

# **ACPLearn: A Deep Neural Network model to predict novel anticancer peptides**

Aditya Kiran Koushik<sup>a</sup>, Katelyn Del Toro<sup>b</sup>, Sandrasegaram Gnanakaran<sup>c</sup>, William Curtis Hines<sup>b</sup>

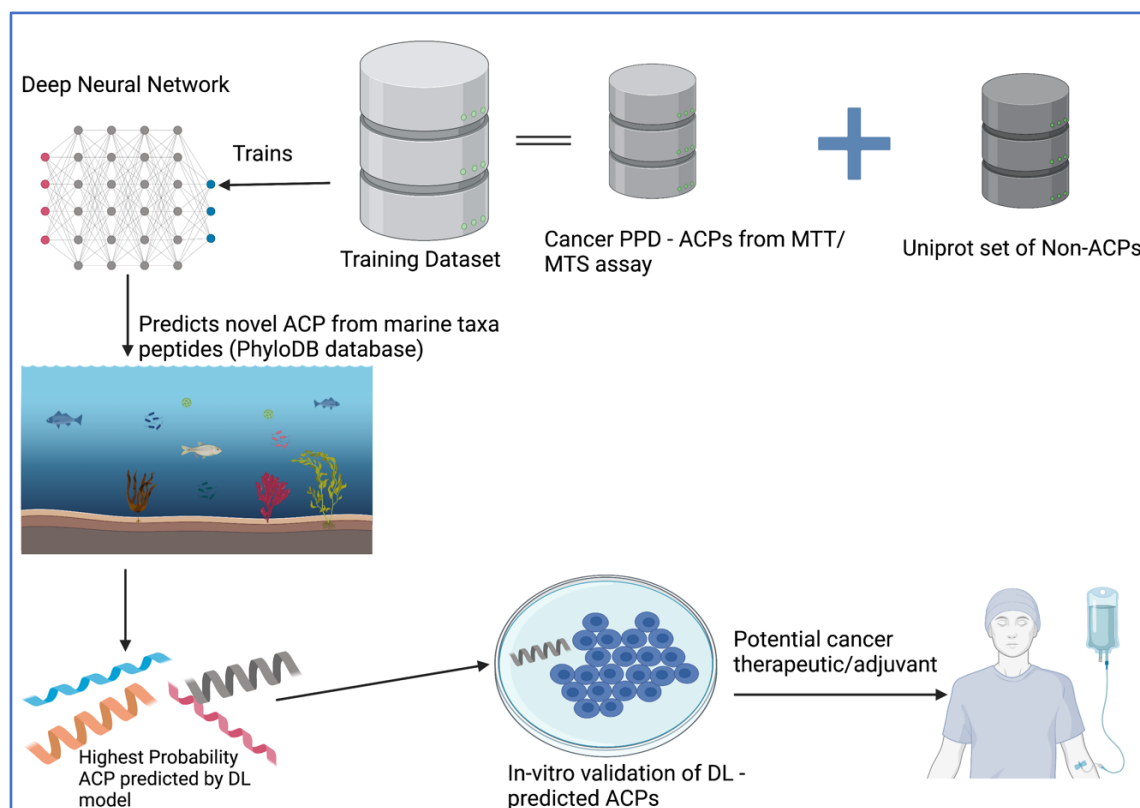
<sup>a</sup>La Cueva High School, Albuquerque, New Mexico, USA.

<sup>b</sup>Department of Biochemistry and Molecular Biology, University of New Mexico School of Medicine, Albuquerque, New Mexico, United States of America.

<sup>c</sup>Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

## Abstract

Anticancer Peptides (ACPs), a class of small peptide molecules, have gained increasing attention in cancer research due to their ability to selectively target and kill cancer cells, sparing normal cells. ACPs, unlike chemotherapy, have less toxicity and fewer side effects, are highly specific to cancer cells, are easy to synthesize and modify, and are cost-effective therapeutics. Unfortunately, the experimental identification of novel ACPs is time-consuming and expensive, therefore computational methods to identify key features of ACPs is promising. Here, a robust deep-learning model was developed that recognized molecular features of ACPs from the CancerPPD database with a sensitivity and specificity of 92% and 99%. Next, 20,000 peptide entries catalogued in the J. Craig Venter Institute PhyloDB database were screened for potentially novel ACPs by the deep learning model, from which the model selected top forty novel ACPs with >90% ACP probability, six of which originated from a species of unicellular algae *P. Antarctica*. MCF-7 Breast Cancer Cell culture validation of the top four novel ACPs exhibit a statically significant cytotoxic effect on cancer cells at concentrations above 10-100  $\mu\text{g/ml}$ . In summary, for the first time, the deep learning approach described here applies learned information on ACPs to make new predictions and would present the first AI (artificial intelligence) predicted novel ACPs to be validated *in vitro*.



**Graphical Abstract of Project**

## Introduction

Cancer is a devastating disease that claims the lives of millions every year. It remains a significant public health concern in the United States, with projections for 2023 estimating nearly two million new cases and over 600,000 deaths.<sup>1</sup> As cancer drug resistance, costliness, and side effects are prevalent in traditional cancer treatments such as radiotherapy and chemotherapy, alternative therapeutic approaches are becoming critically important.<sup>2-4</sup> One such approach is peptide therapeutics, such as Anticancer Peptides (ACPs).<sup>5</sup> ACPs are small peptide molecules, mostly comprised of 10 to 50 amino acids, and are typically derived from antimicrobial peptides.<sup>6</sup> They have received attention due to their selectivity towards cancer cells, ease of synthesis and modification, minimal toxicity toward normal cells, and cost-effectiveness compared to traditional chemotherapy.<sup>7</sup>

Currently, the identification of anticancer peptides in laboratory experiments is both time-consuming and costly, therefore the development of Machine Learning (ML)/Deep Learning (DL) techniques to predict anticancer peptides are becoming critically important.<sup>8</sup> Many ML/DL models exist to accurately identify ACPs, employing various feature extraction and feature engineering methods for the training process. For example, Tyagi et al.,<sup>9</sup> developed a support vector machine (SVM) based model to identify ACPs based on amino acid composition (AAC) and dipeptide composition (DPC) descriptors. Chen et al. additionally presented the iACP prediction tool,<sup>10</sup> which optimized g-gap dipeptide features for higher predictive accuracy.

Many studies have also utilized DL methods for ACP predictions. For example, Yi et al. presented a Long Short-Term Memory (LSTM) neural network called ACP-DL, which encodes ACP sequences using binary feature profile (BFP) and the K-mer sparse metric proposed by You et al.<sup>11,12</sup> In addition, Yu et al. proposed DeepACP, which compares Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures for better predictive performance.<sup>13</sup> Finally, Ahmed et al. proposed ACP-MHCNN, a multiheaded CNN that achieved high performance by extracting sequence, physicochemical, and evolutionary features of ACPs.<sup>14</sup>

Despite the availability of numerous computational models for ACP prediction, to our best knowledge, none have been applied to predict **new anticancer peptides** nor undergone *in vitro validation* of their predictions. Additionally, most ACP prediction models are trained on smaller benchmark datasets with  $n < 1000$  peptides, potentially limiting the applicability of the models. Models that only rely on compositional features during training might exhibit poor accuracy or

limited applicability, given the important role of amino acid positioning on the N-terminus or C-terminus in ACP prediction.<sup>15</sup>

To address these limitations, we present ACPLearn - a deep learning framework designed to predict novel anticancer peptides. We attempt to address current limitations by including a larger training dataset of  $n=1580$ , which is a more suitable sample size for deep learning applications, and by incorporating novel physicochemical peptide descriptors from iFeature, including CKSAAGP (composition of  $k$ -spaced amino acid group pairs) and CTD (composition, transition, distribution features).<sup>16</sup> Finally, we employ the model to predict novel ACPs in a subset of the PhyloDB database of peptides and proteins, which contains data from the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP).

