Naval Research Laboratory
Washington, DC 20375-5320

NRL/6910/MR—2023/1

Controlled Generation of Protein Sequences from Tree–Preserving Embeddings and Generative Deep Learning Models

JEROME ANTHONY E. ALVAREZ

STEM Student Employment Program (SSEP)
Center for Bio/Molecular Science and Engineering Division

SCOTT N. DEAN

Laboratory for Bio/Nano Science and Technology Branch
Center for Bio/Molecular Science and Engineering Division

May 3, 2023

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
03-05-2023	NRL Memorandum Report	04/01/2022 – 04/27/2023

4. TITLE AND SUBTITLE

Controlled Generation of Protein Sequences from Tree–Preserving Embeddings and Generative Deep Learning Models

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jerome Anthony E. Alvarez* and Scott N. Dean

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1S67

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/6910/MR--2023/1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Threat Reduction Agency
8725 John J Kingman Rd Ste 6201
Fort Belvoir, VA 22060

10. SPONSOR / MONITOR'S ACRONYM(S)

DTRA

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Rapid development of novel biomolecular sequences, and in particular proteins, is essential for a range of fields from drug development to food safety, and recent advances of generative deep learning techniques for sequence generation has demonstrated their potential large value for these applications. However, progression from simple exploratory data analysis to controlled generation of novel sequences that resemble those in nature remains cumbersome. Dimensionality reduction relevant to exploratory data analysis, which embeds high-dimensional data in a low-dimensional space while preserving some level of structure in the data, can enable useful clustering. While clustering methods including principal component analysis, multidimensional scaling, and stochastic neighbor embedding are popular in protein sequence analysis, their direct application to sequence datasets generally fail to separate clusters, and the clusters that form do not appear in-line with known attributes of the proteins being analyzed. In contrast, tree-preserving embeddings demonstrate remarkable performance for logical clustering that follow from dendrograms showing sequence-wise distance. Here we developed a new approach utilizing tree-preserving embeddings for controlled sampling in sequence space developed by several different generative models. Without restoring to more complex models, the generated sequences conform to proteins with known attributes along with high sequence similarity. This combined method requires no additional data sources and can deal with a wide range of different protein types, simplifying its use for generation of new proteins and other biomolecular sequences.

15. SUBJECT TERMS

Tree preserving embedding
Generative deep learning
Symmetry
Antibodies
Fluorescent proteins;

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Jerome A. Alvarez
U	U	U	U	21	**19b. TELEPHONE NUMBER** *(include area code)* (202) 767-0394

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

1. Introduction

The high level of interest in protein sequence generation has its basis in their potential wide range of applications, including peptide antibiotics, antibody-based therapeutics, and vaccines. Generation of new sequences is generally required for production of variants that may evade previously developed antibiotic resistance, in the case of antimicrobials, or for optimized activities for antibody therapeutics among other attributes.

Many groups have implemented different techniques for sequence generation, where much of the recent focus has been focused on repurposing and modifying models initially developed for natural language processing (NLP), in particular NLP deep learning models, for application to biological problems. For example, the study by Bowman *et al.* for generation of sentences in a continuous space led to a variety of variational autoencoder (VAE)-based methods for peptide, protein, and DNA sequence generation [1-4]. Similar VAE models have been derived by the conditional text-generating VAE report by Hu *et al.* [5], where Das *et al.* made use of a modified version for generation of antimicrobial peptides [6]. Similarly for generative adversarial networks (GANs), where Tucs *et al.* made use of GANs protein generation [7]. Less complex text generation models such as recurrent neural networks (RNNs) have also been applied widely to protein and peptide sequences generation [8-10].

In the case of VAEs and GANs, one of the major attributes useful for controlled generation is the developed latent space, which allows for generation of objects in continuous space with a degree of control over sampling, and therefore producing sequences with desired characteristics. However, in many cases, this space is insufficiently informed by sequence similarity, where there is risk of sampling in a region of sequence space neighboring proteins that are very different in character from the target characteristics.

While the previous described deep learning techniques have been demonstrated to work, both the computational expense and data requirements of many of these methods are high, while many biological datasets remain incredibly small. The small dataset size necessitates the use of simpler models. Thus, much of sequence generation models' value is closely tied to the attributes of the corresponding protein, and the ability of the model to lend the user a degree of control over its output. The main goal is to generate new protein with specified properties where, for example, a model for producing new fluorescent protein (FP) sequences should reliably produce novel fluorescent proteins of a specific emission wavelength (e.g., blue green) and not a random color. In addition, a variety of other attributes of the generated protein need to be closely monitored even if they are considered as secondary to the primary reason for generating a novel protein, such as solubility, size, and stability. For these reasons, production of a new protein with high sequence similarity to those in nature that have been characterized is highly valuable for experimentalists producing and working with the new proteins, and thus should be included in the generative model pipeline.

In this work we describe pairing simpler methods generative deep learning with tree preserving embedding (TPE), which we demonstrate provides a valuable system for sequence generation, analysis, and visualization. Here we show the value of TPE for generation of proteins, which may be applied widely, based on our proof-of-concept results on fluorescent proteins and single-domain antibodies. This study demonstrates that TPE application on protein sequence datasets is modular and produces viable results complemented by a generative recurrent neural network (RNN) or Sequence-to-Sequence–Long Short-Term Memory (Seq2Seq–LSTM) deep learning models. Sampling outputs show that the generated sequences are unique to naturally occurring proteins, while both resembling the target dataset in physicochemical attributes and retaining strong sequence similarity. This combined method requires no additional data sources and can be applied to a wide range of different protein types. Because it has no inherent properties specific to proteins, it can simplify its use for generation of both new proteins and other biomolecular sequences.

Manuscript approved May 3, 2023.

2. Materials and Methods

2.1 Datasets

Datasets used in this study were obtained from the Fluorescent Proteins Database (FPbase; www.fpbase.org) and Single Domain Antibody Database (sdAb-DB; www.sdab-db.ca). After filtering for incomplete data and further preprocessing, the FPbase dataset consisted of 593 sequences paired with names and their corresponding excitation and emission wavelengths. Single-domain antibodies (sdAb-DB) consisted of 1000 sequences paired with names, target, and source information. All protein sequences were restricted to the 20 natural amino acids and sequences with non-conventional residues were removed. FPbase sequences were removed if no information for excitation and emission wavelengths were available.

2.2 Dimensionality Reduction and Sequence Generation

Prerequisite to TPE application, both FPbase and sdAb-DB sequences were converted to *AAStringSet* objects from the R Bioconductor package. Two-dimensional TPE was implemented and used according to Shieh *et al.* [11]. For input, both datasets were first run through a multiple sequence alignment (MSA) and then converted into a distance matrix using the *stringDist* function from the BioStrings R package. TPE is iteratively run over each sequence and outputs a series of two-dimensional arrays for each of the sequences in the dataset. Principal Components Analysis (PCA) and t-Stochastic Neighbor Embedding (t-SNE) were performed using *prcomp* and *Rtsne* functions, respectively, both with two dimensions as parameters. The various generative models used in this study are as follows:

(a) Sequence-to-Sequence–Long Short-Term Memory (Seq2Seq–LSTM): modified Seq2Seq–LSTM was implemented in-house using R with a Keras backend. The main framework of sequence-to-sequence models are derived from simple question answering queries that may require mapping "*a sequence of words representing the question*" to "*a sequence of words representing the answer*". Furthermore, LSTM architecture solves general sequence to sequence problems [12]. For example, Google's Neural Machine Translation System for translating various languages consists of an LSTM network with 8 encoder and 8 decoder layers using attention and residual connections [13]. The modified Seq2Seq–LSTM model uses 1 layer of LSTM as encoder and with the decoder network sampling on various batch size ranges.

(b) Recurrent Neural Network (RNN): modified RNN originally employed for text generation was implemented in Python using a Keras backend. RNNs are a straightforward adaptation of the standard feed-forward neural network allowing it to model sequential data; its high dimensional hidden state and nonlinear evolution enabling it to integrate information over many iterations for making accurate predictions [14]. RNN dataset train-test splits were 80:20, and the temperature was set to 0.15 for all sequences generation steps.

Both Seq2Seq–LSTM and RNN models were compiled with an Adam optimizer and used categorical cross entropy for the loss metric track over the course of training. Similarly with both models, all duplicate sequences – generated sequences present in the FPbase or sdAb-DB datasets or in the generated sets – were removed prior to analysis.

