Methods and Tools in Antimicrobial Peptide Research. *J. Chem. Inf. Model.* **2021**, *61*, 3172−3196.

(37) Nielsen, M.; Andreatta, M.; Peters, B.; Buus, S. Immunoinformatics: Predicting Peptide-MHC Binding. *Annu. Rev. Biomed. Data Sci.* **2020**, *3*, 191−215.

(38) Aranha, M. P.; Spooner, C.; Demerdash, O.; Czejdo, B.; Smith, J. C.; Mitchell, J. C. Prediction of Peptide Binding to MHC Using Machine Learning with Sequence and Structure-Based Feature Sets. *Biochim. Biophys. Acta, Gen. Subj.* **2020**, *1864*, 129535.

(39) Arif, M.; Ahmed, S.; Ge, F.; Kabir, M.; Khan, Y. D.; Yu, D.-J.; Thafar, M. StackACPred: Prediction of Anticancer Peptides by Integrating Optimized Multiple Feature Descriptors with Stacked Ensemble Approach. *Chemom. Intell. Lab. Syst.* **2022**, *220*, 104458.

(40) Hasan, M. M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: Improved and Robust Prediction of Hemolytic Peptide and Its Activity by Fusing Multiple Feature Representation. *Bioinformatics* **2020**, *36*, 3350−3356.

(41) Lawrence, T. J.; Carper, D. L.; Spangler, M. K.; Carrell, A. A.; Rush, T. A.; Minter, S. J.; Weston, D. J.; Labbé, J. L. AmPEPpy 1.0: A Portable and Accurate Antimicrobial Peptide Prediction Tool. *Bioinformatics* **2021**, *37*, 2058−2060.

(42) Alzahrani, E.; Alghamdi, W.; Ullah, M. Z.; Khan, Y. D. Identification of Stress Response Proteins through Fusion of Machine Learning Models and Statistical Paradigms. *Sci. Rep.* **2021**, *11*, 1−15.

(43) Shah, A. A.; Khan, Y. D. Identification of 4-Carboxyglutamate Residue Sites Based on Position Based Statistical Feature and Multiple Classification. *Sci. Rep.* **2020**, *10*, 16913.

(44) Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N. iPhosS(Deep)-PseAAC: Identify Phosphoserine Sites in Proteins using Deep Learning on General Pseudo Amino Acid Compositions via Modified 5-Steps Rule. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2020**, *19*, 1703−1714.

(45) Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N. Optimization of Serine Phosphorylation Prediction in Proteins by Comparing Human Engineered Features and Deep Representations. *Anal. Biochem.* **2021**, *615*, 114069.

(46) Hussain, W. SAMP-PFPDeep: Improving Accuracy of Short Antimicrobial Peptides Prediction Using Three Different Sequence Encodings and Deep Neural Networks. *Briefings Bioinf.* **2022**, *23*, bbab487.

(47) 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*, 420−423.

(48) Teufel, F.; Almagro Armenteros, J. J.; Johansen, A. R.; Gíslason, M. H.; Pihl, S. I.; Tsirigos, K. D.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models. *Nat. Biotechnol.* **2022**, *40*, 1023−1025.

(49) Zeng, W.; Guo, L.; Xu, S.; Chen, J.; Zhou, J. High-Throughput Screening Technology in Industrial Biotechnology. *Trends Biotechnol.* **2020**, *38*, 888−906.

(50) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. **2017**, arXiv:1705.07874.

(51) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56−67.

(52) Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K. W.; Newman, S. F.; Kim, J.; Lee, S. I. Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749−760.

(53) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. **2018**, ArXiv 2018, cs.LG (1802.03888).

(54) Meyer, A.; Pellaux, R.; Potot, S.; Becker, K.; Hohmann, H.-P.; Panke, S.; Held, M. Optimization of a Whole-Cell Biocatalyst by Employing Genetically Encoded Product Sensors inside Nanolitre Reactors. *Nat. Chem.* **2015**, *7*, 673−678.

(55) Walser, M.; Leibundgut, R. M.; Pellaux, R.; Panke, S.; Held, M. Isolation of monoclonal microcarriers colonized by fluorescent E. coli. *Cytometry, Part A* **2008**, *73A*, 788−798.

(56) Briggs, D. E. Gel-Diffiusion Method for the Assay of α-Amylase. *J. Inst. Brew.* **1962**, *68*, 27−32.

(57) Strahl, H.; Bürmann, F.; Hamoen, L. W. The Actin Homologue MreB Organizes the Bacterial Cell Membrane. *Nat. Commun.* **2014**, *5*, 1−11.

(58) Yu, M. K.; Ma, J.; Fisher, J.; Kreisberg, J. F.; Raphael, B. J.; Ideker, T. Visible Machine Learning for Biomedicine. *Cell* **2018**, *173*, 1562−1565.

(59) Hall, P.; Gill, N.; Schmidt, N. Proposed Guidelines for the Responsible Use of Explainable Machine Learning. **2019**, ArXiv 2019, stat.ML (1906.03533).

(60) Tjalsma, H.; Antelmann, H.; Jongbloed, J. D. H.; Braun, P. G.; Darmon, E.; Dorenbos, R.; Dubois, F.; Westers, H.; Zanen, G.; Quax, W. J.; Kuipers, O. P.; Bron, S.; Hecker, M.; van Dijl, J. F.; van Dijl, J. M. Proteomics of Protein Secretion by Bacillus Subtilis : Separating the "Secrets" of the Secretome. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 207−233.

