UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Químicas e Ingeniería

Título: Theoretical Screening of Therapeutic Peptides with Potential Anticancer Activity

Trabajo de integración curricular presentado como requisito para la obtención del título de Química

Autora

Romero Herdoiza Maylin Fernanda

Tutora

Ph.D Rodríguez Hortensia

Co-tutor

Ph.D Marrero-Ponce Yovani

Urcuquí, diciembre 2021

UNIVERSIDAD
YACHAY
TECH

Urcuquí, 16 de diciembre de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS QUÍMICAS E INGENIERÍA
CARRERA DE QUÍMICA
ACTA DE DEFENSA No. UITEY-CHE-2021-00032-AD

A los 16 días del mes de diciembre de 2021, a las 15:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.
Miembro No Tutor	Dr. FERREIRA DE MENEZES AREIAS , FILIPE MIGUEL , Ph.D.
Tutor	Dra. RODRIGUEZ CABRERA, HORTENSIA MARIA , Ph.D.

El(la) señor(ita) estudiante **ROMERO HERDOIZA, MAYLIN FERNANDA**, con cédula de identidad No. **0704627900**, de la **ESCUELA DE CIENCIAS QUÍMICAS E INGENIERÍA**, de la Carrera de **QUÍMICA**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-39-No.456-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **THEORETICAL SCREENING OF THERAPEUTIC PEPTIDES WITH POTENTIAL ANTICANCER ACTIVITY**, previa a la obtención del título de **QUÍMICO/A**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dra. RODRIGUEZ CABRERA, HORTENSIA MARIA , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Miembro Tribunal De Defensa	Dr. FERREIRA DE MENEZES AREIAS , FILIPE MIGUEL , Ph.D.	10,0
Presidente Tribunal De Defensa	Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.	10,0
Tutor	Dra. RODRIGUEZ CABRERA, HORTENSIA MARIA , Ph.D.	10,0

Lo que da un promedio de: **10 (Diez punto Cero)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

ROMERO HERDOIZA, MAYLIN FERNANDA
Estudiante

Firmado electrónicamente por:
NELSON
FRANCISCO
SANTIAGO VISPO

Dr. SANTIAGO VISPO, NELSON FRANCISCO , Ph.D.
Presidente Tribunal de Defensa

HORTENSIA MARIA
RODRIGUEZ
CABRERA

Firmado digitalmente
por HORTENSIA MARIA
RODRIGUEZ CABRERA
Fecha: 2021.12.17
20:00:23 -05'00'

Dra. RODRIGUEZ CABRERA, HORTENSIA MARIA , Ph.D.
Tutor

FILIPE MIGUEL FERREIRA DE MENEZES AREIAS

Firmado digitalmente por FILIPE MIGUEL FERREIRA DE MENEZES AREIAS
Fecha: 2021.12.16 16:53:43 -05'00'

Dr. FERREIRA DE MENEZES AREIAS , FILIPE MIGUEL , Ph.D.
Miembro No Tutor

Firmado electrónicamente por:
MARIELA SOLEDAD YEPEZ MERLO

YEPEZ MERLO, MARIELA SOLEDAD
Secretario Ad-hoc

Autoría

Yo, **Maylin Fernanda Romero Herdoiza**, con cédula de identidad 0704627900, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así como, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autor(a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, diciembre 2021.

Maylin Fernanda Romero Herdoiza
0704627900

Autorización de publicación

Yo, **Maylin Fernanda Romero Herdoiza**, con cédula de identidad 0704627900,cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.
Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, diciembre 2021.

Maylin Fernanda Romero Herdoiza
0704627900

This page is intentionally left blank.

This work is completely dedicated to my parents:
Juan Romero and Yolanda Herdoiza

Acknowledgements

After five years of university, reaching this point means a lot to me. During the process, I have not only been able to gain knowledge and experience, but I have been able to grow personally, and set new goals for myself. I feel very grateful to all the people who have been part of this process.

First of all, I would like to express my deepest gratitude to my advisors, Ph.D. Yovani Marrero-Ponce and Ph.D. Hortensia Rodríguez. Thank you for all the support you have given me, for introducing me to the world of therapeutic peptides and bioinformatics, for all the learning, and for the trust you have placed in me. I would also extend my gratitude to the researchers at the University of Porto for their support in the use of some of the bioinformatic tools used during this research.

In addition, I would like to thank the School of Chemical Sciences and Engineering, and the professors I have had during the Common Core who have provided me with the knowledge and necessary skills to develop my undergraduate thesis.

I would also like to thank all my friends and colleagues who have accompanied me during these years with their limitless support, motivation, and affection. You have become my second home.

My greatest appreciation to Marlon for always being there for me, supporting me, listening to me, and advising me.

Finally, I want to express my eternal gratitude to my family for all the love, unconditional support, and trust. Especially to my parents, Juan and Yolanda, sisters, Fer, Cris and Lady, and brother Dari, who always do everything to see me grow professionally and personally.

Maylin Romero

This page is intentionally left blank.

Resumen

La inespecificidad de los fármacos quimioterapéuticos y la resistencia a múltiples fármacos (MDR) adquirida por las células cancerígenas generan la necesidad de encontrar alternativas para tratar el cáncer. Los fármacos basados en péptidos son enfoques prometedores en el tratamiento del cáncer, ya que presentan valiosas ventajas como bajo peso molecular, alta especificidad y baja toxicidad. En particular, los péptidos localizadores de tumores (THP) destacan por la capacidad de unirse específicamente a los receptores de las células cancerígenas y a la vasculatura tumoral. Por otro lado, el descubrimiento de fármacos *in silico* ha demostrado ser una forma eficaz y rápida para predecir agentes quimioterapéuticos. Actualmente, hay dos predictores de THP disponibles, TumorHPD y THPep, basados en aprendizaje automático (ML) supervisado. Aquí, se desarrolla una metodología alternativa para descubrir THPs utilizando ciencia de redes y búsqueda por similitud en starPep toolbox (`http://mobiosd-hub.com/starpep/`). Este enfoque se beneficia de la Red de Espacio Químico (CSN). Se diseñaron algunos modelos basados en THPs representativos y no redundantes de la CSN para descubrir nuevos THPs a través de la búsqueda por similitud y fusión de grupos. Su rendimiento se validó con tres conjuntos de datos de referencia de THPs/no-THPs. Se alcanzaron precisiones entre 92.64-99.18% y coeficientes de correlación de Matthews entre 0.894-0.98, superando a los clasificadores de ML. Estos resultados demuestran el potencial de la búsqueda por similitud y la ciencia de redes para la predicción de actividad. Además, el mejor modelo se utilizó para reutilizar péptidos de starPepDB. Se sometieron a una optimización multiobjetivo para mejorar su farmacocinética. Por último, se propone una pequeña biblioteca de péptidos, que consta de 27 THP y 14 péptidos localizadores de tumores anticancerígenos (ACP) putativos. Estos 41 péptidos no han sido relacionados con estas actividades hasta ahora. Por lo tanto, son agentes terapéuticos prometedores para una futura validación experimental.

Palabras clave: cáncer, péptido localizador de tumores, péptido anticancerígeno, descubrimiento de fármacos *in silico*, ciencia de redes, búsqueda por similitud, red de espacio químico, fusión de grupos.

Abstract

Unspecificity of chemotherapeutic drugs and multi-drug resistance (MDR) acquired by cancer cells generate the necessity to find alternatives to treat cancer. Peptide-based drugs are promising approaches in cancer treatments since they present valuable benefits as low molecular weight, high specificity, and low toxicity. Particularly, tumor homing peptides (THPs) are highlighted by their ability to specifically bind towards receptors from cancer cells and tumor vasculature. On the other hand, *in silico* drug discovery has demonstrated being an effective and rapid way to predict chemotherapeutic agents. Currently, there are two available THP predictors, TumorHPD and THPep, based on supervised Machine Learning (ML). Herein, an alternative methodology to discover THPs is developed using network science and similarity searching in starPep toolbox (`http://mobiosd-hub.com/starpep/`). The approach benefits from Chemical Space Network (CSN). Some models were designed based on representative and non-redundant THPs from the CSN to discover novel THPs through similarity searching and group fusion. Their performance was validated with three benchmarking datasets of THPs/non-THPs. Accuracies between 92.64-99.18% and Matthews correlation coefficients between 0.894-0.98 were achieved, outperforming ML classifiers. These results demonstrate the potential of similarity searching and network science for activity prediction. Moreover, the best model was used to repurpose peptides from starPepDB. They were subjected to multi-objective optimization to enhance their pharmacokinetic. Finally, a small peptide library is proposed, consisting of 27 putative THPs and 14 putative tumor homing anticancer peptides (ACPs). These 41 peptides are not related with these activities up to now. Thus, they are promising therapeutic agents for future experimental validation.

Keywords: cancer, tumor homing peptide, anticancer peptide, *in silico* drug discovery, network science, similarity searching, Chemical Space Network, group fusion.

Contents

List of Figures **iii**

List of Tables **v**

1 Introduction **1**

1.1 Scope of Research 2

1.2 Objectives 2

1.2.1 Principal Objective 2

1.2.2 Specific Objectives 2

2 Background Information **3**

2.1 Peptides as Therapeutic Agents 3

2.1.1 Peptides in Cancer Therapy 4

2.1.2 Tumor Homing Peptides (THPs) 6

2.1.3 Cell-Penetrating Peptides (CPPs) 11

2.2 Computational-Aided Drug Discovery 12

2.2.1 Machine Learning (ML) 12

2.2.2 Chemical Similarity Networks 13

2.3 Bioinformatic Tools 18

2.3.1 Databases 18

2.3.2 Web Servers 18

2.3.3 Software for Network Visualization and Analysis 19

3 Experimental Procedure **21**

3.1 Model Selection 21

3.1.1 Data Extraction 21

3.1.2 Similarity Threshold Analysis 21

3.1.3 Network Characterization 23

3.1.4 Similarity Searching Model for THPs Prediction 23

3.2 Potential THPs Prediction 27

3.2.1 Hierarchical Virtual Screening 27

3.2.2 Tumor Homing Optimization 28

3.2.3 Discovery of THP Motifs 29
3.3 Multi-Objective Optimization of THPs 31
3.3.1 Cell-Penetrating Activity 31
3.3.2 Half-Life Time 31
3.3.3 Building Blocks 33
3.3.4 Final Selection of Potential THPs 34

4 Results and Discussion 37
4.1 Model Selection 37
4.1.1 Similarity threshold analysis 37
4.1.2 Network characterization 38
4.1.3 Similarity Searching Model for THPs Prediction 39
4.2 Potential THPs Prediction 43
4.2.1 Hierarchical Virtual Screening 43
4.2.2 Tumor Homing Optimization 44
4.2.3 Discovery of THP motifs 46
4.3 Multi-Objective Optimization of THPs 53
4.3.1 Cell-Penetrating activity 53
4.3.2 Half-Life Time 53
4.3.3 Building Blocks 55
4.3.4 Final Selection of Potential THPs 55

5 Conclusions and Recommendations 60
5.1 Conclusions 60
5.2 Recommendations 60

References 61

Attachments 74

List of Figures

2.1 Peptide formation by an amide bond. Taken from [16]. 3

2.2 Summary of strengths and weaknesses of the application of peptides as therapeutic agents. Taken from [27]. 4

2.3 Comparison between (a) healthy and (b) cancerous cells. Taken from [40]. 5

2.4 Interactions between NGR and RGD from extracellular matrix (ECM) proteins with aminopeptidase N (APN) and integrin, respectively. Taken from [53] 7

2.5 Procedure of *in vivo* phage biopanning. Taken from [54]. 7

2.6 Different pathways of endocytosis. Taken from [100] 11

2.7 Mechanism of internalization of a peptide that contains the R/KXXR/K motif. Taken from [102]. 12

2.8 Chemical space representation as (a) a multi-dimensional coordinated-based map, and as (b) a coordinate-free similarity network. Taken from [117]. 14

3.1 General overview of the experimental procedure. 22

3.2 Schematic representation of the similarity searching process. Q is a peptide from a query dataset, n the number of peptides contained in a query dataset, S is a peptide from the target dataset (Main, Small or Main90 dataset), m is the number of peptides contained in the target dataset (1302, 938 or 619, respectively). 26

3.3 Procedure to optimize tumor homing activity of lead hits. 29

3.4 Procedure to optimize cell-penetrating activity. 31

3.5 Procedure to optimize stability in blood. 32

3.6 Procedure to optimize stability in the gastrointestinal tract. 33

3.7 Binary mutation guide tree used by ROSE to mutate the root peptide. Peptide libraries may be selected either from the internal nodes (peptides closely related to the root) or from terminal nodes/tree leaves (distantly-related to the root). 36

4.1 Density, modularity, and average clustering coefficient (ACC) as a function of similarity threshold of 627 THPs CSN. 38
4.2 CSN of giant component conformed by 528 THPs retrieved from starPepDB. Nodes color represent the community, and size how central the node is. 39
4.3 Degree distribution of the 528 giant components, where k is the vertex degree. 40
4.4 CSN of (a) 99 outliers with a density of 0.30 and (b) 34 remaining outliers obtained after 30% similarity extraction scaffold. Layout: Fruchterman Reingold. 41
4.5 Hierarchical virtual screening for repurposing of peptides from starPepDB. 49
4.6 Heat map of SET 1 (54 lead compounds). 50
4.7 Histogram of pairwise sequence identity of SET 1 (54 lead compounds). . . 50
4.8 3D structure of 27 putative THPs generated with PEP-FOLD 3 and visualized with PyMOL. 57
4.9 3D structure of 15 putative ACPs generated with PEP-FOLD 3 and visualized with PyMOL. 59

List of Tables

2.1 Classical (well-known) tumor homing motifs. **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis). 9

2.2 Webservers used for activity predictions of peptides. Algorithms for classification are Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (SGBoost), Artificial Neural Network (ANN), WEKA (package of classifier algorithms), Determinant Analysis (DA), and (Meta)genomic AMP Classification and Retrieval system (MACREL). . . . 20

3.1 Parameters to be met by peptides in activity predictions. 33

4.1 Global networks properties of CSN of 528 nodes and outliers. Density, number of clusters, and modularity were calculated in starPep toolbox, while average degree, ACC, and diameter in Gephi. 38

4.2 Results from statistical analysis of recovery performance of using models obtained from 30% of scaffold extraction in the CSN of giant components followed by removing nodes with 10% lower centrality than the most central node as queries. Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, PP is the precision of positives, and NP is the precision of negatives. 42

4.3 Results from statistical analysis of recovery performance when sets of 99 and 34 outliers were used as queries. Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives. 43

4.4 Results from statistical analysis of recovery performance when the mixture of sets were used as queries. H is the set obtained when harmonic centrality was calculated, W is the set obtained when the weighted degree was calculated, Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives. 44

4.5 Results from statistical analysis of recovery performance of using models obtained from 30, 40, 50, and 60% of scaffold extraction in the CSN of giant components as queries. Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives. 45
4.6 Results from statistical analysis of recovery performance when a mixture of sets obtained from harmonic and weighted degree using 60% of scaffold extraction, and 99 outliers were used as queries. H is the set obtained when harmonic centrality was calculated, W is the set obtained when the weighted degree was calculated, Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives. 46
4.7 Statistic analysis of 9 best SSMs in Small Dataset as target. 47
4.8 Statistics analysis of 9 best SSMs in Main90 Dataset as the target. 47
4.9 Comparison between the best SSM THP1 to predict THPs and the ML models reported in the literature as tumor homing benchmarking tests. P_{pos} corresponds to the sensibility, and R_{TP}(%) to specificity. 48
4.10 Discovered motifs by Multiple Sequence Alignment (MSA). **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis). . . 48
4.11 Discovered Motifs by STREME.**Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).) 51
4.12 Motifs found in PROSITE. **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis). 51
4.13 Short optimized THPs with 5-8 aa length. 55
4.14 Physicochemical properties of 14 putative ACPs. 58
4.15 Tumor homing and anticancer predictions of 14 putative ACPs. 59

Chapter 1

Introduction

Cancer is the second leading cause of death worldwide [1]. It is treated with radiotherapy, surgery, or systematic therapy [2]. Nevertheless, they carry short- and long-term health effects [2]. Notably, side effects of chemotherapy are caused by their unspecificity towards cancer cells [3]. For that reason, the attempts of the scientific community to find alternatives to currently used drugs do not cease. On this basis, peptides emerge as potential therapeutic agents for cancer treatment.

Peptides are characterized by having a low molecular weight in comparison to proteins and antibodies, short half-life time in the organism, binding to cancer cells selectively, and being non-toxic [4–7]. Hence, peptide-based drugs are opening a new door to an improved cancer diagnosis and treatment. Especially, tumor homing peptides (THPs) are highlighted by their capability to specifically bind to tumor cells and tumor vessels' receptors [8]. Therefore, they can be conjugated to therapeutic agents that present restrictions for their alone application in cancer therapy, acting as drug carriers.

THPs are discovered using *in vitro* and *ex vivo/in vivo* phage display technology [9]. However, wet-lab drug discovery procedures take much time, high investment, and need collaborative work from mastery in different fields [10]. Thus, prior *in silico* studies are employed for drug discovery, reducing resources and time [11]. In this way, short sets of molecules become the candidates for posterior experimental verification. Databases, web servers, and software, mainly based on Machine Learning (ML) approaches, are the bioinformatic tools applied to discover novel drugs [12]. Particularly about THPs, two databases recollect information about the experimentally proved peptides with tumor homing activity: TumorHoPe [13], and starPepDB [14], and two THPs web servers for prediction: TumorHPD [9], and THPep [15], both based on supervised ML approaches. Moreover, TumorHPD, and THPep design new THPs by the generation of random libraries where the peptide undergoes stochastic substitutions in different positions.

1.1 Scope of Research

This work seeks to broaden the chemical space of therapeutics peptides used for chemotherapeutic drug delivery, which contributes to solving one of the major problems found in cancer treatments: side-effects by unspecific targeting.

This study focuses on a set of potential THPs with anticancer activity found by an alternative methodology that combines network science and similarity searching. The main tools used are starPep toolbox software to perform the network analysis, scaffold extractions, and similarity searching; Chemical Space Network (CSN) to represent the chemical space of THPs as a coordinate-free system in starPep toolbox; freely available web servers for activity prediction and design of peptides; and, the evolutionary algorithm ROSE to generate peptide libraries.

1.2 Objectives

1.2.1 Principal Objective

The main objective is to discover potential THPs with a potential anticancer activity using an alternative methodology based on network science and similarity searching.

1.2.2 Specific Objectives

- To design a representative THPs model from starPepDB, which contains experimentally tested peptides.
- To carry out a similarity searching using the THPs model in the starPep toolbox to identify potential THPs from starPepDB.
- To discover new motifs in the set of potential THPs.
- To perform a multi-objective optimization of tumor homing, cell penetrability, anticancer capabilities, and half-time of potential THPs by punctual mutations and shortening sequence using webservers.
- To design THPs with anticancer activity using ROSE, an evolutionary algorithm.

Chapter 2

Background Information

2.1 Peptides as Therapeutic Agents

Peptides are short chains of amino acids joined together by peptide bonds, a covalent amide linkage (Figure 2.1). The peptide sequences are read from amino- to carboxyl-terminus [16]. Peptides have different biological roles in the organism, acting as biological regulators, inhibitors, neurotransmitters, antibiotics, hormones, ion channel ligands, or enzyme substrates [17, 18]. They can be synthesized, obtained from natural sources, or through genetic, recombinant, or chemical libraries [19, 20]. The worldwide methodology applied for peptide synthesis is the solid-phase synthesis, discovered by Merrifield in 1963 [21].

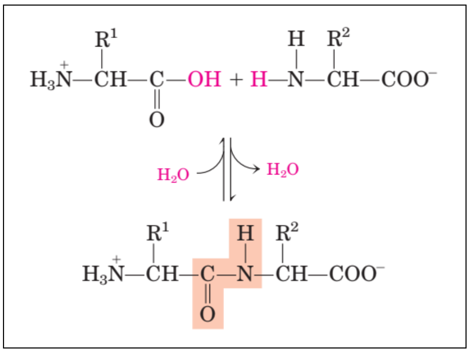


Figure 2.1: Peptide formation by an amide bond. Taken from [16].

Peptides have different biochemical and therapeutic characteristics than small molecules and proteins, making them attractive to the pharmaceutical and biotechnological industry to act as antimicrobials, antivirals, anticancer, cardiovascular agents, and treat diabetes, even vaccines [22, 23]. From 2015 to 2019, 15 peptides or peptide-containing molecules were approved by the U.S. Food and Drug Administration (FDA) as drugs demonstrating the growing interest of the scientific community [24].

Being smaller than proteins allows them to penetrate tissues more easily, have low cost, easier synthesis, and do not require folding to be biologically active [25]. In contrast to small molecules, they have higher specificity and efficacy due to representing the smallest functional part of a protein [19]. Moreover, they are not supposed to interact with the immune system, are biocompatible, have tunable bioactivity, and have low cytotoxicity due to the degradation products being amino acids [4, 17, 25].

The low oral bioavailability and rapid metabolism are significant challenges that peptides must face up to be potential drug candidates [22]. The main reasons are that, in general, their hydrophobicity prevents crossing physiological barriers, and proteases can quickly degrade them in the blood and digestive system [5, 19]. Then, peptides have low stability and short half-life time being removed from the circulation by the kidneys and liver in minutes. Consequently, commercially available therapeutic peptides are administered via subcutaneous, intravenous, or intramuscular injections [26]. Figure 2.2 summarizes the advantages and pitfalls of using therapeutic peptides.

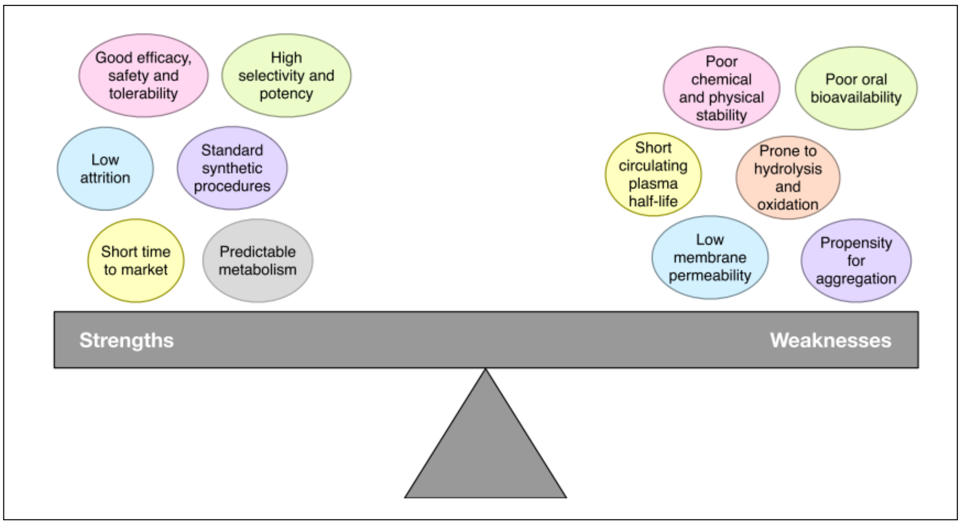

Good efficacy, safety and tolerability
High selectivity and potency
Low attrition
Standard synthetic procedures
Short time to market
Predictable metabolism
Poor chemical and physical stability
Poor oral bioavailability
Short circulating plasma half-life
Prone to hydrolysis and oxidation
Low membrane permeability
Propensity for aggregation
Strengths
Weaknesses

Figure 2.2: Summary of strengths and weaknesses of the application of peptides as therapeutic agents. Taken from [27].

However, studies reveal that some chemical modification and residue mutation improve their stability in plasma [28–30]. Indeed, it is reported that cysteine residues increase half-life by disulfide bond formation of the peptide with plasma albumin [31, 32]. The most common attempts to increase half-life are increasing molecular weight, cyclization, terminal modifications such as PEGylation, or replacing both terminals L-amino acids with D-amino acids [33–35].

2.1.1 Peptides in Cancer Therapy

Cancer is a disease that can be developed in different cell and tissue types. According to the World Health Organization, it is the second leading cause of death worldwide, with approximately 9.6 million deaths (one of six deaths) in 2018 [1]. It is based on the abnormal growth of cells due to an inherited genetic mutation or induced by the environment [3]. Cells are considered cancerous if acquired the following capabilities [36]:

- Generation of their signals and to respond to weak ones that are not identified by normal cells.
- Does not respond towards antiproliferative signals.
- Resistance towards apoptotic signals.
- Replication without limit.
- Angiogenesis, i.e., stimulation of new blood vessel formation to feed them and growth.
- Metastasis and tissue invasion, i.e., spreading the invasion through the body after the localized invasion of tissue.

Tumor blood and lymphatic vasculature differ in biomarkers expression and morphology from normal lymphatic and blood vessels [37, 38]. These differences are known as the "vascular zip codes" [39]. Besides, cancer cells commonly present a higher negatively charged and fluid outer membrane and greater surface area than normal mammalian cells (Figure 2.3) [40]. The high negatively charged membrane is granted by the presence of negative glycoproteins, phosphatidylserines, O-glycosylated mucins, and chaperone proteins. The high fluidity is a consequence of low cholesterol levels. Indeed, as cancer progresses, its membrane fluidity increases. Moreover, cancer cells increase the microvilli, which concede a higher surface.

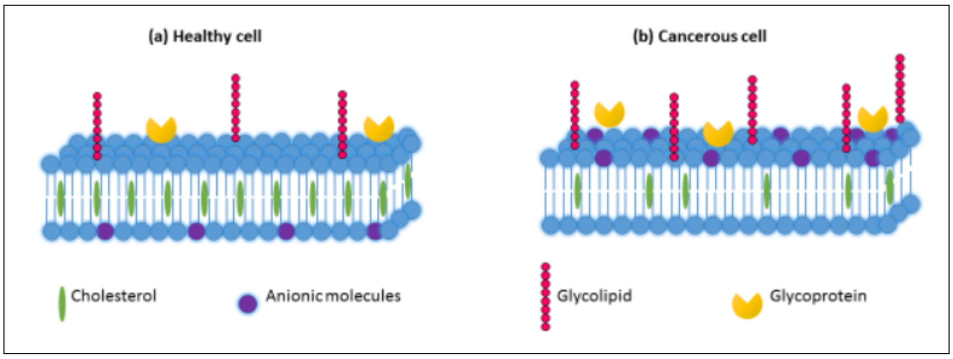


Figure 2.3: Comparison between (a) healthy and (b) cancerous cells. Taken from [40].

Localized cancers are treated with radiotherapy, by surgery, or both. However, in the case of metastatic or advanced cancer, they are treated with chemotherapy [41].

Chemotherapy is also used before local approaches to reduce the tumor size, known as neoadjuvant chemotherapy [42]. The main drawback of chemotherapy is that drugs in clinical use cannot differentiate between healthy and cancer cells, causing adverse side effects in patients [43]. Additionally, cancer cells are generating multi-drug resistance (MDR) [44]. For that reason, in the pharmaceutical industry, there is a necessity to develop new anticancer agents with a different mode of action to fight the current drug resistance of cancer cells without being cytotoxic to healthy cells [3]. Advantageously, peptides present some characteristics such as specificity towards cancer cells and low toxicity in healthy cells, allowing their application in diagnosis, treatment, and prognostic of cancer [6]. Chemotherapy based on peptides can be classified according to their mode of action.

- Mimetic peptides: to influence interactions between molecules that are relevant for cancer viability. In this way, they can induce apoptosis, immune response, tumor regression, inhibition of cancer growth, and angiogenesis [45–47].
- Biomarker peptides: to act as cancer-targeting of molecular imaging techniques in cancer diagnosis, such as magnetic resonance imaging (MRI), single-photon emission computed tomography (SPECT), or positron emission tomography (PET) [45, 48].
- Drug delivery systems: to penetrate biological barriers and/or home tumor cells or vessels to provide selective anticancer drug delivery [45, 46].

The FDA has already approved some anticancer peptides (ACPs) which are in clinical use [45]. In the years from 2015 to 2019, 5 FDA-approved drugs were destined for oncology [24].

2.1.2 Tumor Homing Peptides (THPs)

THPs are short peptides composed of 3-to-15 amino acids. They easily cross membranes by their small length and home tumor cells and vessels, taking advantage of cancer cells and tumor vessels' peculiarities [38]. THPs are widely investigated as drug carriers and for imaging purposes on oncology treatments and diagnosis since they decrease side effects [48, 49]. Moreover, nowadays, the application of nanomaterials to cancer treatment is of great interest, but they present size limitations causing low drug delivery efficiency [50]. Then, the development of peptides-conjugated nanomaterials is a promising drug delivery system.

First-generation of THPs have RGD and NGR motifs. RGD peptides have the characteristic of selectively binding to α integrin receptors of angiogenic blood vessels, metastatic tumor cells, and tumor endothelial cells, while NGR to aminopeptidase N receptors (Figure 2.4) [51, 52].

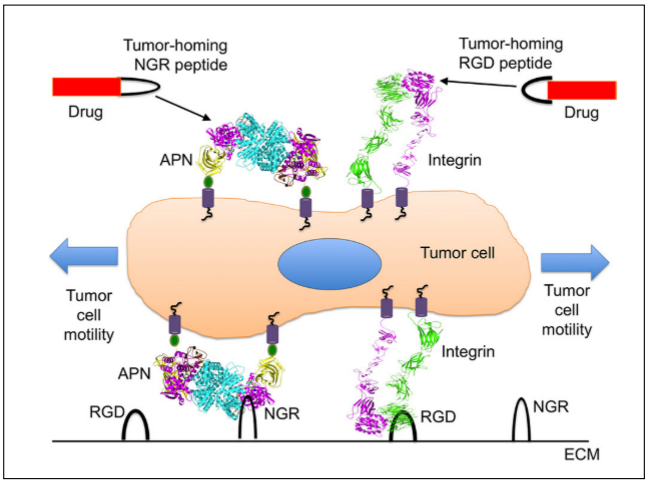


Figure 2.4: Interactions between NGR and RGD from extracellular matrix (ECM) proteins with aminopeptidase N (APN) and integrin, respectively. Taken from [53]

There are non-RGD neither NGR peptides that home tumor vasculature and cancer cells by interactions with other receptors, such as EGFR. The table 2.1 shows some of the reported motifs in THPs.

THPs and their target receptors are commonly identified through *in vitro* and *ex vivo/in vivo* phage display (Figure 2.5).

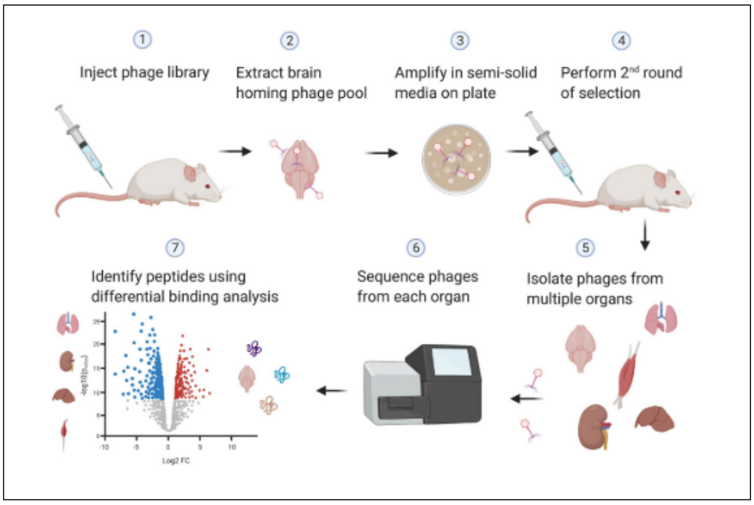

①
Inject phage library
②
Extract brain homing phage pool
③
Amplify in semi-solid media on plate
④
Perform 2nd round of selection
⑤
Isolate phages from multiple organs
⑥
Sequence phages from each organ
⑦
Identify peptides using differential binding analysis
Log2 FC

Figure 2.5: Procedure of *in vivo* phage biopanning. Taken from [54].

Phage display technology generates peptides by random peptide libraries. Random peptide libraries are constructed based on the insertion of random oligonucleotides in the genome of phages which will encode random peptide sequences on their surfaces [55]. The general process is that phages encoding different peptides are injected into the tail of mice and let them circulate through the body for 5 to 15 minutes [51]. Then, the specific sequence binds to the target receptor and is amplified to collect the tumor-specific phages. This process is called biopanning. The main advantages of these libraries are that they can simultaneously contain up to 10^9 variants [56]. Without previous knowledge of existing interactions, it is possible to identify the sequence interacting with a target molecule [57]. However, this task is uphill, and may not translate to humans due to differences between animal model and humans, such as peptide binding and vasculature [58].

Table 2.1: Classical (well-known) tumor homing motifs. **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).

No.	Motif	Frequency**	THP examples	Receptor	Target	Ref.
1	NGR	45/(41)	RGDPAYNGRFL	APN and integrin	Breast cancer cells (MDA-MB-435 and MCF-7) Contains RGD motif.	[59]
			CNGRCVSGCAGRC	APN	Breast cancer (MDA-MB-435), Kaposi's sarcoma, melanoma cells. Angiogenic endothelial cells. Containing VSG motif.	[60–66]
			NGRSL	APN	Neuroblastoma (GI-ME-N, GI-LI-N, HTLA-230, IMR-32, and SH-SY5Y, KS1767). Angiogenic endothelial cells.	[67]
2	RGD	36/(35)	CRGDK	NRP-1 and integrin	Breast cancer (MDA-MB-321, MDA-MB-231). Penetrating peptide	[68, 69]
			RGDGWK	Integrin $\alpha_5\beta_1$	Melanoma (B16F10). Tumor vessels.	[70]
			CEKRGDSVC	Integrins $\alpha_5\beta_3$ and $\alpha_5\beta_5$ NRP-1	Prostate cancer. Endothelial and tumor cells.	[71]
3	RVS	7/(6)	RSGRVSN	EGFR	Ovarian cancer (SKOV3).	[72]
			CRVSRQNKC		Lung (H460), stomach adenocarcinoma cells (SNU484).	[73]
			ARVSFWRYSSFAPTY	Galβ_1 $\rightarrow 3GalNAc\alpha$ disaccharide of T antigen	Breast cancer (MDA-MB-435) cells.	[74]
4	GVS	7/(7)	SKSSGVS		Breast cancer (MDA-MB-435).	[75]
			KGVSLSYRKKGVSLSYR	CXCR4	Competitive inhibitor to SDF-1 in solid tumors. Inhibit metastasis in osteosarcoma, melanoma, prostate, and breast cancer (MDA-MB-435). Cell death of ovarian cancer cells by mitotic catastrophe.	[76–78]
			ATLDGVS	EGFR	Ovarian cancer (SKOV3).	[72]
5	AEGEF	7/(7)	AEGEFIHNRYNRFFYWYGDPAK	HER2 (Human epidermal growth factor receptor 2)	Breast cancer (SKBr3, MCF7), ovarian cancer, melanoma. MeWo cells.	[79, 80]
			AEGEFMYWGDSHWLQYWYEGDPAK			
			AEGEFWGDSHWLQYWYEGDPAK			
6	VSG	5/(3)	RRHSVSG	EGFR	Ovarian cancer (SKOV3).	[72]
			CVSGPRC		Breast cancer.	[75]
7	CSD (CSDxxHxWC)	5/(4)	CSDSWHYWC	VEGFR-3	Tumor metastasis.	[81]
			CSDYNHHWC			
			CSDWQHPWC			
8	WRP	5/(5)	ASSSYPLIHWRPWAR	VEGF-C	Melanoma (B16BL6). Angiogenic endothelial cell. Tumor growth suppresor.	[82, 83]
			IHWRPWAR			
			DRWRPALP			

Table 2.1 (cont.): Classical (well-known) tumor homing motifs. **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).

No.	Motif	Frequency**	THP examples	Receptor	Target	Ref.
9	RPM	5/(14)	CPIEDRPMC	Integrin $\alpha_5\beta_1$	Colorectal cancer cells (HT29).	[84, 85]
			GRRPMKLNKTP		Liver metastatic gastric cancer cells (XGC9811-L).	
			ALRDRPM		Colorectal cancer cells (HT29).	[84]
10	SVR	4/(4)	NSVRGSR	EGFR	Ovarian cancer cells (SKOV3).	[72]
			IASVRWA		Melanoma tumor (B16B15b).	[75]
			CADPNSVRAMC		Cervical cancer (SiHa).	[86, 87]
11	PRP	0/(6)	SVSVGMKPSPRP		Angiogenic endothelial cells.	[86, 87]
			APRPG	VEGF-stimulated HUVECs	Tumor vasculature. Proved in Meth A sarcoma, and Colon 26 NL-17 carcinoma cells.	[82, 83]
			WTHHHSYPRPL			
12	GSL	0/(3)	GSLACQNIVVCVKKQCNALC		Breast carcinoma, Kaposi's sarcoma, and malignant melanoma.	[63]
			CLSGSLSC			
			CGSLVRC			
13	KGD	0/(0)		β_3 integrin	Melanoma cells, inhibit lung matastasis	[88, 89]
14	PSP	6/(6)	SVSVGMKPSPRP	VEGF- stimulated HUVECs	To the tumor neovasculature of both humans and mice. Human lung (H460), colon (HCT116), breast (BT483), prostate (PC3), liver (Mahlavu), and pancreatic (PaCa) cancer.	[90]
15	RGR	1/(1)	CRGRRST	platelet-derived growth factor receptor β (PDGFRβ) expressed	Angiogenic cells from tumor vasculature. NIH-3T3.	[73, 91]

2.1.3 Cell-Penetrating Peptides (CPPs)

Cell-penetrating peptides (CPPs) are short chains of 5-to-30 residues that can internalize into cells' cytosol without cell damage [92]. Trans-activator of transcription (TAT) protein from the human immunodeficiency virus 1 (HIV-1) is the first reported CPP [93]. CPPs are rich in basic amino acids (K, R, H, and Orn) [94]. However, arginine is the one that contributes the most to enhance penetrability [95]. The majority of CPPs are cationic, but they also can be amphipathic or hydrophobic [96, 97].

CPPs are used to transport cargo to targeted cells by conjugation or co-administration. The uptake depends on physicochemical properties of peptide, cell type, cargo, interactions with membranes, concentration, temperature, and peptide to cell ratio [48, 98]. The entry mechanism is energy-dependent endocytosis, direct penetration (energy-independent), or through multiple mechanisms [99]. Endocytosis can occur by phagocytosis, macropinocytosis, caveolae/lipid raft-mediated endocytosis, or clathrin-mediated endocytosis (Figure 2.6) [100, 101]. Direct penetration occurs at higher concentrations of CPPs by electrostatic interactions with membranes and then forming inverted micelles, carpets, or pores (barrel-stave and toroidal models) [92, 96].

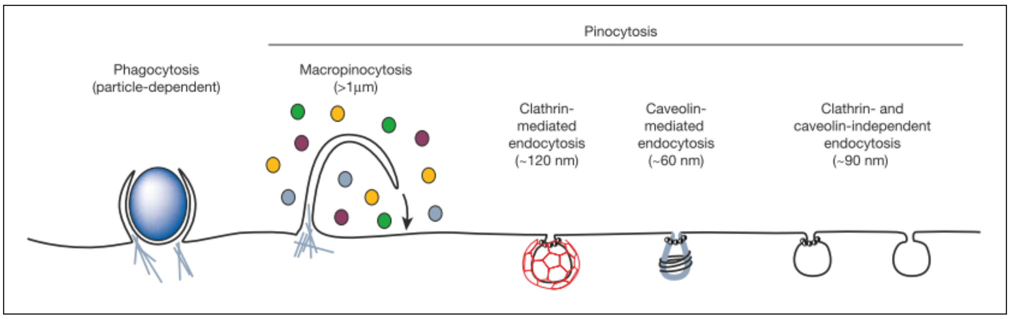


Figure 2.6: Different pathways of endocytosis. Taken from [100]

CendR (R/KXXR/K) represents an important CPP motif where X can be any amino acid different from R or K. It binds to neurophil-1 or -2 (NRP-1 and NRP-2, respectively) and initiates an endocytic transport (CendR pathway), but it is only active when it is located at the C-terminus [71]. When CendR is located into the sequence, protease can cut the sequence letting the motif in the C-terminus when the sequence is bonded to an integrin (Figure 2.7), for example, through the RGD motif, as happens in iRGD tumor-penetrating peptide [39].

The major disadvantage of CPPs is that they are not selective, then they cannot

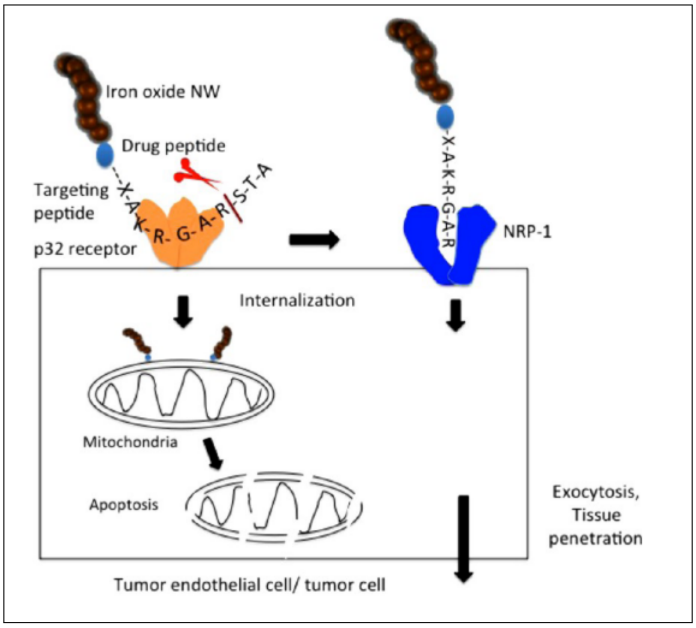


Figure 2.7: Mechanism of internalization of a peptide that contains the R/KXXR/K motif. Taken from [102].

differentiate between cancer and healthy cells [103]. Therefore, systems that penetrate tumor tissue and deliver at tumor-specific sites are desired in chemotherapy [104–107]. In this context, tumor-penetrating peptides have enhanced drug delivery of coupled and non-coupled drugs [108].

2.2 Computational-Aided Drug Discovery

Computer-aided drug discovery is a useful strategy to save time and resources, contributing to the fast introduction of peptides-based drugs in the global market [109]. In this work, some ML-based bioinformatic tools and chemical space networks are employed.

2.2.1 Machine Learning (ML)

ML, a subgroup of artificial intelligence, is the main approach applied for *in silico* drug discovery. ML uses algorithms to build mathematical models from training data sets to perform automated predictions of test sets [110].

Training data sets correspond to unlabeled and labeled data used as the sample sets. Meanwhile, test sets are unknown data sets that are going to be analyzed. The nature of the training data determines the ML model type [110], which can be: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, or transfer learning.

In supervised ML models, training data are labeled, while unsupervised models are unlabeled. Then, unsupervised ML models must find patterns to predict the outcomes. Semi-supervised ML combines parts of supervised and unsupervised models; its training data set contains labeled and unlabeled data, but the amount of unlabeled data is bigger. In reinforcement learning, the training data are used as feedback. Finally, transfer learning techniques consider that data is constantly changing, and data are transferred from one domain to another. By the way, supervised learning is the most widely applied for therapeutic peptides predictions [109].

ML models can classify or regress the training data to classify or regress the test sets, and the model performance will depend on the quality and quantity of training data [109, 110]. Therapeutic peptides models are commonly predicted through classifiers of supervised learning, particularly Random Forest (RF) and Support Vector Machine (SVM) [111].

RF applies classification or regression algorithms and is based on decision trees [112]. SVM classifies unlabeled data. It performs a binary classification using a linear hyperplane to maximize the separation between classes [7, 112]. When the space is not linear, SVM uses a kernel function to construct a linearly separating feature space, such as radial, polynomial, or Gaussian function [7].

2.2.2 Chemical Similarity Networks

The representation of all synthetically and natural molecules is known as chemical space. Nevertheless, as the amount of molecules in the chemical space is vast, small segments of the chemical space are used according to the activity of interest, namely, the biologically relevant chemical space where compounds that participate in biological systems are represented [113].

Chemical space is commonly visualized as a multi-dimensional coordinate system, where numerical features or computational vectors characterize molecules to represent their physicochemical properties, known as molecular descriptors [113]. Each molecular descriptor represents one dimension; thus coordinate-based maps require dimensionality reduction for visualizing in 2- or 3-Dimension maps [114]. This pitfall is known as a course of dimensionality [115]. In this scenario, coordinate-free chemical spaces, such as Chemical Space Networks (CSNs), emerge to visualize the chemical space with lower complexity [116]. Figure 2.8 illustrates the differences between chemical space represented as a coordinate-free system based on similarity and a coordinate-based map.

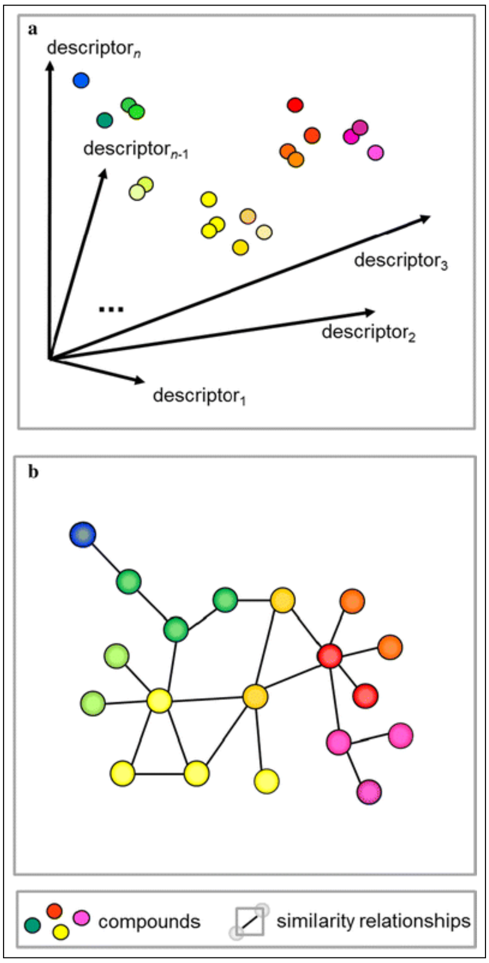


Figure 2.8: Chemical space representation as (a) a multi-dimensional coordinated-based map, and as (b) a coordinate-free similarity network. Taken from [117].

Chemical Space Network (CSN)

CSN is an undirected coordinate-free system (G) where molecules are represented as nodes joined by an edge if they have any similar relationship between their descriptors. Thus, it is defined as $G = (V, E)$, where V is the set of nodes present in network G, and E is the set of overall edges that connects V [118]. The connection between nodes depends on the selected similarity threshold. The similarity metric used in this work is the min-max normalized Euclidean, then the distance between two nodes is based on the Euclidean distance ($d(u, v)$). Two nodes are connected if $s(u, v)$ is equal or greater than the similarity threshold.

Layout algorithms determine the appearance of CSNs, where nodes are considered springs that repulse each other or are attracted by their similarity relationship. Distances between nodes are not based on how related they are but on the applied layout since its algorithm transforms pairwise similarity into the distance [119]. In this work, two algorithms for layout were used, Fruchterman Reingold and Force Atlas 2.

Additionally, some properties, and statistical measurements from networks science are applied to better understand these networks, such as clustering, modularity and centrality [120].

Similarity Threshold

Similarity threshold is an important concept in network science since it defines the network's topology and appearance [120]. It establishes the lower limit value of similarity between node pairs connected by an edge [121]. In other words, if two nodes have an equal or greater similarity value than the established, they are connected.

Node Degree

Node degree or vertex degree is the number of edges bonded to a node [122]. In other words, it represents the number of nodes with which it is attached.

Density

Network density is the ratio between the number of edges present in the network and all possible edges [120]. It depends on the similarity threshold value and determines the properties of the network. Then, network density is given by

$$\rho = \frac{2m_t}{n(n-1)} \tag{2.2.1}$$

where m_t is the number of edges at a threshold value t, and n is the number of nodes of the network. Generally, density decreases as the similarity threshold increases [114].

Clustering

Clustering is a concept with outstanding importance in unsupervised learning [114]. It is based on the division of the graph data into different communities or groups according to the similarity between nodes [120]. Consequently, similar nodes reside in the same community, and nodes from distinct communities are different. In a network, modularity measures how good is the classification of nodes into communities [123]. Modularity can be positive or negative with a maximum value of 1 [124], and is given by

$$Q = \frac{1}{2m_t} \sum_{uv} (a_{uv} - \frac{k_u k_v}{2m_t}) \delta(c_u, c_v) \tag{2.2.2}$$

where a_{uv} is the weight of the edge, i.e., similarity value between node u and node v, k_u is the sum of the weight of edges joined to node u, c_u is the community of u, and $\delta(c_u, c_v)$ is defined as

$$\delta(c_u, c_v) = \begin{cases} 1 & \text{if} \quad c_u = c_v \\ 0 & \text{if} \quad c_u \neq c_v \end{cases} \tag{2.2.3}$$

Modularity represents the number of edges that connect nodes intra-community minus the expected number of edges that are randomly settled in an identical network [124]. Modularity increases as density decreases, then community structures are better resolved. Hence, modularity must be optimized as much as possible to obtain the best partition of the network.

Louvain Clustering algorithm has demonstrated the best accuracy and computing time belong reported algorithms applied for modularity optimization [114, 125]. This algorithm begins assigning all nodes in different communities and consists of two phases [125]. In phase I, one node is moved to the community of its neighbor, and its new modularity value is evaluated. When the movement increases modularity, the node is changed to this community; otherwise, it keeps in the original community. This process is performed with the full nodes until no modularity improvement occurs. In phase II, a new network is constructed based on the resultant communities from the first phase. Finally, the process (phase I and II) is repeated until no modularity changes occur.

Moreover, the network ability to cluster together can be measured through the average clustering coefficient (ACC). The clustering coefficient is the ability to connect two nodes that share a neighbor [120]. Therefore, ACC is a global measurement of the neighborhood connectivity.

Centrality

In network science, centrality is one of the essential measurements since nodes rank according to how representative they are in the network [122, 126]. There are different methods to calculate centrality, but this research is focused on harmonic, community hub-bridge, betweenness, and weighted degree.

- Betweenness centrality: is based on short path lengths. The betweenness centrality of node u is the number of shortest paths between node pairs (without considering node u) that pass through it [127]. It is given by

  $$C_B(u) = \frac{1}{(N-1)(N-2)} \sum_{x \neq u, x \neq v, v \neq u} \frac{SP_{xv}(u)}{SP_{xv}} \tag{2.2.4}$$

  where N is the number of total nodes, $(N-1)(N-2)$ is the number of node pairs excluding node u,$SP_{xv}(u)$ is the number of shortest paths between nodes x and v that cross node u, and SP_{xv} is the total number of shortest paths between nodes x and v.

- Harmonic centrality: is a global centrality measurement based on the distance between two nodes [128]. The harmonic centrality of node u is given by

  $$C_H(u) = \sum_{v \neq u} \frac{1}{d(u,v)} \tag{2.2.5}$$

  where $d(u,v)$ is the distance from node u to node v.

- Weighted degree: is based on the similarity between a node pair, known as the weight of the edge [114]. It is given by the internal and external strength as follows.

  $$k_u^{in} = \sum_{v \in c_u} a_{uv} k_u^{ex} = \sum_{v \notin c_u} a_{uv} \tag{2.2.6}$$

  where k_u^{in} is the internal strength, k_u^{ex} is the external strength, and a_{uv} is the similarity value between u and v.

- Community hub-bridge centrality: is a local centrality measurement based on where the node is located into the community, and nodes can be considered hubs or bridges [114]. Local hub nodes are those that connect various internal nodes, while bridge nodes are those located at the boundary of a community being the attachment between two neighboring communities [129]. Hub-bridge centrality of node u is given by

$$C_{HB}(u) = k_u^{in} * CS(u) + k_u^{ex} * NC(u) \tag{2.2.7}$$

where k_u^{in}, and k_u^{ex} are internal and external strength, respectively, defined as in Weighted degree. $CS(u)$ is the community size of u, and $NC(u)$ is the number of neighboring communities directly attached to u by other nodes from its community.

2.3 Bioinformatic Tools

For the development of this study, some databases, web servers, and software are of particular interest.

2.3.1 Databases

TumorHoPe is a database intended to recollect reported THPs and contains 744 experimentally proved THPs [13]. Furthermore, starPepDB is a graph-based database that contains 45210 reported peptides where 659 corresponds to THPs [14].

2.3.2 Web Servers

To date, TumorHPD [9], and THPep [15] are the only web servers based on ML approaches for predicting of tumor homing activity of peptides.

TumorHPD developed by Sharma et al., was the pioneer web server [9]. Its prediction model is a SVM that uses three input features: amino acid composition, dipeptide composition, and binary profile pattern. The reported accuracy of their predictions is 86.56% [9]. However, the data set used for training and testing contains peptides with high similarity sequences and does not present statistical representations [15]. Additionally, the performance of the SVM model is not well described [15].

Alternatively, Shoombuatong et al. construct THPep where RF is used as prediction model, and amino acid composition, dipeptide composition, and pseudo amino acid composition as features achieving 90.13% of overall accuracy [15]. They removed the sequences with higher than 90% similarity to avoid overestimated predictions.

Moreover, other webservers were used to predict other activities, including cell-penetrating, anticancer, hemolysis, toxicity, antibacterial, among others (see Table 2.2).

2.3.3 Software for Network Visualization and Analysis

StarPep toolbox

StarPep toolbox is a software that uses FASTA files as inputs, and includes the starPepDB. Peptides are represented as nodes joined by an edge if they have any relationship. It can perform querying, filtering, visualization of networks, scaffold extractions, single or multiple queries similarity searching, and analysis of peptides by graph networks [14].

Networks can be built based on the metadata of peptides or based on the similarity between them. In metadata networks, nodes are connected by a specific parameter in common, such as origin, the target which assessed against, functionality, the database where they come from, the cross-reference, N-terminus, C-terminus, or amino acid composition. In similarity networks, peptides are defined by descriptors, such as length, net charge, isoelectric point, molecular weight, Boman index, indexes based on aggregation operators, hydrophobic moment, average hydrophilicity, hydrophobic periodicity, aliphatic index, and instability index. Moreover, networks are visualized using different layouts, such as Fruchterman Reingold or Force Atlas 2.

Networks can be clustered, and communities are optimized using the Louvain method. Moreover, centrality of each node can be measured, particularly, harmonic, community hub-bridge, betweenness, and weighted degree. Centrality is highly important to perform scaffold extractions due to peptides are ranked according to their centrality score, and then redundant sequences are removed, prioritizing the most central. Thus, scaffold extractions depend on the type of centrality applied.

On the other hand, similarity searching, which is the basis of this study, is performed using a set of queries against a target dataset, where different percentage of identity can be applied. The identity score is a number between 0-1, and it is calculated using Smith-Waterman local alignment and Blosum 62 substitution matrix. Multiple queries similarity searching works using the group fusion model.

Gephi

Gephi is open-source software for the visualization and analysis of network graphs. It calculates relevant data from the networks, such as average degree, diameter, radius, density, modularity, clustering coefficient, average clustering coefficient (ACC), average path length, number of edges, and nodes.

Table 2.2: Webservers used for activity predictions of peptides. Algorithms for classification are Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (SGBoost), Artificial Neural Network (ANN), WEKA (package of classifier algorithms), Determinant Analysis (DA), and (Meta)genomic AMP Classification and Retrieval system (MACREL).