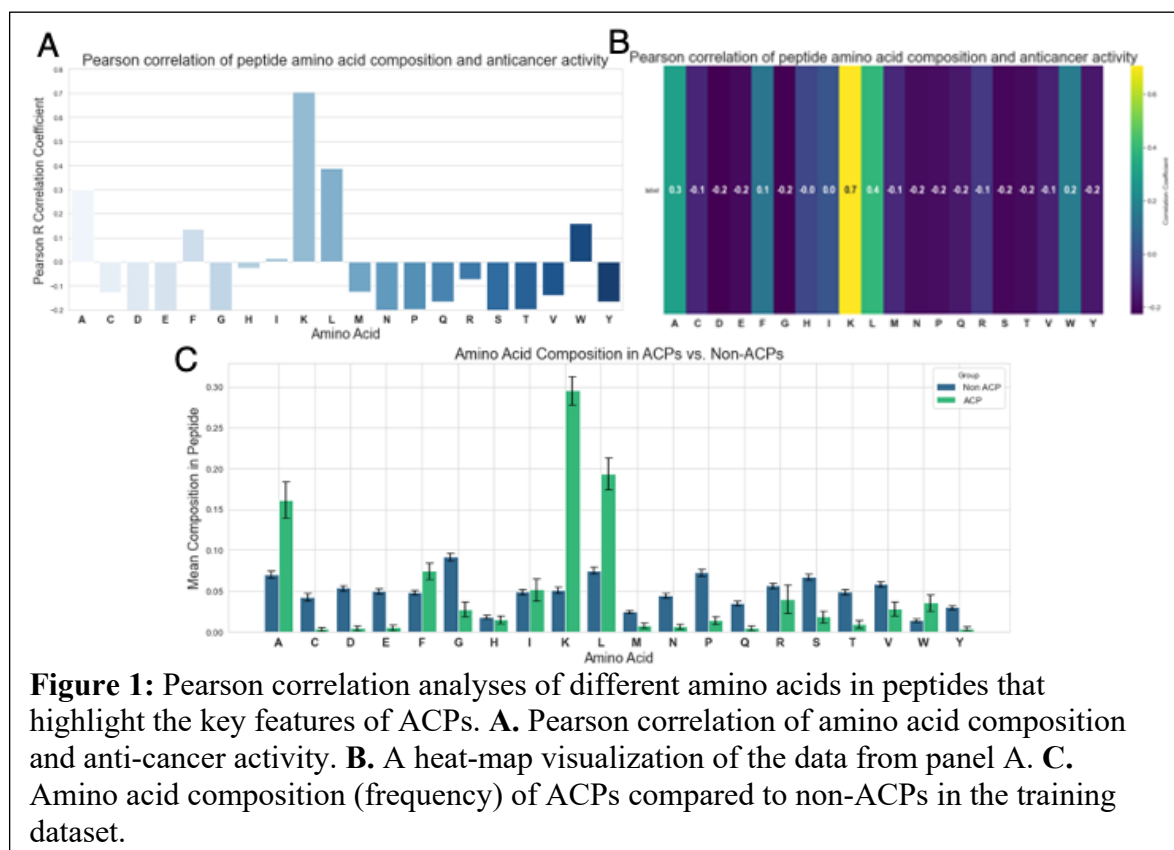
The developed ACPLearn model effectively distinguishes between ACPs and non-ACPs with a sensitivity, specificity, and MCC score of 92%, 99%, and 89% respectively, and outperformed existing models on a benchmark dataset. From the PhyloDB database, the model identified forty novel ACPs, with six originating from the oceanic unicellular algae *P. Antarctica*. This discovery of potential ACPs in a new, oceanic algae highlight the applicability of our model for further discoveries. Finally, we selected the top four ACPs predicted by the model, which were validated *in-vitro* with a cytotoxicity toward MCF-7 breast cancer cells (mean  $IC_{50}$  of 45.4  $\mu\text{g/ml}$ ).

## Results

### *Pearson correlation analyses identified key features and physicochemical properties of ACPs in the training data*

Correlation analysis performed with Pearson Correlation reveals a few key amino acids and features present in ACPs. Amino acids Lysine (K) and Leucine (L) have a strong and moderate positive correlation with ACPs of  $r = 0.7$  and  $r = 0.4$ , respectively (**Figure 1A and 1B**). ACPs with higher lysine composition are likely to be cationic in nature and conform to an alpha-helix, which can allow the peptide to interact with the anionic components of the cancer cell membrane.<sup>7</sup> Two hydrophobic amino acids (Leucine and Alanine) with moderate ACP correlation ( $\sim r = 0.3-0.4$ , **Figure 1A and 1B**) may also play an important role in ACP function for being key part of the alpha-helix, which promotes cell membrane penetrance.<sup>17</sup> Similarly, Phenylalanine (F) with  $r = 0.17$ , may be able to increase affinity for the cancer cell membrane and increase cytotoxicity.<sup>18</sup>

This pattern is similarly observed in **Figure 1C**, where residues Alanine, Phenylalanine, Isoleucine, Lysine, Leucine, and Tryptophan have a higher mean composition in ACPs than non-ACPs, while other residues have a higher composition in non-ACPs. Finally, amino acids other than the ones mentioned above in the AAC descriptor show negative/no correlation with ACPs (**Figure 1A**). However, it is important to consider that the amino acid composition may not be able to effectively capture the full scope of ACPs, and some negatively correlated residues may still play a key role in ACPs based on contextual residues or ACP structure-activity relationship (SAR). For this reason, additional descriptors such as *composition transition and distribution* (CTD) and *composition of k-spaced amino acid group pairs* (CKSAAGP) were considered (**Table 1** and Methods).



Some of the features described in **Table 1** are validated by literature,<sup>7,15,18</sup> such as composition of cationic amino acids or alpha-helix-conforming amino acids. However, some novel properties of ACPs were found through correlation analysis in the current study (**Table 1**). For example, gaps of one amino acid between aliphatic and cationic residues are a strong indicator of

anti-cancer activity ( $r = 0.73$ , **Table 1**). Similarly, gaps of three amino acids between two cationic residues are more frequent in ACPs ( $r = 0.69$ , **Table 1**). In contrast, the hydrophobicity\_ARGP820101.G1 feature shows a negative correlation ( $r = -0.56$ , **Table 1**) indicating that residues with polar side chains are more frequent in non-ACPs. Likewise, non-polar and non-aromatic residues are more frequent in non ACPs ( $r = -0.52$ , **Table 1**). In all, these results indicate that the physicochemical descriptors used to train ACPLearn can extract important properties of ACPs.

