CysPresso: Prediction of cysteine-dense peptide expression in mammalian cells using deep learning protein representations

Sébastien Ouellet[1], Angus Z. Lau[2,3], Larissa Ferguson[4], Tony K.Y. Lim[5]*

Author Information

[1] Ottawa, Ontario, Canada
[2] Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.
[3] Physical Sciences Platform, Sunnybrook Research Institute, Toronto, Ontario, Canada.
[4] Neurobiology Division, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom
[5] Vancouver, British Columbia, Canada
* Corresponding Author (tony.ky.lim@gmail.com)

ORCID: SO (0000-0002-3098-4907), LF(0000-0003-4274-8634), TL(0000-0003-1843-0060)

Abstract

Purpose: While many cysteine-rich peptides (CDPs) have potential therapeutic uses, our ability to identify CDPs that will express recombinantly in mammalian cells without prior experimentation is limited. To address this, we developed a novel machine learning model to predict the recombinant expressibility of CDPs based on primary sequence. *Methods & Results:* Random forest classifiers were trained using various protein representations based on natural language processing models (SeqVec), protein function (proteInfer), and protein structure (AlphaFold2). We found that all of these protein representations were useful at predicting CDP expressibility, though AlphaFold2 representations provided superior model performance. As AlphaFold2 generates multiple representations per peptide, we then sought to optimize the model by comparing its predictive success after training with different AlphaFold2 representation types. We found that concatenating the representations to create a combined AlphaFold2 representation yielded the best performance. To further improve model performance, we partitioned the model between knottins and non-knottins, which improved prediction of expressibility for knottins. Implementation of time-series classification further boosted performance of knottin expressibility prediction. *Conclusion:* We demonstrate that AlphaFold2 representations are suitable for tasks in addition to structure prediction, such as prediction of recombinant expressibility.

Keywords:
Protein representation, recombinant peptide expression, cysteine-dense peptide, knottin, deep learning, AlphaFold

Introduction

Cysteine-dense peptides (CDPs) are highly constrained by disulfide bonds, producing a molecular structure that confers extreme stability against proteolytic, thermal, and chemical degradation [1]. As a result, there is much interest in utilizing CDPs as drug scaffolds [2]. Naturally occurring CDPs have diverse pharmacological and toxicological effects, including activities against ion channels, G-protein coupled receptors and enzymes, as well as cytotoxic and anti-microbial properties [3–6]. One particular class of CDP, the inhibitor cysteine knot peptides—also known as knottins—has received much interest due to their confirmed and potential uses as analgesics, anthelmintics, anti-erectile dysfunction agents, neuroprotectives, antimalarials, antimicrobials, antitumour agents, protease inhibitors, toxins and insecticides [7, 8].

One of the challenges of using CDPs for therapeutic purposes is that they are difficult to synthesize and express. While the disulfide bonds are important for the stability of the molecule, incorrect linkage can lead to misfolded non-native isomers [9–11]. Recent work using high-throughput expression systems to express CDPs in mammalian cells has allowed the screening of hundreds of CDPs for expression [12, 13]. Though this work has led to the identification of a number of CDPs that can be efficiently produced through recombinant expression, our ability to discern the expressibility of CDPs based on amino acid sequences without the need for prior experimentation is extremely limited. Paramount to the design and development of recombinant CDPs for therapeutic purposes is a method to predict which CDPs will express in mammalian cells.

To create such a tool, we developed a machine learning classifier that predicts whether a CDP will be expressible based on its primary sequence. We tested various protein representations to identify the ideal model to utilize with machine learning algorithms to predict expressibility from primary sequences. We tested the performance of representations from an ELMo language model (SeqVec) [14], a deep dilated convolutional network trained on protein functions (ProteInfer) [15], and a neural network trained to predict protein structure (AlphaFold2) [16], and determined that AlphaFold2 representations performed best at predicting expressibility. The AlphaFold2 algorithm generates four protein representations, and we found that combining the representations enhanced model performance. Stratification of the dataset into knottin and non-knottin CDPs improved predictions of expressibility for knottin peptides. Additionally, model performance for knottin peptide expressibility could be further enhanced through application of a time series classification method [17]. Using this method, we created CysPresso, a machine learning model that predicts whether a given CDP is expressible in HEK293 cells. This tool will help increase the success rate of identifying drug-like peptides compatible with recombinant expression. CysPresso is freely available as a Colaboratory notebook at https://github.com/Zebreu/cyspresso.

Methods

Data preprocessing

This study utilized previously published data [12]. In this dataset, CDPs of 30 to 50 amino acids in length were tested for expression using the Daedalus lentivirus production system in HEK293 cells and subsequently purified by reverse phase chromatography (RPC). Products were deemed successfully

expressed if the RPC peak demonstrated a single peak under both oxidizing and reducing conditions, which is characteristic of a single folding state and the absence of heterogeneity.