For each model, three different sampling methods were evaluated for generation of new sequences: 1) random, 2) using clusters from raw distances, and 3) TPE-based method. For random sampling, Seq2Seq–LSTM was trained with the entire dataset, while for RNN no tuning step was performed. For methods two and three, sampling from clusters arrived at via raw distances or TPE. The models were either trained or tuned using sequences from a selected cluster.

2.3 Clustering and Protein Structures

Most data analysis and visualizations were performed using functions in the tidyverse R package, a superset package consisting of dplyr, tidyr, ggplot2, among others [15]. The Peptides package in R was used for calculating protein features [16].

For clustering beyond those described above in *Dimensionality Reduction and Sequence Generation*, both hierarchical clustering using *hclust* function, part of the stats package and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for discovering clusters of similar sequences [17] were used. *Hclust* function's *method* parameter was set to "*complete*" throughout the study. For FPs, the DBSCAN parameters were as follows: *epsilon* was set to 85 and *minimum points* was set to 4. For sdAbs, the parameters were 25 for *epsilon* and 5 for *minimum points*. Otherwise for both *hclust* and *DBSCAN* default settings were used.

For protein structure analysis and prediction, image rendering was visualized through PyMOL software [18] (https://www.pymol.org). Known structures for sdAbs and FPs were downloaded from the Protein Data Bank [19] (https://www.rcsb.org/), and other protein structures were predicted by I-TASSER secondary structure prediction server [20].

Protein classification models was constructed using XGBoost from the xgboost library (Chen and Guestrin, 2016) where each of the sequences were classified into red, yellow, green, blue for FPs and into alpaca, camel/camelid, and llama for sdAbs. The input peptides sequences were encoded numerically to vectors, each amino acid or padding characters – which were appended to the end vector below the maximum length (200) – receiving a unique number. The data was randomly shuffled and split into training and test sets at a 90:10 ratio and cross validation was performed to determine the best model parameterization using repeated shuffle-splits.

3. Results

3.1 Overview

Our method makes use of TPE paired with a sequence-generating model (**Figure 1**). This simple system applies MSA to the input whole dataset of sequences. The sequences are further processed into *AAStringSet* objects to be converted into distance matrix by the *stringDist* function. The distance output is used directly into a dimensionality reduction algorithm – in this case TPE – which yields a two-dimensional array for each of the input sequences. If the sequence generative model has a fine-tuning step, such as in our text-generating RNN, the initial baseline model can be trained at this stage using the full dataset of sequences.

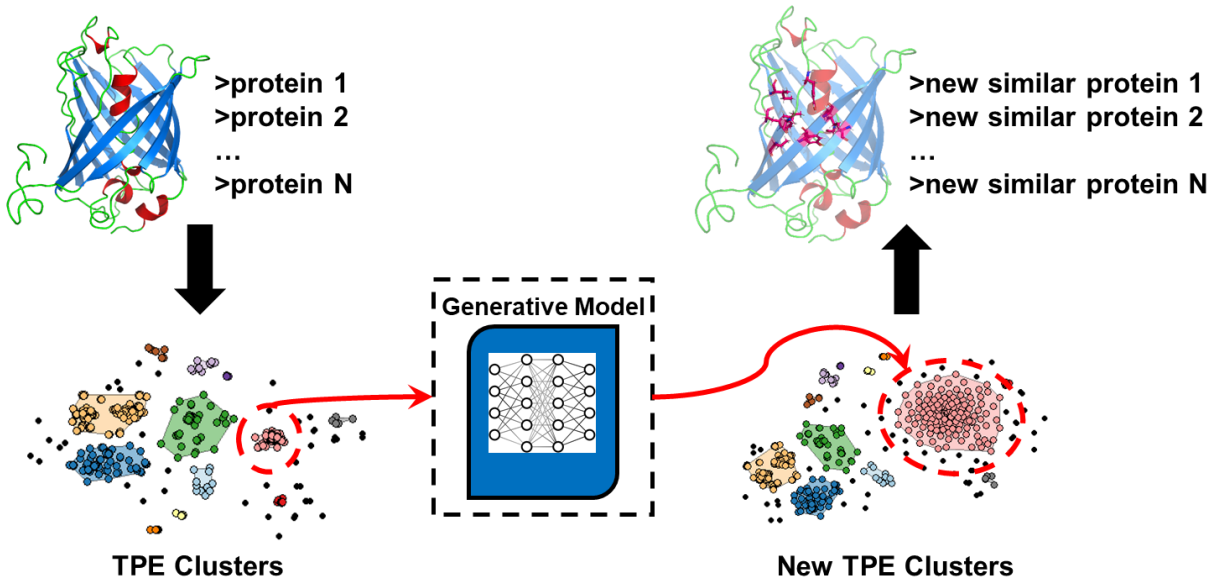

Figure 1 - Schematic of using TPE to direct sequence generation. A dataset consisting of protein sequences are subjected to MSA, clustering, and TPE. The output of TPE is used for identification of viable sequence-similar clusters, which are used to inform the sequence generating model (RNN shown).

Following output from TPE, clustering was performed on two dimensions using DBSCAN which auto-selects clusters where the input is a 2D array and two parameters: *epsilon* and *minimum points*. Both *epsilon* and *minimum points* are specified in the *Methods* section. The clustering by DBSCAN is analyzed by differentiation of attributes of interest (i.e., emission wavelength in FPs). After selecting a cluster of interest, the cluster is then used as the small fine-tuning dataset for the pre-trained model and generated sequences are further assessed.

3.2 TPE-Mediated Clustering

Hierarchical clustering when paired with TPE, rather than using raw distances, produces superior clustering for the randomly sampled datasets evaluated (**Figures 2 and 3**). Computed string distances were then either input directly into a hierarchical clustering function plotted as a dendrogram, or input into TPE and TPE coordinates were then converted to distances and used as input for hierarchical clustering. The circular dendrograms clearly show that the majority of sequences are placed into a single cluster. For both datasets, the raw distances yield clusters of significantly lower quality where the distribution of counts per cluster skews strongly to the largest three clusters, whereas TPE cluster counts are more evenly distributed (**Figure 4**).

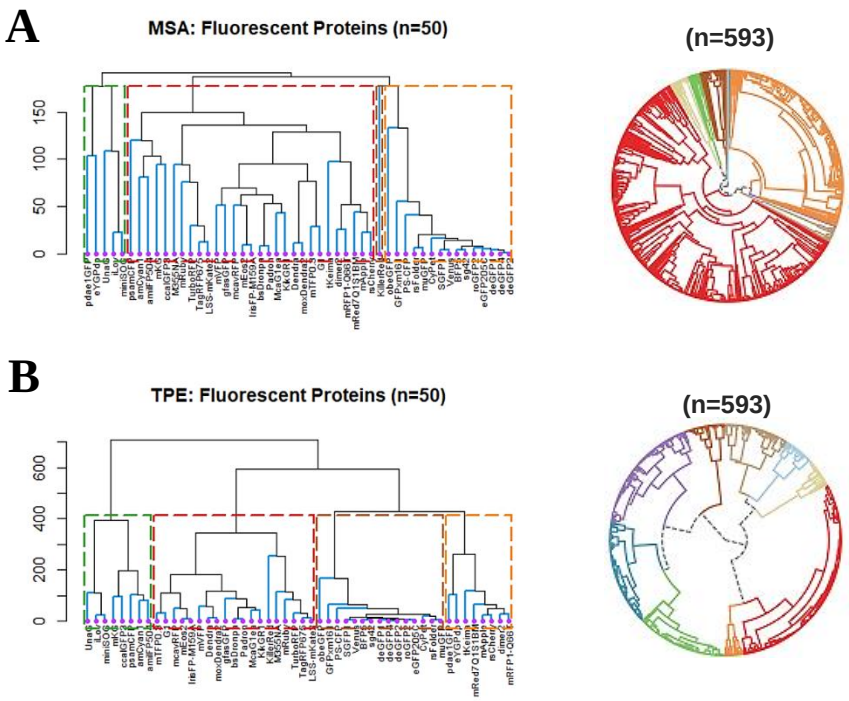

Figure 2. Hierarchical clustering of fluorescent protein sequences using multiple sequence alignment distances and TPE-computed embeddings. A) The multiple sequence alignment dendrogram shows the crowded cluster of sequences, with the majority placed in a single cluster for both datasets (red). B) TPE dendrogram displays a significantly more balanced clusters of FP sequences. Both radial (right panels) MSA- and TPE-clustered dendrograms (left panels) show simple hierarchical clustering of the randomly sampled 50 FP sequences.