Feature	Descriptor	Correlation Coefficient with anti-cancer activity	Explanation
secondarystruct.G1	CTDC	$r = 0.55$	Composition of alpha-helix residues
hydrophobicity_ARGP820101.G1	CTDC	$r = -0.56$	Composition of polar side chain residues
postivecharger.alphaticr.gap1	CKSAAGP	$r = 0.73$	One amino acid gap between cationic and aliphatic residue
hydrophobicity_FASG890101.2.residue100	CTDD	$r = -0.52$	Position on peptide where 100% of non-polar and non-aromatic residues are located before the position
postivecharger.postivecharger.gap3	CKSAAGP	$r = 0.69$	Three amino acid gap between cationic residues

**Table 1:** Few Key Features of Anticancer Peptides found by correlation analyses.

*ACPLearn predicts ACPs with high specificity, sensitivity, precision and MCC score, and outperforms existing models*

**Table 2** describes the model performance for three-fold Cross Validation (CV) on the custom dataset created in this study. The model achieves high performance in discriminating ACP vs non-ACP as all calculated metrics were near or above 90%. It should be noted that the accuracy

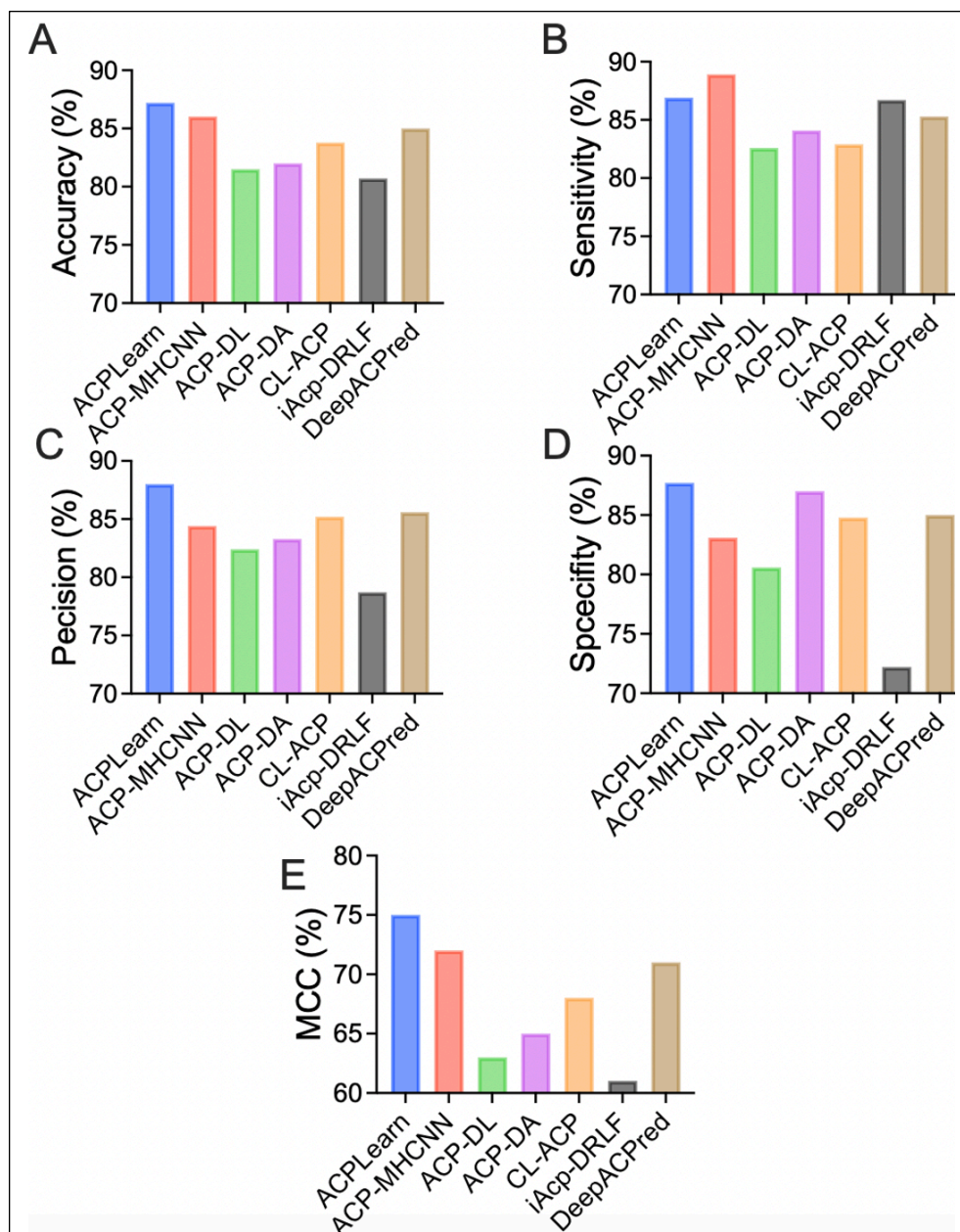
metric was not computed for 3-fold CV as the custom dataset with  $n = 1580$  peptides was imbalanced (148 ACPs and 1432 non-ACPs), so the accuracy metric is not interpretable. Instead, sensitivity, specificity, precision, and Matthew's correlation coefficient (MCC) score were analyzed (see methods for calculations). The model achieved a mean sensitivity of 92.4% with SD of 2.4, 99.1 specificity with SD of 0.5, precision of 88.4 with SD of 3.7, and finally MCC score of 89.2 with SD of 1.3. As detailed in the results, the model performs inference on the PhyloDB data to identify novel ACPs with high probability scores for each of the 3 folds.

	Sensitivity (%)	Specificity (%)	Precision (%)	MCC (%)
Fold 1	92.9	99.0	86.9	89.1
Fold 2	89.8	99.6	92.7	90.5
Fold 3	94.4	98.6	85.8	88.0
Avg $\pm$ STD	92.4 $\pm$ 2.4	99.1 $\pm$ 0.5	88.4 $\pm$ 3.7	89.2 $\pm$ 1.3

**Table 2:** Three-Fold CV Model performance on custom dataset

For more robust evaluation of the model and for benchmarking/comparison purposes, we also evaluated the model on the ACP-740 dataset (see methods) with a five-fold CV. As shown in **Table 3**, the performance of ACPLearn was compared with six other existing ACP prediction models. Out of the 6 models for comparison, ACP-MHCNN,<sup>14</sup> ACP-DL,<sup>11</sup> CL-ACP,<sup>19</sup> and DeepACPred<sup>20</sup> are neural network or deep learning-based methods, while ACP-DA<sup>21</sup> and iACP-DRLF<sup>22</sup> are ML based methods (although iACP-DRLF uses deep representation learning features). ACPLearn **outperforms** all six models in the accuracy, specificity, precision, and MCC score metrics while it outperforms all models in the sensitivity metric except for ACP-MHCNN. The improvement achieved by ACPLearn compared to the other methods range from 1.2% - 6.5%, 0.2% - 4.3%, 0.7% - 15.5%, 2.4% - 9.3%, 3% - 14% for accuracy, sensitivity, specificity, precision, and MCC respectively. These differences can be observed visually in **Figure 2**, where the most significant differences can be observed with the Precision and MCC metrics. This indicates that the model has a low false-positive rate, which might be attributed to the extensive features and descriptors used to train the model including CTD and CKSAAGP, which provide valuable information on distribution of key amino acids, a factor not considered by existing models. These features can also explain the high specificity of 99% observed in **Table 1**, indicating the model is strongly able to distinguish non-ACPs. Overall, these results indicate that ACPLearn has a strong capacity to identify ACPs, making it a valuable tool for the discovery of novel ACPs. Additionally,

compared to other ACP prediction methods, ACPLearn has a better interpretability and simplicity, as the architecture is a simple Deep Neural network and only used five descriptors (CTDC, CTDD, CKSAAGP, *amino acid composition* or AAC, *grouped amino acid composition* or GAAC) were used to encode peptides, and are relatively explainable (methods).