The original dataset contained a list of 1249 CDPs, including UniProt accession number, primary sequence, source organism, and expressibility. Twenty two duplicate entries were removed, leaving 1227 CDPs. CDPs were identified as knottins if the keyword knottin (KW-0960) was mapped to its UniProt entry on the UniProt database [18].

Protein representations

Protein representations, or embeddings, encode proteins into a numeric format compatible with mathematical operations used in machine learning. Representations from SeqVec were generated according to previously described methods [14], generating 1024 per-residue features derived from the final hidden layer of the neural network. Averaging within-feature yielded an embedding of 1024 features per CDP. Similarly, representations from proteInfer were generated [15] to yield 1100 per-residue features, which were then averaged within-feature to yield an embedding of 1100 features per CDP.

To generate AlphaFold2 representations, AlphaFold2 [16] was run using the reduced database setting for Multiple Sequence Alignment (MSA) search on the Compute Canada supercomputer cluster. The five models used during the 14th community wide experiment on the Critical Assessment of techniques for protein Structure Prediction (CASP14) were utilized to generate representations, but only the representation from the model with the highest average predicted Local Distance Difference Test was used. Embeddings from the final hidden layer of the single (384 per-residue features), pair (128 per-residue features after averaging the first dimension), MSA (256 per-residue features) and structure module (384 per-residue features) representations were collected and per-residue features were then averaged. To create the combined AlphaFold2 representation, the single, pair, MSA and structure module features were concatenated, producing feature vectors of 1152 dimensions.

Supervised learning algorithms

Supervised learning was conducted using random forest classification [19]. Hyperparameters were left at default scikit-learn values, other than the n_estimators (number of trees) value which was set to 300. To estimate performance between different protein models, 10-fold cross validation was used, and the same shuffle was utilized across the different protein models. When data stratification was carried out, shuffling the train and test folds was balanced to contain an equal number of expressing and non-expressing non-knottins and knottins among each fold.

Random Convolutional Kernel Transformation

RandOm Convolutional KErnel Transform (ROCKET) is an approach developed for time series classification tasks that uses random convolutional kernels to transform sequential features [18]. To apply ROCKET, we created a concatenation of the four AlphaFold2 representations without any per-residue averaging. Sequences were padded with zeros to ensure that all sequences had equal length, producing a representation of size 50x1152 per CDP. These representations were then transformed by ROCKET,

convolving randomly generated kernels over the representation, producing a new representation of 20000 features per CDP. The ROCKET representations were then classified using a L2-regularized logistic regression model.

SHAP analysis

To investigate the impact of AlphaFold2 representation types and residue positions on the predictions, we trained a random forest classifier on representations of size 50x1152 per CDP and analysed individual predictions over the dataset using SHapley Additive exPlanations (SHAP), a method developed to compute the direct contributions from each feature toward the final class probability [22]. The absolute values of these contributions were summed and visualized to determine which features contribute the most towards expressibility. Utilizing SHAP values, we were able to isolate the contributions per AlphaFold2 representation type as well as the contributions per residue position. To account for differences in sequence length, SHAP values were normalized by the number of sequences of that length. As data transformation (such as ROCKET) breaks direct relationships between residue positions and feature contributions, SHAP analysis was only applied to the random forest model.

Leave-one-out cross validation

To evaluate the quality of the classification in terms of class-specific accuracies, we computed confusion matrices for both knottin and non-knottin models using ROCKET time-series classification with the combined AlphaFold2 representations as described above. A leave-one-out cross validation scheme was followed, training as many model instances as samples per set and aggregating the predicted labels over all samples.

Comparisons of model performance

The models discussed in this paper (except SeqVec) were evaluated with a permutation test over 50 random permutations of the stratified dataset, with a 90-10 split for training and validation sets. The different models were ranked in order of area under the receiver operating characteristic curve (AUC) values, and the average rank out of 50 permutations was determined and plotted on a critical difference diagram using previously described methods [20], with the significance level set to 0.05.

Code

Code used to generate this work is available on github at https://github.com/Zebreu/cyspresso.

Results

Peptide expressibility database

The library used in this study characterized the expression of 1227 CDPs [12]. The lengths of the CDPs varied between 30 to 50 amino acids in length and were derived from a wide range of organisms, including mammals, arachnids, amphibians, reptiles, insects, birds, molluscs, fish, plants, and fungi. Of

these CDPs, 351 were classified as knottins according to the UniProt database (Table 1). The data was slightly skewed, with 58.6% and 47% of the non-knottin and knottin CDPs deemed expressible in HEK293 cells, respectively.

	Successful expression	Unsuccessful expression	Total
Non-knottin CDPs	513 (58.6%)	363 (41.4%)	876
Knottin CDPs	165 (47.0%)	186 (53.0%)	351
			1227

Table 1. Expressibility of recombinant non-knottin and knottin CDPs.
Using a high-throughput expression system in HEK293 cells, the expressibility of 1227 CDPs was determined [12]. Entries were cross referenced with the UniProt database to distinguish knottins from non-knottins.