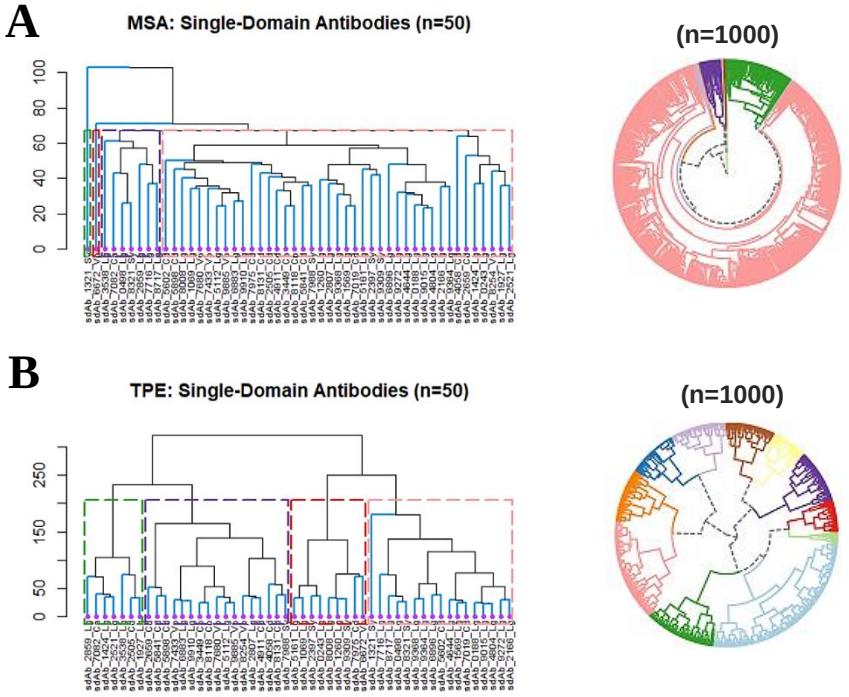

Figure 3. Hierarchical clustering of single-domain antibodies using multiple sequence alignment distances and TPE-computed embeddings. A) Similarly, the resulting dendrogram shows the crowded cluster of sequences, with the majority placed in a single cluster for both datasets (pink). B) TPE dendrogram displays a significantly more balanced clusters of sdAb sequences. Both radial (right panels) MSA- and TPE-clustered dendrograms (left panels) show simple hierarchical clustering of the randomly sampled 50 sdAb sequences.

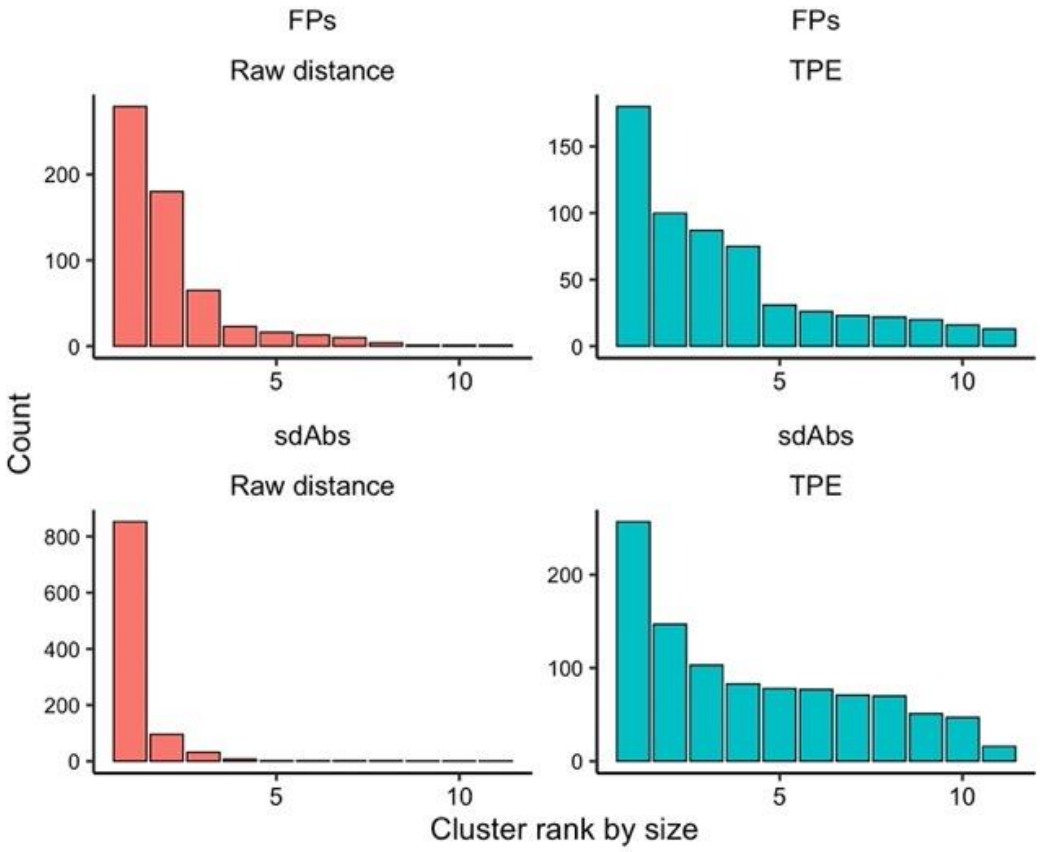

Figure 4. Hierarchical clustering sequences using multiple sequence alignment distances and TPE-computed embeddings. The multiple sequence alignment (MSA) dendrogram (left) shows the crowded cluster of sequences which could be deemed potentially inefficient in characterizing larger amount of data. On the other hand, TPE dendrogram (right) shows a significantly more balance in the relative heights (distances) of clustered sequences.

Embeddings and dendrograms have long been used as complementary representations for dissimilarities [21]. The crowding problem in clustering is usually prevalent in embeddings. One of its causes occur when the intrinsic dimensionality of the data exceeds the embedding dimensionality which results inadequate space in the embedding to allow clusters to separate [11]. With this, the separation of similar sequences is initialized by conducting multiple sequence alignments of protein sequences to discover divergent sequences. Dissimilarity measurements are then clustered by TPE which preserves each sequence's connectedness, and thus grouped into their corresponding amino acid characteristics. Differences in hierarchical clustering between raw multiple sequence alignment distances and TPE-computed embeddings are distinct and distinguishable in grouping protein sequences.

3.3 Preprocessing for Sequence Generation

In further preprocessing, we used DBSCAN [17] for discovering clusters of similar sequences. In this process, TPE-computed embeddings of sequences are processed through fast fixed-radius nearest neighbor computation using a K-dimensional (K-D) tree Euclidean distance [22]. The algorithm starts with an arbitrary point *p* and retrieves all points density-reachable from *p*; if there are no points density-reachable from *p*, DBSCAN visits the next point of the database. The importance of using densities as a clustering scheme is efficient for visualizing dissimilarities between all the sequences available in the dataset. Here, we compared the dimensionality reduction of TPE to PCA and t-SNE, each on two dimensions using alignment-based distances as inputs, and cluster assignments from DBSCAN were applied to TPE coordinates; same coloring from the TPE clustering was applied to each other method (**Figure 5**). Without considering coloring, PCA results for both datasets clearly show inadequate separation, especially for the sdAb-DB dataset where no clear clusters are visible. In contrast, clusters are seen for both t-SNE and TPE results. After further analysis, cluster separation when comparing t-SNE and TPE for using the adjusted Rand Index (ARI) measurement, TPE has the highest score of 0.71 compared to 0.60 for FPs, and the highest score of 0.60 vs. 0.49 for sdAbs. These results suggest that TPE outperforms t-SNE by evenly distributing points for each cluster, while also providing better separation. These results support findings in the original report on TPE [11]. However, the relatively lower resolution of t-SNE does not mean its output is useless, and we will make use of its clustering in addition to TPE in the next section.