**Figure 2:** Performance Comparison for ACP740 dataset across ML and DL ACP predictors (5-Fold CV). A-E. Shows accuracy, sensitivity, precision, specificity and Mathew correlation coefficient (MCC) percentages of different ACP prediction models.



Model	Acc (%)	Sens (%)	Spec (%)	Prec (%)	MCC (%)
<b>ACPLearn (Present Study)</b>	<b>87.2 <math>\pm</math> 2.29</b>	<b>86.9 <math>\pm</math> 5.29</b>	<b>87.7 <math>\pm</math> 5.67</b>	<b>88.0 <math>\pm</math> 4.27</b>	<b>0.75 <math>\pm</math> 4.69</b>
ACP-MHCNN	86.0	88.9	83.1	84.4	0.72
ACP-DL	81.5	82.6	80.6	82.4	0.63
ACP-DA	82.0	84.1	87.0	83.3	0.65
CL-ACP	83.8	82.9	84.8	85.2	0.68
iAcp-DRLF	80.7	86.7	72.2	78.7	0.61
DeepACPred	85.0	85.3	85.0	85.6	0.71

**Table 3:** Performance Comparison for ACP740 dataset across ML and DL ACP predictors (5-Fold CV)

### *ACPLearn discovers 4 novel ACPs*

As mentioned before, ACPLearn was trained on three-folds. For each of the three-folds, the DL model would scan through ~20,000 peptides in PhyloDB to find potentially novel ACP candidates with a predicted probability over 90% (see methods). After three-folds, the model displayed forty total ACP candidates with a probability over 90%. However, for budget restrictions, only four of these peptides could be synthesized for *in vitro* testing. Out of the original four peptides, three peptide fragments were embedded within a few of the original peptides that have a high BLAST similarity (computed with CancerPPD's BLAST program) with existing ACPs, namely KWKKIKFLIVYY, FKHIKKWFI, FLAKKALIHLE, with scores of 26, 17, and 20 respectively. Two of these embedded peptides and two original full-length peptides were selected out of the forty total ACPs predicted by the model for *in-vitro* validation.

We selected fragments KWKKIKFLIVYY and FLAKKALIHLE for having the highest BLAST score of 26 and 20 respectively as two peptides for *in-vitro* validation. To pick the two full peptides, BLAST similarity and net charge were analyzed. Out of 40 peptides, Sequence 3 in **Table 4** had the highest BLAST score of 41, highest net charge of 10.99, and predicted alpha-helical structure which are key properties of ACPs<sup>15,18</sup> (**Table 5**). The second peptide was Sequence 4 due to having the second highest net charge and confirming to helical structure (**Tables 4-5**).

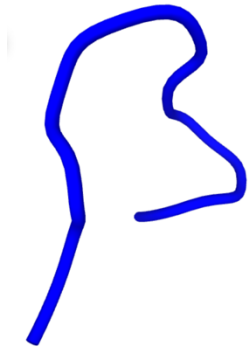
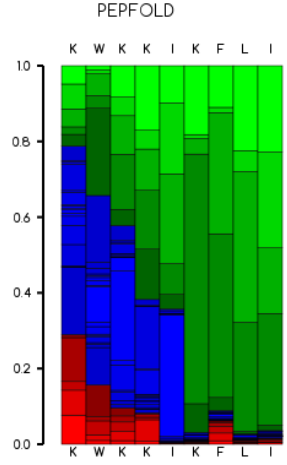

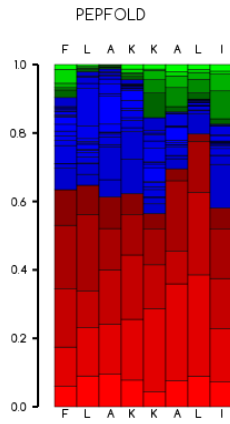
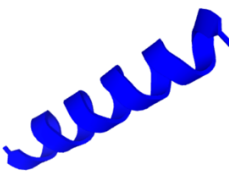
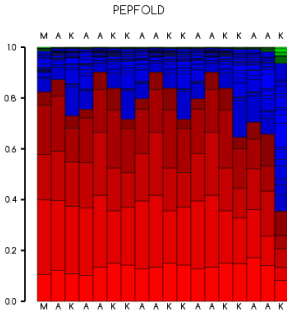
**Table 4** shows information on the four selected peptides as mentioned above. The PEP-Fold 3.5 tool was used to create secondary structure predictions for the peptides as shown in **Table 5**.<sup>23-25</sup> Three of the predicted peptides originate from bacterial species, while sequence four

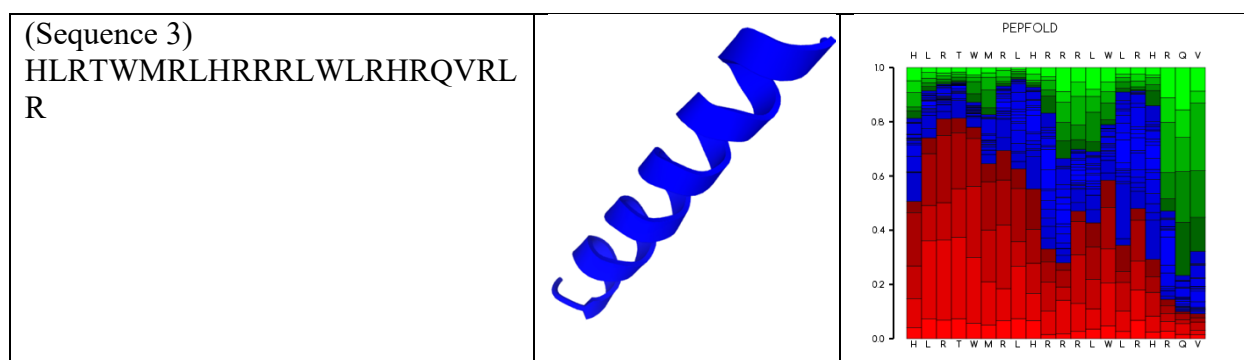
originates from a unicellular algae *Phaeocystis antarctica*, which is found in the Southern Ocean. This algal species also has a unique role in regulating global carbon and sulfur cycles.<sup>26</sup> We hope that further research can be made into the potential anticancer properties of this unicellular algae, as this study provided a first insight.