Protein representation performance

Protein representations of CDPs were generated by obtaining neural embeddings from SeqVec, proteInfer, and AlphaFold2–models that are distinct in how they generate representations. For instance, SeqVec generates protein representations by utilizing the bi-directional language model 'Embeddings from Language Models' to identify relationships based on sequence syntax [14]. On the other hand, proteInfer [15] and AlphaFold2 [16] are neural network models trained to predict protein function and structure, respectively. Embeddings from these models were averaged and used as inputs for random forest algorithms to predict expressibility and tested using a stratified 10-fold cross validation. AUC was used as the evaluation metric (Fig. 1A). As random forest algorithms perform implicit feature selection, the performance of the random forest algorithm becomes a measure of feature performance. Thus, we examined the performance of protein representations generated from primary sequences at predicting expressibility. Out of the different protein representations tested, SeqVec had the least predictive power but still yielded usable models with AUC averaging 0.665. When classifiers were trained using embeddings from proteInfer, performance was improved with classifier AUC averaging 0.696. Interestingly, the best performance was obtained with embeddings from AlphaFold2**.** By using the single representation from AlphaFold2, AUC averaged 0.781 (Fig. 1B).

AlphaFold2 uses four different types of representations for structure predictions [16]. The MSA representation is based on the primary sequence and evolutionary relationships between proteins. The abstract single representation is derived from a linear projection of the first row of the MSA representation. The pair representation is a model of amino acids which are likely in close proximity based on a structure database search. These three representations are updated by the neural network and used as inputs to generate the structure module representation, which is a prediction of atom coordinates

and a per-residue confidence measure. As the four AlphaFold2 representations contain different information (evolutionary relationships, abstract features, known protein structures, and atomic positions), we examined whether any of them provided superior model performance at predicting CDP expressibility, and found that AUC for MSA, single, pair, and structure module representations were similar (0.789, 0.781, 0.790, and 0.788 respectively). We then examined the effect of concatenating the four representations and observed an improvement in model performance (AUC = 0.801) (Fig. 1C).

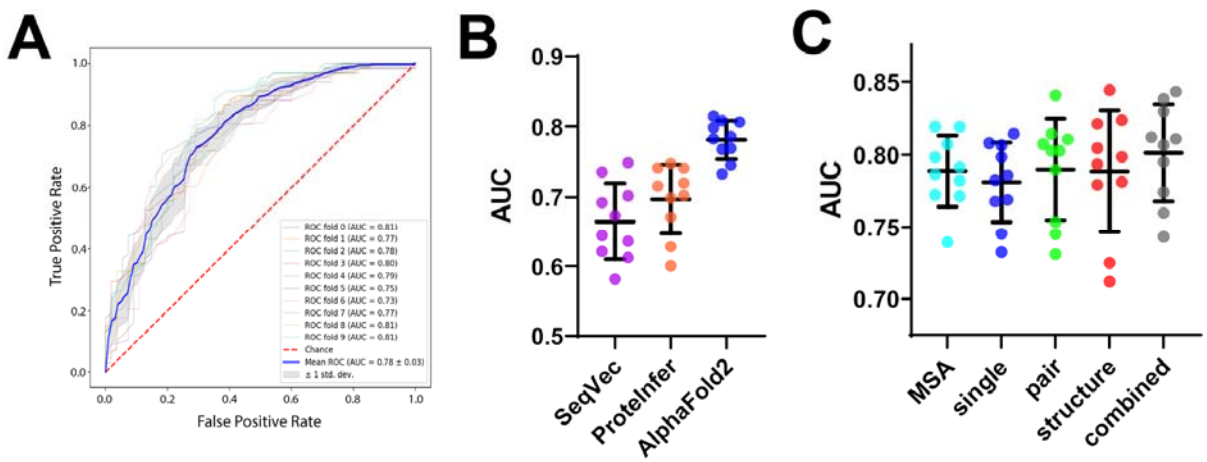

Fig 1. Random forest classifiers trained on SeqVec, proteInfer, and AlphaFold2 protein representations predict recombinant CDP expressibility.

Random forest classifiers were trained using protein representations generated by primary sequences of CDPs and performance was estimated by 10-fold cross validation. AUC was used as the performance metric. **(A)** An example showing receiver operating characteristic curves generated using the single representation from AlphaFold2 neural embeddings. **(B)** The performance of protein representations generated by SeqVec, proteInfer, and AlphaFold2. AlphaFold2 protein representations had the highest predictive performance. **(C)** The predictive performance of neural embeddings from the four representations generated by AlphaFold2 and a concatenated combined representation. The combined representation produced classifiers with the best performance at predicting recombinant CDP expressibility.

Generating a state-of-the-art classifier of recombinant non-knottin and knottin CDP expressibility

Having determined that combined AlphaFold2 protein representations provided the best performance for predicting recombinant CDP expressibility, we then sought to optimize the classification model. To accomplish this, we first examined whether the classifier would benefit from partitioning the data into non-knottin CDPs and knottin CDPs. Intriguingly, when the dataset was subset to knottins, the classifier performance was enhanced, although such an improvement was not observed when the dataset was subset to non-knottin CDPs (Fig. 2A).