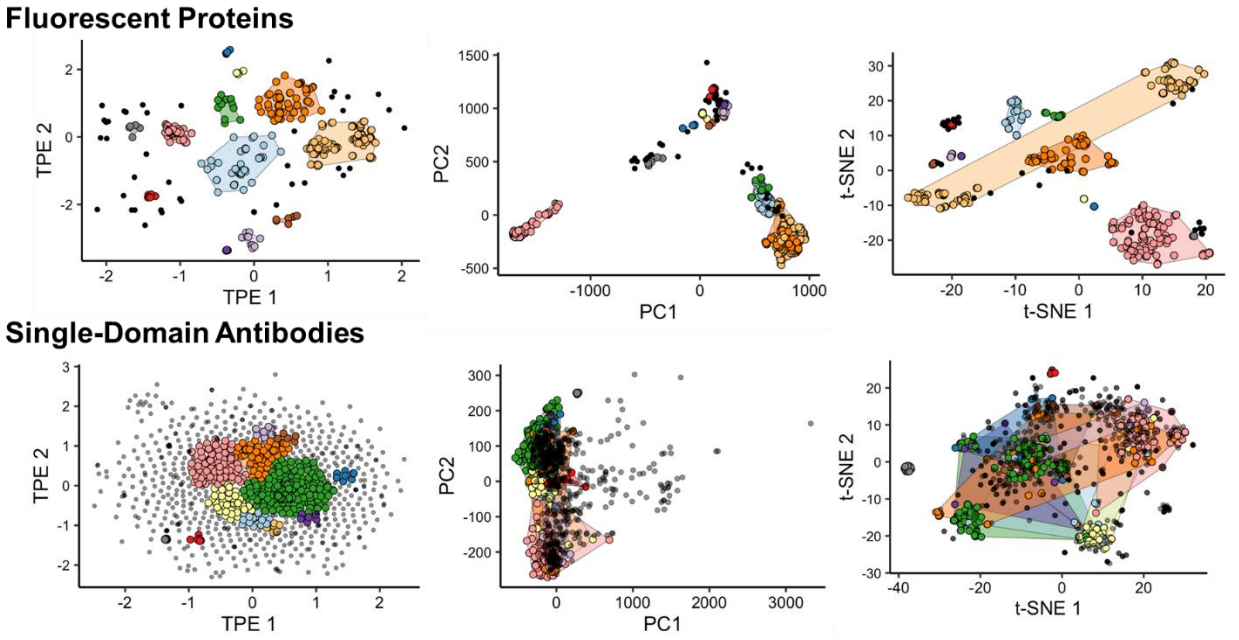

Figure 5. Dimensionality reduction comparison. Proteins, FPs (top) and sdAbs (bottom), are visualized by three different dimensionality reduction methods: TPE, PCA, and t-SNE. Clusters are colored according to DBSCAN-based clustering applied to TPE, which is used to color each plot. Each cluster is encircled at its outermost point. Points that were not clustered by DBSCAN are colored black.

3.4 TPE Comparisons with T-SNE

Initial dimensionality reduction comparison from Figure 4 suggests a slight propensity of t-SNE to perform sufficient clustering of similar sequences. Unlike PCA which is a linear dimensionality reduction technique that maps as many eigenvectors (principal components) as needed for preserving variance [23],

t-SNE behaves similarly to TPE – the embedding of high-dimensional data in a low-dimensional space is nonlinear and the vectors are mapped into two or three dimensions [24].

3.4.1 TPE vs T-SNE: Fluorescent Proteins

Clusters from FPbase sequences can be sufficiently separated with TPE or t-SNE (**Figure 6**). t-SNE separated clusters well for FPs (**Figure 6A**), and this projection was also utilized to highlight the 100 proteins with the lowest emission wavelengths in FPbase, which are visualized as dispersed throughout the space (**Figure 6B**). Conversely, applying the same process to TPE clusters (**Figure 6D-F**), we found that several sequences were not able to cluster (seen as solid black points; **Figure 6D**). For the 100 sequences with the lowest emission wavelength, the results were similar to t-SNE where they were uncontrolled and distributed throughout the clusters (**Figures 6B and 3E**). Striking similarities were also observed for t-SNE and TPE as both methods produced 12 clusters with similar distributions (**Figures 6C and 6F**)

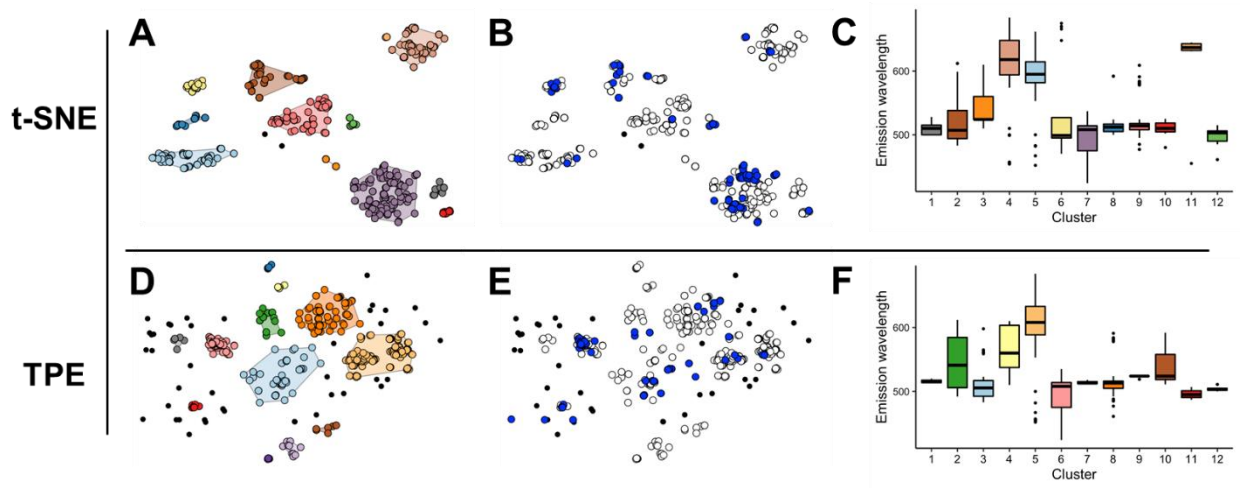

Figure 6. t-SNE and TPE clusters of FP sequences. A) Embedding of FP sequences via t-SNE. B) The same t-SNE embedding with the 100 lowest emission wavelength blue-green FPs highlighted. C) Emission wavelength distributions of the different t-SNE clusters. D) Embedding of FP sequences via TPE. E) The same TPE embedding with the 100 lowest emission wavelength blue-green FPs highlighted. F) Emission wavelength distributions of the different TPE clusters.

3.4.2 TPE vs T-SNE: Single-Domain Antibodies

Consequently, we investigated the identical process but applied to the sdAb-DB dataset for production of novel sdAb sequences (**Figure 7**). t-SNE also separated clusters well for sdAbs (**Figure 7A**), and this projection was also utilized to highlight a random sample of 100 llama antibodies, which are visualized as dispersed throughout the space, but mostly contained in a large cluster (**Figure 7B**). However, alpaca-sourced sdAbs were distributed into several clusters, while llama-sourced sdAbs is predominantly found in a large cluster (**Figure 7C**). Applying the same process to TPE clusters, (**Figure 7D**), we found that unlike the t-SNE results, the 100 llama sdAbs are more dispersed throughout the coordinate space (**Figure 7E**). Finally, alpaca sdAbs were not as dispersed as compared with t-SNE and the majority of llama sdAbs are still grouped in one large cluster (**Figure 7F**).

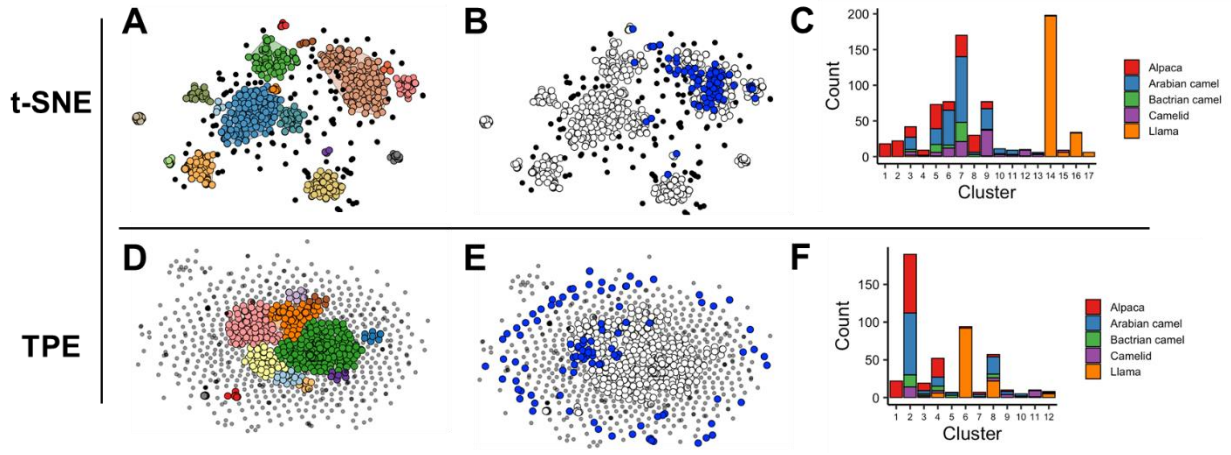

Figure 7. t-SNE and TPE clusters of sdAb sequences. A) Embedding of sdAb sequences via t-SNE. B) The same t-SNE embedding with the 100 random llama-sourced antibodies highlighted. C) Source name distributions of the different t-SNE clusters. D) Embedding of sdAb sequences via TPE. E) The same TPE embedding with the 100 random llama-sourced antibodies highlighted. F) Source name distributions of the different TPE clusters.