No.	Webserver	Predicted activity	Classifier	Ref.
1	TumorHPD	Tumor homing	SVM	[9]
2	THPep	Tumor homing	SVM	[15]
3	AntiCP	Anticancer	SVM	[130]
4	ACPred	Anticancer	SVM and RF	[131]
5	iACP	Anticancer	SVM	[132]
6	ENNAACT	Anticancer	ANN	[133]
7	CellPPD	Cell penetrating	SVM	[134]
8	C2Pred	Cell penetrating	SVM	[135]
9	MLACP	Cell penetrating	RF	[136]
10	CpACpP	Anticancer and cell penetrating	XGBoost, SVM and RF	[137]
11	ToxinPred	Toxic	SVM	[138]
12	HemoPI	Hemolytic	SVM	[139]
13	HemoPred	Hemolytic	RF	[140]
14	PlifePred	Half-life in blood	SVM and WEKA	[28]
15	HLP	Half-life in acid media	SVM	[141]
16	ANuPP	Amyloidogenic	Regression	[142]
17	SGnn	Prion-like domains	ANN	[143]
18	AlgPred2	Allergens	RF	[144]
19	PepSolubility 1.0	Solubility	-	-
20	SolupHred	pH-dependent aggregation	-	[145]
21	IL2Pred	IL-2 induction	RF	[146]
22	IL4pred	IL-4 induction	SVM	[147]
23	IL-10Pred	IL-10 induction	SVM and RF	[148]
24	AIPpred	Inflammatory	RF	[136]
25	ProInflam	Inflammatory	SVM and RF	[149]
26	AntiInflam	Inflammatory	SVM	[150]
27	PRRpred	Pattern recognition receptor	SVM	[151]
28	QSPpred	Quorum sensitivity	SVM	[152]
29	AMPfun	Various (anticancer, antimicrobial, etc)	SVM and RF	[153]
30	CAMPr3	Antimicrobial	SVM, RF, ANN and DA	[154]
31	AxPEP	Antimicrobial	RF	[155]
32	Macrel	Antimicrobial and hemolytic	MACREL	[156]
33	AMPDiscover	Various (antimicrobial, antiviral, etc)	RF and ANN	[157]
34	ClassAMP	Antibacterial, antiviral and antifungal	SVM	[112]
35	iAMpred	Antibacterial, antiviral and antifungal	SVM	[158]
36	Antifp	Antifungal	SVM	[159]
37	Meta-iAVP	Antiviral	RF	[160]
38	AntiTbPred	Antitubercular	SVM	[161]
39	dPABBs	Bio-film active	SVM and WEKA	[162]

Chapter 3

Experimental Procedure

The overall workflow of this study, shown in Figure 3.1, is based on 3 steps: i) selection of the model of representative THPs from starPepDB, ii) prediction of potential THPs, and iii) multi-objective optimization of potential THPs. In the first step, some models of representative THPs from starPepDB were built using different centrality measures to rank the nodes and extract the representative and less redundant sequences by local alignment; then, the best model was selected in accordance with the performance and its capacity to correctly retrieved THPs from well-known THPs databases using similarity searching and group fusion. In the second step, the model was used to perform similarity searching with the aim to repurpose peptides as THPs from starPepDB, and their tumor homing activity was optimized using TumorHPD server. Additionally, sequence motifs were found from the set of potential THPs using multiple sequence alignments, alignment-free methods, and PROSITE server. In the last step, cell-penetrability, anticancer, and stability of potential THPs were optimized by three methodologies: punctual mutations and shortening the sequences in freely available webservers, creating a family of related peptides from a root peptide by applying a probabilistic model of evolution called ROSE, and by the addition of TAT sequence at the C-end.

3.1 Model Selection

3.1.1 Data Extraction

The dataset of reported THPs was extracted from starPepDB in starPep toolbox. All 45120 peptides contained in starPepDB were filtered by the query “Tumor Homing” in the metadata function, where 659 entries were obtained.

3.1.2 Similarity Threshold Analysis

Network analysis of peptides was performed building CSN of 659 THPs in starPep toolbox. In order to choose the appropriate similarity threshold to build the network of THPs, CSNs were built varying 0.05 the cut-off value from 0.10 to 0.90 (17 CSNs in total). Some

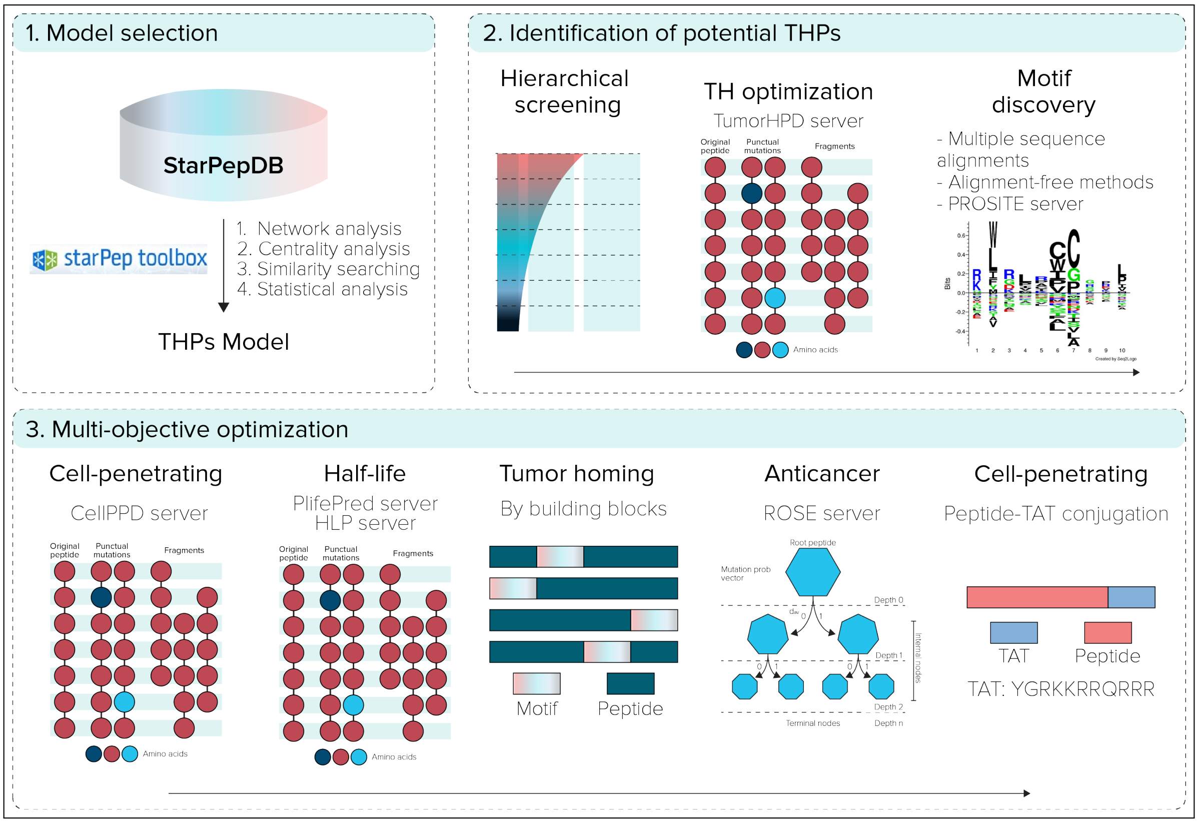


Figure 3.1: General overview of the experimental procedure.

metrics were retrieved from each CSN using starPep toolbox, such as density, number of communities, modularity, and number of singletons.

By default, when CSN is built, nodes with higher than 98% of similarity were eliminated using the local alignment Smith-Waterman algorithm and Blosum 62 substitution matrix. The similarity metric used to establish the pairwise similarity relationships between nodes was the min-max normalized Euclidean. Then, the community hub-bridge centrality was calculated with which outliers, nodes with 0 as vertex degree, were identified and removed, remaining the giant (or connected) components of the CSN, i.e., subgraph where all nodes are connected. After that, the network was clustered and the modularity optimized using the Modularity optimization clustering algorithm which is based on the Louvain method [125].

The network was saved as a Graph ML file to open in Gephi [163] for subsequent calculation of ACC. Finally, density, modularity, and ACC as a function of similarity threshold were graphed in Origin to decide which similarity threshold is better.

3.1.3 Network Characterization

CSN of the giant components using the best similarity threshold is characterized by the number of nodes, edges, outliers, density, number of communities, and modularity, which were parameters obtained from starPep toolbox; ACC, diameter (larger shortest path), average path length, and a total of triangles, which were obtained from Gephi; and the distribution degree. These parameters allow knowing the topology, and structural patterns of the CSN.

For network visualization, Force Atlas 2 was used as a layout algorithm, colors represent different clusters, and node size depends on how central is according to the community hub-bridge centrality. Network visualization aims to obtain an aesthetically pleasing and understandable graph where nodes are not overlapped.

Outliers

CSN of outliers was built with a cut-off of 0.30 to procure an appropriate density and, then, it was clustered. Moreover, a subsequent scaffold extraction was applied based on hub-bridge centrality, and 30% identity by local alignment was applied.

The network was characterized according to the number of nodes, edges, and communities, density, modularity, average degree, ACC, and diameter obtained before scaffold extraction, and the number of nodes and edges obtained after scaffold extraction. For network visualization, Fruchterman Reingold was used as a layout algorithm, colors represent different clusters, and node size depends on how central is according to hub-bridge.

3.1.4 Similarity Searching Model for THPs Prediction

In this study, the proposed method for discovering potential THPs was based on similarity searching. For that reason, multiple query similarity searching models (SSMs) composed by several queries of the most important and less redundant nodes of CSN and a similarity threshold were tested against datasets that contain well-known THPs/non-THPs through similarity searching. The recoveries from the similarity searching were statistically evaluated to choose the best model which was used to identify potential THPs.

Centrality analysis

The most influential nodes were used to find the new potential THPs, and centrality is the key parameter that provides this information. Thus, the four available centrality types in starPep toolbox, weighted degree, community hub-bridge, betweenness, and harmonic, were calculated and normalized using the min-max method. Then, redundant peptides were removed applying scaffold extraction, where peptides were ranked based on the scores obtained after centrality calculation, and using as similarity identity cutoff 30% based on local alignment algorithm Smith-Waterman and Blosum 62 substitution matrix. Subsequently, nodes with 10% lower centrality than the most central node were removed in each metric.

On the other hand, harmonic and weighted degree were calculated, normalized, and redundant peptides were removed applying 4 different percentages of similarity identity, 30, 40, 50, and 60%.

Query datasets (reference sequences)

The retrieved sets after applying scaffold extractions at each centrality measure and the two sets of outliers were used as queries. Additionally, combinations of outliers with sets obtained from centrality-based scaffold extractions, and combinations between sets obtained from scaffold extractions performed using different centrality metrics, were used as queries. In total, 22 sets of most influential nodes were used as queries, where 12 sets came from each applied percentage of scaffold extraction, 2 sets of outliers, and 8 sets came from the combination between sets.

Target Databases

Three training datasets that consider well-known THPs and randomly generated non-THPs [27] were used as the target or calibration for the recovery. THPep and TumorHPD employ these datasets for training their supervised ML classifiers [9, 27].

- Main dataset: 651 experimentally validated THPs and 651 random non-THPs. They were collected from TumorHoPe[13], and the literature [9].
- Small dataset: 469 experimentally validated THPs and 469 random non-THPs. They are peptides derived from the Main dataset with a length of 4-to-10 aa residues.

- Main90 dataset: 176 THPs and 443 non-THPs. They are peptides from the Main dataset with equal or lower than 90% of sequence similarity.

Main and Small datasets were retrieved from Ref. [9], while Main90 from Ref. [27].

Group fusion

Group fusion is based on the variation of a query (reference molecule), but keeping constant the identity measure [164]. The identity score is calculated to each peptide from target dataset varying the queries. The fusion group's algorithm associates a fused score to each target peptide, i.e., the maximum similarity (MAX-SIM) score from all obtained identity scores varying the query. Therefore, considering peptide S from target dataset, reference peptide Q from queries, and the identity score I(S,Q) the MAX-SIM score obtained, the algorithm assigns I(S,Q) as the fused score to peptide S. The identity scores were calculated with the Smith-Waterman local alignment algorithm with Blosum62 substitution matrix, and is a number between 0-1, being 1 the maximum similarity score.

Retrospective Similarity Searching

Main Dataset was imported to starPep toolbox, and the similarity searching based on local alignment Smith-Waterman and Blosum 62 substitution matrix were performed using the "Multiple query sequences" option of the software and the sets obtained from and 30% of scaffold extraction followed by removing nodes with 10% lower centrality than the most central node as queries. During the similarity searching the group fusion is applied by default, and results were ranked according to the fused score corresponding to the MAX-SIM value. Subsequently, seven different percentages of identity (similarity thresholds), 30, 40, 50, 60, 70, 80, and 90%, were tested, where peptides with similarity scores equal or higher than the applied threshold were retrieved as predicted THPs. The rescued nodes, i.e., predicted THPs, were statistically evaluated to validate the prediction. The procedure is illustrated in Figure 3.2. Here, it was possible to identify the two centrality measures and percentages of sequence identity with the best performance.

Then, similarity searching was performed using only sets of the best two centrality measures as queries: harmonic and weighted degree, and 30, 40, 50, 60, and 70% of identity. In Small and Main90 datasets, only sets of harmonic and weighted degrees were used, applying 40, 50, and 60% of identity for recovery.

In total 98 different SSM were evaluated. Figure 3.2 illustrates how similarity searching works.

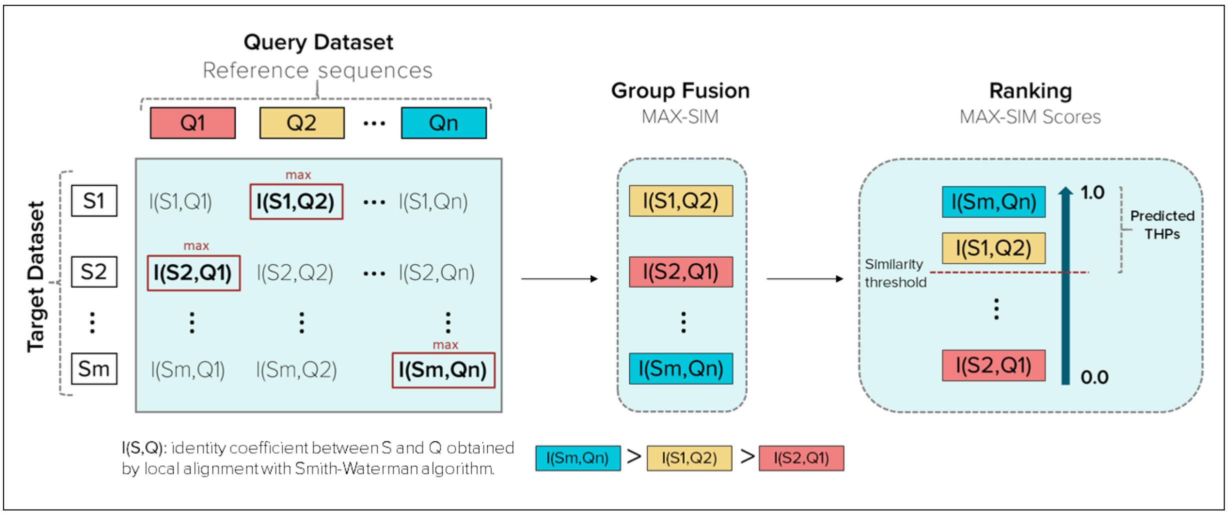


Figure 3.2: Schematic representation of the similarity searching process. Q is a peptide from a query dataset, n the number of peptides contained in a query dataset, S is a peptide from the target dataset (Main, Small or Main90 dataset), m is the number of peptides contained in the target dataset (1302, 938 or 619, respectively).

Statistical Analysis

The ability of the SSMs to predict THPs was validated by the measurement of accuracy (Ac), kappa (κ), recall (R), precision (P), Matthews correlation coefficient (MCC), and false accept rate (FAR%) using the following formulas.

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1.1}$$

$$\kappa = \frac{Po - Pc}{1 - Pc} \tag{3.1.2}$$

$$R_{TP} = \frac{TP}{TP + FN} \tag{3.1.3}$$

$$R_{TN} = \frac{TN}{TN + FP} \tag{3.1.4}$$

$$P_{pos} = \frac{TP}{TP + FP} \tag{3.1.5}$$

$$P_{neg} = \frac{TN}{TN + FN} \tag{3.1.6}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{3.1.7}$$

$$FAR\% = \frac{FP}{FP + TN} * 100 \tag{3.1.8}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, R_{TP} is the recall of true positive or sensitivity, R_{TN} is the recall of true negative or specificity, P_{pos} is the precision of positive predictions, P_{neg} is the precision of negative predictions, Po is the relative observed agreement between the observers equal to the Ac formula, and Pc is the expected chance agreement calculated by the formula $Pc = \frac{(TP+FP)*(TP+FN)+(FN+TN)*(FP+TN)}{(TP+TN+FP+FN)^2}$.

Finally, the best 9 SSMs were compared and ranked using the Friedman test-based analysis performed in KEEL [165], open-source software from Java. The Friedman test identified the best model based on the statistical metrics previously shown [166]. Moreover, it allows us to compare the models and determine if the difference between them is statistically significant and not due to chance. The confusion or classification matrix of the best model was constructed. Additionally, the best models were compared with reported ML models used for THPs prediction, TumorHPD and THPep, by using the same 3 calibration datasets.

3.2 Potential THPs Prediction

3.2.1 Hierarchical Virtual Screening

Pipeline Prospective Screening

The first step to identify potential THPs was to carry out a drug repurposing in the starPep toolbox, which is to find new target activities of known molecules [167]. Currently, this alternative methodology to discover drugs is widely applied due to reduced approval time for their clinical use [168]. In this sense, peptides from starPepDB were repurposing as THPs.

First, peptides without reported TH activity and toxicity with a sequence length between 3 to 25 residues were filtered from the chemical space of starPepDB. Secondly, peptides with higher than 95% of sequence similarity by local alignment Smith-Waterman and Blosum 62 substitution matrix were removed using the Scaffold extraction option. Thirdly, multiple query similarity searching was performed using the best SSM, obtained in the previous section, as the query against the chemical space of non-THPs, non-toxic, and non-redundant peptides with a length of 3-25 aa (amino acids), using 60% as similarity threshold. In the recovered set, peptides with a similarity score of 1 were removed.

Activity Prediction

Peptides with reported tumor homing activity in the literature were removed since the main objective of this study is to identify novel THPs. Then, theoretical activities of virtual hits were predicted using webservers 1-3, 7, 11, and 12 from Table 2.2, to corroborate their potential as THPs and prioritize those that do not harm healthy cells. The activities of interest were tumor homing, anticancer, cell-penetrating, toxicity, and hemolysis. The SVM thresholds used were 0.30 in servers 1, 3, and 7, and 0 in server 11.

Redundancy Reduction by Network Analysis

CSN of hits was built, clustered, and the modularity was optimized using the Louvain method in starPep toolbox. Then, harmonic and weighted degree centralities were calculated to perform a scaffold extraction using a 60% identity as threshold.

Visual Mining

The neighborhood of well-known THPs of each potential THPs was visualized using starPep toolbox. CSN of 659 THPs in starPepDB was built using 0.60 as cut-off, clustered, and optimized modularity. Hits obtained in the previous step after scaffold extraction were embedded into the CSN of 659 THPs to study the neighborhood of each peptide. Hence, the 3-nearest neighbors from 659 THPs which are directly attached to each hit, were visualized. When two peptides had the same two or three neighbors, one of them was prioritized, choosing the one with better-predicted activities.

3.2.2 Tumor Homing Optimization

Lead hits obtained from hierarchical virtual screening are peptides from starPepDB with a natural or designed activity different from tumor homing. This is why their tumor homing action should be enhanced. Lead hits were optimized by punctual amino acid mutations using the “Designing of Tumor Homing Peptides” module of TumorHPD (`https://webs.iiitd.edu.in/raghava/tumorhpd/peptide.php`) (Figure 3.3). Moreover, lead and mutated sequences were shortened into fragments 5, 10, and 15 residues in length using the same server.

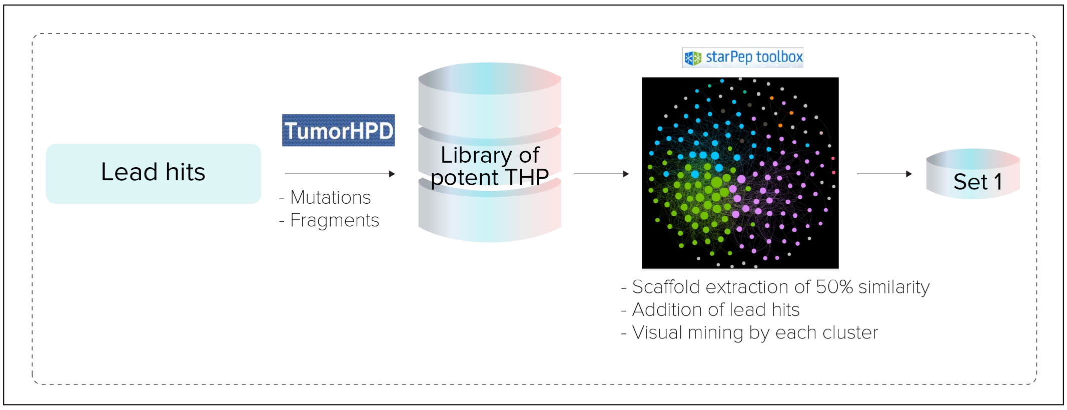


Figure 3.3: Procedure to optimize tumor homing activity of lead hits.

The selected optimized sequences, which must show a higher tumor homing activity score than parent hits, were analyzed through CSN in starPep toolbox using 0.60 as similarity threshold to built the network. Besides, tumor homing, toxicity, hemolytic, anticancer, and cell-penetrability were predicted using servers 2, 3, 7, 11, and 12 from Table 2.2. Redundant sequences with higher than 50% of similarity were removed by scaffold extraction.

Finally, the optimized sequences and parent hits were merged, and its CSN was built using 0.50 of cut-off and clustered. Moreover, harmonic centrality was calculated. Each cluster was analyzed separately in order to prioritize the most central, potent, non-toxic, and non-hemolytic potential THPs.

The heat map and histogram of pairwise sequence identity of lead compounds were constructed to study their structural diversity.

3.2.3 Discovery of THP Motifs

Multiple Sequence Alignments

The resulting potential THPs were hard-to-align sequences because of their short length and variability, they were grouped into seven clusters according to the neighborhood in the CSN. Given that clusters 1 and 5 were underrepresented by 2 peptides each, they were fused in a cluster labeled 1-5. Thus, peptide clusters (2-4, 1-5, and singletons) were aligned independently by using multiple sequence alignments (MSA), publicly available at `https://www.ebi.ac.uk/Tools/msa/`. Four different MSA algorithms were applied with their default parameters to determine consensus motifs within each cluster.

1. Clustal-Omega v 1.2.4 [169].
2. MAFFT (Multiple Alignment using Fast Fourier Transform) v7.487 with the iterative refinement FFT-NS-i option [170].
3. MUSCLE (Multiple Sequence Comparison by Log- Expectation) v3.8 [171].
4. T-Coffee (Tree-based Consistency Objective Function for Alignment Evaluation) v1.83 [172].

The resulting MSAs were employed to extract the conserved motifs by considering the consensus sequences estimation from the programs Jalview v2.11.1.4 [173], EMBOSS Cons v6.6.0 (`https://www.ebi.ac.uk/Tools/msa/emboss_cons/`) and Seq2Logov2.1 (`http://www.cbs.dtu.dk/biotools/Seq2Logo/`) [174].

Alignment-Free Method

Peptides were analyzed in STREME [175] (Sensitive, Thorough, Rapid, Enriched Motif Elicitation) to discover fixed-length patterns (ungapped motifs) that were enriched with respect to a control set generated by shuffling input peptides [173]. The analyses were performed via its webserver `https://meme-suite.org/meme/tools/streme`, by considering both total peptides and by each cluster. The motif width was set between 3-5 amino acids length. STREME applies a statistical test at p-value threshold = 0.05 to determine the enrichment of motifs in the input peptides compared to the control set.

Motif Search in PROSITE

Peptides were queried by the Motif Search tool (`https://www.genome.jp/tools/motif/`), integrated into the GenomeNet Suite (`https://www.genome.jp/`). PROSITE Pattern and PROSITE Profile libraries were only considered for the motif search.

3.3 Multi-Objective Optimization of THPs

3.3.1 Cell-Penetrating Activity

The penetration ability of 54 THPs was optimized by amino acid mutations using "Design Cell-Penetrating Peptide & Generate Its Mutants" module from CellPPD (`https://webs.iiitd.edu.in/raghava/cellppd/submission.php`). Toxicity, tumor homing, anticancer, and hemolytic activities of optimized sequences were also predicted using the servers 1-3, 7, 11, and 12 from Table 2.2. Then, 54 THPs were combined with optimized sequences, and its CSN was built in starPep using 0.65 as threshold and clustered. Harmonic centrality was calculated, followed by the scaffold extraction of sequences with lower than 90% similarity. Then, a set of sequences was selected, analyzing the neighborhood of each cluster and prioritizing optimized sequences with higher TH activity, non-toxic, and non-hemolytic. Finally, multiple reference similarity searching was performed using the THPs model. Figure 3.4 shows the overall procedure for the optimization.

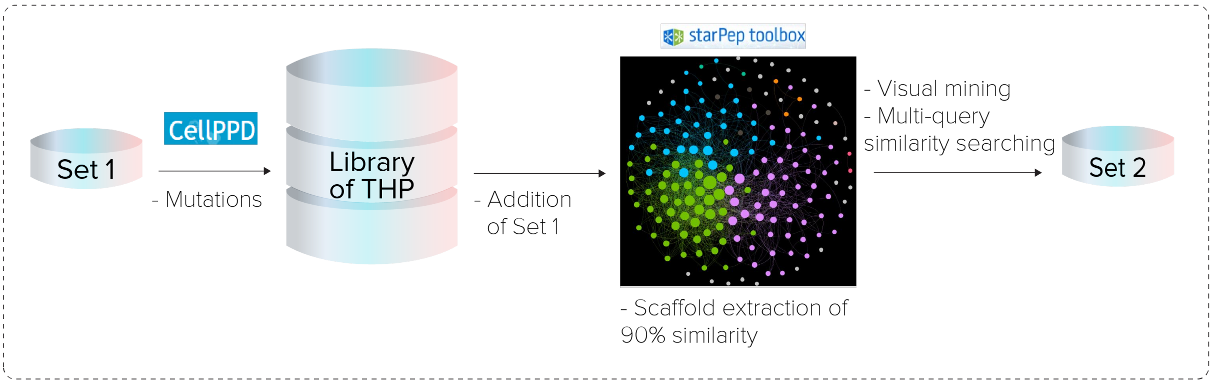


Figure 3.4: Procedure to optimize cell-penetrating activity.

3.3.2 Half-Life Time

Half-life in the blood of sequences obtained after cell-penetrating optimization was optimized by punctual amino acid mutations and shortening in fragments of 5 and 10 residues using the "Analog Generation" module from PlifePred (`https://webs.iiitd.edu.in/raghava/plifepred/analog.php`). Toxicity, tumor homing, cell-penetrating, anticancer, and hemolytic activities of optimized sequences were predicted using the servers 1-3, 7, 11, and 12 from Table 2.2. Then, multiple sequences searching was performed using the model of THPs in starPep toolbox. On the other hand, CSN of unrecovered sequences from multiple searching was built using 0.65 as threshold and clustered. Harmonic centrality was calculated, and sequences with higher than 60% similarity by local alignment were removed. Figure 3.5 shows the overall procedure for the optimization.

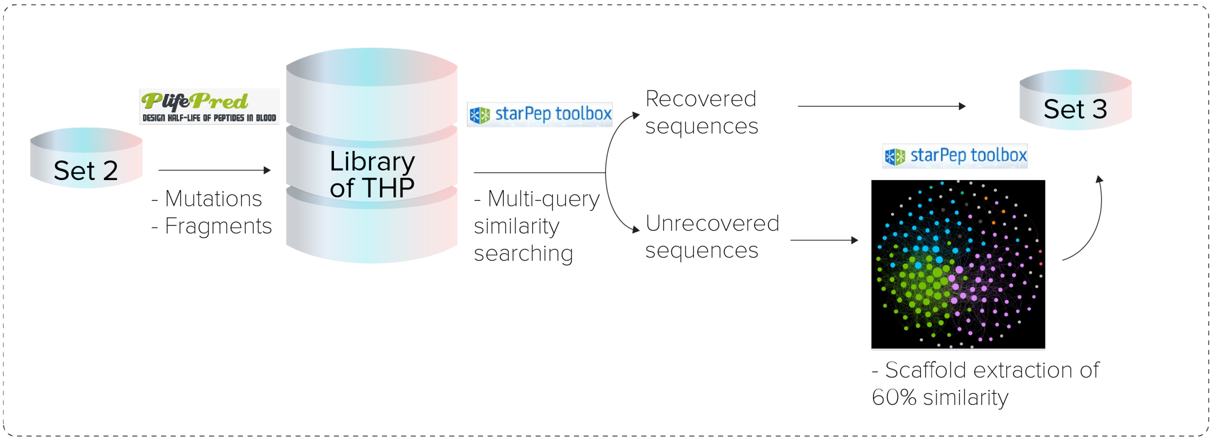


Figure 3.5: Procedure to optimize stability in blood.

Subsequently, half-life in an intestinal-like environment was optimized by punctual amino acids mutations and shortening in fragments of 5, 10 and 15 residues using the "Submission form for Designing Stable Peptide" module from HLP (`http://crdd.osdd.net/raghava/hlp/pep_both.htm`). Toxicity, tumor homing, cell-penetrating, anticancer, hemolytic activities, and half-time in the blood of optimized sequences were predicted using the servers 1-3, 7, 11, 12, and 14 from Table 2.2. Then, multiple sequences searching was performed with the model of THPs in starPep toolbox. In this case, the CSN of both recovered and unrecovered sequences was built using 0.65 as threshold. Finally, harmonic centrality was calculated, and sequences with higher than 90 and 65% similarity by local alignment were removed, respectively. Figure 3.6 shows the overall procedure for the optimization.

To deeply characterize the optimized THPs, more activities were predicted, such as allergen reaction using AlgPred2, aggregation-prone regions (APR) using AnuPP, and hemolysis using another server, HemoPred. In those sequences with unfavorable predicted activities, they were replaced with a better variant obtained by punctual mutations. They were filtered to keep a stronghold of the best performing multi-target sequences following the prediction specified in Table 3.1. Finally, multiple sequence searching was performed using 60% of identity with the model of THPs. Therefore, potential peptides are tumor

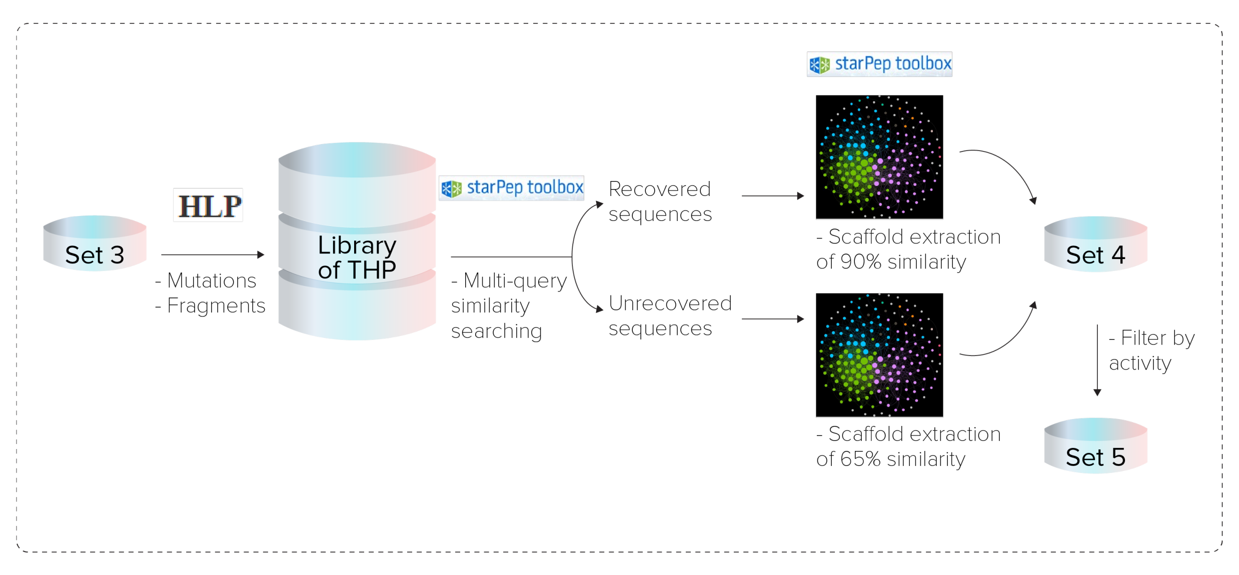


Figure 3.6: Procedure to optimize stability in the gastrointestinal tract.

homing active by all available models to predict THPs, the developed here, THPep, and TumorHPD.

Table 3.1: Parameters to be met by peptides in activity predictions.

Server	**Prediction**
TumorHPD	Score >2
THPep	TH
AntiCP	Score >0.60
ToxinPred	Non-toxic
HemoPred	Non-hemolytic
AlgPred	Score <0.40
ANuPP	no APR
HL in blood	>800 s

3.3.3 Building Blocks

Using three sets of motifs, new sequences were designed by building blocks.

1. Discovered motifs: Tables 4.10, 4.11, and 4.12.
2. Reported motifs: Table 2.1.
3. Short sequences from optimized hits: Table 4.13.

As two motifs found in PROSITE (Table 4.12) have 10 and 7 aa, respectively, motif 1 was divided into two sequences of 5 aa each, and only the last 4 aa of motif 2 were considered as a motif since the first 4 aa was also found as a motif by STREME (Table 4.11). Motifs were added in *N*-terminus, *C*-terminus, and inserted into three random positions in the sequences from SET 4 and also in motif sequences. Then, their activities (tumor homing, anticancer, cell-penetrating, toxicity, hemolysis, allergen, and half-life in blood) were predicted using the servers 1-3, 7, 11, 12, 14, 16, and 18 from Table 2.2, and sequences were filtered considering the parameters shown in Table 3.1.

CSN of the obtained sequences was built and clustered in starPep toolbox. Harmonic centrality was calculated, followed by a scaffold extraction of sequences with lower than 70% similarity by local alignment. Toxic sequences were identified and removed by the multi-sequence searching using as query a set of representative 105 venom peptides (Attachments **A**) obtained from starPepDB and 50% of identity as a cut-off. Then, TH active by the developed model and ACPs were identified by two independently multiple query similarity searching. One was performed using 60% of identity and the model of THPs as query, and the second was performed using 50% of identity with a set of representative 162 ACPs (Attachments **B**) obtained from starPepDB. Both recoveries were joined. In unrecovered peptides, CSN was constructed and clustered. Then, hub-bridge centrality was calculated, followed by a scaffold extraction of sequences with lower than 50% similarity.

On the other hand, active CPPs by the two models of CellPPD from the library of building blocks were filtered. Their CSN was built and clustered in starPep toolbox, followed by the scaffold extraction of sequences with lower than 50% similarity.

3.3.4 Final Selection of Potential THPs

Sequences obtained from the activity optimization (SET 5) and building blocks were joined. CSN was built and clustered, followed by a scaffold extraction of 70% similarity based on harmonic centrality in starPep toolbox. Then, a multiple sequence search was performed using 60% of identity with the THPs model and an ACPs model found in Attachments **B**. In unrecovered peptides, CSN was constructed and clustered. Then, hub-bridge centrality was calculated, followed by a scaffold extraction of sequences with higher than 50% similarity.

Finally, the set of potential THPs was reduced to 27 sequences in order to provide a pool of the most potent sequences considering other predicted activities, such as solubility (server 20 from Table 2.2). The biological profile of 27 sequences was characterized by predictions using the remaining servers from Table 2.2.

De novo Design of ACPs

An alternative optimization methodology was performed to improve tumor homing activity, penetrability, solubility but most importantly, anticancer activity while maintaining low toxicity and hemolysis. For this purpose, 14 sequences with a higher compromise between their activities were chosen and some of them were optimized using ROSE (`https://bibiserv.cebitec.uni-bielefeld.de/rose`) [176].

The ROSE program is an algorithm that creates a family of related peptides from a root peptide by applying a probabilistic model of evolution. The algorithm inserts, deletes, and substitutes amino acid residues from the sequences guided by the topology and branch lengths of a predefined evolutionary tree. ROSE was calibrated, hence the generated peptides retained at least 60% of identity with the corresponding root sequence. ROSE's internal parameters were tuned as follows: the binary mutation guide trees with 1023 nodes and depth k = 9, and average distance (d_{av}) of 5–20 PAMs. The diversity of the resulting peptides also depends on the root sequence, which is represented by a mutation probability vector. Each position/residue in the vector is weighted by variability or conservation degree shown in the sequence consensus, where the "zero" value indicates no mutations (high conservation degree), while the "one" value represents high mutation probability. Figure 3.7 illustrates the binary mutation guide tree used by ROSE.

Then, the well-known cell-penetrating sequence TAT (YGRKKRRQRRR) was added to the C-end of 14 peptides to evaluate if their anticancer activity is kept while penetrability increases. TAT was added directly and via a non-steric hindrance amino acid, A. Moreover, their 3D structures were generated using PEP-FOLD 3.0 [177], to study whether TAT affects the structural conformation of the sequence, resulting in loss of activity. Finally, the sequences were fully characterized using all of the servers listed in Table 2.2, and their activities were compared.

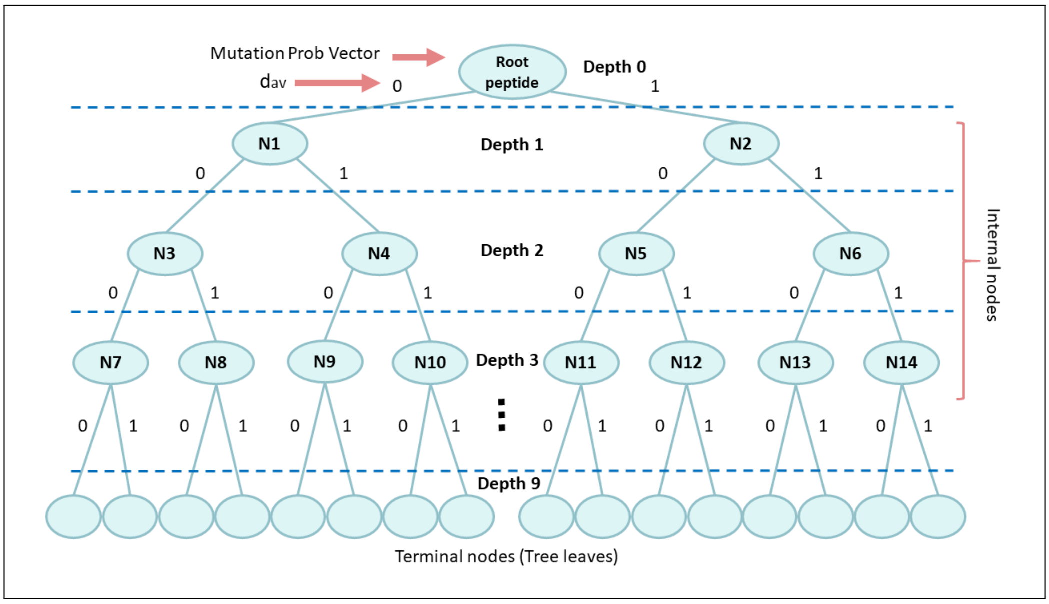


Figure 3.7: Binary mutation guide tree used by ROSE to mutate the root peptide. Peptide libraries may be selected either from the internal nodes (peptides closely related to the root) or from terminal nodes/tree leaves (distantly-related to the root).

Chapter 4

Results and Discussion

4.1 Model Selection

4.1.1 Similarity threshold analysis

From the set of 659 THPs retrieved from starPepDB, 627 peptides were filtered with lower than 98% similarity by local alignment. Before building CSN of 627 peptides, the adequate similarity threshold was chosen. This step is non-trivial since it is the parameter that defines the topology and networks parameters, such as density, modularity, ACC, and singletons [121]. Hence, the appropriate cutoff to build the CSN was determined based on how density, modularity, ACC, and singletons change varying the similarity cutoff. Attachments **C** shows the obtained parameters at each cutoff.

The graph of density, modularity, and ACC as a function of the similarity threshold (Figure 4.1) shows that density is lower at a higher similarity threshold, and ACC follows the same pattern until 0.65 similarity threshold. On the contrary, as the similarity threshold increases, modularity increases, and clustering is optimized.

A well-defined network needs a compromise between the density, modularity, and ACC parameters, but also the number of outliers because they are atypical peptides with particular properties. Networks with very low density result in too many outliers (Attachments **C**), while networks with very high density show a massive connection. In both cases, information is lost and interpretation becomes difficult. Literature reports that, generally, the best density percentages are around 1% or 2.5% due to displaying high modularity and allowing an adequate interpretation of the network [121]. As modularity indicates the existence of community structures, the ideal value must show an equilibrium between non-clustered network, and a network with artificial clusters due to too high modularity value. Based on that, the selected similarity threshold was 0.60 where CSN shows the best parameters and connectivity: 2.3% of density, 0.47 of modularity, 0.428 of ACC, and 99 outliers (15.8% of overall nodes). Therefore, the giant components of the network were 528 nodes.

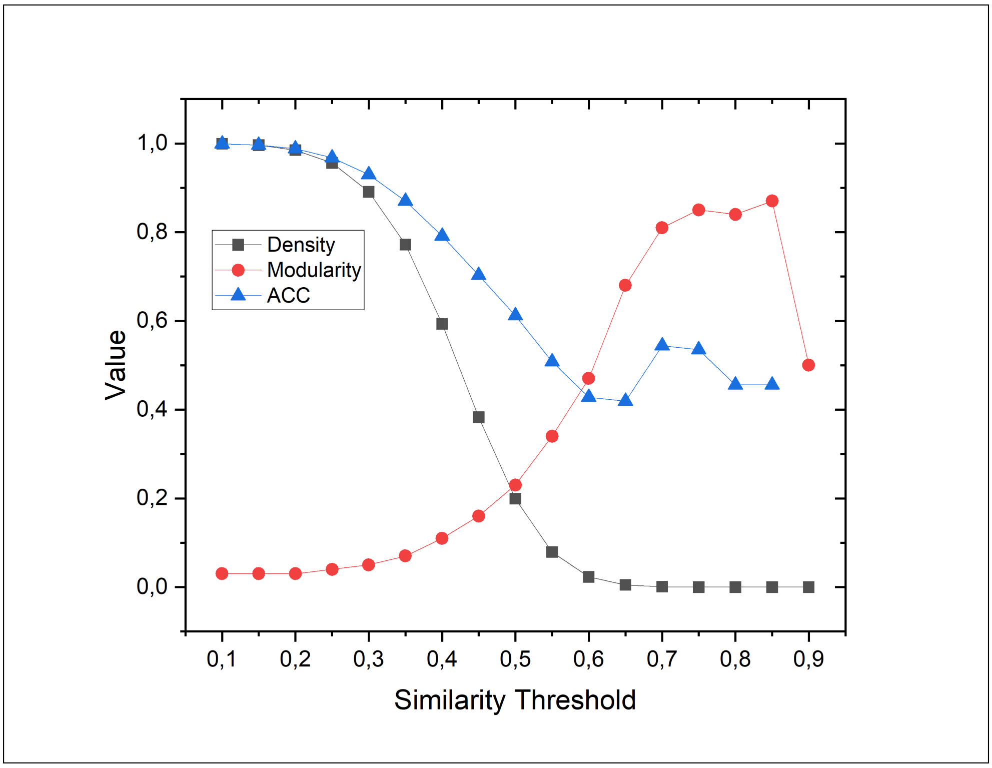


Figure 4.1: Density, modularity, and average clustering coefficient (ACC) as a function of similarity threshold of 627 THPs CSN.

4.1.2 Network characterization

To get a general comprehension of CSNs of the giant component (Figure 4.2) and outliers (Figure 4.4), some parameters were calculated (Table 4.1): density, number of clusters, modularity, average degree, ACC, and diameter.

Additionally, the degree of distribution of the giant components is shown in Figure 4.3. It gives some information about the structure of the CSN. In this case, it can be observed that the degree of distribution is concentrated in the nodes with low vertex degrees, but it has a tail associated to the nodes with higher vertex degrees that are in lower proportion. The nodes with higher degrees correspond to the most central nodes, which, as can be seen in Figure 4.2, are few.

Table 4.1: Global networks properties of CSN of 528 nodes and outliers. Density, number of clusters, and modularity were calculated in starPep toolbox, while average degree, ACC, and diameter in Gephi.

	Nodes	Edges	Density	Clusters	Modularity	Average degree	ACC	Diameter	Nodes after scaffold extraction	Edges after scaffold extraction
THPs	528	4452	0.023	10	0.47	16.864	0.428	8	-	-
Outliers	99	2691	0.891	3	0.13	54.364	0.733	3	34	384

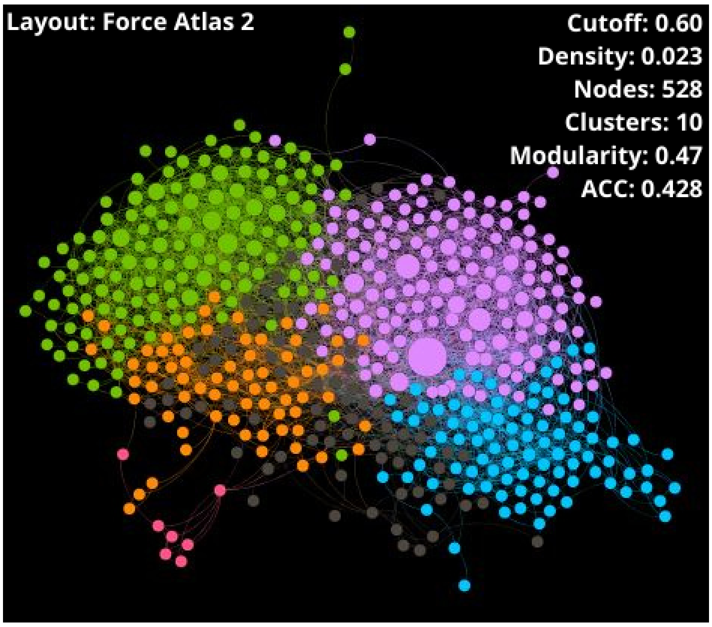


Figure 4.2: CSN of giant component conformed by 528 THPs retrieved from starPepDB. Nodes color represent the community, and size how central the node is.

Outliers are relevant THPs because they present particular characteristics that 528 nodes do not have. CSN of singletons was built using 0.30 of similarity threshold (Figure 4.4a). Then, sequences with higher than 30% similarity by local alignment were removed based on hub-bridge centrality ranking, where 34 outliers with orthogonal sequences were obtained (Figure 4.4b).

4.1.3 Similarity Searching Model for THPs Prediction

Centrality is the key parameter to build the model by which the novel THPs are going to be proposed since it allows the identification of the most influential sequences of the giant components. Moreover, outliers are nodes with unique properties that enriches the model of influential sequences. Therefore, sets from centrality measurements and sets of outliers represented the chemical space of THPs and were used as queries to perform the similarity searching against the target datasets. In total 98 different SSMs were generated, that were based on 22 query sets and similarity thresholds between 0.3 and 0.9.

Table 4.2 shows the results of the similarity searching using the sets obtained after applying 30% of scaffold extraction in the CSN of giant components followed by removing nodes with 10% lower centrality than the most central node as queries against the Main dataset. There, 7 different percentages of identity (30, 40, 50, 60, 70, 80, and 90%) were applied.

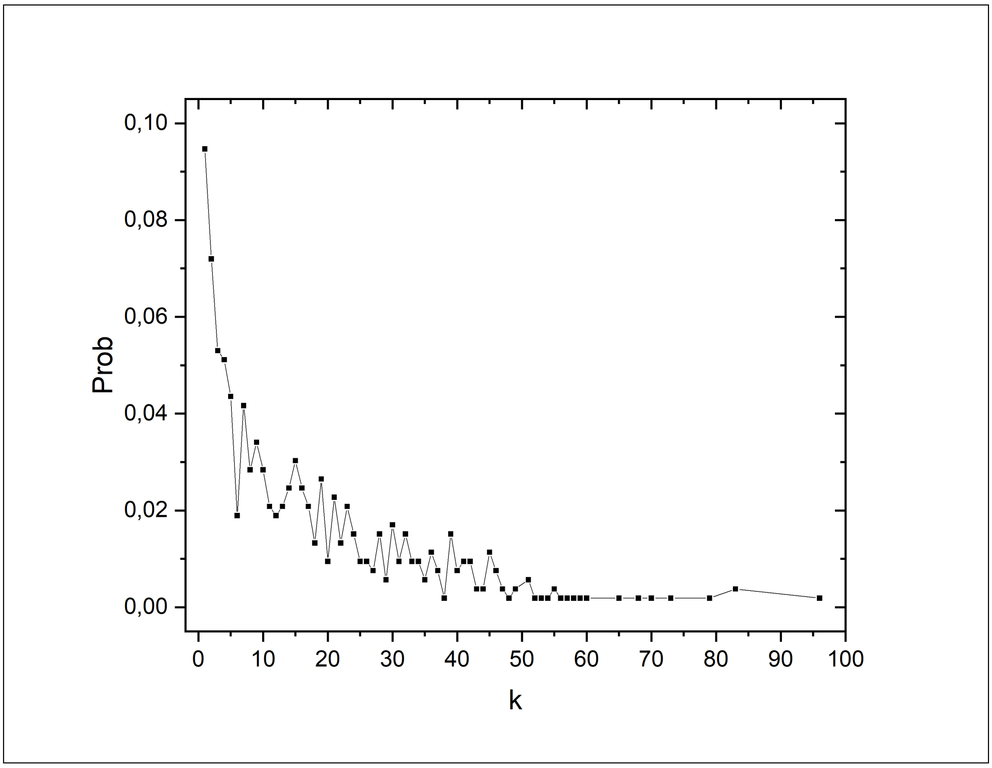
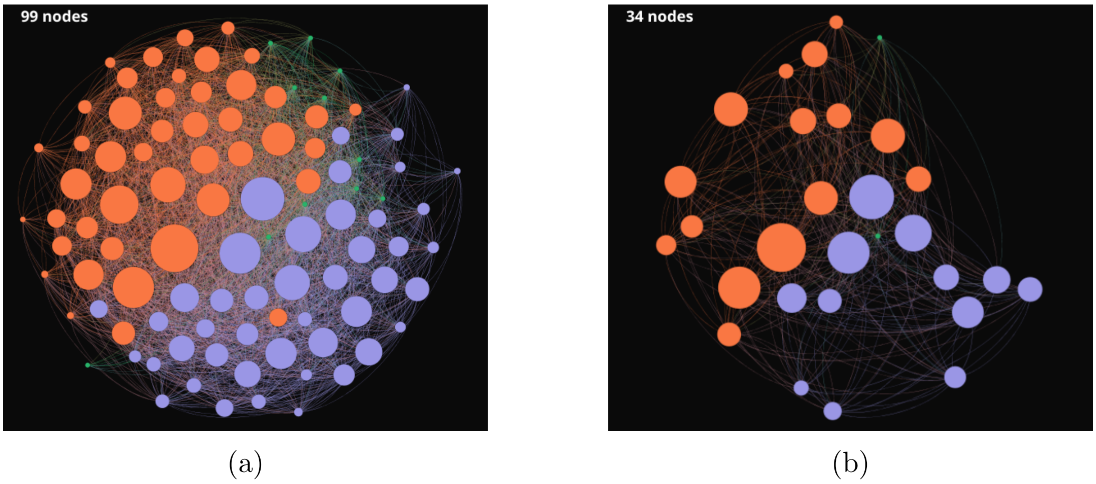


Figure 4.3: Degree distribution of the 528 giant components, where k is the vertex degree.

Results, shown in Table 4.2, indicated that harmonic centrality and weighted degree using 30, 40, 50, 60, and 70% of identity have a better performance with accuracy between 56-58%, recovering between 86-116 nodes. However, these results were not satisfactory considering that the Main dataset consists of 651 active nodes, and the recoveries contained low positive, and many FP nodes, which were reflected in the low R_{TP} and P_{neg} obtained, respectively.

On the other hand, using the sets of outliers as queries, results were also unsatisfactory due to R_{TP} and P_{neg} are low (Table 4.3). However, comparing the two sets of outliers, the set of 99 outliers showed better performance than 34 outliers, and similar statistics as previously obtained using harmonic and weighted degree. This behavior was expected because the outliers are unique sequences with high structural diversity whose similarity to the 528 THPs is less than 60%, so using a set with a higher number of outliers allows recovering more sequences by local alignment.

In order to increase the number of nodes used as queries, sets obtained from the harmonic and weighted degree, and 99 outliers were joined and used as queries (Table 4.4). However, results showed that R_{TP} and P_{neg} remained low.



(a) (b)

Figure 4.4: CSN of (a) 99 outliers with a density of 0.30 and (b) 34 remaining outliers obtained after 30% similarity extraction scaffold. Layout: Fruchterman Reingold.

Then, sets using 30, 40, 50, and 60% of scaffold extraction based on harmonic and weighted degree were obtained and used as queries. Statistics showed better recovery using 60% of scaffold extraction, where R_{TP}, P_{neg}, kappa, and a number of recovery increase significantly (Table 4.5).

In the last attempt to maximize the model's performance, sets were obtained from harmonic and weighted degree using 60% of scaffold extraction, and outliers were joined and used as queries in the similarity searching (Table 4.6). Table 4.6 shows that the best performance was achieved using the union of harmonic, weighted degree and 99 outliers sets as queries, in total 479 queries. Moreover, the best percentage of identity at which a compromise of all statistical parameters was achieved with 60%. All statistical parameters showed values greater than 0.88 (4.6).

In general, it is observed that the best performance of query datasets followed the tendency of $weighteddegree > harmonic > hub - bridge > betweenness > singletons$. Although, the combination of query datasets from different centrality types exceeds the performance of the sets obtained with only one centrality measure. Moreover, the addition of the sets of outliers improved the performance of the combination sets since it generates the complete representation of the chemical space of THPs.

On the other hand, the performance of the 9 best SSMs was validated in Small and Main90 Datasets. The models used as queries were the union of the set of harmonic with outliers, the set of weighted degree with outliers, and the sets of harmonic, weighted and 99 outliers, all using 60% of scaffold extraction by local alignment, and 40, 50, and 60%

Table 4.2: Results from statistical analysis of recovery performance of using models obtained from 30% of scaffold extraction in the CSN of giant components followed by removing nodes with 10% lower centrality than the most central node as queries. Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, PP is the precision of positives, and NP is the precision of negatives.

Centrality	**Nodes**	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
Weighted degree	54	30	0.588	765	537	0.175	0.177	0.998	0.991	0.548
		40	0.587	764	538	0.174	0.175	0.998	0.991	0.548
		50	0.585	762	540	0.171	0.172	0.998	0.991	0.547
		60	0.582	758	544	0.164	0.164	1	1	0.545
		70	0.566	737	565	0.132	0.132	1	1	0.535
		80	0.558	726	576	0.115	0.115	1	1	0.531
		90	0.549	715	587	0.098	0.098	1	1	0.526
Hub-bridge	31	30	0.564	734	568	0.127	0.129	0.998	0.988	0.534
		40	0.563	733	569	0.126	0.127	0.998	0.988	0.534
		50	0.562	732	570	0.124	0.126	0.998	0.988	0.533
		60	0.562	732	570	0.124	0.124	1	1	0.533
		70	0.547	712	590	0.094	0.094	1	1	0.525
		80	0.538	701	601	0.077	0.077	1	1	0.52
		90	0.531	692	610	0.063	0.063	1	1	0.516
Betweenness	25	30	0.56	729	573	0.12	0.121	0.998	0.988	0.532
		40	0.558	727	575	0.117	0.118	0.998	0.987	0.531
		50	0.558	727	575	0.117	0.118	0.998	0.987	0.531
		60	0.558	727	575	0.117	0.117	1	1	0.531
		70	0.54	703	599	0.08	0.08	1	1	0.521
		80	0.531	692	610	0.063	0.063	1	1	0.516
		90	0.525	684	618	0.051	0.051	1	1	0.513
Harmonic	63	30	0.574	747	555	0.147	0.151	0.997	0.98	0.54
		40	0.573	746	556	0.146	0.149	0.997	0.98	0.539
		50	0.572	745	557	0.144	0.146	0.998	0.99	0.539
		60	0.568	739	563	0.135	0.135	1	1	0.536
		70	0.566	737	565	0.132	0.132	1	1	0.535
		80	0.565	735	567	0.129	0.129	1	1	0.534
		90	0.559	728	574	0.118	0.118	1	1	0.531

identity in the similarity searching. Tables 4.7 and 4.8 show the results obtained and validate the performance of the models.

The best 9 SSMs were compared and ranked using the Friedman test by comparing the multiple statistical metrics of each SSM on the three target datasets (details in Attachments **D**). According to the test, the best SSM is the set **CSN-TH-0.60Sc-479-H+W+s-0.6-583**. It is composed by the union of nodes with identity lower than 60% from the global centrality harmonic with those obtained applying weighted degree and the set of 99 outliers, in total 479 nodes. The best percentage of identity used to carry out the similarity searching was 60%. The confusion matrices of the better SSM (THP1) are

Table 4.3: Results from statistical analysis of recovery performance when sets of 99 and 34 outliers were used as queries. Ac is the accuracy, R$_{TP}$ is the recall of true positives, R$_{TN}$ is the recall of true negatives, P$_{pos}$ is the precision of positives, and P$_{neg}$ is the precision of negatives.