Peptide Sequence predicted by ACPLearn from PhyloDB (>90% probability)	Net Charge	Average Hydropathy	Species of Origin
(Sequence 4) KWKKIKFLIVYY	+4.00	-0.45 (hydrophobic)	<i>Mycoplasma fermentans</i> (bacteria)
(Sequence 1) FLAKKALIHLE	+2.09	-0.70 (hydrophobic)	<i>Acinetobacter baumannii</i> (bacteria)
(Sequence 2): MAKAAKKAACKAAK KAAKKKK	+10.99	1.3 (hydrophilic)	<i>Hyphomicrobium</i> MC1 (bacteria)
(Sequence 3) HLRTWMRLHRRRLWLRHRQVRLR	+9.27	0.29 (hydrophilic)	<i>Phaeocystis antarctica</i> (algae)

**Table 4:** Information of four novel ACPs predicted by ACPLearn

**Table 5** shows the PEP-Fold secondary structure prediction of the four ACPs candidates. As mentioned in the Introduction, most known Anticancer Peptides are alpha-helical since their amino acids are mostly cationic.<sup>17,18</sup> Three of the four ACP candidates exhibit an alpha-helical structure in the prediction, with Sequence 4 being the exception and conforming to a random coil structure, as denoted by the high probability green columns. Although Sequence 4 may not conform to an alpha-helix, it still is highly cationic and moderately hydrophobic which is why it may have been picked up by the model. Sequence 4 also has a high composition of Lysine residues, which was an important descriptor for ACPs in the training set (**Figure 1**). Lysine is also predominant in peptides with anti-cancer activity.<sup>27</sup> Sequence 2 and 3 show stronger helix probabilities above 50% for nearly all residues of the sequence. Although the predicted structure of Sequence 4 shows a helical shape, many residues of the sequence show below a 50% probability and ending residues may conform to a random coil.

Peptide Sequence	Structure	Probabilities (Red – Helix, Blue – Sheet, Green - Coil)
(Sequence 4) KWKKIKFLIVYY		
(Sequence 1) FLAKKALIHFL		
(Sequence 2): MAKAAKKAACKAA KKAACKKK		



**Table 5:** Secondary Structure Predictions of top four ACPs predicted by ACPLearn

The four peptides predicted from the model were synthesized and tested against MCF-7 breast cancer cell line (see Methods). Testing these peptides *in vitro* with the MTT assay reveals that the peptides can exert a significant cytotoxicity (visually lower cell density – **Figure 3**; lower absorbance – **Figure 4**) towards MCF-7 breast cancer cells compared to control.

#### ***Dose-dependent cytotoxic effect of four ACPs predicted by ACPLearn on MCF-7 cancer cells***

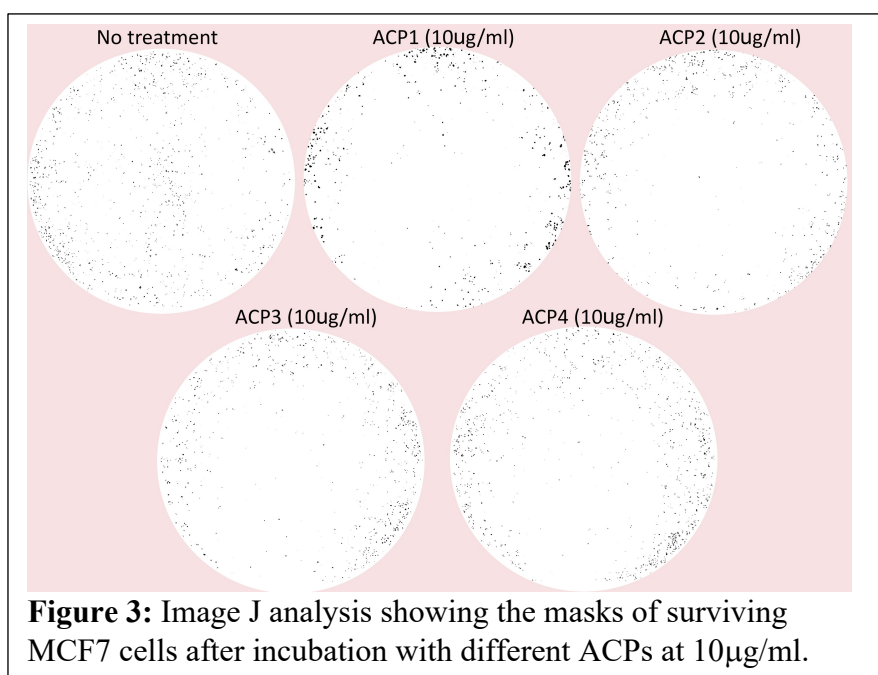
As shown in **Figure 4A**, ACP #1 had the lowest absorbance (mean  $\pm$  SEM of  $0.4 \pm 0.018$ ) at 0.1  $\mu\text{g/ml}$ . ACP #3 was the most cytotoxic at 1  $\mu\text{g/ml}$  (mean  $\pm$  SEM of  $0.30 \pm 0.009$ ), and ACP #2 was the most cytotoxic at higher doses of 100 and 1000  $\mu\text{g/ml}$  of  $0.18 \pm 0.00088$  and  $0.089 \pm 0.0003$ , respectively. However, it should be noted that all 4 peptides share a similar overall cytotoxicity at concentrations above 1 mg/ml and all deliver a lethal dose at 1 mg/ml (1000  $\mu\text{g/ml}$ ). The inferred  $\text{IC}_{50}$  values for ACP 1, 2, 3, and 4 are 63.7, 27.1, 39.8, and 50.9 mg/ml respectively, with a lower value indicating higher cytotoxicity.

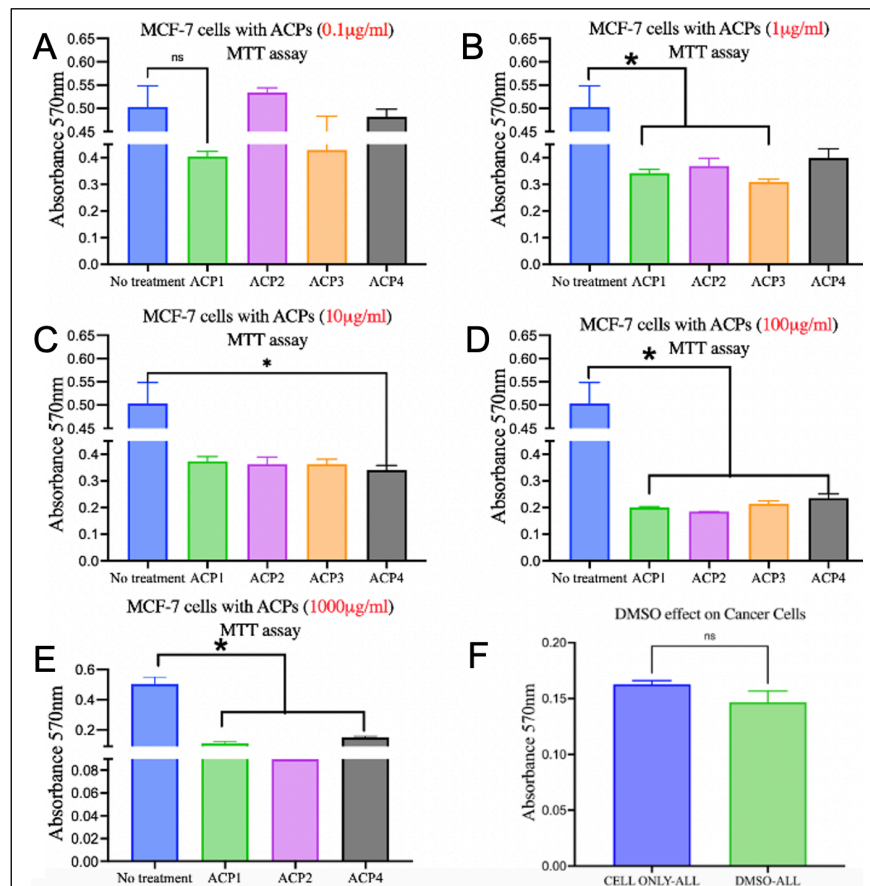
From the data, the peptide fragments (Sequence 1 and 4) showed the higher  $\text{IC}_{50}$  value compared to the full peptides (Sequence 2 and 3) predicted by the model, suggesting that the full predicted peptides showed a higher cytotoxicity. However, this difference could also be attributed to the fact that the peptide fragments were hydrophobic and moderately cationic while the full sequences were highly cationic and hydrophilic. Because this is a preliminary assay, it would require further experimentation to explain the difference in  $\text{IC}_{50}$  values and overall cytotoxicity.