To further improve classifier performance, a time series classification method was utilized. The neural embeddings generated by the models used in this study produce high dimensional feature vectors describing each residue in the peptide. The features were initially averaged to compress the sequence of amino acid residue features into a summary representation of a fixed size per peptide that was then used

to train the random forest classifiers. While averaging preserves information from each residue, information such as order and sequential relationships are lost. Therefore, we hypothesized that by using time series methods, sequential relationships could be captured to improve classifier performance. As a result, we implemented a time series transformation on each combined AlphaFold2 representation [17]. Before implementing ROCKET, classifier performance was estimated to be 0.827 ± 0.065 (Fig. 2A). Using the same cross validation folds, we found that employing ROCKET improved classifier performance for knottins and resulted in an estimated AUC of 0.865 ± 0.043 (Fig. 2B). Curiously, while the prediction of knottin expressibility was improved by the use of ROCKET, it appeared to have no impact on the prediction of non-knottin expressibility (Fig. 2B).

The proteInfer, individual AlphaFold2, combined AlphaFold2, and ROCKET models were compared by permutation test. Model performance was ranked according to AUC. For predicting the expressibility of non-knottin CDPs, the top ranked model was the combined AlphaFold2 representation without ROCKET time series transformation, though it was not ranked significantly better than the MSA, pair or ROCKET models (Figure 2C). On the other hand, for predicting the expressibility of knottin CDPs, the top ranked model was the combined AlphaFold2 representation with ROCKET time series transformation, which ranked significantly better than all of the other models (Figure 2D).

The finalized model, CysPresso, which utilizes partitioning and ROCKET time series classification, was evaluated by leave-one-out-cross-validation. Using this approach, classifier accuracy was determined to be 0.742 for non-knottin CDPs (Figure 2E, Table 2) and 0.789 for knottins (Figure 2F, Table 2).

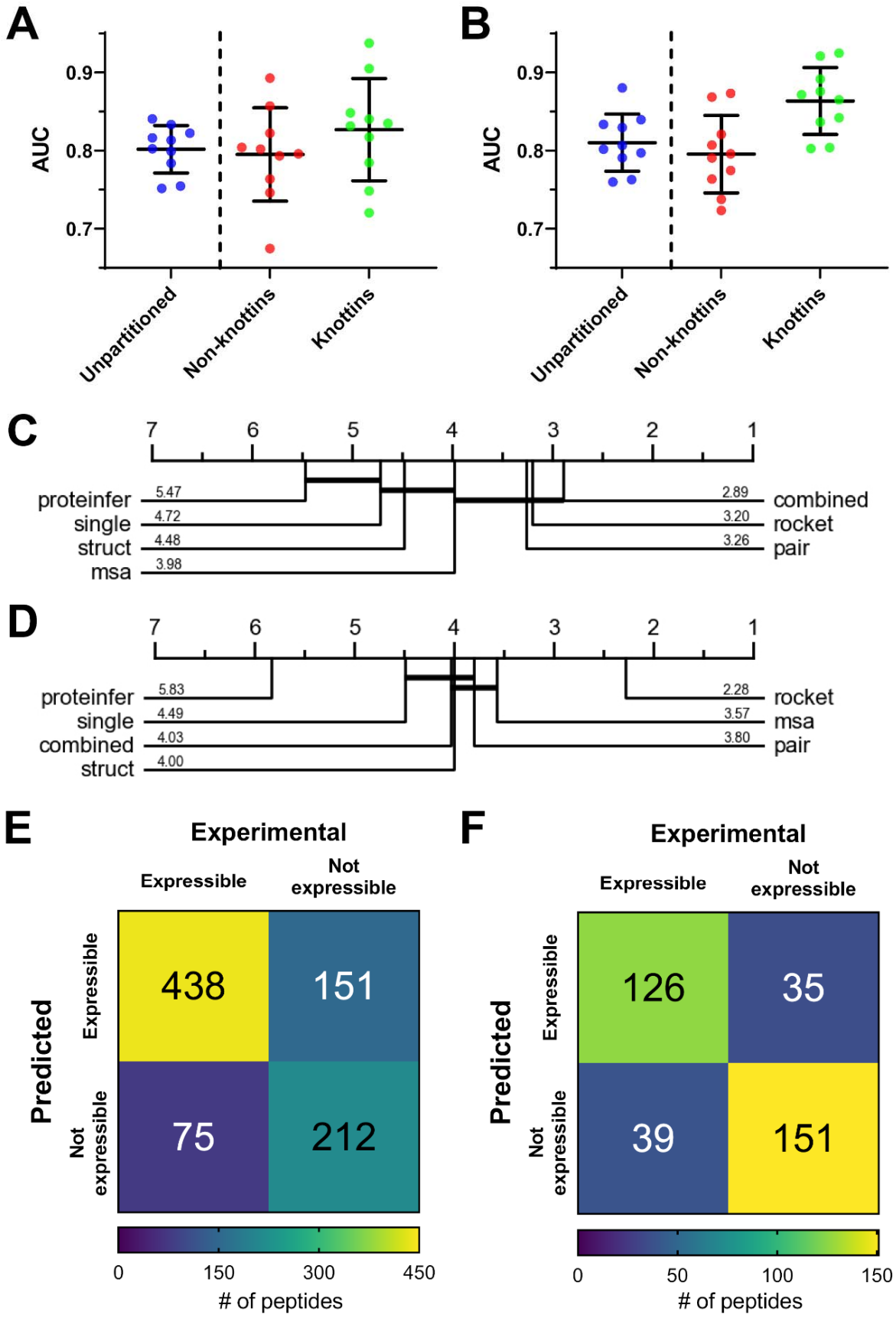
A
AUC
0.9
0.8
0.7
Unpartitioned
Non-knottins
Knottins
B
AUC
0.9
0.8
0.7
Unpartitioned
Non-knottins
Knottins
C
7
6
5
4
3
2
1
proteinfer 5.47
single 4.72
struct 4.48
msa 3.98
2.89 combined
3.20 rocket
3.26 pair
D
7
6
5
4
3
2
1
proteinfer 5.83
single 4.49
combined 4.03
struct 4.00
2.28 rocket
3.57 msa
3.80 pair
E
Experimental
Expressible
Not expressible
Predicted
Expressible
Not expressible
438
151
75
212
0
150
300
450
# of peptides
F
Experimental
Expressible
Not expressible
Predicted
Expressible
Not expressible
126
35
39
151
0
50
100
150
# of peptides