	Nodes	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
Outliers	99	30	0.585	762	540	0.171	0.174	0.997	0.983	0.547
		40	0.578	752	550	0.155	0.158	0.997	0.981	0.542
		50	0.578	752	550	0.155	0.158	0.997	0.981	0.542
		60	0.577	751	551	0.154	0.157	0.997	0.981	0.542
		70	0.577	751	551	0.154	0.157	0.997	0.981	0.542
		80	0.576	750	552	0.152	0.154	0.998	0.99	0.541
		90	0.576	750	552	0.152	0.152	1	1	0.541
Outliers (30%Sc)	34	30	0.528	688	614	0.057	0.058	0.998	0.974	0.515
		40	0.526	685	617	0.052	0.054	0.998	0.972	0.513
		50	0.526	685	617	0.052	0.054	0.998	0.972	0.513
		60	0.526	685	617	0.052	0.054	0.998	0.972	0.513
		70	0.526	685	617	0.052	0.054	0.998	0.972	0.513
		80	0.526	685	617	0.052	0.052	1	1	0.513
		90	0.526	685	617	0.052	0.052	1	1	0.513

shown in Attachments **E**. It can be seen that the prediction of the model was not random due to MCC was much greater than 0 [178]. Moreover, the performance of statistical metrics showed good results with accuracy, recall, precision, and kappa statistic values higher than 0.85.

Finally, Friedman test of the THP1 versus the reported models used in TumorHPD[9] and THPep[15] servers revealed that the similarity searching methodology to discover potential THPs is superior (details in Attachments **F**). These ML results present a weak predictive ability, where accuracy is 86.56% and 90.13%, and maximal MCC is 0.70 and 0.76, respectively [9, 15]. The test found significant differences between THP1 and the ML models from TumorHPD and THPep servers. Table 4.9 shows the comparison between them on all benchmarking datasets.

4.2 Potential THPs Prediction

4.2.1 Hierarchical Virtual Screening

Molecules to be repurposed using the THP1 were a stronghold of peptides from the chemical space of starPepDB. Starting from 45120 peptides, and after applying the previously

Table 4.4: Results from statistical analysis of recovery performance when the mixture of sets were used as queries. H is the set obtained when harmonic centrality was calculated, W is the set obtained when the weighted degree was calculated, Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives.

	Nodes	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
H+W	77	30	0.606	789	513	0.212	0.215	0.997	0.986	0.559
		40	0.605	788	514	0.21	0.214	0.997	0.986	0.559
		50	0.604	787	515	0.209	0.21	0.998	0.993	0.558
W+S	167	30	0.664	865	437	0.329	0.333	0.995	0.986	0.599
		40	0.664	864	438	0.327	0.332	0.995	0.986	0.598
		50	0.662	862	440	0.324	0.329	0.995	0.986	0.597
H+S	153	30	0.651	848	454	0.303	0.309	0.994	0.98	0.59
		40	0.651	847	455	0.301	0.307	0.994	0.98	0.589
		50	0.65	846	456	0.3	0.304	0.995	0.985	0.589
H+W+S	176	30	0.683	889	413	0.366	0.372	0.994	0.984	0.613
		40	0.682	888	414	0.364	0.37	0.994	0.984	0.612
		50	0.681	887	415	0.363	0.367	0.995	0.988	0.611

explained filters and performing the similarity searching, 43 lead hits were retrieved (Attachments **G**). Figure 4.5 shows the step-by-step hierarchical virtual screening. Until today, these repurposed sequences do not have reported tumor homing activity, demonstrating their high potential as tumor homing agents.

4.2.2 Tumor Homing Optimization

A library of 180 sequences (Attachments **H**) was obtained from optimization of 43 hits in TumorHPD with a higher TH score than the originals, non-toxicity, and less hemolytic activity. Mutations enriched the sequences in W and C, where mainly, G and V residues from originals were mutated to W, and R, K, and also some W to C. Studies report that the presence of W contributes positively to the intracellular translocation of peptides [179]. Moreover, it was reported that W enhances the stability of peptides in serum and salt [180].

41 peptides from the library were prioritized by studying their CSN where 50% scaffold extraction by local alignment was accomplished. To perform the scaffold extraction, the sequences were clustered and ranked according to the global harmonic centrality, and only the most central sequences with a similarity between them lower than 50% were kept. 41 sequences have higher predicted TH activity by TumorHPD than original peptides

Table 4.5: Results from statistical analysis of recovery performance of using models obtained from 30, 40, 50, and 60% of scaffold extraction in the CSN of giant components as queries. Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives.

	% Sc	**Nodes**	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
Harmonic	30	58	30	0.561	731	571	0.123	0.124	0.998	0.988	0.533
			40	0.561	730	572	0.121	0.123	0.998	0.988	0.532
			50	0.56	729	573	0.12	0.121	0.998	0.988	0.532
	40	140	30	0.658	857	445	0.316	0.323	0.994	0.981	0.595
			40	0.658	857	445	0.316	0.321	0.995	0.986	0.594
			50	0.659	858	444	0.318	0.32	0.998	0.995	0.595
	50	251	30	0.763	993	309	0.525	0.533	0.992	0.986	0.68
			40	0.763	994	308	0.527	0.533	0.994	0.989	0.68
			50	0.765	996	306	0.53	0.533	0.997	0.994	0.681
	60	368	30	0.859	1118	184	0.717	0.727	0.991	0.987	0.784
			40	0.859	1119	183	0.719	0.727	0.992	0.99	0.784
			50	0.862	1122	180	0.724	0.727	0.997	0.996	0.785
Weighted degree	30	60	30	0.589	767	535	0.178	0.18	0.998	0.992	0.549
			40	0.588	766	536	0.177	0.178	0.998	0.991	0.549
			50	0.588	765	537	0.175	0.177	0.998	0.991	0.548
	40	140	30	0.657	855	447	0.313	0.32	0.994	0.981	0.594
			40	0.657	855	447	0.313	0.318	0.995	0.986	0.593
			50	0.657	856	446	0.315	0.316	0.998	0.995	0.594
	50	255	30	0.761	991	311	0.522	0.525	0.997	0.994	0.677
			40	0.76	989	313	0.519	0.525	0.994	0.988	0.677
			50	0.5	651	651	0	0	1	0	0.5
	60	370	30	0.859	1119	183	0.719	0.728	0.991	0.988	0.785
			40	0.86	1120	182	0.72	0.728	0.992	0.99	0.785
			50	0.863	1123	179	0.725	0.728	0.997	0.996	0.786

with scores between 0.39 and 1.92. Moreover, they are anticancer and have less toxicity and hemolytic activity. 12 of 41 sequences come from fragments of original sequences of 5, 10, and 15 lengths; 15 obtained after 4 punctual mutations of originals; and 14 from fragments of mutated sequences of 5, 10, and 15 lengths. Two of 41 peptides, CNGRCGGKLA and WCAMS, are part of reported THPs, which validates the novel methodology to discover potential THPs described here. CNGRCGGKLA is the N-end of CNGRCGGKLAKLAKKLAKLAK peptide which contains NGR TH motif and a disulfide bridge that gives stability. CNGRCGGKLAKLAKKLAKLAK binds to CD13 of tumor cells acting as ACP and THP [181]. While WCAMS is the C-end of KLWCAMS peptide that homes mouse B16B15b melanoma [75].

From the combination of 43 lead and 41 optimized hits, 54 peptides (SET 1) were selected. Sequences from SET 1 present a diverse molecular structure, low toxicity, and

Table 4.6: Results from statistical analysis of recovery performance when a mixture of sets obtained from harmonic and weighted degree using 60% of scaffold extraction, and 99 outliers were used as queries. H is the set obtained when harmonic centrality was calculated, W is the set obtained when the weighted degree was calculated, Ac is the accuracy, R_{TP} is the recall of true positives, R_{TN} is the recall of true negatives, P_{pos} is the precision of positives, and P_{neg} is the precision of negatives.

	Nodes	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
W+H	380	30	0.867	1129	173	0.734	0.743	0.991	0.988	0.794
		40	0.868	1130	172	0.736	0.743	0.992	0.99	0.795
		50	0.87	1133	169	0.74	0.743	0.997	0.996	0.795
		60	0.87	1133	169	0.74	0.74	1	1	0.794
		70	0.849	1106	196	0.699	0.699	1	1	0.769
H+S	467	30	0.932	1214	88	0.865	0.877	0.988	0.986	0.889
		40	0.933	1215	87	0.866	0.877	0.989	0.988	0.89
		50	0.935	1218	84	0.871	0.877	0.994	0.993	0.89
		60	0.935	1218	84	0.871	0.874	0.997	0.996	0.888
		70	0.913	1189	113	0.826	0.829	0.997	0.996	0.854
W+S	469	30	0.933	1215	87	0.866	0.879	0.988	0.986	0.891
		40	0.934	1216	86	0.868	0.879	0.989	0.988	0.891
		50	0.936	1219	83	0.873	0.879	0.994	0.993	0.891
		60	0.937	1220	82	0.874	0.877	0.997	0.997	0.89
		70	0.915	1191	111	0.829	0.833	0.997	0.996	0.856
H+W+S	479	30	0.941	1225	77	0.882	0.894	0.988	0.986	0.903
		40	0.942	1226	76	0.883	0.894	0.989	0.988	0.903
		50	0.944	1229	73	0.888	0.894	0.994	0.993	0.904
		60	0.945	1230	72	0.889	0.892	0.997	0.997	0.903
		70	0.923	1202	100	0.846	0.849	0.997	0.996	0.869

hemolytic activity, and most of them also show potential anticancer activity (Attachments **I**). The sequence diversity of the lead peptides was evaluated by using all vs. all global alignment where pairwise sequence identities were calculated. As Figures 4.6 and 4.7 show, they all mostly displayed sequence identities lower than 30% indicating structural singularity. Among the 54 lead hits, only one sequence has the well-known NGR motif. Therefore, SET 1 is composed of new structural entities within the known structural space of the THPs.

4.2.3 Discovery of THP motifs

As a consequence of the structural diversity of SET 1, the discovery of motifs accounting for the TH activity is not a straightforward task. In this sense, sensitive multiple sequence alignment (MSA) tools and alignment-free (AF) approaches (e.g., STREME) were applied to unravel new TH motifs.

Table 4.7: Statistic analysis of 9 best SSMs in Small Dataset as target.

	Nodes	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
H+S	467	40	0.917	860	78	0.834	0.838	0.996	0.995	0.86
		50	0.916	859	79	0.832	0.836	0.996	0.995	0.858
		60	0.914	857	81	0.827	0.832	0.996	0.995	0.855
W+S	469	40	0.92	863	75	0.84	0.844	0.996	0.995	0.865
		50	0.92	863	75	0.84	0.844	0.996	0.995	0.865
		60	0.919	862	76	0.838	0.842	0.996	0.995	0.863
H+W+S	479	40	0.928	870	68	0.855	0.859	0.996	0.995	0.876
		50	0.928	870	68	0.855	0.859	0.996	0.995	0.876
		60	0.926	869	69	0.853	0.857	0.996	0.995	0.875

Table 4.8: Statistics analysis of 9 best SSMs in Main90 Dataset as the target.

	Nodes	**% Id**	**Ac**	**Correct class.**	**Incorrect class.**	κ	$\mathbf{R}_{TP}$	$\mathbf{R}_{TN}$	$\mathbf{P}_{pos}$	$\mathbf{P}_{neg}$
H+S	467	40	0.985	600	9	0.964	0.983	0.986	0.966	0.993
		50	0.99	603	6	0.976	0.983	0.993	0.983	0.993
		60	0.992	604	5	0.98	0.983	0.995	0.989	0.993
W+S	469	40	0.98	597	12	0.952	0.966	0.986	0.966	0.986
		50	0.984	599	10	0.96	0.966	0.991	0.977	0.986
		60	0.987	601	8	0.968	0.966	0.995	0.988	0.986
H+W+S	479	40	0.985	600	9	0.964	0.983	0.986	0.966	0.993
		50	0.989	602	7	0.972	0.983	0.991	0.977	0.993
		60	0.992	604	5	0.98	0.983	0.995	0.989	0.993

To make possible the application of MSA algorithms for motif identification, the resulting 54 lead THPs were mapped onto CSN space to identify putative communities. These networks communities were considered clusters containing related peptides. Finally, 6 clusters were conformed with 14, 10, 8, 4, 10, and 8 members, respectively. The last cluster grouped the singletons (peptides identified as atypical in the CSN).

Clustal-Omega [169], MAFFT [170], MUSCLE [171], and T-Coffee [172] which are MSA algorithms developed after the classical ClustalW were applied, so that they can deal with hard-to-align sequences shown in each cluster, and thus to detect any conserved signature or motif. Since each MSA has implemented a different algorithm to improve alignment quality, their altogether consideration for the estimation of consensus regions helped us to identify TH motifs by using the Jalview, EMBOSS Cons and Seq2Logo programs (Attachtments **J**). As the EMBOSS Cons, gives a more legible output, only displaying high scored amino acids/positions (capital letters), less scored but positive residues (lower-case letters), and non-consensus positions (x), were selected as

Table 4.9: Comparison between the best SSM THP1 to predict THPs and the ML models reported in the literature as tumor homing benchmarking tests. P_{pos} corresponds to the sensibility, and R_{TP}(%) to specificity.

Dataset	**Method**	**Ac(%)**	**P_{pos}(%)**	**R_{TP}(%)**	**MCC**
Main	TumorHPD	86.56	80.63	89.71	0.7
	THPep	86.1	87.07	85.18	0.72
	THP1	**94.47**	**99.66**	**89.25**	**0.894**
Small	TumorHPD	81.88	73.13	90.92	0.65
	THPep	83.37	81.24	85.81	0.67
	THP1	92.64	99.5	85.71	**0.861**
Main90	TumorHPD	89.66	83.64	80.68	0.74
	THPep	90.8	91.8	87.97	0.77
	THP1	**99.18**	**98.86**	**98.3**	**0.98**

primary source to set consensus/conserved regions. The non-consensus positions were estimated using default parameters by visual inspection of the corresponding positions in the Jalview program [173] and in the Seq2Logo [174]. Table 4.10 depicts the consensus motifs, unraveled by each MSA algorithm.

Table 4.10: Discovered motifs by Multiple Sequence Alignment (MSA). **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).

No.	**Motif**	**EMBOSS concensus**	**Cluster**	**Cluster size**	**MSA Method**	**Frequency****
1	wwW	wwW	2	14	CLUSTALW-O	1/(1)
		xxW			MAFFT	0/(0)
2	C[fl][rg][vl]rW	CxxxrW	3	10	MAFFT	0/(0)
3	C[gpi][gs]cR	CxxxR			MUSCLE	0/(0)
4	[rkl]GLC	RGlc	4	8	CLUSTALW-O	0/(0)
		kGLC			MAFFT	0/(0)
		xGLc			MUSCLE	0/(0)
5	c[wp]kG	cwkG	1+5	4	CLUSTALW-O MUSCLE	0/(0)
		cxkG			T-Coffee	0/(1)
6	Not found	Non-concensus	6	10	CLUSTALW-O MUSCLE MAFFT T-Coffee	0
7	l[rp][cw]c	lxxc	Singletons	8	MUSCLE	0/(0)

None of the motifs found by MSA have been reported as TH motifs. However, one of the motifs from No.3 “CxxxR”, “CGGCR” contains “CXXC” motif which is the active site of thioredoxin (Trx), a relevant protein in mammalian cells that act as an antioxidant and participates in programmed cell death inhibition and cell growth, commonly used as a target for cancer treatments [182, 183]. Moreover, “CIGCR” (from No.3 “CxxxR”) is

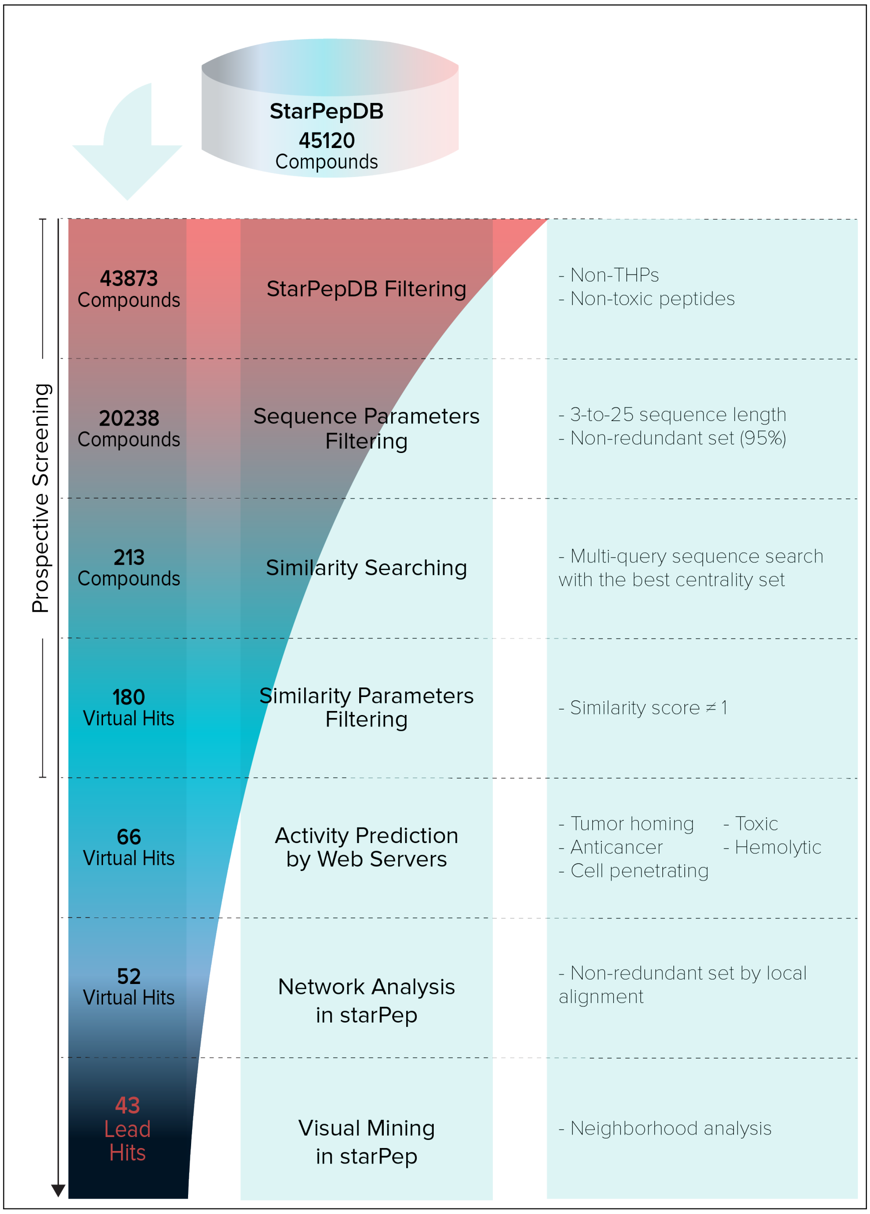


Figure 4.5: Hierarchical virtual screening for repurposing of peptides from starPepDB.

a motif from Epstein-Barr nuclear antigen 1 (EBNA1) epitope which binds to G protein-coupled receptor in pregnant women, related to pre-eclampsia [184], and "CWKG" (No.5) is contained in a nanoscale molecular platform used as drug delivery system in chemotherapy to enhance the conjugation of mitomycin C to the carrier [185].

On the other hand, the AF approach STREME was used to find unaligned patterns ranging 3-5 aa length within the overall 54 peptides and within each peptide cluster. STREME has been recently reported as the most accurate and sensitive algorithm among its competing state-of-art partners [175]. Unlike previous algorithms [186–188], STREME uses a position weight matrix (PWM) to count position matches efficiently for a motif

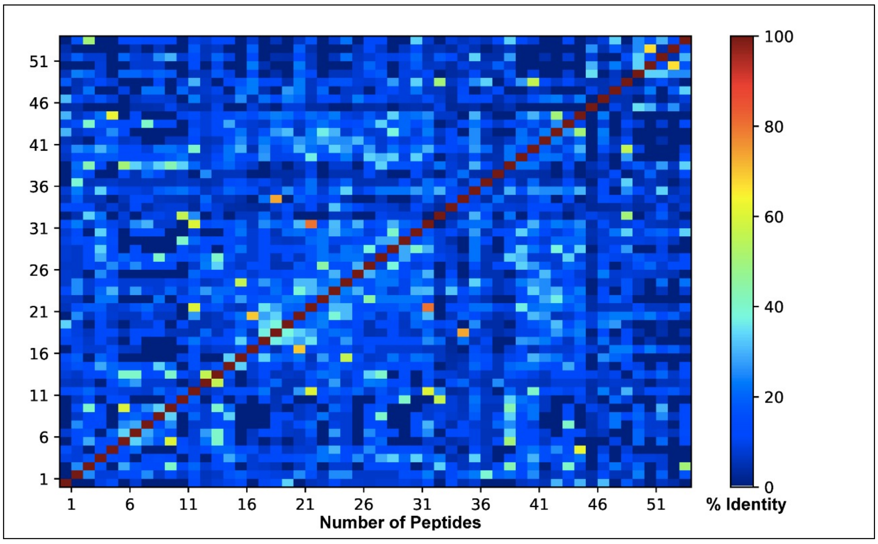


Figure 4.6: Heat map of SET 1 (54 lead compounds).

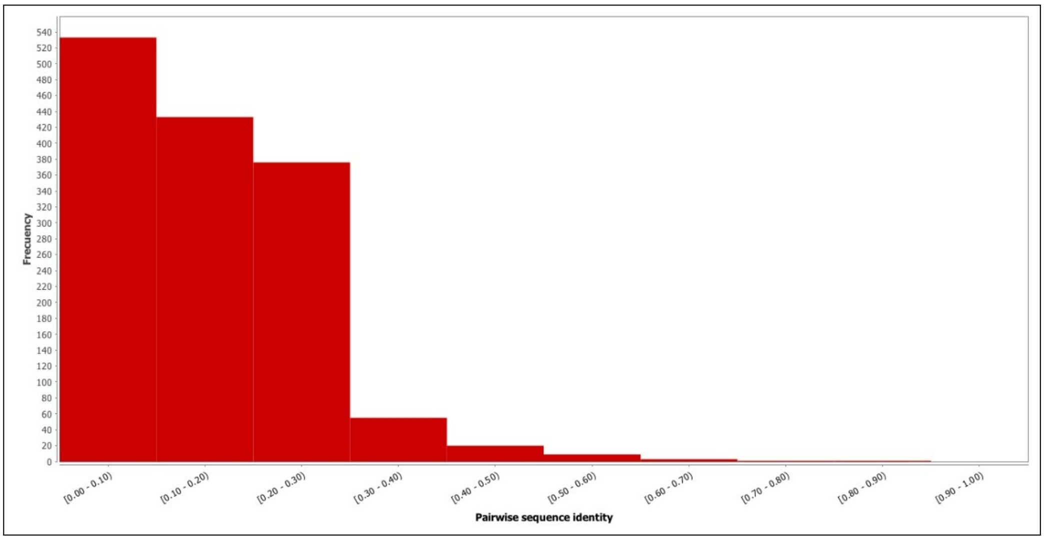


Figure 4.7: Histogram of pairwise sequence identity of SET 1 (54 lead compounds).

candidate against a Markov model built up from shuffling the same input set (control sequences). Table 4.11 displays the enriched motifs discriminating the 54 lead peptides against the control sequences. The same search was also performed by considering each cluster content. Motifs appearing in more than 20% of the query sequences are listed according to their statistical significance (score).

One of the motifs discovered by STREME had been reported as tumor homing, “WRP” which interacts with VEGF-C (Table 2.1) [82, 83]. Moreover, other found motifs were reported but not as TH, such as “WRPW”, “PRW”, “WKG”, and “PSHL”. “WRPW”, which contains “WRP” motif, is the binding site of 7 Enhancer of split E(spl) basic helix-loop-helix (bHLH) and Hairy proteins to the WD40 domain of corepressor

Table 4.11: Discovered Motifs by STREME.**Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).)

No.	Motif	Cluster	Cluster size	Matches in positive seqs.	Matches in control seq.	Sites (%)	Score	Frequency**
1	WRP	2	14	7	1	50	1.6e-002	5/(5)
2	WVL			5	1	35.7	8.2e-002	0/(0)
3	WS[YR]			3	0	21.4	1.1e-001	1/(1)Y
4	WWWM			3	0	21.4	1.1e-001	0/(0)
5	CFRV	3	10	3	0	30	1.1e-001	1/(1)
6	HWK			2	0	20	2.4e-001	0/(0)
7	PRW			2	0	20	2.4e-001	3/(3)
8	CN[WG]	4	8	3	0	37.5	1.0e-001	34/(32)G
9	WARG			3	0	37.5	1.0e-001	0/(0)
10	GIG			2	0	25.0	2.3e-001	5/(4)
11	WKG	1-5	4	3	1	75.0	2.4e-001	0/(0)
12	KNKHK	6	10	3	0	30.0	1.1e-001	0/(0)
13	PSHL			3	0	30.0	1.1e-001	0/(0)
14	LRLRI	Singletons	8	2	0	25.0	2.3e-001	1/(1)
15	CC[CQ]			3	1	37.5	2.8e-001	0/(0)
16	LSP	All sequences	54	11	1	20.4	3.4e-003	3/(3)
17	WSYG			7	0	13.0	8.2e-003	0/(0)
18	WRPW			5	0	9.3	3.2e-002	2/(2)

protein Groucho-TLE [189]. "PRW" is part of a biocatalyst, where it is conjugated to a lipid by an ester or amide bond [190]. "WKG" is a ribosomally synthesized and post-translationally modified peptide [191]. Moreover, "PSHL" is a tetrapeptide that affects the maturation and activation of HIV-1 protease (PR) [192].

Lastly, 54 lead THPs were queried against PROSITE's pattern and profile databases by using the search engine Motif Search of the GenomeNet suite [193]. Only two query peptides had significant matches with motifs found in Gonadotropin-releasing hormones (GnRH) and Bombesin-like peptides (Table 4.12).

Table 4.12: Motifs found in PROSITE. **Taken from TumorHoPe (outside parenthesis), and starPepDB (inside parenthesis).

No.	Motif found	Hit Peptide	Accesion	Match with	Signature	Related Seqs.	Frequency**
1	QHWSYGLRPG	starPep_07237	PS00473	Q[HY][FYW]Sx(4)PG	Gonadotropin-releasing hormones	67	1/(1)QHWSY
2	WARGHFM	starPep_10020	PS00257	WAxG[SH][LF]M	Bombesin-like peptides	36	0/(0)

These two peptide signatures and their receptors are involved in neuroendocrine signaling pathways associated with physiological states and tumors. GnRH is the hypothalamic decapeptide that plays a key role in the control of women's reproductive cycle. GnRH binds to specific receptors on the pituitary gonadotrophic cells, but it also is expressed in other reproductive organs, e.g. ovaries, and tumors derived from the ovaries. It has been shown GnRH is involved in the regulation of proliferation and metastasis of ovarian cancer either by indirect signaling pathway or direct interaction with the GnRH receptors placed at the surface of ovarian cancer cells [194].

Bombesin-like peptides were initially discovered from the frog skin, where they are secreted from cutaneous glands as a means of communication and defense. They were later found to be widely distributed in mammalian neural and endocrine cells represented by the neuromedin B (NMB) and the gastrin-releasing peptide (GRP), respectively. Bombesin-like peptide receptors are G protein-coupled and have seven membrane-spanning domains, so they are involved in signal transduction pathways [195]. Growing evidence shows bombesin-like peptides and their receptors play important roles in both physiological conditions and diseases. In fact, an abnormal expression of bombesin receptors has been observed in several types of tumors, which has motivated the development of more specific and safer bombesin-derivatives for tumors diagnosis and therapy [103].

The motif search by using different approaches may render a diversity of outcomes. However, some hits shared by different search approaches can support the reliability of the findings. For example, one motif "WSY" found by the PROSITE search was also encountered by STREME, an algorithm that works regardless of database and sequence similarity. Some of the motifs estimated by MSA algorithms were also identified by the AF approach STREME such as "WWW" and "WKG". All motifs were searched against TH databases, TumorHoPe, and starPepDB, in order to discriminate the possible new signatures from the existing ones (Table 2.1). New motifs appear at very low frequency contained within THPs (last column of Tables 4.10, 4.11 and 4.12), except "CNG" found by STREME, which appears 34 times in TumorHoPe and 32 in starPepDB. However, "CNG" has not been reported as TH motif.

4.3 Multi-Objective Optimization of THPs

4.3.1 Cell-Penetrating activity

Sequences from SET 1 show potent TH activity and singularity, but only 7 of them are permeable into cells. Improving their permeability was important to enhance their therapeutic activity due to it facilitates drug targeting. Thus, a library of 150 sequences (Attachments **K**) was obtained by punctual aa mutations, mainly to positively charged R or K residues, using CellPPD.

SET 1 and the library of 150 optimized sequences were combined and reduced by scaffold extractions and similarity searching. The stronghold is a SET 2 of 42 hits with TH scores between 0.19-3.61 according to TumorHPD, non-toxic by all models of ToxinPred, anticancer by at least one of the AntiCP models, and non-hemolytic by at least three models of HemoPI, where 34 hits are CPPs by at least one of CellPPD. Attachments **L** shows their predicted activities.

It was difficult to achieve cell-penetrating peptides according to both SVM models of CellPPD server while keeping the tumor homing activity, low toxicity and hemolysis due to mutations change considerably their structure affecting activity. Therefore, an alternative cell-penetrating optimization was performed based on the conjugation of the sequence with a well-known cell-penetrating sequence, such as TAT. The conjugation was applied as final step, after the selection of putative THPs, thus it is explained in detail later.

4.3.2 Half-Life Time

Half-life is a highly relevant parameter of therapeutic peptides since it governs the drug's pharmacokinetics, in consequence, the activity [196]. It directly influences the bioavailability, biodistribution, and necessary dosage of the drug. It is known that peptides show a short half-life in the gastrointestinal tract due to protease cleavage, which is the main reason why the preferred route of administration is parental [197]. Nevertheless, peptides present a short half-life in circulation as a result of enzymatic degradation but also renal clearance [33]. Therefore, it is essential to evaluate the theoretical half-life in the blood and digestive system of THPs and try to improve their stability, mainly by aa mutations.

Half-life in the blood of 42 sequences was improved by punctual mutations and shortening in fragments of 5 and 10 residues in PlifePred. A library of 206 sequences (Attachments **M**) with tumor homing activity, non-toxic, and non-hemolytic was obtained. After the application of scaffold extractions and similarity searching in the peptides from the library, 59 sequences (SET 3) with higher stability in blood than originals were retrieved. Sequences from SET 3 have predicted half-life time in blood between 13 to 33 minutes.

On the other hand, the stability of 59 sequences in the gastrointestinal tract was optimized using HLP by punctual mutations. The selection of sequences was based on those with predicted stability labeled as high (higher than 1 second) or normal (0.1-1 second) according to the server. Sequences with similar or higher TH scores than originals, anticancer activity, non-toxic and non-hemolytic were filtered, resulting in a library of 250 sequences (Attachments **N**). From the library, 78 sequences were retrieved (SET 4) after scaffold extractions and similarity searching. These sequences are TH active by the two servers (THP and TumorHPD) with a score between 0.46-3.30, non-toxic, the majority are anticancer by almost one SVM model of AntiCP, have a half-life in blood between 11-30 minutes with normal or high (0.799-6.599 seconds) gastrointestinal stability. Moreover, sequences from SET 4 were characterized by predicting other properties, including aggregation, and allergenic reaction (Attachments **O**).

To find a compromise between blood and gastrointestinal half-life time was not easy since neutral residues (E, S, T, and G) stabilized peptides in the gastrointestinal [141], while they decrease half-life in blood [28, 198]. Moreover, the proposed THPs was short sequences and small structural alterations modify their activity.

Although the stability of the sequences has been slightly increased, predicted half-life times are not adequate, considering that the chemotherapeutic drug needs a half-life of more than hours to days to decrease the number of doses administered to the patient [199]. Thus, parental administration is preferable.

On the other hand, the majority of sequences lost penetrability. The main reason was cell penetrability improves by increasing positively charged residues such as R and K, but with increasing charge, the sequence is more prone to degradation by the action of proteases [141].

Finally, the 78 sequences were reduced to 13 lead peptides (SET 5) (highlighted sequences in Attachments **O**) that accomplished SVM scores shown in Table 3.1. Compared to the previously obtained SET 1, sequences from SET 5 have a higher potential to be THPs because scoring thresholds were more stringent. Nevertheless, the half-life is still low and a limiting factor in their pharmacokinetics. Thus, other optimization routes should be sought, such as increasing molecular weight or conjugation with stability enhancers [200].

4.3.3 Building Blocks

In a final attempt to enhance the tumor homing activity of SET 4, the reported motifs (Table 2.1), discovered in this study (Tables 4.10, 4.11 and 4.12), and the short sequences of SET 4 shown in Table 4.13 were attached and inserted to the 78 sequences from SET 4 and to motifs (Tables 2.1, 4.10, 4.11 and 4.12).

Table 4.13: Short optimized THPs with 5-8 aa length.

No	Sequence	SVM Score
1	WPGCHSWA	3.12
2	CSKGC	3.01
3	CRPGC	2.94
4	CRCGF	2.92
5	PYWLP	2.82
6	CPCKL	2.65
7	WRQLPWFG	2.52
8	PLSWA	2.5
9	AFPSWRM	2.36
10	AMDSRWM	2.22
11	YWRGF	2.07
12	ECGFW	2.03

In total, a library of 7923 sequences was built by the addition of motifs in the N-end, C-end, and 3 random positions. This large library was reduced to 62 lead peptides (SET 6), where all are TH active by the ML servers, and 13 are TH active by the THP1 model.

On the other hand, the library was filtered keeping only the non-redundant sequences with cell-penetrating activity, deriving 10 peptides (SET 7).

4.3.4 Final Selection of Potential THPs

At this point, there are 3 sets of lead sequences with optimized activities (SET 5, 6, and 7), totaling 85 sequences. As they were obtained in different steps, they presented redundancy among the sequences. Therefore, redundant sequences were removed only from SET 5 and 6 by scaffold extraction, resulting in a set of 39 lead sequences. SET 7 was not reduced because they were sequences with high cell-penetrating potential according to both SVM models from CellPPD.

Finally, the 39 lead sequences were combined with SET 7, and peptides with the highest TH activity (highest ML scores) and the trade-off between all the different predicted activities were filtered. The result was a set of positively charged 27 potential THPs (SET: Putative THPs).

Predicted activities are shown in Attachments **P**. The range of size of 27 putative THPs is between 7-11 aa residues. In general, their physicochemical properties show a low score of hydropathicity meaning that they were more hydrophilic, positively influencing solubility [201]. They are tumor homing according to the prediction of all SVM methods with a score between 1.16-3.34. Additionally, they are non-toxic, non-hemolytic, and they are anticancer according to all SVM models. According to the immunogenicity of the compounds, the scores that indicate how much allergic reaction they produce is less than 0.416, and they do not induce IL-10, an interleukin that has both tumor-inhibitory and tumor-promoting activity [202]. The scores that determine IL-4 induction are low, however, the production of this type of interleukin has an antiangiogenic effect and favors tumor cell growth inhibition [203]. On the contrary, predictions show that all the molecules are inducers of IL-12, an interleukin widely studied for its immunotherapy potential in cancer treatments because it mediates tumor regression [203]. In addition, they may exhibit other activities such as antimicrobial or antiviral agents.

Putative THPs showed simple structures and were classified as random coiled or single helix (Figure 4.8). The next step would be to test their tumor homing activity experimentally, to corroborate their potential.

De novo Design of ACPs

From the 27 sequences, 14 sequences with higher anticancer scores, and commitment to TH scores, low toxicity, low hemolysis, and high solubility were prioritized. 8 of them were mutated by ROSE to build an extended peptide library as a source of finding new ACPs with enhanced cell permeability. Then, the resulting libraries were screened to identify putative ACPs by using the several ML-based programs trained to identify some

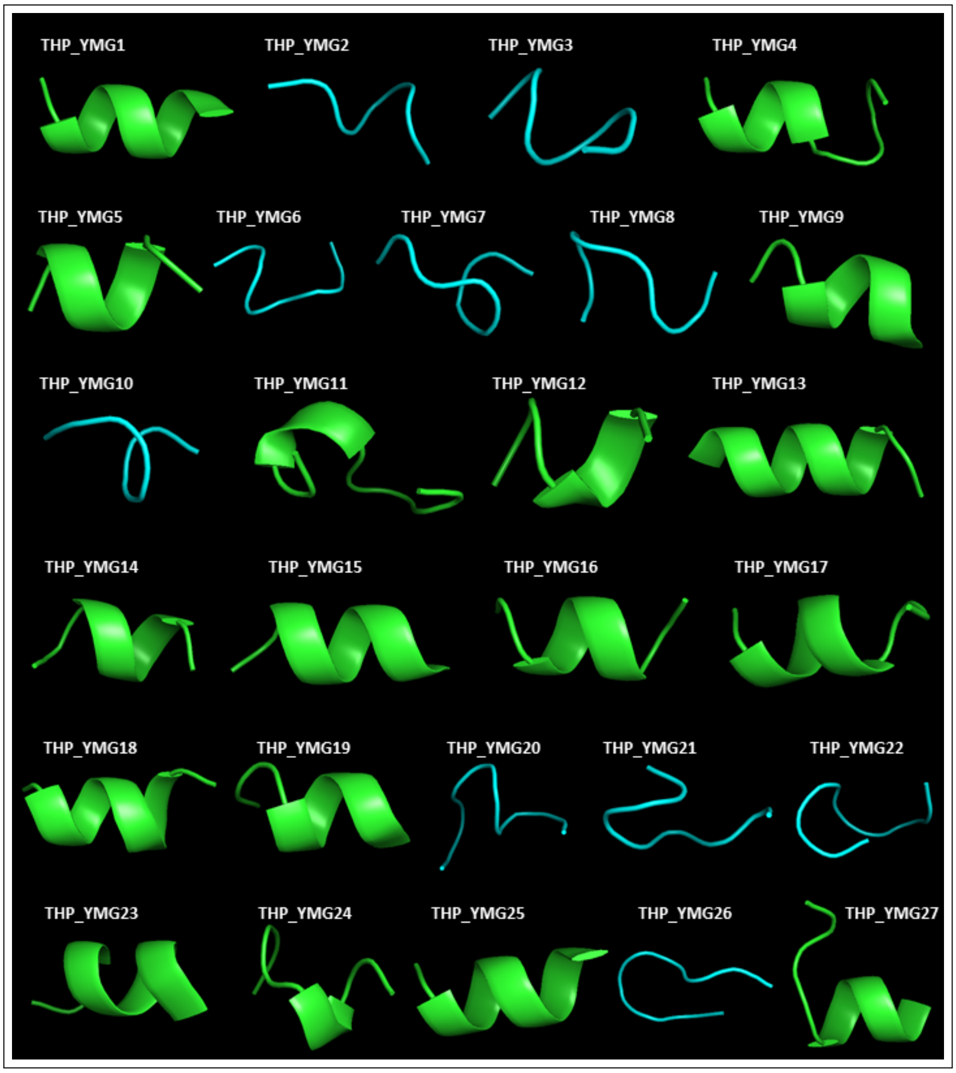


Figure 4.8: 3D structure of 27 putative THPs generated with PEP-FOLD 3 and visualized with PyMOL.

therapeutic peptides (Table 2.2). Finally, among the top-ranked candidates, i.e., those with higher anticancer scores according to AntiCP, lower human toxicity, and higher % of sequence similarity to ancestor peptides, the most potent and orthogonal 8 sequences were selected.

In general, the application of evolutionary approaches has been devoted to the optimization steps of peptide drugs. Here, a different approach for the design of bioactive peptides is proposed, which also leverages ML models and evolutionary algorithms but in a different mode. The strategy repurposes the simulation of sequences evolution to the rational generation of diversity-oriented peptide libraries that are subsequently explored with ML models of several pharmacological and ADME-TOX endpoints. This is achieved by applying a flexible evolutionary algorithm, as implemented in ROSE, that comprises parameters such as average genetic distance, tree topology, and insertion and deletion events, among others. The advantage of using evolutionary algorithms to build libraries of candidates lies in the application of previous knowledge on the sites/residues that account for biological activity when mutations are performed. Thus, a consensus (root) peptide with its corresponding conservation scoring profile can be used to assign different mutation rates to each position in the sequence.

On the other hand, the sequence diversity of the peptides in the library can be controlled by evaluating the ROSE output with an all vs all global alignment [204]. Here, ROSE parameters were calibrated to produce peptide libraries with an overall 60% of identity by using the software starPep toolbox. All these evolutionary considerations provide rationality to the generation of peptide libraries. Thus, the probability to find new biologically relevant peptides is higher than approaches using stochastic mutations. The resulting peptide library was screened with several web servers to identify putative ACPs and, at the same time, to diminish the likelihood of action with the human counterpart as well ADME properties like cell permeability.

Table 4.14 shows the physicochemical properties of ACPs. Moreover, Table 4.15 summarizes the obtained scores for the 14 putative ACPs and Figure 4.9 shows their 3D structures generated with PEP-FOLD 3. Notably, ACP-YMG1 and ACP-YMG10, which is derived from one of the motif WRP shown in Table 2.1, is predicted to potentially bind VEGF-C. Similarly, ACP-YMG12 is predicted to bind to the tumor neovasculature, when precisely this peptide originates from the motif PSP. The other structures do not have known motifs, so *a priori* it is not possible to determine what they would be bound to and require experimental studies. Putative ACPs also showed a simple structure that can be classified as the random coiled or single helix.

Table 4.14: Physicochemical properties of 14 putative ACPs.

CodeID	**Sequence**	**Length**	**Hydrophobicity**	**Steric hindrance**	**Sidebulk**	**Hydropathicity**	**Amphipathicity**	**Hydrophilicity**	**Net Hydrogen**	**Charge**	**pI**	**Mol wt**
ACP_YMG1	WRPWPSHL	8	-0.14	0.38	0.38	-1.08	0.43	-0.64	0.89	1.5	10.11	1191.53
ACP_YMG2	EKFWPRSG	8	-0.3	0.53	0.53	-1.42	0.82	0.38	1	1	9.1	1063.3
ACP_YMG3	PRWPLSWA	8	-0.07	0.44	0.44	-0.52	0.27	-0.64	0.78	1	10.11	1083.37
ACP_YMG4	HHGTPRWC	8	-0.25	0.37	0.37	-1.33	0.59	-0.31	0.89	2	8.61	1096.37
ACP_YMG5	PSPAFKWW	8	0.01	0.46	0.46	-0.57	0.41	-0.72	0.56	1	9.11	1204.51
ACP_YMG6	AMYWRGFWWP	10	0.05	0.54	0.54	-0.36	0.22	-1.25	0.73	1	9.1	1496.9
ACP_YMG7	CTGCQNWWM	9	-0.03	0.57	0.57	-0.3	0.12	-1.01	0.7	0	5.82	1259.64
ACP_YMG8	WWYWRGFWM	9	0.08	0.55	0.55	-0.51	0.25	-1.67	0.9	1	9.1	1548.99
ACP_YMG9	LPWCKRLRT	9	-0.34	0.51	0.51	-0.6	0.86	0.06	1.2	3	10.87	1273.7
ACP_YMG10	WRPGSWAKQALKSI	14	-0.17	0.54	0.54	-0.62	0.74	-0.11	0.93	3	11.17	1741.29
ACP_YMG11	CYLHSSSCGSCHNCK	15	-0.17	0.5	0.5	-0.31	0.41	-0.29	0.69	2	8.17	1757.22
ACP_YMG12	SNWWRLKT	8	-0.3	0.52	0.52	-1.27	0.68	-0.28	1.33	2	11.01	1191.48
ACP_YMG13	GKWARGW	7	-0.19	0.53	0.53	-1.15	0.77	-0.16	1	2	11.01	1046.31
ACP_YMG14	RQRICVPRR	9	-0.65	0.58	0.58	-1.19	1.1	0.79	1.8	4	12.01	1339.76

On the other hand, when adding TAT and A-TAT, permeability was enhanced. How-

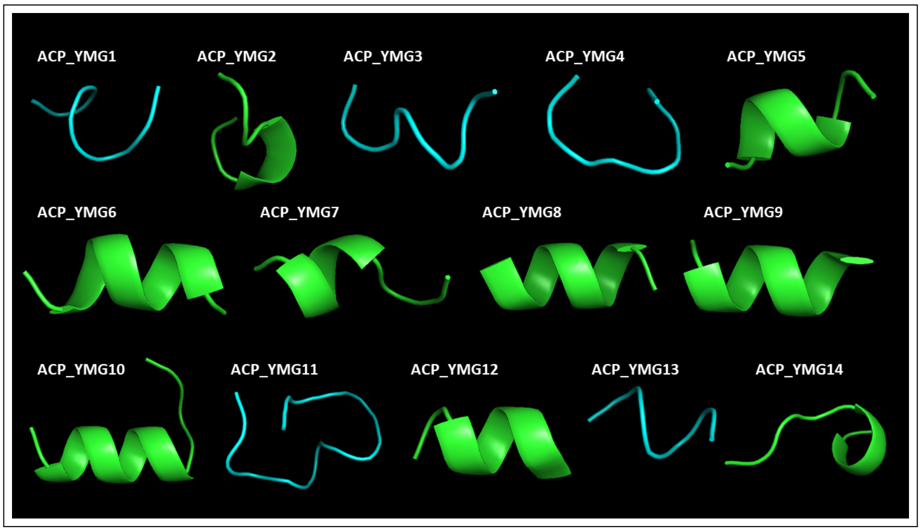


Figure 4.9: 3D structure of 15 putative ACPs generated with PEP-FOLD 3 and visualized with PyMOL.

ever, the predicted tumor homing action was considerably decreased. When comparing the 3D structures of sequences with and without TAT, it could be observed that in some sequences the peptide conformational structure changed subtly when A-TAT was added and not when only TAT was added, but different domains are not formed. Therefore, the peptides are expected to maintain the predicted activities when bound to TAT. Nevertheless, it requires further studies.

Table 4.15: Tumor homing and anticancer predictions of 14 putative ACPs.

CodeID	**Sequence**	**TumorHPD SVM Score**	**THPep**	**AntiCP SVM Scores**		**iACP Prob**	**CpACpP Prediction**						**ACPred Prob**	**ENNAACT ACP Prob**	**AMPfun Score**
ACP_YMG1	WRPWPSHL	3.28	THP	1.11	0.78	0.147606	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.977	0.776	0.4404
ACP_YMG2	EKFWPRSG	0.39	THP	1.06	0	0.965416	non-ACP	non-CpACP	ACP	CpACP	ACP	CpACP	0.988	0.095	0.4725
ACP_YMG3	PRWPLSWA	3.14	THP	1.02	0.87	0.959603	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.819	0.095	0.385
ACP_YMG4	HHGTPRWC	2.07	THP	1.02	1.2	0.959603	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.965	0.775	0.5314
ACP_YMG5	PSPAFKWW	2.07	THP	1.05	0.88	0.765597	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.989	0.574	0.3503
ACP_YMG6	AMYWRGFWWP	3.02	THP	1.04	1.24	0.525148	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.98	0.585	0.6715
ACP_YMG7	CTGCQNWWM	2.34	THP	1.16	0.91	0.701894	ACP	non-CpACP	ACP	non-CpACP	ACP	non-CpACP	0.954	0.653	0.4235
ACP_YMG8	WWYWRGFWM	2.78	THP	0.76	0.81	0.959603	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.989	0.993	0.6427
ACP_YMG9	LPWCKRLRT	1.63	THP	1.09	1.41	0.482859	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.084	0.998	0.3624
ACP_YMG10	WRPGSWAKQALKSI	0.19	THP	0.85	1.01	0.874125	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.984	0.85	0.0375
ACP_YMG11	CYLHSSSCGSCHNCK	2.51	THP	0.64	0.98	0.682829	ACP	non-CpACP	ACP	non-CpACP	ACP	non-CpACP	0.994	0.736	0.5205
ACP_YMG12	SNWWRLKT	1.49	THP	0.9	0.23	0.959603	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.957	0.483	0.3282
ACP_YMG13	GKWARGW	2.06	THP	-0.04	1.41	0.985151							0.987	0.997	0.2661
ACP_YMG14	RQRICVPRR	0.33	THP	0.81	0.44	0.835487	ACP	CpACP	ACP	CpACP	ACP	CpACP	0.908	0.453	0.4703

Chapter 5

Conclusions and Recommendations

5.1 Conclusions

In this study, a novel methodology based on network science and similarity searching was proposed and applied to explore the chemical space of THPs and discover potential THPs. Statistically, the performance of the strategy transcends current supervised ML approaches used in THPs predictions, demonstrating the potential of this alternative unsupervised approach. Hence, *in silico* predictions using the model based on representative THPs in conjunction with TumorHPD and THPep give high reliability to discover potential THPs. Herein, 54 lead compounds are repurposed as potential THPs that were obtained using the method in the starPep toolbox, followed by activity optimization using TumorHPD. In the set, novel motifs with tumor homing activity are proposed. Moreover, 54 lead molecules were subjected to punctual mutations and sequence shortening in order to find molecules with greater stability, and to enhance their tumor homing activity, identifying 27 putative THPs. In addition, a *de novo* design of ACPs was described, using evolutionary algorithms to find sequences that concentrate in tumor tissue, and have anticancer activity at the same time, where 14 ACPs were derived. The two sets of 27 THPs and 14 ACPs present a diversity structure, and would evolve the currently known chemical space of THPs.

5.2 Recommendations

This study is based on *in silico* approaches, consequently, biological assays are required to validate the tumor homing and anticancer activity. Once the activity of the peptides has been validated, it is recommended to optimize their pharmacokinetics, particularly their stability in blood, using other methodologies, such as PEGylation. On the other hand, the good performance of the methodology for predicting peptide activity based on similarity searching and network science suggests its application for the prediction of other endpoints in peptides, e.g. antimicrobial activity, toxicity, hemolytic, or anticancer.

References

(1) World Health Organization (WHO). Cancer.

(2) Miller, K. D.; Nogueira, L.; Mariotto, A. B.; Rowland, J. H.; Yabroff, K. R.; Alfano, C. M.; Jemal, A.; Kramer, J. L.; Siegel, R. L. *CA: A Cancer Journal for Clinicians* **2019**, *69*, 363–385.

(3) Hoskin, D. W.; Ramamoorthy, A. *Biochimica et Biophysica Acta - Biomembranes* **2008**, *1778*, 357–375.

(4) Loffet, A. *Journal of Peptide Science* **2002**, *8*, 1–7.

(5) Segura-Campos, M.; Chel-Guerrero, L.; Betancur-Ancona, D.; Hernandez-Escalante, V. M. *Food Reviews International* **2011**, *27*, 213–226.

(6) Wu, D.; Gao, Y.; Qi, Y.; Chen, L.; Ma, Y.; Li, Y. *Cancer Letters* **2014**, *351*, 13–22.

(7) Wei, G.; Wang, Y.; Huang, X.; Hou, H.; Zhou, S. *Small Methods* **2018**, *2*, 1–16.

(8) Khandia, R.; Sachan, S.; K. Munjal, A.; Tiwari, R.; Dhama, K. In *Topics in Anti-Cancer Research*; BENTHAM SCIENCE PUBLISHERS: 2016, pp 43–86.

(9) Sharma, A.; Kapoor, P.; Gautam, A.; Chaudhary, K.; Kumar, R.; Chauhan, J. S.; Tyagi, A.; Raghava, G. P. *Scientific Reports* **2013**, *3*, 1–7.

(10) Cui, W.; Aouidate, A.; Wang, S.; Yu, Q.; Li, Y.; Yuan, S. *Frontiers in Pharmacology* **2020**, *11*, 1–14.

(11) Li, X.; Cai, H.; Wu, X.; Li, L.; Wu, H.; Tian, R. *Frontiers in Chemistry* **2020**, *8*, 1–19.

(12) Tu, M.; Cheng, S.; Lu, W.; Du, M. *TrAC Trends in Analytical Chemistry* **2018**, *105*, 7–17.

(13) Kapoor, P.; Singh, H.; Gautam, A.; Chaudhary, K.; Kumar, R.; Raghava, G. P. *PLoS ONE* **2012**, *7*, DOI: `10.1371/journal.pone.0035187`.

(14) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. *Bioinformatics* **2019**, *35*, ed. by Wren, J., 4739–4747.

(15) Shoombuatong, W.; Schaduangrat, N.; Pratiwi, R.; Nantasenamat, C. *Computational Biology and Chemistry* **2019**, *80*, 441–451.

(16) Nelson, D. L.; Cox, M. M., *Principles of Biochemistry*, New York, 2013.

(17) Tesauro, D.; Accardo, A.; Diaferia, C.; Milano, V.; Guillon, J.; Ronga, L.; Rossi, F. *Molecules* **2019**, *24*, 351.

(18) Fosgerau, K.; Hoffmann, T. *Drug Discovery Today* **2015**, *20*, 122–128.

(19) Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. *Drug Discovery Today* **2010**, *15*, 40–56.

(20) Chen, L.; Patrone, N.; Liang, J. F. *Biomacromolecules* **2012**, *13*, 3327–3333.

(21) Merrifield, R. B. *Journal of the American Chemical Society* **1963**, *85*, 2149–2154.

(22) Lau, J. L.; Dunn, M. K. *Bioorganic & Medicinal Chemistry* **2018**, *26*, 2700–2707.

(23) Albericio, F.; Kruger, H. G. *Future Medicinal Chemistry* **2012**, *4*, 1527–1531.

(24) De la Torre, B. G.; Albericio, F. *Molecules* **2020**, *25*, 2019–2021.

(25) Ladner, R. C.; Sato, A. K.; Gorzelany, J.; De Souza, M. *Drug Discovery Today* **2004**, *9*, 525–529.

(26) Pichereau, C.; Allary, C. *EBR - European Biopharmaceutical Review* **2005**, 88–93.

(27) Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. *EXCLI Journal* **2018**, *17*, 734–752.

(28) Mathur, D.; Singh, S.; Mehta, A.; Agrawal, P.; Raghava, G. P. *PLoS ONE* **2018**, *13*, 1–10.

(29) Dwyer, J. J.; Wilson, K. L.; Davison, D. K.; Freel, S. A.; Seedorff, J. E.; Wring, S. A.; Tvermoes, N. A.; Matthews, T. J.; Greenberg, M. L.; Delmedico, M. K. *Proceedings of the National Academy of Sciences* **2007**, *104*, 12772–12777.

(30) Nguyen, L. T.; Chau, J. K.; Perry, N. A.; de Boer, L.; Zaat, S. A. J.; Vogel, H. J. *PLoS ONE* **2010**, *5*, ed. by Vij, N., e12684.

(31) Pang, H. B.; Braun, G. B.; She, Z. G.; Kotamraju, V. R.; Sugahara, K. N.; Teesalu, T.; Ruoslahti, E. *Journal of Controlled Release* **2014**, *175*, 48–53.

(32) Wu, Y.-L.; Huang, J.; Xu, J.; Liu, J.; Feng, Z.; Wang, Y.; Lai, Y.; Wu, Z.-R. *Regulatory Peptides* **2010**, *164*, 83–89.

(33) Werle, M.; Bernkop-Schnürch, A. *Amino Acids* **2006**, *30*, 351–367.

(34) Jenssen, H.; Aspmo, S. I. In *Methods in Molecular Biology*, 2008; Vol. 494, pp 177–186.

(35) Hamamoto, K.; Kida, Y.; Zhang, Y.; Shimizu, T.; Kuwano, K. *Microbiology and Immunology* **2002**, *46*, 741–749.

(36) Hanahan, D.; Weinberg, R. A. *Cell* **2000**, *100*, 57–70.

(37) Ruoslahti, E. *Advanced Drug Delivery Reviews* **2017**, *110-111*, 3–12.

(38) Sharma, M.; El-Sayed, N. S.; Do, H.; Parang, K.; Tiwari, R. K.; Aliabadi, H. M. *Scientific Reports* **2017**, *7*, 1–14.

(39) Alberici, L.; Roth, L.; Sugahara, K. N.; Agemy, L.; Kotamraju, V. R.; Teesalu, T.; Bordignon, C.; Traversari, C.; Rizzardi, G. P.; Ruoslahti, E. *Cancer Research* **2013**, *73*, 804–812.

(40) Kunda, N. K. *Drug Discovery Today* **2020**, *25*, 238–247.

(41) Huang, W.; Seo, J.; Willingham, S. B.; Czyzewski, A. M.; Gonzalgo, M. L.; Weissman, I. L.; Barron, A. E. *PLoS ONE* **2014**, *9*, ed. by Afarinkia, K., e90397.

(42) Hosseinzadeh, E.; Banaee, N.; Nedaie, H. A. *Current Cancer Therapy Reviews* **2017**, *13*, 17–27.

(43) Amit, D.; Hochberg, A. *Journal of Translational Medicine* **2010**, *8*, 134.

(44) Gatti, L.; Zunino, F. In *Chemosensitivity*; Humana Press: New Jersey, 2005; Vol. 111, pp 127–148.

(45) Kurrikoff, K.; Aphkhazava, D.; Langel, Ü. *Current Opinion in Pharmacology* **2019**, *47*, 27–32.

(46) Xiao, Y.-F.; Jie, M.-M.; Li, B.-S.; Hu, C.-J.; Xie, R.; Tang, B.; Yang, S.-M. *Journal of Immunology Research* **2015**, *2015*, 1–13.