**Figure 4B – 4F** compares absorbance of four ACPs with the no treatment/control group at four concentrations. At 0.1  $\mu\text{g/ml}$ , none of the four ACPs showed a significant difference ( $p < 0.05$ ) from the control group (**Figure 4B**). At 1  $\mu\text{g/ml}$ , ACPs 1, 2, and 3 showed a significant difference ( $p < 0.01$ ) from the control group. At 10  $\mu\text{g/ml}$ , only ACP4 showed a significant difference and all

ACPs showed a significant difference ( $p < 0.001$ ) at 100  $\mu\text{g/ml}$  and 1000  $\mu\text{g/ml}$  (**Figure 4B**). It should be noted that in **Figure 4F**, the ACP3 group is missing. This is because of experimental error, and unfortunately no valid data was produced for ACP3 at 1000  $\mu\text{g/ml}$ , and therefore the cytotoxicity of ACP3 at 1000  $\mu\text{g/ml}$  cannot be analyzed. Finally, to rule out the effect of DMSO itself, cells were treated with DMSO at same volumes (1ml, 2.5ml, 10ml, and 25ml) as those with ACPs. A  $t$ -test between absorbance of all untreated cells ( $n=11$  in total) and DMSO wells at different concentrations ( $n=8$ ) did not show a significant difference in MTT assay ( $p=0.08$ ). As another step to rule out DMSO's effect, comparison between untreated cells ( $n=11$ ) and highest volume (10ml and 25ml DMSO) treated cells also did not show significant difference in MTT absorbance ( $p=0.1407$ ). Together, these results suggest that the DMSO effects are minimal, if any, in cancer cell toxicity.

In conclusion, all ACPs, show a significant decrease and therefore high overall cytotoxicity when compared to absorbance of untreated cells (shown in black), suggesting that ACPs predicted by our ACPLearn Deep Neural Network framework could show expected anti-cancer effects. The main purpose of the *in vitro* experiment was to validate the model's predictive ability, but this would require a far greater sample of predicted ACP candidates for testing, and more robust experimentation. Instead, this preliminary experiment serves to provide a first insight into how computational models for ACP prediction could be validated *in vitro* or *in vivo*.





**Figure 4:** Performance Comparison for ACP740 dataset across ML and DL ACP predictors (5-Fold CV). **A-F** shows the comparisons of different ACPs at increasing concentrations on MCF-7 survival based on MTT cell survival assay. Unpaired *t* test; mean  $\pm$  sem (n=3; \*p<0.05; ns=not significant).

## Discussion and Conclusion

In this study, I developed the ACPLearn Deep Neural Network framework for predicting novel ACPs from marine taxa. Through rigorous testing on benchmark datasets, the model consistently outperformed several other ML and DL prediction models in Accuracy, Sensitivity, Specificity, Precision, and MCC metrics. Intriguingly, among the model's predictions in PhyloDB, peptides from *P. Antarctica* and other bacterial species emerged as potential ACP candidates. Specifically, the peptide derived from *P. Antarctica* presents as a particularly promising result, suggesting the need for more in-depth exploration of this species for potential anticancer properties. To recap, of the forty ACP candidates predicted by the model from a substantial set of ~20,000 peptides from PhyloDB, four were chosen for *in vitro* validation. The transition of these computational predictions to the laboratory revealed significant cytotoxicity of the four predicted peptides towards MCF-7 breast cancer cells, demonstrating the efficacy of the model. However, the model's limitations became evident.

A deeper understanding of how these peptides induce cell death requires functional assays. Predictive validation cannot rest on a handful of peptides, and without controlling variables such as peptide length or hydrophobicity, it remains speculative to ascertain factors leading to decreased cytotoxicity. Despite these challenges, this research lays a foundation by identifying these potential variables. Due to inherent complexities such as peptide shape, fragment/full peptide distinction, peptide hydropathy, BLAST score, and charge, direct comparisons between peptides become challenging, so instead we benchmarked against the control/untreated group. While our main goal was to validate the model, the presence of uncontrolled variables highlights the complexities in selecting peptides using these criteria. A significant limitation to acknowledge is the interpretability challenge posed by deep neural network models. Even though ACPLearn may be simpler compared to models like ACP-MHCNN, ACP-DL, DeepACP, etc. in its architecture and functions, its predictions cannot be directly interpreted nor justified. Future endeavors may lean towards interpretable models like SHAP or LIME to understand the decision-making processes of such networks. As the fusion of bioinformatics and experimental biology gains traction, tools like ACPLearn stand at the frontier of an accelerated and cost-effective drug discovery trajectory in oncology, especially with its discovery of potential anticancer properties of *P. Antarctica*.

## Methods

### Data Collection

In total, there were three separate datasets used in this study: A custom created training dataset to train the model, the existing ACP-740 dataset to benchmark the model with existing ones, and finally the PhyloDB database to apply the model and find new ACPs.

A dataset of both ACPs and non-ACPs was compiled to train the deep neural network (DNN) ACPLearn model. ACPs were compiled from the CancerPPD database,<sup>28</sup> a comprehensive resource of anticancer peptides and proteins derived from various natural sources, including plants, animals, and microorganisms. ACPs in CancerPPD were filtered for only ACPs validated through the MTT assay, and using CD-HIT webserver<sup>29</sup>, peptides with above 90% similarity were removed. The filtering resulted in a total of 148 ACPs out of 624 peptides in CancerPPD.

The dataset of non-ACPs used in this study was adopted from Li et al's work on AMP prediction<sup>30</sup> by querying SwissProt protein database.<sup>31</sup> The training dataset of 1432 non-ACPs is found in this GitHub link: <https://github.com/bcgsc/AMPlify>. The ACP and non-ACP datasets were compiled to creating a final custom training dataset with n = 1580 peptides in total.

The ACP-740 dataset was used for benchmarking the model. The ACP-740 dataset was introduced in a previous study.<sup>11</sup> Initially, a total of 388 experimentally validated anticancer peptides (ACPs) were curated, with 138 ACPs sourced from 12 and 250 ACPs from 30. Furthermore, 456 antimicrobial peptides (AMPs) lacking anticancer activity were collected as negative samples, comprising of 206 AMPs from 12 and 250 AMPs from 30. To ensure diversity within the dataset and avoid redundancy, a CD-HIT approach was applied, removing 12 similar positive samples and 92 similar negative samples. This process was in accordance with the methodology followed in previous studies.<sup>11</sup> Consequently, the final version of the ACP-740 dataset was comprised of 740 samples, consisting of 376 positive samples (ACP) and 364 negative samples (AMP).