Fig 2. Dataset partitioning and time series classification improves prediction of knottin expressibility.
(A) Splitting the dataset into non-knottin and knottin partitions improves prediction of knottin but not non-knottin CDP expressibility when using random forest classifiers.
(B) Transforming AlphaFold2 protein representations with ROCKET time-series classification further improves prediction of knottin but not non-knottin CDP expressibility.
(C) Mean AUC rank for various models at predicting non-knottin expressibility (50 permutations). Training random forest algorithms on the combined AlphaFold2 representation provided the best performance. Ranks that are not significantly different are connected by horizontal lines.
(D) Mean AUC rank for various models at predicting knottin expressibility (50 permutations). Utilizing ROCKET on the combined AlphaFold2 representation provided the best performance. Ranks that are not significantly different are connected by horizontal lines.
(E) Confusion matrix of the final machine learning model for non-knottin CDPs evaluated by leave-one-out cross validation.
(F) Confusion matrix of the final machine learning model for knottins evaluated by leave-one-out cross validation.

Measure	Non-knottin CDP model	Knottin model
Sensitivity	0.853	0.764
Specificity	0.584	0.812
Precision	0.743	0.783
Accuracy	0.742	0.789
F1 Score	0.794	0.773

Table 2. Classifier performance metrics for CysPresso.

Model interpretability

We then posed the question: Which features are most important for determining expressibility in non-knottin and knottin CDPs? To answer this, we calculated random forest feature contributions by SHAP analysis [21]. Surprisingly, this analysis revealed that the region most important for the prediction of expressibility in non-knottins was near the C-terminus, whereas amino acid residues closer to the N-terminus were deemed more important for predicting expressibility in knottins (Fig. 3A). We then examined the relative importance of the different AlphaFold2 representations in the prediction of expressibility and found that the structure module representation was most important for predicting non-knottin expressibility, while the abstract single representation was most important for predicting knottin expressibility (Fig. 3B). This difference provides a strong rationale for partitioning the dataset into non-knottin and knottin subgroups for improved model performance.