(47) Reche, P. A.; Fernandez-Caldas, E.; Flower, D. R.; Fridkis-Hareli, M.; Hoshino, Y. *Journal of Immunology Research* **2014**, *2014*, 1–2.

(48) Danhier, F.; Le Breton, A.; Préat, V. *Molecular Pharmaceutics* **2012**, *9*, 2961–2973.

(49) Langel, Ü. *Cell-Penetrating Peptides: Methods and Protocols* **2015**, *1324*, 1–468.

(50) Hu, C.; Chen, X.; Huang, Y.; Chen, Y. *Scientific Reports* **2018**, *8*, 1–14.

(51) Laakkonen, P.; Vuorinen, K. *Integrative Biology* **2010**, *2*, 326–337.

(52) Elsabahy, M.; Shrestha, R.; Clark, C.; Taylor, S.; Leonard, J.; Wooley, K. L. *Nano Letters* **2013**, *13*, 2172–2181.

(53) Liu, C.; Yang, Y.; Chen, L.; Lin, Y. L.; Li, F. *Journal of Biological Chemistry* **2014**, *289*, 34520–34529.

(54) Pleiko, K.; Põšnograjeva, K.; Haugas, M.; Paiste, P.; Tobi, A.; Kurm, K.; Riekstina, U.; Teesalu, T. *Nucleic Acids Research* **2021**, *49*, E38–E38.

(55) D'Onofrio, N.; Caraglia, M.; Grimaldi, A.; Marfella, R.; Servillo, L.; Paolisso, G.; Balestrieri, M. L. *Biochimica et Biophysica Acta - Reviews on Cancer* **2014**, *1846*, 1–12.

(56) Nemudraya, A. A.; Richter, V. A.; Kuligina, E. V. *Acta naturae* **2016**, *8*, 48–57.

(57) Arap, M. A. *Genetics and Molecular Biology* **2005**, *28*, 1–9.

(58) Wu, C.-H.; Liu, I.-J.; Lu, R.-M.; Wu, H.-C. *Journal of Biomedical Science* **2016**, *23*, 8.

(59) Ahmed, S.; Mathews, A. S.; Byeon, N.; Lavasanifar, A.; Kaur, K. *Analytical Chemistry* **2010**, *82*, 7533–7541.

(60) Meng, J.; Nan, M.; Yan, Z.; Han, W.; Zhang, Y. *Journal of Biochemistry* **2006**, *140*, 299–304.

(61) Aoki, Y.; Hosaka, S.; Kawa, S.; Kiyosawa, K. *Cancer Gene Therapy* **2001**, *8*, 783–787.

(62) Pasqualini, R.; Koivunen, E.; Kain, R.; Lahdenranta, J.; Stryhn, A.; Ashmun, R. A.; Shapiro, L. H.; Arap, W. **2000**, *60*, 722–727.

(63) Arap, W. *Science* **1998**, *279*, 377–380.

(64) Dijkgraaf, I.; Van de Vijver, P.; Dirksen, A.; Hackeng, T. M. *Bioorganic & Medicinal Chemistry* **2013**, *21*, 3555–3564.

(65) Ruoslahti, E. I.; Pasqualini, R.; Arap, W.; Bredesen, D. E.; Ellerby, H. M. Chimeric prostate-homing peptides with pro-apoptotic activity, 2001.

(66) Jullienne, B.; Vigant, F.; Muth, E.; Chaligné, R.; Bouquet, C.; Giraudier, S.; Perricaudet, M.; Benihoud, K. *Gene Therapy* **2009**, *16*, 1405–1415.

(67) Pastorino, F.; Brignole, C.; Marimpietri, D.; Cilli, M.; Gambini, C.; Ribatti, D.; Longhi, R.; Allen, T. M.; Corti, A.; Ponzoni, M. *Cancer Research* **2003**, *63*, 7400–7409.

(68) Kumar, A.; Ma, H.; Zhang, X.; Huang, K.; Jin, S.; Liu, J.; Wei, T.; Cao, W.; Zou, G.; Liang, X. J. *Biomaterials* **2012**, *33*, 1180–1189.

(69) Liu, D.; Wang, C.; Yang, J.; An, Y.; Yang, R.; Teng, G. *ACS Omega* **2020**, *5*, 9316–9323.

(70) Samanta, S.; Sistla, R.; Chaudhuri, A. *Biomaterials* **2010**, *31*, 1787–1797.

(71) Sugahara, K. N.; Teesalu, T.; Karmali, P. P.; Kotamraju, V. R.; Agemy, L.; Girard, O. M.; Hanahan, D.; Mattrey, R. F.; Ruoslahti, E. *Cancer Cell* **2009**, *16*, 510–520.

(72) Kolonin, M. G. et al. *Cancer Research* **2006**, *66*, 34–40.

(73) He, X.; Na, M.-h.; Kim, J.-S.; Lee, G.-Y.; Park, J. Y.; Hoffman, A. S.; Nam, J.-o.; Han, S.-e.; Sim, G. Y.; Oh, Y.-k.; Kim, I.-S.; Lee, B.-h. *Molecular Pharmaceutics* **2011**, *8*, 430–438.

(74) Peletskaya, E. N.; Glinsky, V. V.; Glinsky, G. V.; Deutscher, S. L.; Quinn, T. P. *Journal of Molecular Biology* **1997**, *270*, 374–384.

(75) Ruoslahti, E.; Pasqualini, R. Tumor Homing molecules, conjugates derived therefrom, and methods of using same, 1998.

(76) Wang, F.; Li, Y.; Shen, Y.; Wang, A.; Wang, S.; Xie, T. *International journal of molecular sciences* **2013**, *14*, 13447–13462.

(77) Kwong, J.; Kulbe, H.; Wong, D.; Chakravarty, P.; Balkwill, F. *Molecular Cancer Therapeutics* **2009**, *8*, 1893–1905.

(78) Wong, D.; Kandagatla, P.; Korz, W.; Chinni, S. R. *BMC Urology* **2014**, *14*, DOI: `10.1186/1471-2490-14-12`.

(79) Nazemian, M.; Hojati, V.; Zavareh, S.; Madanchi, H.; Hashemi-Moghaddam, H. *International Journal of Peptide Research and Therapeutics* **2020**, *26*, 259–269.

(80) Urbanelli, L.; Ronchini, C.; Fontana, L.; Menard, S.; Orlandi, R.; Monaci, P. *Journal of Molecular Biology* **2001**, *313*, 965–976.

(81) Qin, X.; Wan, Y.; Li, M.; Xue, X.; Wu, S.; Zhang, C.; You, Y.; Wang, W.; Jiang, C.; Liu, Y.; Zhu, W.; Ran, Y.; Zhang, Z.; Han, W.; Zhang, Y. *Journal of Biochemistry* **2007**, *142*, 79–85.

(82) Asai, T.; Nagatsuka, M.; Kuromi, K.; Yamakawa, S.; Kurohane, K.; Ogino, K.; Tanaka, M.; Taki, T.; Oku, N. *FEBS Letters* **2002**, *510*, 206–210.

(83) Oku, N.; Asai, T.; Watanabe, K.; Kuromi, K.; Nagatsuka, M.; Kurohane, K.; Kikkawa, H.; Ogino, K.; Tanaka, M.; Ishikawa, D.; Tsukada, H.; Momose, M.; Nakayama, J.; Taki, T. *Oncogene* **2002**, *21*, 2662–2669.

(84) Kelly, K. A.; Jones, D. A. *Neoplasia* **2003**, *5*, 437–444.

(85) Lee, Y. M.; Lee, D.; Kim, J.; Park, H.; Kim, W. J. *Journal of Controlled Release* **2015**, *205*, 172–180.

(86) Gray, B. P.; Brown, K. C. *Chemical Reviews* **2014**, *114*, 1020–1081.

(87) Brown, K. *Current Pharmaceutical Design* **2010**, *16*, 1040–1054.

(88) Hwang, Y. J.; Myung, H. *Frontiers in Microbiology* **2020**, *11*, DOI: `10.3389/fmicb.2020.491001`.

(89) Dabrowska, K et al. *Acta virologica* **2004**, *48*, 241–8.

(90) Lee, T. Y.; Lin, C. T.; Kuo, S. Y.; Chang, D. K.; Wu, H. C. *Cancer Research* **2007**, *67*, 10959–10965.

(91) Herringson, T. P.; Altin, J. G. *Journal of Drug Targeting* **2011**, *19*, 681–689.

(92) Ruczynski, J.; Wierzbicki, P. M.; Kogut-Wierzbicka, M.; Mucha, P.; Siedlecka-Kroplewska, K.; Rekowski, P. *Folia Histochemica et Cytobiologica* **2015**, *52*, 257–269.

(93) Brunel, F. M.; Liu, F.; Mayer, J. P. *Successful Drug Discovery* **2019**, *4*, 3–34.

(94) Araste, F.; Abnous, K.; Hashemi, M.; Taghdisi, S. M.; Ramezani, M.; Alibolandi, M. *Journal of Controlled Release* **2018**, *292*, 141–162.

(95) Desale, K.; Kuche, K.; Jain, S. *Biomaterials Science* **2021**, *9*, 1153–1188.

(96) Gessner, I.; Neundorf, I. *International Journal of Molecular Sciences* **2020**, *21*, 1–21.

(97) Myrberg, H.; Zhang, L.; Mäe, M.; Langel, Ü. **2008**, 70–75.

(98) Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nature Materials* **2009**, *8*, 543–557.

(99) Guo, Z.; Peng, H.; Kang, J.; Sun, D. *Biomedical Reports* **2016**, *4*, 528–534.

(100) Schmid, S. L.; Conner, S. D. *Nature* **2003**, *422*, 37–44.

(101) Ma, D. X.; Shi, N. Q.; Qi, X. R. *International Journal of Pharmaceutics* **2011**, *419*, 200–208.

(102) Sharma, S.; Kotamraju, V. R.; Mölder, T.; Tobi, A.; Teesalu, T.; Ruoslahti, E. *Nano Letters* **2017**, *17*, 1356–1364.

(103) Guo, M.; Qu, X.; Qin, X. Q. *Current Opinion in Endocrinology, Diabetes and Obesity* **2015**, *22*, 3–8.

(104) Liang, D. S.; Su, H. T.; Liu, Y. J.; Wang, A. T.; Qi, X. R. *Biomaterials* **2015**, *71*, 11–23.

(105) Hu, Q.; Gao, X.; Gu, G.; Kang, T.; Tu, Y.; Liu, Z.; Song, Q.; Yao, L.; Pang, Z.; Jiang, X.; Chen, H.; Chen, J. *Biomaterials* **2013**, *34*, 5640–5650.

(106) Feng, X.; Jiang, D.; Kang, T.; Yao, J.; Jing, Y.; Jiang, T.; Feng, J.; Zhu, Q.; Song, Q.; Dong, N.; Gao, X.; Chen, J. *ACS Applied Materials and Interfaces* **2016**, *8*, 17817–17832.

(107) Miao, D.; Jiang, M.; Liu, Z.; Gu, G.; Hu, Q.; Kang, T.; Song, Q.; Yao, L.; Li, W.; Gao, X.; Sun, M.; Chen, J. *Molecular Pharmaceutics* **2014**, *11*, 90–101.

(108) Teesalu, T.; Sugahara, K. N.; Ruoslahti, E. *Frontiers in Oncology* **2013**, *3 AUG*, 1–8.

(109) Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. *Medicinal Research Reviews* **2020**, *40*, 1276–1314.

(110) Zhang, X.-D., *Chapter 6 Machine Learning*; 13, 2017; Vol. 45, pp 40–48.

(111) Attique, M.; Farooq, M. S.; Khelifi, A.; Abid, A. *IEEE Access* **2020**, *8*, 148570–148594.

(112) Joseph, S.; Karnik, S.; Nilawe, P.; Jayaraman, V. K.; Idicula-Thomas, S. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2012**, *9*, 1535–1538.

(113) Dobson, C. M. *Nature* **2004**, *432*, 824–828.

(114) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; García-Jacas, C. R.; Chavez, E.; Beltran, J. A.; Guillen-Ramirez, H. A.; Brizuela, C. A. *Scientific Reports* **2020**, *10*, 18074.

(115) Li, W.; Tan, S.; Xing, Y.; Liu, Q.; Li, S.; Chen, Q.; Yu, M.; Wang, F.; Hong, Z. *Molecular Pharmaceutics* **2018**, *15*, 1505–1514.

(116) Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2014**, *28*, 795–802.

(117) Vogt, M.; Stumpfe, D.; Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 191–208.

(118) Bondy, J. A.; Murty, U. S. R., *Graph Theory*; Graduate Texts in Mathematics, Vol. 244; Springer London: London, 2008.

(119) de la Vega de León, A.; Bajorath, J. *F1000Research* **2016**, *5*, 2634.

(120) Zwierzyna, M.; Vogt, M.; Maggiora, G. M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 113–125.

(121) Zahoránszky-Kohalmi, G.; Bologa, C. G.; Oprea, T. I. *Journal of Cheminformatics* **2016**, *8*, 1–17.

(122) Newman, M., *Networks*; 1; Oxford University Press: 2010; Vol. 15, pp 583–605.

(123) Newman, M. E. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **2004**, *70*, 9.

(124) Newman, M. E. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 8577–8582.

(125) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, *2008*, P10008.

(126) Lü, L.; Chen, D.; Ren, X. L.; Zhang, Q. M.; Zhang, Y. C.; Zhou, T. *Physics Reports* **2016**, *650*, 1–63.

(127) Pfeiffer, J. J.; Neville, J. **2011**.

(128) Boldi, P.; Vigna, S. *Internet Mathematics* **2014**, *10*, 222–262.

(129) Csermely, P.; Korcsmáros, T.; Kiss, H. J.; London, G.; Nussinov, R. *Pharmacology and Therapeutics* **2013**, *138*, 333–408.

(130) Tyagi, A.; Kapoor, P.; Kumar, R.; Chaudhary, K.; Gautam, A.; Raghava, G. P. *Scientific Reports* **2013**, *3*, 1–8.

(131) Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. *Molecules* **2019**, *24*, 1973.

(132) Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K. C. *Oncotarget* **2016**, *7*, 16895–16909.

(133) Timmons, P. B.; Hewage, C. M. *Biomedicine & Pharmacotherapy* **2021**, *133*, 111051.

(134) Gautam, A.; Chaudhary, K.; Kumar, R.; Sharma, A.; Kapoor, P.; Tyagi, A.; Raghava, G. P. S. *Journal of Translational Medicine* **2013**, *11*, 74.

(135) Tang, H.; Su, Z.-D.; Wei, H.-H.; Chen, W.; Lin, H. *Biochemical and Biophysical Research Communications* **2016**, *477*, 150–154.

(136) Manavalan, B.; Subramaniyam, S.; Shin, T. H.; Kim, M. O.; Lee, G. *Journal of Proteome Research* **2018**, *17*, 2715–2726.

(137) Nasiri, F.; Atanaki, F. F.; Behrouzi, S.; Kavousi, K.; Bagheri, M. *ACS Omega* **2021**, *6*, 19846–19859.

(138) Gupta, S.; Kapoor, P.; Chaudhary, K.; Gautam, A.; Kumar, R.; Raghava, G. P. S. *PLoS ONE* **2013**, *8*, ed. by Patterson, R. L., e73957.

(139) Chaudhary, K.; Kumar, R.; Singh, S.; Tuknait, A.; Gautam, A.; Mathur, D.; Anand, P.; Varshney, G. C.; Raghava, G. P. S. *Scientific Reports* **2016**, *6*, 22843.

(140) Win, T. S.; Malik, A. A.; Prachayasittikul, V.; S Wikberg, J. E.; Nantasenamat, C.; Shoombuatong, W. *Future Medicinal Chemistry* **2017**, *9*, 275–291.

(141) Sharma, A.; Singla, D.; Rashid, M.; Raghava, G. P. S. *BMC Bioinformatics* **2014**, *15*, 1–8.

(142) Prabakaran, R.; Rawat, P.; Kumar, S.; Michael Gromiha, M. *Journal of Molecular Biology* **2021**, *433*, 166707.

(143) Iglesias, V.; Santos, J.; Santos-Suárez, J.; Pintado-Grima, C.; Ventura, S. *Frontiers in Molecular Biosciences* **2021**, *8*, DOI: `10.3389/fmolb.2021.718301`.

(144) Sharma, N.; Patiyal, S.; Dhall, A.; Pande, A.; Arora, C.; Raghava, G. P. S. *Briefings in Bioinformatics* **2021**, *22*, DOI: `10.1093/bib/bbaa294`.

(145) Pintado, C.; Santos, J.; Iglesias, V.; Ventura, S. *Bioinformatics* **2021**, *37*, ed. by Arne, E., 1602–1603.

(146) Lathwal, A.; Kumar, R.; Kaur, D.; Raghava, G. P. S. *bioRxiv* **2021**, *302*, DOI: `10.1101/2021.06.20.449146`.

(147) Dhanda, S. K.; Gupta, S.; Vir, P.; Raghava, G. P. S. *Clinical and Developmental Immunology* **2013**, *2013*, 1–9.

(148) Nagpal, G.; Usmani, S. S.; Dhanda, S. K.; Kaur, H.; Singh, S.; Sharma, M.; Raghava, G. P. S. *Scientific Reports* **2017**, *7*, 42851.

(149) Gupta, S.; Madhu, M. K.; Sharma, A. K.; Sharma, V. K. *Journal of Translational Medicine* **2016**, *14*, 178.

(150) Gupta, S.; Sharma, A. K.; Shastri, V.; Madhu, M. K.; Sharma, V. K. *Journal of Translational Medicine* **2017**, *15*, 7.

(151) Kaur, D.; Arora, C.; Raghava, G. P. S. *Frontiers in Immunology* **2020**, *11*, DOI: `10.3389/fimmu.2020.00071`.

(152) Rajput, A.; Gupta, A. K.; Kumar, M. *PLOS ONE* **2015**, *10*, ed. by Kurgan, L., e0120066.

(153) Chung, C.-R.; Kuo, T.-R.; Wu, L.-C.; Lee, T.-Y.; Horng, J.-T. *Briefings in Bioinformatics* **2020**, *21*, 1098–1114.

(154) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. *Nucleic Acids Research* **2016**, *44*, D1094–D1097.

(155) Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H. K.; Wong, K. H.; Siu, S. W. *Molecular Therapy - Nucleic Acids* **2020**, *20*, 882–894.

(156) Santos-Júnior, C. D.; Pan, S.; Zhao, X.-M.; Coelho, L. P. *PeerJ* **2020**, *8*, e10555.

(157) Pinacho-Castellanos, S. A.; García-Jacas, C. R.; Gilson, M. K.; Brizuela, C. A. *Journal of Chemical Information and Modeling* **2021**, *61*, 3141–3157.

(158) Meher, P. K.; Sahu, T. K.; Saini, V.; Rao, A. R. *Scientific Reports* **2017**, *7*, 42362.

(159) Agrawal, P.; Bhalla, S.; Chaudhary, K.; Kumar, R.; Sharma, M.; Raghava, G. P. *Frontiers in Microbiology* **2018**, *9*, 1–13.

(160) Schaduangrat, N.; Nantasenamat, C.; Prachayasittikul, V.; Shoombuatong, W. *International Journal of Molecular Sciences* **2019**, *20*, 5743.

(161) Usmani, S. S.; Bhalla, S.; Raghava, G. P. S. *Frontiers in Pharmacology* **2018**, *9*, DOI: `10.3389/fphar.2018.00954`.

(162) Sharma, A.; Gupta, P.; Kumar, R.; Bhardwaj, A. *Scientific Reports* **2016**, *6*, 21839.

(163) Bastian, M.; Heymann, S.; Jacomy, M *International AAAI Conference on Weblogs and Social Media* **2009**, 361–362.

(164) Willett, P. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(165) Triguero, I.; González, S.; Moyano, J. M.; García, S.; Alcalá-Fdez, J.; Luengo, J.; Fernández, A.; del Jesús, M. J.; Sánchez, L.; Herrera, F. *International Journal of Computational Intelligence Systems* **2017**, *10*, 1238.

(166) Iman, R. L.; Davenport, J. M. *Communications in Statistics - Theory and Methods* **1980**, *9*, 571–595.

(167) Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; Norris, A.; Sanseau, P.; Cavalla, D.; Pirmohamed, M. *Nature Reviews Drug Discovery* **2019**, *18*, 41–58.

(168) Lee, W. H.; Loo, C. Y.; Ghadiri, M.; Leong, C. R.; Young, P. M.; Traini, D. *Advanced Drug Delivery Reviews* **2018**, *133*, 107–130.

(169) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. *Molecular Systems Biology* **2011**, *7*, DOI: `10.1038/msb.2011.75`.

(170) Katoh, K.; Misawa, K.; Kuma, K. I.; Miyata, T. *Nucleic Acids Research* **2002**, *30*, 3059–3066.

(171) Edgar, R. C. *Nucleic Acids Research* **2004**, *32*, 1792–1797.

(172) Notredame, C.; Higgins, D. G.; Heringa, J. *Journal of Molecular Biology* **2000**, *302*, 205–217.

(173) Xu, J.; Li, F.; Leier, A.; Xiang, D.; Shen, H.-H.; Marquez Lago, T. T.; Li, J.; Yu, D.-J.; Song, J. *Briefings in Bioinformatics* **2021**, *22*, 1–22.

(174) Thomsen, M. C. F.; Nielsen, M. *Nucleic Acids Research* **2012**, *40*, 281–287.

(175) Bailey, T. L. *Bioinformatics* **2021**, *37*, ed. by Birol, I., 2834–2840.

(176) Stoye, J.; Evers, D.; Meyer, F. *Bioinformatics* **1998**, *14*, 157–163.

(177) Lamiable, A.; Thévenet, P.; Rey, J.; Vavrusa, M.; Derreumaux, P.; Tufféry, P. *Nucleic Acids Research* **2016**, *44*, W449–W454.

(178) Chicco, D.; Tötsch, N.; Jurman, G. *BioData Mining* **2021**, *14*, 13.

(179) Jobin, M.-L.; Blanchet, M.; Henry, S.; Chaignepain, S.; Manigand, C.; Castano, S.; Lecomte, S.; Burlina, F.; Sagan, S.; Alves, I. D. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2015**, *1848*, 593–602.

(180) Chu, H. L.; Yip, B. S.; Chen, K. H.; Yu, H. Y.; Chih, Y. H.; Cheng, H. T.; Chou, Y. T.; Cheng, J. W. *PLoS ONE* **2015**, *10*, 1–14.

(181) Ellerby, H. M.; Arap, W.; Ellerby, L. M.; Kain, R.; Andrusiak, R.; Rio, G. D.; Krajewski, S.; Lombardo, C. R.; Rao, R.; Ruoslahti, E.; Bredesen, D. E.; Pasqualini, R. *Nature Medicine* **1999**, *5*, 1032–1038.

(182) Bayse, C. A.; Pollard, D. B. *Journal of Peptide Science* **2019**, *25*, 16–22.

(183) Lee, S.; Kim, S. M.; Lee, R. T. *Antioxidants and Redox Signaling* **2013**, *18*, 1165–1207.

(184) Elliott, S. E.; Parchim, N. F.; Kellems, R. E.; Xia, Y.; Soffici, A. R.; Daugherty, P. S. *Clinical Immunology* **2016**, *168*, 64–71.

(185) Ohta, T.; Hashida, Y.; Yamashita, F.; Hashida, M. *Biological & Pharmaceutical Bulletin* **2016**, *39*, 1687–1693.

(186) Bailey, T. L. *Bioinformatics* **2011**, *27*, 1653–1659.

(187) Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; Glass, C. K. *Molecular Cell* **2010**, *38*, 576–589.

(188) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. *Nucleic Acids Research* **2009**, *37*, 202–208.

(189) Jennings, B. H.; Pickles, L. M.; Wainwright, S. M.; Roe, S. M.; Pearl, L. H.; Ish-Horowicz, D. *Molecular Cell* **2006**, *22*, 645–655.

(190) Castelletto, V.; Edwards-Gayle, C. J.; Hamley, I. W.; Pelin, J. N.; Alves, W. A.; Aguilar, A. M.; Seitsonen, J.; Ruokolainen, J. *ACS Applied Bio Materials* **2019**, *2*, 3639–3647.

(191) Benjdia, A.; Berteau, O. *Frontiers in Chemistry* **2021**, *9*, 1–16.

(192) Yu, F.-H.; Huang, K.-J.; Wang, C.-T. *Journal of Virology* **2017**, *91*, 1–14.

(193) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. *Nucleic Acids Research* **2002**, *30*, 42–46.

(194) Ohlsson, B. *Frontiers in Endocrinology* **2017**, *8*, 1–7.

(195) Spindel, E. R. In *Handbook of Biologically Active Peptides*, Kastin, A. J., Ed., Second Edi; Academic Press: Boston, 2013, pp 326–330.

(196) Cavaco, M.; Valle, J.; Flores, I.; Andreu, D.; A. R. B. Castanho, M. *Clinical and Translational Science* **2021**, *14*, 1349–1358.

(197) Haggag, Y. A. *Biomedical Journal of Scientific & Technical Research* **2018**, *8*, 6659–6662.

(198) Morozumi, N.; Sato, S.; Yoshida, S.; Yamaki, A.; Furuya, M.; Inomata, N.; Ohnuma, N.; Minamitake, Y.; Ohsuye, K.; Kangawa, K. *Peptides* **2012**, *33*, 279–284.

(199) AlQahtani, A. D.; O'Connor, D.; Domling, A.; Goda, S. K. *Biomedicine and Pharmacotherapy* **2019**, *113*, 108750.

(200) Bruno, B. J.; Miller, G. D.; Lim, C. S. *Therapeutic Delivery* **2013**, *4*, 1443–1467.

(201) Matsui, D.; Nakano, S.; Dadashipour, M.; Asano, Y. *Scientific Reports* **2017**, *7*, 1–12.

(202) Sheikhpour, E.; Noorbakhsh, P.; Foroughi, E.; Farahnak, S.; Nasiri, R.; Neamatzadeh, H. *Reports of Biochemistry and Molecular Biology* **2017**, *7*, 30–37.

(203) Kajiwara, A.; Doi, H.; Eguchi, J.; Ishii, S.; Hiraide-Sasagawa, A.; Sakaki, M.; Omori, R.; Hiroishi, K.; Imawari, M. *Oncology Reports* **2012**, *27*, 1765–1771.

(204) Needleman, S. B.; Wunsch, C. D. *Journal of Molecular Biology* **1970**, *48*, 443–453.

Attachments

A. FASTA of a set of representative 105 venom peptides obtained from starPepDB.

```
>starPep_42302
VRDAYIAKNYNCVYECFRDSYCNDLCTKNGASSGYCQWAGKYGNACWCYALPDNVPIRVPGKCH
>starPep_26952
KKNGYAVDSSGKVAECLFNNYCNNECTKVYYADKGYCCLLKCYCFGLADDKPVLDIWDSTKNYCDVQIIDLS
>starPep_36890
RKCLIKYSQANESSKTCPSGQLLCLKKWEIGNPSGKEVKRGCVATCPKPWKNEIIQCCAKDKCNA
>starPep_35339
QAVGLPHGFCIQCNRKTWSNCSIGHRCLPYHMTCYTLYKPDENGEMKWAVKGCARMCPTAKSGERVKCCTGASCNSD
>starPep_01487
KSCCPNTTGRNIYNTCRFAGGSRERCAKLSGCKIISASTCPSDYPK
>starPep_11356
LVKCRGTSDCGRPCQQQTGCPNSKCINRMCKCYGC
>starPep_08992
DCGHLHDPCPNDRPGHRTCCIGLQCRYGKCLVRV
>starPep_40522
SVNPCCDPVICKPRDGEHCISGPCCNNCKFLNSGTICQRARGDGNHDYCTGITTDCPRNRYN
>starPep_17417
DCVRFWGKCSQTSDCCPHLACKSKWPRNICVWDGSV
>starPep_10426
IPYCGQTGAECYSWCIKQDLSKDWCCDFVKDIRMNPPADKCP
>starPep_24098
GTYCIELGERCPNPREGDWCCHKCVPEGKRFYCRDQ
>starPep_08211
ADDDCLPRGSKCLGENKQCCKGTTCMFYANRCVGV
>starPep_20284
FRGLAKLLKIGLKSFARVLKKVLPKAAKAGKALAKSLADENAIRQQNQ
>starPep_14045
AACKCDDEGPDIRTAPLTGTVDLGSCNAGWEKCASYYTIIADCCRKKK
>starPep_09101
DLWQFGKMILKVAGKLPFPYYGAYGCYCGWGGRGKPKDPTDRCCFVHDCC
>starPep_18579
EDPLYCQAIGCPTLYSEANLAVSKECRDQGKLGDDFHRCCEEQCGSTTPASA
>starPep_09273
EPDEICRARMTHKEFNYKSNVCNGCGDQVAACEAECFRNDVYTACHEAQK
>starPep_04906
ACVGDGQRCASWSGPYCCDGYYCSCRSMPYCRCRNNS
>starPep_28739
LKCYQHGKVVTCHRDMKFCYHNTGMPFRNLKLILQGCSSSCSETENNKCCSTDRCNK
>starPep_32830
MNSSKLIRMLEEDGWRLVRVTGSHHHFKHPKKPGLVTVPHPKKDLPIGTVKSIQKSAGL
>starPep_03482
MKLQNTLILIGCLFLMGAMIGDAYSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
>starPep_42730
VVIGQRCYRSPDCYSACKKLVGKATGKCTNGRCDC
>starPep_09985
GPSFCKADEKPCEYHADCCNCCLSGICAPSTNWILPGCSTSSFFKI
>starPep_41131
TPFAIKCATDADCSRKCPGNPPCRNGFCACT
>starPep_09189
ECLGFGKGCNPSNDQCCKSSNLVCSRKHRWCKYEI
>starPep_09703
GCMKEYCAGQCRGKVSQDYCLKHCKCIPR
>starPep_27368
KNRPTFCNLLPETGRCNALIPAFYYNSHLHKCQKFNYGGCGGNANNFKTIDECQRTCAAKYGRSS
>starPep_09902
GLIHKVTKVQQLCAFNQDMAGWCEKSCQAAEGKNGYCHGTKCKCGKPLSYRRK
>starPep_01641
ADDKNPLEEFRETNYEVFLEIAKNGLKATSNPKRVVIVGAGMAGLSAAY
>starPep_24157
GVIPKKIWETVCPTVEPWAKKCSGDIATYIKRECGKL
>starPep_36339
RDGYPLASNGCKFGCSGLGENNPTCNHVCEKKAGSDYGYCYAWTCYCEHVAEGTVLWGDSGTGPCRS
>starPep_01371
GKFSVFSKILRSIAKVFKGVGKVRKQFKTASDLDKNQ
>starPep_16015
CAKKRNWCGKNEDCCCPMKCIYAWYNQQGSCQTTITGLFKKC
>starPep_24272
GWCGDPGATCGKLRLYCCSGFCDCYTKTCKDKSSA
>starPep_17657
DFPLSKEYESCVRPRKCKPPLKCNKAQICVDPNKGW
>starPep_15767
AVITGACERDLQCGKGTCCAVSLWIKSVRVCTPVGTSGEDCHPASHKIPFSGQRMHHTCPCAPNLACVQTSPKKFKCLSKS
```

A. (Cont.) FASTA of a set of representative 105 venom peptides obtained from starPepDB.

```
>starPep_22698
GKRPRPVMCQCVDTTNGGVRLDAVTRAACSIDSFIDGYYTEKDGFCRAKYSWDLFTSGQFYQACLRYSHAGTNCQPDPQYE
>starPep_13668
WLGCARVKEACGPWEWPCCSGLKCDGSECHPQ
>starPep_20467
FVQHRPRDCESINGVCRHKDTVNCREIFLADCYNDEQKCCRK
>starPep_09596
FPRPRICNLACRAGIGHKYPFCHCR
>starPep_06936
MKTQFAIFLITLVLFQMFSQSDAIFKAIWSGIKSLFGKRGLSDLDDLDESFDGEVSQADIDFLKELMQ
>starPep_40558
SWDSIWKSAKNKMDKIMRQKVAKWMAKKEGKSVEEVQAKVDAMSKKDIRMHVISHYGKKAFEQLSKSLEE
>starPep_00486
GRGREFMSNLKEKLSGVKEKMKNS
>starPep_00538
RICRRRSAGFKGPCVSNKNCAQVCMQEGWGGGNCDGPLRRCKCMRRC
>starPep_31970
MISMLRCTFFFVSVILITSYFVTPTMSIKCNRKRHVIKPHICRKICGKNG
>starPep_02059
ADPTFGFTPLGLSEKANLQIMKAYD
>starPep_24080
GTTCYCGKTIGIYWFGTKTCPSNRGYTGSCGYFLGICCYPVD
>starPep_09161
DVTFSLLGANTKSYAAFITNFRKDVASEKK
>starPep_33838
NCVANILNINEAVIATGCVPAGGELRIFVGSSHSYLIKATSSCGLSLTNQVFINGESVQSGGRC
>starPep_02272
IWLTALKFLGKNLGKHLAKQQLAKL
>starPep_00607
FHPSLWVLIPQYIQLIRKILKSG
>starPep_11966
MTKQSIVIVLFAAIAMMACLQRVTAEPAPEPIAAPIAEPYANPEAIASPEAKDLHTVVSAILQALGKK
>starPep_06767
MAQDIISTIGDLVKWIIDTVNKFTKK
>starPep_40811
TDDESGNKCAKTKRRENVCRVCGNRSGNDEYYSECCESDYRYHRCLDLLRN
>starPep_13871
YCQKWMWTCDEERKCCEGLVCRLWCKRIINM
>starPep_00286
GFFALIPKIISSPLFKTLLSAVGSALSSSGEQE
>starPep_41864
VIIYELNLQGTTKAQYSTILKQLRDDIKDPNLXYGXXDYS
>starPep_09560
FLPLLILGSLLMTPPVIQAIHDAQR
>starPep_02364
PNPKVFFDMTIGGQSAGRIVMEEYA
>starPep_05690
GLKDWWNKHKDKIVEVVKDSGKAGLNAA
>starPep_05884
HGEGTFTSDLSKQMEEEAVRLFIEWLKNGGPSSGAPPPS
>starPep_35905
QPQSHSIELDEVSKEAASTRAALTSNL
>starPep_13368
VAVKATTTEEETEIPAK
>starPep_32019
MKDLMSLVIAPIFVGLVLEMISRVLDEEDDSRK
>starPep_09722
GEEELQENQELIREKSN
>starPep_10044
GSPRTEYEACRVRCQVAEHGVERQRRCQQVCEKRLREREGRRE
>starPep_09011
DDRRSPLEECFQQNDYEEFLEIARNSQLYQESLREDSSYHLSFIESLKSDALFSYEKKFWEADGIHGGKVINDLSLIHDLPKREIQALCYPSIKK
>starPep_01642
ADDRNPLEQCFRETDYEEFLEIARNNLKATSNPKHVVIVGAGMAGLSAAYVLSGGGHQVTV
>starPep_01891
KVCRQRSAGFKGPCVSDKNCAQVCLQEGWGGGNCDGPFRRCKCIRQC
>starPep_03728
AAPCFCSGKPGRGDLWILRGTCPGGYGYTSNCYKWPNICCYPH
>starPep_03969
FVQHRPRDCESINGVCRHKDTVNCREIFLADCYNDGQKCCRK
>starPep_08260
AEKDCIAPGAPCFGTDKPCCNPRAWCSSYANKCL
>starPep_09697
GCGGLMAGCDGKSTFCCSGYNCSPTWKWCVYARP
>starPep_09702
GCLGEGEKCADWSGPSCCDGFYCSCRSMPYCRCRNNS
>starPep_11811
MKTQFAILLVALVLFQMFAQSDAILGKIWEGIKSLFGKRGLSDLDGLDELFDGEISKADRDFLRELMR
>starPep_14332
ADCNGACSPFEVPPCRSRDCRCVPIGLFVGFCIHPTG
>starPep_16975
CSCNDINDKECMYFCHQDVIWDEP
>starPep_18084
DRDSCVDKSRCAKYGYYQECQDCCKNAGHNGGTCMFFKCKCA
>starPep_18467
ECLEIFKACNPSNDQCCKSSKLVCSRKTRWCKYQI
```

A. (Cont.) FASTA of a set of representative 105 venom peptides obtained from starPepDB.

```
>starPep_11744
MKFLVNVALVFYGRVHFLHLCVHFLHLWAPEPEPAPEAEAEADAEADPEAGIGAVLKVLTTGLPALISWIKRKRQQG
>starPep_16979
CSCTDMSDLECMNFCHKDVIWINRN
>starPep_13928
YKQCHKKGGHCFPKEKICIPPSSDLGKMDCRWKWKCCKKGSG
>starPep_39618
SGPADCCRMKECCTDRVNECLQRYSGREDKFVSFCYQEATVTCGSFNEIVGCCYGYQMCMIRVVKPNSLSGAHEACKTVSCGNPCA
>starPep_21163
GEATTIWGVGADEAIDKGTPSKNDLQNMSADLAKNGFKGHQGVACSTVKDGNKDVYMIKFSLAGGSNDPGGSPCSDD
>starPep_09979
GPMRIPEKHRIVREYIRKFGLQLNEFVQETENAWYYIKNIRKKVHEVKKDPGLLKYPVKP
>starPep_32131
MKISQVFIFVFLLMISVAWANEAYEEESNYLSERFDADVEEITPEFRGIRCPKSWKCKAFKQRVLKRLLAMLRQHAF
>starPep_23891
GSCVPVDQPCSLNTQPCCDDATCTQERNENGHTVYYCRA
>starPep_24192
GVPCRCDSDGPSVHGNTLSGTVWVGSCASGWHKCNDEYNIAYECCKE
>starPep_25916
ISIDPPCRFCYHRDGSGNCVYDAYGCGAV
>starPep_29538
LTCVKSNSIWFPTSEDCPDGQNLCFKRWQYISPRMYDFTRGCAATCPKAEYRDVINCCGTDKCNK
>starPep_34640
NSVNPCCDPQTCKPIEGKHCISGPCCENCYFLRSGTICQRARGDGNNDYCTGITPDCPRNRYN
>starPep_36661
RICYSHKASLPRATKTCVENTCYKMFIRTHRQYISERGCGCPTAMWPYQTECCKGDRCNK
>starPep_37544
RPTDIKCSESYQCFPVCKSRFGKTNGRCVNGFCDCF
>starPep_39452
SECVENGGFCPDPEKMGDWCCGRCIRNECRNG
>starPep_41191
TPYPVNCKTDRDCVMCGLGISCKNGYCQGCT
>starPep_44620
YKQCHKKGGHCFPKEKICIPPSSDFGKMDCRWRWKCCKKRSGK
>starPep_32546
MKYFVIALALAVALVCIAESTAYEVNEELENELDDLDDAAWLAVAEELQGLEDFEESRGLFGKLIKKFGRKAISYAVKKARGKN
>starPep_14712
AKACTPLLHDCSHDRHSCCRGDMFKYVCDCFYPEGEDKTEVCSCQQPKSHKIAEKIIDKAKTTL
>starPep_32810
MNSKIFAVLLLLAFLSCVLSDQYCPKSSITACKKMNIRNDCCKDDDCTGGSWCCATPCGNFCKYPTDRPGGKRAAGGKSCKTGYVYY
>starPep_28501
LFECSFSCEIEKEGDKPCKKKKCKGGWKCKFNMCVKV
>starPep_06838
MGAALKMTIFLLIVACAMIATTEAAVRIGPCDQVCPRIVPERHECCRAHGRSGYAYCSGGGMYCN
>starPep_02749
MEKIANAVKSAIEAGQNQDWTKLGTSILDIVSNGVTELSKIFGF
>starPep_35451
QEDGEIVCGEDDPCGTQICECDKAAAICFRNSMDT
>starPep_00644
FSFKRLKGFAKKLWNSKLARKIRTKGLKYVKNFAKDMLSEGEEAPPAAEPPVEAPQ
>starPep_02421
VFHAYSARGVRNNYKSAVGPADWVISAVRGFIHG
```

B. FASTA of a set of representative 162 ACPs obtained from starPepDB.

```
>starPep_05497
GFKDLLKGAAKALKKTVLF
>starPep_05855
GWRKWIKKATHVGKHIGKAALDAYI
>starPep_03042
FLGALFKVASKVLPSVKCAITKKC
>starPep_00126
GLFGKLIKKFGRKAISYAVKKARGKH
>starPep_06208
KILRGVSKKIMRTFLRRISKDILTGKK
>starPep_09845
GIPCGESCVFIPCLTSAIGCSCKSKVCYRN
>starPep_11176
LKCNKLVPLFYKTCPAGKNL
>starPep_18164
DTAVTGLASPLSTGKILDQKAYSCANRLIVLCIENSFMTDARK
>starPep_24426
GYNYAKKLANLAKKFANALW
>starPep_09764
GFWSSVWDGAKNVGTAIIKNAKVCVYAVCVSHK
>starPep_03287
IKIPSFFRNILKKVGKEAVSLIAGALKQS
>starPep_00640
FLSLIPHIVSGVASIAKHF
>starPep_00315
GLFDIVKKIAGHIASSI
>starPep_32958
MQFITDLIKKAVDVFKGLFGNK
>starPep_00361
KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAL
>starPep_24256
GVWGIAKIAGKVLGNILPHVFSSNQS
>starPep_00657
GFLGILFHGVHHGRKKALHMNSERRS
>starPep_00807
KSSAYSLQMGATAIKQVKKLFKKWGW
>starPep_22576
GKEFKRIVWLSKTAKKL
>starPep_18824
EKSSRPEFYKVILGAHEEYIRG
>starPep_27320
KNECLWTDMLSNFGYPGYQSKHYACIRQKG
>starPep_14350
ADMDFTGIAESIIKKIKETNAKPPA
>starPep_12134
PAWRKAFRWAARMLKKAA
>starPep_36015
QRTESIIHRALYYDLIS
>starPep_13195
TFRAFLSSRLQDLYSIVRRADRAAV
>starPep_21198
GEILCNLCTGLINTLENLLTTKRKRQQ
>starPep_34474
NPEKALEKLIAIQKAIKGMLNGWFTGVGFRRKR
>starPep_10217
HLRRINKLLTRIGLYRHAFG
>starPep_12142
PDEDAINNALNKVCSTGRRQRSICKQLLKK
>starPep_24842
HTHQDFQPVLHLVALNTPLSGGMRGIR
>starPep_18008
DPFFKVPVNKLAAVSNFGYDLYRVRSSMSPTTN
>starPep_07864
VLLVTLTRLHQRGVIYRKWRHFSGRKYR
>starPep_41343
TTITGKKCQSWAAMFPHRHSKT
>starPep_07104
MWKEFHNVLSSGQLLADKRWARWYNRW
>starPep_07120
NLVSALIEGRKYLKNVLKKLNRLKEKNKAKNSKENN
>starPep_07882
VNWKKXLGKXIKXVK
>starPep_10591
KIAKVALAKLGIGAVLKVLTTGL
>starPep_11276
LPRRNRWSKIWKKVVTVFS
>starPep_12079
NRFTARFRRTPWRLCLQFRQ
>starPep_13296
TRWLWLLRGGLKAAGWGIRAHLNRNQ
```

B. (Cont.) FASTA of a set of representative 162 ACPs obtained from starPepDB.

```
>starPep_02535
FLHHIVGLIHHGLSLFGDRAD
>starPep_00758
GWKKWFTKGERLSQRHFA
>starPep_00419
FLGALIKGAIHGGRFIHGMIQNHH
>starPep_09994
GQVWEATATVNAIRGSVTPAVSQFNARTAD
>starPep_35821
QMIVIELGTNPLKSSGIENGAFQGMK
>starPep_00249
ACGILHDNCVYVPAQNPCCRGLQCRYGKCLVQV
>starPep_26183
KAKAKAVSRSARAGLQFPVGRIHRHLK
>starPep_36692
RIIDLLWRVWRPQKPKFVTVWVR
>starPep_41266
TRSRWRRFIRGAGRFARRYGWRIA
>starPep_02289
KRKCPKTPFDNTPGAWFAHLILGC
>starPep_14847
ALARQPLTGSPPNERAFFCSSLRR
>starPep_00719
GLLSVLGSVVKHVIPHVVPVIAEHL
>starPep_14190
AAPFLECQGRQGTCHFFAN
>starPep_03327
KKCKFFCKVKKKIKSIGFQIPIVSIPFK
>starPep_41021
TKWTPCSRTCGMGISNRV
>starPep_23911
GSGSGSGSLKKIFKKPMVIGVTIPF
>starPep_19944
FLKDHRISTFKNWPF
>starPep_02569
GFKRIVQRIKDFLRNLV
>starPep_01434
GTGLPMSERRKIMLMMR
>starPep_10463
ITCPQVTQSLAPCVPYLISG
>starPep_11901
MRGIRGADFQAFQQARAVGLAGTFR
>starPep_04365
LALERRSGWLRLFGLKPRRKH
>starPep_13155
TAGIKLTVPIEKFPVTTQTFWG
>starPep_30757
MFSPILSLEIILALATLQSVFAQPVICTTVGSAAEGS
>starPep_00686
GKGRWLERIGKAGGIIIGGALDHL
>starPep_15907
AWYRGAAPPKQEFLDIEDP
>starPep_12249
PRFWEYALRLME
>starPep_03733
ACVNQCPDAIDRFIVKDKGCHGVEKKYYKQVYVACMNGQHLYCRTEWGGPCQL
>starPep_35471
QEPHRHSIFTPQTNPRADLEKN
>starPep_36506
RGFTKMPHVQIHTEASESL
>starPep_04324
KRMGIFHLFWAGLRKLGNLIKNKIQQGIENFLG
>starPep_01246
ATPATPTVAQFVIQGSTICLVC
>starPep_29515
LSSTCILVLVKDILVLVVKEILVLVVKDKPI
>starPep_40880
TFKRKNGSRKNGHRPGGYSLIALGNKKVLKAPYMESI
>starPep_26006
ITMQGIQGQKIRMIMF
>starPep_25650
IMRIKQGQIGQMTI
>starPep_00911
ANDPQCLYGNVAAKF
>starPep_26645
KIKSCYYLPCFVTS
>starPep_18575
EDMNQKLFDLRGKFKRPPLRRVRMSADAML
>starPep_02982
CVLIGQRCDNDRGPRCCSGQGNCVPLPFLGGVCAV
>starPep_02368
QICKAPSQTFPGLCFMDSSCRKYCIKEKFTGGHCSKLQRKCLCTKPC
>starPep_27510
KRFKQDGGWSHWSPWSSC
```

B. (Cont.) FASTA of a set of representative 162 ACPs obtained from starPepDB.

```
>starPep_41295
TSLDASIIWAMMQN
>starPep_43002
WGRAFSAGVHRLANGGNG
>starPep_16327
CDSDSDITWDQLWDLMK
>starPep_00564
YRGGYTGPIPRPPPIGRPPFRPVCNACYRLSVSDARNCCIKFGSCCHLVK
>starPep_32937
MPTWAWWLFLVLLLALWAPARG
>starPep_07884
VNWXXILGXIIXVVX
>starPep_29189
LPGLTGSKGVRGISGLPGFSG
>starPep_24115
GVDITVIRPNH
>starPep_21830
GFHDHGPCDPPSHK
>starPep_05790
GRKKRRQRRRGGWMWVTNLRTD
>starPep_16457
CGGYSGGWHRLRSTSYRCG
>starPep_13094
STRUCTUREGIVEN
>starPep_04732
RWFKIQMQIRRWKNKK
>starPep_08575
AXQNMEILEXTPLTXVX
>starPep_42725
VVGSPSAQDEASPL
>starPep_22164
GHRATSDLASTGEESQD
>starPep_16341
CELDENNTPMC
>starPep_40547
SVSRAGSPSGGPFC
>starPep_41784
VGTDFSGNDDISDVQK
>starPep_37972
RRPKGRAMRREKQRPSDKPRR
>starPep_33129
MRLLVLSSLLCILLLCFSIFSTEGKRRPAKAWSGRRTRLCCHRVPSPNSTNLKGHHVRLC
KPCKLEPEPRLWVVPGALPQV
>starPep_44212
XXLIXVWAXGFXXAXXLFXGIG
>starPep_43649
WYTXXXTWXWXY
>starPep_43904
XIXIILPPLPII
>starPep_09828
GIIKKIIIKKIIIKKIIIKKI
>starPep_28706
LIXFXPX
>starPep_09658
FXYWKXT
>starPep_00089
SWLSKTAKKLENSAKKRISEGIAIAIQGGPR
>starPep_00224
RIIDLLWRVRRPQKPKFVTVWVR
>starPep_00260
DTHFPICIFCCGCCHRSKCGMCCKT
>starPep_00775
ILGPVLGLVSDTLDDVLGIL
>starPep_03342
KLKNFAKGVAQSLLNKASCKLSGQC
>starPep_03814
CETPSKHFNGLCIRSSNCASVCHGEHFTDGRCQGVRRRCMCLKPC
>starPep_03965
FVKLKKILNIINSIFKK
>starPep_04262
KKALKHALAKWLPALKALAHKLAKK
>starPep_04575
NFAEIFAAVNKLIKQGVVKG
>starPep_04632
QRSVSNAATRVSRTGRSRWRDVSRNFMRR
>starPep_05117
CHTNGGYCVRAICPPSARRPGSCFPEKNPCCKYM
>starPep_06981
MPKEKVFLKIEKMGRNIRN
```

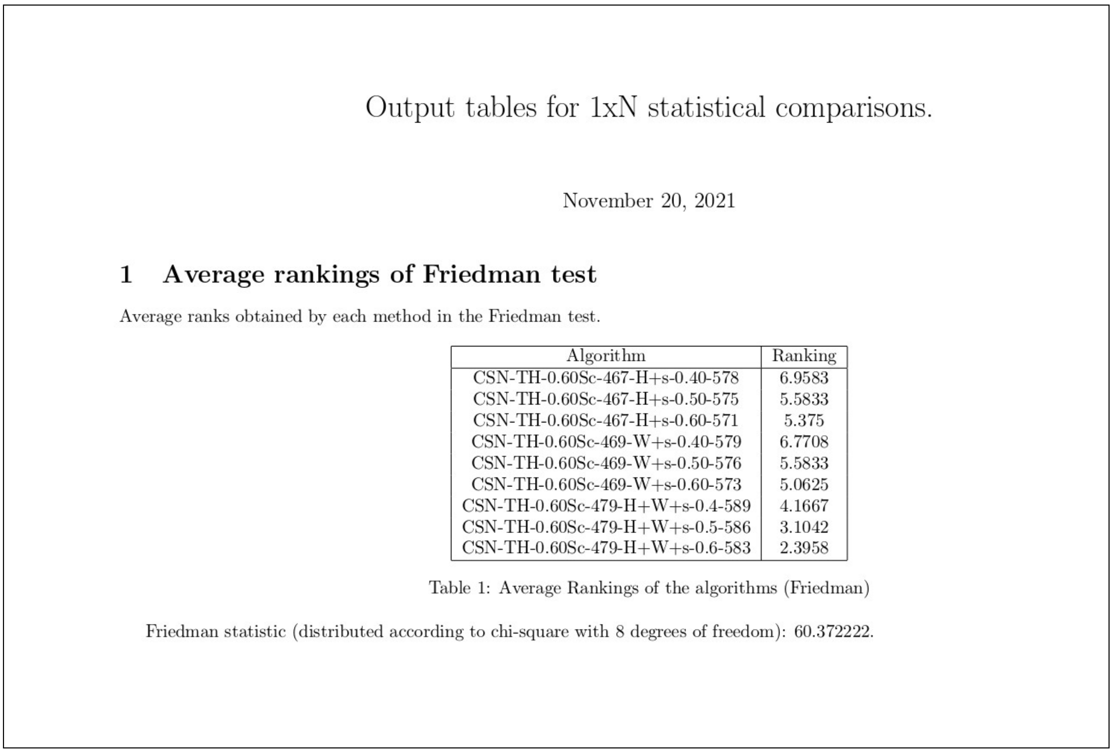
B. (Cont.) FASTA of a set of representative 162 ACPs obtained from starPepDB.

```
>starPep_16391
CGESCVFIPCISSVIGCACKSKVCYKNGSIP
>starPep_18108
DRSTREPIYMSTI
>starPep_18157
DSSPVSTEQLAPTA
>starPep_19548
FFSLIPKLVKGLISAFK
>starPep_20122
FLSLIPAAISAVSALANHF
>starPep_24834
HTASDAAAAAALTAANAAAAAAAASMA
>starPep_39171
SAPFIECHGRGTCNYYANS
>starPep_41038
TLPFAYCNIHQVCHYAQRNDRSYWL
>starPep_24565
HGLGHGHEQQHGLGHGHKFKLDDDLEHQGGHVLD
>starPep_30471
MDSNKDERAYAQWVIIILHNVGSSPFKIANLGLSWGKLYADGNKDKEVYP
>starPep_21042
GCRRLCWKQRCVTYCRGR
>starPep_41838
VIFEWTLLQVLSESDQDQSLEVFLT
>starPep_09784
GGVCPKILKKCRRDSDCPGACICRGNGYCGSGSD
>starPep_04761
SKWQHQQDSCRKQLQGVNLTPCEKHIMEKIQGRGDDDDDDDDD
>starPep_23735
GRFKRFRKKLKRLWHKVGPFVGPILHY
>starPep_18139
DSEGWKVQPNINRDQDGNTAGSVRVQKQLGNHEVHAGASRVFSGPNRGGPSYNVGATFNW
>starPep_01120
ITSISLCTPGCKTGALMGCNMKTATCNCSIHVSK
>starPep_26494
KGIRGYKGGYCKGAFKQTCKCY
>starPep_10026
GSEGPLKPGARIFSFDGKDVLRHPT
>starPep_39822
SKRKSRPVSVKTFEDIPLEEP
>starPep_33979
NGREACLDPEAPMVQKIVQKMLKG
>starPep_41197
TQQAFQKFLAAVTSALGKQYH
>starPep_05425
FSPQMLQDIIEKKTKIL
>starPep_21918
GFRKRFNKLVKKVKHTIKETANVSKDVAIVAGSGVAVGAAM
>starPep_29425
LRSRGELVAKFLAGEQSPEDYVAE
>starPep_18829
EKYEGKISKTMSGLDCQAWDS
>starPep_28608
LHCPALVTYNTDTFESMPNPEGRYTFGASCV
>starPep_00270
FLIGMTQGLICLITRKC
>starPep_17955
DLWIRETLTSPKSLTG
>starPep_00021
ACYCRIPACIAGERRYGTCIYQGRLWAFCC
>starPep_02292
KSCCPNTTGRNIYNTCRLTGSSRETCAKLSGCKIISASTCPSNYPK
>starPep_00334
GLMSSIGKALGGLIVDVLKPKTPAS
>starPep_26158
IYSFDGRDIMTDPSWPQKVIWHGSSPHGVRLVDNYCEAWRTA
>starPep_29236
LPRFSTMPFIYCNINEVCHY
>starPep_03897
FCTCNVKGFNAKNKRGIIYP
>starPep_01521
MRKEFHNVLSSGQLLADKRPARDYNRK
>starPep_08214
ADDKNPLEECFRETDYEEFLEIARNGLKATSNPKRVV
>starPep_34689
NVLLSPLSVATALSALSLGAEQRTES
>starPep_15086
ANIKLSVQMKLFKRHLKWKIIVKLNDGRELSLDA
>starPep_40840
TEENRELVSELKRP
>starPep_00124
GKPRPYSPRPTSHPRPIRV
```

C. CSN parameters of similarity threshold analysis.

Similarity threshold	Density	Communities	Modularity	Singletons	ACC
0.1	0.999	3	0.03	0	0.999
0.15	0.996	3	0.03	0	0.996
0.2	0.985	3	0.03	0	0.988
0.25	0.956	3	0.04	0	0.968
0.3	0.891	3	0.05	0	0.93
0.35	0.772	3	0.07	0	0.87
0.4	0.593	3	0.11	0	0.791
0.45	0.383	4	0.16	2	0.703
0.5	0.199	4	0.23	3	0.612
0.55	0.079	6	0.34	20	0.508
0.6	0.023	10	0.47	99	0.428
0.65	0.005	34	0.68	238	0.419
0.7	0.001	38	0.81	449	0.544
0.75	0	21	0.85	548	0.535
0.8	0	13	0.84	587	0.456
0.85	0	9	0.87	606	0.456
0.9	0	2	0.5	623	-

D. Output from Friedman test where 9 best SSMs were compared.



Output tables for 1xN statistical comparisons.