To predict novel ACPs, a subset of peptides from PhyloDB was compiled. PhyloDB Peptide Database from Marine Animals is curated by J. Craig Venter Institute and is a specialized resource for the identification and characterization of peptides derived from marine organisms. This database (version 1.075) consists of 24,509,327 peptides from 19,962 viral, 230 archaeal, 4910 bacterial, and 894 eukaryotic taxa. The PhyloDB dataset for prediction was compiled by



downloading the original dataset of 24 million peptides and filtering the database for peptides between 7 and 25 amino acids in length (7 was chosen since it is a minimum requirement for some iFeature analysis tools and 25 was the maximum peptide length of ACPs in the training set). The filter resulted in a final prediction dataset of 19,729 peptides.

### Feature Engineering via iFeature

To create numerical representations of peptides that can be used in the training process, the iFeature library was utilized. The following will describe some key physicochemical descriptors that were used in this study.

**AAC (amino acid composition):** The composition of the 20 standard amino acids inside of an ACP are calculated as a percentage.

$$f(t) = \frac{N(t)}{N}, t \in \{A, D, D, \dots, Y\} \quad (1)$$

**Equation 1:** AAC features are created by calculating the percentage composition of each amino acid in total peptide.

**CTDC (Composition, transition, distribution, composition):** Similar to AAC, but the composition of amino acids is under a certain category (composition of polar amino acids, alpha helical amino acids, positively charged, hydrophobic, etc.).

$$C(r) = \frac{N(r)}{N}, r \in [polar, neutral, hydrophobic] \quad (2)$$

**Equation 2:** CTDC features are created by calculating percentages of polar, neutral, and hydrophobic, and other physicochemical properties in the peptide.

**CTDD (Composition, transition, distribution, distribution):** Describes not only composition, but distribution of certain amino acids throughout peptide. Describes the corresponding fraction of groups in the entire sequence, where the first residue of a given group is located, and where 25, 50, 75 and 100% of the occurrences are located.

**CKSAAGP (Composition of k-spaced amino acid group pairs):** This feature calculates the frequency of amino acid group pairs separated by any k-residues.

$$\left( \frac{N_{g1g1}}{N_{total}}, \frac{N_{g1g2}}{N_{total}}, \frac{N_{g1g3}}{N_{total}}, \dots, \frac{N_{g5g5}}{N_{total}} \right)_{25} \quad (3)$$

**Equation 3:** g-subscript represent grouped pairs of amino acid depending on the gap value, which is taken as a percentage of the total.

**GAAC (Grouped amino acid composition):** Consists of 5 features represented by g1 (Aliphatic amino acids: GAVLMI), g2 (Aromatic amino acids: FYW), g3 (Positively charged: KRH), g4 (Negatively charged: DE), and g5 (Uncharged: STCPNQ). The features are given by the number of amino acids from a given group divided by the total number of amino acids in the peptide (Equation 4).

$$f(g) = \frac{N(g)}{N}, g \in \{g1, g2, g3, g4, g5\} \quad (4)$$

$$N((g_t)) = \sum N(t), \quad t \in g$$

**Equation 4:** GAAC features are created by calculating the percentage composition of each group of amino acids in total peptide.

AAC, GAAC, and CTDC descriptors provide strong compositional features that allow the model to recognize key amino acid/amino acid groups found in ACPs. CTDD and CKSAAGP features provide important information about the distribution and location of amino acids throughout the peptide. After featurization was complete, the training data had 384 features, with n=384 neurons for neural network input.

### ACPLearn Architecture

To train ACPLearn, the input features of AAC, GAAC, CKSAAGP, CTDD, and CTDC are fed into the neural network with a dimension of n=384. The data was scaled using the scikit-learn standard scaler and normalized.

The Deep Neural network consisted of 3 hidden layers and 3 dropout layers - the first two dropout layers were initialized with dropout rate greater than 0.5 to prevent overfitting, and the final dense layer used a sigmoid activation function to condense the raw neuron output as an ACP probability between 0 and 1 as shown by Equation 5.

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}. \quad (5)$$

**Equation 5:** Sigmoid Activation Function used in DNN.

The model was compiled with binary cross entropy loss function with the ADAM optimizer. Binary cross entropy, or log loss, compares predicted probability with the true class

label, in this case either a 1 (ACP) or 0 (non-ACP). The probability is then penalized based on the distance from the true label.

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (6)$$

Equation 6 describes the loss function where  $y$  is the true label and  $p(y)$  is the predicted probability of a given peptide being an ACP.

### Performance Metrics and Training

The model was evaluated on Accuracy, Sensitivity, Specificity, Precision, and finally MCC score.

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (7)$$

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Equation 7: Sensitivity, Specificity, Precision, and MCC calculations

Sensitivity measures the ability of the model to predict ACPs. It is arguably the most important since the model must be able to predict novel ACPs in the PhyloDB database. Specificity measures the ability of the model to predict non-ACPs or be able to distinguish ACPs from non-ACPs. Precision refers to the accuracy rate of the model to predict the positive class, in this case ACPs.

Finally, MCC measures the correlation between predicted and actual binary classifications, accounting for true positives, true negatives, false positives, and false negatives. It ranges from -1 (completely incorrect) to 1 (perfectly correct), with 0 indicating no correlation.

The above metrics were calculated for each fold after the model was trained with a 3-fold cross validation (CV). For each fold, the model is trained on two folds, evaluated on the final fold, and then makes predictions on the PhyloDB set of 19,729 peptides. The PhyloDB set underwent

the same processes as the training data, including featurization with five descriptors, scaling, and normalization. After all peptides in PhyloDB are given a predicted probability, peptides with a probability score greater than 0.9 in all three folds were selected for further analysis. Because predictions are made for each fold, a lower fold number of 3 was chosen so the DNN would have more instances of ACP classes per fold to learn from. In addition, 5-fold-CV, a more conventional evaluation method, was utilized on the benchmark dataset.

## Implementation

The model was created using TensorFlow with Keras backend in Jupyter Notebook and Python. The iFeature program (<https://github.com/Superzchen/iFeature>) was used for feature extraction from the training dataset and performed with the Mac OS terminal.

## MTT Assay

Forty ACPs were predicted in total with a probability score above 90%, and from this set of peptides we picked the top four for cytotoxicity analysis. The anti-cancer activity of the top four peptides were tested on MCF-7 human breast cancer cell line. This choice was made as MCF-7 is a common experimental cell line used to test the activity of ACPs (alongside HeLA), with 237 entries in CancerPPD.<sup>28</sup>

First, all four peptides were custom synthesized from Genscript® at 4 mg each, with a purity above 90%. Next, the MCF-7 cells (HTB-22 line form ATCC®) was maintained in DMEM with 10% fetal bovine serum at 37°C in an atmosphere of 5% CO<sub>2</sub> and 95% air until the treatment. Cells were grown in 96-well format (at ~5,000 cells/well) as recommended by Roche for MTT cell toxicity assay. Peptides were diluted in DMSO as concentrated stocks (4mg/ml) and serially diluted prior to treatment. Most of the previously reported ACPs have had anti-cancer toxic activity at a IC<sub>50</sub> of 5-100mg/ml for synthetic peptides similar in length to the one predicted above. Based on this, the MCF-7 cells were treated with predicted ACPs at four concentrations (0.1µg/ml, 1µg/ml, 10µg/ml, 100µg/ml; 1mg/ml) with 3 replicates/trials for 72 h. After 3 days of peptide treatment, 10ml of MTT labeling reagent was added to each well and incubated for 4h at 37°C. Next 100ml of solubilization buffer (from the MTT kit) was added into each well and incubated overnight at 37°C. The plate was read in an ELISA plate reader to measure absorbance of the formazan product between 570nm. Reference wavelength was >650nm (per the manufactures’

instructions – Roche's MTT assay kit). The colorimetric absorbance data on cell viability was analyzed and interpreted.

The MTT assay was used to measure cellular metabolic activity as an indicator of cell viability, proliferation and cytotoxicity. This assay is based on the reduction of a yellow tetrazolium salt (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide or MTT) to purple formazan crystals by metabolically active cells. The insoluble formazan crystals are dissolved using a solubilization solution and the resulting-colored solution is quantified by measuring absorbance at 570 nanometers using a multi-well spectrophotometer. The higher the absorbance, the greater the number of viable, metabolically active cells (shown by absorbance in Figure 20). Together, it suggests that a lower absorbance indicates less metabolically active cells in the well and therefore more cancer cells being killed by the ACPs.

## References

- 1 Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J Clin* **73**, 17-48, doi:10.3322/caac.21763 (2023).
- 2 Chakraborty, S. & Rahman, T. The difficulties in cancer treatment. *Ecancermedicalscience* **6**, ed16, doi:10.3332/ecancer.2012.ed16 (2012).
- 3 Hauner, K., Maisch, P. & Retz, M. [Side effects of chemotherapy]. *Urologe A* **56**, 472-479, doi:10.1007/s00120-017-0338-z (2017).
- 4 Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nat Rev Cancer* **13**, 714-726, doi:10.1038/nrc3599 (2013).
- 5 Craik, D. J., Fairlie, D. P., Liras, S. & Price, D. The future of peptide-based drugs. *Chem Biol Drug Des* **81**, 136-147, doi:10.1111/cbdd.12055 (2013).
- 6 Mader, J. S. & Hoskin, D. W. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. *Expert Opin Investig Drugs* **15**, 933-946, doi:10.1517/13543784.15.8.933 (2006).
- 7 Chiangjong, W., Chutipongtanate, S. & Hongeng, S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application (Review). *Int J Oncol* **57**, 678-696, doi:10.3892/ijo.2020.5099 (2020).
- 8 Hu, Y. *et al.* Application of Machine Learning Approaches for the Design and Study of Anticancer Drugs. *Curr Drug Targets* **20**, 488-500, doi:10.2174/1389450119666180809122244 (2019).
- 9 Tyagi, A. *et al.* In silico models for designing and discovering novel anticancer peptides. *Sci Rep* **3**, 2984, doi:10.1038/srep02984 (2013).
- 10 Chen, W., Ding, H., Feng, P., Lin, H. & Chou, K. C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* **7**, 16895-16909, doi:10.18632/oncotarget.7815 (2016).
- 11 Yi, H. C. *et al.* ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. *Mol Ther Nucleic Acids* **17**, 1-9, doi:10.1016/j.omtn.2019.04.025 (2019).
- 12 Zhu-Hong, Y., MengChu, Z., Xin, L. & Shuai, L. Highly Efficient Framework for Predicting Interactions Between Proteins. *IEEE Trans Cybern* **47**, 731-743, doi:10.1109/TCYB.2016.2524994 (2017).
- 13 Yu, L., Jing, R., Liu, F., Luo, J. & Li, Y. DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm. *Mol Ther Nucleic Acids* **22**, 862-870, doi:10.1016/j.omtn.2020.10.005 (2020).
- 14 Ahmed, S. *et al.* ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci Rep* **11**, 23676, doi:10.1038/s41598-021-02703-3 (2021).
- 15 Huang, K. Y. *et al.* Identification of subtypes of anticancer peptides based on sequential features and physicochemical properties. *Sci Rep* **11**, 13594, doi:10.1038/s41598-021-93124-9 (2021).

- 16 Chen, Z. *et al.* iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res* **50**, W434-W447, doi:10.1093/nar/gkac351 (2022).
- 17 Dai, Y. *et al.* Pro-apoptotic cationic host defense peptides rich in lysine or arginine to reverse drug resistance by disrupting tumor cell membrane. *Amino Acids* **49**, 1601-1610, doi:10.1007/s00726-017-2453-y (2017).
- 18 Dennison, S. R., Whittaker, M., Harris, F. & Phoenix, D. A. Anticancer alpha-helical peptides and structure/function relationships underpinning their interactions with tumour cell membranes. *Curr Protein Pept Sci* **7**, 487-499, doi:10.2174/138920306779025611 (2006).
- 19 Wang, H., Zhao, J., Zhao, H., Li, H. & Wang, J. CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model. *BMC Bioinformatics* **22**, 512, doi:10.1186/s12859-021-04433-9 (2021).
- 20 Lane, N. a. K., I. DeepACPpred: A Novel Hybrid CNN-RNN Architecture for Predicting Anti-Cancer Peptides. *Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020)* (2020).
- 21 Chen, X. G., Zhang, W., Yang, X., Li, C. & Chen, H. ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation. *Front Genet* **12**, 698477, doi:10.3389/fgene.2021.698477 (2021).
- 22 Lv, Z. B., Cui, F. F., Zou, Q., Zhang, L. C. & Xu, L. Anticancer peptides prediction with deep representation learning features. *Briefings in Bioinformatics* **22**, doi:ARTN bbab008 10.1093/bib/bbab008 (2021).
- 23 Lamiable, A. *et al.* PEP-FOLD3: faster structure prediction for linear peptides in solution and in complex. *Nucleic Acids Research* **44**, W449-W454, doi:10.1093/nar/gkw329 (2016).
- 24 Shen, Y. M., Maupetit, J., Derreumaux, P. & Tufféry, P. Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction. *J Chem Theory Comput* **10**, 4745-4758, doi:10.1021/ct500592m (2014).
- 25 Thévenet, P. *et al.* PEP-FOLD: an updated structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Research* **40**, W288-W293, doi:10.1093/nar/gks419 (2012).
- 26 Bertrand, E. M. *et al.* Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. (Raw data at <https://allenlab.ucsd.edu/data/>). *Proc Natl Acad Sci U S A* **112**, 9938-9943, doi:10.1073/pnas.1501615112 (2015).
- 27 Shoombuatong, W., Schaduengrat, N. & Nantasenamat, C. Unraveling the Bioactivity of Anticancer Peptides as Deduced from Machine Learning. *Excli J* **17**, 734-752, doi:10.17179/excli2018-1447 (2018).
- 28 Tyagi, A. *et al.* CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* **43**, D837-843, doi:10.1093/nar/gku892 (2015).
- 29 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).
- 30 Li, C. *et al.* AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* **23**, 77, doi:10.1186/s12864-022-08310-4 (2022).

- 31 UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531, doi:10.1093/nar/gkac1052 (2023).