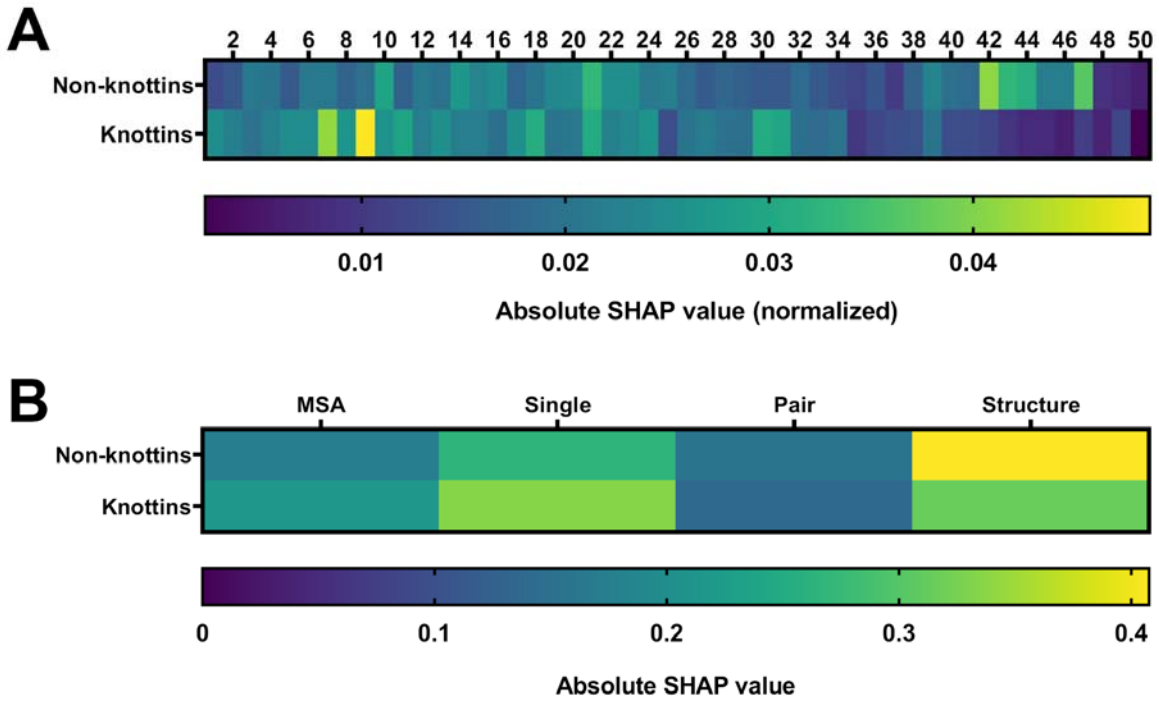

Fig 3. Features important for predicting expressibility in non-knottins are different from the features that are important for predicting expressibility in knottin peptides.
This figure shows absolute SHAP values, which show how much a feature affects prediction, calculated for different AlphaFold2 features.
(A) The relative importance of amino acid position for prediction of expressibility in non-knottin and knottin peptides. In non-knottins, a region near the C-terminus (position 42 to 47) was identified as playing an important role in determining expressibility. In knottin peptides, the features at amino acid position 7-9 near the N-terminus were most important for expressibility.
(B) The relative importance of the different AlphaFold2 representations for expressibility prediction of non-knottin and knottin peptides. For non-knottins, the structure representation module representation was most important for prediction of expressibility. On the other hand, in knottin peptides, the abstract single representation was most important for prediction of expressibility.

Discussion

The expressibility of recombinant peptides in human cells can be predicted from primary sequences using machine learning methods.

Peptide therapeutics occupy 5% of the global pharmaceutical market, and their approval rate has steadily increased over the last few decades [22]. Presently, peptide therapeutics face two major challenges: the poor yield of peptides from traditional solid-phase synthesis [23] and the intrinsically short plasma half-life of peptides [24]. As a result of recent advances, recombinant peptide production is now a cost effective, simple, and high-yield production platform for peptide drugs [25]. Our tool helps predict which peptides are compatible with recombinant expression, reducing the amount of time-consuming empirical experimentation required to identify expressible CDPs by allowing laboratory efforts to focus on the

candidates most likely to express. Furthermore, CysPresso may be useful in the identification of CDPs that are compatible with *in situ* expression technologies such as viral gene therapy and mRNA-based therapeutics, a method that may extend the *in vivo* activity of peptide therapeutics.

Machine learning algorithms have been used for the prediction of recombinant protein solubility in bacterial expression systems such as *E. coli* and *B. subtilis*. Recent examples for sequence-based prediction of protein solubility include SoluProt which achieves an AUC of 0.62 [26], SoDoPe which achieves an AUC of 0.71 [27], SKADE which achieves an AUC of 0.82 [28], and UniRep-RP which achieves an AUC of 0.64 [29]. Our model performance of 0.87 provides comparable performance for expressibility of CDPs in mammalian cells, and is to our knowledge the first study to model difficult-to-express peptide expression in mammalian cells.

AlphaFold2 representations retain information that is complementary to prediction tasks beyond protein structure.

In this study, we tested various protein representations and evaluated their usefulness at predicting expressiblity of difficult to express peptides. We began our investigation by utilizing neural embeddings from protein language models such as SeqVec and ProteInfer, which are based on statistical models and generated by training large scale sequence datasets. We also utilized AlphaFold2 neural embeddings, which include several representations based on evolutionary, positional, structural, and simulated relationships. AlphaFold2 protein representations provided better predictive performance than the purely language based models. This may be due to the structural information contained within AlphaFold2 representations, as the expressibility of a peptide is determined not only by its amino acid sequence but also by its tertiary and quaternary structure.

While AlphaFold2 accurately predicts protein structures, the suitability of AlphaFold2 neural embeddings for use in general functional prediction tasks is not well established. We find that AlphaFold2 neural embeddings are indeed suitable for general purpose tasks such as the prediction of expressibility of difficult-to-express peptides such as CDPs. Our results agree with work from other groups that have shown that AlphaFold2 produces a general-purpose protein representation that can be used for functional predictions [30].