November 20, 2021

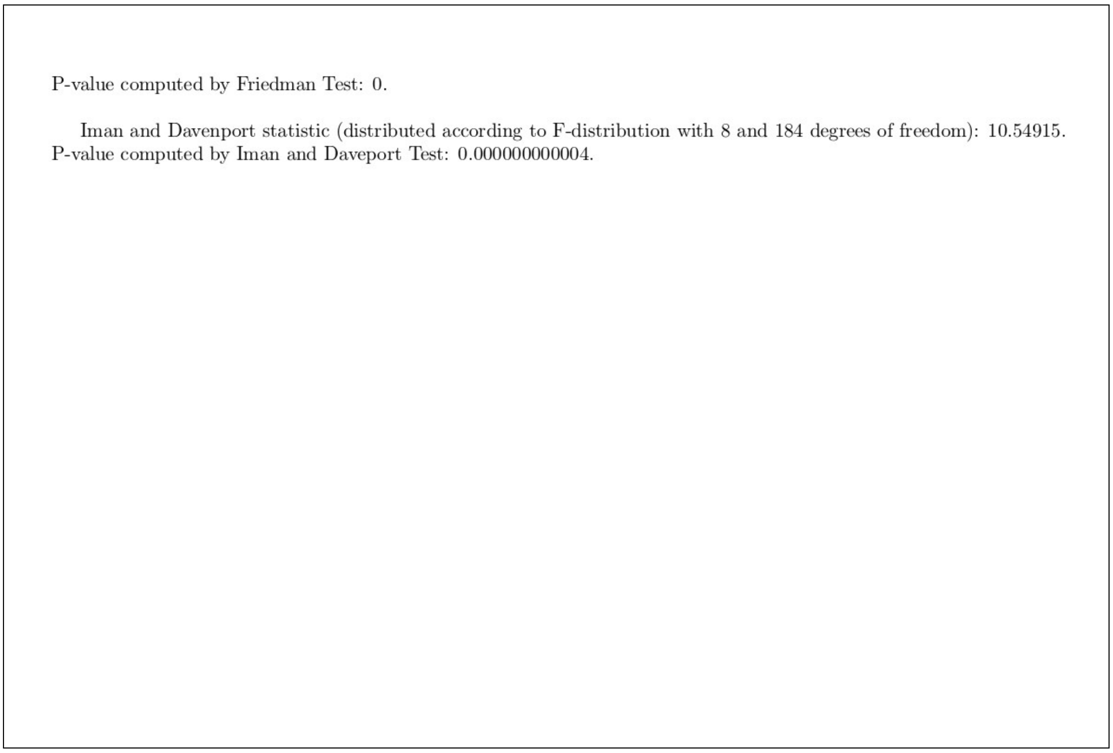
1 Average rankings of Friedman test

Average ranks obtained by each method in the Friedman test.

Algorithm	Ranking
CSN-TH-0.60Sc-467-H+s-0.40-578	6.9583
CSN-TH-0.60Sc-467-H+s-0.50-575	5.5833
CSN-TH-0.60Sc-467-H+s-0.60-571	5.375
CSN-TH-0.60Sc-469-W+s-0.40-579	6.7708
CSN-TH-0.60Sc-469-W+s-0.50-576	5.5833
CSN-TH-0.60Sc-469-W+s-0.60-573	5.0625
CSN-TH-0.60Sc-479-H+W+s-0.4-589	4.1667
CSN-TH-0.60Sc-479-H+W+s-0.5-586	3.1042
CSN-TH-0.60Sc-479-H+W+s-0.6-583	2.3958

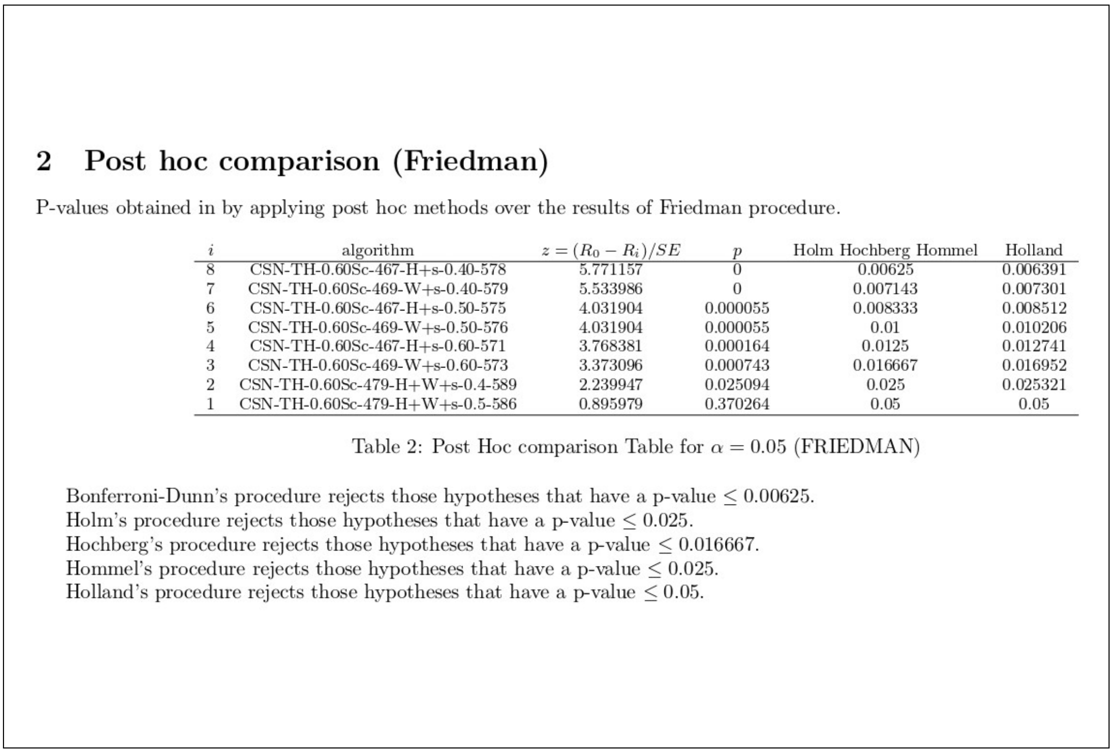
Table 1: Average Rankings of the algorithms (Friedman)

Friedman statistic (distributed according to chi-square with 8 degrees of freedom): 60.372222.



P-value computed by Friedman Test: 0.

Iman and Davenport statistic (distributed according to F-distribution with 8 and 184 degrees of freedom): 10.54915. P-value computed by Iman and Daveport Test: 0.000000000004.



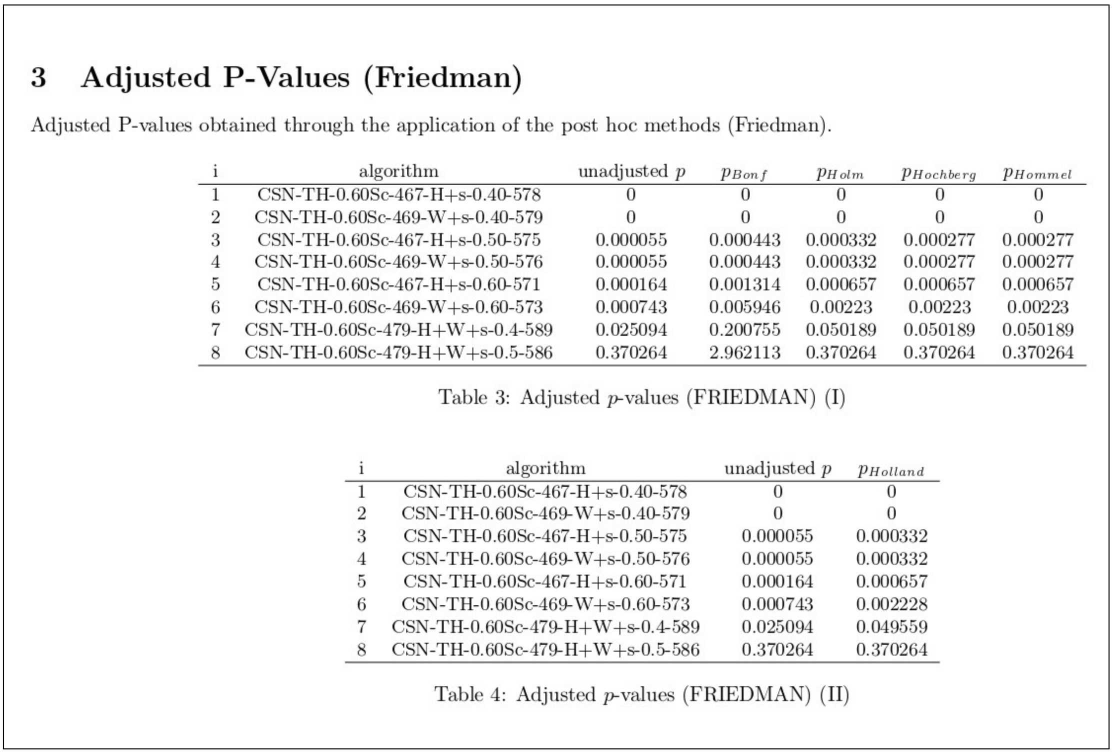
2 Post hoc comparison (Friedman)

P-values obtained in by applying post hoc methods over the results of Friedman procedure.

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm Hochberg Hommel	Holland
8	CSN-TH-0.60Sc-467-H+s-0.40-578	5.771157	0	0.00625	0.006391
7	CSN-TH-0.60Sc-469-W+s-0.40-579	5.533986	0	0.007143	0.007301
6	CSN-TH-0.60Sc-467-H+s-0.50-575	4.031904	0.000055	0.008333	0.008512
5	CSN-TH-0.60Sc-469-W+s-0.50-576	4.031904	0.000055	0.01	0.010206
4	CSN-TH-0.60Sc-467-H+s-0.60-571	3.768381	0.000164	0.0125	0.012741
3	CSN-TH-0.60Sc-469-W+s-0.60-573	3.373096	0.000743	0.016667	0.016952
2	CSN-TH-0.60Sc-479-H+W+s-0.4-589	2.239947	0.025094	0.025	0.025321
1	CSN-TH-0.60Sc-479-H+W+s-0.5-586	0.895979	0.370264	0.05	0.05

Table 2: Post Hoc comparison Table for $\alpha = 0.05$ (FRIEDMAN)

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.00625.
Holm's procedure rejects those hypotheses that have a p-value ≤ 0.025.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.016667.
Hommel's procedure rejects those hypotheses that have a p-value ≤ 0.025.
Holland's procedure rejects those hypotheses that have a p-value ≤ 0.05.



3 Adjusted P-Values (Friedman)

Adjusted P-values obtained through the application of the post hoc methods (Friedman).

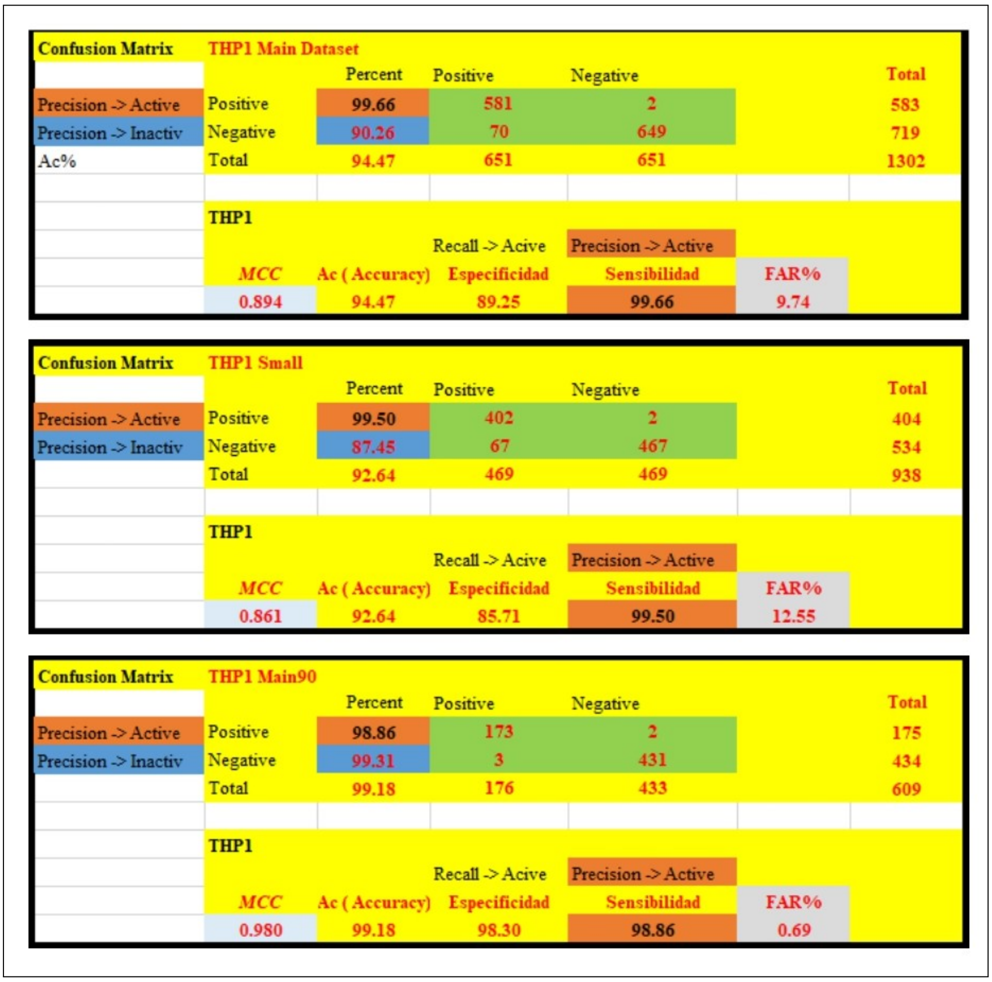
i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	$p_{Hochberg}$	p_{Hommel}
1	CSN-TH-0.60Sc-467-H+s-0.40-578	0	0	0	0	0
2	CSN-TH-0.60Sc-469-W+s-0.40-579	0	0	0	0	0
3	CSN-TH-0.60Sc-467-H+s-0.50-575	0.000055	0.000443	0.000332	0.000277	0.000277
4	CSN-TH-0.60Sc-469-W+s-0.50-576	0.000055	0.000443	0.000332	0.000277	0.000277
5	CSN-TH-0.60Sc-467-H+s-0.60-571	0.000164	0.001314	0.000657	0.000657	0.000657
6	CSN-TH-0.60Sc-469-W+s-0.60-573	0.000743	0.005946	0.00223	0.00223	0.00223
7	CSN-TH-0.60Sc-479-H+W+s-0.4-589	0.025094	0.200755	0.050189	0.050189	0.050189
8	CSN-TH-0.60Sc-479-H+W+s-0.5-586	0.370264	2.962113	0.370264	0.370264	0.370264

Table 3: Adjusted p-values (FRIEDMAN) (I)

i	algorithm	unadjusted p	$p_{Holland}$
1	CSN-TH-0.60Sc-467-H+s-0.40-578	0	0
2	CSN-TH-0.60Sc-469-W+s-0.40-579	0	0
3	CSN-TH-0.60Sc-467-H+s-0.50-575	0.000055	0.000332
4	CSN-TH-0.60Sc-469-W+s-0.50-576	0.000055	0.000332
5	CSN-TH-0.60Sc-467-H+s-0.60-571	0.000164	0.000657
6	CSN-TH-0.60Sc-469-W+s-0.60-573	0.000743	0.002228
7	CSN-TH-0.60Sc-479-H+W+s-0.4-589	0.025094	0.049559
8	CSN-TH-0.60Sc-479-H+W+s-0.5-586	0.370264	0.370264

Table 4: Adjusted p-values (FRIEDMAN) (II)

E. Confusion matrices of the best SSM THP1 against Main, Small, and Main90 datasets.

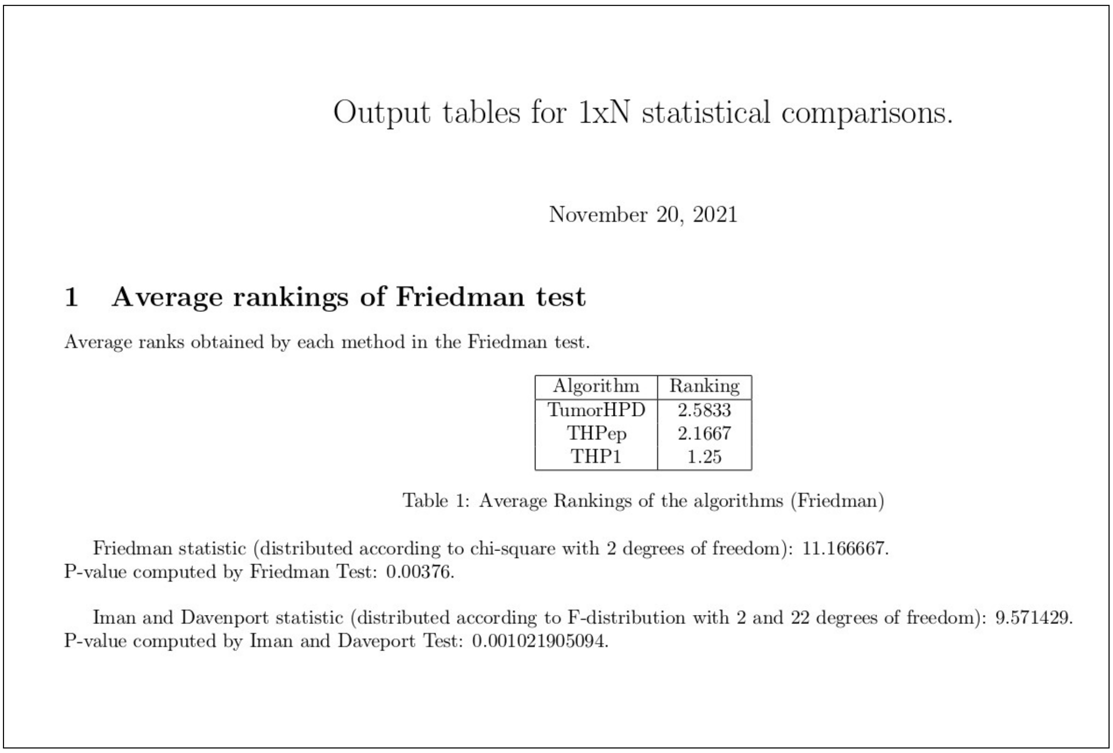


Confusion Matrix	THP1 Main Dataset					
		Percent	Positive	Negative		Total
Precision -> Active	Positive	99.66	581	2		583
Precision -> Inactiv	Negative	90.26	70	649		719
Ac%	Total	94.47	651	651		1302
	THP1					
			Recall -> Acive	Precision -> Active		
	MCC	Ac (Accuracy)	Especificidad	Sensibilidad	FAR%	
	0.894	94.47	89.25	99.66	9.74	

Confusion Matrix	THP1 Small					
		Percent	Positive	Negative		Total
Precision -> Active	Positive	99.50	402	2		404
Precision -> Inactiv	Negative	87.45	67	467		534
	Total	92.64	469	469		938
	THP1					
			Recall -> Acive	Precision -> Active		
	MCC	Ac (Accuracy)	Especificidad	Sensibilidad	FAR%	
	0.861	92.64	85.71	99.50	12.55	

Confusion Matrix	THP1 Main90					
		Percent	Positive	Negative		Total
Precision -> Active	Positive	98.86	173	2		175
Precision -> Inactiv	Negative	99.31	3	431		434
	Total	99.18	176	433		609
	THP1					
			Recall -> Acive	Precision -> Active		
	MCC	Ac (Accuracy)	Especificidad	Sensibilidad	FAR%	
	0.980	99.18	98.30	98.86	0.69	

F. Output from Friedman test where THP1 was compared with literature models.



Output tables for 1xN statistical comparisons.

November 20, 2021

1 Average rankings of Friedman test

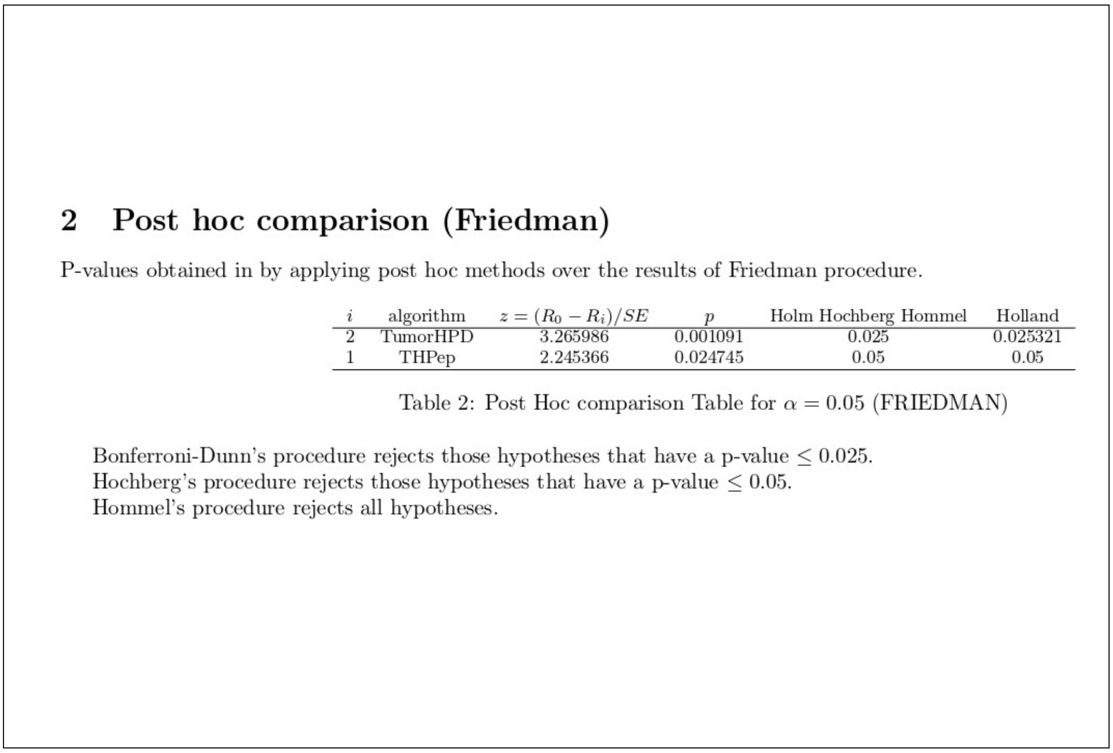
Average ranks obtained by each method in the Friedman test.

Algorithm	Ranking
TumorHPD	2.5833
THPep	2.1667
THP1	1.25

Table 1: Average Rankings of the algorithms (Friedman)

Friedman statistic (distributed according to chi-square with 2 degrees of freedom): 11.166667. P-value computed by Friedman Test: 0.00376.

Iman and Davenport statistic (distributed according to F-distribution with 2 and 22 degrees of freedom): 9.571429. P-value computed by Iman and Daveport Test: 0.001021905094.



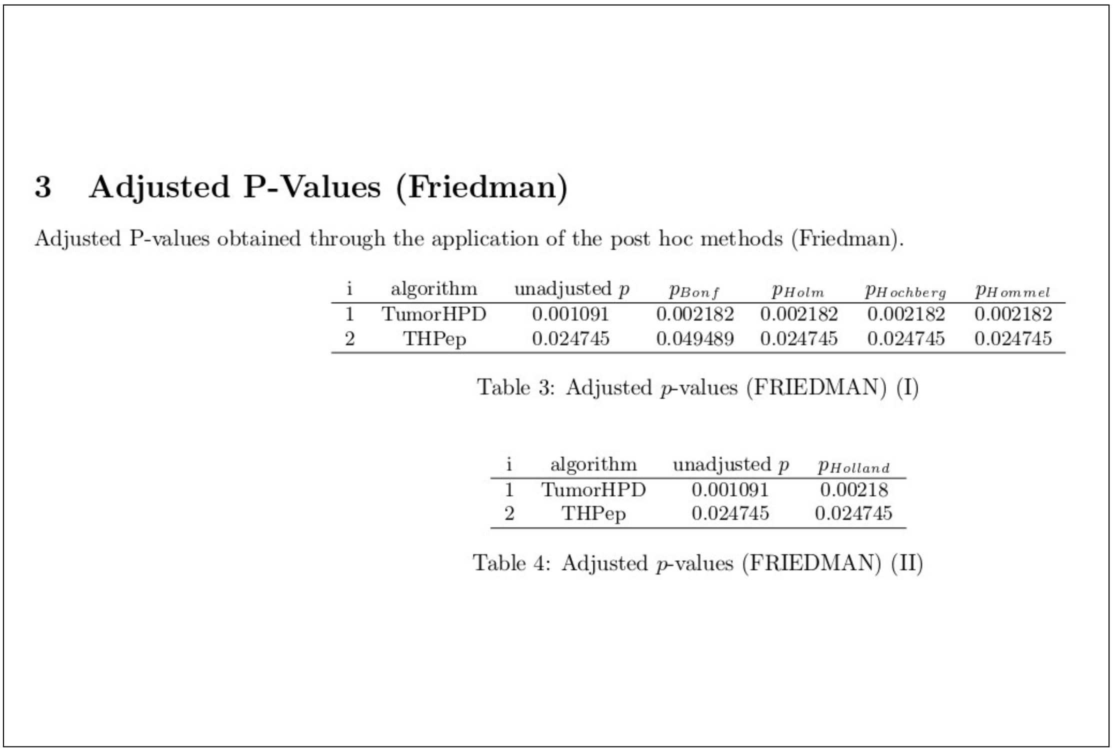
2 Post hoc comparison (Friedman)

P-values obtained in by applying post hoc methods over the results of Friedman procedure.

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm Hochberg Hommel	Holland
2	TumorHPD	3.265986	0.001091	0.025	0.025321
1	THPep	2.245366	0.024745	0.05	0.05

Table 2: Post Hoc comparison Table for $\alpha = 0.05$ (FRIEDMAN)

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.025.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.05.
Hommel's procedure rejects all hypotheses.



3 Adjusted P-Values (Friedman)

Adjusted P-values obtained through the application of the post hoc methods (Friedman).

i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	$p_{Hochberg}$	p_{Hommel}
1	TumorHPD	0.001091	0.002182	0.002182	0.002182	0.002182
2	THPep	0.024745	0.049489	0.024745	0.024745	0.024745

Table 3: Adjusted p-values (FRIEDMAN) (I)

i	algorithm	unadjusted p	$p_{Holland}$
1	TumorHPD	0.001091	0.00218
2	THPep	0.024745	0.024745

Table 4: Adjusted p-values (FRIEDMAN) (II)

G. Predicted activities of 43 repurposed peptides obtained from hierarchical virtual screening of peptides from starPepDB.*Originally, these peptides contained a X aa, which was changed by an aa that gave them greater tumor homing potential.

ID	Sequence	TumorHPD		THPep	AntiCP		CellPPD	ToxinPred							
		SVM Score			SVM Score			SVM Score		SVM Score		SVM Score		SVM Score	
starPep_27924	KWCFRVAYRGISYRRCR	0.62	THP	THP	1	Anticp	CPP	-1.45	Non-Toxin	-1.45	Non-Toxin	-0.73	Non-Toxin	-0.73	Non-Toxin
starPep_43589	WWWKNKGKKNGKH	0.98	THP	THP	0.56	Anticp	CPP	-0.32	Non-Toxin	-0.32	Non-Toxin	-0.04	Non-Toxin	-0.04	Non-Toxin
starPep_24644	HKHGHGHLKHKNKLKKNGKH	0.37	THP	THP	0.38	Anticp	CPP	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin
starPep_02029	TPFKLSLHL	0.33	THP	THP	0.17	Anticp	Non-CPP	-1.02	Non-Toxin	-1.02	Non-Toxin	-1.29	Non-Toxin	-1.29	Non-Toxin
starPep_07234	QGRLGTQWAVGHLM	0	THP	THP	-0.29	Non-Anticp	Non-CPP	-1.48	Non-Toxin	-1.48	Non-Toxin	-1.21	Non-Toxin	-1.21	Non-Toxin
starPep_43502	WWAMKWIRV	1.58	THP	THP	1.15	Anticp	CPP	-0.49	Non-Toxin	-0.49	Non-Toxin	-0.81	Non-Toxin	-0.81	Non-Toxin
starPep_13108	SVSWGMKPSPRQ	0.6	THP	THP	-0.46	Non-Anticp	Non-CPP	-1.4	Non-Toxin	-1.4	Non-Toxin	-1.76	Non-Toxin	-1.76	Non-Toxin
starPep_27446	KQCISLKGICKDLACT	0.38	THP	THP	1.8	Anticp	Non-CPP	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.02	Non-Toxin	-0.02	Non-Toxin
starPep_27346	KNKGKKWWW	1.04	THP	THP	0.69	Anticp	CPP	-0.6	Non-Toxin	-0.6	Non-Toxin	-0.44	Non-Toxin	-0.44	Non-Toxin
starPep_26052	IVLVRRWPK	-0.33	Non-THP	THP	0.54	Anticp	CPP	-0.9	Non-Toxin	-0.9	Non-Toxin	-0.99	Non-Toxin	-0.99	Non-Toxin
starPep_14535	AGIRRPPGFSPLRIA	0.46	THP	THP	0.15	Anticp	Non-CPP	-0.81	Non-Toxin	-0.81	Non-Toxin	-0.36	Non-Toxin	-0.36	Non-Toxin
starPep_16575	CKGKGAKAARSGKC	1.07	THP	THP	1.81	Anticp	CPP	1	Toxin	1	Toxin	1	Toxin	1	Toxin
starPep_10105	GWAGWLLSPRGSRPSWGP	2.2	THP	THP	0.86	Anticp	CPP	-1.03	Non-Toxin	-1.03	Non-Toxin	-1.05	Non-Toxin	-1.05	Non-Toxin
starPep_35988	QRNKGLRHH	-0.37	Non-THP	THP	0.33	Anticp	CPP	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin
starPep_10014	GRRSTHWRI	0.64	THP	THP	-0.03	Non-Anticp	CPP	-0.92	Non-Toxin	-0.92	Non-Toxin	-1.51	Non-Toxin	-1.51	Non-Toxin
starPep_07641	RSQMQDGQLQSCCQELQNVEEQCQC	0.28	THP	THP	-1.03	Non-Anticp	Non-CPP	0.24	Toxin	0.24	Toxin	-0.37	Non-Toxin	-0.37	Non-Toxin
starPep_18023	DPSFNSWG	0.51	THP	THP	-0.47	Non-Anticp	Non-CPP	-1.09	Non-Toxin	-1.09	Non-Toxin	-0.99	Non-Toxin	-0.99	Non-Toxin
starPep_25472	ILPVKWPWWPWRR	2.12	THP	THP	1.16	Anticp	CPP	-0.56	Non-Toxin	-0.56	Non-Toxin	-0.57	Non-Toxin	-0.57	Non-Toxin
starPep_43120	WKGRWYKTT	0.71	THP	THP	0.5	Anticp	CPP	-0.05	Non-Toxin	-0.05	Non-Toxin	-0.39	Non-Toxin	-0.39	Non-Toxin
starPep_13030	SPRGSRPSWGPTDPRRRS	1.04	THP	THP	-0.1	Non-Anticp	CPP	-1.29	Non-Toxin	-1.29	Non-Toxin	-1.72	Non-Toxin	-1.72	Non-Toxin
starPep_04689	RLRLRIGRR	1.14	THP	THP	-0.04	Non-Anticp	CPP	-1.22	Non-Toxin	-1.22	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin
starPep_15346	AQPSFAF	0.67	THP	THP	-0.3	Non-Anticp	Non-CPP	-1.2	Non-Toxin	-1.2	Non-Toxin	-0.88	Non-Toxin	-0.88	Non-Toxin
starPep_05157	DGPKKKKKKSPSKSSG	-0.19	Non-THP	non-THP	-0.12	Non-Anticp	CPP	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.92	Non-Toxin	-0.92	Non-Toxin
starPep_07335	RCICRLGIC	2.77	THP	THP	1.57	Anticp	CPP	-0.65	Non-Toxin	-0.65	Non-Toxin	-0.3	Non-Toxin	-0.3	Non-Toxin
starPep_08545	AVESTVATLEASPEVIESPPE	-1.71	Non-THP	non-THP	-1.27	Non-Anticp	Non-CPP	-0.96	Non-Toxin	-0.96	Non-Toxin	-1.07	Non-Toxin	-1.07	Non-Toxin
starPep_41900	VIWRWRKFY	0.36	THP	THP	1.2	Anticp	CPP	-0.58	Non-Toxin	-0.58	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin
starPep_05293	FFRNLWKGAKAAFRAGHAAWRA	0.07	THP	THP	0.35	Anticp	CPP	-0.74	Non-Toxin	-0.74	Non-Toxin	-1.28	Non-Toxin	-1.28	Non-Toxin
starPep_36476	RFWVRGRRS	0.66	THP	THP	0.11	Anticp	CPP	-0.98	Non-Toxin	-0.98	Non-Toxin	-1.43	Non-Toxin	-1.43	Non-Toxin
starPep_17042*	VLIWC	1.97	THP	THP	0.83	Anticp	Non-CPP	-0.41	Non-Toxin	-0.41	Non-Toxin	-0.32	Non-Toxin	-0.32	Non-Toxin
starPep_12276	PTSNHSPTSCPPTCPGYRWMCLRRF	1.71	THP	THP	0.94	Anticp	Non-CPP	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.68	Non-Toxin	-0.68	Non-Toxin
starPep_10092	GVGSPYVSRLLGICL	0.18	THP	non-THP	0.63	Anticp	Non-CPP	-0.91	Non-Toxin	-0.91	Non-Toxin	-1.13	Non-Toxin	-1.13	Non-Toxin
starPep_07237	QHWSYGLRPG	1.5	THP	THP	0.29	Anticp	Non-CPP	-0.9	Non-Toxin	-0.9	Non-Toxin	-0.74	Non-Toxin	-0.74	Non-Toxin
starPep_12415	QSFGNQWARGHFM	0.37	THP	THP	-0.53	Non-Anticp	Non-CPP	-1.17	Non-Toxin	-1.17	Non-Toxin	-0.83	Non-Toxin	-0.83	Non-Toxin
starPep_08820	CPSHLDAFC	1.97	THP	THP	0.37	Anticp	Non-CPP	-1.04	Non-Toxin	-1.04	Non-Toxin	-0.9	Non-Toxin	-0.9	Non-Toxin
starPep_01400	GLLSGVLGVGKKVDCGLSGLC	0.28	THP	non-THP	1.26	Anticp	Non-CPP	-1.09	Non-Toxin	-1.09	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin
starPep_43956*	LWRPP	2.9	THP	THP	1.13	Anticp	Non-CPP	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.29	Non-Toxin	-0.29	Non-Toxin
starPep_42404	VRLRIRSAVIRA	-0.21	Non-THP	non-THP	-0.47	Non-Anticp	Non-CPP	-0.96	Non-Toxin	-0.96	Non-Toxin	-1.45	Non-Toxin	-1.45	Non-Toxin
starPep_12257	PRPGPIYY	1.15	THP	THP	1.02	Anticp	Non-CPP	-1.1	Non-Toxin	-1.1	Non-Toxin	-0.49	Non-Toxin	-0.49	Non-Toxin
starPep_18019	DPPFSPRL	2.05	THP	THP	-0.08	Non-Anticp	Non-CPP	-1	Non-Toxin	-1	Non-Toxin	-0.46	Non-Toxin	-0.46	Non-Toxin
starPep_01691	EGGGPQWAVGHFM	-0.07	Non-THP	THP	0.16	Anticp	Non-CPP	-1.03	Non-Toxin	-1.03	Non-Toxin	-1.01	Non-Toxin	-1.01	Non-Toxin
starPep_13827*	CFRVC	2.34	THP	THP	1.19	Anticp	Non-CPP	-1.13	Non-Toxin	-1.13	Non-Toxin	-0.46	Non-Toxin	-0.46	Non-Toxin
starPep_16808	CNGRCGGKLAKLAKKLAKLAK	-0.03	Non-THP	non-THP	1.74	Anticp	CPP	-0.07	Non-Toxin	-0.07	Non-Toxin	0.34	Toxin	0.34	Toxin
starPep_29033	LLLNKKGKNKHKGHGHGHKH	-0.02	Non-THP	THP	0.58	Anticp	Non-CPP	-1.2	Non-Toxin	-1.2	Non-Toxin	-1.03	Non-Toxin	-1.03	Non-Toxin

G. (cont.) Predicted activities of 43 repurposed peptides obtained from hierarchical virtual screening of peptides from starPepDB.*Originally, these peptides contained a X aa, which was changed by an aa that gave them greater tumor homing potential.

ID	Sequence	HemoPI				
		SVM Score 1	**SVM Score 2**	**SVM Score 3**	**SVM Score 3**	**SVM Score 4**
starPep_01400	GLLSGVLGVGKKVDCGLSGLC	0.84	0.72	0.58	0.58	0.67
starPep_01691	EGGGPQWAVGHFM	0	0.06	0.49	0.49	0.45
starPep_02029	TPFKLSLHL	0.6	0.59	0.5	0.5	0.53
starPep_04689	RLRLRIGRR	0.96	0.79	0.48	0.48	0.42
starPep_05157	DGPKKKKKKSPSKSSG	0.54	0.49	0.49	0.49	0.44
starPep_05293	FFRNLWKGAKAAFRAGHAAWRA	0.73	0.94	0.49	0.49	0.46
starPep_07234	QGRLGTQWAVGHLM	0.24	0.25	0.49	0.49	0.49
starPep_07237	QHWSYGLRPG	0.16	0.21	0.48	0.48	0.41
starPep_07335	RCICRLGIC	1	0.79	0.49	0.49	0.44
starPep_07641	RSQMQDGQLQSCCQELQNVEEQCQC	0	0.15	0.49	0.15	0.44
starPep_08545	AVESTVATLEASPEVIESPPE	0	0	0.48	0.48	0.4
starPep_08820	CPSHLDAFC	0.54	0.5	0.49	0.49	0.43
starPep_10014	GRRSTHWRI	0.85	0.75	0.49	0.49	0.44
starPep_10092	GVGSPYVSRLLGICL	0.63	0.58	0.51	0.51	0.43
starPep_10105	GWAGWLLSPRGSRPSWGP	0.6	0.53	0.49	0.49	0.4
starPep_12257	PRPGPIYY	0.09	0.31	0.49	0.49	0.43
starPep_12276	PTSNHSPTSCPPTCPGYRWMCLRRF	0.43	0.49	0.49	0.49	0.39
starPep_12415	QSFGNQWARGHFM	0.11	0.16	0.49	0.49	0.49
starPep_13030	SPRGSRPSWGPTDPRRRS	0.29	0.72	0.49	0.49	0.43
starPep_13108	SVSWGMKPSPRQ	0.14	0.22	0.48	0.48	0.38
starPep_13827	RWCFRVCYGCCR	1	0.99	0.61	0.61	0.58
starPep_14535	AGIRRPPGFSPLRIA	0.33	0.34	0.48	0.48	0.34
starPep_15346	AQPSFAF	0.02	0.24	0.49	0.49	0.44
starPep_16575	CKGKGAKAARSGKC	1	1	0.48	0.48	0.39
starPep_16808	CNGRCGGKLAKLAKKLAKLAK	1	1	0.42	0.42	0.41
starPep_17042	CTDYVLIWC	0.92	0.78	0.49	0.49	0.46
starPep_18019	DPPFSPRL	0	0.24	0.49	0.49	0.42
starPep_18023	DPSFNSWG	0.13	0.22	0.49	0.49	0.4
starPep_24644	HKHGHGHLKHKNKLKKNGKH	0.77	0.62	0.49	0.49	0.44
starPep_25472	ILPVKWPWWPWRR	0.91	0.86	0.65	0.65	0.72
starPep_26052	IVLVRRWPK	0.92	0.81	0.4	0.4	0.31
starPep_27346	KNKGKKWWW	1	0.68	0.49	0.49	0.44
starPep_27446	KQCISLKGICKDLACT	1	0.92	0.49	0.49	0.55
starPep_27924	KWCFRVAYRGISYRRCR	1	1	0.53	0.53	0.44
starPep_29033	LLLNKKGKNKHKGHGHGHKH	0.82	0.69	0.49	0.15	0.45
starPep_35988	QRNKGLRHH	0.32	0.36	0.49	0.49	0.4
starPep_36476	RFWVRGRRS	1	0.94	0.49	0.49	0.41
starPep_41900	VIWRWRKFY	1	1	0.49	0.49	0.5
starPep_42404	VRLRIRSAVIRA	0.8	0.71	0.48	0.48	0.43
starPep_43120	WKGRWYKTT	0.81	0.63	0.49	0.49	0.43
starPep_43502	WWAMKWIRV	1	0.78	0.49	0.49	0.46
starPep_43589	WWWKNKGKKNGKH	0.92	0.69	0.49	0.49	0.45
starPep_43956	KWDPPPPSPP	0.28	0.47	0.49	0.49	0.44

H. FASTA of 180 THPs derived from 43 lead hits.

```
>starPep_43956_L5
WDPPP
>starPep_43956_It4_3_4L_1_5
WWLLRPPSPP
>starPep_43956_It4_3_4L_1_6
WWLLPRPSPP
>starPep_43956_It4_3_4W_1_5
LWLWRPPSPP
>starPep_43956_It4_3_5_1_4
WWLRLPPSPP
>starPep_13827_L5_1
XRWCF
>starPep_13827_L5_2
CFRVC
>starPep_13827_L5_3
WCFRV
>starPep_13827_L10
RWCFRVCYXG
>starPep_14535_L5_3
IRRPP
>starPep_14535_It4_3_14W_2_4
AWWWRPPGFSPLRWA
>starPep_14535_It4_3_14W_8_4_L10
WWRPPWFSPL
>starPep_14535_It4_3_14W_8_5_L10_1
WRWPPWFSPL
>starPep_14535_It4_3_14W_8_5_L10_2
WPPWFSPLRW
>starPep_15346_It4_1W_2_5_7
WHPSWAM
>starPep_15346_It4_6_2_5_7
AHPSWWM
>starPep_15346_It4_1W_2_5_7_L5_1
WHPSW
>starPep_15346_It4_6_2_5_7_L5_1
HPSWW
>starPep_15346_It4_6_2_5_7_L5_2
PSWWM
>starPep_16575_L5_1
CKGKG
>starPep_16575_It4_6_2_8_4_L5
CGCKC
>starPep_16575_It4_6_2_8_7_L5
KGCCC
>starPep_16808_L5_1
CNGRC
>starPep_16808_L5_2
NGRCG
>starPep_16808_L5_3
RCGGK
>starPep_16808_L10
CNGRCGGKLA
>starPep_17042_L5_1
LIWCX
>starPep_17042_L5_2
VLIWC
>starPep_17042_It4_2_5_7_4_L5_1
CCGVL
>starPep_17042_It4_2_5_7_4_L5_2
CCCGV
>starPep_18019_It4_1W_2_4_7
WWPYSPHL
>starPep_18019_It4_1W_6_4_7
WPPYSWHL
>starPep_18019_It4_1W_2_4_7_L5
WWPYS
>starPep_25472_It4_5L_4_1_12
ALPYLWPWWPWSR
>starPep_25472_It4_5L_4_1_12_L5
YLWPW
>starPep_25472_It4_5L_4_1_12_L10
YLWPWWWPWSR
>starPep_25472_It4_5Y_4_1_12_L10
LYWPWWWPWSR
>starPep_26052_L5_2
VRRWP
>starPep_26052_It4_9W_1_5_4
WVLCSRWPW
>starPep_26052_It4_9W_1_5_4_L5_1
LCSRW
>starPep_26052_It4_9W_1_5_4_L5_2
WVLCS
>starPep_27346_L5_2
KKWWW
>starPep_27346_It4_1_3_7_6_L5
CNCGK
>starPep_27346_It4_1_5_7_6_L5_1
CNKGC
>starPep_27346_It4_3_5_7_6_L5
KNCGC
>starPep_27446_L5_1
KGICK
>starPep_27446_L5_2
QCISL
>starPep_27446_L10
LKGICKDLAC
>starPep_27446_L15
KQCISLKGICKDLAC
>starPep_27924_L5_1
KWCFR
>starPep_27924_L5_2
WCFRV
>starPep_27924_L15
WCFRVAYRGISYRRC
>starPep_27924_It3_1C_6_9_14_L5_2
WCFRC
>starPep_29033_L5_3
HGHKH
>starPep_29033_L10
KGKNKHKGHG
>starPep_29033_L15
KKGKNKHKGHGHGHK
>starPep_29033_It4_1_2_3_11_L5_1
CCCNK
>starPep_35988_L5
GLRHH
>starPep_35988_It4_4C_8_9_2_L5_1
CGLRC
>starPep_35988_It4_4C_8_9_7_L5_2
GLCCC
>starPep_35988_It4_4C_8_2_1C_L5
WCNCG
>starPep_36476_It4_1_5_7_2C
WCWVWGLRS
>starPep_36476_It4_1_5_7_2C_L5_1
CWVWG
>starPep_36476_It4_1_5_7_2C_L5_2
WVWGL
>starPep_36476_It4_1_5_7_2A_L5
AWVWG
>starPep_41900_L5_4
WRKFY
```

H. (cont.) FASTA of 180 THPs derived from 43 lead hits.

```
>starPep_01400_L5_2
LSGLC
>starPep_01400_It4_14_2_3_4_L5_1
CCGVL
>starPep_01400_It4_14_7_2_6_L5
CLCGV
>starPep_01691_It4_9_6_3_4
EGWWPWWAWGHFM
>starPep_01691_It4_9_6_3_10
EGWGPWWAWWHFM
>starPep_01691_It4_9_6_2_4_L10
WPWWAWGHFM
>starPep_01691_It4_9_6_3_4_L5
WAWGH
>starPep_01691_It4_9_6_3_4_L10
WWPWWAWGHF
>starPep_08820_L5
CPSHL
>starPep_08820_It4_6W_8W_2_4_L5
CSRLW
>starPep_08820_It4_6W_8C_2_4_L5_2
CWSRL
>starPep_08820_It4_6C_8_2_4_L5_1
LCAWC
>starPep_08820_It4_6C_8_2_4_L5_2
WSRLC
>starPep_07237_It4_1_6W_10_8C
WHWSYWLCPC
>starPep_07237_It4_1_6C_10_8C
WHWSYCLCPW
>starPep_07237_It4_1_6W_10_8C_L5_1
WHWSY
>starPep_07237_It4_1_6W_10_8C_L5_3
WLCPC
>starPep_07237_It4_1_6C_10_8C_L5_1
HWSYC
>starPep_02029_It4_4W_5_3_7
TPWWWSYHL
>starPep_02029_It3_4W_7_3_9
TPWWLSWHY
>starPep_02029_It4_4W_5_3_7_L5
WWSYH
>starPep_02029_It4_4W_5_3_9_L5
WWSLH
>starPep_02029_It3_4W_7_3_9_L5_1
WLSWH
>starPep_04689_L5_1
LRLRI
>starPep_04689_It4_1_3_5_8_L5_1
CIGCR
>starPep_04689_It4_1_3_5_8_L5_2
CLCIG
>starPep_04689_It4_1_5_8_9_L5_2
LCIGC
>starPep_05157_It4_1C_10_12_4_L5_1
CGPCK
>starPep_05157_It4_1C_10_12_4_L5_2
CPCKS
>starPep_05157_It4_1C_10_12_4_L10
CGPCKKKKKC
>starPep_05157_It4_1C_10_12_6_L10
CGPKKCKKKC
>starPep_05157_It4_1C_10_14_4_L5
CPSKC
>starPep_05293_It4_7_10_9_L5
WWAFR
>starPep_05293_It4_7_10_9_L10
LWWGWWWAFR
>starPep_05293_It4_7_10_11_12_L5
LWWGA
>starPep_05293_It4_7_10_11_15_L10_1
LWWGAWWAFR
>starPep_05293_It4_7_10_11_15_L10_2
RNLWWGAWWA
>starPep_07234_It4_10W_1_7_6_L5_1
RLGCW
>starPep_07234_It4_10W_1_7_6_L5_2
LGCWW
>starPep_07335_It4_3W_1_8_L5
CWCRL
>starPep_07641_L5_1
CCQEL
>starPep_07641_It4_3_5_8_14_L5
CCCEL
>starPep_08545_It4_3_10_14_17_L5_1
WASPW
>starPep_08545_It4_3_10_14_17_L5_2
LWASP
>starPep_08545_It4_10_14_17_21_L15
ATLWASPWVIWSPPW
>starPep_10014_L5_3
THWRI
>starPep_10014_It4_9C_3_8_5_L5_1
CSLHW
>starPep_10014_It4_9C_3_8_5_L5_2
SLHWC
>starPep_10092_It4_13W_2_1_3
WWWSPYVSRLLGWCL
>starPep_10092_It4_13W_2_1_3_L5
WWSPY
>starPep_10092_It4_13W_7_1_3_L10
PYWSRLLGWC
>starPep_10092_It4_13W_7_1_12_L10_1
PYWSRLLWWC
>starPep_10092_It4_13W_7_1_12_L10_2
SPYWSRLLWW
>starPep_10105_It4_1_4_11_17
WWAWWLLSPRHSRPSWYP
>starPep_10105_It4_1_4_11_17_L10
WWAWWLLSPR
>starPep_10105_It4_1_4_11_17_L15
WAWWLLSPRHSRPSW
>starPep_10105_It4_1_11_4_17_L15
WWAHWLLSPRWSRPS
>starPep_12257_It4_4W_6_1_8
WRPWPLYF
>starPep_12257_It4_4L_6_1_7
WRPLPWFY
>starPep_12257_It4_4W_6_1_7_L5_1
WPLFY
>starPep_12257_It4_4W_6_1_8_L5
WPLYF
>starPep_12257_It4_4L_6_1_7_L5
LPWFY
>starPep_12276_It4_2_8_13_4_L5_1
PWSWH
>starPep_12276_It4_2_8_13_4_L5_2
WHSPW
>starPep_12276_It4_2_8_13_18_L10_1
CPGYWWMCLR
>starPep_12276_It4_2_8_13_18_L10_2
WCPGYWWMCL
>starPep_12415_L5_1
WARGH
>starPep_12415_It4_1_6_3_12C
WSWGNWWARGHCM
>starPep_12415_It4_1_6_3_5_L5_2
CWWAR
>starPep_12415_It4_1_6_3_5_L10_1
SWGCWWARGH
>starPep_12415_It4_1_6_3_5_L10_2
WSWGCWWARG
>starPep_13030_L5_1
RPSWG
>starPep_13030_It4_13W_12_1_3_L5_1
PLWPR
>starPep_13030_It4_13W_12_1_3_L5_2
WGPLW
>starPep_13030_It4_13W_12_1_6_L10
WPSWGPLWPR
>starPep_13108_L5_1
SVSWG
>starPep_13108_It4_12_7W_5_2_L5
HMWPS
>starPep_13108_It4_12_7W_5_11_L10
SWHMWPSPHW
```

H. (cont.) FASTA of 180 THPs derived from 43 lead hits.

```
>starPep_18019_It4_1W_3_4_7_L5_2
WPWYS
>starPep_18019_It4_1W_6_4_7_L5
PYSWH
>starPep_18023_It4_1H_8_5_4C
HPSCWSWH
>starPep_18023_It4_1W_5_8_4C
WPSCHSWH
>starPep_18023_It4_1H_8_5_4C_L5_1
HPSCW
>starPep_18023_It4_1H_8_5_4C_L5_2
CWSWH
>starPep_18023_It4_1W_5_8_4C_L5
WPSCH
>starPep_24644_L10_1
HKHGHGHLKH
>starPep_24644_L10_2
HGHGHLKHKN
>starPep_24644_It4_8C_14_3_5_L5_1
HKCGC
>starPep_24644_It4_8C_14_3_5_L5_3
KCGCG
>starPep_24644_It4_8C_14_3_5_L10
CGCGHCKHKN
>starPep_43502_It4_5S_7_9_1_L5_2
AMSWC
>starPep_43502_It4_5S_7_9_2_L5_1
WCAMS
>starPep_43589_L5_1
WWWKN
>starPep_43589_L5_4
KGKKN
>starPep_43589_It4_4_6_8_5_L5
CGCKN
>starPep_01400_L5_1
CGLSG
>starPep_13108_It4_12_7H_5_11_L10
SWWMHPSPHW
>starPep_41900_It4_7C_1_2_8
CLWRWRCGY
>starPep_41900_It4_7C_1_2_4_L5_1
CLWRW
>starPep_41900_It4_7C_1_2_4_L5_2
LWRWC
>starPep_42404_It4_1W_9_2_5_L5
WRSAW
>starPep_42404_It4_1W_9_2_10_L5_1
RSAWW
>starPep_42404_It4_1W_9_2_10_L5_2
SAWWR
>starPep_43120_L5_1
WKGRW
>starPep_43120_L5_2
KGRWY
>starPep_43120_It4_2C_8_9_7L_L5_2
RWYLC
>starPep_43502_It4_5S_7_9_1
CWAMSWCRC
>starPep_43502_It4_5C_7_9_1
CWAMCWSRC
>starPep_43502_It4_5S_7_9_1_L5_1
CWAMS
>starPep_07234_It4_1W_7_6_10_L10_1
WGRLGWWWAC
>starPep_07234_It4_1W_7_6_10_L10_2
RLGWWWACGH
>starPep_07335_L5
CICRL
>starPep_07335_It4_3C_1_8_L5
CCRLG
>starPep_07335_It4_8_1_3_L5_1
CRLGC
>starPep_07335_It4_8_1_3_L5_2
RLGCC
```

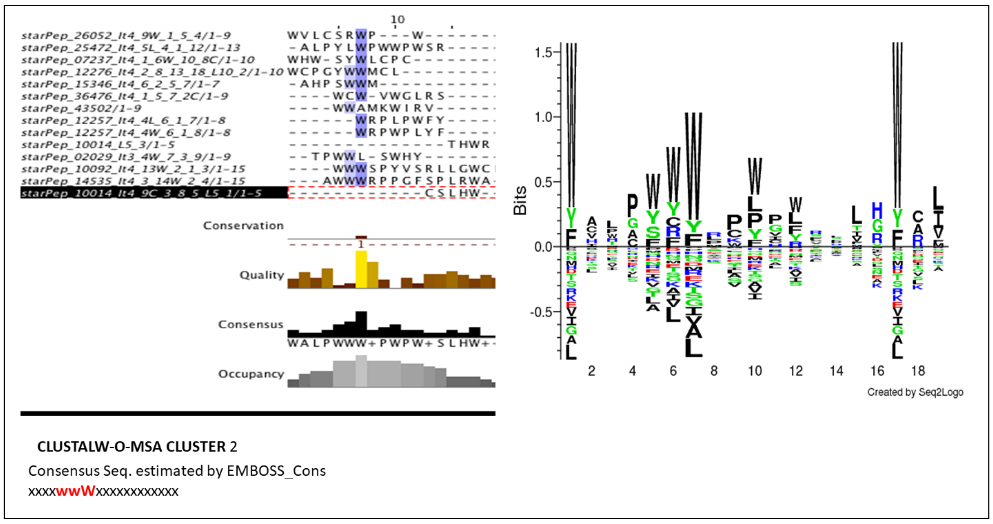
I. Predicted activities of SET 1, conformed by 54 lead THPs.