3.5 Sequence Generation with Deep Learning Models

The desired group of sequences to generate from in the production of novel proteins involved a series of processes: 1) *dimensionality reduction through TPE* – this step provides a selection of several clusters of related proteins; 2) *protein cluster selection* – in this study, the selected clusters from both TPE and t-SNE projections are the most populous clusters of similar sequences; and 3) *sequence generation* – the dataset of the selected protein cluster is then trained through RNN and Seq2Seq–LSTM to produce novel proteins that exhibit similar attributes (i.e., physicochemical properties, defined regions, or protein-specific characteristics). To compare the sequence generation results between fluorescent proteins and single-domain antibodies, and to have both relatively balanced datasets, we have reduced the sdAb dataset into 500 sequences for model training. This reduced dataset consists of naturally occurring proteins, excluding the synthetic sequences. Using the three most populous clusters of sequences as input, both Seq2Seq–LSTM- and RNN-generated FPs were then re-embedded using TPE. Increased cluster sizes were observed and the newly generated proteins were able to cluster with their parent sequences with the exception of one cluster discussed below (**Figures 8 and 9**).

3.5.1 Generated Fluorescent Proteins

Three selected populous FP clusters were mostly composed of red, cerulean, and yellow to green fluorescent proteins, respectively (**Figure 8A**; highlight colors not related to protein fluorescence). After sequence generation, both Seq2Seq–LSTM and RNN models were able to increase the size of the clusters of the same mapping of the parent sequences (**Figures 8B-C**). However, not all generated sequences were able to cluster at all (data not shown) due to higher dissimilarity by sequence distance. The sets of generated FPs that were able to cluster cannot be inspected individually due to their large amount, but a randomly selected FP generated from Seq2Seq–LSTM display similar 3D characteristics is discussed in section 3.7.

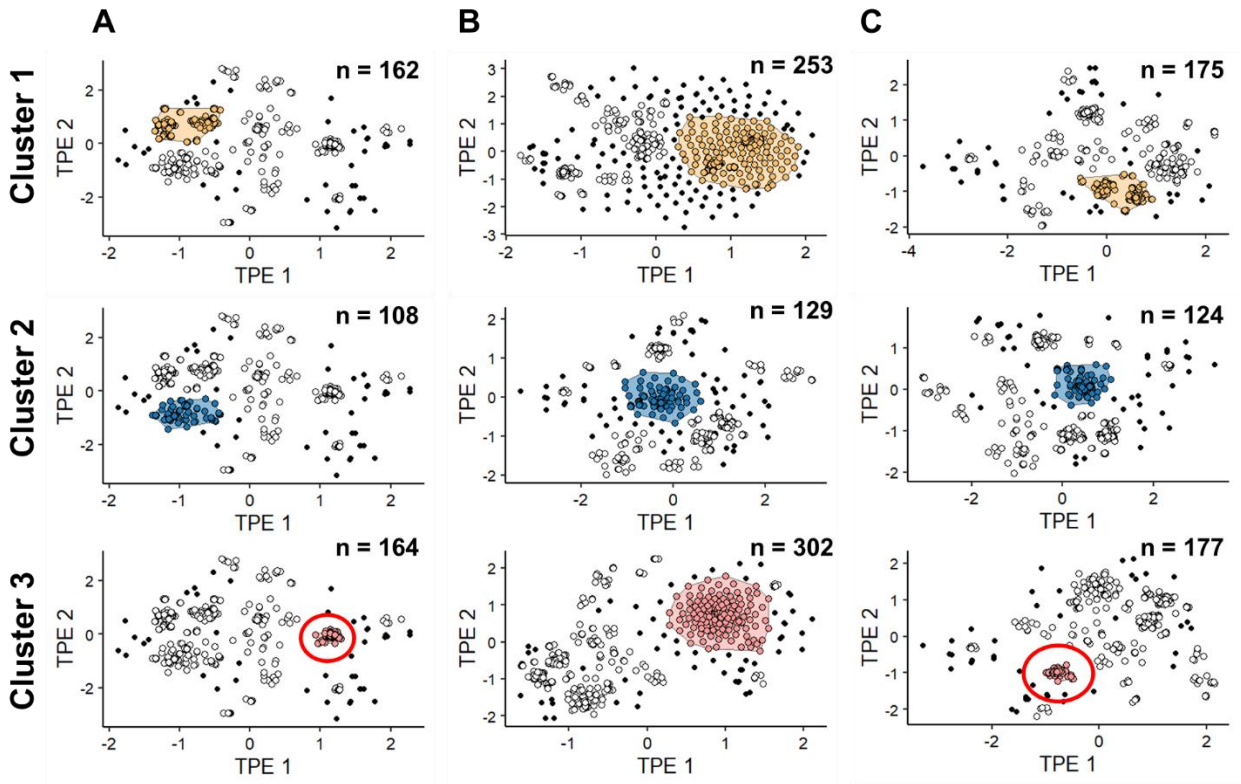

Figure 8. Generation of FP sequences. A) Embedding of FP sequences via TPE with highlighted populous clusters. B) Generation of FP sequences via Seq2Seq–LSTM, and C) via RNN. Red circles were added to support visualization.

3.5.2 Generated Single-Domain Antibodies

Three selected populous sdAb clusters were composed of multiple-sourced sdAb sequences, with the exception of Cluster 3 which consists of llama-sourced sdAbs (**Figure 9A**). After sequence generation, both Seq2Seq–LSTM and RNN models were also able to increase the size of the clusters of the same mapping of the parent sequences (**Figures 9B-C**). Similar with FPs, not all generated sequences were able to cluster due to higher dissimilarity by sequence distance included on the solid black points. On a specific note, RNN-generated sdAbs on Cluster 2 were not able to cluster with their parent sequences (**Figure 9C**). This indicates a change of mapping using TPE after sequence generation which could be a limiting factor.

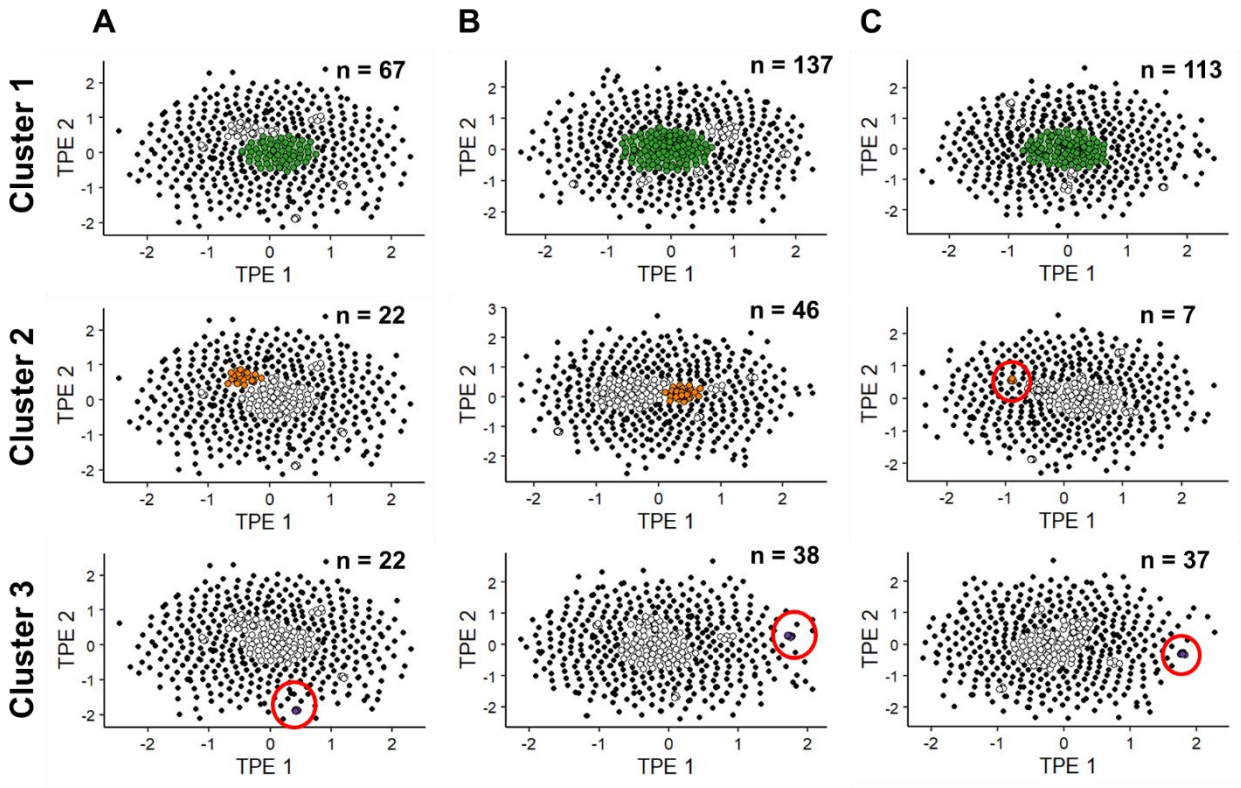

Figure 9. Generation of sdAb sequences. A) Embedding of sdAb sequences via TPE with highlighted populous clusters. B) Generation of sdAb sequences via Seq2Seq–LSTM, and C) via RNN. Red circles were added to support visualization.

3.6 Physicochemical Comparison of Generated Sequences

As initial means of determining whether the generated proteins were similar to the sequences that the models were trained on, we delved into the physicochemical characteristics of the novel sequences to compare distributions: charge, hydrophobicity, and sequence length. These attributes were chosen since they can be applied not only to FPs and sdAbs, but also to other potential proteins, such as antimicrobial peptides and membrane-spanning proteins where metrics such as hydrophobic moment at 100° are useful but not useful for the former.

For this comparison, we retained the three different populous clusters to evaluate relative to the distributions of measures from sequences in the parent database for both Seq2Seq and RNN. Each distribution calculated for generated sequences versus database sequences is statistically indifferent (**Figure 10**; $p>0.05$; Welch's t-test). This holds for each of the cluster-attribute pairings. Importantly, in most cases the distribution is narrower around the median for the generated sequences than for the database sequences, which is expected because the generated sequences were constrained within the selected sequence space.

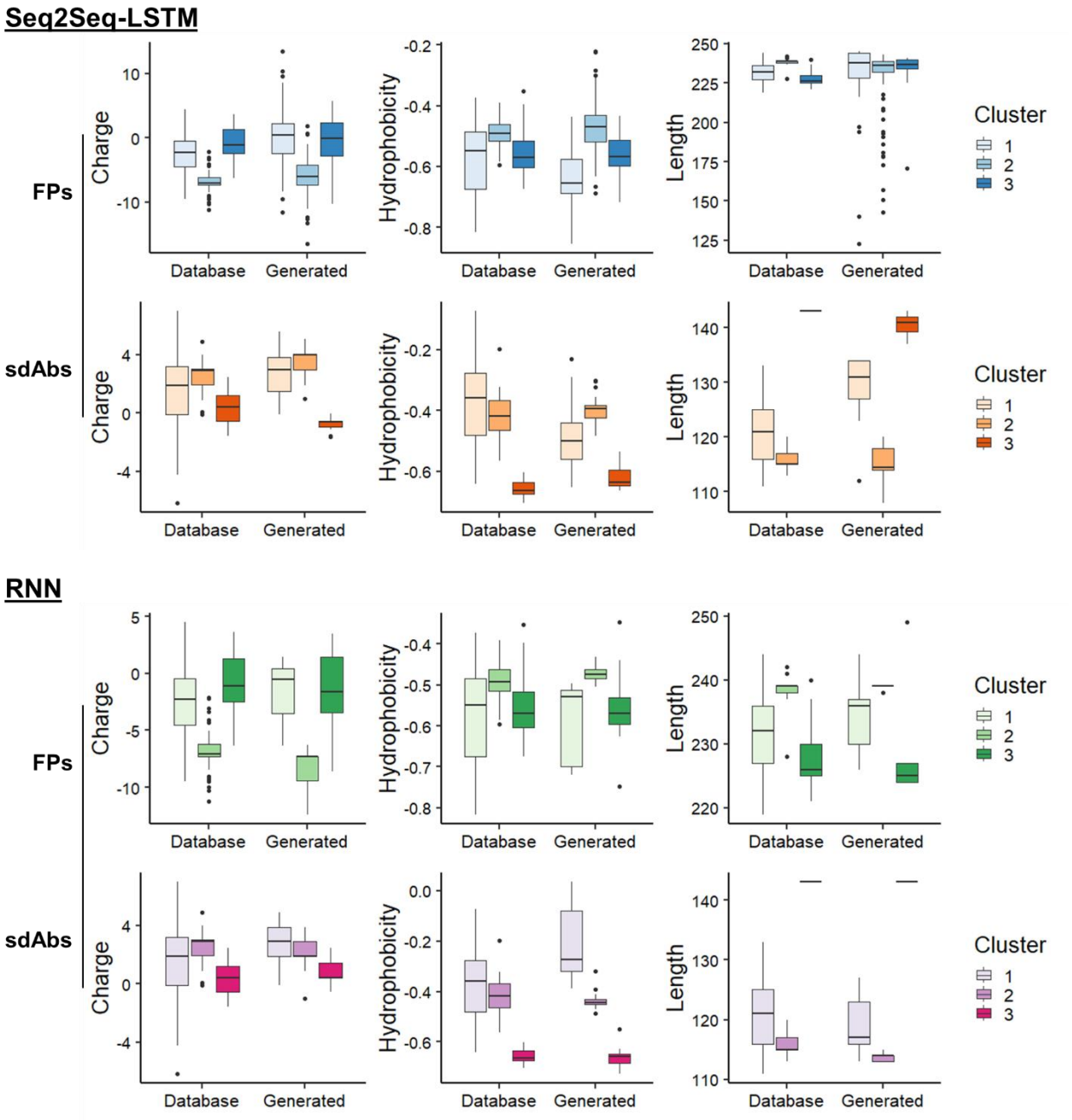

Figure 10. Physicochemical characteristics of generated sequences. Charge, hydrophobicity, and sequence length calculated for both the TPE-based cluster used for training model and the generated sequences. Both generated sequences from the two deep learning models – Seq2Seq (top) and RNN (bottom) – exhibit similar characteristics from their parent sequences. However, distribution for length deviated minimally for cluster 2 of Seq2Seq generated sequences exhibited by outliers (solid black points).

Seq2Seq results for the generation of FPs are shown to be largely similar with the database sequences, especially for hydrophobicity and charge. However, the distribution for length deviated minimally for Cluster 2 as shown on Figure 9. Furthermore, even though Seq2Seq model was also able to produce viable sdAbs, not all generated sequences were able to cluster with their parent sequences which suggest that the Seq2Seq model used in this study may require some additional tuning for better performance. However, these aspects of the results should be unrelated to the dimensionality reduction aspects of this study.

3.7 Protein-Specific Attributes

After analysis of sequence distance and physicochemical characteristics, we investigated whether the proteins contained conserved domains and other attributes important for regular function as FPs and sdAbs. We found that all FP sequences generated were classified as the Green Fluorescent Protein (GFP) superfamily in BLASTp, suggesting the automated domain detection function found significant similarity to known to the protein family. This was unsurprising given the significant sequence similarity of the TPE-RNN/Seq2Seq generated sequences with known sequences.

In particular, we closely looked at a random selection of generated sequences (**Figure 11**). Here we compared the 3D structure of avGFP – a distant ancestor of many of the fluorescent proteins found in the selected cluster using I-TASSER. AvGFP (PDB ID: 1EMA) exhibits the canonical β-barrel structure that surrounds a chromophore which is Ser65-Tyr66-Gly67, required for the native protein fold for both formation and fluorescence emission [25]. Looking at the predicted structure of the randomly selected sample, the synthetic protein exhibits similar β-barrel architecture of 1EMA – the protein fold consists of an 11-stranded β-barrel sheets around a coaxial helix. Extra helices were formed on positions Val60-Thr61-Leu62 and Asp134-Iso135-Leu136. The suspected chromophore site is still seen with a substitution of Ala63-Tyr64-Gly65. Performing an MSA using Clustal Omega on 13 Seq2Seq- and 13 RNN-generated proteins for the same cluster, we saw that all but one has viable chromophore residues (**Figure 12**). And as previously seen in Figure 5, it is again observed that Seq2Seq, relative to RNN, yield higher sequence variability under these configurations, with more variance both within and outside the chromophore region of the generated sequences.

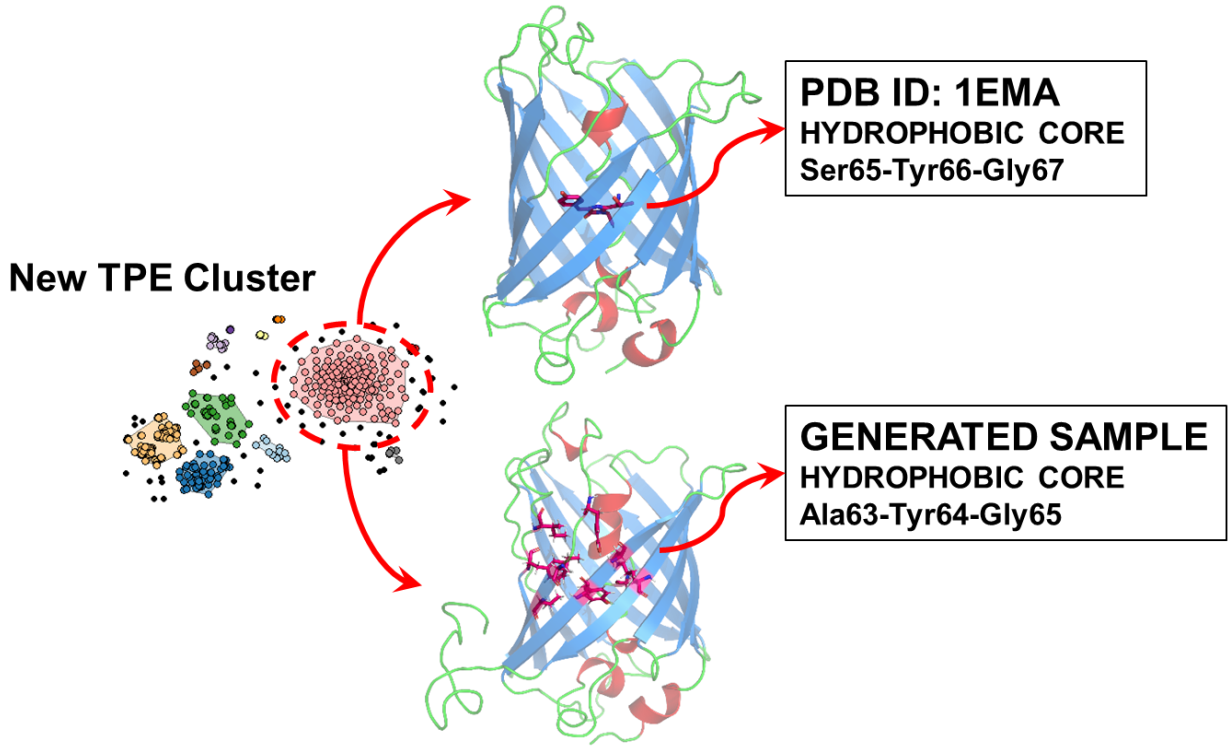

Figure 11. Secondary structures of 1EMA and randomly selected sample from the same cluster. (Top) *avGFP* (PDB ID: 1EMA) exhibits a β-barrel structure that surrounds a chromophore (highlighted as magenta sticks). (Bottom) Predicted structure of the randomly selected generated fluorescent protein sequence. The synthetic protein exhibits similar β-barrel architecture of 1EMA where the protein fold consists of an 11-stranded β-barrel sheets around a coaxial helix. Extra helices were formed on positions Val60-Thr61-Leu62 and Asp134-Iso135-Leu136. The predicted binding site of the randomly selected sequence is highlighted as magenta sticks. Suspected chromophore site is seen with a substitution, Ala63-Tyr64-Gly65. Helices are colored as red, sheets as blue, and loops as green.

Similar to FPs, all generated sdAb sequences were identified by BLASTp as within the Ig superfamily and more specifically as contain an IgV_H domain, suggesting all were sufficiently antibody-like according to the automated BLASTp classification process. Similarly, we found that all sdAb sequences generated by both Seq2Seq and RNN models were likely viable antibodies based on conserved regions using MSA. Specifically, among the llama-like sequences generated, following an MSA using Clustal Omega on RNN-generated sequences from a fine-tuned model trained on a specific TPE-based cluster, we saw that all sequences were >95% identical except for the CDR3 region responsible for target binding (**Figure 13**), which was expected given nearly sequences within the cluster were sourced from llamas.

Figure 12. Alignment of generated FP sequences. Selection of RNN-generated sequences and random selection of Seq2Seq sequences were aligned using Clustal Omega. Chromophore residues highlighted with box.

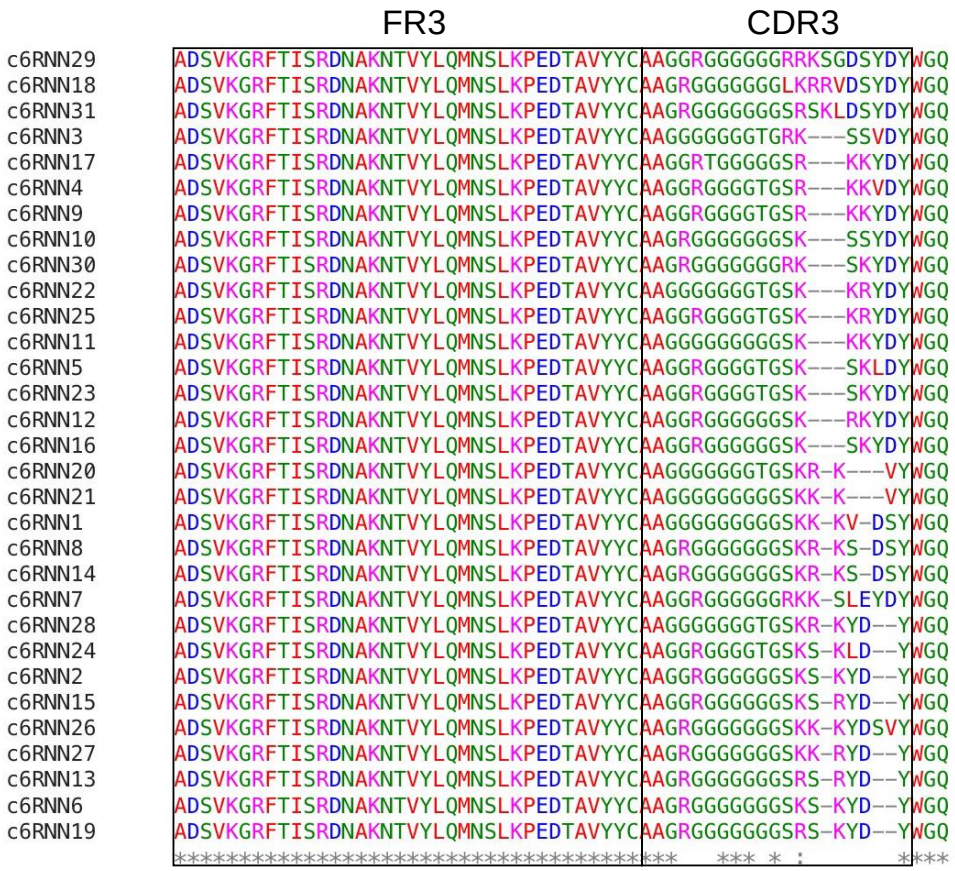

Figure 13. Alignment of generated sdAbs. Selection of RNN-generated sequences were aligned using Clustal Omega. FR3 and CDR3 portions are highlighted.

4. Discussion

FPs such as GFP from the Pacific Northwest jellyfish *Aequorea Victoria* has generated intense interest as a marker for gene expression [25] and it has been used in many imaging studies in cells and biosensors [26, 27] over the past several decades. For instance in Förster Resonance Energy Transfer experiments using fluorescent proteins has been used for detection of protein-protein interactions, enzyme activities, and small molecules in the intracellular environment [28]. FPs importance to the study of biological systems has spurred the continuous effort to discover its mutations for improving its characteristics. Specifically, the mutation of *avGFP*, *mGreenLantern*, has been used for neuronal imaging since it outperforms other GFPs on spectroscopic brightness, chemical and thermodynamic stability, compatibility with commercial antibodies, and whole-brain tissue clearing [29]. Interest in discovering more fluorescent proteins is still desirable since its discovery due to their differences in detectability, fluorescence emissions, and imaging resolutions in various cellular systems.

Similarly, for sdAbs, although conventional poly- and monoclonal antibodies are indispensable reagents in basic research, diagnostics, and therapeutics, their shortcomings including batch-to-batch variability of the polyclonals, and very high cost and time-consuming production of monoclonals has pushed scientists toward continued search for development of new sdAb sequences with the similar favorable properties of the larger counterparts, but with improvements. These improvements generally stem from their smaller size, allowing for efficient penetration into solid tumors and can allow for passage

through the blood brain barrier for treatment of various maladies [30]. For these, reasons generation of both sdAbs and FPs was focused on in this proof-of-concept study.

Many studies have utilized generative deep learning models and related techniques for production of novel peptide and protein sequences, with particular success in robust, easily evaluated systems like antimicrobial peptides. This study seeks to solve issues in generation of proteins in fragile systems, such as FPs and sdAbs, where small differences can lead to very different attributes of the new protein. The clearest example of this in FPs wherein some of the lowest emission wavelength (blue green) and highest emission wavelength (red) proteins are placed near one another, even in a sequence distance-informed projection such as TPE. This is a fragile system such that there exists a very high risk of sampling in a region of sequence space neighboring proteins that are very different in character from the target characteristics. Since sampling from that output could vary widely (e.g., obtaining a red when the desired is green), and based on the results, this is the outcome if one would sample from either all points with low emission or by using t-SNE clusters. Here, we have shown that TPE-based clustering informed generation allows reliable output from a region of sequential space with no/low neighbors with unwanted attributes, and therefore low likelihood of stumbling upon a variant with surprising properties. Since at the early phase of using the TPE clustering in evaluating where to sample from, this enabled us to visualize which regions are proximally viable reference for sequence generation, thus avoiding issues only seen at the stage of expensive and time-consuming protein production.

TPE enables this evaluation by revealing clusters while not forcing clustering where it does not exist, all in a sequence-distance informed process. Most dimensionality reduction methods can struggle because of the crowding problem, especially on datasets such as those investigated in this study where some of the data points differ by a single mutation. In their original study, Shieh *et al.* found that TPE was able to separate clusters of interest extremely well compared to both PCA and t-SNE [11], succeeding at this operation due to its incorporation of the single linkage dendrogram in the embedding and by preserving the clustering at all resolutions. In addition, although t-SNE in this study has the appearance of producing well-formed clusters, particularly for FPs, many of these are in effect artificial upon further inspection where no clear attribute of the protein is identified for members of the cluster. Moreover, it is well-known that force-based methods such as t-SNE can find clusters in datasets where none are present [31]. However, it is important to note that TPE scales with cubic time complexity (O^3), making it a significantly slower dimensionality reduction algorithm when the dataset is large.

Using TPE, we expect to better inform sampling or fine-tune models trained on systems less robust to mutations, such as FPs and sdAbs. Although beyond the scope of this study, we believe this to be the case relative to generative models of higher complexity as well, such as VAEs and GANs. For methods similar to VAE, where it is enabled to move around in the latent space to sample from different regions for producing novel sequences, it is unclear and unlikely whether a neighbor in latent space is simultaneously a neighbor in sequence space, and whether nearby sequences have very different attributes from the target characteristics. Using TPE-based clustering with a good-performing model, such as RNN to generate novel sequences, has a higher guarantee that the sequences produced will conform to existing members of that cluster, provided that sufficient numbers of members exist and no large gaps are present. Overall, we find this method to be of great promise for a wide range of generative models for biomolecular design and since this combined method requires no additional data sources beyond the provided dataset of sequences and can deal with a wide range of different protein types. Since it has no inherent properties specific to proteins, it has clear streamlined use for generation of both new proteins and other biomolecular sequences.

Acknowledgements

We acknowledge funding support through base funds of the Naval Research Laboratory (WU# 1V33) and funds from the Defense Threat Reduction Agency (HDTRA1033536).

References

1. Bowman, S.R., et al., *Generating sentences from a continuous space.* arXiv preprint arXiv:1511.06349, 2015.
2. Dean, S.N. and S.A. Walper, *Variational Autoencoder for Generation of Antimicrobial Peptides.* ACS Omega, 2020.
3. Dean, S.N., et al., *PepVAE: Variational Autoencoder Framework for Antimicrobial Peptide Generation and Activity Prediction.* Front Microbiol, 2021. **12**: p. 725727.
4. Linder, J., et al., *A generative neural network for maximizing fitness and diversity of synthetic DNA and protein sequences.* Cell systems, 2020. **11**(1): p. 49-62. e16.
5. Hu, Z., et al. *Toward controlled generation of text*. in *International conference on machine learning*. 2017. PMLR.
6. Das, P., et al., *Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences.* arXiv preprint arXiv:1810.07743, 2018.
7. Tucs, A., et al., *Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks.* 2020.
8. Müller, A.T., J.A. Hiss, and G. Schneider, *Recurrent neural network model for constructive peptide design.* Journal of chemical information and modeling, 2018. **58**(2): p. 472-479.
9. Nagarajan, D., et al., *Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria.* Journal of Biological Chemistry, 2018. **293**(10): p. 3492-3509.
10. Min, S., et al., *Pre-training of deep bidirectional protein sequence representations with structural information.* IEEE Access, 2021. **9**: p. 123912-123926.
11. Shieh, A.D., T.B. Hashimoto, and E.M. Airoldi, *Tree preserving embedding.* Proceedings of the National Academy of Sciences, 2011. **108**(41): p. 16916-16921.
12. Sutskever, I., O. Vinyals, and Q.V. Le, *Sequence to sequence learning with neural networks.* Advances in neural information processing systems, 2014. **27**.
13. Wu, Y., et al., *Google's neural machine translation system: Bridging the gap between human and machine translation.* arXiv preprint arXiv:1609.08144, 2016.
14. Sutskever, I., J. Martens, and G. Hinton, *Generating text with recurrent neural networks*, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 2011, Omnipress: Bellevue, Washington, USA. p. 1017–1024.
15. Wickham, H., et al., *Welcome to the Tidyverse.* Journal of open source software, 2019. **4**(43): p. 1686.
16. Osorio, D., P. Rondón-Villarreal, and R. Torres, *Peptides: a package for data mining of antimicrobial peptides.* Small, 2015. **12**: p. 44-444.
17. Schubert, E., et al., *DBSCAN revisited, revisited: why and how you should (still) use DBSCAN.* ACM Transactions on Database Systems (TODS), 2017. **42**(3): p. 1-21.
18. DeLano, W.L., *PyMOL*. 2002.

19. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallographica Section D: Biological Crystallography, 1998. **54**(6): p. 1078-1084.

20. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nature methods, 2015. **12**(1): p. 7-8.

21. Shepard, R.N., *Multidimensional scaling, tree-fitting, and clustering.* Science, 1980. **210**(4468): p. 390-398.

22. Hahsler, M., M. Piekenbrock, and D. Doran, *dbscan: Fast Density-Based Clustering with R.* Journal of Statistical Software, 2019. **91**(1): p. 1 - 30.

23. F.R.S., K.P., *LIII. On lines and planes of closest fit to systems of points in space.* Philosophical Magazine Series 1, 1901. **2**: p. 559-572.

24. Hinton, G.E. and S.T. Roweis. *Stochastic Neighbor Embedding*. in *NIPS*. 2002.

25. Ormö, M., et al., *Crystal structure of the Aequorea victoria green fluorescent protein.* Science, 1996. **273**(5280): p. 1392-1395.

26. Ai, H.-w., et al., *Fluorescent protein FRET pairs for ratiometric imaging of dual biosensors.* Nature Methods, 2008. **5**(5): p. 401-403.

27. Trigo-Mourino, P., et al., *Dynamic tuning of FRET in a green fluorescent protein biosensor.* Science Advances, 2019. **5**(8): p. eaaw4988.

28. Ding, Y., et al., *Förster Resonance Energy Transfer-Based Biosensors for Multiparameter Ratiometric Imaging of Ca2+ Dynamics and Caspase-3 Activity in Single Cells.* Analytical Chemistry, 2011. **83**(24): p. 9687-9693.

29. Campbell, B.C., et al., *mGreenLantern: a bright monomeric fluorescent protein with rapid expression and cell filling properties for neuronal imaging.* Proceedings of the National Academy of Sciences, 2020. **117**(48): p. 30710-30721.

30. de Marco, A., *Biotechnological applications of recombinant single-domain antibody fragments.* Microbial cell factories, 2011. **10**(1): p. 1-14.

31. Venna, J., et al., *Information retrieval perspective to nonlinear dimensionality reduction for data visualization.* Journal of Machine Learning Research, 2010. **11**(2).