Different variables determine the expressibility of non-knottin peptides and knottin peptides

Our analysis demonstrated that expressibility is determined by different protein features in knottin and non-knottin CDPs. Knottins are characterized by six cysteine residues making up three interwoven disulfide bridges that form a unique pseudoknot with conserved secondary structure [2]. It is hypothesized that during folding, the C-terminus of the peptide first forms a conserved □-hairpin, allowing disulfide bridges to form between cysteines II–V and III–VI. Folding is then completed by the N-terminal loop swinging around to create the cysteine I–IV disulfide bridge [31]. Since our data suggests that the N-terminal region is important for the prediction of expressibility, it is likely that the success of the final N-terminal folding step is required for successful expression of knottin peptides.

Our results also demonstrated that the various AlphaFold2 representations hold differential predictive value. For non-knottins, the structure representation was most important for predicting expressibility, while for knottins, the abstract single representation was most predictive. We speculate that this may be due to the fact that the non-knottin CDPs in this study was a heterogeneous group that included cyclotides, hitchins, growth factor cysteine knots and non-knotted CDPs [12], while the knottin peptides in this study were a homogenous group that all had the conserved inhibitor cysteine knot structure characteristic of knottins. As a result, structurally diverse non-knottin CDPs benefited more from structural information, whereas structural information was less important at providing features that determine expressibility in the structurally related knottins group. In this case, primary sequence relationships captured in the MSA representation provided more meaningful predictive value.

Time series classification improves AlphaFold2 feature performance for prediction of knottin expressibility

The goal of time series classification is to predict the class that a new sample belongs to using labeled ordered training data. While the data used with time series algorithms are typically sequences of discrete-time data, any data registered with some notion of ordering is compatible [32]. As AlphaFold2 and many other protein models such as SeqVec and ProteInfer produce such ordered data (ie. a neural embedding for each amino acid residue), we were able to cast our classification problem as a time series classification problem. We find that applying a time series transformation such as ROCKET improves the capture of ordered features relevant for the prediction of peptide expressibility when examining knottin peptides. Thus far, almost all studies that use protein representations for supervised machine learning tasks generate embeddings by taking the average of the pool of activations from the last hidden layer of the model [33]. Our results suggest that the use of time series transformations may provide a better summary of the protein representation by reducing the amount of data lost by averaging.

Study limitations

We analyzed a previously published dataset [12] that examined the expressibility of CDPs in HEK293 cells using a high-throughput Daedalus lentivirus transduction system. In that study, CDPs were considered successfully expressed if the peptide product was free of heterogeneity, as determined by reverse phase chromatography. It is important to note that, while reverse phase chromatography is a rigorous tool for ensuring chemical and conformational homogeneity, it is possible that the expressed products lack biological activity or that heterogeneous products deemed unexpressible may in fact be biologically functional. Additionally, the Daedalus lentiviral expression system that generated the dataset requires proteolytic cleavage to free the peptide from a mouse Igκ peptide sequence. Thus, it is possible our results are not generalizable to peptides expressed by other methods.

It is also important to note that the dataset used in this study is likely to be affected by selection biases. In ideal circumstances, the expression of CDPs would have been carried out with randomization to reduce the possibility of coverage bias and sampling biases. While CDPs were expressed from a variety of sources, randomization was not carried out during the selection of CDPs to screen for expressibility [12]. As a result, future work should aim to evaluate CysPresso with randomly sampled real-world data.

Conclusion

In this work, we used supervised machine learning to develop a tool to predict whether a CDP is expressible in a mammalian cell expression system. To accomplish this, we tested the performance of various protein representations and found that AlphaFold2 representations, which are typically used to generate predictions of protein structures, can also be used to predict expressibility. By combining the four AlphaFold2 representations, partitioning the model, and utilizing time series transformation, we further improved the performance of the expressibility classifier to yield an approach with state-of-the-art performance for predicting CDP expressibility in mammalian cells.