ID	Sequence	THP Model	TumorHPD		THPep	AntiCP		CellPPD	ToxinPred							
			SVM Score			SVM Score			SVM Score		SVM Score		SVM Score		SVM Score	
starPep_24644	HKHGHGHLKHKNKLKKNGKH	1	0.37	THP	THP	0.38	Anticp	CPP	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin
starPep_43502	WWAMKWIRV	1	1.58	THP	THP	1.15	Anticp	CPP	-0.49	Non-Toxin	-0.49	Non-Toxin	-0.81	Non-Toxin	-0.81	Non-Toxin
starPep_13108	SVSWGMKPSPRQ	1	0.6	THP	THP	-0.46	Non-Anticp	Non-CPP	-1.4	Non-Toxin	-1.4	Non-Toxin	-1.76	Non-Toxin	-1.76	Non-Toxin
starPep_35988	QRNKGLRHH	1	-0.37	Non-THP	THP	0.33	Anticp	CPP	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin
starPep_04689	RLRLRIGRR	1	1.14	THP	THP	-0.04	Non-Anticp	CPP	-1.22	Non-Toxin	-1.22	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin
starPep_36476	RFWVRGRRS	1	0.66	THP	THP	0.11	Anticp	CPP	-0.98	Non-Toxin	-0.98	Non-Toxin	-1.43	Non-Toxin	-1.43	Non-Toxin
starPep_12276	PTSNHSPTSCPPTCPGYRWMCLRRF	1	1.71	THP	THP	0.94	Anticp	Non-CPP	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.68	Non-Toxin	-0.68	Non-Toxin
starPep_10092	GVGSPYVSRLLGICL	1	0.18	THP	non-THP	0.63	Anticp	Non-CPP	-0.91	Non-Toxin	-0.91	Non-Toxin	-1.13	Non-Toxin	-1.13	Non-Toxin
starPep_07237	QHWSYGLRPG	1	1.5	THP	THP	0.29	Anticp	Non-CPP	-0.9	Non-Toxin	-0.9	Non-Toxin	-0.74	Non-Toxin	-0.74	Non-Toxin
starPep_12415	QSFGNQWARGHFM	1	0.37	THP	THP	-0.53	Non-Anticp	Non-CPP	-1.17	Non-Toxin	-1.17	Non-Toxin	-0.83	Non-Toxin	-0.83	Non-Toxin
starPep_08820	CPSHLDAFC	1	1.97	THP	THP	0.37	Anticp	Non-CPP	-1.04	Non-Toxin	-1.04	Non-Toxin	-0.9	Non-Toxin	-0.9	Non-Toxin
starPep_43956	KWDPPPPSPP	1	1.25	THP	THP	0.72	Anticp	CPP	-0.3	Non-Toxin	-0.3	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin
starPep_13827	RWCFRVCYGCCR	1	2.79	THP	THP	1.75	Anticp	Non-CPP	1.14	Toxin	1.14	Toxin	1.28	Toxin	1.28	Toxin
starPep_14535_It4_3_14W_2_4	AWWWRPPGFSPLRWA	0	3	THP	THP	1.26	Anticp	Non-CPP	-0.79	Non-Toxin	-0.79	Non-Toxin	-0.69	Non-Toxin	-0.69	Non-Toxin
starPep_25472_It4_5L_4_1_12	ALPYLWPWWPWSR	0	3.55	THP	THP	1.14	Anticp	Non-CPP	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.38	Non-Toxin	-0.38	Non-Toxin
starPep_15346_It4_6_2_5_7	AHPSWWM	0	3.5	THP	THP	0.65	Anticp	Non-CPP	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin
starPep_16808_L10	CNGRCGGKLA	1	2.3	THP	THP	1.42	Anticp	Non-CPP	-0.33	Non-Toxin	-0.33	Non-Toxin	-0.06	Non-Toxin	-0.06	Non-Toxin
starPep_17042_L5_1	LIWC	0	2.39	THP	THP	1.14	Anticp	Non-CPP	-0.59	Non-Toxin	-0.59	Non-Toxin	-0.36	Non-Toxin	-0.36	Non-Toxin
starPep_17042_It4_2_5_7_4_L5_1	CCGVL	1	3.22	THP	THP	1.36	Anticp	Non-CPP	-0.04	Non-Toxin	-0.04	Non-Toxin	-0.26	Non-Toxin	-0.26	Non-Toxin
starPep_18023_It4_1W_5_8_4C	WPSCHSWH	0	3.58	THP	THP	0.55	Anticp	Non-CPP	-0.79	Non-Toxin	-0.79	Non-Toxin	-0.5	Non-Toxin	-0.5	Non-Toxin
starPep_24644_It4_8C_14_3_5_L10	CGCGHCKHKN	0	2.84	THP	THP	1.13	Anticp	Non-CPP	-0.84	Non-Toxin	-0.84	Non-Toxin	-0.61	Non-Toxin	-0.61	Non-Toxin
starPep_26052_It4_9W_1_5_4	WVLCSRWPW	0	3.28	THP	THP	0.93	Anticp	Non-CPP	-0.46	Non-Toxin	-0.46	Non-Toxin	-0.33	Non-Toxin	-0.33	Non-Toxin
starPep_27346_It4_1_5_7_6_L5_1	CNKGC	0	3.42	THP	THP	1.49	Anticp	Non-CPP	-0.34	Non-Toxin	-0.34	Non-Toxin	-0.13	Non-Toxin	-0.13	Non-Toxin
starPep_27446_L15	KQCISLKGICKDLAC	1	0.58	THP	THP	1.81	Anticp	Non-CPP	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.21	Non-Toxin	-0.21	Non-Toxin
starPep_27924_L15	WCFRVAYRGISYRRC	1	1.16	THP	THP	1.13	Anticp	Non-CPP	-1.63	Non-Toxin	-1.63	Non-Toxin	-0.94	Non-Toxin	-0.94	Non-Toxin
starPep_29033_L10	KGKNKHKGHG	0	1.07	THP	THP	0.33	Anticp	Non-CPP	-0.69	Non-Toxin	-0.69	Non-Toxin	-0.64	Non-Toxin	-0.64	Non-Toxin
starPep_29033_L15	KKGKNKHKGHGHGHK	1	1.03	THP	THP	0.39	Anticp	Non-CPP	-1.16	Non-Toxin	-1.16	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin
starPep_35988_It4_4C_8_9_7_L5_2	GLCCC	1	3.18	THP	THP	1.02	Anticp	Non-CPP	-0.32	Non-Toxin	-0.32	Non-Toxin	-0.24	Non-Toxin	-0.24	Non-Toxin
starPep_36476_It4_1_5_7_2C	WCWVWGLRS	0	3.3	THP	THP	1.06	Anticp	Non-CPP	-0.38	Non-Toxin	-0.38	Non-Toxin	-1.16	Non-Toxin	-1.16	Non-Toxin
starPep_41900_It4_7C_1_2_8	CLWRWRCGY	0	3.51	THP	THP	1.92	Anticp	CPP	-0.3	Non-Toxin	-0.3	Non-Toxin	-0.22	Non-Toxin	-0.22	Non-Toxin
starPep_42404_It4_1W_9_2_5_L5	WRSAW	0	2.89	THP	THP	0.59	Anticp	Non-CPP	-0.75	Non-Toxin	-0.75	Non-Toxin	-0.73	Non-Toxin	-0.73	Non-Toxin
starPep_43120_L5_1	WKGRW	0	1.84	THP	THP	1.18	Anticp	Non-CPP	-0.69	Non-Toxin	-0.69	Non-Toxin	-0.52	Non-Toxin	-0.52	Non-Toxin
starPep_43502_It4_5C_7_9_1	CWAMCWSRC	0	3.71	THP	THP	1.37	Anticp	Non-CPP	-0.28	Non-Toxin	-0.28	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin
starPep_43502_It4_5S_7_9_2_L5_1	WCAMS	1	3.35	THP	THP	0.43	Anticp	Non-CPP	-0.46	Non-Toxin	-0.46	Non-Toxin	-0.6	Non-Toxin	-0.6	Non-Toxin
starPep_01400_L5_1	CGLSG	1	1.42	THP	THP	1.35	Anticp	Non-CPP	-0.89	Non-Toxin	-0.89	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin
starPep_08820_L5	CPSHL	1	2.58	THP	THP	0.56	Anticp	Non-CPP	-0.91	Non-Toxin	-0.91	Non-Toxin	-0.61	Non-Toxin	-0.61	Non-Toxin
starPep_07237_It4_1_6W_10_8C	WHWSYWLCPC	0	3.77	THP	THP	1.29	Anticp	Non-CPP	-0.96	Non-Toxin	-0.96	Non-Toxin	-0.67	Non-Toxin	-0.67	Non-Toxin
starPep_07237_It4_1_6C_10_8C	WHWSYCLCPW	0	3.77	THP	THP	1.29	Anticp	Non-CPP	-0.47	Non-Toxin	-0.47	Non-Toxin	-0.04	Non-Toxin	-0.04	Non-Toxin
starPep_02029_It3_4W_7_3_9	TPWWLSWHY	0	3.56	THP	THP	0.53	Anticp	Non-CPP	-1	Non-Toxin	-1	Non-Toxin	-0.51	Non-Toxin	-0.51	Non-Toxin
starPep_04689_L5_1	LRLRI	1	1.66	THP	THP	0.4	Anticp	Non-CPP	-0.93	Non-Toxin	-0.93	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin
starPep_04689_It4_1_3_5_8_L5_1	CIGCR	1	2.85	THP	THP	1.56	Anticp	Non-CPP	-0.42	Non-Toxin	-0.42	Non-Toxin	0	Non-Toxin	0	Non-Toxin
starPep_05157_It4_1C_10_12_4_L5_2	CPCKS	1	2.47	THP	THP	1.48	Anticp	Non-CPP	-0.59	Non-Toxin	-0.59	Non-Toxin	-0.13	Non-Toxin	-0.13	Non-Toxin
starPep_05293_It4_7_10_11_12_L5	LWWGA	1	2.59	THP	THP	1.27	Anticp	Non-CPP	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin
starPep_07335_It4_3C_1_8_L5	CCRLG	0	3.61	THP	THP	1.45	Anticp	Non-CPP	-0.4	Non-Toxin	-0.4	Non-Toxin	-0.33	Non-Toxin	-0.33	Non-Toxin
starPep_07641_L5_1	CCQEL	1	3.04	THP	THP	0.39	Anticp	Non-CPP	-0.07	Non-Toxin	-0.07	Non-Toxin	-0.14	Non-Toxin	-0.14	Non-Toxin
starPep_10014_L5_3	THWRI	0	1.23	THP	THP	0.66	Anticp	Non-CPP	-0.89	Non-Toxin	-0.89	Non-Toxin	-0.83	Non-Toxin	-0.83	Non-Toxin
starPep_10014_It4_9C_3_8_5_L5_1	CSLHW	0	3.28	THP	THP	0.85	Anticp	Non-CPP	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin
starPep_10092_It4_13W_2_1_3	WWWSPYVSRLLGWCL	1	3.23	THP	THP	0.73	Anticp	Non-CPP	-0.41	Non-Toxin	-0.41	Non-Toxin	-0.95	Non-Toxin	-0.95	Non-Toxin
starPep_10105_It4_1_11_4_17_L15	WWAHWLLSPRWSRPS	0	3.31	THP	THP	0.55	Anticp	Non-CPP	-1.15	Non-Toxin	-1.15	Non-Toxin	-1.38	Non-Toxin	-1.38	Non-Toxin
starPep_12257_It4_4W_6_1_8	WRPWPLYF	0	3.43	THP	THP	1.44	Anticp	Non-CPP	-0.32	Non-Toxin	-0.32	Non-Toxin	-0.28	Non-Toxin	-0.28	Non-Toxin
starPep_12257_It4_4L_6_1_7	WRPLPWFY	0	3.43	THP	THP	1.44	Anticp	Non-CPP	-0.87	Non-Toxin	-0.87	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin
starPep_12257_It4_4W_6_1_8_L5	WPLYF	1	2.95	THP	THP	0.77	Anticp	Non-CPP	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.42	Non-Toxin	-0.42	Non-Toxin
starPep_12276_It4_2_8_13_18_L10_2	WCPGYWWMCL	0	3.38	THP	THP	1.56	Anticp	Non-CPP	-0.04	Non-Toxin	-0.04	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin
starPep_12415_It4_1_6_3_12C	WSWGNWWARGHCM	0	3.23	THP	THP	0.9	Anticp	Non-CPP	-1.08	Non-Toxin	-1.08	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin

I. (cont.) Predicted activities of SET 1, conformed by 54 lead THPs.

ID	Sequence	HemoPI				
		SVM Score 1	SVM Score 2	SVM Score 3	SVM Score 3	SVM Score 4
starPep_24644	HKHGHGHLKHKNKLKKNGKH	0.49	0.49	0.62	0.77	0.44
starPep_43502	WWAMKWIRV	0.49	0.49	0.78	1	0.46
starPep_13108	SVSWGMKPSPRQ	0.48	0.48	0.22	0.14	0.38
starPep_35988	QRNKGLRHH	0.49	0.49	0.36	0.32	0.4
starPep_04689	RLRLRIGRR	0.48	0.48	0.79	0.96	0.42
starPep_36476	RFWVRGRRS	0.49	0.49	0.94	1	0.41
starPep_12276	PTSNHSPTSCPPTCPGYRWMCLRRF	0.49	0.49	0.49	0.43	0.39
starPep_10092	GVGSPYVSRLLGICL	0.51	0.51	0.58	0.63	0.43
starPep_07237	QHWSYGLRPG	0.48	0.48	0.21	0.16	0.41
starPep_12415	QSFGNQWARGHFM	0.49	0.49	0.16	0.11	0.49
starPep_08820	CPSHLDAFC	0.49	0.49	0.5	0.54	0.43
starPep_43956	KWDPPPPSPP	0.49	0.49	0.47	0.28	0.44
starPep_13827	RWCFRVCYGCCR	0.61	0.61	0.99	1	0.58
starPep_14535_It4_3_14W_2_4	AWWWRPPGFSPLRWA	0.67	0.65	0.49	0.49	0.43
starPep_25472_It4_5L_4_1_12	ALPYLWPWWPWSR	0.7	0.7	0.5	0.5	0.52
starPep_15346_It4_6_2_5_7	AHPSWWM	0.42	0.45	0.49	0.49	0.43
starPep_16808_L10	CNGRCGGKLA	0.91	0.71	0.48	0.48	0.38
starPep_17042_L5_1	LIWC	1	0.77	0.49	0.49	0.44
starPep_17042_It4_2_5_7_4_L5_1	CCGVL	1	0.73	0.49	0.49	0.44
starPep_18023_It4_1W_5_8_4C	WPSCHSWH	0.73	0.57	0.49	0.49	0.44
starPep_24644_It4_8C_14_3_5_L10	CGCGHCKHKN	0.93	0.67	0.49	0.49	0.41
starPep_26052_It4_9W_1_5_4	WVLCSRWPW	1	1	0.49	0.49	0.47
starPep_27346_It4_1_5_7_6_L5_1	CNKGC	0.97	0.66	0.49	0.49	0.43
starPep_27446_L15	KQCISLKGICKDLAC	1	0.91	0.48	0.48	0.5
starPep_27924_L15	WCFRVAYRGISYRRC	1	1	0.51	0.51	0.44
starPep_29033_L10	KGKNKHKGHG	0.67	0.57	0.49	0.49	0.44
starPep_29033_L15	KKGKNKHKGHGHGHK	0.71	0.58	0.49	0.15	0.44
starPep_35988_It4_4C_8_9_7_L5_2	GLCCC	0.95	0.58	0.49	0.49	0.44
starPep_36476_It4_1_5_7_2C	WCWVWGLRS	1	1	0.49	0.49	0.44
starPep_41900_It4_7C_1_2_8	CLWRWRCGY	1	1	0.49	0.49	0.48
starPep_42404_It4_1W_9_2_5_L5	WRSAW	0.88	0.65	0.49	0.49	0.44
starPep_43120_L5_1	WKGRW	0.97	0.67	0.49	0.49	0.46
starPep_43502_It4_5C_7_9_1	CWAMCWSRC	1	0.88	0.49	0.49	0.44
starPep_43502_It4_5S_7_9_2_L5_1	WCAMS	0.86	0.63	0.49	0.49	0.44
starPep_01400_L5_1	CGLSG	0.93	0.67	0.49	0.49	0.45
starPep_08820_L5	CPSHL	0.44	0.41	0.49	0.49	0.44
starPep_07237_It4_1_6W_10_8C	WHWSYWLCPC	1	0.83	0.49	0.49	0.44
starPep_07237_It4_1_6C_10_8C	WHWSYCLCPW	1	0.83	0.49	0.49	0.44
starPep_02029_It3_4W_7_3_9	TPWWLSWHY	0.68	0.61	0.49	0.49	0.43
starPep_04689_L5_1	LRLRI	0.87	0.76	0.49	0.49	0.42
starPep_04689_It4_1_3_5_8_L5_1	CIGCR	1	0.75	0.49	0.49	0.44
starPep_05157_It4_1C_10_12_4_L5_2	CPCKS	1	0.73	0.49	0.49	0.44
starPep_05293_It4_7_10_11_12_L5	LWWGA	0.96	0.66	0.49	0.49	0.44
starPep_07335_It4_3C_1_8_L5	CCRLG	1	0.74	0.49	0.43	
starPep_07641_L5_1	CCQEL	0.44	0.46	0.49	0.49	0.44
starPep_10014_L5_3	THWRI	0.78	0.65	0.49	0.49	0.44
starPep_10014_It4_9C_3_8_5_L5_1	CSLHW	1	0.72	0.49	0.49	0.44
starPep_10092_It4_13W_2_1_3	WWWSPYVSRLLGWCL	1	0.91	0.49	0.49	0.44
starPep_10105_It4_1_11_4_17_L15	WWAHWLLSPRWSRPS	0.8	0.71	0.49	0.49	0.43
starPep_12257_It4_4W_6_1_8	WRPWPLYF	0.61	0.67	0.51	0.51	0.6
starPep_12257_It4_4L_6_1_7	WRPLPWFY	0.61	0.67	0.51	0.51	0.6
starPep_12257_It4_4W_6_1_8_L5	WPLYF	0.48	0.5	0.49	0.49	0.45
starPep_12276_It4_2_8_13_18_L10_2	WCPGYWWMCL	1	0.84	0.49	0.49	0.44
starPep_12415_It4_1_6_3_12C	WSWGNWWARGHCM	0.86	0.67	0.49	0.49	0.42

J. Jalview, EMBOSS Cons and Seq2Logo results of MSA.

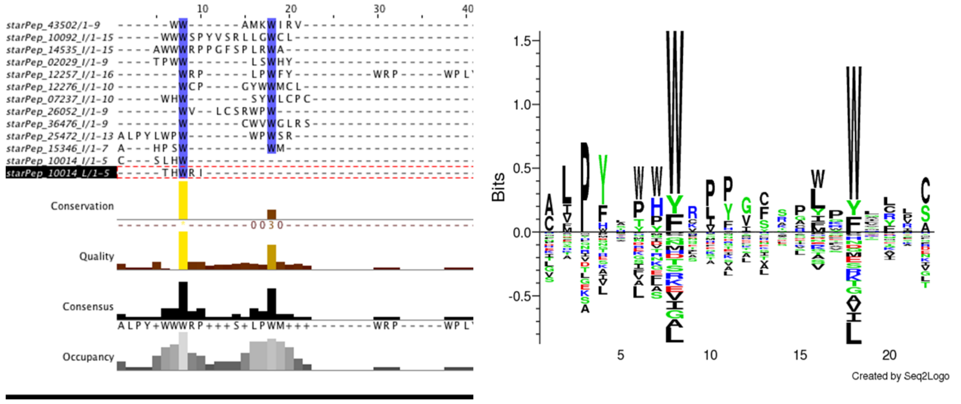


```
                                               10
starPep_26052_It4_9W_1_5_4/1-9              WVLCSRWP---W--------
starPep_25472_It4_5L_4_1_12/1-13            -ALPYLWPWWPWSR-----
starPep_07237_It4_1_6W_10_8C/1-10           WHW-SYWLCPC---------
starPep_12276_It4_2_8_13_18_L10_2/1-10      WCPGYWWMCL----------
starPep_15346_It4_6_2_5_7/1-7               -AHPSWWM------------
starPep_36476_It4_1_5_7_2C/1-9              ----WCW-VWGLRS-----
starPep_43502/1-9                           ----WWAMKWIRV------
starPep_12257_It4_4L_6_1_7/1-8              ------WRPLPWFY-----
starPep_12257_It4_4W_6_1_8/1-8              ------WRPWPLYF-----
starPep_10014_L5_3/1-5                      --------------THWR
starPep_02029_It3_4W_7_3_9/1-9              --TPWWL-SWHY-------
starPep_10092_It4_13W_2_1_3/1-15            ----WWWSPYVSRLLGWC
starPep_14535_It4_3_14W_2_4/1-15            ---AWWWRPPGFSPLRWA
starPep_10014_It4_9C_3_8_5_L5_1/1-5         -------------CSLHW--
```

CLUSTALW-O-MSA CLUSTER 2

Consensus Seq. estimated by EMBOSS_Cons

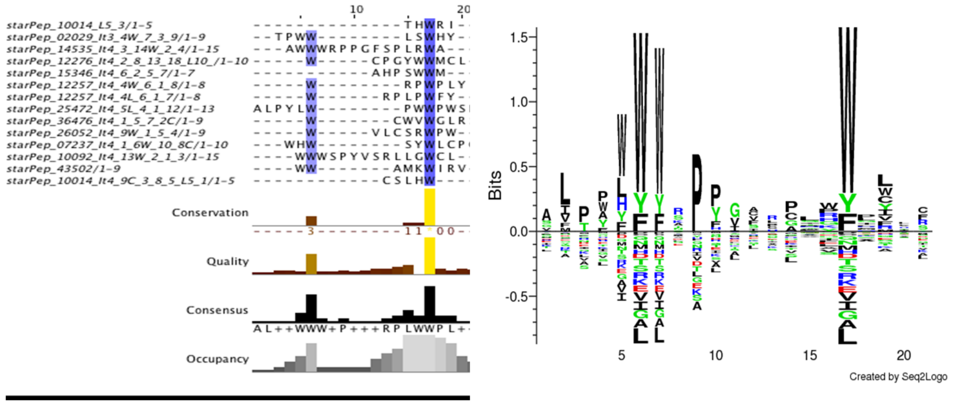
xxxxwwWxxxxxxxxxxxx



MAFFT-MSA CLUSTER 2

Consensus Seq. estimated by EMBOSS_Cons

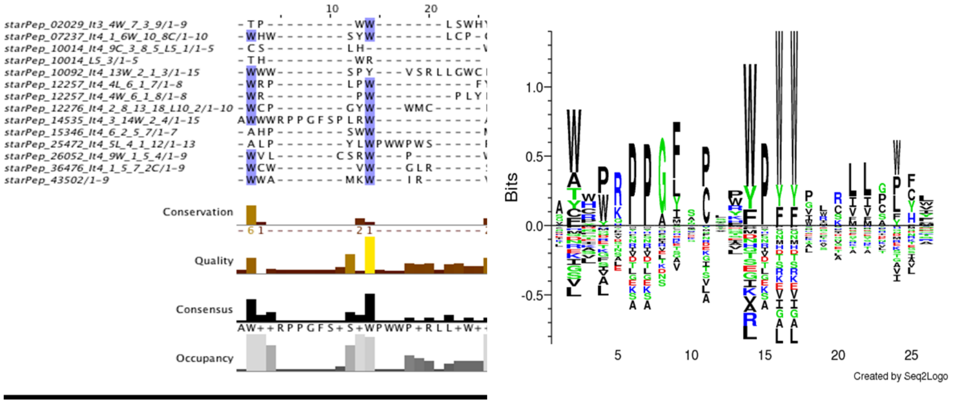
xxxxxxx**W**xxxxxxxxx**W**xxxx (pos 8 y 18)



MUSCLE-MSA CLUSTER 2

Consensus Seq. estimated by EMBOSS_Cons

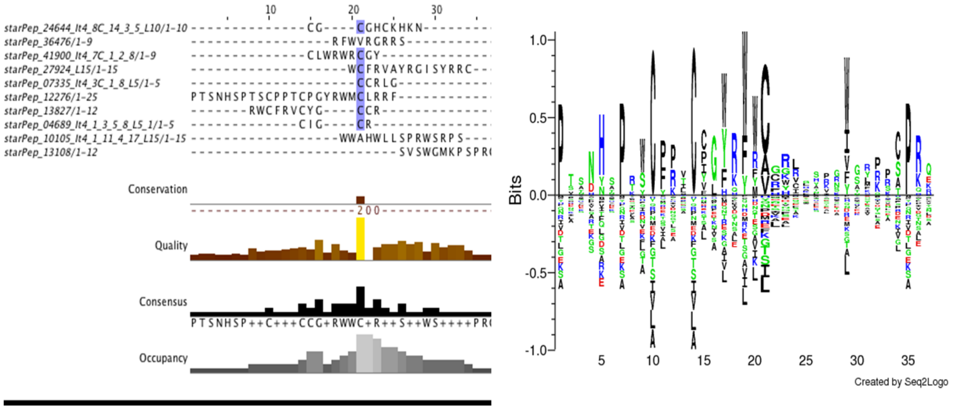
xxxxx**W**xxxxxxxxxx**W**xxxx (pos 6 – 17)



Tcoffe-MSA CLUSTER 2

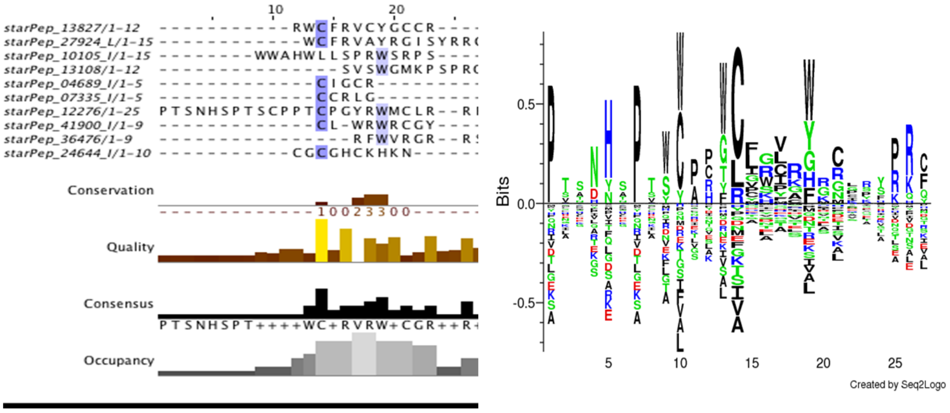
Consensus Seq. estimated by EMBOSS_Cons

x**W**xxxxxxxxxxx**W**xxxxxxxxxxxxx (ps 2 and 14)



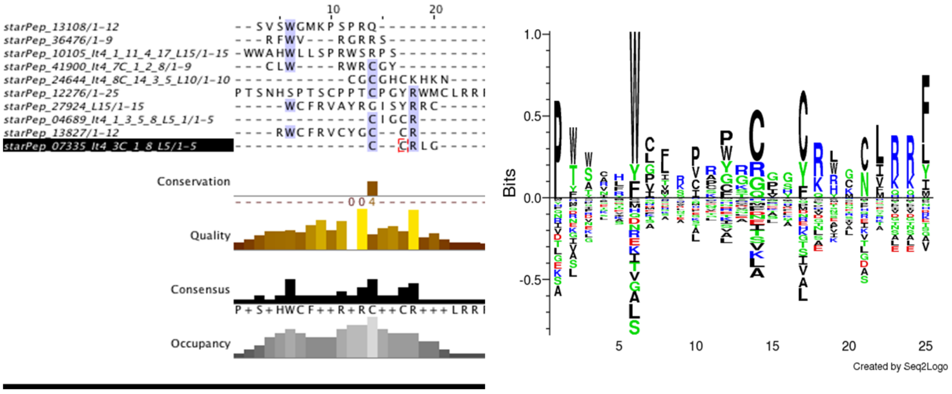
CLUSTALW-O-MSA CLUSTER 3

Consensus Seq. estimated by EMBOSS_Cons
nnnnnnnnnnnnnnnnnnnn**C**nnnnnnnnnnnnnnnn



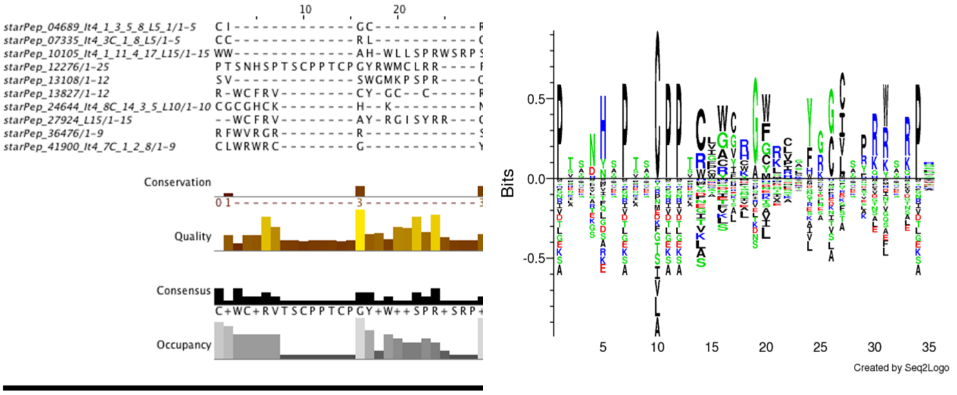
MAFFT-MSA CLUSTER 3

Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxxxxxxxx**C**xxx**rW**xxxxxxxxx



MUSCLE-MSA CLUSTER 3

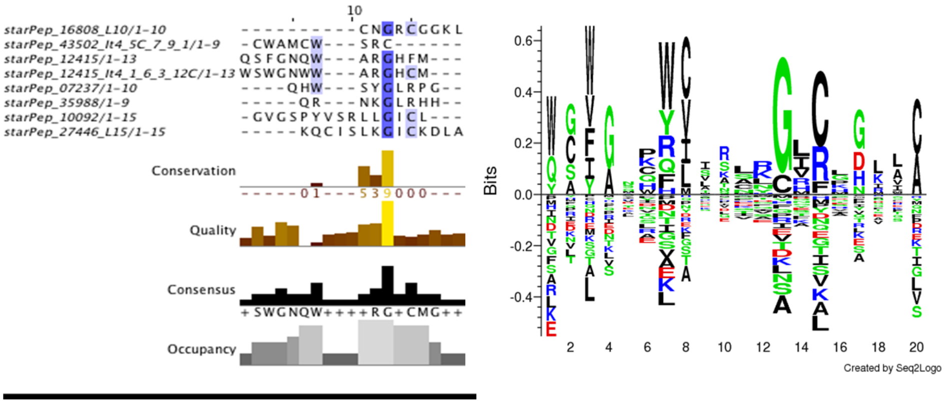
Consensus Seq. estimated by EMBOSS_Cons
xxxxx**W**xxxxxxxx**C**xxx**R**xxxxxxxx



Tcoffe-MSA CLUSTER 3

Consensus Seq. estimated by EMBOSS_Cons

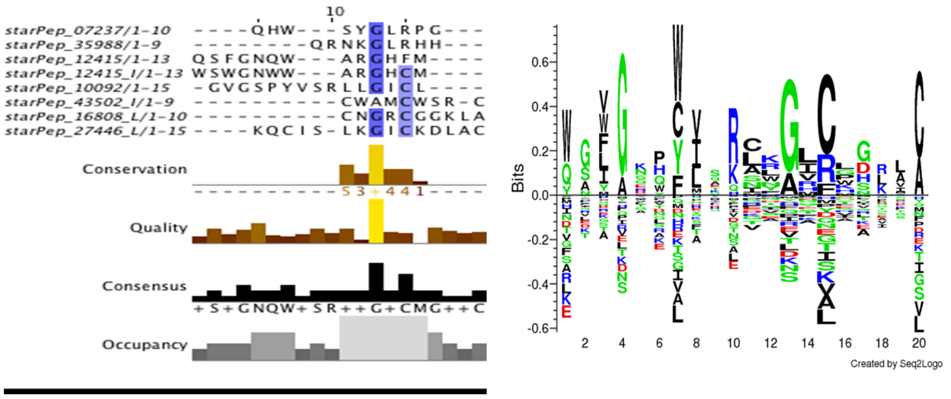
xxxxxxxxxxxxx**c**xxxxx**wR**xxxxxxxxxxxxxx



CLUSTALW-O-MSA CLUSTER 4

Consensus Seq. estimated by EMBOSS_Cons

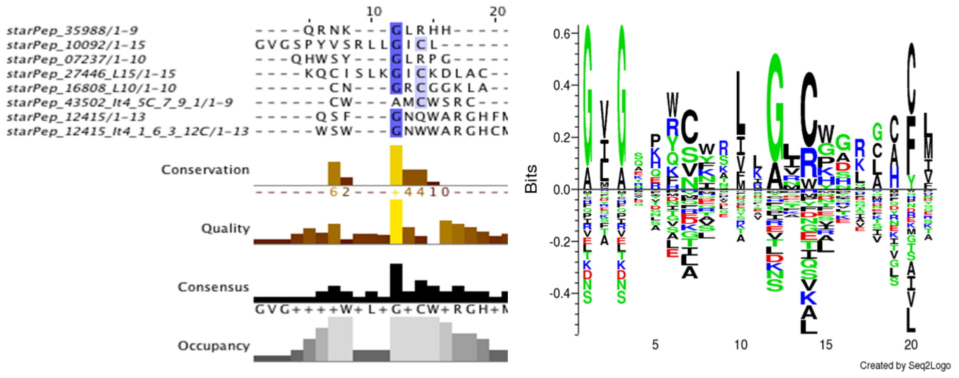
xxxxxxWxxx**SRGl**cxxxxx



MAFFT-MSA CLUSTER 4

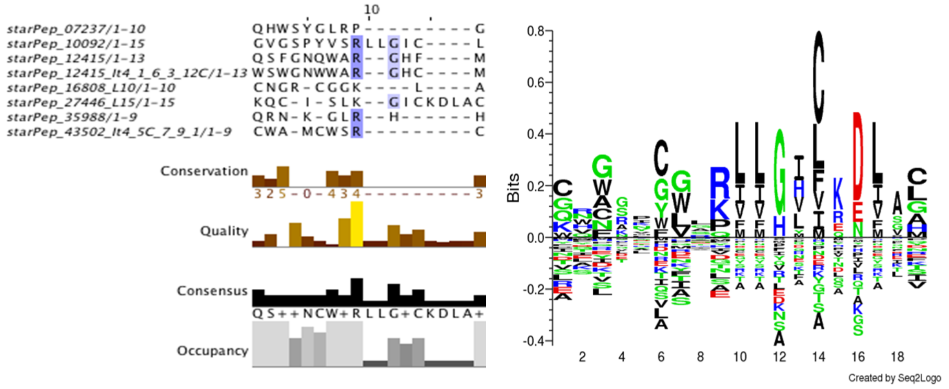
Consensus Seq. estimated by EMBOSS_Cons

xxxxxxwxxxx**kGLC**xxxxx



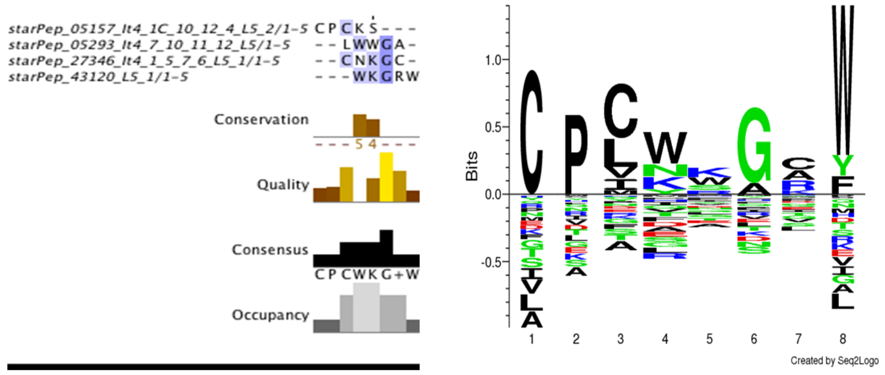
MUSCLE-MSA CLUSTER 4

Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxwxxx**GLc**xxrxxxx



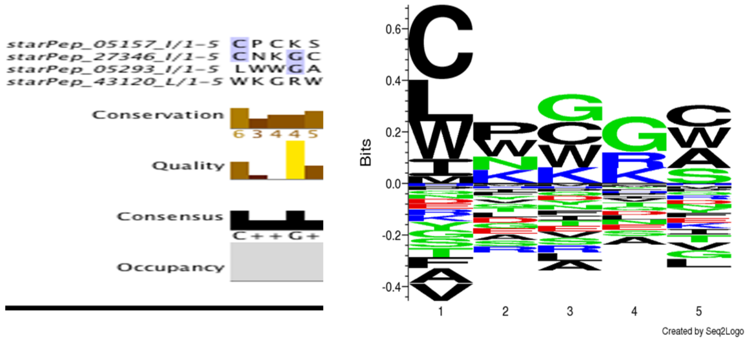
Tcoffe-MSA CLUSTER 4

Consensus Seq. estimated by EMBOSS_Cons
qxxxxxxa**Rxxg**xxxxxxx



CLUSTALW-O-MSA CLUSTER 1-5

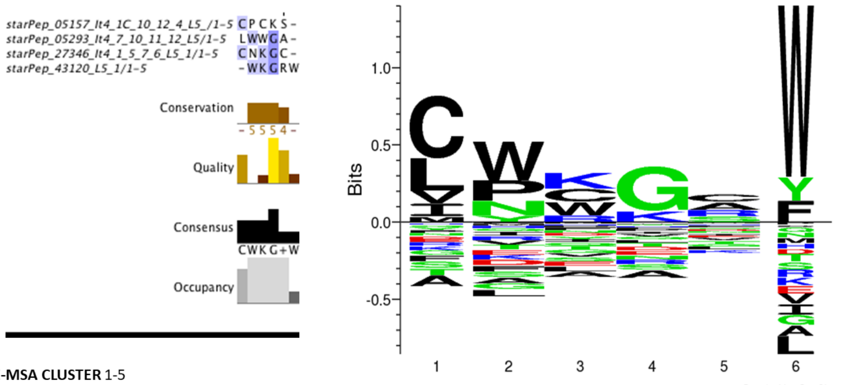
Consensus Seq. estimated by EMBOSS_Cons
xxcw**kG**xx



```
starPep_05157_I/1-5  CPCKS
starPep_27346_I/1-5  CNKGC
starPep_05293_I/1-5  LWWGA
starPep_43120_L/1-5 WKGRW
```

MAFFT-MSA CLUSTER 1-5

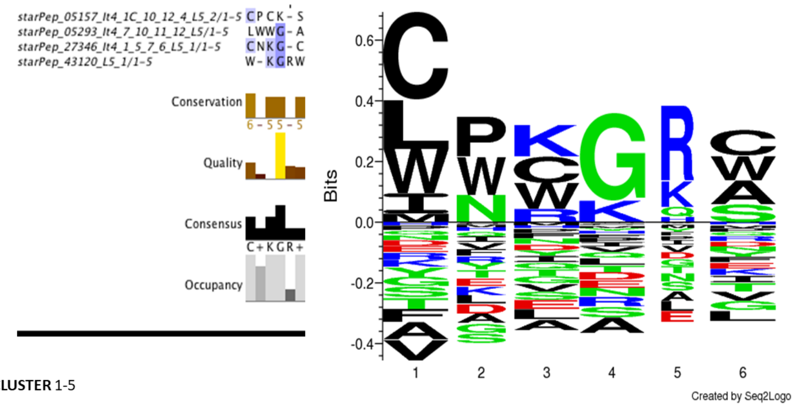
Consensus Seq. estimated by EMBOSS_Cons
cxxga



```
starPep_05157_It4_1C_10_12_4_L5_/1-5 CPCKS-
starPep_05293_It4_7_10_11_12_L5/1-5  LWWGA-
starPep_27346_It4_1_5_7_6_L5_1/1-5   CNKGC-
starPep_43120_L5_1/1-5               -WKGRW
```

MUSCLE-MSA CLUSTER 1-5

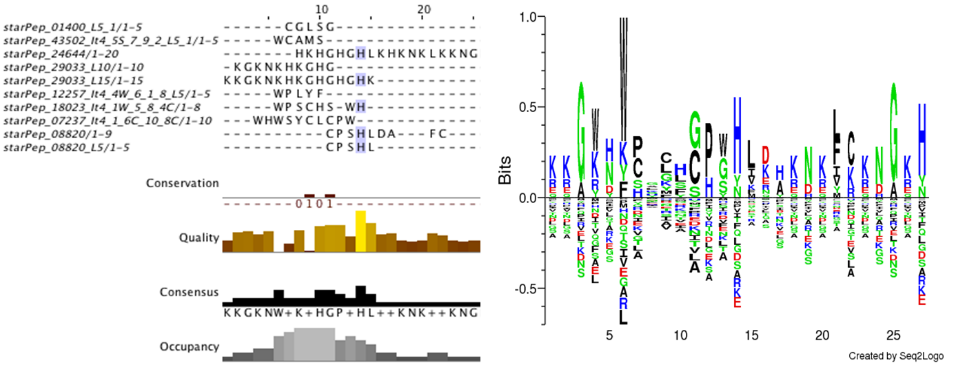
Consensus Seq. estimated by EMBOSS_Cons
cw**kGa**x



```
starPep_05157_It4_1C_10_12_4_L5_2/1-5 CPCK-S
starPep_05293_It4_7_10_11_12_L5/1-5   LWWG-A
starPep_27346_It4_1_5_7_6_L5_1/1-5    CNKG-C
starPep_43120_L5_1/1-5                W-KGRW
```

Tcoffe-MSA CLUSTER 1-5

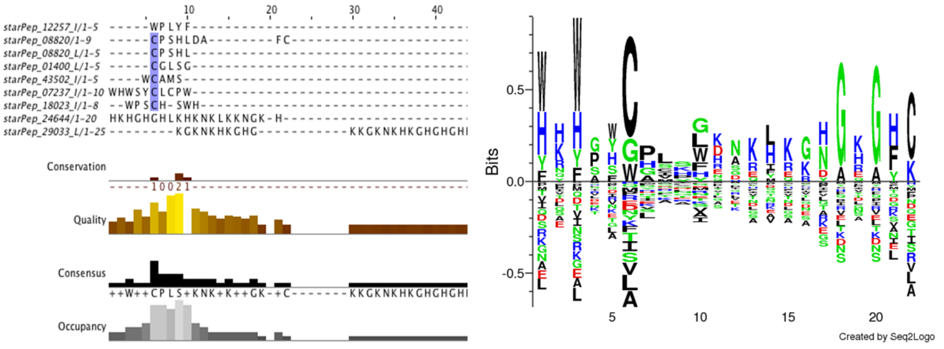
Consensus Seq. estimated by EMBOSS_Cons
cx**kGxa**



```
                                          10        20
starPep_01400_L5_1/1-5                -------CGLSG---------------
starPep_43502_It4_5S_7_9_2_L5_1/1-5   -----WCAMS-----------------
starPep_24644/1-20                    -------HKHGHGHLKHKNKLKKNG
starPep_29033_L10/1-10                -KGKNKHKGHG----------------
starPep_29033_L15/1-15                KKGKNKHKGHGHGHK------------
starPep_12257_It4_4W_6_1_8_L5/1-5     -----WPLYF-----------------
starPep_18023_It4_1W_5_8_4C/1-8       -----WPSCHS-WH-------------
starPep_07237_It4_1_6C_10_8C/1-10     ---WHWSYCLCPW--------------
starPep_08820/1-9                     ----------CPSHLDA---FC---
starPep_08820_L5/1-5                  ----------CPSHL------------
```

CLUSTALW-O-MSA CLUSTER 6

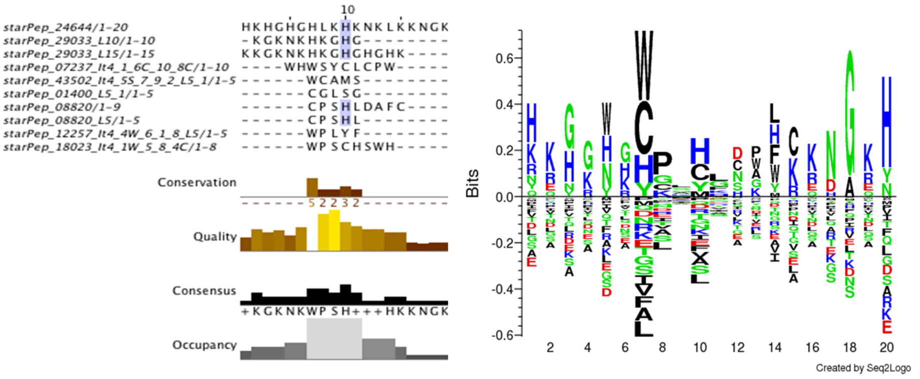
Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxxxxxxxhxxxxxxxxxxxxx



```
                               10        20        30        40
starPep_12257_I/1-5     -----WPLYF---------------------------------
starPep_08820/1-9       -----CPSHLDA--------FC---------------------
starPep_08820_L/1-5     -----CPSHL---------------------------------
starPep_01400_L/1-5     -----CGLSG---------------------------------
starPep_43502_I/1-5     ----WCAMS----------------------------------
starPep_07237_I/1-10    WHWSYCLCPW---------------------------------
starPep_18023_I/1-8     --WPSCH-SWH--------------------------------
starPep_24644/1-20      HKHGHGHLKHKNKLKKNGK-H----------------------
starPep_29033_L/1-25    --------KGKNKHKGHG------------KKGKNKHKGHGHGHI
```

MAFFT-MSA CLUSTER 6

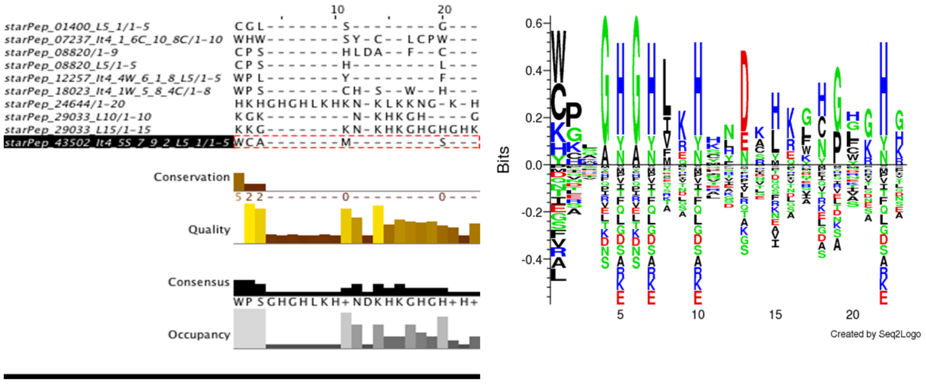
Consensus Seq. estimated by EMBOSS_Cons
xxxxxCxxxxxxxxxxxxxxxxx



```
                                                10
starPep_24644/1-20                    HKHGHGHLKHKNKLKKNGK
starPep_29033_L10/1-10                -KGKNKHKGHG--------
starPep_29033_L15/1-15                KKGKNKHKGHGHGHK----
starPep_07237_It4_1_6C_10_8C/1-10     ----WHWSYCLCPW-----
starPep_43502_It4_5S_7_9_2_L5_1/1-5   ------WCAMS--------
starPep_01400_L5_1/1-5                ------CGLSG--------
starPep_08820/1-9                     ------CPSHLDAFC----
starPep_08820_L5/1-5                  ------CPSHL--------
starPep_12257_It4_4W_6_1_8_L5/1-5     ------WPLYF--------
starPep_18023_It4_1W_5_8_4C/1-8       ------WPSCHSWH-----
```

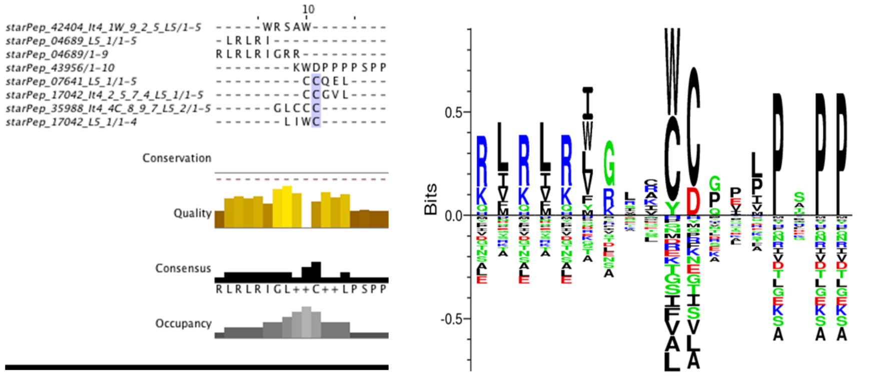
MUSCLE-MSA CLUSTER 6

Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxxxHxxxxxxxxxx



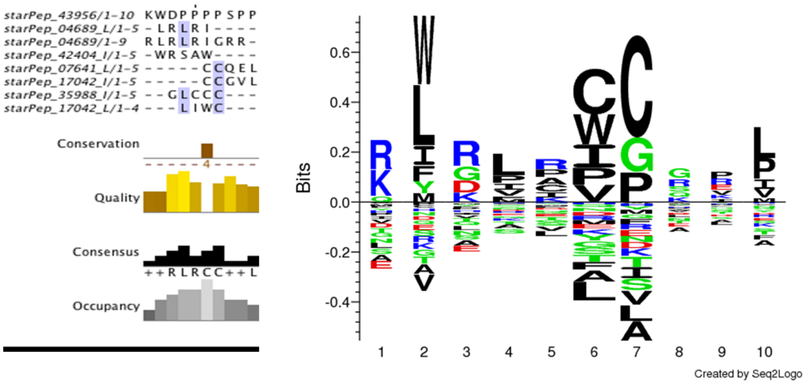
Tcoffe-MSA CLUSTER 6

Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxxxxxxxxxxxxxxxxx



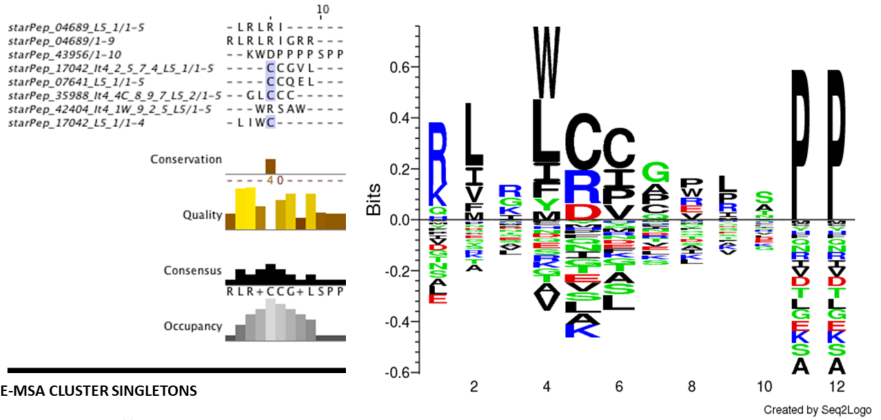
CLUSTALW-O-MSA CLUSTER SINGLETONS

Consensus Seq. estimated by EMBOSS_Cons
nnnnnnnnnn**c**nnnnnnn



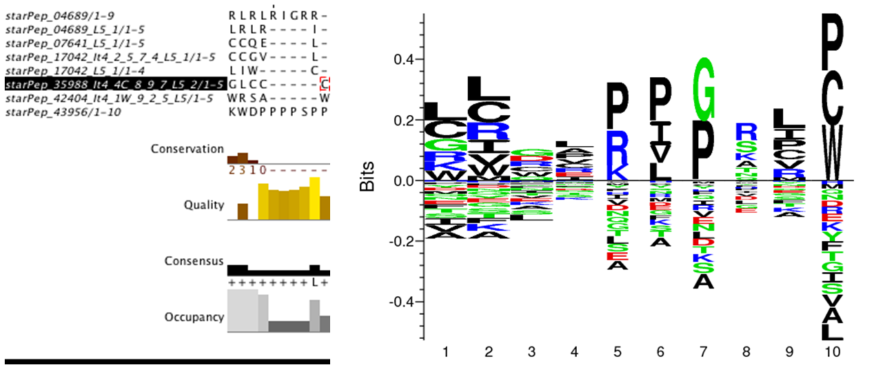
MAFFT-MSA CLUSTER SINGLETONS

Consensus Seq. estimated by EMBOSS_Cons
xxxlxxcxxx



MUSCLE-MSA CLUSTER SINGLETONS

Consensus Seq. estimated by EMBOSS_Cons
xxxxcxxxxxxxx



Tcoffe-MSA CLUSTER SINGLETONS

Consensus Seq. estimated by EMBOSS_Cons
xxxxxxxxxx

K. FASTA file of 150 cell-penetrating sequences derived from 54 lead hits.

```
>starPep_24644_CPP1
HKHGHGHLKHKNKLKKKGKH
>starPep_24644_CPP2
HKHKHGHLKHKNKLKKNGKH
>starPep_24644_CPP3
HKHGHGHLKHKNKLKKNKKH
>starPep_24644_CPP4
HKHGHGHLKHKNKLKKKKKH
>starPep_43502_CPP1
WWKKKWKKK
>starPep_43502_CPP2
CWAKKWKKK
>starPep_43502_CPP3
CWKKKWKKK
>starPep_13108_CPP1
SVPWRMKPSPRQ
>starPep_13108_CPP2
SVSWRMKPSPRQ
>starPep_13108_CPP3
SVPWGMKPSPRQ
>starPep_13108_CPP4
SVRWWMKPSPRQ
>starPep_13108_CPP5
SVRWGWKPSPRQ
>starPep_35988_CPP1
SRRHRSRHH
>starPep_35988_CPP2
SRRARSRHH
>starPep_35988_CPP3
SRRRRSRHH
>starPep_04689_CPP1
RLRLRRRRR
>starPep_04689_CPP2
RLRRRQRRR
>starPep_04689_CPP3
RGRRRIRRR
>starPep_04689_CPP4
RLRRRRRRR
>starPep_04689_CPP5
RRRRRIRRR
>starPep_36476_CPP1
RRWRRRRRS
>starPep_36476_CPP2
RFRRRRRRS
>starPep_36476_CPP3
RFKRRRRRR
>starPep_12276_CPP1
KTRNHSPTSCPPTCPRYRWMCLRRR
>starPep_12276_CPP2
KTRNHRPTSCPPTCPGYRWMCLRRF
>starPep_12276_CPP3
KTRNHSPTSCPPTCPRYRWMCLRRF
>starPep_10092_CPP1
GVGSRRRSRLLGICL
>starPep_10092_CPP2
GVGSRRRSRRLGICL
>starPep_10092_CPP3
GVGSPRRRRRLGICL
>starPep_10092_CPP4
GVGSPRRSRLLGICL
>starPep_10092_CPP5
GVRSRRRSRRLGICL
>starPep_07237_CPP1
QHWSRRLRPG
>starPep_07237_CPP2
QHWSYRLRPR
>starPep_07237_CPP3
QRWSRRLRPG
>starPep_12415_CPP1
QSFRNQWARRHFM
>starPep_12415_CPP2
QSFRNQWRRRHFM
>starPep_12415_CPP3
QSFRRQWARRHFM
>starPep_08820_CPP1
CRRRRDRGC
>starPep_08820_CPP2
GRRRRDRGC
>starPep_08820_CPP3
CRRRRNRGC
>starPep_43956_CPP1
KWRPPPPSPP
>starPep_43956_CPP2
KWRPPPPPPP
>starPep_43956_CPP3
KWRPPPPRPP
>starPep_13827_CPP1
RWCFRRCRGRCR
>starPep_13827_CPP2
RWCFRRCRRCCR
>starPep_13827_CPP3
RWCRRRCRGCCR
>starPep_14535_It4_3_14W_2_4_CPP1
AWWWRPPRFSPLRWA
>starPep_14535_It4_3_14W_2_4_CPP2
AWWWRPPGFRPLRWA
>starPep_14535_It4_3_14W_2_4_CPP3
AWWWRPPRFRPLRWA
>starPep_25472_It4_5L_4_1_12_CPP1
ALPRLWPWWPWSR
>starPep_25472_It4_5L_4_1_12_CPP2
ALPYRWPWWPWSR
>starPep_25472_It4_5L_4_1_12_CPP3
ALPYLWPWRPWSR
>starPep_15346_It4_6_2_5_7_CPP1
ARRRRWW
>starPep_15346_It4_6_2_5_7_CPP2
GRRRRWM
>starPep_15346_It4_6_2_5_7_CPP3
ARRRRWC
>starPep_16808_L10_CPP1
CNGRCRGKLR
>starPep_16808_L10_CPP2
CNGRCRGKLK
>starPep_16808_L10_CPP3
CNGRRRGKLA
>starPep_17042_L5_1_CPP1
RRWR
>starPep_17042_L5_1_CPP2
RIRR
>starPep_17042_It4_2_5_7_4_L5_1_CPP1
SLPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2
CLPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP3
WLPVM
>starPep_18023_It4_1W_5_8_4C_CPP1
GRKKRRWC
>starPep_18023_It4_1W_5_8_4C_CPP2
GRKKRRWP
>starPep_24644_It4_8C_14_3_5_L10_CPP1
CGCGHKKHKK
>starPep_24644_It4_8C_14_3_5_L10_CPP2
KGCGHCKHKK
>starPep_24644_It4_8C_14_3_5_L10_CPP3
CGCKHCKHKK
>starPep_26052_It4_9W_1_5_4_CPP1
RRRRRRWPW
>starPep_26052_It4_9W_1_5_4_CPP2
GRRRRRWPW
>starPep_26052_It4_9W_1_5_4_CPP3
WRRRRRWRW
>starPep_27346_It4_1_5_7_6_L5_1_CPP1
RRKRR
>starPep_27346_It4_1_5_7_6_L5_1_CPP2
RRRGR
>starPep_27446_L15_CPP1
KQCISRKRICKRLAC
>starPep_27446_L15_CPP2
KQCISRKRICKKLAC
>starPep_27446_L15_CPP3
KQCISRKGICKKLAC
>starPep_27924_L15_CPP1
WCFRVRYRGRSYRRC
>starPep_27924_L15_CPP2
WCFRRRYRGISYRRC
>starPep_27924_L15_CPP3
WCFRRRYRRISYRRC
>starPep_29033_L10_CPP1
KGKNKHKGKK
>starPep_29033_L10_CPP2
KGKNKHKKHK
>starPep_29033_L10_CPP3
KGKNKHKKKK
>starPep_29033_L15_CPP1
KKGKNKHKKHGHGHK
>starPep_29033_L15_CPP2
KKGKNKHKRHKHKKK
>starPep_29033_L15_CPP3
KKGKNKHKRKKHKHK
>starPep_35988_It4_4C_8_9_7_L5_2_CPP1
KLKKK
>starPep_35988_It4_4C_8_9_7_L5_2_CPP2
KKCKK
>starPep_36476_It4_1_5_7_2C_CPP1
RCWRRRLRR
>starPep_36476_It4_1_5_7_2C_CPP2
RRWRRRLRS
>starPep_36476_It4_1_5_7_2C_CPP3
RCWRRRRRS
>starPep_41900_It4_7C_1_2_8_CPP1
RLWRRRRGR
>starPep_41900_It4_7C_1_2_8_CPP2
RLWRRRRGG
>starPep_41900_It4_7C_1_2_8_CPP3
RLRRRRRGR
>starPep_42404_It4_1W_9_2_5_L5_CPP1
RRRAR
>starPep_42404_It4_1W_9_2_5_L5_CPP2
RRRRW
>starPep_43120_L5_1_CPP1
RKGRR
>starPep_43120_L5_1_CPP2
RKRRR
>starPep_43120_L5_1_CPP3
RRGRR
>starPep_43502_It4_5C_7_9_1_CPP1
CRARRRRRC
>starPep_43502_It4_5C_7_9_1_CPP2
CRRRRWRRC
>starPep_43502_It4_5C_7_9_1_CPP3
GRARRRRRC
>starPep_43502_It4_5S_7_9_2_L5_1_CPP1
KKALK
>starPep_43502_It4_5S_7_9_2_L5_1_CPP2
VCALK
>starPep_01400_L5_1_CPP1
RRRSR
>starPep_01400_L5_1_CPP2
RQRSR
>starPep_08820_L5_CPP1
CPKKK
>starPep_08820_L5_CPP2
CWKKK
>starPep_07237_It4_1_6W_10_8C_CPP1
WHWSRWLCPC
>starPep_07237_It4_1_6W_10_8C_CPP2
WHWSYWLRPC
>starPep_07237_It4_1_6W_10_8C_CPP3
WRWSRWLCPC
>starPep_07237_It4_1_6C_10_8C_CPP1
WHWSRRLCPW
>starPep_07237_It4_1_6C_10_8C_CPP2
WHWSRCLRPW
>starPep_07237_It4_1_6C_10_8C_CPP3
WHWSRCLCPR
>starPep_02029_It3_4W_7_3_9_CPP1
SRRWRSRHH
>starPep_02029_It3_4W_7_3_9_CPP2
TRRHRSRHH
>starPep_02029_It3_4W_7_3_9_CPP3
GRRWRSRHH
>starPep_04689_L5_1_CPP1
RRLRR
>starPep_04689_L5_1_CPP2
RRWRR
>starPep_04689_L5_1_CPP3
RRPRR
>starPep_04689_It4_1_3_5_8_L5_1_CPP1
RIRRR
>starPep_05157_It4_1C_10_12_4_L5_2_CPP2
CKKKK
>starPep_05293_It4_7_10_11_12_L5_CPP1
LLWRA
>starPep_05293_It4_7_10_11_12_L5_CPP2
GLWRA
>starPep_05293_It4_7_10_11_12_L5_CPP3
RLWRA
>starPep_07335_It4_3C_1_8_L5_CPP1
KKRKK
>starPep_07335_It4_3C_1_8_L5_CPP2
KCKKK
>starPep_10014_L5_3_CPP1
RRWRI
>starPep_10014_L5_3_CPP2
RRRRI
>starPep_10014_It4_9C_3_8_5_L5_1_CPP1
RSRHH
>starPep_10014_It4_9C_3_8_5_L5_1_CPP2
SSRHH
>starPep_10014_It4_9C_3_8_5_L5_1_CPP3
SRRHH
>starPep_10092_It4_13W_2_1_3_CPP1
WWWSPRVSRLLGWCL
>starPep_10092_It4_13W_2_1_3_CPP2
WWWSPRRSRLLGWCL
>starPep_10092_It4_13W_2_1_3_CPP3
WWWSRRVSRLLGWCL
>starPep_10105_It4_1_11_4_17_L15_CPP1
WWAHWRLSPRWSRPS
>starPep_10105_It4_1_11_4_17_L15_CPP2
WWAHWLLSPRWSRPR
>starPep_10105_It4_1_11_4_17_L15_CPP3
WWAHWRLSPRWSRPR
>starPep_12257_It4_4W_6_1_8_CPP1
WRPRRRYR
>starPep_12257_It4_4W_6_1_8_CPP2
WRPRRRRR
>starPep_12257_It4_4W_6_1_8_CPP3
GRPRRRYR
>starPep_12257_It4_4L_6_1_7_CPP1
WRPKWRFK
>starPep_12257_It4_4L_6_1_7_CPP2
WRFKWWFK
>starPep_12257_It4_4W_6_1_8_L5_CPP1
RPRRR
>starPep_12257_It4_4W_6_1_8_L5_CPP2
RRRYR
>starPep_12276_It4_2_8_13_18_L10_2_CPP1
WCPRRWWMCL
>starPep_12276_It4_2_8_13_18_L10_2_CPP2
WCPRRWWRCL
>starPep_12276_It4_2_8_13_18_L10_2_CPP3
WCRRRWWMCL
>starPep_12415_It4_1_6_3_12C_CPP1
WSWGNWWARRHCM
>starPep_12415_It4_1_6_3_12C_CPP2
WSWRNWWARRHCM
>starPep_12415_It4_1_6_3_12C_CPP3
WSWGNWWRRRHCM
```

L. Predicted activities of SET 2, conformed by 42 lead THPs with optimized cell-penetrating activity.

ID	Sequence	PlifePred	TumorHPD		THPep		AntiCP			ToxinPred								CellPPD			
		Half-time (seconds)	SVM Score		SVM Score	SVM Score 1		SVM Score 2		SVM Score 1		SVM Score 2		SVM Score 3		SVM Score 4		SVM Score 1		SVM Score 2	
starPep_43502	WWAMKWIRV	849.01	1.58	THP	THP	0.5	Anticp	1.15	Anticp	-0.49	Non-Toxin	-0.49	Non-Toxin	-0.81	Non-Toxin	-0.81	Non-Toxin	-0.24	Non-CPP	0.09	CPP
starPep_13108_CPP1	SVPWRMKPSPRQ	868.51	0.86	THP	THP	0.8	Anticp	-0.07	Non-Anticp	-1.24	Non-Toxin	-1.24	Non-Toxin	-1.46	Non-Toxin	-1.46	Non-Toxin	0.09	CPP	0.39	CPP
starPep_13108_CPP4	SVRWWMKPSPRQ	999.11	1.07	THP	THP	0.56	Anticp	-0.01	Non-Anticp	-1.47	Non-Toxin	-1.47	Non-Toxin	-1.77	Non-Toxin	-1.77	Non-Toxin	0.09	CPP	0.4	CPP
starPep_13108_CPP5	SVRWGWKPSPRQ	1012.91	0.79	THP	THP	0.66	Anticp	0.28	Anticp	-1	Non-Toxin	-1	Non-Toxin	-1.39	Non-Toxin	-1.39	Non-Toxin	0.09	CPP	0.37	CPP
starPep_04689	RLRLRIGRR	845.01	1.14	THP	THP	0.94	Anticp	-0.04	Non-Anticp	-1.22	Non-Toxin	-1.22	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.24	Non-CPP	1.04	CPP
starPep_12276	PTSNHSPTSCPPTCPGYRWMCLRRF	841.71	1.71	THP	THP	0.8	Anticp	0.94	Anticp	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.68	Non-Toxin	-0.68	Non-Toxin	0.09	CPP	-0.24	Non-CPP
starPep_12276_CPP1	KTRNHSPTSCPPTCPRYRWMCLRRR	924.81	1.37	THP	THP	0.8	Anticp	1.08	Anticp	-0.22	Non-Toxin	-0.22	Non-Toxin	-0.41	Non-Toxin	-0.41	Non-Toxin	0.09	CPP	0.38	CPP
starPep_07237_CPP2	QHWSYRLRPR	955.31	1.58	THP	THP	0.9	Anticp	0.6	Anticp	-1.4	Non-Toxin	-1.4	Non-Toxin	-1.1	Non-Toxin	-1.1	Non-Toxin	0.09	CPP	0.34	CPP
starPep_08820	CPSHLDAFC	689.51	1.97	THP	THP	0.99	Anticp	0.37	Anticp	-1.04	Non-Toxin	-1.04	Non-Toxin	-0.9	Non-Toxin	-0.9	Non-Toxin	-0.24	Non-CPP	-0.46	Non-CPP
starPep_43956	KWDPPPPSPP	835.11	1.25	THP	THP	0.84	Anticp	0.74	Anticp	-0.3	Non-Toxin	-0.3	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin	0.09	CPP	0.15	CPP
starPep_43956_CPP3	KWRPPPPRPP	841.01	1.55	THP	THP	0.69	Anticp	0.94	Anticp	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.12	Non-Toxin	-0.12	Non-Toxin	0.09	CPP	0.62	CPP
starPep_16808_L10	CNGRCGGKLA	874.81	2.3	THP	THP	0.74	Anticp	1.42	Anticp	-0.33	Non-Toxin	-0.33	Non-Toxin	-0.06	Non-Toxin	-0.06	Non-Toxin	0.09	CPP	-0.19	Non-CPP
starPep_17042_L5_1	LIWC	832.81	2.39	THP	THP	0.95	Anticp	1.37	Anticp	-0.59	Non-Toxin	-0.59	Non-Toxin	-0.36	Non-Toxin	-0.36	Non-Toxin	-0.24	Non-CPP	-0.24	Non-CPP
starPep_17042_It4_2_5_7_4_L5_1	CCGVL	834.11	3.22	THP	THP	1.21	Anticp	1.36	Anticp	-0.04	Non-Toxin	-0.04	Non-Toxin	-0.26	Non-Toxin	-0.26	Non-Toxin	-0.24	Non-CPP	-0.27	Non-CPP
starPep_17042_It4_2_5_7_4_L5_1_CPP3	WLPVM	834.71	1.49	THP	THP	1.09	Anticp	0.16	Anticp	-0.73	Non-Toxin	-0.73	Non-Toxin	-0.29	Non-Toxin	-0.29	Non-Toxin	-0.24	Non-CPP	0.14	CPP
starPep_29033_L15_CPP1	KKGKNKHKKHGHGHK	823.51	1.02	THP	THP	0.93	Anticp	0.34	Anticp	-1.16	Non-Toxin	-1.16	Non-Toxin	-0.93	Non-Toxin	-0.93	Non-Toxin	0.09	CPP	0.35	CPP
starPep_35988_It4_4C_8_9_7_L5_2	GLCCC	834.81	3.18	THP	THP	-0.19	Non-Anticp	1.02	Anticp	-0.32	Non-Toxin	-0.32	Non-Toxin	-0.24	Non-Toxin	-0.24	Non-Toxin	-0.24	Non-CPP	-0.25	Non-CPP
starPep_35988_It4_4C_8_9_7_L5_2_CPP2	KKCKK	834.71	1.07	THP	THP	0.91	Anticp	0.68	Anticp	-0.74	Non-Toxin	-0.74	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin	-0.24	Non-CPP	0.14	CPP
starPep_43502_It4_5S_7_9_2_L5_1	WCAMS	830.41	3.35	THP	THP	0.85	Anticp	0.43	Anticp	-0.46	Non-Toxin	-0.46	Non-Toxin	-0.6	Non-Toxin	-0.6	Non-Toxin	-0.24	Non-CPP	-0.26	Non-CPP
starPep_43502_It4_5S_7_9_2_L5_1_CPP1	KKALK	823.31	0.19	THP	non-THP	0.95	Anticp	0.88	Anticp	-1.13	Non-Toxin	-1.13	Non-Toxin	-0.71	Non-Toxin	-0.71	Non-Toxin	-0.24	Non-CPP	0.06	CPP
starPep_01400_L5_1	CGLSG	829.91	1.42	THP	THP	0.7	Anticp	1.35	Anticp	-0.89	Non-Toxin	-0.89	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin	-0.24	Non-CPP	-0.29	Non-CPP
starPep_08820_L5_CPP1	CPKKK	833.31	1.07	THP	THP	0.69	Anticp	0.39	Anticp	-0.83	Non-Toxin	-0.83	Non-Toxin	-0.65	Non-Toxin	-0.65	Non-Toxin	-0.24	Non-CPP	0.04	CPP
starPep_04689_It4_1_3_5_8_L5_1	CIGCR	832.81	2.85	THP	THP	1.25	Anticp	1.56	Anticp	-0.42	Non-Toxin	-0.42	Non-Toxin	0	Non-Toxin	0	Non-Toxin	-0.24	Non-CPP	-0.24	Non-CPP
starPep_05157_It4_1C_10_12_4_L5_2	CPCKS	833.51	2.47	THP	THP	0.85	Anticp	1.48	Anticp	-0.59	Non-Toxin	-0.59	Non-Toxin	-0.13	Non-Toxin	-0.13	Non-Toxin	-0.24	Non-CPP	-0.24	Non-CPP
starPep_12415_CPP3	QSFRRQWARRHFM	883.21	0.63	THP	THP	0.85	Anticp	0.13	Anticp	-0.92	Non-Toxin	-0.92	Non-Toxin	-0.89	Non-Toxin	-0.89	Non-Toxin	0.98	CPP	0.45	CPP
starPep_14535_It4_3_14W_2_4_CPP1	AWWWRPPRFSPLRWA	847.31	2.93	THP	THP	0.79	Anticp	1.16	Anticp	-0.98	Non-Toxin	-0.98	Non-Toxin	-1.13	Non-Toxin	-1.13	Non-Toxin	0.98	CPP	0.21	CPP
starPep_25472_It4_5L_4_1_12_CPP2	ALPYRWPWWPWSR	841.41	3.34	THP	THP	0.92	Anticp	1.26	Anticp	-0.56	Non-Toxin	-0.56	Non-Toxin	-0.38	Non-Toxin	-0.38	Non-Toxin	0.98	CPP	0.29	CPP
starPep_15346_It4_6_2_5_7	AHPSWWM	837.81	3.5	THP	THP	0.58	Anticp	0.65	Anticp	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin	0.69	CPP	-0.26	Non-CPP
starPep_17042_It4_2_5_7_4_L5_1_CPP2	CLPVM	832.71	0.71	THP	THP	1.2	Anticp	0.36	Anticp	-0.53	Non-Toxin	-0.53	Non-Toxin	-0.16	Non-Toxin	-0.16	Non-Toxin	0.69	CPP	0.14	CPP
starPep_18023_It4_1W_5_8_4C	WPSCHSWH	834.51	3.58	THP	THP	0.8	Anticp	0.55	Anticp	-0.79	Non-Toxin	-0.79	Non-Toxin	-0.5	Non-Toxin	-0.5	Non-Toxin	0.69	CPP	-0.26	Non-CPP
starPep_18023_It4_1W_5_8_4C_CPP1	GRKKRRWC	927.61	0.69	THP	THP	0.4	Anticp	0.34	Anticp	-0.42	Non-Toxin	-0.42	Non-Toxin	-0.62	Non-Toxin	-0.62	Non-Toxin	0.69	CPP	0.22	CPP
starPep_24644_It4_8C_14_3_5_L10	CGCGHCKHKN	836.71	2.84	THP	THP	0.65	Anticp	1.24	Anticp	-0.84	Non-Toxin	-0.84	Non-Toxin	-0.61	Non-Toxin	-0.61	Non-Toxin	0.98	CPP	-0.18	Non-CPP
starPep_27346_It4_1_5_7_6_L5_1	CNKGC	834.21	3.42	THP	THP	0.79	Anticp	1.49	Anticp	-0.34	Non-Toxin	-0.34	Non-Toxin	-0.13	Non-Toxin	-0.13	Non-Toxin	0.69	CPP	-0.25	Non-CPP
starPep_29033_L15_CPP2	KKGKNKHKRHKHKKK	792.21	0.79	THP	THP	0.82	Anticp	0.37	Anticp	-1.08	Non-Toxin	-1.08	Non-Toxin	-0.77	Non-Toxin	-0.77	Non-Toxin	0.98	CPP	0.95	CPP
starPep_02029_It3_4W_7_3_9	TPWWLSWHY	844.71	3.56	THP	THP	0.73	Anticp	0.53	Anticp	-1	Non-Toxin	-1	Non-Toxin	-0.51	Non-Toxin	-0.51	Non-Toxin	0.69	CPP	0	CPP
starPep_02029_It3_4W_7_3_9_CPP1	SRRWRSRHH	835.11	0.61	THP	THP	1.02	Anticp	0.12	Anticp	-0.8	Non-Toxin	-0.8	Non-Toxin	-1.1	Non-Toxin	-1.1	Non-Toxin	0.69	CPP	1.08	CPP
starPep_05293_It4_7_10_11_12_L5	LWWGA	834.81	2.59	THP	THP	0.34	Anticp	1.27	Anticp	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin	0.69	CPP	-0.25	Non-CPP
starPep_07335_It4_3C_1_8_L5	CCRLG	833.71	3.61	THP	THP	0.99	Anticp	1.45	Anticp	-0.4	Non-Toxin	-0.4	Non-Toxin	-0.33	Non-Toxin	-0.33	Non-Toxin	0.69	CPP	-0.24	Non-CPP
starPep_07641_L5_1	CCQEL	834.41	3.04	THP	THP	0.87	Anticp	0.39	Anticp	-0.07	Non-Toxin	-0.07	Non-Toxin	-0.14	Non-Toxin	-0.14	Non-Toxin	0.69	CPP	-0.27	Non-CPP
starPep_10014_L5_3	THWRI	842.61	1.23	THP	THP	0.71	Anticp	0.66	Anticp	-0.89	Non-Toxin	-0.89	Non-Toxin	-0.83	Non-Toxin	-0.83	Non-Toxin	0.69	CPP	-0.24	Non-CPP
starPep_10014_It4_9C_3_8_5_L5_1	CSLHW	813.51	3.28	THP	THP	0.68	Anticp	0.85	Anticp	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin	0.69	CPP	-0.26	Non-CPP
starPep_12257_It4_4L_6_1_7	WRPLPWFY	837.51	3.43	THP	THP	1.17	Anticp	1.44	Anticp	-0.87	Non-Toxin	-0.87	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin	0.69	CPP	-0.25	Non-CPP

M. FASTA of 206 sequences stable in blood.

```
>starPep_12415_CPP3_H1
QSFLRQWARRHFM
>starPep_12415_CPP3_H2
QSFRLQWARRHFM
>starPep_12415_CPP3_H3
QSFRRQWALRHFM
>starPep_12415_CPP3_H4
QSFRRQWARLHFM
>starPep_12415_CPP3_H5
QSFYRQWARRHFM
>starPep_14535_It4_3_14W_2_4_CPP1_HL5_1
SPLRW
>starPep_14535_It4_3_14W_2_4_CPP1_HL5_2
PLRWA
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_1
WWRPPRFSPL
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2
WRPPRFSPLR
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_3
RPPRFSPLRW
>starPep_14535_It4_3_14W_2_4_CPP1_H3
AWWWRQPRFSPLRWA
>starPep_14535_It4_3_14W_2_4_CPP1_H4
AWWWRPPRFSQLRWA
>starPep_14535_It4_3_14W_2_4_CPP1_H5
AQWWRPPRFSPLRWA
>starPep_25472_It4_5L_4_1_12_CPP2_HL5
PYRWP
>starPep_25472_It4_5L_4_1_12_CPP2_H1
ALPYRQPWWPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H2
ALPYRWPQWPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H3
ALPYRWPWQPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H4
ALPYRWPWWPQSR
>starPep_25472_It4_5L_4_1_12_CPP2_H5
ALPYRIPWWPWSR
>starPep_15346_It4_6_2_5_7_H1
AHPSWGM
>starPep_15346_It4_6_2_5_7_H2
AHPSLWM
>starPep_15346_It4_6_2_5_7_H4
AHPSRWM
>starPep_15346_It4_6_2_5_7_H5
AHPSWRM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H1
SLPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H2
CNPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H3
CDPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H4
CLPVR
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H5
CLPRM
>starPep_18023_It4_1W_5_8_4C_HL5_1
WPSCH
>starPep_18023_It4_1W_5_8_4C_HL5_2
PSCHS
>starPep_18023_It4_1W_5_8_4C_HL5_3
SCHSW
>starPep_18023_It4_1W_5_8_4C_HL5_4
CHSWH
>starPep_18023_It4_1W_5_8_4C_H1
WPSCGSWH
>starPep_18023_It4_1W_5_8_4C_H2
WPSCTSWH
>starPep_18023_It4_1W_5_8_4C_H3
WPSCHSWT
>starPep_18023_It4_1W_5_8_4C_H4
WPSCASWH
>starPep_18023_It4_1W_5_8_4C_H5
WPSCHSWA
>starPep_18023_It4_1W_5_8_4C_CPP1_H1
GTKKRRWC
>starPep_18023_It4_1W_5_8_4C_CPP1_H2
GRKKTRWC
>starPep_18023_It4_1W_5_8_4C_CPP1_H3
GRKKRTWC
>starPep_24644_It4_8C_14_3_5_L10_H1
CGCGRCKHKN
>starPep_24644_It4_8C_14_3_5_L10_H2
CGCGHCKRKN
>starPep_27346_It4_1_5_7_6_L5_1_H1
RNKGC
>starPep_27346_It4_1_5_7_6_L5_1_H2
CNKGR
>starPep_27346_It4_1_5_7_6_L5_1_H3
CRKGC
>starPep_27346_It4_1_5_7_6_L5_1_H4
WNKGC
>starPep_27346_It4_1_5_7_6_L5_1_H5
CNKGW
>starPep_29033_L15_CPP2_HL5_1
KKGKN
>starPep_29033_L15_CPP2_HL5_2
KGKNK
>starPep_29033_L15_CPP2_HL5_3
KNKHK
>starPep_29033_L15_CPP2_HL5_4
HKHKK
>starPep_29033_L15_CPP2_HL5_5
KHKKK
>starPep_29033_L15_CPP2_HL10_1
GKNKHKRHKH
>starPep_29033_L15_CPP2_HL10_2
KNKHKRHKHK
>starPep_29033_L15_CPP2_HL10_3
NKHKRHKHKK
>starPep_29033_L15_CPP2_HL10_4
KHKRHKHKKK
>starPep_29033_L15_CPP2_H1
KKHKNKHKRHKHKKK
>starPep_29033_L15_CPP2_H2
KKGKNKHKHHKHKKK
>starPep_29033_L15_CPP2_H3
KKKKNKHKRHKHKKK
>starPep_29033_L15_CPP2_H4
KKGKKKHKRHKHKKK
>starPep_29033_L15_CPP2_H5
KKGKNKHKKHKHKKK
>starPep_02029_It3_4W_7_3_9_H1
TPRWLSWHY
>starPep_02029_It3_4W_7_3_9_H2
TPWRLSWHY
>starPep_02029_It3_4W_7_3_9_H3
TPWWLSRHY
>starPep_02029_It3_4W_7_3_9_H4
TPQWLSWHY
>starPep_02029_It3_4W_7_3_9_H5
TPWWLSQHY
>starPep_02029_It3_4W_7_3_9_CPP1_HL5_1
SRRWR
>starPep_02029_It3_4W_7_3_9_CPP1_HL5_2
RRWRS
>starPep_02029_It3_4W_7_3_9_CPP1_HL5_3
RWRSR
>starPep_02029_It3_4W_7_3_9_CPP1_HL5_4
WRSRH
>starPep_02029_It3_4W_7_3_9_CPP1_H1
QRRWRSRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H2
SRRWRQRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H3
SRRWRSRQH
>starPep_02029_It3_4W_7_3_9_CPP1_H4
SRRWRSRHQ
>starPep_05293_It4_7_10_11_12_L5_H2
LRWGA
>starPep_05293_It4_7_10_11_12_L5_H3
LWRGA
>starPep_05293_It4_7_10_11_12_L5_H4
LSWGA
>starPep_07641_L5_1_H1
WCQEL
>starPep_07641_L5_1_H2
CWQEL
>starPep_07641_L5_1_H3
NCQEL
>starPep_07641_L5_1_H4
CNQEL
>starPep_07641_L5_1_H5
ICQEL
>starPep_10014_L5_3_H1
TLWRI
>starPep_10014_L5_3_H2
TSWRI
>starPep_10014_L5_3_H3
TQWRI
>starPep_10014_L5_3_H4
TAWRI
>starPep_10014_It4_9C_3_8_5_L5_1_H1
RSLHW
>starPep_10014_It4_9C_3_8_5_L5_1_H2
SSLHW
>starPep_10014_It4_9C_3_8_5_L5_1_H3
CSLSW
>starPep_10014_It4_9C_3_8_5_L5_1_H4
PSLHW
>starPep_10014_It4_9C_3_8_5_L5_1_H5
TSLHW
>starPep_12257_It4_4L_6_1_7_HL5_1
WRPLP
>starPep_12257_It4_4L_6_1_7_HL5_2
RPLPW
>starPep_12257_It4_4L_6_1_7_H1
PRPLPWFY
>starPep_12257_It4_4L_6_1_7_H2
WRPLPPFY
>starPep_12257_It4_4L_6_1_7_H3
WRSLPWFY
>starPep_12257_It4_4L_6_1_7_H4
WRPLSWFY
>starPep_12257_It4_4L_6_1_7_H5
WRQLPWFY
>starPep_43502_HL5_3
AMKWI
>starPep_43502_H1
SWAMKWIRV
>starPep_43502_H2
WSAMKWIRV
>starPep_43502_H3
WWAMKSIRV
>starPep_13108_CPP1_HL10
VPWRMKPSPR
>starPep_13108_CPP1_H1
SVLWRMKPSPRQ
>starPep_13108_CPP1_H2
SVPWRMKLSPRQ
>starPep_13108_CPP1_H3
SVPWRMKPSLRQ
>starPep_13108_CPP1_H4
SVTWRMKPSPRQ
>starPep_13108_CPP1_H5
SVPWRMKTSPRQ
>starPep_13108_CPP4_H1
SVRWWMKLSPRQ
>starPep_13108_CPP4_H2
SVRWWMKPSLRQ
>starPep_13108_CPP4_H3
SVRWWMKTSPRQ
>starPep_13108_CPP4_H4
SVRWWMKPSTRQ
>starPep_13108_CPP4_H5
SVRWWMKISPRQ
>starPep_13108_CPP5_H1
SVRWGWKLSPRQ
>starPep_13108_CPP5_H2
SVRWGWKPSLRQ
>starPep_13108_CPP5_H3
SVRWGWKTSPRQ
>starPep_13108_CPP5_H4
SVRWGWKPSTRQ
>starPep_13108_CPP5_H5
SVRWGWKASPRQ
>starPep_04689_HL5
RIGRR
>starPep_04689_H3
RQRLRIGRR
>starPep_04689_H4
RLRQRIGRR
>starPep_12276_H1
PTSNHSPTSWPPTCPGYRWMCLRRF
>starPep_12276_H2
PTSNHSPTSCPPTWPGYRWMCLRRF
>starPep_12276_H3
PTSNHSPTSCPPTCPGYRWMWLRRF
>starPep_12276_H4
WTSNHSPTSCPPTCPGYRWMCLRRF
>starPep_12276_H5
PTSNHSWTSCPPTCPGYRWMCLRRF
>starPep_12276_CPP1_H1
KTRNHSQTSCPPTCPRYRWMCLRRR
>starPep_12276_CPP1_H2
KTRNHSPTSCQPTCPRYRWMCLRRR
>starPep_12276_CPP1_H3
KTRNHSPTSCPQTCPRYRWMCLRRR
>starPep_12276_CPP1_H5
KTRNHSPTSQPPTCPRYRWMCLRRR
>starPep_07237_CPP2_H1
QHWSYRLRPK
>starPep_07237_CPP2_H2
QHWSYRLKPR
>starPep_07237_CPP2_H3
QHWSYKLRPR
>starPep_07237_CPP2_H4
QHWSYRLRPT
>starPep_07237_CPP2_H5
QHWSYRLTPR
```

M. (cont.) FASTA of 206 sequences stable in blood.

```
>starPep_08820_HL5_1
CPSHL
>starPep_08820_HL5_2
PSHLD
>starPep_08820_HL5_5
LDAFC
>starPep_08820_H1
GPSHLDAFC
>starPep_08820_H2
CPSHLDAFG
>starPep_08820_H3
CPSHLGAFC
>starPep_08820_H4
CPSHGDAFC
>starPep_08820_H5
CPSHLDAGC
>starPep_43956_HL5
KWDPP
>starPep_43956_H1
KWDRPPPSPP
>starPep_43956_H2
KWDPRPPSPP
>starPep_43956_H3
KWDPPRPSPP
>starPep_43956_H4
KWDPPPRSPP
>starPep_43956_H5
KWDPPPPSRP
>starPep_43956_CPP3_HL5
KWRPP
>starPep_43956_CPP3_H1
KWRYPPPRPP
>starPep_43956_CPP3_H2
KWRPYPPRPP
>starPep_43956_CPP3_H3
KWRPPYPRPP
>starPep_43956_CPP3_H4
KWRPPPYRPP
>starPep_43956_CPP3_H5
KWRPPPPRYP
>starPep_16808_L10_H5
FNGRCGGKLA
>starPep_17042_L5_1_H1
LIWT
>starPep_17042_L5_1_H2
LIWA
>starPep_17042_L5_1_H4
LIWQ
>starPep_17042_It4_2_5_7_4_L5_1_H4
ECGVL
>starPep_17042_It4_2_5_7_4_L5_1_H5
CEGVL
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H1
WLPVS
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H2
WLSVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H3
SLPVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H4
WLPVR
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H5
WLPKM
>starPep_29033_L15_CPP1_HL5_1
GKNKH
>starPep_29033_L15_CPP1_HL5_2
KHKKH
>starPep_29033_L15_CPP1_HL5_3
HKKHG
>starPep_29033_L15_CPP1_HL5_4
KKHGH
>starPep_29033_L15_CPP1_HL5_5
KHGHG
>starPep_29033_L15_CPP1_HL5_6
HGHGH
>starPep_29033_L15_CPP1_HL5_7
GHGHK
>starPep_29033_L15_CPP1_HL10_1
GKNKHKKHGH
>starPep_29033_L15_CPP1_HL10_2
KNKHKKHGHG
>starPep_29033_L15_CPP1_HL10_3
NKHKKHGHGH
>starPep_29033_L15_CPP1_HL10_4
KHKKHGHGHK
>starPep_29033_L15_CPP1_H1
HKGKNKHKKHGHGHK
>starPep_29033_L15_CPP1_H2
KHGKNKHKKHGHGHK
>starPep_29033_L15_CPP1_H3
KKGHNKHKKHGHGHK
>starPep_29033_L15_CPP1_H4
KKGKHKHKKHGHGHK
>starPep_29033_L15_CPP1_H5
KKGKNHHKKHGHGHK
>starPep_35988_It4_4C_8_9_7_L5_2_CPP2_H1
CKCKK
>starPep_35988_It4_4C_8_9_7_L5_2_CPP2_H5
KKCKC
>starPep_43502_It4_5S_7_9_2_L5_1_H1
WCAGS
>starPep_43502_It4_5S_7_9_2_L5_1_H2
WGAMS
>starPep_43502_It4_5S_7_9_2_L5_1_H3
WLAMS
>starPep_43502_It4_5S_7_9_2_L5_1_H4
GCAMS
>starPep_43502_It4_5S_7_9_2_L5_1_H5
WKAMS
>starPep_01400_L5_1_H1
CALSG
>starPep_01400_L5_1_H2
CGLSA
>starPep_01400_L5_1_H3
CPLSG
>starPep_01400_L5_1_H4
CGLSP
>starPep_01400_L5_1_H5
CGPSG
>starPep_08820_L5_CPP1_H1
CPGKK
>starPep_08820_L5_CPP1_H2
CPKGK
>starPep_08820_L5_CPP1_H3
CPKKG
>starPep_08820_L5_CPP1_H4
CPRKK
>starPep_08820_L5_CPP1_H5
CPKRK
>starPep_04689_It4_1_3_5_8_L5_1_H4
AIGCR
>starPep_04689_It4_1_3_5_8_L5_1_H5
CIGAR
>starPep_05157_It4_1C_10_12_4_L5_2_H1
GPCKS
>starPep_05157_It4_1C_10_12_4_L5_2_H2
CPGKS
>starPep_05157_It4_1C_10_12_4_L5_2_H3
SPCKS
>starPep_05157_It4_1C_10_12_4_L5_2_H4
CPSKS
>starPep_05157_It4_1C_10_12_4_L5_2_H5
CPCKT
```

N. FASTA of 250 sequences stable in gastrointestinal tract.

```
>starPep_43502_H2_G5_1
WSAMK
>starPep_43502_H2_G5_2
SAMKW
>starPep_43502_H2_G_1
WSAMKWIPV
>starPep_43502_H2_G_2
WSAMKWITV
>starPep_43502_H2_G_3
WSAMKWIRY
>starPep_13108_CPP1_H3_G5
WRMKP
>starPep_13108_CPP1_H3_G_1
SIPWRMKPSLRQ
>starPep_13108_CPP1_H3_G_2
SVPWRMKPSLPQ
>starPep_13108_CPP1_H3_G_3
SLPWRMKPSLRQ
>starPep_13108_CPP1_H3_G_4
SLPWRMKPSLNQ
>starPep_04689_H3_G5_1
RQRLR
>starPep_04689_H3_G5_2
QRLRI
>starPep_04689_H3_G_1
RQRLPIGRR
>starPep_04689_H3_G_2
RQRLRPGRR
>starPep_12276_H5_G5_1
PGYRW
>starPep_12276_H5_G5_2
CPGYR
>starPep_12276_H5_G10_4
PTSNHSWTSC
>starPep_12276_H5_G15_2
PPTCPGYRWMCLRRF
>starPep_12276_H5_G_1
PTSNHSWTSCPPTCPGGRWMCLRRF
>starPep_12276_H5_G_2
PTSGHSWTSCPPTCPGYRWMCLRRF
>starPep_12276_H5_G_3
PTSNHSWTSCPPGCPGYRWMCLRGF
>starPep_07237_CPP2_H2_G_1
QHWSYRLLPR
>starPep_07237_CPP2_H2_G_2
QHWSYRLPPR
>starPep_07237_CPP2_H2_G_3
QHDSYRLLPR
>starPep_07237_CPP2_H2_G_4
QHDSYRLPPR
>starPep_08820_H3_G5
CPSHL
>starPep_08820_H3_G_1
CPSHLPAFC
>starPep_08820_H3_G_3
CPSHLGFFC
>starPep_08820_H3_G_4
CPSHLPGFC
>starPep_43956_CPP3_H1_G5
WRYPP
>starPep_43956_CPP3_H1_G_1
KWRYPLPRPP
>starPep_43956_CPP3_H1_G_2
KWRYPPPRLP
>starPep_43956_CPP3_H1_G_3
KWRYPPPLPP
>starPep_43956_CPP3_H1_G_4
KWRYPLPRLP
>starPep_43956_CPP3_H1_G_5
KWRYPLPLPP
>starPep_16808_L10_H5_G_1
FNGRCGGKLP
>starPep_16808_L10_H5_G_2
FPGRCGGKLA
>starPep_17042_It4_2_5_7_4_L5_1_H4_G_1
ECGVG
>starPep_17042_It4_2_5_7_4_L5_1_H4_G_2
ECGFG
>starPep_17042_It4_2_5_7_4_L5_1_H4_G_3
QCGFL
>starPep_17042_It4_2_5_7_4_L5_1_H4_G_4
ECGFW
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H1_G_1
WLPGS
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H1_G_2
WLPNS
>starPep_17042_It4_2_5_7_4_L5_1_CPP3_H1_G_3
WLPTS
>starPep_29033_L15_CPP1_H5_G5_1
GHGHK
>starPep_29033_L15_CPP1_H5_G5_2
HGHGH
>starPep_29033_L15_CPP1_H5_G10_1
NHHKKHGHGH
>starPep_29033_L15_CPP1_H5_G10_2
HHKKHGHGHK
>starPep_43502_It4_5S_7_9_2_L5_1_H4_G_1
GCQMS
>starPep_43502_It4_5S_7_9_2_L5_1_H4_G_2
GCAGS
>starPep_43502_It4_5S_7_9_2_L5_1_H4_G_3
GCAMW
>starPep_43502_It4_5S_7_9_2_L5_1_H4_G_4
GFAMW
>starPep_01400_L5_1_H4_G_1
CGLLP
>starPep_01400_L5_1_H4_G_2
CGLPP
>starPep_01400_L5_1_H4_G_3
CGFSP
>starPep_01400_L5_1_H4_G_4
CGLNP
>starPep_01400_L5_1_H4_G_5
CGDLP
>starPep_05157_It4_1C_10_12_4_L5_2_H5_G_1
CPCKL
>starPep_12415_CPP3_H4_G5_1
WARLH
>starPep_12415_CPP3_H4_G5_2
QWARL
>starPep_12415_CPP3_H4_G5_3
SFRRQ
>starPep_12415_CPP3_H4_G10_4
FRRQWARLHF
>starPep_12415_CPP3_H4_G_1
QSFRRQWARLGFM
>starPep_12415_CPP3_H4_G_2
QSFRRQWARLPFM
>starPep_12415_CPP3_H4_G_3
QSIRRQWARLPFM
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G5
FSPLR
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G_1
WRPPGFSPLR
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G_2
WRPPRFLPLR
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G_3
WRPPRFSPLP
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G_4
WLPPRFSPLR
>starPep_14535_It4_3_14W_2_4_CPP1_HL10_2_G_5
WRPPRFGPLR
>starPep_25472_It4_5L_4_1_12_CPP2_H2_G_1
ALPYRWPQWPWSI
>starPep_25472_It4_5L_4_1_12_CPP2_H2_G_2
ALPYRWPQLPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H2_G_3
ALPYRLPQWPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H2_G_4
ALPYRWPGWPWSR
>starPep_25472_It4_5L_4_1_12_CPP2_H2_G_5
ALPYRWNQWPWSR
>starPep_15346_It4_6_2_5_7_H4_G5
PSRWM
>starPep_15346_It4_6_2_5_7_H4_G_1
ALPSRWM
>starPep_15346_It4_6_2_5_7_H4_G_2
AFPSRWM
>starPep_15346_It4_6_2_5_7_H4_G_3
AMDSRWM
>starPep_15346_It4_6_2_5_7_H4_G5_1
AHPSW
>starPep_15346_It4_6_2_5_7_H4_G5_2
HPSWR
>starPep_15346_It4_6_2_5_7_H4_G5_3
PSWRM
>starPep_15346_It4_6_2_5_7_H4_G1_1
ALPSWRM
>starPep_15346_It4_6_2_5_7_H4_G1_2
AHPSWQM
>starPep_15346_It4_6_2_5_7_H4_G1_3
AHPSWAM
>starPep_15346_It4_6_2_5_7_H4_G1_4
AFPSWRM
>starPep_15346_It4_6_2_5_7_H4_G1_5
AHTSWRM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H3_G_1
CDPGM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H3_G_2
CDSVM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H3_G_3
CDPQM
>starPep_17042_It4_2_5_7_4_L5_1_CPP2_H3_G_4
CGPVM
>starPep_18023_It4_1W_5_8_4C_H5_G5_1
WPSCH
>starPep_18023_It4_1W_5_8_4C_H5_G5_2
SCHSW
>starPep_18023_It4_1W_5_8_4C_H5_G5_3
CHSWA
>starPep_18023_It4_1W_5_8_4C_H5_G
WPGCHSWA
>starPep_24644_It4_8C_14_3_5_L10_H2_G_4
CGCGDCKRKN
>starPep_27346_It4_1_5_7_6_L5_1_H3_G_1
CRPGC
>starPep_27346_It4_1_5_7_6_L5_1_H3_G_3
CRFGC
>starPep_27346_It4_1_5_7_6_L5_1_H3_G_4
CSKGC
>starPep_27346_It4_1_5_7_6_L5_1_H3_G_5
CRCGF
>starPep_02029_It3_4W_7_3_9_H5_G5_1
TPWWL
>starPep_02029_It3_4W_7_3_9_H5_G5_2
PWWLS
>starPep_02029_It3_4W_7_3_9_H5_G_1
TPWWLNQHY
>starPep_02029_It3_4W_7_3_9_H5_G_2
TPWWLPQHY
>starPep_02029_It3_4W_7_3_9_H5_G_3
TPWWLSIHY
>starPep_02029_It3_4W_7_3_9_H5_G_4
LPWWLSQHY
>starPep_02029_It3_4W_7_3_9_H5_G_5
TPWWLSGHY
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G5_1
QRRWR
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G5_2
RRWRS
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G_1
QMRWRSRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G_2
QRRWTSRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G_3
QCRWRSRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G_4
QMRWDSRHH
>starPep_02029_It3_4W_7_3_9_CPP1_H1_G_5
QCRWDSRHH
>starPep_05293_It4_7_10_11_12_L5_H3_G_3
LWPGF
>starPep_05293_It4_7_10_11_12_L5_H3_G_5
YWRGF
>starPep_07641_L5_1_H5_G_1
ICIEL
>starPep_07641_L5_1_H5_G_2
ICQEE
>starPep_07641_L5_1_H5_G_3
ICQER
>starPep_07641_L5_1_H5_G_4
ECKL
>starPep_10014_L5_3_H4_G_1
TAWSI
>starPep_10014_L5_3_H4_G_2
TSWRI
>starPep_10014_L5_3_H4_G_3
TAWRY
>starPep_10014_L5_3_H4_G_4
TSWSI
>starPep_12257_It4_4L_6_1_7_H5_G_1
WRQLPWFG
```

N. (cont.) FASTA of 250 sequences stable in gastrointestinal tract.

```
>starPep_12257_It4_4L_6_1_7_H5_G_3
WRQLPAFY
>starPep_12257_It4_4L_6_1_7_H5_G_4
WRGLPWFY
>starPep_12257_It4_4L_6_1_7_H5_G_5
WRLLPGFY
>starPep_43502_HL5_3_G_1
AMWWI
>starPep_43502_HL5_3_G_2
AMTWI
>starPep_43502_HL5_3_G_3
AMRWI
>starPep_43502_HL5_3_G_5
AFKWW
>starPep_43502_HL5_3_G_7
AFGWY
>starPep_43502_H3_G_3
WWAMKSIRY
>starPep_43502_H3_G_4
WWAMTSIRV
>starPep_43502_H3_G_5
WWAMKSIWV
>starPep_43502_H3_G_9
WWAMTSIES
>starPep_43502_H3_G_10
WWAMTSIEA
>starPep_13108_CPP1_HL10_G5
MKPSP
>starPep_13108_CPP1_HL10_G_1
LPWRMKPSPR
>starPep_13108_CPP1_HL10_G_2
VPWRMKPLPR
>starPep_13108_CPP1_HL10_G_3
VPWRMKPSPN
>starPep_13108_CPP1_HL10_G_4
VPWRMLPSIR
>starPep_13108_CPP1_H1_G5_1
WRMKP
>starPep_13108_CPP1_H1_G5_2
KPSPR
>starPep_13108_CPP1_H1_G5_3
PSPRQ
>starPep_13108_CPP1_H1_G10_1
SVLWRMKPSP
>starPep_13108_CPP1_H1_G10_2
VLWRMKPSPR
>starPep_13108_CPP1_H1_G10_3
LWRMKPSPRQ
>starPep_13108_CPP1_H1_G_1
SILWRMKPSPRQ
>starPep_13108_CPP1_H1_G_2
SVLWRMKPLPRQ
>starPep_13108_CPP1_H1_G_3
SILWRMKPLPRQ
>starPep_13108_CPP1_H1_G_4
SILWRMKLLPRQ
>starPep_13108_CPP1_H4_G5
RMKPS
>starPep_13108_CPP1_H4_G10_1
TWRMKPSPRQ
>starPep_13108_CPP1_H4_G10_2
VTWRMKPSPR
>starPep_13108_CPP1_H4_G10_3
SVTWRMKPSP
>starPep_13108_CPP1_H4_G_1
SITWRMKPSPRQ
>starPep_13108_CPP1_H4_G_2
SVTWRMKPSPNQ
>starPep_13108_CPP1_H4_G_3
SGTWRMKPSPRQ
>starPep_13108_CPP1_H4_G_4
SITWRMKPLPRQ
>starPep_12276_H1_G5_1
PGYRW
>starPep_12276_H1_G5_2
CPGYR
>starPep_12276_H1_G5_4
PTSNH
>starPep_12276_H1_G5_5
HSPTS
>starPep_12276_H1_G5_6
TSNHS
>starPep_12276_H1_G10_4
TSNHSPTSWP
>starPep_12276_H1_G15_1
TSWPPTCPGYRWMCL
>starPep_12276_H1_G15_2
HSPTSWPPTCPGYRW
>starPep_12276_H1_G15_3
SWPPTCPGYRWMCLR
>starPep_12276_H1_G15_4
WPPTCPGYRWMCLRR
>starPep_12276_H1_G15_5
PPTCPGYRWMCLRRF
>starPep_12276_H2_G5
WPGYR
>starPep_12276_H2_G10_1
TWPGYRWMCL
>starPep_12276_H2_G10_2
PTWPGYRWMC
>starPep_12276_H2_G10_3
PPTWPGYRWM
>starPep_12276_H2_G10_4
CPPTWPGYRW
>starPep_12276_H3_G15_3
CPPTCPGYRWMWLRR
>starPep_12276_H4_G_1
WTSNHSPTSCPPTCPGGRWMCLRRF
>starPep_12276_H4_G_2
WTSNHSPTSCPPGCPGYRWMCLRGF
>starPep_12276_H4_G_3
WTSNHSPTSCPPGCPGGRWMCLRGF
>starPep_12276_CPP1_H2_G5_1
RYRWM
>starPep_12276_CPP1_H2_G5_3
YRWMC
>starPep_12276_CPP1_H2_G5_4
PRYRW
>starPep_12276_CPP1_H2_G10_1
TRNHSPTSCQ
>starPep_12276_CPP1_H2_G10_2
RNHSPTSCQP
>starPep_12276_CPP1_H2_G10_3
NHSPTSCQPT
>starPep_12276_CPP1_H2_G10_4
HSPTSCQPTC
>starPep_12276_CPP1_H2_G15_1
TSCQPTCPRYRWMCL
>starPep_12276_CPP1_H2_G15_2
PTSCQPTCPRYRWMC
>starPep_12276_CPP1_H2_G15_3
NHSPTSCQPTCPRYR
>starPep_12276_CPP1_H2_G15_4
HSPTSCQPTCPRYRW
>starPep_12276_CPP1_H2_G15_5
QPTCPRYRWMCLRRR
>starPep_12276_CPP1_H2_G15_6
SCQPTCPRYRWMCLR
>starPep_12276_CPP1_H2_G_1
KTRGHSPTSCQPTCPRYRWMCLRRG
>starPep_12276_CPP1_H2_G_2
KTRNHSPTSCQPTCPRGRWMCLRRG
>starPep_12276_CPP1_H2_G_3
KTRNHSPTSCQPTCPRYRWMCLRCG
>starPep_12276_CPP1_H2_G_4
KTRNHSPTSCQPTCPRYRWMCLCRG
>starPep_12276_CPP1_H3_G5
SPTSC
>starPep_12276_CPP1_H3_G10_1
RNHSPTSCPQ
>starPep_12276_CPP1_H3_G10_2
NHSPTSCPQT
>starPep_12276_CPP1_H3_G10_3
HSPTSCPQTC
>starPep_12276_CPP1_H3_G10_4
TRNHSPTSCP
>starPep_12276_CPP1_H3_G_2
KTRNHSPTSCPQTCPRYRWMCLRGR
>starPep_12276_CPP1_H3_G_4
KTGNHSPTSCPQTCPRYRWMCLRRR
>starPep_12276_CPP1_H3_G_5
KTRNHSPTSCPQTCPRYRWMCLRRC
>starPep_12276_CPP1_H3_G_6
KTRNHSPTSCPQTCPRYRWMCLCRR
>starPep_12276_CPP1_H3_G_7
KTRNHSPTSCPQTCPRYCWMCLRRR
>starPep_25472_It4_5L_4_1_12_CPP2_HL5_G_2
PYWLP
>starPep_12276_CPP1_H5_G5
CPRYR
>starPep_12276_CPP1_H5_G_1
KTRNHSPTSQPPTCPRYRWMCLRRC
>starPep_12276_CPP1_H5_G_2
KTRNHSPTSQPPTCPRYRWMCLRCR
>starPep_12276_CPP1_H5_G_3
KTRNHSPTSQPPTCPRYRWMCLCRR
>starPep_12276_CPP1_H5_G_4
KTRNHSPTSQPPTCPRYCWMCLRRR
>starPep_07237_CPP2_H1_G5_1
QHWSY
>starPep_07237_CPP2_H1_G5_2
HWSYR
>starPep_07237_CPP2_H1_G5_3
WSYRL
>starPep_07237_CPP2_H1_G5_4
YRLRP
>starPep_07237_CPP2_H1_G_2
QHWSYRLPPK
>starPep_07237_CPP2_H1_G_3
QHWSYRLRPG
>starPep_07237_CPP2_H1_G_4
QHWSYRLRPE
>starPep_07237_CPP2_H1_G_5
QHWSYRLLPG
>starPep_07237_CPP2_H3_G_1
QHWSYKLLPR
>starPep_07237_CPP2_H3_G_2
QHWSYKLPPR
>starPep_07237_CPP2_H4_G_1
QHWSYRLLPT
>starPep_07237_CPP2_H4_G_2
QHWSYRLPPT
>starPep_08820_H5_G_1
CPSHLDSGC
>starPep_08820_H5_G_2
CPSHLPAGC
>starPep_08820_H5_G_3
CPSHLDPGC
>starPep_29033_L15_CPP1_HL5_5_G
KCGHG
>starPep_29033_L15_CPP1_HL5_5_G_2
KCGFG
>starPep_29033_L15_CPP1_HL5_5_G_3
KCGPG
>starPep_29033_L15_CPP1_HL5_5_G_4
KCGDG
>starPep_29033_L15_CPP1_HL5_5_G_5
KCGTG
>starPep_29033_L15_CPP1_HL5_6_G
HGSGH
>starPep_29033_L15_CPP1_HL10_2_G_1
KNKHKKPGHG
>starPep_29033_L15_CPP1_HL10_2_G_2
KNKHKKHGPG
>starPep_29033_L15_CPP1_HL10_3_G_1
NKHKKPGHGH
>starPep_29033_L15_CPP1_HL10_3_G_2
NKHKKHGPGH
>starPep_29033_L15_CPP1_HL10_4_G_1
KHKKPGHGHK
>starPep_29033_L15_CPP1_HL10_4_G_2
KHKKHGPGHK
>starPep_29033_L15_CPP1_H1_G_1
HPGKNKHKKHGHGHK
>starPep_29033_L15_CPP1_H1_G5
GKNKHKKHGH
>starPep_29033_L15_CPP1_H2_G_1
KPGKNKHKKHGHGHK
>starPep_29033_L15_CPP1_H2_G_2
KHGKNKHKKPGHGHK
>starPep_29033_L15_CPP1_H2_G_3
KHGKNKHKKHGPGHK
>starPep_14535_It4_3_14W_2_4_CPP1_HL5_1_G_1
SPLPW
>starPep_14535_It4_3_14W_2_4_CPP1_HL5_2_G
PLSWA
>starPep_25472_It4_5L_4_1_12_CPP2_HL5_G_1
PYRLP
```

O. Predicted activities of SET 4, conformed by 78 lead THPs with optimized gastrointestinal stability. 13 sequences from SET 4 are highlighted.

	PlifePred	TumorHPD		THPep	AntiCP				ToxinPred								CellPPD					AMPfun	AlgPred2		AnuPP
Sequence	Half-time (seconds)	SVM Score		SVM Score	SVM Score 1		SVM Score 2		SVM Score 1		SVM Score 2		SVM Score 3		SVM Score 4		SVM Score 1		SVM Score 2		HemoPred	Anticancer	Score		APR?
QHWSYRLPPT	1070.41	2.35	THP	THP	0.78	Anticp	0.02	Anticp	-1.02	Non-Toxin	-1.02	Non-Toxin	-1.09	Non-Toxin	-1.09	Non-Toxin	0.05	CPP	0.05	CPP	non-hemolytic	0.1844	0.345	Allergen	False
VPWRMLPSIR	993.71	1.31	THP	THP	1	Anticp	-0.03	Non-Anticp	-1.1	Non-Toxin	-1.1	Non-Toxin	-1.42	Non-Toxin	-1.42	Non-Toxin	0.05	CPP	-0.03	Non-CPP	non-hemolytic	0.384	0.282	Non-Allergen	False
TSWPPTCPGARWRWL	917.51	2.43	THP	THP	0.71	Anticp	1.03	Anticp	-0.6	Non-Toxin	-0.6	Non-Toxin	-0.83	Non-Toxin	-0.83	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	hemolytic	0.0625	0.299	Non-Allergen	False
ALPYRWPQLPWSR	907.61	2.9	THP	THP	0.93	Anticp	0.77	Anticp	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.72	Non-Toxin	-0.72	Non-Toxin	0.05	CPP	0.18	CPP	non-hemolytic	0.1029	0.331	Allergen	False
WRPPRFSPLP	897.11	2.61	THP	THP	0.86	Anticp	0.86	Anticp	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.75	Non-Toxin	-0.75	Non-Toxin	0.05	CPP	0.3	CPP	non-hemolytic	0.2074	0.318	Allergen	False
SITWRMKPLPRQ	1775.91	0.7	THP	THP	0.76	Anticp	-0.25	Non-Anticp	-1.69	Non-Toxin	-1.69	Non-Toxin	-2.07	Non-Toxin	-2.07	Non-Toxin	0.05	CPP	0.32	CPP	non-hemolytic	0.0547	0.284	Non-Allergen	False
AFPSWRM	829.71	2.36	THP	THP	0.4	Anticp	-0.02	Non-Anticp	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.74	Non-Toxin	-0.74	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.306	0.277	Non-Allergen	False
KWRYPLPRPP	1286.91	2.07	THP	THP	0.79	Anticp	1.2	Anticp	-0.07	Non-Toxin	-0.07	Non-Toxin	-0.25	Non-Toxin	-0.25	Non-Toxin	0.05	CPP	0.57	CPP	non-hemolytic	0.2701	0.316	Allergen	False
CPSHLDWSC	665.41	2.76	THP	THP	1.12	Anticp	0.6	Anticp	-0.99	Non-Toxin	-0.99	Non-Toxin	-0.87	Non-Toxin	-0.87	Non-Toxin	-0.26	Non-CPP	-0.34	Non-CPP	non-hemolytic	0.2443	0.401	Allergen	False
WRLRAQFY	1009.51	1.85	THP	THP	0.51	Anticp	0.7	Anticp	-0.88	Non-Toxin	-0.88	Non-Toxin	-1.18	Non-Toxin	-1.18	Non-Toxin	-0.26	Non-CPP	-0.23	Non-CPP	non-hemolytic	0.4294	0.287	Non-Allergen	False
NKHKKPGHGH	834.91	0.84	THP	THP	0.25	Anticp	0.23	Anticp	-1.14	Non-Toxin	-1.14	Non-Toxin	-0.69	Non-Toxin	-0.69	Non-Toxin	0.05	CPP	-0.1	Non-CPP	non-hemolytic	0.3139	0.379	Allergen	False
WLPWPFRAY	835.71	3.3	THP	THP	0.8	Anticp	1.52	Anticp	-0.96	Non-Toxin	-0.96	Non-Toxin	-0.8	Non-Toxin	-0.8	Non-Toxin	-0.26	Non-CPP	0.03	CPP	non-hemolytic	0.3484	0.305	Allergen	False
QSIRRQWARLPFM	1334.01	1.1	THP	THP	0.63	Anticp	-0.04	Non-Anticp	-0.84	Non-Toxin	-0.84	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin	0.05	CPP	0.32	CPP	non-hemolytic	0.151	0.332	Allergen	False
QMRWRSRHH	841.41	0.8	THP	THP	0.88	Anticp	0.39	Anticp	-0.98	Non-Toxin	-0.98	Non-Toxin	-1.92	Non-Toxin	-1.92	Non-Toxin	-0.26	Non-CPP	0.81	CPP	non-hemolytic	0.2471	0.299	Non-Allergen	False
AMDSRWM	843.11	2.22	THP	THP	0.84	Anticp	-0.44	Non-Anticp	-1.11	Non-Toxin	-1.11	Non-Toxin	-1.02	Non-Toxin	-1.02	Non-Toxin	-0.26	Non-CPP	-0.24	Non-CPP	non-hemolytic	0.2295	0.273	Non-Allergen	False
WRQLPWFG	868.91	2.52	THP	THP	1.06	Anticp	1.12	Anticp	-1	Non-Toxin	-1	Non-Toxin	-0.67	Non-Toxin	-0.67	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.4834	0.378	Allergen	False
CARGNCKRGG	884.91	2.52	THP	THP	0.23	Anticp	1	Anticp	-0.61	Non-Toxin	-0.61	Non-Toxin	-0.67	Non-Toxin	-0.67	Non-Toxin	-0.26	Non-CPP	-0.12	Non-CPP	non-hemolytic	0.4627	0.44	Allergen	False
CGLNP	822.81	1.77	THP	THP	0.85	Anticp	0.85	Anticp	-0.78	Non-Toxin	-0.78	Non-Toxin	-0.56	Non-Toxin	-0.56	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	non-hemolytic	0.5638	0.436	Allergen	-
WPGCHSWA	884.51	3.12	THP	THP	1.1	Anticp	1.57	Anticp	-0.75	Non-Toxin	-0.75	Non-Toxin	-0.47	Non-Toxin	-0.47	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.4931	0.396	Allergen	False
WWHSTSWRV	841.41	3.15	THP	THP	0.36	Anticp	-0.08	Non-Anticp	-1.36	Non-Toxin	-1.36	Non-Toxin	-1.29	Non-Toxin	-1.29	Non-Toxin	-0.26	Non-CPP	-0.09	Non-CPP	non-hemolytic	0.4146	0.278	Non-Allergen	False
RQRLPIGRR	867.51	1.16	THP	THP	1.03	Anticp	0.16	Anticp	-1.02	Non-Toxin	-1.02	Non-Toxin	-0.76	Non-Toxin	-0.76	Non-Toxin	-0.26	Non-CPP	0.31	CPP	non-hemolytic	0.4788	0.416	Allergen	False
CDSVM	816.41	1.04	THP	THP	1.18	Anticp	-0.67	Non-Anticp	-0.54	Non-Toxin	-0.54	Non-Toxin	-0.25	Non-Toxin	-0.25	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	non-hemolytic	0.2877	0.353	Allergen	-
CGFSP	857.91	1.48	THP	THP	0.92	Anticp	1.28	Anticp	-0.78	Non-Toxin	-0.78	Non-Toxin	-0.25	Non-Toxin	-0.25	Non-Toxin	-0.26	Non-CPP	-0.28	Non-CPP	hemolytic	0.4621	0.4	Allergen	-
FPGRCGGKLA	1265.21	1.1	THP	THP	0.91	Anticp	1.35	Anticp	-0.58	Non-Toxin	-0.58	Non-Toxin	-0.35	Non-Toxin	-0.35	Non-Toxin	0.05	CPP	-0.26	Non-CPP	non-hemolytic	0.2072	0.439	Allergen	False
AFKWW	834.91	1.71	THP	THP	0.51	Anticp	1.17	Anticp	-0.82	Non-Toxin	-0.82	Non-Toxin	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	hemolytic	0.3034	0.462	Allergen	-
PLSWA	835.51	2.5	THP	THP	0.8	Anticp	0.42	Anticp	-1	Non-Toxin	-1	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.3942	0.36	Allergen	-
CRPGC	823.21	2.94	THP	THP	1.19	Anticp	1.66	Anticp	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.2	Non-Toxin	-0.2	Non-Toxin	-0.26	Non-CPP	-0.23	Non-CPP	non-hemolytic	0.6089	0.412	Allergen	-
PYWLP	835.91	2.82	THP	THP	0.99	Anticp	0.98	Anticp	-0.88	Non-Toxin	-0.88	Non-Toxin	-0.64	Non-Toxin	-0.64	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.8403	0.403	Allergen	-
TAWRY	841.01	1.59	THP	THP	0.59	Anticp	0.94	Anticp	-0.65	Non-Toxin	-0.65	Non-Toxin	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.26	Non-CPP	-0.24	Non-CPP	non-hemolytic	0.273	0.31	Allergen	-
ECGFW	834.31	2.03	THP	THP	0.95	Anticp	0.48	Anticp	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.46	Non-Toxin	-0.46	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	hemolytic	0.456	0.465	Allergen	-
CDPGM	837.31	1.35	THP	THP	0.88	Anticp	0.18	Anticp	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.22	Non-Toxin	-0.22	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.5954	0.374	Allergen	-
GFAMW	835.51	1.67	THP	THP	-0.8	Non-Anticp	0.42	Anticp	-0.9	Non-Toxin	-0.9	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.4382	0.359	Allergen	-
CGDLP	843.71	1.33	THP	THP	1.24	Anticp	0.22	Anticp	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	non-hemolytic	0.5289	0.385	Allergen	-
KCGFG	837.81	1.65	THP	THP	1.26	Anticp	1.17	Anticp	-0.83	Non-Toxin	-0.83	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.5114	0.473	Allergen	-
QCGFL	822.51	1.49	THP	THP	0.97	Anticp	0.31	Anticp	-0.85	Non-Toxin	-0.85	Non-Toxin	-0.37	Non-Toxin	-0.37	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	hemolytic	0.5666	0.422	Allergen	-
WLPNS	842.71	2.09	THP	THP	0.58	Anticp	-0.46	Non-Anticp	-0.71	Non-Toxin	-0.71	Non-Toxin	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.4917	0.375	Allergen	-
YWRGF	835.11	2.07	THP	THP	0.23	Anticp	1.02	Anticp	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.68	Non-Toxin	-0.68	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	hemolytic	0.6917	0.302	Allergen	-
ICQER	832.71	1.46	THP	THP	1.07	Anticp	-0.07	Non-Anticp	-0.8	Non-Toxin	-0.8	Non-Toxin	-0.48	Non-Toxin	-0.48	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.4488	0.395	Allergen	-

O. (cont.) Predicted activities of SET 4, conformed by 78 lead THPs with optimized gastrointestinal stability. 13 sequences from SET 4 are highlighted.

Sequence	PlifePred	TumorHPD		THPep	AntiCP				ToxinPred								CellPPD				HemoPred	AMPfun	AlgPred2		AnuPP
	Half-time (seconds)	SVM Score		SVM Score	SVM Score 1		SVM Score 2		SVM Score 1		SVM Score 2		SVM Score 3		SVM Score 4		SVM Score 1		SVM Score 2			Anticancer	Score		APR?
CSKGC	834.21	3.01	THP	THP	0.97	Anticp	1.72	Anticp	-0.51	Non-Toxin	-0.51	Non-Toxin	-0.2	Non-Toxin	-0.2	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.5193	0.411	Allergen	-
AMWWP	834.81	2.79	THP	THP	1.12	Anticp	1.05	Anticp	-0.79	Non-Toxin	-0.79	Non-Toxin	-0.59	Non-Toxin	-0.59	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.7749	0.412	Allergen	-
CRCGF	811.81	2.92	THP	THP	0.88	Anticp	1.31	Anticp	-0.71	Non-Toxin	-0.71	Non-Toxin	-0.26	Non-Toxin	-0.26	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.4826	0.405	Allergen	-
CGPVM	828.51	1	THP	THP	1.31	Anticp	0.95	Anticp	-0.69	Non-Toxin	-0.69	Non-Toxin	-0.42	Non-Toxin	-0.42	Non-Toxin	-0.26	Non-CPP	-0.21	Non-CPP	non-hemolytic	0.4002	0.394	Allergen	-
GCQMS	810.21	1.18	THP	THP	-0.4	Non-Anticp	-0.12	Non-Anticp	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.56	Non-Toxin	-0.56	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	hemolytic	0.4668	0.388	Allergen	-
TAWSI	846.11	1.42	THP	THP	0.82	Anticp	0.17	Anticp	-0.91	Non-Toxin	-0.91	Non-Toxin	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.26	Non-CPP	-0.26	Non-CPP	non-hemolytic	0.5965	0.314	Allergen	-
SPTSC	834.41	1.38	THP	THP	0.83	Anticp	0.5	Anticp	-0.88	Non-Toxin	-0.88	Non-Toxin	-0.47	Non-Toxin	-0.47	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.3398	0.365	Allergen	-
HGSGH	834.81	1.36	THP	THP	0.55	Anticp	0.78	Anticp	-0.78	Non-Toxin	-0.78	Non-Toxin	-0.49	Non-Toxin	-0.49	Non-Toxin	-0.26	Non-CPP	-0.28	Non-CPP	non-hemolytic	0.4809	0.364	Allergen	-
LCGVG	834.81	1.75	THP	THP	1.27	Anticp	1.05	Anticp	-0.89	Non-Toxin	-0.89	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.26	Non-CPP	-0.28	Non-CPP	hemolytic	0.7435	0.384	Allergen	-
GCRGS	814.31	2.41	THP	THP	-0.97	Non-Anticp	1.05	Anticp	-0.6	Non-Toxin	-0.6	Non-Toxin	-0.28	Non-Toxin	-0.28	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.2971	0.372	Allergen	-
CPCKL	834.41	2.65	THP	THP	0.95	Anticp	1.59	Anticp	-0.63	Non-Toxin	-0.63	Non-Toxin	-0.13	Non-Toxin	-0.13	Non-Toxin	-0.26	Non-CPP	-0.24	Non-CPP	non-hemolytic	0.4697	0.434	Allergen	-
ICIEL	834.71	1.3	THP	THP	0.89	Anticp	0.32	Anticp	-0.71	Non-Toxin	-0.71	Non-Toxin	-0.37	Non-Toxin	-0.37	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	non-hemolytic	0.5471	0.326	Allergen	-
KTRNHSPTSCPMACPRHRWNCWHHC	830.01	2.42	THP	THP	0.8	Anticp	1.22	Anticp	-0.14	Non-Toxin	-0.14	Non-Toxin	-0.61	Non-Toxin	-0.61	Non-Toxin	-0.26	Non-CPP	0.12	CPP	non-hemolytic	0.0881	0.254	Non-Allergen	False
WWWCHSPTSQPPTCPRDRNMNRCRR	933.11	2	THP	THP	0.73	Anticp	1.01	Anticp	-0.52	Non-Toxin	-0.52	Non-Toxin	-0.78	Non-Toxin	-0.78	Non-Toxin	0.23	CPP	0.23	CPP	non-hemolytic	0.103	0.262	Non-Allergen	False
WARNHSPTSCNPTQPRYCWHNHRNG	884.91	1.89	THP	THP	0.76	Anticp	0.62	Anticp	-0.51	Non-Toxin	-0.51	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.26	Non-CPP	-0.04	Non-CPP	hemolytic	0.0581	0.304	Allergen	False
LTRNHSHTSRLPTNPRCSHKWQAWC	1459.91	1.91	THP	THP	0.75	Anticp	0.43	Anticp	-0.28	Non-Toxin	-0.28	Non-Toxin	-0.66	Non-Toxin	-0.66	Non-Toxin	-0.03	Non-CPP	-0.03	Non-CPP	hemolytic	0.0367	0.265	Non-Allergen	False
SIPWRMKPSLRQ	1287.71	0.78	THP	THP	0.64	Anticp	-0.42	Non-Anticp	-1	Non-Toxin	-1	Non-Toxin	-1.48	Non-Toxin	-1.48	Non-Toxin	0.05	CPP	0.29	CPP	non-hemolytic	0.0752	0.298	Non-Allergen	False
SITWRMKPSPRQ	1212.51	0.46	THP	THP	0.73	Anticp	-0.66	Non-Anticp	-1.55	Non-Toxin	-1.55	Non-Toxin	-2.05	Non-Toxin	-2.05	Non-Toxin	0.05		0.31	CPP	non-hemolytic	0.1558	0.287	Non-Allergen	False
SVPWRMKPSLPQ	901.91	1.13	THP	THP	0.67	Anticp	-0.37	Non-Anticp	-1.18	Non-Toxin	-1.18	Non-Toxin	-1.65	Non-Toxin	-1.65	Non-Toxin	0.05	CPP	0.17	CPP	non-hemolytic	0.0519	0.302	Allergen	False
SVTWRMKPSPNQ	1131.91	0.47	THP	THP	0.75	Anticp	-1.21	Non-Anticp	-1.18	Non-Toxin	-1.18	Non-Toxin	-1.77	Non-Toxin	-1.77	Non-Toxin	0.05	CPP	0.12	CPP	non-hemolytic	0.1514	0.283	Non-Allergen	False
LPWRMKPSPR	905.91	1.92	THP	THP	0.84	Anticp	0.57	Anticp	-1.31	Non-Toxin	-1.31	Non-Toxin	-1.61	Non-Toxin	-1.61	Non-Toxin	0.05	CPP	0.23	CPP	non-hemolytic	0.3411	0.264	Non-Allergen	False
QHWSYKLPPR	1154.31	1.62	THP	THP	0.68	Anticp	0.2	Anticp	-0.67	Non-Toxin	-0.67	Non-Toxin	-0.96	Non-Toxin	-0.96	Non-Toxin	0.05	CPP	0.16	CPP	non-hemolytic	0.1056	0.333	Allergen	False
SLPWRMKPSLNQ	1174.41	0.96	THP	THP	0.67	Anticp	-0.7	Non-Anticp	-1.25	Non-Toxin	-1.25	Non-Toxin	-1.61	Non-Toxin	-1.61	Non-Toxin	0.05	CPP	0.2	CPP	non-hemolytic	0.075	0.291	Non-Allergen	False
QHWSYRLRPE	1230.41	1.43	THP	THP	0.85	Anticp	0.08	Anticp	-1.15	Non-Toxin	-1.15	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin	0.05	CPP	0	Non-CPP	non-hemolytic	0.2905	0.346	Allergen	False
WTSNHSPTSCRSNCPRYRWMCRRNF	863.31	1.77	THP	THP	0.79	Anticp	0.68	Anticp	-0.58	Non-Toxin	-0.58	Non-Toxin	-0.85	Non-Toxin	-0.85	Non-Toxin	-0.26	Non-CPP	-0.09	Non-CPP	hemolytic	0.0488	0.255	Non-Allergen	False
SGTWRMKPSPRQ	1152.81	0.62	THP	THP	0.88	Anticp	-0.42	Non-Anticp	-1.3	Non-Toxin	-1.3	Non-Toxin	-1.52	Non-Toxin	-1.52	Non-Toxin	0.05	CPP	0.24	CPP	non-hemolytic	0.2584	0.25	Non-Allergen	False
HSPTSWPPTCPGYRW	858.91	1.89	THP	THP	0.73	Anticp	0.81	Anticp	-0.51	Non-Toxin	-0.51	Non-Toxin	-0.4	Non-Toxin	-0.4	Non-Toxin	0.05	CPP	-0.05	Non-CPP	non-hemolytic	0.2825	0.315	Allergen	False
VLWRMKPSPR	987.31	0.89	THP	THP	0.72	Anticp	0.14	Anticp	-1.35	Non-Toxin	-1.35	Non-Toxin	-1.59	Non-Toxin	-1.59	Non-Toxin	0.05	CPP	0.25	CPP	non-hemolytic	0.4058	0.269	Non-Allergen	False
WASNHAMGSCACRMPKGRKHMRECRC	785.41	1.64	THP	THP	0.73	Anticp	0.5	Anticp	-0.12	Non-Toxin	-0.12	Non-Toxin	-0.48	Non-Toxin	-0.48	Non-Toxin	-0.26	Non-CPP	-0.2	Non-CPP	non-hemolytic	0.05	0.264	Non-Allergen	False
HHGTNKHCKHNHSHC	834.61	1.9	THP	THP	0.77	Anticp	0.56	Anticp	-0.02	Non-Toxin	-0.02	Non-Toxin	-0.22	Non-Toxin	-0.22	Non-Toxin	-0.26	Non-CPP	-0.06	Non-CPP	hemolytic	0.0498	0.409	Allergen	False
WSAMKWITV	946.61	1.11	THP	THP	0.79	Anticp	0.01	Anticp	-0.69	Non-Toxin	-0.69	Non-Toxin	-1.03	Non-Toxin	-1.03	Non-Toxin	-0.26	Non-CPP	-0.12	Non-CPP	hemolytic	0.3768	0.285	Non-Allergen	False
WSAMKWIPV	913.51	1.19	THP	THP	0.67	Anticp	0.23	Anticp	-0.51	Non-Toxin	-0.51	Non-Toxin	-0.73	Non-Toxin	-0.73	Non-Toxin	-0.26	Non-CPP	-0.06	Non-CPP	non-hemolytic	0.386	0.337	Allergen	False
WWAMKSIWV	844.41	1.95	THP	THP	0.46	Anticp	0.54	Anticp	-0.93	Non-Toxin	-0.93	Non-Toxin	-1.14	Non-Toxin	-1.14	Non-Toxin	-0.26	Non-CPP	-0.12	Non-CPP	non-hemolytic	0.4153	0.282	Non-Allergen	False
WSAMPWIRY	909.11	2.53	THP	THP	0.93	Anticp	0.49	Anticp	-0.43	Non-Toxin	-0.43	Non-Toxin	-0.63	Non-Toxin	-0.63	Non-Toxin	-0.26	Non-CPP	-0.15	Non-CPP	non-hemolytic	0.1963	0.286	Non-Allergen	False
WWAMKSIRY	1006.61	1.46	THP	THP	0.38	Anticp	0.62	Anticp	-0.77	Non-Toxin	-0.77	Non-Toxin	-0.86	Non-Toxin	-0.86	Non-Toxin	-0.26	Non-CPP	-0.05	Non-CPP	non-hemolytic	0.4149	0.25	Non-Allergen	False
AMRWI	844.81	1.74	THP	THP	0.94	Anticp	0.53	Anticp	-0.84	Non-Toxin	-0.84	Non-Toxin	-0.8	Non-Toxin	-0.8	Non-Toxin	-0.26	Non-CPP	-0.25	Non-CPP	non-hemolytic	0.4291	0.243	Non-Allergen	-
AFGWY	823.41	1.56	THP	THP	0.69	Anticp	0.97	Anticp	-0.82	Non-Toxin	-0.82	Non-Toxin	-0.33	Non-Toxin	-0.33	Non-Toxin	-0.26	Non-CPP	-0.27	Non-CPP	non-hemolytic	0.5166	0.405	Allergen	-
FYGRCGGRLP	828.81	2.09	THP	THP	1.04	Anticp	0.98	Anticp	-0.72	Non-Toxin	-0.72	Non-Toxin	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.26	Non-CPP	-0.19	Non-CPP	non-hemolytic	0.397	0.398	Allergen	False
RQRLR	849.91	1.12	THP	THP	1.1	Anticp	0.18	Anticp	-0.91	Non-Toxin	-0.91	Non-Toxin	-0.49	Non-Toxin	-0.49	Non-Toxin	-0.26	Non-CPP	-0.02	Non-CPP	non-hemolytic	0.7244	0.382	Allergen	-
QRLRI	836.91	1.49	THP	THP	0.6	Anticp	0.42	Anticp	-1.06	Non-Toxin	-1.06	Non-Toxin	-0.71	Non-Toxin	-0.71	Non-Toxin	-0.26	Non-CPP	-0.22	Non-CPP	non-hemolytic	0.4002	0.346	Allergen	-

P. Physicochemical properties of SET 8, conformed by 27 lead THPs.

ID	Sequence	Length	Hydrophobicity	Steric hindrance	Sidebulk	Hydropathicity	Amphipathicity	Hydrophilicity	Net Hydrogen	Charge	pI	Mol wt
THP_YMG1	LPWCLRLRI	9	-0.09	0.51	0.51	0.69	0.49	-0.56	0.9	2	10.38	1282.81
THP_YMG2	NGRCWKG	7	-0.35	0.58	0.58	-1.39	0.77	0.23	1.12	2	9.55	877.1
THP_YMG3	WRPWPSHL	8	-0.14	0.38	0.38	-1.08	0.43	-0.64	0.89	1.5	10.11	1191.53
THP_YMG4	WSYWRQLPWFG	11	-0.03	0.53	0.53	-0.68	0.31	-1.11	0.92	1	9.1	1582.96
THP_YMG5	WRPLSWAP	8	-0.07	0.44	0.44	-0.52	0.27	-0.64	0.78	1	10.11	1109.41
THP_YMG6	PLSWPRWA	8	-0.07	0.44	0.44	-0.52	0.27	-0.64	0.78	1	10.11	1083.37
THP_YMG7	PRWPLSWA	8	-0.07	0.44	0.44	-0.52	0.27	-0.64	0.78	1	10.11	1083.37
THP_YMG8	HHGTPRWC	8	-0.25	0.37	0.37	-1.33	0.59	-0.31	0.89	2	8.61	1096.37
THP_YMG9	WSPYWLPR	8	-0.1	0.46	0.46	-0.87	0.27	-0.84	0.89	1	9.1	1260.58
THP_YMG10	RGDLRWC	7	-0.39	0.56	0.56	-0.94	0.61	0.35	1.25	1	8.6	1008.28
THP_YMG11	CGCGSCRSCR	10	-0.32	0.57	0.57	-0.13	0.45	0.24	0.91	2	8.67	1187.52
THP_YMG12	LRCWSRC	7	-0.35	0.52	0.52	-0.24	0.61	-0.11	1.25	2	9.1	1026.35
THP_YMG13	CNWWRLRAQFY	11	-0.22	0.57	0.57	-0.68	0.51	-0.71	1.25	2	9.55	1706.12
THP_YMG14	PSPAFKWW	8	0.01	0.46	0.46	-0.57	0.41	-0.72	0.56	1	9.11	1204.51
THP_YMG15	PYWARGWLP	9	-0.02	0.48	0.48	-0.56	0.25	-0.84	0.7	1	9.1	1242.58
THP_YMG16	TARGLCWRY	9	-0.23	0.54	0.54	-0.42	0.49	-0.34	1.1	2	9.55	1288.62
THP_YMG17	TAPYWLPWRY	10	-0.05	0.49	0.49	-0.65	0.22	-1.01	0.82	1	8.93	1515.88
THP_YMG18	AMYWRGFWWP	10	0.05	0.54	0.54	-0.36	0.22	-1.25	0.73	1	9.1	1496.9
THP_YMG19	WWWMGCRGS	9	-0.03	0.55	0.55	-0.44	0.25	-0.92	0.8	1	8.6	1255.57
THP_YMG20	CPGCRHGSGH	10	-0.21	0.44	0.44	-0.86	0.49	0.03	0.64	2	8.4	1147.41
THP_YMG21	HGSWRPWGH	9	-0.18	0.39	0.39	-1.59	0.54	-0.45	0.9	2	10.11	1256.5
THP_YMG22	WRPWSPTSC	9	-0.18	0.46	0.46	-0.93	0.25	-0.46	0.9	1	8.6	1222.52
THP_YMG23	HHWARGSHC	9	-0.24	0.35	0.35	-1.19	0.68	-0.31	0.9	2.5	8.61	1193.46
THP_YMG24	CPGCRWWWM	9	-0.02	0.52	0.52	-0.23	0.25	-1.05	0.7	1	8.39	1355.81
THP_YMG25	WWYWRGFWM	9	0.08	0.55	0.55	-0.51	0.25	-1.67	0.9	1	9.1	1548.99
THP_YMG26	RQRLPIGRR	9	-0.64	0.57	0.57	-1.52	1.1	0.86	1.8	4	12.48	1307.71
THP_YMG27	QHWSYKLPPR	10	-0.34	0.5	0.5	-1.75	0.88	-0.15	1.2	2.5	10.01	1311.65

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

	Sequence	TumorHPD		THPep	AntiCP				CpACpP									CellPPD				MLACP			
		SVM Score		SVM Score	SVM Score 1		SVM Score 2		SVM_ACP	SVM_CPP	SVM_CpACP	RF_ACP	RF_CPP	RF_CpACP	XGB_ACP	XGB_CPP	XGB_CpACP	SVM Score 1		SVM Score 1			Prob.	Uptake Eff.	Prob.
THP_YMG1	LPWCLRLRI	2.32	THP	THP	1.36	Anticp	1.18	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.22	Non-CPP	0.28	Non-CPP	CPP	0.7643	Low	0.4549
THP_YMG2	NGRCWKG	2.32	THP	THP	1.15	Anticp	1.1	Anticp										-0.18	Non-CPP	-0.22	Non-CPP	CPP	0.7751	Low	0.3296
THP_YMG3	WRPWPSHL	3.28	THP	THP	1.11	Anticp	0.78	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.29	Non-CPP	-0.25	Non-CPP	CPP	0.8171	Low	0.371
THP_YMG4	WSYWRQLPWFG	2.81	THP	THP	0.96	Anticp	0.76	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	0.01	Non-CPP	-0.1	Non-CPP	CPP	0.7123	Low	0.3444
THP_YMG5	WRPLSWAP	3.14	THP	THP	0.91	Anticp	0.87	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.12	Non-CPP	-0.25	Non-CPP	CPP	0.7651	Low	0.2131
THP_YMG6	PLSWPRWA	3.14	THP	THP	1.13	Anticp	0.87	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.26	Non-CPP	-0.25	Non-CPP	CPP	0.7651	Low	0.2131
THP_YMG7	PRWPLSWA	3.14	THP	THP	1.02	Anticp	0.87	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.11	Non-CPP	-0.25	Non-CPP	CPP	0.7651	Low	0.2131
THP_YMG8	HHGTPRWC	2.07	THP	THP	1.02	Anticp	1.2	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.26	Non-CPP	-0.25	Non-CPP	CPP	0.7577	Low	0.4167
THP_YMG9	WSPYWLPR	3.34	THP	THP	0.94	Anticp	0.86	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.15	Non-CPP	-0.25	Non-CPP	CPP	0.8122	Low	0.2947
THP_YMG10	RGDLRWC	2.32	THP	THP	1.4	Anticp	0.76	Anticp										-0.21	Non-CPP	-0.22	Non-CPP	CPP	0.6981	High	0.6024
THP_YMG11	CGCGSCRSCR	3.19	THP	THP	0.99	Anticp	1.45	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.13	Non-CPP	-0.22	Non-CPP	CPP	0.6307	Low	0.3324
THP_YMG12	LRCWSRC	3	THP	THP	1.15	Anticp	1.39	Anticp										-0.21	Non-CPP	-0.25	Non-CPP	CPP	0.8172	High	0.6999
THP_YMG13	CNWWRLRAQFY	2.27	THP	THP	1.04	Anticp	1.1	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	0.05	Non-CPP	0.2	Non-CPP	CPP	0.7114	Low	0.4958
THP_YMG14	PSPAFKWW	2.07	THP	THP	1.05	Anticp	0.88	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.25	Non-CPP	-0.26	Non-CPP	CPP	0.7569	Low	0.2157
THP_YMG15	PYWARGWLP	3	THP	THP	0.97	Anticp	1.55	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.26	Non-CPP	0.23	Non-CPP	CPP	0.6944	Low	0.2535
THP_YMG16	TARGLCWRY	2.07	THP	THP	0.7	Anticp	1.14	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.24	Non-CPP	-0.06	Non-CPP	CPP	0.7784	Low	0.3654
THP_YMG17	TAPYWLPWRY	2.67	THP	THP	0.72	Anticp	1.32	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.06	Non-CPP	0.02	Non-CPP	CPP	0.6803	Low	0.3384
THP_YMG18	AMYWRGFWWP	3.02	THP	THP	1.04	Anticp	1.24	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.18	Non-CPP	-0.05	Non-CPP	CPP	0.6154	Low	0.28
THP_YMG19	WWWMGCRGS	3.14	THP	THP	0.77	Anticp	0.93	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.25	Non-CPP	-0.24	Non-CPP	CPP	0.6684	Low	0.1729
THP_YMG20	CPGCRHGSGH	2.25	THP	THP	1.28	Anticp	1.24	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.64	Non-CPP	-0.34	Non-CPP	CPP	0.6261	Low	0.3322
THP_YMG21	HGSWRPWGH	2.47	THP	THP	1.03	Anticp	0.78	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.28	Non-CPP	0.48	CPP	CPP	0.7325	Low	0.3061
THP_YMG22	WRPWSPTSC	2.45	THP	THP	1.29	Anticp	0.74	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.25	Non-CPP	0.2	Non-CPP	CPP	0.6978	Low	0.1483
THP_YMG23	HHWARGSHC	2.21	THP	THP	0.8	Anticp	1.07	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.26	Non-CPP	0.31	CPP	CPP	0.7687	Low	0.361
THP_YMG24	CPGCRWWWM	3.24	THP	THP	1.22	Anticp	1.57	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.25	Non-CPP	0.42	CPP	CPP	0.644	Low	0.4308
THP_YMG25	WWYWRGFWM	2.78	THP	THP	0.76	Anticp	0.81	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.23	Non-CPP	0.3	CPP	CPP	0.667	Low	0.3574
THP_YMG26	RQRLPIGRR	1.16	THP	THP	1.03	Anticp	0.16	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	-0.26	Non-CPP	0.31	CPP	CPP	0.9809	Low	0.446
THP_YMG27	QHWSYKLPPR	1.62	THP	THP	0.68	Anticp	0.2	Anticp	ACP	CPP	CpACP	ACP	CPP	CpACP	ACP	CPP	CpACP	0.05	CPP	0.16	CPP	CPP	0.7826	Low	0.3066

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

	Sequence	ToxinPred								HemoPred	Macrel				PlifePred	PepSolubility	ANuPP	
		SVM Score 1		SVM Score 2		SVM Score 3		SVM Score 4			AMP?	Prob		Hemo Prob	Half-life (s)		APR?	Score
THP_YMG1	LPWCLRLRI	-0.37	Non-Toxin	-0.37	Non-Toxin	-0.43	Non-Toxin	-0.43	Non-Toxin	non-hemolytic	False	0.356	NonHemo	0.485	862.81	not soluble	False	1.59
THP_YMG2	NGRCWKG	-0.35	Non-Toxin	-0.35	Non-Toxin	-0.34	Non-Toxin	-0.34	Non-Toxin	non-hemolytic	False	0.347	NonHemo	0.317	900.91	not soluble	False	0.47
THP_YMG3	WRPWPSHL	-0.54	Non-Toxin	-0.54	Non-Toxin	-0.62	Non-Toxin	-0.62	Non-Toxin	non-hemolytic	False	0.149	NonHemo	0.406	846.51	not soluble	False	0.94
THP_YMG4	WSYWRQLPWFG	-1.22	Non-Toxin	-1.22	Non-Toxin	-1.17	Non-Toxin	-1.17	Non-Toxin	non-hemolytic	False	0.069	NonHemo	0.416	933.21	not soluble	False	2.29
THP_YMG5	WRPLSWAP	-1.03	Non-Toxin	-1.03	Non-Toxin	-0.67	Non-Toxin	-0.67	Non-Toxin	non-hemolytic	False	0.149	NonHemo	0.465	845.31	not soluble	False	1
THP_YMG6	PLSWPRWA	-1.21	Non-Toxin	-1.21	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin	non-hemolytic	False	0.079	NonHemo	0.347	845.31	not soluble	False	1
THP_YMG7	PRWPLSWA	-1.21	Non-Toxin	-1.21	Non-Toxin	-0.98	Non-Toxin	-0.98	Non-Toxin	non-hemolytic	False	0.129	NonHemo	0.356	845.31	not soluble	False	1.06
THP_YMG8	HHGTPRWC	-0.7	Non-Toxin	-0.7	Non-Toxin	-0.7	Non-Toxin	-0.7	Non-Toxin	non-hemolytic	False	0.228	NonHemo	0.307	836.11	not soluble	False	0.75
THP_YMG9	WSPYWLPR	-1.09	Non-Toxin	-1.09	Non-Toxin	-1.1	Non-Toxin	-1.1	Non-Toxin	non-hemolytic	False	0	NonHemo	0.475	847.11	not soluble	False	1.22
THP_YMG10	RGDLRWC	-0.55	Non-Toxin	-0.55	Non-Toxin	-0.62	Non-Toxin	-0.62	Non-Toxin	non-hemolytic	False	0.257	NonHemo	0.129	818.81	not soluble	False	0.67
THP_YMG11	CGCGSCRSCR	-0.41	Non-Toxin	-0.41	Non-Toxin	-0.35	Non-Toxin	-0.35	Non-Toxin	non-hemolytic	False	0.257	NonHemo	0.426	829.91	not soluble	False	1.37
THP_YMG12	LRCWSRC	-0.91	Non-Toxin	-0.91	Non-Toxin	-0.42	Non-Toxin	-0.42	Non-Toxin	non-hemolytic	False	0.218	NonHemo	0.495	831.01	not soluble	False	0.56
THP_YMG13	CNWWRLRAQFY	-0.54	Non-Toxin	-0.54	Non-Toxin	-1.14	Non-Toxin	-1.14	Non-Toxin	non-hemolytic	False	0.188	NonHemo	0.426	1032.01	not soluble	False	2.1
THP_YMG14	PSPAFKWW	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.93	Non-Toxin	-0.93	Non-Toxin	non-hemolytic	False	0.188	NonHemo	0.347	833.81	not soluble	False	1.11
THP_YMG15	PYWARGWLP	-0.97	Non-Toxin	-0.97	Non-Toxin	-0.93	Non-Toxin	-0.93	Non-Toxin	non-hemolytic	False	0.198	NonHemo	0.297	839.31	not soluble		
THP_YMG16	TARGLCWRY	-0.37	Non-Toxin	-0.37	Non-Toxin	-0.49	Non-Toxin	-0.49	Non-Toxin	non-hemolytic	False	0.158	NonHemo	0.317	970.01	not soluble	False	1.19
THP_YMG17	TAPYWLPWRY	-0.54	Non-Toxin	-0.54	Non-Toxin	-0.68	Non-Toxin	-0.68	Non-Toxin	non-hemolytic	False	0.168	NonHemo	0.277	824.51	not soluble	False	2.13
THP_YMG18	AMYWRGFWWP	-0.86	Non-Toxin	-0.86	Non-Toxin	-1.43	Non-Toxin	-1.43	Non-Toxin	non-hemolytic	False	0.248	NonHemo	0.366	839.11	not soluble	False	2.04
THP_YMG19	WWWMGCRGS	-0.83	Non-Toxin	-0.83	Non-Toxin	-0.84	Non-Toxin	-0.84	Non-Toxin	non-hemolytic	False	0.089	NonHemo	0.396	829.01	not soluble	False	1.4
THP_YMG20	CPGCRHGSGH	-0.83	Non-Toxin	-0.83	Non-Toxin	-0.42	Non-Toxin	-0.42	Non-Toxin	non-hemolytic	False	0.228	NonHemo	0.426	813.71	not soluble	False	1.41
THP_YMG21	HGSWRPWGH	-0.65	Non-Toxin	-0.65	Non-Toxin	-0.61	Non-Toxin	-0.61	Non-Toxin	non-hemolytic	False	0.168	NonHemo	0.347	833.21	not soluble	False	1.2
THP_YMG22	WRPWSPTSC	-0.74	Non-Toxin	-0.74	Non-Toxin	-0.54	Non-Toxin	-0.54	Non-Toxin	non-hemolytic	False	0.03	NonHemo	0.248	858.41	not soluble	False	1.21
THP_YMG23	HHWARGSHC	-0.64	Non-Toxin	-0.64	Non-Toxin	-0.58	Non-Toxin	-0.58	Non-Toxin	non-hemolytic	False	0.238	NonHemo	0.297	834.01	not soluble	False	1.35
THP_YMG24	CPGCRWWWM	-0.91	Non-Toxin	-0.91	Non-Toxin	-0.45	Non-Toxin	-0.45	Non-Toxin	non-hemolytic	False	0.109	NonHemo	0.386	835.41	not soluble	False	1.42
THP_YMG25	WWYWRGFWM	-1.13	Non-Toxin	-1.13	Non-Toxin	-1.2	Non-Toxin	-1.2	Non-Toxin	non-hemolytic	False	0.238	NonHemo	0.317	834.91	not soluble	False	1.93
THP_YMG26	RQRLPIGRR	-1.02	Non-Toxin	-1.02	Non-Toxin	-0.76	Non-Toxin	-0.76	Non-Toxin	non-hemolytic	False	0.287	Hemo	0.634	867.51	not soluble	False	1.2
THP_YMG27	QHWSYKLPPR	-0.67	Non-Toxin	-0.67	Non-Toxin	-0.96	Non-Toxin	-0.96	Non-Toxin	non-hemolytic	False	0.178	NonHemo	0.426	1154.31	not soluble	False	1.71

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

ID	Sequence	AlgPred2		IL2Pred		IL-10Pred		IL4pred		AIPpred		PRR			
		ML Score		Score		Score		Score			Prob	RF Score		SVM Score	
THP_YMG1	LPWCLRLRI	0.381	Allergen	0.806	IL-2 inducer	-0.055212105	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5326	0.44	Non-Pattern Recognition Receptor	0.54	Pattern Recognition Receptor
THP_YMG2	NGRCWKG	0.393	Allergen	0.69	IL-2 inducer	-0.660145052	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5209	0.58	Pattern Recognition Receptor	0.64	Pattern Recognition Receptor
THP_YMG3	WRPWPSHL	0.312	Allergen	0.928	IL-2 inducer	-0.273554559	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5163	0.47	Non-Pattern Recognition Receptor	0.54	Pattern Recognition Receptor
THP_YMG4	WSYWRQLPWFG	0.378	Allergen	0.716	IL-2 inducer	0.179412966	IL10 non-inducer	0.27	IL4 inducer	AIP	0.407	0.46	Non-Pattern Recognition Receptor	0.46	Non-Pattern Recognition Receptor
THP_YMG5	WRPLSWAP	0.27	Non-Allergen	0.736	IL-2 inducer	-0.379288801	IL10 non-inducer	0.28	IL4 inducer	Non-AIP	0.3302	0.46	Non-Pattern Recognition Receptor	0.43	Non-Pattern Recognition Receptor
THP_YMG6	PLSWPRWA	0.27	Non-Allergen	0.674	IL-2 inducer	-0.451490516	IL10 non-inducer	0.28	IL4 inducer	AIP	0.3651	0.46	Non-Pattern Recognition Receptor	0.43	Non-Pattern Recognition Receptor
THP_YMG7	PRWPLSWA	0.27	Non-Allergen	0.674	IL-2 inducer	-0.451490516	IL10 non-inducer	0.28	IL4 inducer	AIP	0.3651	0.46	Non-Pattern Recognition Receptor	0.43	Non-Pattern Recognition Receptor
THP_YMG8	HHGTPRWC	0.283	Non-Allergen	0.714	IL-2 inducer	-0.676544903	IL10 non-inducer	0.28	IL4 inducer	AIP	0.4256	0.48	Non-Pattern Recognition Receptor	0.51	Pattern Recognition Receptor
THP_YMG9	WSPYWLPR	0.305	Allergen	0.858	IL-2 inducer	-0.136880401	IL10 non-inducer	0.26	IL4 inducer	AIP	0.5186	0.42	Non-Pattern Recognition Receptor	0.41	Non-Pattern Recognition Receptor
THP_YMG10	RGDLRWC	0.363	Allergen	0.838	IL-2 inducer	-0.764152544	IL10 non-inducer	0.28	IL4 inducer	Non-AIP	0.3372	0.52	Pattern Recognition Receptor	0.42	Non-Pattern Recognition Receptor
THP_YMG11	CGCGSCRSCR	0.373	Allergen	0.894	IL-2 inducer	-1.140580429	IL10 non-inducer	0.28	IL4 inducer	AIP	0.4628	0.48	Non-Pattern Recognition Receptor	0.49	Non-Pattern Recognition Receptor
THP_YMG12	LRCWSRC	0.32	Allergen	0.984	IL-2 inducer	-0.505084432	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5256	0.52	Pattern Recognition Receptor	0.5	Pattern Recognition Receptor
THP_YMG13	CNWWRLRAQFY	0.352	Allergen	0.868	IL-2 inducer	0.050828942	IL10 non-inducer	0.27	IL4 inducer	AIP	0.5186	0.45	Non-Pattern Recognition Receptor	0.4	Non-Pattern Recognition Receptor
THP_YMG14	PSPAFKWW	0.38	Allergen	0.86	IL-2 inducer	-0.294363533	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5209	0.47	Non-Pattern Recognition Receptor	0.46	Non-Pattern Recognition Receptor
THP_YMG15	PYWARGWLP	0.285	Non-Allergen	0.75	IL-2 inducer	-0.101615187	IL10 non-inducer	0.28	IL4 inducer	AIP	0.4256	0.46	Non-Pattern Recognition Receptor	0.34	Non-Pattern Recognition Receptor
THP_YMG16	TARGLCWRY	0.357	Allergen	0.402	IL-2 non-inducer	-0.174363382	IL10 non-inducer	0.24	IL4 inducer	AIP	0.5233	0.43	Non-Pattern Recognition Receptor	0.33	Non-Pattern Recognition Receptor
THP_YMG17	TAPYWLPWRY	0.29	Non-Allergen	0.724	IL-2 inducer	-0.063655925	IL10 non-inducer	0.27	IL4 inducer	AIP	0.4884	0.44	Non-Pattern Recognition Receptor	0.36	Non-Pattern Recognition Receptor
THP_YMG18	AMYWRGFWWP	0.258	Non-Allergen	0.454	IL-2 non-inducer	0.013567739	IL10 non-inducer	0.27	IL4 inducer	AIP	0.5721	0.42	Non-Pattern Recognition Receptor	0.38	Non-Pattern Recognition Receptor
THP_YMG19	WWWMGCRGS	0.317	Allergen	0.806	IL-2 inducer	-0.610054741	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5628	0.5	Pattern Recognition Receptor	0.54	Pattern Recognition Receptor
THP_YMG20	CPGCRHGSGH	0.349	Allergen	0.85	IL-2 inducer	-0.642309999	IL10 non-inducer	0.27	IL4 inducer	AIP	0.3884	0.47	Non-Pattern Recognition Receptor	0.52	Pattern Recognition Receptor
THP_YMG21	HGSWRPWGH	0.312	Allergen	0.852	IL-2 inducer	-0.629949166	IL10 non-inducer	0.26	IL4 inducer	AIP	0.4535	0.5	Pattern Recognition Receptor	0.62	Pattern Recognition Receptor
THP_YMG22	WRPWSPTSC	0.298	Non-Allergen	0.898	IL-2 inducer	-0.526674545	IL10 non-inducer	0.19	Non IL4 inducer	AIP	0.4884	0.48	Non-Pattern Recognition Receptor	0.58	Pattern Recognition Receptor
THP_YMG23	HHWARGSHC	0.277	Non-Allergen	0.832	IL-2 inducer	-0.593692415	IL10 non-inducer	0.28	IL4 inducer	AIP	0.4814	0.53	Pattern Recognition Receptor	0.54	Pattern Recognition Receptor
THP_YMG24	CPGCRWWWM	0.323	Allergen	0.74	IL-2 inducer	-0.80685999	IL10 non-inducer	0.28	IL4 inducer	AIP	0.5419	0.44	Non-Pattern Recognition Receptor	0.48	Non-Pattern Recognition Receptor
THP_YMG25	WWYWRGFWM	0.254	Non-Allergen	0.758	IL-2 inducer	-0.161368748	IL10 non-inducer	0.26	IL4 inducer	AIP	0.6512	0.38	Non-Pattern Recognition Receptor	0.42	Non-Pattern Recognition Receptor
THP_YMG26	RQRLPIGRR	0.416	Allergen	0.558	IL-2 inducer	0.429411895	IL10 non-inducer	0.25	IL4 inducer	AIP	0.5256	0.5	Pattern Recognition Receptor	0.34	Non-Pattern Recognition Receptor
THP_YMG27	QHWSYKLPPR	0.333	Allergen	0.862	IL-2 inducer	-0.353241621	IL10 non-inducer	0.1	Non IL4 inducer	AIP	0.4163	0.44	Non-Pattern Recognition Receptor	0.42	Non-Pattern Recognition Receptor

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

ID	Sequence	QSPpred												dPABBS			
		Comp	Binary	Physico	Hybrid	KNN Comp	KNN Binary	KNN Physico	KNN Hybrid	RF Comp	RF Binary	RF Physico	RF Hybrid	SVM score	Prediction - dPABBs	WEKA Probability	Prediction - dPABBs
THP_YMG1	LPWCLRLRI	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	-0.47	Biofilm-inactive	0.74	Biofilm-active
THP_YMG2	NGRCWKG	QSP	QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.33	Biofilm-inactive	0.71	Biofilm-active
THP_YMG3	WRPWPSHL	QSP	QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	0.02	Biofilm-inactive	0.29	Biofilm-inactive
THP_YMG4	WSYWRQLPWFG	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.21	Biofilm-inactive	0.46	Biofilm-inactive
THP_YMG5	WRPLSWAP	QSP	QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.01	Biofilm-inactive	0.33	Biofilm-inactive
THP_YMG6	PLSWPRWA	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.01	Biofilm-inactive	0.33	Biofilm-inactive
THP_YMG7	PRWPLSWA	QSP	Non-QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	-0.01	Biofilm-inactive	0.33	Biofilm-inactive
THP_YMG8	HHGTPRWC	Non-QSP	Non-QSP	QSP	Non-QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	0.17	Biofilm-inactive	0.55	Biofilm-active
THP_YMG9	WSPYWLPR	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	QSP	QSP	QSP	-0.26	Biofilm-inactive	0.23	Biofilm-inactive
THP_YMG10	RGDLRWC	QSP	Non-QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.48	Biofilm-inactive	0.5	Biofilm-active
THP_YMG11	CGCGSCRSCR	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	Non-QSP	QSP	-0.74	Biofilm-inactive	0.06	Biofilm-inactive
THP_YMG12	LRCWSRC	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.27	Biofilm-inactive	0.49	Biofilm-inactive
THP_YMG13	CNWWRLRAQFY	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	Non-QSP	QSP	QSP	QSP	QSP	0.43	Biofilm-inactive	0.67	Biofilm-active
THP_YMG14	PSPAFKWW	QSP	QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	0	Biofilm-inactive	0.38	Biofilm-inactive
THP_YMG15	PYWARGWLP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.03	Biofilm-inactive	0.54	Biofilm-active
THP_YMG16	TARGLCWRY	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	0.39	Biofilm-inactive	0.71	Biofilm-active
THP_YMG17	TAPYWLPWRY	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	0.44	Biofilm-inactive	0.66	Biofilm-active
THP_YMG18	AMYWRGFWWP	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	QSP	QSP	QSP	0.47	Biofilm-inactive	0.52	Biofilm-active
THP_YMG19	WWWMGCRGS	QSP	Non-QSP	QSP	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	-0.48	Biofilm-inactive	0.2	Biofilm-inactive
THP_YMG20	CPGCRHGSGH	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	QSP	-0.6	Biofilm-inactive	0.22	Biofilm-inactive
THP_YMG21	HGSWRPWGH	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	QSP	QSP	QSP	-0.13	Biofilm-inactive	0.31	Biofilm-inactive
THP_YMG22	WRPWSPTSC	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.71	Biofilm-inactive	0.19	Biofilm-inactive
THP_YMG23	HHWARGSHC	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	QSP	Non-QSP	0.43	Biofilm-inactive	0.49	Biofilm-inactive
THP_YMG24	CPGCRWWWM	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	-0.16	Biofilm-inactive	0.5	Biofilm-active
THP_YMG25	WWYWRGFWM	QSP	QSP	QSP	QSP	Non-QSP	QSP	QSP	Non-QSP	QSP	QSP	QSP	QSP	0.16	Biofilm-inactive	0.54	Biofilm-active
THP_YMG26	RQRLPIGRR	QSP	Non-QSP	Non-QSP	QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	QSP	Non-QSP	0.67	Biofilm-active	0.26	Biofilm-active
THP_YMG27	QHWSYKLPPR	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	Non-QSP	QSP	QSP	Non-QSP	0.38	Biofilm-inactive	-0.2	Biofilm-inactive

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

ID	Sequence	AMPDiscover										ClassAMP	CAMPr3				AMPClassifier			
		RF					ANN						ANN	DA	RF	SVM	kNN	SVM	RF	XGBoost
THP_YMG1	LPWCLRLRI	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	AMP	AMP	AMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG2	NGRCWKG	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	NAMP	AMP	NAMP	AMP	AMP	AMP	AMP	AMP
THP_YMG3	WRPWPSHL	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	NAMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG4	WSYWRQLPWFG	AMP	ABP	AFP	nonAPP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	AMP	AMP	NAMP	non-AMP	AMP	non-AMP	AMP
THP_YMG5	WRPLSWAP	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	NAMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG6	PLSWPRWA	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	NAMP	AMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG7	PRWPLSWA	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	NAMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG8	HHGTPRWC	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	NAMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG9	WSPYWLPR	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	NAMP	AMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG10	RGDLRWC	AMP	nonABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	NAMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG11	CGCGSCRSCR	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	AMP	AMP	AMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG12	LRCWSRC	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	AMP	AMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG13	CNWWRLRAQFY	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	AMP	NAMP	NAMP	non-AMP	AMP	non-AMP	AMP
THP_YMG14	PSPAFKWW	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	AMP	AMP	NAMP	AMP	AMP	AMP	AMP	AMP
THP_YMG15	PYWARGWLP	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	NAMP	AMP	NAMP	NAMP	non-AMP	AMP	AMP	non-AMP
THP_YMG16	TARGLCWRY	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	AMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG17	TAPYWLPWRY	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	NAMP	AMP	NAMP	NAMP	AMP	AMP	non-AMP	AMP
THP_YMG18	AMYWRGFWWP	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	AMP	AMP	AMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG19	WWWMGCRGS	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	AMP	AMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG20	CPGCRHGSGH	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	AMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG21	HGSWRPWGH	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	NAMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG22	WRPWSPTSC	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	NAMP	NAMP	NAMP	non-AMP	AMP	AMP	AMP
THP_YMG23	HHWARGSHC	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antifungal	NAMP	NAMP	NAMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG24	CPGCRWWWM	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	AMP	AMP	AMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG25	WWYWRGFWM	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antibacterial	AMP	AMP	AMP	NAMP	AMP	AMP	AMP	AMP
THP_YMG26	RQRLPIGRR	AMP	ABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	NAMP	AMP	AMP	AMP	AMP	AMP	AMP
THP_YMG27	QHWSYKLPPR	AMP	nonABP	AFP	APP	AVP	AMP	ABP	AFP	APP	AVP	Antiviral	NAMP	NAMP	NAMP	NAMP	AMP	AMP	AMP	AMP

P. (cont.) Predicted activities of SET 8, conformed by 27 lead THPs.

ID	Sequence	AMPfun								Meta-IAVP	iAMpred			AxPEP	Antifp		AntiTbPred	
		AMP	Antiparasitic	Antiviral	Anticancer	Targeting mammals	Anti-fungal	Targeting Gram (-)	Targeting Gram (+)		Score antibacterial	Score antiviral	Score antifungal	Score Antimicrobial	Score	Prediction	Score	Score
THP_YMG1	LPWCLRLRI	0.9844	0.5182	0.8667	0.4245	0.3333	0.5375	0.9167	0.8363	AVP	0.055	0.628	0.071	0.71	-0.8118754	Non-Antifungal	0.93772771	0.7280815
THP_YMG2	NGRCWKG	0.9862	0.2727	0.6	0.3636	0.0667	0.6723	0.775	0.6915	Non-AVP	0.813	0.658	0.505	0.32	0.79718338	Antifungal	-0.23979716	1.1932261
THP_YMG3	WRPWPSHL	0.9763	0.1545	0.925	0.4404	0.0667	0.5453	0.6583	0.6642	Non-AVP	0.783	0.406	0.527	0.74	-0.62622508	Non-Antifungal	-0.44607806	0.76633401
THP_YMG4	WSYWRQLPWFG	0.996	0.1091	0.6333	0.6447	0.1	0.4749		0.6539	Non-AVP	0.828	0.388	0.394	0.63	-0.23566232	Non-Antifungal	0.78007734	2.0333341
THP_YMG5	WRPLSWAP	0.9938	0.0545	0.7833	0.5128	0.0667	0.5525	0.6083	0.7273	AVP	0.783	0.229	0.361	0.85	-0.61259939	Non-Antifungal	0.22572689	0.79420516
THP_YMG6	PLSWPRWA	0.99	0.0455	0.7167	0.4194	0.0667	0.3872	0.775	0.6996	AVP	0.777	0.224	0.354	0.51	-0.20985209	Non-Antifungal	0.22719556	0.80155775
THP_YMG7	PRWPLSWA	0.995	0.1273	0.7167	0.385	0.0667	0.5236	0.6917	0.7847	AVP	0.777	0.223	0.354	0.63	-0.33683856	Non-Antifungal	0.21691015	0.79944312
THP_YMG8	HHGTPRWC	0.9659	0.2273	0.6333	0.5314	0.1	0.7236	0.5583	0.5811	Non-AVP	0.758	0.686	0.51	0.45	0.25321613	Non-Antifungal	-0.067360765	0.70995073
THP_YMG9	WSPYWLPR	0.9825	0.1636	0.775	0.578	0.1	0.4922	0.6917	0.6286	AVP	0.78	0.33	0.578	0.76	-0.05488184	Non-Antifungal	-0.44186145	1.0573971
THP_YMG10	RGDLRWC	0.9638	0.3455	0.775	0.3859	0.1	0.5933	0.5833	0.5287	AVP	0.243	0.539	0.208	0.32	-0.32753846	Non-Antifungal	0.34355081	0.66435081
THP_YMG11	CGCGSCRSCR	0.979	0.2	0.5833	0.5676	0.0333	0.6502	0.8	0.7469	Non-AVP	0.874	0.898	0.799	0.66	1.1961358	Antifungal	-0.67749451	0.56310059
THP_YMG12	LRCWSRC	0.99	0.2636	0.7667	0.4995	0.1	0.8337	0.7167	0.6923	Non-AVP	0.481	0.772	0.399	0.28	0.24398566	Antifungal	-0.82271693	1.0600832
THP_YMG13	CNWWRLRAQFY	0.9867	0.3727	0.6167	0.4197	0.0667	0.4771	0.6	0.5417	AVP	0.647	0.507	0.167	0.5	-0.90660105	Non-Antifungal	1.3840615	2.074092
THP_YMG14	PSPAFKWW	0.9867	0.2909	0.7583	0.3503	0.1333	0.5026	0.7167	0.5767	AVP	0.818	0.61	0.383	0.41	-0.41760293	Non-Antifungal	-0.062949073	0.65767449
THP_YMG15	PYWARGWLP	0.9817	0.1273	0.7583	0.6639	0.2	0.4754	0.6917	0.5523	Non-AVP	0.751	0.379	0.456	0.55	-0.86812146	Non-Antifungal	1.0873826	1.4211057
THP_YMG16	TARGLCWRY	0.9863	0.1273	0.725	0.444	0.1333	0.676	0.7333	0.6556	AVP	0.482	0.566	0.569	0.4	-0.52801402	Non-Antifungal	2.6482984	1.8209758
THP_YMG17	TAPYWLPWRY	0.99	0.0091	0.8	0.644	0.2	0.3619	0.7	0.6267	AVP	0.727	0.461	0.518	0.48	-0.74428438	Non-Antifungal	0.53912696	1.1989435
THP_YMG18	AMYWRGFWWP	0.9891	0.2545	0.7833	0.6715	0	0.4247	0.675	0.6045	AVP	0.776	0.672	0.389	0.57	-0.3361888	Non-Antifungal	1.0619798	1.5119185
THP_YMG19	WWWMGCRGS	0.9716	0.3636	0.575	0.4076	0	0.6874	0.7417	0.6765	Non-AVP	0.711	0.638	0.44	0.54	0.33264286	Antifungal	-0.32409069	0.95468052
THP_YMG20	CPGCRHGSGH	0.9832	0.2545	0.6	0.3414	0.1	0.7226	0.775	0.6803	AVP	0.891	0.776	0.786	0.66	0.44008087	Antifungal	-0.26021683	0.33392327
THP_YMG21	HGSWRPWGH	0.98	0.1182	0.8083	0.3211	0.1333	0.5176	0.7167	0.6482	Non-AVP	0.72	0.567	0.488	0.69	0.024978864	Non-Antifungal	-0.1633616	0.82942079
THP_YMG22	WRPWSPTSC	0.9863	0.1364	0.9	0.3045	0.0333	0.5182	0.6333	0.5477	Non-AVP	0.767	0.317	0.457	0.73	0.091045668	Non-Antifungal	-0.76873198	0.60133054
THP_YMG23	HHWARGSHC	0.97	0.2727	0.6583	0.4811	0.1	0.7814	0.75	0.7936	Non-AVP	0.736	0.739	0.467	0.59	-0.14851988	Non-Antifungal	-0.098896869	0.69554753
THP_YMG24	CPGCRWWWM	0.9841	0.2182	0.725	0.3506	0	0.6013	0.7917	0.7895	Non-AVP	0.705	0.717	0.477	0.47	0.070877086	Antifungal	-0.07697802	0.9764879
THP_YMG25	WWYWRGFWM	0.973	0.4	0.7917	0.6427	0.0333	0.6068	0.7667	0.677	Non-AVP	0.666	0.609	0.433	0.53	-0.79464479	Non-Antifungal	-0.13798628	0.97381843
THP_YMG26	RQRLPIGRR	0.9907	0.5455	0.7417	0.4788	0.0667	0.7151	0.7417	0.6352	AVP	0.894	0.573	0.663	0.65	-0.30205662	Non-Antifungal	0.55902755	1.049842
THP_YMG27	QHWSYKLPPR	0.9793	0.0909	0.5667	0.1056	0.1333	0.5046	0.65	0.5732	Non-AVP	0.324	0.106	0.293	0.46	0.14825637	Antifungal	-1.0168668	1.0873239