(61) Choo, K. H.; Ranganathan, S. Flanking Signal and Mature Peptide Residues Influence Signal Peptide Cleavage. *BMC Bioinf.* **2008**, *9*, S15.

(62) Zanen, G.; Houben, E. N. G.; Meima, R.; Tjalsma, H.; Jongbloed, J. D. H.; Westers, H.; Oudega, B.; Luirink, J.; van Dijl, J. M.; Quax, W. J. Signal Peptide Hydrophobicity Is Critical for Early Stages in Protein Export by Bacillus Subtilis. *FEBS J.* **2005**, *272*, 4617−4630.

(63) Dyrløv Bendtsen, J.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340*, 783−795.

(64) Reymond, J.-L.; Fluxà, V. S.; Maillard, N. Enzyme Assays. *Chem. Commun.* **2008**, *1*, 34−46.

(65) Markel, U.; Essani, K. D.; Besirlioglu, V.; Schiffels, J.; Streit, W. R.; Schwaneberg, U. Advances in Ultrahigh-Throughput Screening for Directed Enzyme Evolution. *Chem. Soc. Rev.* **2020**, *49*, 233−262.

(66) Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D. Speeding up Enzyme Discovery and Engineering with Ultrahigh-Throughput Methods. *Curr. Opin. Struct. Biol.* **2018**, *48*, 149−156.

(67) Naseri, G.; Koffas, M. A. G. Application of Combinatorial Optimization Strategies in Synthetic Biology. *Nat. Commun.* **2020**, *11*, 2446.

(68) Gunst, R. F.; Mason, R. L. Fractional Factorial Design. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1*, 234−244.

(69) Wu, Z.; Yang, K. K.; Liszka, M. J.; Lee, A.; Batzilla, A.; Wernick, D.; Weiner, D. P.; Arnold, F. H. Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* **2020**, *9*, 2154−2161.

(70) Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions. *Nat. Methods* **2011**, *8*, 785−786.

(71) Käll, L.; Krogh, A.; Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **2004**, *338*, 1027−1036.

(72) Sauer, C.; Ver Loren van Themaat, E.; Boender, L. G. M.; Groothuis, D.; Cruz, R.; Hamoen, L. W.; Harwood, C. R.; van Rij, T. Exploring the Nonconserved Sequence Space of Synthetic Expression Modules in Bacillus Subtilis. *ACS Synth. Biol.* **2018**, *7*, 1773−1784.

(73) Macaluso, A.; Mettus, A. M. Efficient Transformation of Bacillus Thuringiensis Requires Nonmethylated Plasmid DNA. *J. Bacteriol.* **1991**, *173*, 1353−1356.

(74) Kawamura, F.; Doi, R. H. Construction of a Bacillus Subtilis Double Mutant Deficient in Extracellular Alkaline and Neutral Proteases. *J. Bacteriol.* **1984**, *160*, 442−444.

(75) Lam, K. H. E.; Chow, K. C.; Wong, W. K. R. Construction of an Efficient Bacillus Subtilis System for Extracellular Production of Heterologous Proteins. *J. Biotechnol.* **1998**, *63*, 167−177.

(76) Hambraeus, G.; von Wachenfeldt, C.; Hederstedt, L. Genome-Wide Survey of MRNA Half-Lives in Bacillus Subtilis Identifies Extremely Stable MRNAs. *Mol. Genet. Genomics* **2003**, *269*, 706−714.

(77) McKenzie, T.; Hoshino, T.; Tanaka, T.; Sueoka, N. The Nucleotide Sequence of PUB110: Some Salient Features in Relation to Replication and Its Regulation. *Plasmid* **1986**, *15*, 93−103.

(78) Carl, U. D.; Batz, L.; Schuchardt, I.; Germeroth, L.; Schmidt, T. G. M. StarGate: A High-Capacity Expression Cloning System to Speed-Up Biopharmaceutical Development. In *Modern Biopharmaceuticals: Recent Success Stories*; Knäblein, J., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim (DE), 2013; pp 147−164.

(79) Femmer, C.; Bechtold, M.; Held, M.; Panke, S. In vivo directed enzyme evolution in nanoliter reactors with antimetabolite selection. *Metab. Eng.* **2020**, *59*, 15−23.

(80) Andrews, S. FASTQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

(81) Gaspar, J. M. NGmerge: Merging Paired-End Reads via Novel Empirically-Derived Models of Sequencing Errors. *BMC Bioinf.* **2018**, *19*, 536.

(82) Zhang, J.; Kobert, K.; Flouri, T.; Stamatakis, A. PEAR: A Fast and Accurate Illumina Paired-End ReAd MergeR. *Bioinformatics* **2014**, *30*, 614−620.

(83) Gordon, A.; Hannon, G. J. Gordon FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit.

(84) Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet J.* **2011**, *17*, 10.

(85) Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. https://sourceforge.net/projects/bbmap/.

(86) Frey, B. J.; Dueck, D. Clustering by Passing Messages between Data Points. *Science* **2007**, *315*, 972−976.

(87) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825.

(88) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.

(89) Hiromasa, K. Python implementation of Kennard-Stone algorithm. https://github.com/hkaneko1985/kennardstonealgorithm (accessed Jan 22, 2019).

(90) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(91) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3−42.

(92) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10−20.

(93) Millman, K. J.; Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **2011**, *13*, 9−12.