References

1. Northfield SE, Wang CK, Schroeder CI, Durek T, Kan M-W, Swedberg JE, Craik DJ (2014) Disulfide-rich macrocyclic peptides as templates in drug design. Eur J Med Chem 77:248–257
2. Gracy J, Chiche L (2011) Structure and Modeling of Knottins, a Promising Molecular Scaffold for Drug Discovery. Current Pharmaceutical Design 17:4337–4350
3. Molesini B, Treggiari D, Dalbeni A, Minuz P, Pandolfini T (2017) Plant cystine-knot peptides: pharmacological perspectives. British Journal of Clinical Pharmacology 83:63–70
4. Dongol Y, Cardoso FC, Lewis RJ (2019) Spider Knottin Pharmacology at Voltage-Gated Sodium Channels and Their Potential to Modulate Pain Pathways. Toxins (Basel) 11:E626
5. Scott A, Weldon S, Taggart CC (2011) SLPI and elafin: multifunctional antiproteases of the WFDC family. Biochem Soc Trans 39:1437–1440
6. Muratspahić E, Koehbach J, Gruber CW, Craik DJ (2020) Harnessing cyclotides to design and develop novel peptide GPCR ligands. RSC Chem Biol 1:177–191
7. Gracy J, Le-Nguyen D, Gelly J-C, Kaas Q, Heitz A, Chiche L (2008) KNOTTIN: the knottin or inhibitor cystine knot scaffold in 2007. Nucleic Acids Res. https://doi.org/10.1093/nar/gkm939
8. Postic G, Gracy J, Périn C, Chiche L, Gelly J-C (2018) KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. Nucleic Acids Res 46:D454–D458
9. Reinwarth M, Glotzbach B, Tomaszowski M, Fabritz S, Avrutina O, Kolmar H (2013) Oxidative Folding of Peptides with Cystine-Knot Architectures: Kinetic Studies and Optimization of Folding Conditions. ChemBioChem 14:137–146
10. Reinwarth M, Nasu D, Kolmar H, Avrutina O (2012) Chemical Synthesis, Backbone Cyclization and Oxidative Folding of Cystine-knot Peptides — Promising Scaffolds for Applications in Drug Design. Molecules 17:12533–12552
11. Rivera-de-Torre E, Rimbault C, Jenkins TP, Sørensen CV, Damsbo A, Saez NJ, Duhoo Y, Hackney CM, Ellgaard L, Laustsen AH (2022) Strategies for Heterologous Expression, Synthesis, and Purification of Animal Venom Toxins. Front Bioeng Biotechnol 9:811905
12. Correnti CE, Gewe MM, Mehlin C, et al (2018) Screening, large-scale production and structure-based classification of cystine-dense peptides. Nat Struct Mol Biol 25:270–278
13. Crook ZR, Sevilla GP, Friend D, et al (2017) Mammalian display screening of diverse cystine-dense peptides for difficult to drug targets. Nat Commun 8:2244
14. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B (2019) Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20:723
15. Sanderson T, Bileschi ML, Belanger D, Colwell LJ (2021) ProteInfer: deep networks for protein functional inference. 2021.09.20.461077
16. Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589

17. Dempster A, Petitjean F, Webb GI (2020) ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Disc 34:1454–1495
18. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480–D489
19. Ho TK (1995) Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. pp 278–282 vol.1
20. Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. J Mach Learn Res 7:1–30
21. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30:
22. Muttenthaler M, King GF, Adams DJ, Alewood PF (2021) Trends in peptide drug discovery. Nat Rev Drug Discov 20:309–325
23. Isidro-Llobet A, Kenworthy MN, Mukherjee S, Kopach ME, Wegner K, Gallou F, Smith AG, Roschangar F (2019) Sustainability Challenges in Peptide Synthesis and Purification: From R&D to Production. J Org Chem 84:4615–4628
24. Kontermann RE (2011) Strategies for extended serum half-life of protein therapeutics. Curr Opin Biotechnol 22:868–876
25. Ajingi YS, Rukying N, Aroonsri A, Jongruja N (2022) Recombinant Active Peptides and their Therapeutic Functions. Current Pharmaceutical Biotechnology 23:645–663
26. Hon J, Marusiak M, Martinek T, Kunka A, Zendulka J, Bednar D, Damborsky J (2021) SoluProt: prediction of soluble protein expression in Escherichia coli. Bioinformatics 37:23–28
27. Bhandari BK, Gardner PP, Lim CS (2020) Solubility-Weighted Index: fast and accurate prediction of protein solubility. Bioinformatics 36:4691–4698
28. Raimondi D, Orlando G, Fariselli P, Moreau Y (2020) Insight into the protein solubility driving forces with neural attention. PLoS Comput Biol 16:e1007722
29. Martiny H-M, Armenteros JJA, Johansen AR, Salomon J, Nielsen H (2021) Deep protein representations enable recombinant protein expression prediction. Computational Biology and Chemistry 95:107596
30. Hu M, Yuan F, Yang KK, Ju F, Su J, Wang H, Yang F, Ding Q (2022) Exploring evolution-based & -free protein language models as protein function predictors. https://doi.org/10.48550/arXiv.2206.06583
31. McCarthy S, Robinson J, Thalassinos K, Tabor AB (2020) A Chemical Biology Approach to Probing the Folding Pathways of the Inhibitory Cystine Knot (ICK) Peptide ProTx-II. Frontiers in Chemistry 8:228
32. Gamboa JCB (2017) Deep Learning for Time-Series Analysis. https://doi.org/10.48550/arXiv.1701.01887
33. Yang KK, Wu Z, Bedbrook CN, Arnold FH (2018) Learned protein embeddings for machine learning. Bioinformatics 34:2642–2648

Statements & Declarations

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval / Consent to Participate / Consent to Publish

This study does not involve human subjects, vertebrate animals, or human embryonic stem cells.

Author Contributions

Sébastien Ouellet: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Angus Z. Lau:** Investigation, Resources. **Larissa Ferguson:** Conceptualization, Writing - Review & Editing, Visualization. **Tony K.Y. Lim:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization.