Elif Çisem ÇÖLDÜR

MASTER OF SCIENCE THESIS

2024-ANTALYA

T.C.

AKDENIZ UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

DEPARTMENT OF GENE AND CELL THERAPY

CELL PENETRATING PEPTIDE PREDICTION AND VALIDATION BY MACHINE LEARNING FOR EFFICIENT DELIVERY OF GENE THERAPY VECTORS

Elif Çisem ÇÖLDÜR

MASTER OF SCIENCE THESIS

2024-ANTALYA

T.C.
AKDENIZ UNIVERSITY
INSTITUTE OF HEALTH SCIENCES
DEPARTMENT OF GENE AND CELL THERAPY

CELL PENETRATING PEPTIDE PREDICTION AND VALIDATION BY MACHINE LEARNING FOR EFFICIENT DELIVERY OF GENE THERAPY VECTORS

Elif Çisem ÇÖLDÜR

MASTER OF SCIENCE THESIS

ADVISOR
Assoc. Prof. Dr. Devrim DEMİR DORA

2024-ANTALYA

To the Institute of Health Sciences;

This study has been approved by our jury as a Master of Science Thesis in the Department of Gene and Cell Therapy, Gene and Cell Therapy Master of Science program. 14/02/2024

Signatures

Thesis Advisor : Assoc. Prof. Dr. Devrim DEMİR DORA
Akdeniz University

Member : Prof. Dr. Ahter Dilşad ŞANLIOĞLU
Akdeniz University

Member : Asst. Prof. Dr. Ongun Mehmet SAKA
Ankara University

This thesis has been deemed by the jury appointed by the Administrative Board of the Institute and accepted by the Administrative Board of the Institute with decision dated ……/……./……. and numbered ……/……..

Institute Director

DECLARATION

I affirm that this thesis is the result of my individual effort, free from any unethical conduct throughout the entire process, from planning to completion. I have obtained all information included in this thesis in compliance with ethical principles, and I have duly credited all external information and opinions that were not generated within the scope of this thesis. Furthermore, I have appropriately documented and listed all such references in the bibliography.

Elif Çisem ÇÖLDÜR

Thesis Advisor
Assoc. Prof. Dr. Devrim DEMİR DORA

ACKNOWLEDGEMENTS

I would like to give my warmest thanks to my supervisor Assoc. Prof. Dr. Devrim DEMİR DORA who made this work possible,

I extend my sincerest gratitude to Prof. Dr. Salih ŞANLIOĞLU, who has consistently provided support throughout my academic journey until today,

I convey my deepest appreciation to Yaren SEVER and MSc. Zeynep Güneş TEPE for their invaluable guidance, support, and encouragement throughout the research process. Their expertise and insights have played a crucial role in shaping this thesis,

I am also grateful to Gizem TUTKUN for her constructive feedback and valuable suggestions, which significantly contributed to the enhancement of this work,

I express my sincere thanks to the esteemed members of my thesis defense committee, Prof. Dr. Ahter Dilşad ŞANLIOĞLU, Assoc. Prof. Dr. Devrim DEMİR DORA, and Asst. Prof. Dr. Ongun Mehmet SAKA,

A special acknowledgment goes to my family for their unwavering support, understanding, and encouragement during the challenging phases of this academic journey.

ABSTRACT

Objective: Predicting cell-penetrating peptides (CPP) through machine learning is vital for advancing gene therapy. This study aims to use a model based on peptide chain sequences and physicochemical properties to distinguish between CPPs and non-CPPs, crucial for improving the targeted delivery of therapeutic genes.

Method: Approximately 1400 different peptide chains were utilized as data, incorporating features such as amino acid composition, length, entropy, and CTDC, along with peptide sequences as inputs. The output is binary, with a result of "1" indicating a CPP and "0" denoting a non-CPP. The Random Forest model was chosen as the machine learning model for classification. Predictions were made for different peptides not present in the dataset. In the experimental section of the study, the reowned CPP protamine was employed, demonstrating its ability to enhance cellular internalization.

Results: The model achieved a training accuracy of 91.96% and a validation accuracy of 86.62%, demonstrating its effectiveness in predicting peptide classification. The developed algorithm has accurately predicted nearly all of the different peptide samples that were not present in the dataset. Experimental validation in the study confirmed the CPP property of protamine. It was determined that the utilization of protamine resulted in a remarkable 64% increase in transduction efficiency. This underscores the effectiveness of protamine in facilitating the entry of therapeutic agents into cells, showcasing its potential utility in enhancing various biomedical applications, including gene delivery.

Conclusion: Through this study, a novel algorithm has been developed to predict CPPs, paving the way for its application in gene therapy research. The algorithm enhances the identification of CPPs, facilitating targeted delivery of therapeutic genes and potentially improving the precision of gene-based treatments. Its efficiency in discovering CPPs may expedite the development of innovative gene delivery methods, advancing the field of gene therapy.

Keywords: CPP, *in silico*, machine learning, gene therapy

ÖZET

Amaç: Bu tez çalışmasının amacı, gen tedavisi uygulamalarında kritik bir öneme sahip olan hücre zarına penetre olabilen peptitlerin (CPP) makine öğrenme modeli kullanılarak CPP olma veya CPP olmama özelliğinin tahmin edilmesidir. CPP'ler terapötik materyalin hücre içine etkin bir şekilde ulaştırılmasında büyük öneme sahiptir. Bu bağlamda, bu çalışmada geliştirilen yeni algoritma kullanılarak gen tedavisinde kullanılabilecek yeni CPP'lerin kısa sürede verimli şekilde elde edilmesi amaçlanmaktadır.

Yöntem: Veri seti yaklaşık olarak 1400 farklı peptit zinciri kullanılarak oluşturulmuştur. Literatürdeki çalışmalarda önemi gösterilmiş olan amino asit kompozisyonu, uzunluk, entropi ve CTDC gibi fizikokimyasal özellikler kullanılarak makine öğrenme modeli girdileri oluşturulmuştur. CPP olan peptitler algortima ile "1" olarak sınıflandırılırken CPP olmayanlar "0" olarak sınıflandırılmıştır. Makine öğrenme modeli olarak Random Forest modeli seçilmiştir. Veri setinde bulunmayan farklı peptitler için tahminler yapılmıştır. Çalışmanın deneysel bölümünde, CPP olduğu bilinen protamin kullanılarak protaminin genetik materyalin hücre içine girşini arttırdığı gösterilmiştir.

Bulgular: Tez kapsamında geliştirilmiş olan model, %91,96 eğitim doğruluğu ve %86,62 doğrulama doğruluğu ile peptit sınıflandırmasını etkili bir şekilde gerçekleştirmiştir. Ayrıca, algoritma, veri setinde bulunmayan peptit örneklerinin büyük bir kısmını doğru bir şekilde tahmin etmiştir. Yapılan deneylerde, protaminin CPP özelliğine sahip olduğu gösterilmiş ve protamin kullanımının transdüksiyon veriminde %64'lük anlamlı bir artış sağladığı belirlenmiştir.

Sonuç: Bu çalışma, CPP tahmini yapan bir algoritma geliştirerek, *in vitro* deneyler öncesinde genetik materyalin taşıma verimliliğini hızlı ve etkili bir şekilde planlamayı mümkün kılmaktadır. Bu sayede, maddi kazancı artırmayı ve zaman kaybını minimize etmeyi hedeflemektedir. Ayrıca, gelecekte hücre içine aktarılmak istenen çeşitli materyallerin taşıma verimliliklerini tahmin etmek için yeni makine öğrenme algoritmalarına rehberlik edebilecek bir örnek çalışma olmuştur.

Anahtar Kelimeler: CPP, *in silico*, makine öğrenmesi, gen tedavisi

TABLE OF CONTENTS

ÖZET i

ABSTRACT ii

TABLE OF CONTENTS iii

LIST OF TABLES vi

INDEX OF FIGURES vii

ICONS and ABBREVIATIONS ix

1. INTRODUCTION 1

2. REVIEW OF THE LITERATURE 2

2.1 Gene Therapy: Definition, Significance, and Challenges 2

2.2 Objectives of Gene Therapy 4

2.3 The Significance of Intracellular Gene Transfer 4

2.4 Gene Delivery Methods: Viral and Non-viral Approaches 5

2.5 Advantages and Disadvantages of Gene Delivery Methods 7

2.6 Utilization of GFP Plasmids in Gene Therapy 8

2.7 Cell-Penetrating Peptides (CPPs): Key Players in Enhancing Gene Delivery 9

2.8 Protamine: A Comprehensive Insight into its Composition and CPP Attributes 11

2.9 A General Overview of Machine Learning 13

2.10 Steps in Machine Learning 15

2.11 Understanding Random Forest and Alternative Machine Learning Models 17

2.12 The Definitions and Importance of Features in Machine Learning Model 19

2.13 Understanding Overfitting, Feature Importance, and Underfitting in Machine Learning 28

2.14 Exploring Key Evaluation Metrics in Machine Learning: Accuracy, Recall, and F1 Score 29

2.15 The Use of Machine Learning in Genetic Studies 30

2.16 Insights into CPP: A Perspective from Machine Learning Studies 31

3. MATERIALS and METHODS 33

3.1 Machine Learning Steps 35

3.1.1. Data Collection 35

3.1.2 Data Preprocessing 35

3.1.3 Data Analysis 39

3.1.4 Choosing a Model 43

3.1.5 Training the Model 44

3.1.6 Evaluating the Model 45

3.1.7 Feature Importance 45

3.1.8 Model Optimization 46

3.1.9 Prediction 47

3.2 Experimental Part 48

3.2.2 Cell Culture Material & Plasmids 48

3.3 Maintenance and Propagation of Mammalian Suspension Cell Lines 50

3.4 Lentiviral Transfection/Transduction with Spinoculation Protocol 50

3.4.1 Production of Lentivirus 50

3.4.2 Lentiviral Transduction 50

4. RESULTS 52

4.1 Results of Machine Learning Algorithm 52

4.2 Results of Experimental Part 54

5. DISCUSSION 57

6. CONCLUSION AND SUGGESTIONS 66

REFERENCES 67

RESUME 77

LIST OF TABLES

Table 1 A comparison table for viral and non-viral delivery systems used in gene therapy 7

Table 2 The categorization of amino acids into three groups including Group1, Group 2, Group 3 24

Table 3 CTDC categorization chart 25

Table 4 The table showing a portion of the dataset used in this study 36

Table 5 List of features used as input in this study 38

Table 6 Peptide sequences not found in dataset subjected to prediction processes using a machine learning model 48

Table 7 Cell lines, cell culture media and equipment 48

Table 8 Training accuracy and validation accuracy values of the initial model and model with feature selection 52

Table 9 The results of the performance metrics 53

Table 10 Prediction results of peptides that were not exist in dataset 53

INDEX OF FIGURES

Figure 1 A cell with a mutation can be treated by replacing the mutated/non-functioning gene with a functioning gene 2

Figure 2 Schema of two distinct technique of gene therapy such as *in vivo* and *ex vivo.* 3

Figure 3 The depiction of the behavior of CPPs in both extracellular and intracellular environments. 10

Figure 4 The diagram of the benefits of machine learning 15

Figure 5 The diagram illustrating the seven stages of the machine learning process 17

Figure 6 Demonstration of random forest model workflow 18

Figure 7 The four distinct applications of the AAIndex. 22

Figure 8 The workflow of the machine learning aspect in this study 34

Figure 9 The distribution of CPP and non-CPP data in this study 39

Figure 10 The graph shows the distribution of sequences in group 3 as CPPs or non-CPPs, based on their normalized van der Waals values 40

Figure 11 The relation between being CPP/non-CPP and hydrophobicity 41

Figure 12 The correlation between polarizability-G3 and classification as CPP/non-CPP 42

Figure 13 The influence of aspartic acid (D) on the penetration capability of peptides 42

Figure 14 The graph of the impact of the amount of proline in a peptide sequence on whether the peptide is classified as CPP or non-CPP 43

Figure 15 The division of training and test data 44

Figure 16 The graph of feature importance technique 46

Figure 17 The summary of the model optimization 47

Figure 18 The plasmid map of PTRIP-SFFV-GFPEV 49

Figure 19 Comparison of control and protamine complex group under fluorescent microscopy 54

Figure 20 Transduction efficiencies of non-transfected, non-stimulated and with protamine 55

Figure 21 The bright-field and fluorescent microscopy images of the control and protamine groups 56

ICONS AND ABBREVIATIONS

CPP	: Cell Penetrating Peptide
HIV	: Human Immunodeficiency Virus
DNA	: Deoxyribonucleic Acid
RNA	: Ribonucleic Acid
SIN	: Self-inactivating
CRISPR-Cas9	: Clustered Regularly Interspaced Short Palindromic Repeats CRISPR associated protein 9
GFP	: Green Fluorescent Protein
SFFV	: Spleen Focus Forming Virus
siRNA	: Small Interfering RNA
PEG	: Polyethylene Glycol
LMWP	: Low Molecular Weight Protamine
mRNA	: Messenger RNA
GPU	: Graphics Processing Unit
ROC	: Receiver Operating Characteristic
CTDC	: Composition/Transition/Distribution - Compostion
AUC-ROC	**:** Area Under the Receiver Operating Characteristic Curve
MHC	: Major Histocompatibility Complex
SUMO	: Small Ubiquitin-related Modifier
DBGET	: Integrated Database Retrieval System
LinkDB	: Database of Link Information

G	: Glycine
A	: Alanine
V	: Valine
L	: Leucine
I	: Isoleucine
P	: Proline
F	: Phenylalanine
Y	: Tyrosine
W	: Trytophan
S	: Serine
T	: Threonine
C	: Cysteine
M	: Methionine
N	: Asparagine
Q	: Glutamine
D	: Aspartic Acid
E	: Glutamic Acid
K	: Lysine
R	: Arginine
H	: Histidine

1. INTRODUCTION

In the existing literature, several studies have investigated the utilization of cell-penetrating peptides (CPPs) for delivering nucleic acids, siRNA, plasmid DNA, drugs, and peptides with therapeutic potential (Ye et al., 2017; Zhu Z. et al., 2018; Kim G. C. et al., 2021). Reviewing these studies reveals that carrier systems modified with CPPs exhibit higher efficacy and lower toxicity compared to unmodified carrier systems (Jeong et al., 2016; Porosk et al., 2019; Liu Y. et al., 2021; Nam et al., 2023). Hence, CPPs play a pivotal role in the success of gene therapies.

The goal of this study is to employ a machine learning model to classify peptides as CPP/non-CPP according to their amino acid sequences and various physicochemical properties. By conducting *in silico* predictions before *in vitro* studies, the identification of peptides with low efficacy becomes possible, allowing for their exclusion from experimental use. Peptides identified as highly effective can then undergo further evaluation in *in vitro* studies, providing a time- and cost-effective contribution to research.

To experimentally validate the algorithm's accuracy, the study utilized protamine, a well-known CPP, to assess its impact on cellular uptake and penetration.

2. REVIEW OF THE LITERATURE

2.1. Gene Therapy: Definition, Significance, and Challenges

Gene therapy represents a field that has witnessed significant advancements in recent years, offering a promising avenue for the treatment of genetic diseases (Friedmann, 1992). It is an innovative therapeutic approach that targets genetic disorders through the correction, modification, or integration of altered genes into the genetic material. The primary goal of this approach is to rectify the genetic abnormalities at the root of the disease or mitigate their effects, ultimately aiming for the treatment of the condition. Gene therapy holds potential not only in addressing monogenic diseases (those caused by a single gene) but also in the treatment of complex diseases, marking it as a versatile and groundbreaking field in medical research (Gonçalves et al., 2017).

Figure 1 shows a diagram of a cell with a functional gene and a cell with a non-functioning gene. This is a representation of the potential of gene therapy to correct genetic defects.

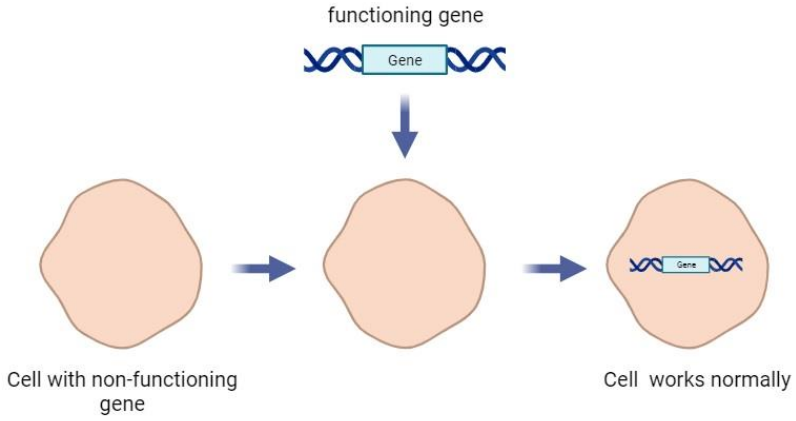


Figure 1. A cell with a mutation can be treated by replacing the mutated/non-functioning gene with a functioning gene (Shahryari et al., 2021)

Although gene therapy holds great potential in the treatment of genetic diseases, it faces some important challenges. The first of these is the difficulty of effectively reaching the cells and tissues targeted by gene therapy. Viral or non-viral carrier systems are generally used for efficient gene transfer within the body. However, studies are continuing on the

safety, effectiveness and specificity of these carrier systems (Katz et al., 2013; Sung et al., 2019)

Another challenge is that gene therapy can be associated with unwanted side effects and immune responses. Genes brought into the body from outside have the potential to be recognized and attacked by the immune system. Therefore, immune system responses must be considered and managed to ensure the safety and efficacy of gene therapy applications (Soofiyani et al., 2013)

Another important challenge of gene therapy is the correct regulation or replacement of the targeted gene. Gene editing is a process that requires accurate targeting and specificity. New gene editing techniques, such as the developed CRISPR-Cas9, offer the potential to effect targeted genetic changes. However, further studies are needed on the efficacy, accuracy and risk of false targeting of these techniques (Frangoul et al., 2021; Zhang et al., 2021)

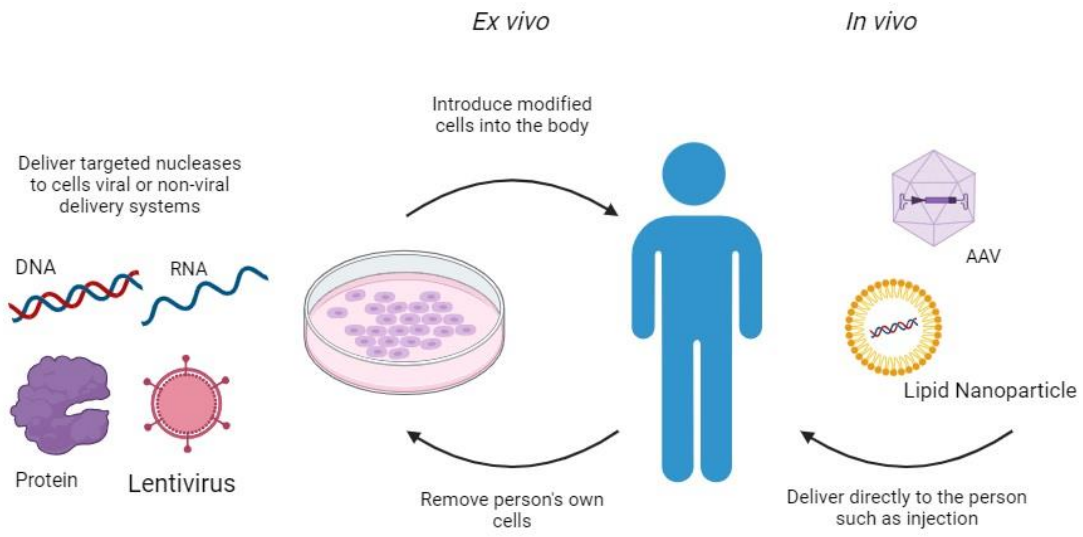


Figure 2. Schema of two distinct technique of gene therapy such as *in vivo* and *ex vivo* (Kumar et al., 2016; Shahryari et al, 2021)

Figure 2 depicts two main approaches to gene therapy: *ex vivo* and *in vivo*. In *in vivo* delivery, genes are directly introduced into the body, often through injection, targeting them to specific tissues or organs. However, during *ex vivo* delivery, cells are first

extracted from the body, modified with the desired genes in a controlled environment, and then reintroduced into the body. The image additionally illustrates the various types of genetic material used in therapy, not just DNA but also RNA and proteins. This highlights the versatility of gene therapy in treating diverse genetic and disease conditions (Kumar et al, 2016).

2.2 Objectives of Gene Therapy

Gene therapy is an approach employed in the treatment of genetic diseases, encompassing a range of objectives. These objectives encompass the correction of genetic abnormalities, regulation of gene expression, silencing or modification of genes (Gonçalves et al, 2017). The fundamental aim of gene therapy is to treat genetic diseases by restoring the function of the targeted gene or enabling its normal expression. Furthermore, it seeks to rectify abnormal phenotypic traits resulting from genetic disorders and alleviate associated symptoms. Gene therapy is regarded as a promising approach in the treatment of hereditary diseases, and various research initiatives and clinical trials centered on gene therapy continue to advance (Kotterman et al., 2014; Duan, 2018)

2.3 The Significance of Intracellular Gene Transfer

Gene therapy represents a promising approach in the treatment of genetic diseases, with the transfer of genes into cells forming the cornerstone of this therapeutic strategy. The transfer of genes into cells provides opportunities for a healthy gene to replace a missing or defective one or to correct unwanted genetic mutations (Bulaklak et al., 2020). This process is facilitated through tools known as gene therapy vectors. Gene therapy vectors are instrumental in ensuring the gene reaches the target cells and aids in its transportation into the cell. These vectors often take the form of modified viruses or structures like plasmid DNA (Thomas et al., 2003). The transfer of genes into cells is an effective strategy that holds the potential to achieve promising results in the treatment of genetic diseases by modifying the genetic material of target cells (Kato et al., 2011). This approach underscores the crucial significance of accurate gene transfer into cells for the successful implementation of gene therapy.

2.4 Gene Transfer Methods: Viral and Non-viral Approaches

Gene delivery systems that used for gene therapy includes different strategies to deliver genes into the target cells. These systems are categorized into two groups which are viral delivery systems and non-viral delivery systems. Viral delivery systems utilize the variety of viruses, but non-viral systems use distinct formulations without any virus (Nayerossadat et al., 2012; Cring et al., 2022)

Viral gene delivery methods use genetically modified viruses to deliver genetic material to target cells. In this method, tools called gene therapy vectors ensure that the gene is delivered to the target cell. Viruses such as retroviruses, adenoviruses, and lentiviruses are viral gene delivery methods commonly used as gene therapy vectors. Viral gene delivery methods generally show high efficiency and can ensure the integration of the gene into desired cells (Slade, 2001). However, these methods also have some disadvantages, for example, immune responses can be induced, and genetic material can be integrated into random sites (Park S. Y. et al., 2019).

Lentiviruses, such as human immunodeficiency virus (HIV), serve as robust tools in gene therapy due to unique attributes. These viruses, belonging to the Retroviridae family, possess the capability to integrate genetic material into the host genome, ensuring stable and prolonged expression of the transgene (Durand et al., 2011). The integration feature is particularly advantageous for applications demanding sustained therapeutic gene expression. Lentiviral vectors demonstrate infectivity in both dividing and non-dividing cells, expanding their utility across diverse cell types and tissues. Their relatively large packaging capacity enables the delivery of substantial transgenes, accommodating complex therapeutic payloads (Milone et al., 2018). Over time, advancements in vector design, including the development of self-inactivating (SIN) vectors, have contributed to the improved safety profile of lentiviral vectors, minimizing the risk of insertional mutagenesis (Arsenijevic et al., 2022).

In the realm of lentiviral research, Jurkat and THP-1 cells stand out as valuable models for studying viral vector-mediated gene transfer. Jurkat cells, derived from human T lymphocytes, and THP-1 cells, a monocytic leukemia cell line, provide unique insights

into the efficiency, tropism, and safety of lentiviral vectors [Larson S, bingham]. Jurkat cells, being a T-cell line, are particularly relevant for immunological research, while THP-1 cells, representing monocytes, contribute to understanding vector interactions within the mononuclear phagocyte system (Stokes et al., 2016; Carrasco-Padilla et al., 2023)

Non-viral gene transfer methods, on the other hand, transport genetic material to target cells using different tools instead of viruses. Examples of non-viral gene transfer methods include strategies such as lipofection, electroporation, microinjection, and the use of ionic polymers Lipofection enables lipid-based carrier molecules to deliver the gene to target cells. Electroporation facilitates the passage of genetic material into cells by temporarily disrupting cell membranes. The microinjection method involves the direct injection of genetic material into the target cell using microinjectors. Ionic polymers, on the other hand, protect and transport genetic material to facilitate gene transfer (Ramamoorth et al., 2015). Safer alternatives to viral gene transfer methods may be considered by non-viral gene transfer, but lower efficiency is often exhibited by them, and challenges may be posed in terms of gene integration into the target cell (Al-Dosari et al., 2009).

In the landscape of gene therapy, the choice between viral and non-viral delivery systems plays a pivotal role in determining the success and applicability of therapeutic interventions. A comparative analysis of these delivery systems reveals distinct features that impact their efficiency, safety, and versatility. The following table encapsulates key attributes, advantages, and limitations, providing a comprehensive overview for informed decision-making in the design and implementation of gene therapy strategies (Al-Dosari et al, 2009; Nayerossadat et al, 2012; Pranatharthiharan et al., 2013).

Table 1. A comparison table for viral and non-viral delivery systems used in gene therapy (Park S. Y. et al, 2019)

Features	Viral Delivery Systems	Non-Viral Delivery Systems
Delivery Efficiency	High	Lower
Targeting Capability	Yes	Generally Weaker
Broad Applicability	Limited (Dependent on cell type)	More Widespread
Immunogenicity	High (Especially in viral vectors)	Low
Gene Size Capacity	Large (Suitable for transporting large genes)	Suitable for Small to Medium-sized Genes
Integration Risk	Yes (Especially in retroviral vectors)	No
Toxicity	Low (However, potential immune responses pose a risk)	Low (Although some lipid-based vectors may exhibit toxicity)
Long-Term Expression	Yes (Permanent expression is possible)	Generally Shorter-Term Expression
Cost	Higher (Production of viral vectors is costly)	Lower

2.5 Advantages and Disadvantages of Gene Delivery Methods

The efficacy and safety of gene therapy hinge significantly on the performance of gene delivery systems. These systems offer distinct advantages, including targeted delivery, prolonged expression, reduced immunogenicity, and versatility. However, they are not without challenges, as they may induce immunogenic responses, exhibit limited payload capacity, pose integration risks, and result in transient expression. Achieving a delicate balance between these advantages and disadvantages is essential for optimizing the outcomes of gene therapy interventions (Gonçalves et al, 2017).

Advantages of gene delivery systems can be explained as follows (Jafarlou et al., 2016):

1.Targeted Delivery: Gene delivery systems allow for the targeted delivery of therapeutic genes to specific cell types, minimizing off-target effects and enhancing treatment precision.

2.Prolonged Expression: Certain gene delivery systems enable sustained and prolonged expression of therapeutic genes, ensuring a more durable therapeutic effect.

3.Reduced Immunogenicity: Advances in gene delivery technologies have led to the development of vectors with reduced immunogenicity, minimizing the risk of immune responses against the delivered genes.

4.Versatility: Gene delivery systems offer versatility by accommodating various types of genetic material, including DNA, RNA, and gene-editing tools such as CRISPR-Cas9.

The drawbacks of gene delivery systems can be delineated as follows (Kohn et al., 2023):

1.Immunogenic Responses: Some gene delivery vectors can elicit immune responses, leading to potential adverse effects and limiting the therapeutic efficacy.

2.Limited Payload Capacity: Certain delivery systems have limited capacity to carry larger therapeutic genes, restricting the range of treatable conditions.

3.Integration Risks: Retroviral vectors, commonly used in gene therapy, may pose a risk of random genomic integration, potentially leading to insertional mutagenesis and oncogenesis.

4.Transient Expression: Some delivery systems result in transient expression of therapeutic genes, necessitating repeated administrations to maintain therapeutic effects.

2.6 Utilization of GFP Plasmids in Gene Therapy

GFP (Green Fluorescent Protein) plasmids are instrumental in gene therapy, providing a non-invasive means to monitor gene expression and cellular behavior in real-time (Ma et al., 2011). Through the integration of GFP into plasmid vectors, researchers gain the ability to visualize transgene expression, enhancing the assessment of gene delivery efficiency and spatial-temporal patterns (Grotjohann et al., 2011). In the context of gene therapy, GFP plasmids often serve as reporter constructs, allowing for the tracking of therapeutic gene expression within target cells or tissues (Soboleski et al., 2005). This real-time monitoring capability contributes to a deeper understanding of transgene dynamics, potential off-target effects, and overall treatment efficacy (Soboleski et al, 2005; Shields et al., 2019). The versatility of GFP as a marker enables its application in

various gene therapy modalities, both viral vector-mediated and non-viral methods, advancing therapeutic strategies (Wahlfors et al., 2001).

The pTRIP-SFFV-GFPEV plasmid is a specialized vector designed for gene expression studies and gene therapy applications. Typically, it includes components like spleen-focus forming virus (SSFV) promoter that controls the manifestation of the Green Fluorescent Protein (GFP) (Burny et al., 1998; Liu E. et al., 2011). This plasmid configuration is tailored for controlled and efficient gene expression, making it suitable for experimental settings to investigate cellular processes, assess transduction efficiency, and optimize gene delivery strategies (Merten et al., 2016). The plasmid design includes elements for stability, propagation, and antibiotic selection in host cells, ensuring reliable and reproducible results in diverse gene therapy experiments (Mairhofer et al., 2008).

2.7 Cell-Penetrating Peptides (CPPs): Key Players in Enhancing Gene Delivery

Cell-penetrating peptides (CPPs) stand as a unique class of short peptides, typically ranging from 5 to 30 amino acids, renowned for their extraordinary ability to traverse cellular membranes and facilitate the intracellular delivery of diverse cargo molecules. Initially observed in viral proteins with membrane-translocating capabilities, CPPs have emerged as pivotal players in biomedicine, holding immense promise for overcoming cellular barriers in therapeutic applications. The diverse sequences of CPPs, derived from various sources, allow for customization to accommodate specific cargo molecules or target cell types (Lee et al., 2019; Gessner et al., 2020).

The intricate mechanisms governing CPP cellular uptake remain a focus of ongoing research, with proposed pathways including endocytosis, direct translocation, and other processes influenced by the physicochemical properties of both the CPP and its cargo (Dixon et al., 2016; Tashima, 2017). Particularly significant in the field of gene therapy, CPPs demonstrate an exceptional ability to efficiently transport genetic material—ranging from DNA to RNA and siRNA—across cellular membranes. This capability is instrumental for introducing therapeutic genes into target cells, offering potential treatments for a diverse spectrum of genetic disorders (Alhakamy et al., 2013; Taylor et al., 2020). Additionally, ongoing investigations aim to modify CPPs to achieve cell-type

specificity, enhancing the precision of cargo delivery and minimizing off-target effects (Martín et al., 2010; Zahid et al., 2015). Despite their promise, challenges such as potential toxicity, immunogenicity, and the need for optimization in specific applications persist, underscoring the ongoing refinement of CPPs in therapeutic contexts (Kristensen et al., 2016).

Figure 3 illustrates the various types of cargo that can form complexes with CPPs for transfer into the intracellular space of cells.

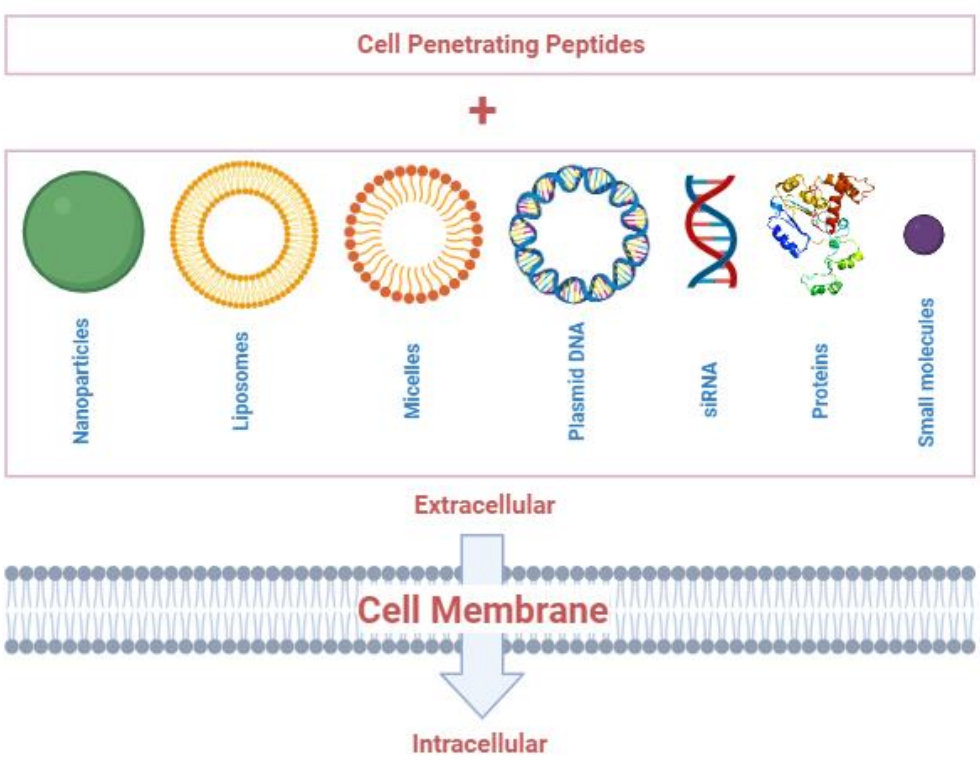


Figure 3. The depiction of the behavior of CPPs in both extracellular and intracellular environments. In the extracellular environment, CPPs can bind to other carrier molecules, such as nanoparticles, liposomes, micelles, plasmid DNA, or siRNA. These carriers help CPPs to pass through the cell membrane (Xu et al., 2019).

Once inside the cell, CPPs can bind to intracellular targets. For example, a CPP can bind to a protein or DNA molecule. This can change or disrupt the function of the protein or DNA molecule (Gautam et al., 2016).

Examples of CPPs Used in Studies:

TAT Peptide: The TAT peptide, derived from the transactivator of transcription protein of the human immunodeficiency virus (HIV), is one of the most well-known CPPs. It has been extensively used for intracellular delivery of various cargoes (Kim A. et al., 2012).

Penetratin: Penetratin is derived from the third helix of the Antennapedia homeodomain protein. It has been employed in studies for the intracellular delivery of therapeutic molecules (Zhu X. et al., 2014).

Arginine-Rich Peptides: Certain short arginine-rich peptides, not necessarily derived from specific proteins, have been investigated as CPPs due to their positive charge and ability to facilitate cellular uptake (Liu B. R. et al., 2015).

Synthetic Peptides: Various synthetic peptides, designed to mimic the properties of natural CPPs, have been developed for enhanced cargo delivery into cells. These may include peptides with diverse sequences optimized for specific applications (Kauffman et al., 2018).

It's important to note that the choice of CPP depends on the cargo, the target cells, and the specific requirements of the study. Researchers continue to explore and design novel CPPs to improve the efficiency and specificity of intracellular delivery in various therapeutic and research contexts.

2.8 Protamine: A Comprehensive Insight into its Composition and CPP Attributes

Protamine, traditionally known for its role in sperm DNA compaction, has been investigated for its potential as a CPP. Most insights into the structural intricacies of protamines and protamine–DNA complexes have been gleaned from the analysis of fish protamines Salmine and Clupeine, along with placental mammal protamines P1 and P2. A typical P1 protamine molecule, consisting of 49 or 50 amino acids, exhibits a tripartite structure: a central DNA-binding domain rich in arginine, flanked by short peptide chains containing cysteine residues. The variability in the number of cysteine residues is evident across species. Central DNA-binding domains encompass anchoring sequences, featuring

3–11 consecutive arginine residues that facilitate peptide-DNA binding. These unique sequences share similarities in size and composition with entire sequences of various fish protamines (Ruseska et al., 2021).

Within the context of medical applications, a crucial consideration revolves around the toxicity associated with protamine. It is essential to highlight that the efficacy, tolerance, and toxicity of protamine are significantly influenced by derivatization. Common modifications involve the creation of sulfate or chloride salts, the reduction of arginine molecules to diminish positive charges (resulting in low-molecular-weight protamine-LMWP), and the attachment of molecules like polyethylene glycol (PEG) (Hofmann et al., 2013; He et al., 2014; González-Aramundiz et al., 2018; Szerlauth et al., 2020). Protamine sulfate, granted approval for medical use in the USA since 1969, boasts a monograph in both the European Pharmacopoeia and the USP. Derived from the sperm of Salmonidae or Culpeidae, protamine sulfate is now also produced through recombinant methods (Pawelczyk et al., 2006). Its primary application in surgery as an antidote against heparin overdoses is supplemented by broader utility. Researchers explore its roles as a CPP and as a constituent in drug delivery systems, encompassing nanoparticles and liposomes (Ranasinghe et al.). In 2001, Byun's research group unveiled LMWP as a peptide fragment resulting from the enzymatic digestion of native protamine sulfate using thermolysin (Byun et al., 1999). This method allows for efficient and rapid production of LMWP, providing the benefits of cost-effectiveness and shortened manufacturing periods. With an extensive portfolio of over 30 published papers delineating the properties and applications of LMWP, researchers proposed diminished toxicity and a reduced immune response when utilizing LMWP as a heparin antidote compared to native protamine. Furthermore, LMWP exhibited notable efficacy as a gene carrier *in vitro* (Clark et al., 2001; Zarei et al., 2021).

Over the years, numerous protamine products have been readily available in the market, solidifying its status as a well-established pharmaceutical ingredient [Panzer, jha k]. To trace its varied applications chronologically, protamine initially found use in the treatment of diabetes mellitus. The combination of protamine and insulin results in an extended insulin effect, leading to reduced blood glucose levels in patients (Renner et al., 1986;

McCall et al., 2013). Subsequently, it was discovered that protamine could neutralize the anticoagulant effects of heparin, positioning it as an antidote in cardiac or vascular surgery to prevent postoperative bleeding events (Ranasinghe et al; Kunz et al., 2018). Among the most notable revelations is the potential use of protamine as a delivery system for biomolecules, such as CPPs for *in vivo* gene transport. Researchers predominantly concentrate on protamine's properties related to cell penetration and nucleus targeting (Weide et al., 2009; Kurrikoff et al., 2016; Yu S. et al., 2021; Timotievich et al., 2023). Additionally, various research groups are exploring the incorporation of protamine into different nanosized formulations to enhance cell penetration (Arbab et al., 2004; Ruseska et al, 2021; Robla et al., 2023). Another field of significant interest, especially in these challenging pandemic times, involves the utilization of protamine in mRNA vaccines (Jarzebska et al., 2021; Wang Y. et al., 2021b).

As a CPP, protamine has been explored for its capacity to efficiently deliver genetic material such as DNA or RNA into target cells. The positive charge of protamine aids in overcoming the negatively charged cell membrane, promoting cellular uptake. Researchers have also modified protamine to enhance its cell-type specificity and optimize cargo delivery, making it a potentially valuable tool in gene delivery applications (Yu X. et al., 2016; Wu et al., 2023).

2.9 A General Overview of Machine Learning

Machine learning is a rapidly developing and advancing research field since 2017. Closely related to disciplines such as artificial intelligence and data analytics, machine learning focuses on the ability of computer systems to learn from data and improve through experience (Choi et al., 2020). Deep learning has made a huge impact in the field of machine learning during this period. The development of algorithms related to artificial neural networks and the advancement of graphics processing unit (GPU) technology have increased and expanded the achievements of deep learning (Wang M. et al., 2021a). Deep learning, with its ability to model complex data structures and detect complex patterns, has revolutionized many application areas such as image processing, natural language processing and automated driving (Khan et al., 2023).

Another important development is the increase in big data sources and the widespread use of cloud computing infrastructure after 2017. Analysis of large-scale datasets has improved the performance and exploration capabilities of machine learning (Gill et al., 2008; Saczynski et al., 2013). Also, advances in cloud computing platforms have enabled rapid deployment and scaling of machine learning algorithms (Sarkar et al., 2023). The combination of big data and cloud computing technologies has allowed machine learning to be used in real-time and scalable applications. For example, applications such as providing personalized recommendations in online commerce or assisting in disease diagnosis in the health sector have become possible with these developments.

Finally, since 2017, it seems that ethical and safety issues in machine learning have become more important. Ethical issues related to the use of machine learning models, algorithmic fairness, data privacy and bias are among the focus areas of researchers and industry (Lo Piano, 2020). In particular, discussions on the transparency, reliability and explainability of the decision-making processes of deep learning-based systems are increasing (Belle et al., 2021). In this context, the social, legal and ethical aspects of machine learning require researchers to take a broader perspective.

Machine learning is a powerful tool that can be used to improve the efficiency and accuracy of a wide range of tasks and it also have many advantages (Sun, 2021). Figure 4 demonstrates six advantages of machine learning such as less time consumption and continuous improvement.

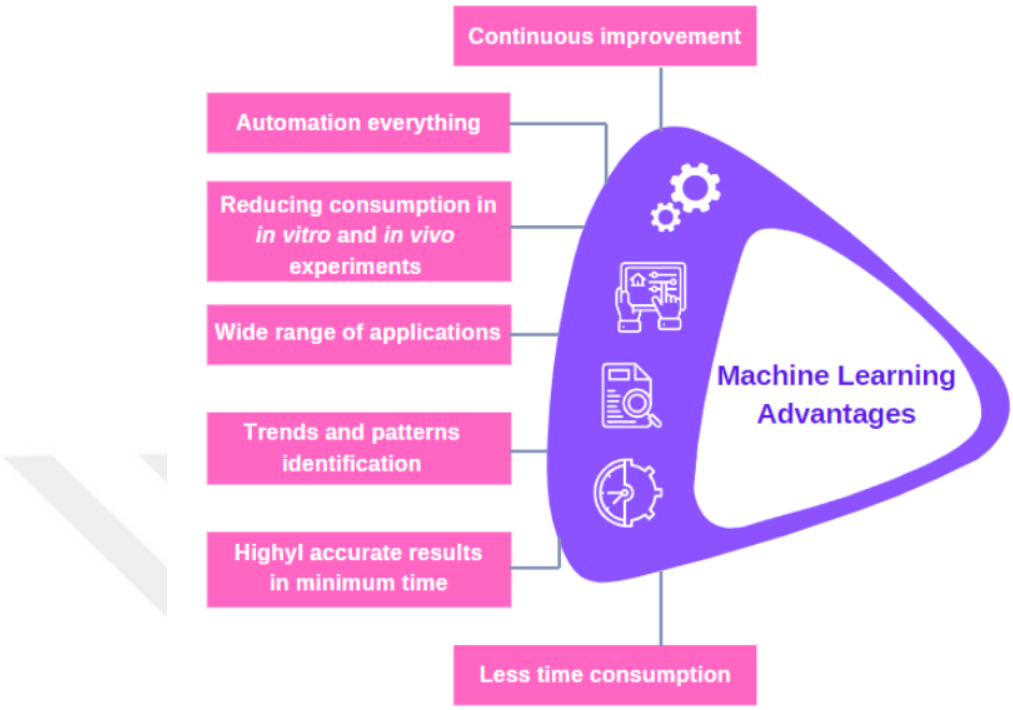


Figure 4. The diagram of the benefits of machine learning. The seven advantages of machine learning are summarized (Sun, 2021).

2.10 Steps in Machine Learning

While developing a machine learning model, the essential steps can be examined through seven key steps as follows (Luo et al., 2016; Badillo et al., 2020):

1. Collecting Data: The initial step in the machine learning journey involves acquiring relevant and diverse datasets. Data collection sets the foundation for subsequent stages, influencing the model's performance and generalization.

2. Preparing the Data: Data preparation involves cleaning, transforming, and organizing the collected data to make it suitable for model training. This step ensures that the input data is in a format that allows the model to learn effectively.

3. Choosing a Model: Selecting an appropriate machine learning model is a critical decision that depends on the nature of the problem and the characteristics of the data. The

choice may involve considerations such as supervised or unsupervised learning, regression or classification, and the model's complexity.

4. Training the Model: Model training is the phase where the selected algorithm learns patterns and relationships from the prepared data. This step involves adjusting the model's parameters to minimize the difference between predicted and actual outcomes.

5. Evaluating the Model: To assess the model's performance and generalization capabilities, rigorous evaluation metrics are employed. Common metrics include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve.

6. Parameter Tuning: Fine-tuning the model's hyperparameters is essential to optimize its performance further. This iterative process involves adjusting parameters to achieve the best possible results and avoid overfitting or underfitting.

7. Predictions: Once the model has been trained, evaluated, and fine-tuned, it is ready to make predictions on new, unseen data. This final step demonstrates the practical application of the machine learning model in real-world scenarios.

In Figure 5, seven key steps of machine learning are summarized. The steps of machine learning, depicted in a detailed diagram, commence with ‘Collecting Data’ and culminate in ‘Predictions’.

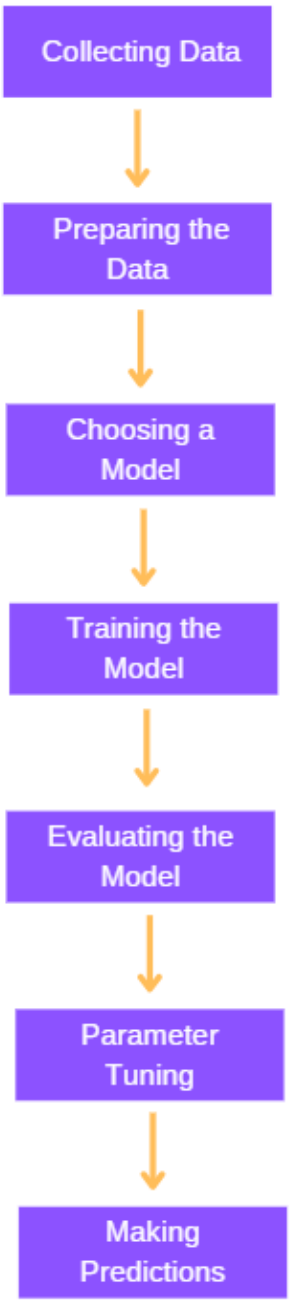


Figure 5. The diagram illustrating the seven stages of the machine learning process (Luo et al, 2016; Badillo et al, 2020).

2.11 Understanding Random Forest and Alternative Machine Learning Models

Machine learning covers a wide field with various models and algorithms. First, supervised learning models can be considered. These are trained with the labeled dataset

and then used to classify or predict new data. For example, Support Vector Machines are a classification model that optimizes data to best distinguish between hyperplanes (Kremer et al., 2014). In addition to this decision trees are also a widely used model. It represents the data with a tree structure and makes classification or estimation by dividing according to a feature at each node (Rigatti, 2017).

The Random Forest model is a widely used method in the field of machine learning and is basically based on decision trees. Random Forest combines multiple decision trees to create a more powerful model. This model trains each tree using random sample selections and random feature selections (Rigatti, 2017; Hu et al., 2023). In the estimation phase, each tree makes separate estimates, and then the average or majority of these estimates are taken to form the final estimate. While the Random Forest model is known for its ability to reduce the risk of data overfitting, it offers advantages such as high accuracy and feature importance ranking. Thanks to these features, successful results have been obtained in many application areas such as classification, regression and feature selection. (Rigatti, 2017; Nguyen et al., 2021).

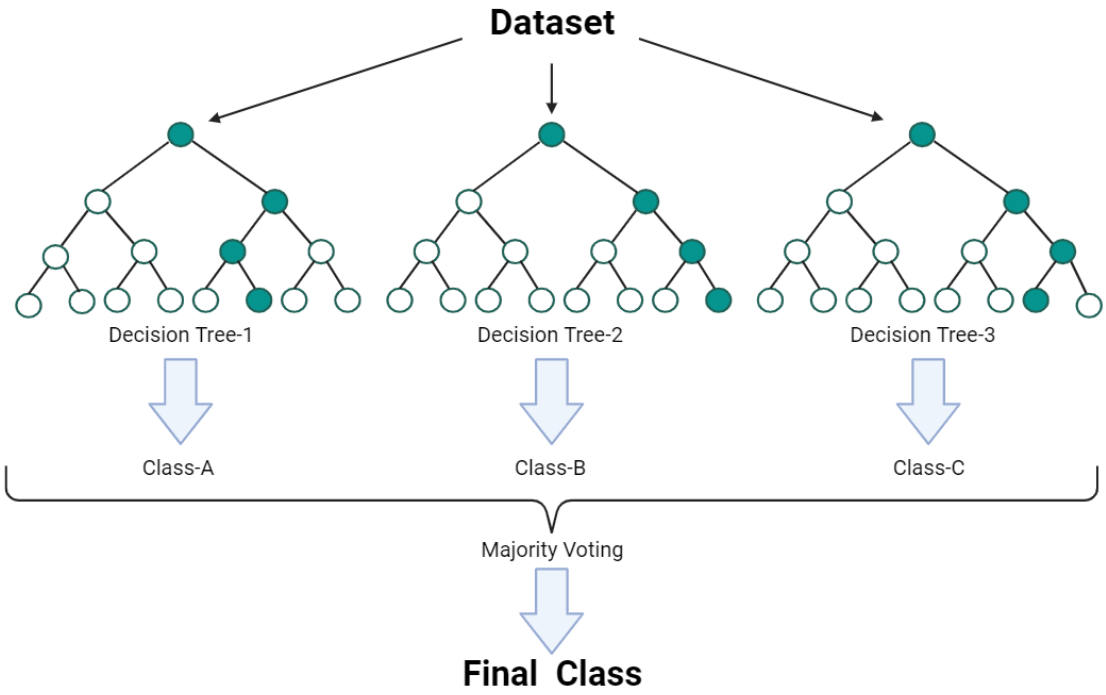


Figure 6. Demonstration of random forest model workflow (Yang et al., 2019).

The Random Forest model is widely preferred and successfully employed in genetic studies across various application domains such as genetic variation analysis, biomedical data integration (Goldstein et al., 2011).

The Random Forest model serves as an effective tool for examining genetic variations. For instance, Riley-Gills and colleuages aim to employ a Random Forest approach to identify and characterize modifiers of the immune tumor microenvironment within loci from genome-wide association studies for breast cancer incidence, revealing potential therapeutic targets and providing a robust framework for functional fine-mapping of risk loci. (riley-gills). Similarly, Chen and colleagues (2021) adopted the Random Forest to find multiple inherited diseases (Chen N. et al., 2021).

The Random Forest model is also used to integrate different types of biomedical data and define subtypes of complex diseases. For example, Chen and their team (2018) combined genetic data with clinical data to distinguish different subtypes of Alzheimer's disease using the Random Forest model (Chen L. et al., 2015). Likewise, Quist and associates (2021) study to create a computational framework that could integrate a variety of data types into a meaningful classification in cancer by using Random Forest (Quist et al., 2021).

These examples underscore the versatility and efficacy of the Random Forest model in genetic research. The model provides researchers with a valuable tool for tasks such as genetic data analysis, classification, and enhancing our understanding of diseases (Goldstein et al, 2011; Chen N. et al, 2021; Quist et al, 2021).

2.12 The Definitions and Importance of Features in Machine Learning Model

Machine learning uses different features that related to many properties of data and can be important for the prediction. To get accured predictions, features should be chosen logically. To make a prediction about any amino acid sequence is CPP or non-CPP, distinct features can be used. Some properties that may be important for proteins are lenght, amino acid composition, entropy and CTDC (Composition/Transition/Distribution – Composition). CTDC consists of variety of features like hydrophobicty, charge,

secondary structure, normalized Van der Walls, solvent accessibility, polarity and polarizability which can be all found in AAindex1 (Gu et al., 2020; Kabir et al., 2022).

Protein length is a fundamental feature of proteins that can be measured in a variety of ways. These different measures can be used to provide insights into protein function, structure, and evolution. Length of protein is one of the basic pieces of information about proteins. Number of amino acids is determined as length. Length can give some ideas about protein. To explain, different features of protein length involve measuring protein length in different ways. For example, minimum protein length is the shortest amino acid sequence in a protein. Maximum protein length is the longest amino acid sequence in a protein. Average protein length is the average of all amino acid sequences in a protein. These features can be used to classify proteins, predict protein function, and determine the evolutionary relationships of proteins (Chang et al., 2006).

Amino acid composition is the measurement of the amount and ratio of amino acids in a protein. It is one of the most basic features used to define a protein. Amino acid composition can provide important information about the function and structure of a protein. For example, some amino acids are associated with specific protein functions. For example, glutamic acid and aspartic acid are found in acidic regions of proteins and help in the catalysis of enzymes. Different features of amino acid composition involve measuring the amount and ratio of amino acids in different ways. For example, amino acid content is the percentage of the amount for each amino acid type in a protein. Amino acid majority is the most abundant amino acid in a protein. Amino acid diversity is the number of different amino acid types in a protein. Amino acid composition features can be used to classify proteins, predict protein function, and determine the evolutionary relationships of proteins (Roy et al., 2009; Saravanan et al., 2015; Lv et al., 2019).

Entropy is a statistical measure that quantifies the disorder or unpredictability of a system. In proteins, entropy can be used to measure the disorder of the distribution of amino acids in a protein. Entropy feature is calculated from the probability distribution of amino acids in a protein. This distribution can be represented as a list of probabilities for each amino acid type in a protein. Entropy feature can provide information about the function and structure of proteins. For example, if the entropy of a protein is high, this suggests that the

protein may have a more disordered structure. This could mean that the protein is more flexible or less active. Entropy feature can be used to classify proteins, predict protein function, and determine the evolutionary relationships of proteins (Pritišanac et al., 2019; Anashkina et al., 2021).

CTDC categorized amino acids into different groups according to their physicochemical properties by using AAindex1. AAindex serves as a comprehensive database comprising numerical indices that represent diverse physicochemical and biochemical attributes of amino acids and amino acid pairs. The recent expansion of AAindex introduces a new section featuring protein contact potentials. The database now encompasses three sections: AAindex1 for amino acid indices, AAindex2 for amino acid substitution matrices, and AAindex3 for statistical protein contact potentials (Dubchak et al., 1995; Kawashima et al., 2008).

Protein structures and functions are intricately shaped by the physicochemical and biochemical properties of the 20 naturally occurring amino acids constituting proteins. The investigation of these amino acid properties, represented by sets of 20 numerical values known as amino acid indices, has been a focal point in numerous experiments and theoretical studies. Initially compiled by Nakai et al. and expanded by Tomii and Kanehisa, AAindex is a continuously updated database that has grown to encompass 544 amino acid indices (Nakai et al., 1988; Kawashima et al, 2008).

The utility of AAindex extends to diverse bioinformatics research areas, such as predicting protein subcellular localization, assessing immunogenicity of MHC class I binding peptides, analyzing protein SUMO modification sites, and examining coordinated substitutions in multiple alignments of protein sequences. Moreover, AAindex has been utilized in the development of derivative databases and web tools for visualizing relationships among its entries. In Figure 7, application areas of AAindex is demonstrated (Liu B. et al., 2007; Tantoso et al., 2008).

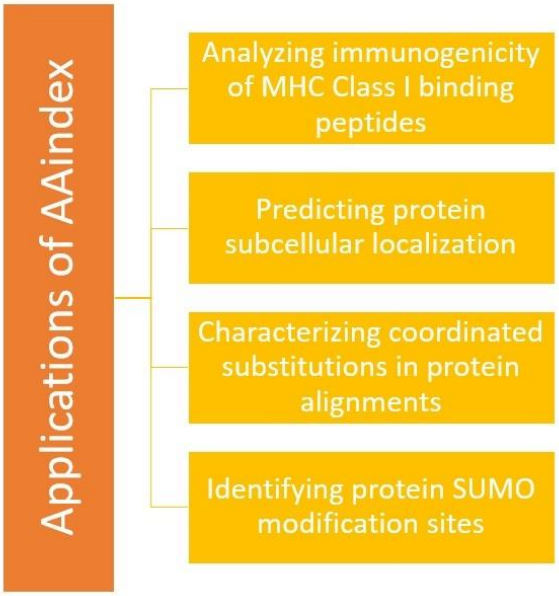


Figure 7. The four distinct applications of the AAIndex. The AAindex has been widely utilized in various bioinformatics studies. The availability and organization of this data have significantly contributed to our understanding of protein structure, function, and evolution (Liu B. et al, 2007; Kawashima et al, 2008; Tantoso et al, 2008).

In 2005, Pokarowski et al. compared various protein pairwise contact potentials, revealing connections to amino acid substitution matrices. Building on this, the present work introduces AAindex3, a new section compiling data on statistical protein contact potentials. This addition enhances AAindex's role as a valuable resource in bioinformatics, facilitating a better understanding of amino acid exchangeability and residue–residue interactions in proteins (Kawashima et al, 2008).

The AAindex database, currently at version 9.0, is released annually and comprises three sections: AAindex1, AAindex2, and AAindex3. AAindex1 contains 544 amino acid indices with detailed information, including accession numbers, descriptions, reference details, and numerical values for the properties of the 20 amino acids. Cross-links with correlation coefficients of 0.8 or higher connect entries with similar properties (Kawashima et al, 2008).

To visualize relationships among amino acid indices, a minimum spanning tree has been constructed, reflecting Tomii et al.'s classification. AAindex2 features amino acid substitution matrices, while AAindex3, the latest addition, includes statistical protein contact potentials. The continuous updates and expansions of AAindex solidify its position as a valuable tool in the bioinformatics study of proteins (Tomii et al., 1996)

All data are sourced from published literature and are accessible through the GenomeNet's DBGET/LinkDB system (Kawashima et al., 2000).

The Protlearn library is a Python library that contains information about the structure and feature of proteins. One of these functions is CTDC, which is a feature that represents the amino acid composition of proteins and the transitions in this composition. CTDC dataset divides amino acids into 3 groups (Group1, Group2, Group3) based on their physicochemical properties like solvent accessibility, hydrophobicity, charge, normalized Van der Walls, polarity, secondary structure and polarizibility. Hydrophobicity has 7 subgroups according to different studies. In table 2, the main categories and groups are given that are utilized in CTDC feature (Dubchak et al, 1995; Dubchak et al., 1999).

Table 2. The categorization of amino acids into three groups including Group1, Group 2, Group 3. In this table, members of the each group are given according to different properties that are hydrophobicity, normalized Van der Waals, polarity and polarizability (Dubchak et al, 1999)

Property	Group 1	Group 2	Group 3
Hydrophobicity	Polar R, K, E, D, Q, N	Neutral G, A, S, T, P, H, Y	Hydrophobic C, V, L, I, M, F, W
Normalized Van der Waals volume	0 - 2.78 G, A, S, C, T, P, D	2.95 - 4.0 N, V, E, Q, I, L	4.43 - 8.08 M, H, K, F, R, Y, W
Polarizability	0 - 0.108 G, A, S, D, T	0.128 - 0.186 C, P, N, V, E, Q, I, L	0.219 - 0.409 K, M, H, F, R, Y, W
Polarity	4.9 - 6.2 L, I, F, W, C, M, V, Y	8.0 - 9.2 P, A, T, G, S	10.4 - 13.0 H, Q, R, K, N, E, D

Furthermore, hydrophobicity is divided into seven groups according to distinct studies and Table 3 is obtained. These seven hydrophobicity groups can be found in AAindex-1.

Table 3. CTDC categorization chart. Hydrophobicity is divided into seven subgroups according to different researches. Since each of the 13 properties has 3 groups, the total dimension of this descriptor is 39 (https://www.genome.jp/aaindex/)

Category	Group 1	Group 2	Group 3
Hydrophobicity_ARGP820101	QSTNGDE	RAHCKMV	LYPFIW
Hydrophobicity_CASG920101	KDEQPSRNTG	AHYMLV	FIWC
Hydrophobicity_ENGD860101	RDKENQHYP	SGTAW	CVLIMF
Hydrophobicity_FASG890101	KERSQD	NTPG	AYHWVMFLIC
Hydrophobicity_PONP930101	KPDESNQT	GRHA	YMFWLCVI
Hydrophobicty_PRAM900101	RKEDQN	GASTPHY	CLVIMFW
Hydrophobicity_ZIMJ680101	QNGSWTDERA	HMCKV	LPFYI
Normalized Van der Waals volume	GASTPDC	NVEQIL	MHKFRYW
Polarity	LIFWCMVY	PATGS	HQRKNED
Polarizability	GASDT	CPNVEQIL	KMHFRYW
Charge	KR	ANCQGHILMFPSTWYV	DE
Secondary structure	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	ALFCGIVW	RKQEND	MSPTHY

Hydrophobicity sub-groups can be accessed from the AAindex-1 database. Upon logging in to the AAindex-1 website (https://www.genome.jp/aaindex/), one of the group names from the table is entered into the search bar. The results are then displayed. A table with numerical hydrophobicity values for each amino acid, as determined by researchers, will be shown by the database. The results for each group are based on the research conducted by different researchers. If a visualization of the data is desired, Excel or programming languages like Python can be used to generate bar charts, heatmaps, or other visualizations that effectively represent these hydrophobicity values. For instance, the PRAM900101 entry in the AAindex-1 database was obtained by Prabhakaran, M. in 1990. This study highlights two key aspects of signal peptides: their crucial hydrophobic core and potential recognition features. The hydrophobic core, composed mainly of leucine, valine, phenylalanine, and alanine, is essential for membrane interaction and translocation. Additionally, signal peptides exhibit intriguing recognition features, such as a preferred basic amino acid at a specific position, equal affinity for α-helical and β-sheet conformations within the hydrophobic regions, and distinct cleavage site characteristics based on size, mass, and acidic charge preference. These findings significantly advance our understanding of signal and nascent peptide properties and their intricate arrangements. This knowledge can illuminate protein translocation mechanisms and potentially guide the design of powerful signal peptides for protein engineering applications (Prabhakaran, 1990).

For each physicochemical property, the frequency of each class is calculated across all sequences in the dataset by using an equation (Dubchak et al, 1995; Dubchak et al, 1999)

$$c(i) = \frac{n_i}{N} \; i = 1,2,3$$

n_i shows the frequency of group and *N* denotes the length of sequence. For example, the frequency of “ARKLY” sequence can be calculated as following steps:

Group the protein's amino acids. In this case, since our protein contains "ARKLY", the amino acids are grouped as follows:

Group 1: Alanine (A)

Group 2: Arginine (R)

Group 3: Lysine (K)

Group 2: Leucine (L)

Group 2: Tyrosine (Y)

Calculate the number of amino acids in each group. In this case, the number of amino acids in each group is as follows:

Group 1: 1

Group 2: 2

Group 3: 1

Divide the number of amino acids in each group by the total number of amino acids in the protein. In this case, the frequencies for each group are as follows:

Group 1: 1/5 = 0.2

Group 2: 2/5 = 0.4

Group 3: 1/5 = 0.2

Therefore, the frequencies for "ARKLY" are as follows:

H frequency: 0.2

C frequency: 0.4

P frequency: 0.2

These frequencies provide a representation of the protein's amino acid composition. For example, these values indicate that the protein's hydrophilic amino acid content is 40%.

The frequencies for the example shown in the figures are calculated as follows:

H frequency: Since there is 1 hydrophilic amino acid (A) in the "ARKLY" protein, the H frequency is 1/5 = 0.2.

C frequency: Since there are 2 hydrophobic amino acids (R, L, Y) in the "ARKLY" protein, the C frequency is 2/5 = 0.4.

P frequency: Since there is 1 polar amino acid (K) in the "ARKLY" protein, the P frequency is 1/5 = 0.2.

2.13 Understanding Overfitting, Feature Importance, and Underfitting in Machine Learning

While developing a machine learning model, one may encounter a phenomenon known as overfitting, where the model learns the training data too closely, capturing noise or idiosyncrasies as if they were genuine patterns (Ying, 2019). In contrast to overfitting, underfitting occurs when a model is too simple to capture the underlying patterns in the data. As a result, it performs poorly on both the training set and new data. Underfit models lack the capacity to grasp the complexities present in the dataset. Addressing underfitting involves increasing the model's complexity, incorporating more relevant features, or using more advanced algorithms (Khalaf Jabbar Rafiqul Zaman Khan, 2015). Ensuring the model has the capability to capture the intricacies of the data is crucial to overcoming underfitting. Several techniques can be employed to address overfitting, and one such approach is the consideration of feature importance. Feature importance serves as a pivotal strategy in mitigating overfitting by assessing the significance of each feature or variable in the dataset. By understanding the relative importance of features, one can make informed decisions about feature selection, leading to more robust and generalizable machine learning models (Altmann et al., 2010; Huda et al., 2019).

Overfitting is a common challenge in machine learning models that occurs when a model learns the training data too well, capturing noise or random fluctuations in the data as if they were genuine patterns. In other words, the model becomes too complex and starts fitting not only the underlying patterns but also the noise inherent in the training dataset. This can lead to poor generalization performance, where the model performs well on the

training data but fails to generalize effectively to new, unseen data. Overfitting can be mitigated through techniques such as regularization, cross-validation, and using larger and more diverse datasets (Ying, 2019).

Feature importance refers to the measure of the contribution of each feature (or variable) in a dataset to the predictive performance of a machine learning model. Understanding feature importance is crucial for identifying which features have the most significant impact on the model's predictions. Various algorithms and techniques can be employed to assess feature importance, such as decision trees, random forests, and permutation importance. Analyzing feature importance aids in feature selection, model interpretation, and the identification of key variables driving the model's predictions. This is essential for creating more interpretable and efficient models, especially in complex datasets with numerous variables (Altmann et al, 2010; Huda et al, 2019).

2.14 Exploring Key Evaluation Metrics in Machine Learning: Accuracy, Recall, and F1 Score

There are some metrics which are used in model evaluation step while developing a machine learning algorithm. In the realm of machine learning model evaluation, three key metrics—F1 score, accuracy, and recall—are commonly employed to assess the performance of classification models (Handelman et al., 2019).

The F1 score is a metric that combines precision and recall into a single value, offering a balanced measure of a model's performance. Precision represents the accuracy of positive predictions, while recall gauges the model's ability to capture all relevant instances of a particular class. The F1 score is particularly useful in scenarios where there is an imbalance between the number of positive and negative instances, providing a comprehensive evaluation of a model's effectiveness (Handelman et al, 2019).

Accuracy is a fundamental metric that measures the overall correctness of a model's predictions by considering both true positives and true negatives. While accuracy provides a straightforward measure of a model's performance, it may not be suitable for imbalanced datasets where the distribution of classes is uneven. In such cases, accuracy might be

misleading, and complementary metrics like precision, recall, and the F1 score become crucial for a more nuanced assessment (M et al., 2015; Handelman et al, 2019).

Recall, also known as sensitivity or true positive rate, quantifies the ability of a model to correctly identify all relevant instances of a particular class. It is especially crucial in scenarios where missing positive instances can have significant consequences, such as in medical diagnoses or fraud detection. Recall is calculated by dividing the number of true positives by the sum of true positives and false negatives (Vujović, 2021).

The selection of appropriate evaluation metrics depends on the specific characteristics of the dataset and the objectives of the machine learning task. Understanding the nuances of F1 score, accuracy, and recall allows for a more comprehensive evaluation of a model's performance in diverse contexts (Zhou et al., 2021).

2.15 The Use of Machine Learning in Genetic Studies

Following 2017, the realm of machine learning has gained significant prominence within the field of genetic studies. The application of machine learning techniques has emerged as a potent instrument for the examination and elucidation of genetic information (Shen et al., 2022; Roman-Naranjo et al., 2023). To illustrate, deep learning methodologies have been harnessed to scrutinize genomic data, yielding impressive outcomes in realms including the discernment of genetic anomalies, prognostication of disease susceptibility, and advancement of pharmaceutical exploration (Libbrecht et al., 2015; Alharbi et al., 2023). Furthermore, the domain of machine learning has contributed to the comprehension of biological processes, such as gene expression, protein interaction networks, and gene regulation, by simplifying the intricacies inherent in the analysis of genetic data (Balogh et al., 2022; Ghazal et al., 2022).

Machine learning techniques have been prominently employed within the domain of genetic studies, deploying a diverse array of algorithms and methods. For instance, gene expression data is subjected to rigorous analysis via classification algorithms, while the intricate task of identifying gene variations finds its solution in the application of sophisticated deep learning models. Simultaneously, the intricate landscape of gene

regulation is navigated through the strategic implementation of clustering techniques (Zrimec et al., 2020; Alharbi et al, 2023).

Beyond these specific applications, the imprint of machine learning in genetics extends to encompass a wider spectrum of activities. Techniques geared towards data size reduction assume significance, alongside endeavors aimed at unraveling the underlying structural architectures of genetic networks. Additionally, the integration of disparate genetic data sources garners attention as an enabler for more efficient data processing and robust modeling (Jayanthi et al., 2018; Alley et al., 2020).

The cumulative impact of these research endeavors is profound. They facilitate the revelation of pivotal insights within genetic research, nurturing advancements in the realm of disease diagnosis and treatment. Furthermore, these efforts lay the foundation for the evolution of personalized medicine approaches, ushering in a new era of tailored healthcare interventions.

2.16 Insights into CPP: A Perspective from Machine Learning Studies

In recent years, the application of machine learning techniques has significantly contributed to the comprehensive understanding and utilization of CPPs. Researchers have employed various machine learning algorithms to discern the intricate patterns governing the efficiency and specificity of CPPs in cellular uptake. For instance, predictive models leveraging support vector machines and neural networks have been developed to elucidate the sequence-structure relationships of CPPs, unveiling key features that influence their ability to penetrate cell membranes (A. Dobchev et al., 2010; Wei et al., 2017a). These models not only aid in predicting novel CPPs but also offer insights into the underlying physicochemical properties governing their functionality.

Moreover, machine learning approaches have been instrumental in optimizing CPPs for targeted delivery. By integrating experimental data and computational predictions, researchers have fine-tuned CPP sequences to enhance cell-type specificity, thereby minimizing off-target effects and improving the precision of cargo delivery (Porosk et al., 2021). The synergy between experimental data and machine learning algorithms provides

a powerful framework for accelerating the design and development of CPPs with tailored functionalities.

Despite these strides, challenges persist, including the need for larger and more diverse datasets, as well as the consideration of dynamic cellular environments. Ongoing research endeavors are leveraging machine learning to address these challenges, paving the way for further advancements in the field of CPPs and their applications in drug delivery, gene therapy, and beyond (De Jong et al., 2020; Park H. et al., 2023).

3. MATERIALS AND METHODS

The methodology employed in this study encompasses a comprehensive exploration of machine learning steps, where approximately 1400 distinct peptide sequences undergo a meticulous analysis based on a myriad of diverse features. As illustrated in the accompanying Figure 8, the method unfolds through a series of structured steps, each designed to extract essential insights and patterns from peptide sequences. The investigated features span a spectrum, including but not limited to length, entropy, amino acid composition, and others, contributing to a nuanced understanding of the inherent characteristics of each peptide.

The central objective of the method is to develop an algorithm capable of predicting whether a given peptide sequence functions as a CPP or falls into the non-CPP category. To achieve this, the study systematically applies machine learning techniques, involving data collection, preprocessing, model selection, and evaluation. The interplay of these steps is pivotal in the creation of a robust algorithm that can discern CPPs from non-CPPs with a high degree of accuracy.

This study adds value to the field by not just looking at a large set of peptide sequences but also by choosing and combining important features to build the model. The strength of our approach lies in its capability to notice small differences in the peptide sequences. This ability includes creating a program that goes beyond sorting and provides detailed information about the specific traits of each peptide.

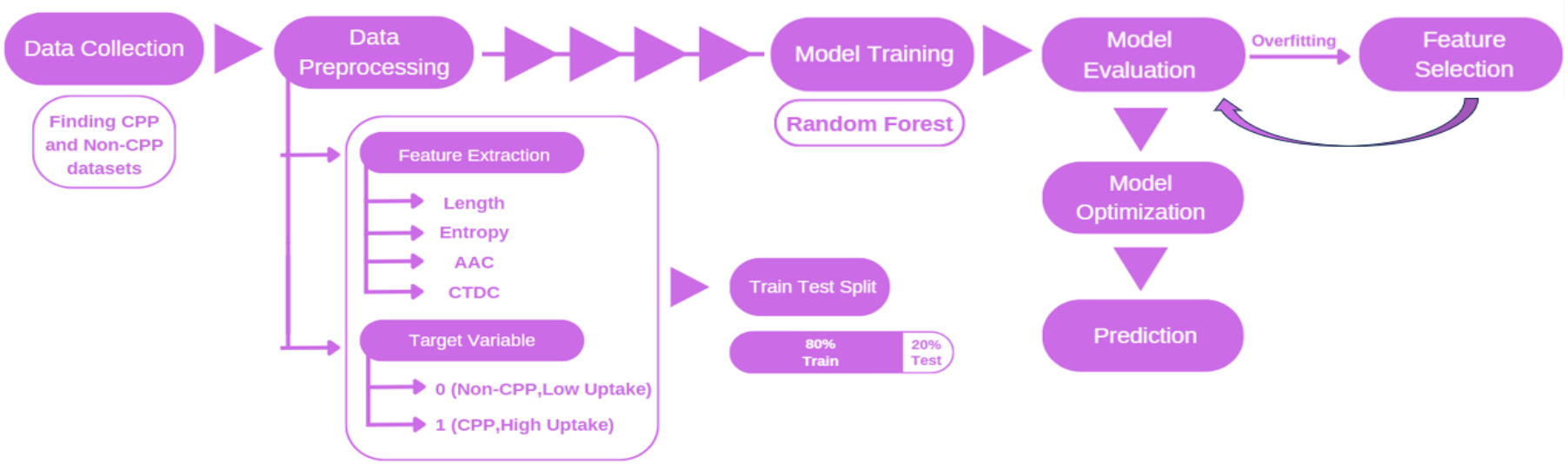


Figure 8. The workflow of the machine learning aspect in this study. The workflow begins with the data collection step on the left and ends with the prediction step, continuing in the direction of the arrows. The figure, which summarizes the steps taken in the study, shows that after model evaluation was performed due to overfitting, the feature importance technique was used and then the model evaluation was performed again.

3.1 Machine Learning Steps

3.1.1. Data Collection

In this study, the data collection process was initiated to develop an algorithm for predicting whether a given peptide is a Cell Penetrating Peptide (CPP) or not, using a Random Forest model. The first step involved assembling a comprehensive dataset that would serve as the foundation for training and testing the algorithm.

To construct this dataset, two primary sources of data were utilized. The first source consisted of peptides with a known status as either CPPs or non-CPPs. The second data source included peptides categorized by their uptake efficiency, specifically, high uptake efficiency and low uptake efficiency (Gautam et al., 2013). These peptides were classified as high uptake efficiency CPPs and low uptake efficiency CPPs. By merging this data source with the previously created CPP and non-CPP dataset, a unified dataset was formed that contains approximately 1400 rows.

This meticulous data collection process ensured that the algorithm had access to a diverse and well-structured dataset, comprising both CPP and non-CPP peptides, as well as information on uptake efficiency. This comprehensive dataset was essential for training and evaluating the Random Forest model, which played a pivotal role in the development of the predictive algorithm.

3.1.2 Data Preprocessing

Upon merging the CPP and non-CPP data sources with the high uptake and low uptake CPP data sources, a unified dataset was obtained. In order to prepare this dataset for machine learning analysis, it was imperative to perform necessary data preprocessing steps. This involved encoding the characteristics of each peptide into a format suitable for machine learning algorithms.

In this process, peptides possessing the combined attributes of being CPPs and exhibiting high uptake efficiency were labeled as "1," while non-CPP peptides with low uptake efficiency were labeled as "0." Additionally, for each peptide sequence, a set of relevant features was calculated and integrated into the dataset to form the input table for the

machine learning model. In table 4, a dataset that was obtained after data preprocessing is given.

Table 4. A table showing a portion of the dataset used in this study. Data including different sequences which are a variety of CPPs and non-CPPs. After data preprocessing procedure, CPPs are converted to "1" and Non–CPPs are "0". This table is obtained by using an algorithm during this study.

	Sequence	Result
0	AAAAARRRIRKQAHAHSK	1
1	AAVACRICMRNFSTRQARRNHRRRHRR	1
2	SGIEEELQKRLEF	0
3	VNMSSFFLMLVYYVTY	0
4	LQDISRAGKNVFRNV	0
...	...	...
1411	AGYLLGKINLKALAALAKKILTYADFIASGRTGRRNAI	1
1412	AGYLLGKLKALAALAKKIL	1
1413	LQPARTLAKKRVVVKRPDYAPPLAGLTT	0
1414	IEKEGFLDFDVAIATPDMMRIIGRLGKILGPRG	0
1415	IRMHN	0

1416 rows × 2 columns

To derive these informative features, the Protlearn library, a Python library specialized in extracting sequence-based features, was employed. For each peptide sequence, a set of distinctive features was computed using the Protlearn library, contributing to the construction of an input table specifically tailored for machine learning applications. Features that were used in this study can be listed as:

1. Length: The length of a protein sequence determines the number of amino acids in a protein sequence. It can be used to distinguish between different types of proteins, such as enzymes and structural proteins

2. Amino Acid Composition: The amino acid composition of a protein sequence is a more detailed representation of the sequence. By using this feature, the freaquency of each amino acid in a sequence can be understood. It can be used to identify the functional properties of a protein, such as its solubility or its binding affinity.

3. Entropy: Entropy is a measure of the disorder in a system. The entropy of a protein sequence can be used to measure the complexity of the sequence.

4. CTDC (Compositon/Transition/Distribution - Composition)

Amino acids are divided into three groups in the CTDC according to their attributes. Table 5 provides a comprehensive overview of the general properties used in CTDC and the properties of the amino acids in the groups.

This meticulous data preprocessing stage ensured that the dataset was transformed into a suitable format for training and evaluating the Random Forest model, enabling the algorithm to effectively learn and make predictions based on the given peptide sequences and their associated features.

Table 5 presents an overview of selected physicochemical properties utilized as inputs for the machine learning algorithm. These properties, such as sequence length, entropy, and hydrophobicity, are extracted for each peptide sequence and fed into the model for training. The algorithm is expected to learn the underlying relationships between these features and the classification of a sequence as either a CPP or non-CPP.

Table 5. List of features used as input in this study. The table shows the physicochemical properties of 1416 peptide sequences. The properties are listed in the 61 columns, and each sequence is represented by a row. It is a valuable resource for investigating the relationship between physicochemical properties and CPP activity

	Length	Entropy	Hydrophobicity_ARGP820101-G1	Hydrophobicity_ARGP820101-G2	Hydrophobicity_ARGP820101-G3	Hydrophobicity_CASG920101-G1	Hydrophobicity_CASG920101-G2	Hydrophobicity_CASG920101-G3	Hydrophobicity_ENGD860101-G1	Hydrophobicity_ENGD860101-G2	...	L	K	M	F	P	S	T	W
0	18.0	2.411509	0.111111	0.833333	0.055556	0.444444	0.500000	0.055556	0.500000	0.444444	...	0.000000	0.000000	0.000000	0.055556	0.222222	0.055556	0.000000	0.000000
1	27.0	3.006025	0.185185	0.740741	0.074074	0.555556	0.296296	0.148148	0.555556	0.222222	...	0.037037	0.074074	0.000000	0.037037	0.370370	0.037037	0.037037	0.037037
2	13.0	2.931209	0.538462	0.153846	0.307692	0.692308	0.153846	0.153846	0.538462	0.153846	...	0.000000	0.000000	0.000000	0.076923	0.076923	0.076923	0.000000	0.000000
3	16.0	2.905639	0.250000	0.312500	0.437500	0.250000	0.625000	0.125000	0.250000	0.187500	...	0.125000	0.062500	0.000000	0.000000	0.000000	0.125000	0.062500	0.187500
4	15.0	3.506891	0.400000	0.400000	0.200000	0.600000	0.266667	0.133333	0.466667	0.200000	...	0.000000	0.133333	0.000000	0.066667	0.133333	0.066667	0.000000	0.133333
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1411	38.0	3.293594	0.263158	0.394737	0.342105	0.447368	0.421053	0.131579	0.315789	0.394737	...	0.000000	0.052632	0.000000	0.000000	0.078947	0.026316	0.052632	0.000000
1412	19.0	2.294274	0.105263	0.473684	0.421053	0.315789	0.631579	0.052632	0.263158	0.368421	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1413	28.0	3.270942	0.214286	0.464286	0.321429	0.571429	0.428571	0.000000	0.464286	0.285714	...	0.000000	0.000000	0.142857	0.035714	0.107143	0.000000	0.107143	0.107143
1414	33.0	3.544881	0.333333	0.303030	0.363636	0.545455	0.242424	0.212121	0.363636	0.242424	...	0.060606	0.000000	0.060606	0.000000	0.090909	0.000000	0.030303	0.030303
1415	5.0	2.321928	0.200000	0.600000	0.200000	0.400000	0.400000	0.200000	0.600000	0.000000	...	0.200000	0.200000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000

1416 rows × 61 columns

3.1.3 Data Analysis

Data analysis was done after data preprocessing to collect, clean, and analyze data to extract meaningful information. Many graphics were drawn to identify any possible patterns and trends in data. Graphics of data analysis were used to make some predictions of future outcomes of our data. Distribution of CPPs and non-CPPs in data is given in Figure 9.

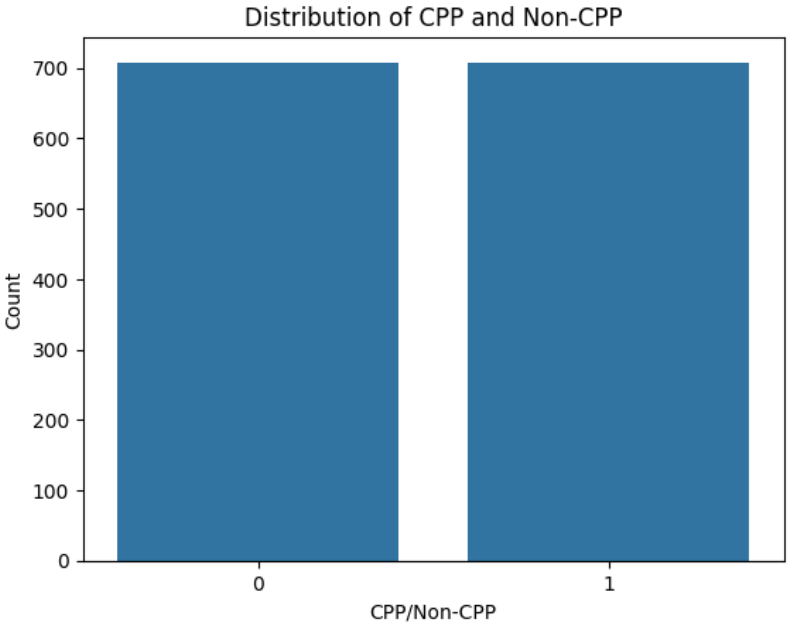


Figure 9. The distribution of CPP and non-CPP data in this study. The same amount of CPPs and non-CPPs are found in data as nearly 700 in each group. The x-axis represents the number of sequences, and the y-axis represents the percentage of sequences that are CPP. The figure shows that the distribution of CPP and non-CPP is balanced, with approximately 50% of the sequences being CPP and 50% being non-CPP.

This Figure 9 is important because it means that the evaluation results of the machine learning model are more reliable. If the distribution of the target variable is not balanced, the model may be biased towards predicting one class over the other. In this case, the model would be more likely to predict that a sequence is CPP if it is already biased towards predicting CPP.

The balanced distribution of the target variable in this dataset suggests that the model is not biased and that the evaluation results are accurate. This is a good sign for the overall reliability of the model.

As shown in Figure 10, normalized van der Waals values of sequences in Group 3 are positively correlated with the probability of being CPPs.

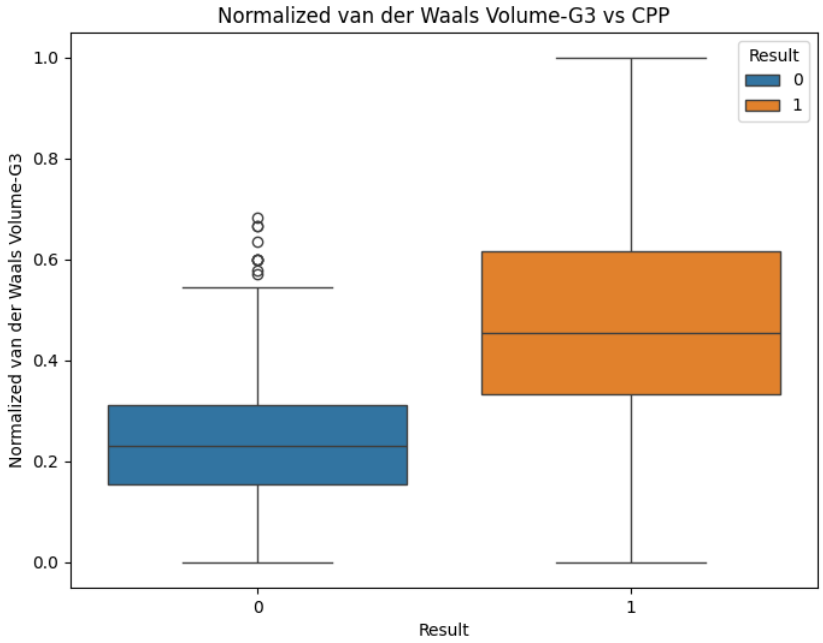


Figure 10. The graph shows the distribution of sequences in group 3 as CPPs or non-CPPs, based on their normalized van der Waals values.

Amino acid sequences in Group 1 have hydrophilic properties. On the other hand, amino acids in Group 2 are neutral and Group 3 amino acids are hydrophobic. It is well-established that hydrophobic peptides exhibit increased propensity for cell membrane penetration. Consequently, amino acids within Group 1 possessing higher hydrophobicity are more likely to facilitate CPP functionality according to Figure 11.

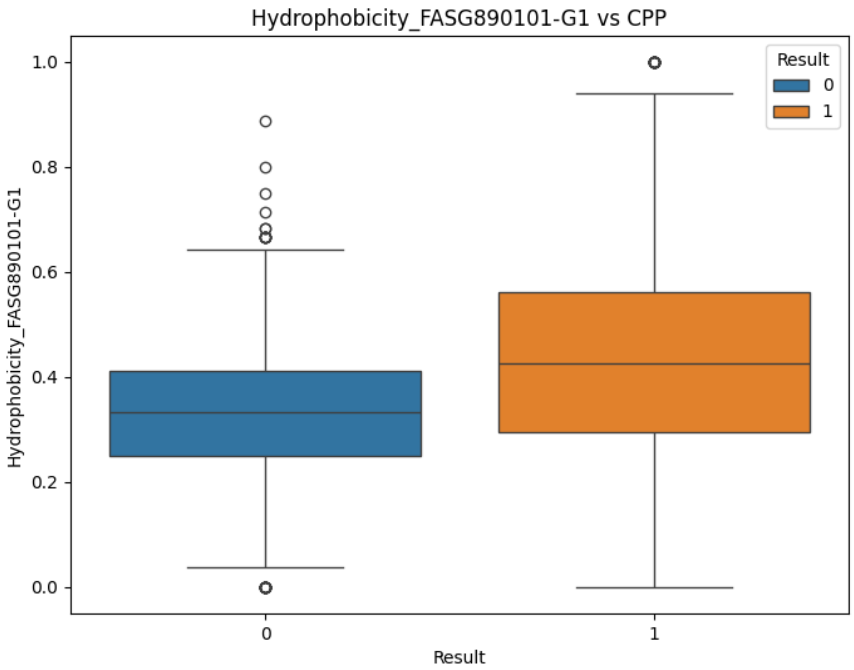


Figure 11. The relation between being CPP/non-CPP and hydrophobicity. The likelihood of a peptide exhibiting CPP behavior in Group 1 increases with increasing hydrophobicity. Amino acids with higher hydrophobicity are therefore more likely to be CPPs than those with lower hydrophobicity (Tsuchiya et al., 2023).

The graph in Figure 12 shows that CPPs tend to have higher polarizability than non-CPPs. This is because CPPs need to be able to interact with the membranes of cells. The higher polarizability of CPPs allows them to easily interact with the charged residues on the cell membrane.

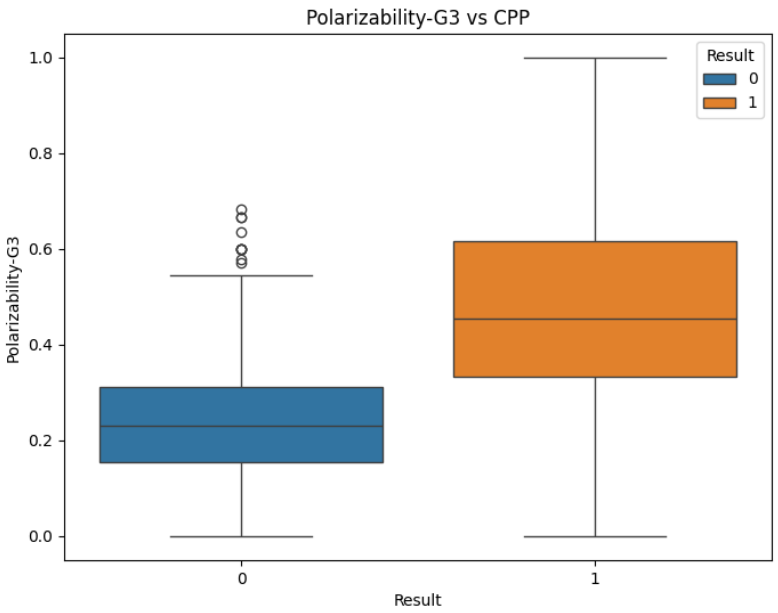


Figure 12. The correlation between polarizability-G3 and classification as CPP/non-CPP. The graph shows the relationship between the polarizability of a peptide sequence and its CPP activity. Polarizability quantifies the susceptibility of a molecule to deformation under the effect of an external electric field (Gäding et al., 2022)

The graph in Figure 13 shows the effect of aspartic acid (D) in a peptide sequence.

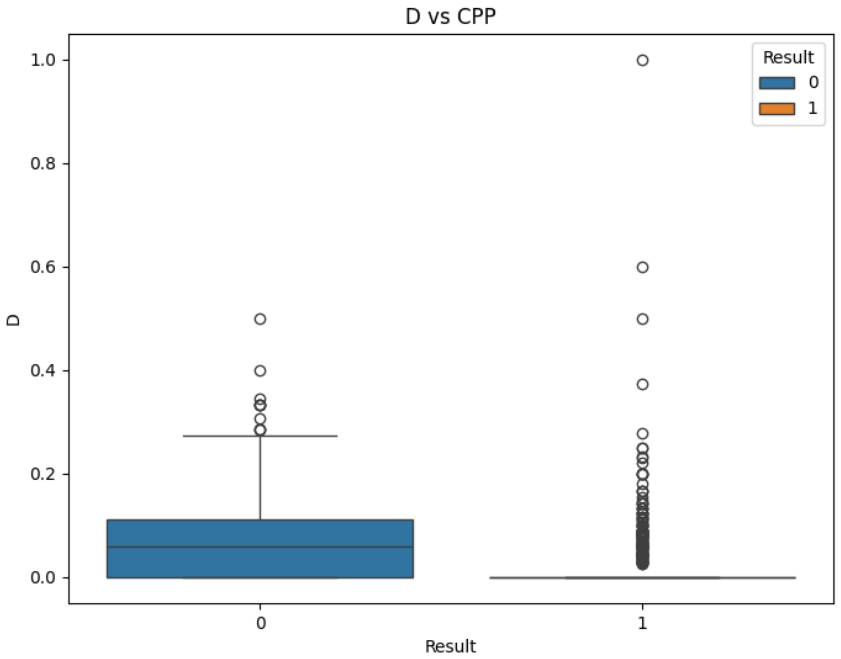


Figure 13. The influence of aspartic acid (D) on the penetration capability of peptides.The effect of aspartic acid on peptides penetration ability. The graph shows that non-CPPs tend to have higher D percentages than

CPPs. This is because D residues are negatively charged (Răcuciu et al., 2022). However, CPPs need to be able to interact with the negatively charged residues on the cell membrane, so generally peptides positively charged are more likely to be a CPP.

In Figure 14, the relationship between the percentage of proline (P) residues in a peptide sequence and its CPP activity is illustrated.

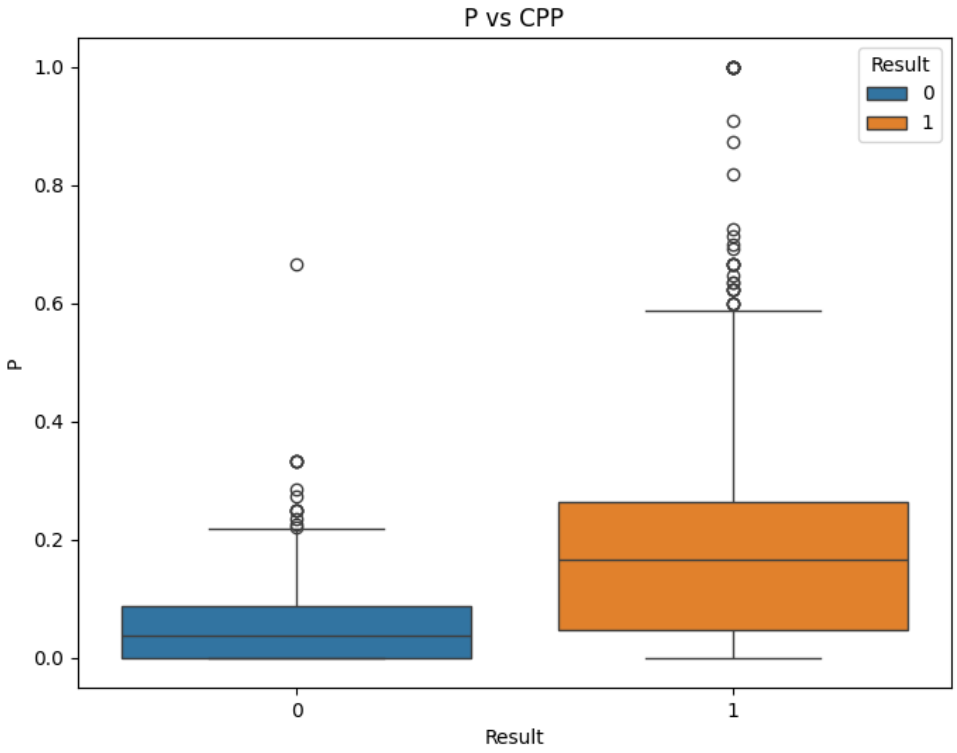


Figure 14. The graph of the impact of the amount of proline in a peptide sequence on whether the peptide is classified as CPP or non-CPP. The graph shows that CPPs tend to have higher proline percentages than non-CPPs. This is because proline residues are rigid and inflexible, which can beneficial to stabilize the structure of the peptide. In addition, proline is a member of amino acid with non-polar R groups. CPPs need to have a stable structure in order to penetrate the cell membrane (Pujals et al., 2008)

3.1.4 Choosing a Model

For the task of predicting whether a given peptide sequence belongs to the category of CPP or non-CPP, previous studies have frequently employed the Random Forest model (yadahalli, favorite,nasiri). The rationale behind the selection of Random Forest for this specific classification task is rooted in its inherent suitability for such problems. As a classification problem, the task involves assigning peptides into two distinct categories, namely "1" and "0," corresponding to CPP and non-CPP labels. Random Forest's ability

to handle classification tasks, particularly when dealing with labeled data such as the binary encoding of sequences into these categories, makes it a logical and effective choice for modeling and prediction in this context. Therefore, the application of the Random Forest model aligns with the requirements of the study, offering a robust approach for addressing the CPP prediction problem.

3.1.5 Training the Model

This section serves as a pivotal phase in the development of our CPP prediction algorithm, facilitating an in-depth assessment of its performance and predictive efficacy.

The first step is to randomly divide the available dataset into two distinct segments. The first segment, referred to as "Training," is allocated to train the algorithm, while the second segment, termed "Test," is reserved for the evaluation of the model's predictions. In this study, dataset is divided into 80% training and 20% test sets. The training set was used to train the algorithm to distinguish between CPPs and non-CPPs. The test set was used to evaluate the accuracy of the algorithm's predictions.

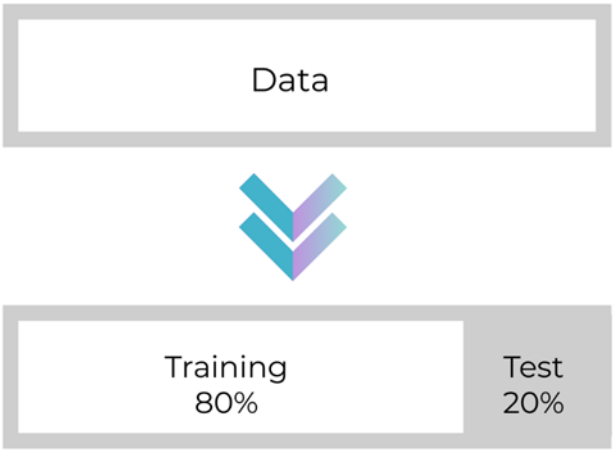


Figure 15. The division of training and test data. Total data is divided into two sets which are training as 80% of the data and test as 20% of the data

During the training, the algorithm learned the features of CPP and non-CPP peptides in the training set. These features can include the length of the peptides, their amino acid composition, their entropy, and so on. The algorithm used these features to learn how to distinguish between CPP and non-CPP peptides.

3.1.6 Evaluating the Model

The model was evaluated after training the model by using test sets. Mainly, performance and accuracy of the model is determined in this step. Accuracy, precision, recall, F1-score were used as evaluation metrics.

After the model evaluation process, training accuracy (100%) and validation accuracy (84.75%) were obtained. Model achieves very high accuracy on the training data, so this indicates the overfitting. To reduce overfitting and eliminate this problem, feature importance technique was used.

3.1.7 Feature Importance

While incorporating numerous features into a machine learning model can enhance its performance, it can also introduce noise and potentially lead to overfitting. To mitigate this risk, feature importance analysis was conducted during the study, and the top 15 features were selected for further utilization. In Figure 17, the list of features was given and were lined up according to their importance.

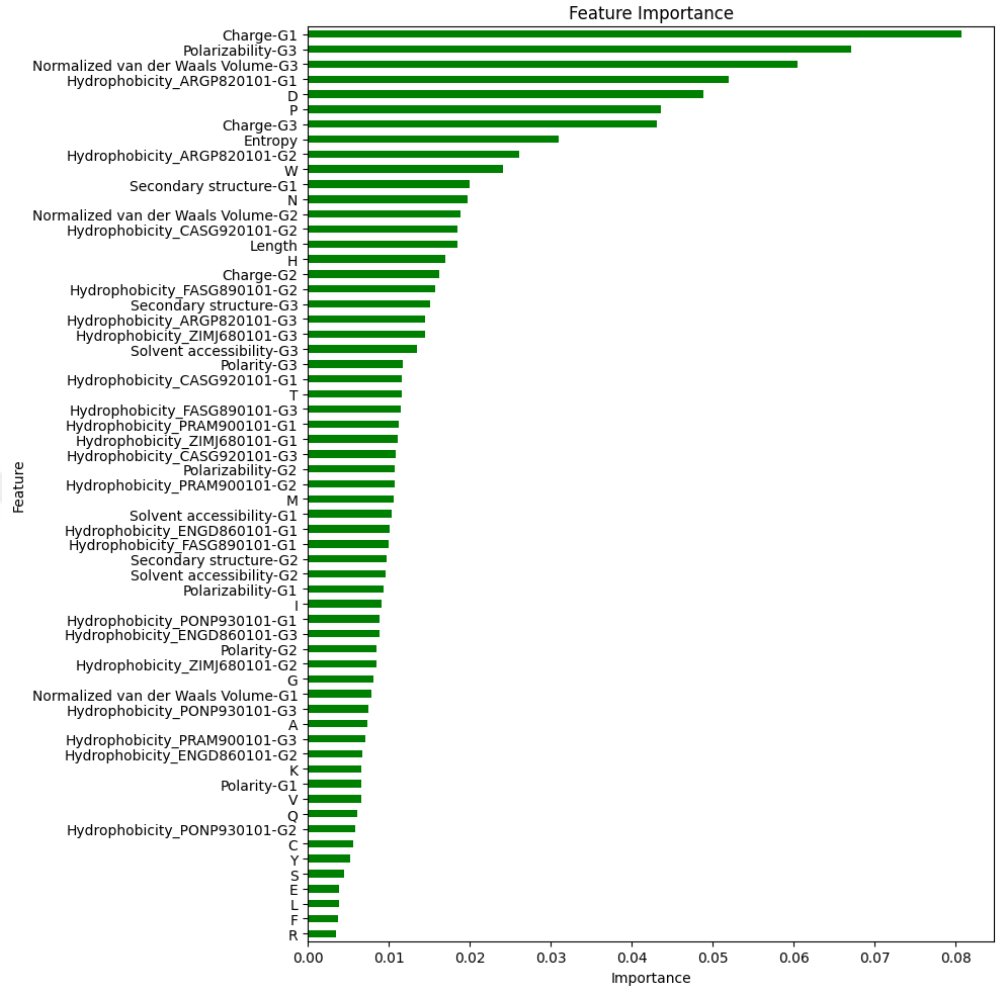


Figure 16. The graph of feature importance technique. The most important feature was Charge-G1. The study was continued using the features selected between charge-g1 and length.

After this process is completed, training accuracy and validation accuracy are calculated again and the overfitting issue has been resolved.

3.1.8 Model Optimization

Hyperparameters were tuned using Grid Search to optimize the Random Forest model, and the parameters with the best performance were selected. During model optimization process, AUC-ROC, Accuracy, F1, Precision, and Recall were calculated.

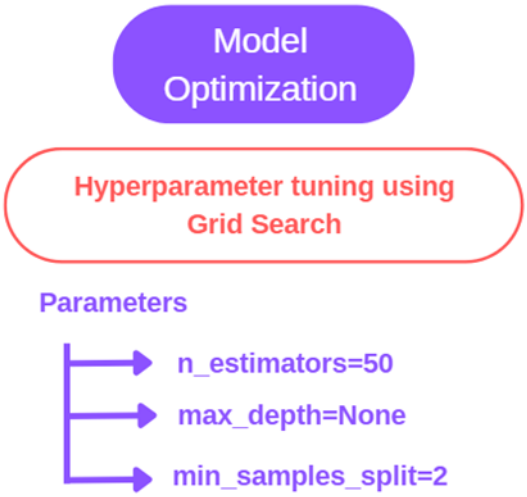


Figure 17. The summary of the model optimization. The best result was obtained with these parameters: max_depth=None, min_samples_split=2, n_estimators=50.

3.1.9 Prediction

Prediction within the machine learning model represents the practical application of the algorithm, where its acquired knowledge and trained parameters are put to use. In our specific study, the primary objective is to provide the algorithm with a peptide sequence and request it to make a binary classification, determining whether the sequence falls into the category of CPP or non-CPP.

To execute this task, the algorithm utilizes the sequence's intrinsic features and learned patterns to make a prediction. When provided with a peptide sequence as input, the model employs its trained parameters and discriminative abilities to assign a binary label, either "CPP" or "non-CPP," based on its interpretation of the sequence's characteristics.

The list of distinct amino acid sequences that were not included in data were used to make predictions that are demonstrated in Table 6. The best-performing parameters were used for prediction.

Table 6. Peptide sequences not found in dataset subjected to prediction processes using a machine learning model

Peptides
RPPGFSPFR
WSYGLRPG
AHALCLTERQIKSNRRMKWKKEN
DFDMLRCMLGRVYRPCWQV
KKLSECLKRIGDELDS
AVALLPAVLLALLAKNNLKDCGLF
DHYACVSSGGQCYLEYSACPIFTKIQCGETCYRAGGKAACCK*

*Protamine sequence

3.2 Experimental Part

3.2.1 Cell Culture Materials & Plasmids

The cell lines, solutions, and equipment used in cell culture experiments are shown in Table 7.

Table 7. Cell lines, cell culture media and equipment

Product Name	**Catalog number**	**Manufacturer**
HEK293 cells	CRL-1517	ATCC
Jurkat cells	CCL-240	ATCC
THP-1 cells	TIB-202	ATCC
RPMI 1640 medium w/stable glutamine	L0498-500	Biowest, USA
Fetal Bovine Serum, European Grade, South America origin	BI0-007-1A	Biological Industries, Israel
Opti-MEM™ I Reduced Serum Medium	31985062	Thermo Scientific, USA
Lipofectamine 2000 reagent	11668019	Thermo Scientific, USA
Protamine Sulphate (Promin)	8699844750961	Vem İlaç, İstanbul
CytoFLEX S	A00-1-1102	Beckman Coulter, USA
NovoCyte® Flow Cytometer	NovoCyte 3000	ACEA Biosciences, USA

In this study, three different plasmids were employed: pCMV-VSV-G, PAXZ plasmid, and PTRIP-SFFV-GFPEV plasmid. These plasmids are provided by Bilkent University. pCMV-VSV-G and PAXZ plasmids are utilized in the construction of lentiviral vectors. The plasmid map of PTRIP-SFFV-GFPEV is illustrated in Figure 18.

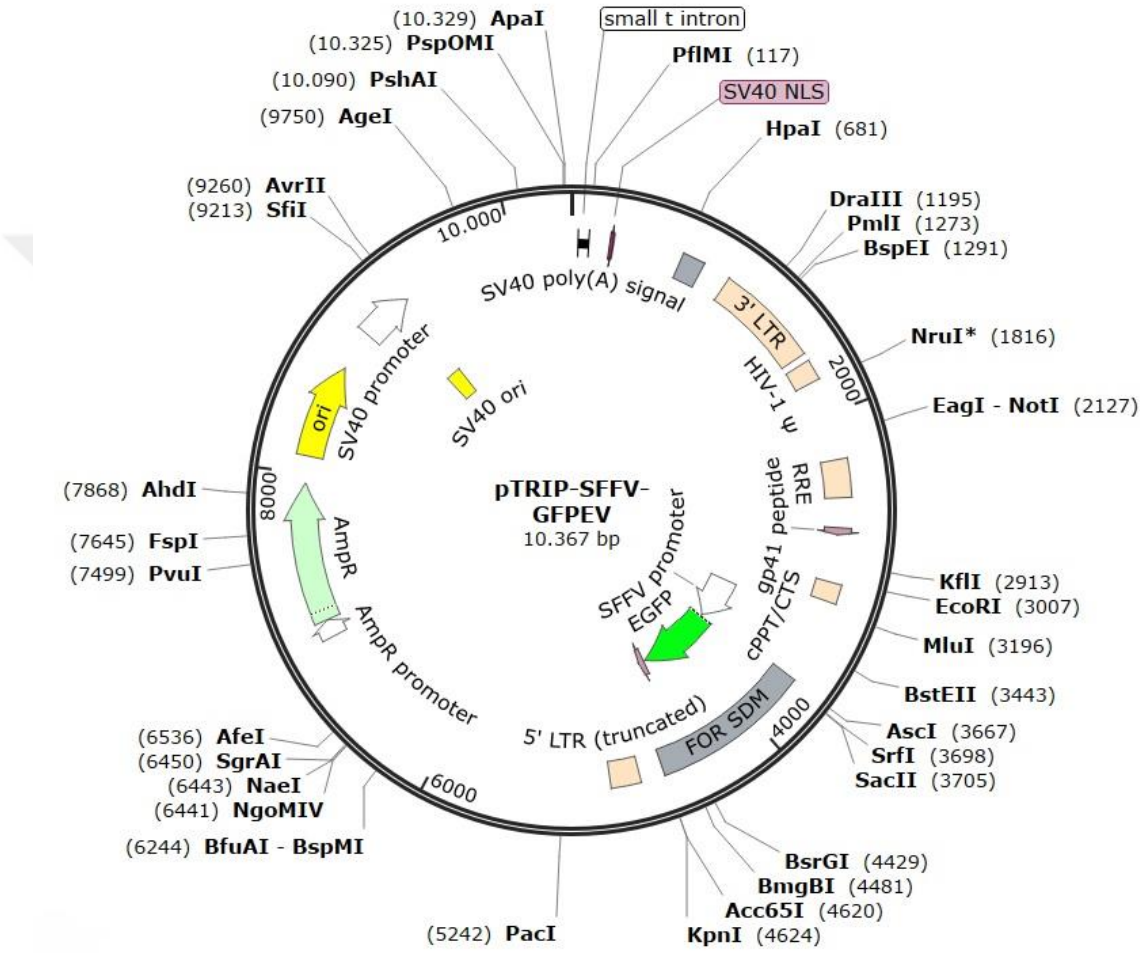


Figure 18. The plasmid map of PTRIP-SFFV-GFPEV. This figure shows a circular map of the PTRIP-SFFV-GFPEV plasmid. The plasmid is 10.367 kilobase pairs (kb) in length and contains a number of different genes and regulatory elements. The green arrow indicates the location of the green fluorescent protein (GFP) gene. GFP is a protein that emits green light when exposed to blue light. It is used as a marker gene in this research.

3.3 Maintenance and Propagation of Mammalian Suspension Cell Lines

Suspension cell lines were maintained in an incubator supplied with 5% CO_2 at 37°C. Jurkat and THP-1 cell lines were cultured in RPMI 1640 growth medium supplemented with 10% (v/v) FBS in T25 culture flasks and passaged every 3 days with 1:2 (v/v) ratio. Passaging was performed by adding fresh medium onto the flask containing suspension cells. Jurkat cells were maintained at a density of 2 x 10^6 cells/mL and THP-1 cells were maintained at a density of 3 x 10^6 cells/mL.

3.4. Lentiviral Transfection/Transduction with Spinoculation Protocol

3.4.1 Production of Lentivirus

HEK293 cells were seeded at a density of 3 x 10^5 cells/mL and incubated at 37°C, 5% CO_2 for 16 hours. Then, HEK293 cells were transfected with 1.5 µg desired lentiviral plasmid (pTRIP-GFP), 0.75 µg VSVG and 0.75 µg PAX2 plasmids in the presence of Lipofectamine 2000 reagent. Briefly, all plasmids were transferred into a 1.5 mL centrifuge tube and diluted in 50 µL OptiMEM. In another centrifuge tube, 6 µL of Lipofectamine 2000 transfection reagent was diluted in 50 µL Opti-MEM. The diluted reagent mixture was added onto the diluted plasmids and mixed by pipetting. The mixture was incubated at room temperature for 20 minutes. Next, the mixture was added onto each well, drop by drop. Cells were incubated at 37°C, 5% CO_2 for 6 hours. Then, the medium was removed and fresh 3 mL medium was added to the wells. After that cells were incubated at 37°C, 5% CO_2 for 24 hours. After that, Lentivirus containing supernatant was collected into an empty 15 mL falcon tube. The supernatant was filtered through 0.45 micron filters into an empty 15 mL falcon tube. If indicated, 8 µg/mL protamine was added to the filtered viral supernatant.

The plate containing HEK293 cells were treated with antiviral disinfectant and placed under UV.

3.4.2 Lentiviral Transduction

Suspension cell lines were seeded at a density of 2 x 10^6 cells/mL and 150 uL of viral supernatant-protamine mixture was added onto each well. The plate was spinoculated by centrifugation at 1200g, 25°C for 2 hours. After spinoculation, the cell pellets were remained undisturbed for 48 hours and incubated at 37°C for 48 hours. Then, cells were transferred to 24-well plate and the volume was adhusted to 1 mL.

After 4 days, transduction efficiency was determined by the amount of GFP expression either measured by flow cytometry or visualized with Fluorescent microscope. For flow cytometry, cells were collected in an 1.5 mL centrifuge tube and centrifuged at 300g for 5 minutes. Cells were resuspended in PBS supplemented with 10% FBS and acquired on CytoFLEX Flow Cytometer. Analysis were performed with CytoFLEX software. For fluorescent microscope, cells were visualized at 4x magnification using Inverted Fluorescent Microscope DMi8 S with either bright-field or excitation filter with central wavelength of 469 nm.

4. RESULTS

4.1 Results of Machine Learning Algorithm

In this study, a machine learning model has been developed to predict whether a peptide is a CPP or a non-CPP based on specific characteristics of the peptide. Various physicochemical properties such as peptide length, polarity, charge, and structural features have been considered to assess the peptides' ability to penetrate the cell. The Random Forest model was employed as the chosen machine learning algorithm. During the model evaluation step, the results of training accuracy and validation accuracy initially indicated the presence of overfitting. To address this issue, a feature importance analysis was conducted. The table below presents the training accuracy and validation accuracy results when overfitting was observed, along with the corresponding values after conducting feature importance analysis.

Table 8. Training accuracy and validation accuracy values of the initial model and model with feature selection. Training accuracy and validation accuracy values while overfitting was observed are given in the first column. After this situation was solved, training accuracy and validation accuracy values in the second column were obtained.

Initial model	Model with feature selection
Training accuracy= 100%	Training accuracy=91.96113074204946%
Validation accuracy= 84.75%	Validation accuracy=86.619718309859%

After that model optimization was done and results of AUC-ROC, accuracy, F1, precision, and recall were presented in Table 9.

Table 9. The results of the performance metrics. Calculated key performance metrics including AUC-ROC, Accuracy, F1, Precision, and Recall are given.

AUC-ROC	test=0.963
Accuracy	test=0.910
F1 score	test=0.908
Precision	test=0.933
Recall	test=0.884
Total time	0.1s

In the final step of developing the machine learning model, the prediction phase, the model was evaluated by making predictions for different peptide sequences not present in the dataset. Table 10 presents the various peptide sequences that were used for predictions along with the corresponding results generated by our model for these peptides.

Table 10. Prediction results of peptides that were not exist in dataset.

Peptide	CPP/ Non-CPP	Prediction Result
RPPGFSPFR	0	0
WSYGLRPG	0	0
AHALCLTERQIKSNRRMKWKKEN	0	0
DFDMLRCMLGRVYRPCWQV	0	1
KKLSECLKRIGDELDS	0	0
AVALLPAVLLALLAKNNLKDCGLF	1	1
DHYACVSSGGQCYLEYSACPIFTKIQCGETCYRAGGKAACCK*	1	1

*Protamine sequence

The model has made predictions for different peptides not present in the dataset, including those with known classifications based on prior research as either CPP or non-CPP. In all

cases, the model provided accurate predictions. One of the peptide sequences that aimed to predict is protamine, a known CPP. Experimental evidence confirming the CPP nature of this peptide has been demonstrated in the experimental section of the study.

4.2 Results of Experimental Part

In the experimental section of the study, the well-known CPP protamine was utilized. The impact of protamine on the penetration into THP-1 and Jurkat cells, which are generally more challenging to infect when forming a complex with lentivirus, was evaluated. Measurements were conducted using fluorescence microscopy to assess cellular uptake. The Protamine-virus complex, incorporating a GFP plasmid, emits fluorescence, and an increase in fluorescence is expected with enhanced entry into the cell.

Figure 19 illustrates the difference between the control group and the group utilizing Protamine in THP-1. The increased fluorescence in the Protamine-treated group indicates an augmentation in cellular entry.

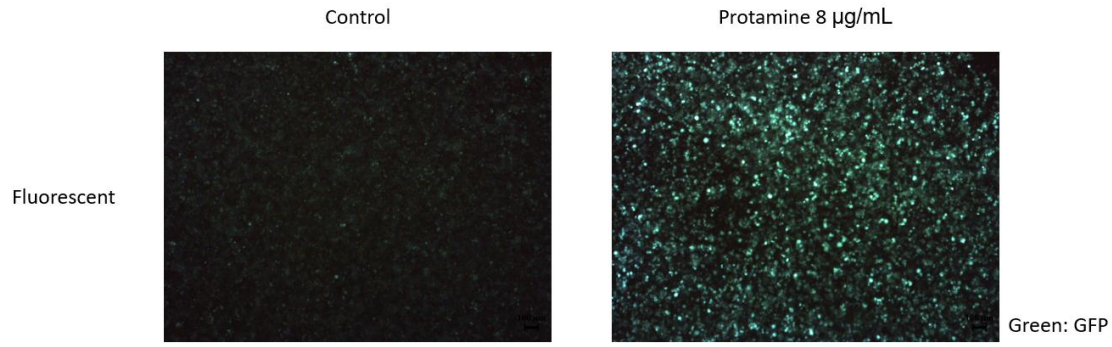


Figure 19. Comparison of control and protamine complex group under fluorescent microscopy.

In Figure 20, transduction efficiencies were evaluated in non-transfected, non-stimulated and with protamine. When Protamine was utilized, the transduction efficiency was measured at 85%, whereas in the absence of Protamine, it was recorded at 52%. Consequently, it was determined that the use of protamine increased the transduction efficiency by 64%.

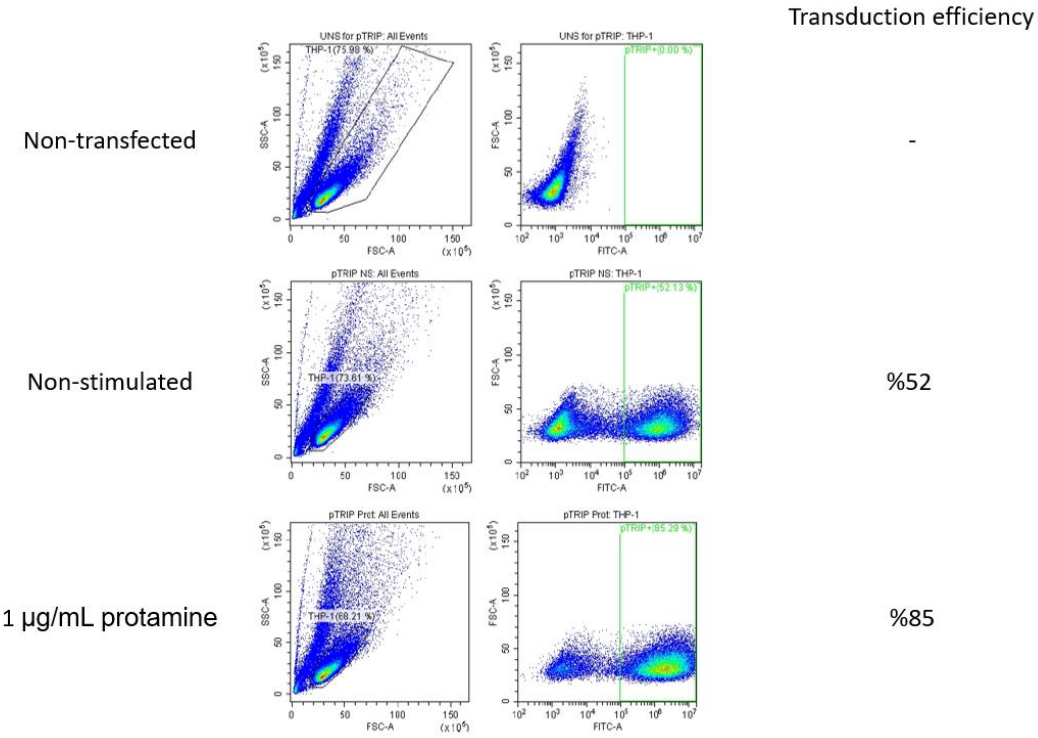


Figure 20. Transduction efficiencies of non-transfected, non-stimulated and with protamine.

In experiments conducted with Jurkat cells, the use of Protamine was observed to enhance cellular entry. The outcomes of bright-field and fluorescence microscopy experiments are depicted in Figure 21.

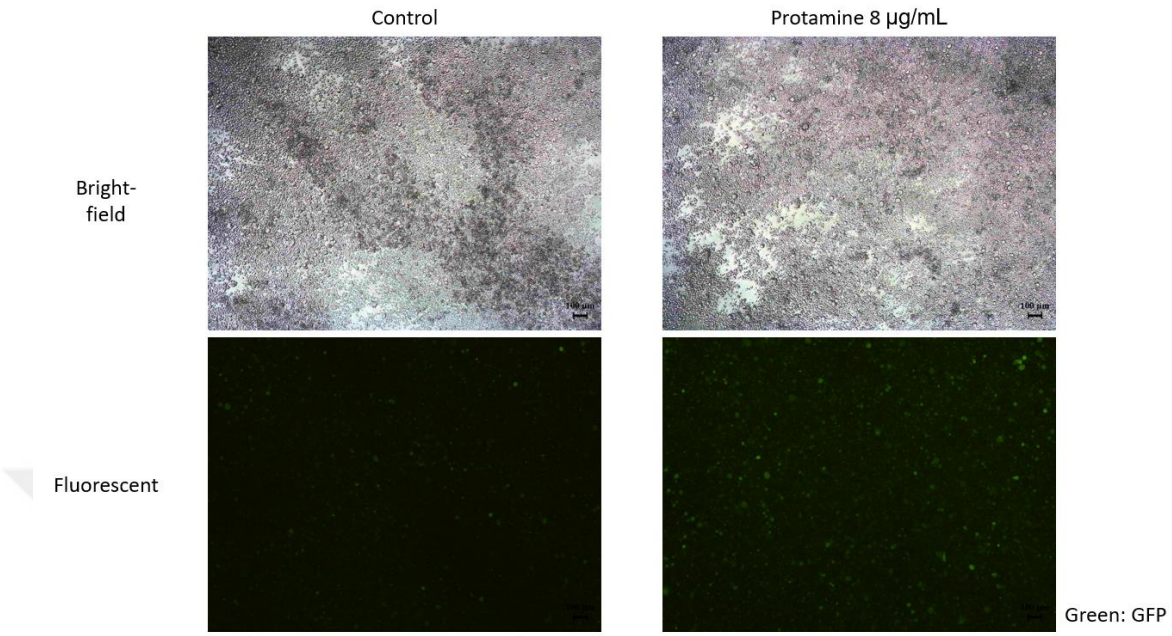


Figure 21. The bright-field and fluorescent microscopy images of the control and protamine groups. Upon examining the results of fluorescence emission, it is apparent that the intensity of green fluorescence increases when protamine is utilized. This observation signifies an augmentation in cellular entry when protamine is applied.

5. DISCUSSION

The study aims to develop an algorithm capable of predicting whether a given peptide sequence functions as a CPP or falls into the non-CPP category. This is achieved through a series of structured steps, including data collection, preprocessing, model selection, and evaluation. Approximately 1400 distinct peptide sequences undergo meticulous analysis based on a myriad of diverse features, such as length, entropy, amino acid composition, and others. The methodology unfolds through a comprehensive exploration of machine learning steps, culminating in the creation of a robust algorithm that can discern CPPs from non-CPPs with a high degree of accuracy. The study adds value to the field by not only looking at a large set of peptide sequences but also by choosing and combining important features to build the model. The balanced distribution of the target variable in the dataset suggests that the model is not biased, indicating the overall reliability of the model. Additionally, the study employs feature importance analysis and model optimization techniques to enhance the model's performance and address overfitting issues. Overall, the study's methodology and findings contribute to a nuanced understanding of the inherent characteristics of each peptide and hold significance in the field of machine learning and peptide research.

The methodology employed in this study differs from other approaches to predicting Cell Penetrating Peptides (CPPs) in several key ways:

1. Data Preprocessing and Balancing: The study performs necessary data preprocessing steps, including encoding the characteristics of each peptide into a format suitable for machine learning algorithms. Furthermore, the study ensures a balanced distribution of CPP and non-CPP sequences in the dataset, which is crucial for the reliability of the model's evaluation results.

2. Comprehensive Feature Analysis: The study conducts a meticulous analysis of diverse features, including length, entropy, amino acid composition, and others, contributing to a nuanced understanding of the inherent characteristics of each peptide. This comprehensive

feature analysis allows for a detailed exploration of the specific traits of each peptide, going beyond simple sorting and providing detailed information about the sequences.

3. Utilization of Machine Learning Techniques: The study systematically applies machine learning techniques, involving data collection, preprocessing, model selection, and evaluation. By leveraging machine learning algorithms, the study aims to develop a robust algorithm capable of accurately discerning CPPs from non-CPPs.

4. Model Selection and Evaluation: The study carefully selects the Random Forest model for the classification task, rooted in its inherent suitability for such problems. The rationale behind the selection of Random Forest is its ability to handle classification tasks, particularly when dealing with labeled data such as the binary encoding of sequences into CPP and non-CPP categories. Additionally, the study evaluates the model's predictions using a test set, ensuring the accuracy and reliability of the algorithm.

Overall, the methodology employed in this study stands out due to its comprehensive feature analysis, systematic application of machine learning techniques, careful model selection and evaluation, and emphasis on data preprocessing and balancing. These aspects distinguish the study's approach to predicting CPPs from other methods and contribute to its significance in the field of machine learning and peptide research.

In the context of machine learning algorithms, the predicted outcomes for protamine, which were not present in the dataset, were experimentally validated. As demonstrated in the experimental section, the results confirmed the accuracy of the predictions.

In the literature, numerous studies have been conducted to predict whether a peptide is a CPP or a non-CPP. In the year 2017, Wei and colleagues developed an online tool named CPPred-RF in their study, contributing to this area of research. The main focus of this study is to propose a novel predictor for identifying CPPs and their uptake efficiency using a two-layer prediction framework based on the random forest algorithm. The article discusses the integration of multiple sequence-based feature descriptors, a two-step feature selection strategy, and the use of a high-quality benchmark dataset to improve the feature representation capability and performance of the predictor. The proposed CPPred-

RF is shown to be competitive with existing methods and is the first predictor that can predict CPPs and their uptake efficiency simultaneously. When comparing this study with the study of Wei and colleagues, several parallels and distinctions become apparent. Notably, both investigations share the overarching objective of developing a predictive model to discern CPPs. This common aim is reinforced by the adoption of machine learning methodologies, specifically the random forest algorithm, for the classification task in both instances. Additionally, a shared approach is observed in the utilization of sequence-based feature descriptors to unravel unique information embedded in CPPs. Both studies further affirm the efficacy of their respective models through comprehensive evaluations, employing the jackknife methodology on benchmark datasets. These commonalities underscore a cohesive alignment in the fundamental goals, methodologies, and evaluative strategies between the two studies. In addition to these similarities, when delving into the distinctions between these two studies, two notable differences come to light. Firstly, the study of Wei and colleagues employs a two-step feature selection strategy to enhance the capability of representing features, whereas our study takes a detailed approach, analyzing various features like length, entropy, and amino acid composition. This nuanced analysis contributes to a more detailed understanding of the inherent characteristics of each peptide. Secondly, the study of Wei and colleagues goes a step further by establishing an online webserver that implements CPPred-RF, offering free access to the public. In contrast, our study does not provide a corresponding webserver for the algorithm. These differences underline the varied approaches in feature selection strategies and accessibility provisions between the two studies (Wei et al., 2017b).

Another study was done by Chen et al. to make predictions for CPP using a machine learning model. The study focuses on predicting and analyzing CPPs using a novel method based on pseudo-amino acid composition (PseAAC) and a random forest model. There are several similarities between this study and Chen's study. oth investigations share a common focus on the prediction of CPPs through the application of machine learning methodologies, particularly employing the random forest model for classification tasks. Additionally, both studies employ feature encoding techniques to represent peptide sequences; the present study utilizes the Protlearn library for sequence-based feature

extraction, while study of Chen et al. relies on pseudo-amino acid composition (PseAAC), resulting in a 270-dimensional vector. Feature selection methods are also a shared aspect, with the present study meticulously analyzing diverse features and the study of Chen et al. incorporating minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) methods. Evaluation methods are consistent in both studies, utilizing cross-validation techniques to ensure the robustness and reliability of classification and prediction models. Divergences arise in the feature encoding approach, where this study incorporates various features such as length, entropy, and amino acid composition, while the study of Chen et al. focuses on PseAAC, generating 270 features per peptide. Regarding model construction, while both studies employ the random forest model for prediction, the present study aims to predict whether a peptide sequence functions as a CPP or falls into the non-CPP category, whereas the study of Chen et al. specifically targets the prediction of cell-penetrating properties of peptides (Chen L. et al, 2015).

In 2020, Fu et al. published a study introducing StackCPPred, a method designed for predicting Cell Penetrating Peptides (CPPs) and their uptake efficiency. Both Fu et al.'s study and the present research employ diverse feature extraction methods to construct sets of features for prediction. This study utilizes features such as length, entropy, and amino acid composition. In contrast, Fu et al.'s study introduces RECM and combines it with other feature extraction methods, including PseAAC, PSSM-composition, and discrete wavelet transform (DWT), aiming to enhance the accuracy of predictions. Differences emerge when evaluating the machine learning approach. This study employs the random forest model to predict CPPs, whereas Fu et al.'s study utilizes a stacking-based machine learning approach, specifically StackCPPred, for the prediction of CPPs and their uptake efficiency. Furthermore, Fu et al.'s study assesses the performance of StackCPPred using benchmark datasets CPP924 and CPPsite3. The results are then compared with state-of-the-art predictors, demonstrating superior performance across sensitivity, specificity, accuracy, and Matthew’s correlation coefficient (MCC) values (Fu et al., 2020).

There are numerous prior works highlighting the significance of the selected features in making predictions for CPPs using the developed machine learning model in this study.

In a study conducted in 2021, the authors introduced a novel framework named BChemRF-CPPred, employing machine learning algorithms to predict cell-penetrating peptides (CPPs) based on their chemical properties. The investigation into the influence of entropy on CPPs led the authors to analyze the correlation between the polar surface area (PSA) and the permeability of CPPs. Their findings indicated that CPPs with high permeability generally exhibit lower PSA values, implying a more compact and less polar structure. This observation aligns with the concept that flexible molecules can establish intrachain hydrogen bonds, reducing their surface polarity and enhancing their ability to permeate cell membranes (de Oliveira et al., 2021).

In another study published in 2018, a comparative analysis of 474 sequence motifs for CPPs was undertaken by the researchers to assess their capacity for transporting cargo into cells. Significant differences in transport performance were identified, with strong correlations being observed between uptake efficiency, sequence length, and positive net charge. Specifically, it was noted that peptides with a positive net charge exhibited a significantly superior ability to penetrate the cell membrane compared to those with a negative or no net charge. Additionally, peptides surpassing 12 amino acid residues in sequence length demonstrated notably higher uptake levels than their shorter counterparts. It is suggested by the researchers that a certain peptide length and a positive net charge are deemed necessary for optimal CPP performance, although they are not considered the sole requirements. Valuable information for researchers and practitioners in the field of molecular medicine aiming to develop CPP-based delivery systems is provided by this study (Ramaker et al., 2018).

In another study conducted in 2022, the impact of peptide cargo characteristics, such as net charge and length, on the internalization efficiency of the CPP/cargo complex was investigated by the authors. The study revealed a significant influence of the net charge of cargoes on the uptake of the CPP/cargo complex. Cargoes with a net positive charge enhanced the overall uptake of the complexes, while those with net neutral and negative charges diminished uptake. Conversely, the net length of the cargo did not exhibit a significant effect on the uptake of the CPP/cargo complexes. To illustrate the impact of length and charge on cell-penetrating peptides, the authors conducted experiments

employing both structured and unstructured CPPs with cargoes featuring varying net charges and lengths. The findings indicated that positively charged cargoes enhanced the uptake efficiency of the CPP/cargo complex, while neutrally charged cargoes exhibited diminished uptake levels. Additionally, when comparing cargoes with a net charge of +2 but different lengths, no significant difference in uptake was observed, suggesting that charge, rather than length, played a crucial role in the uptake of the complex (Hymel et al., 2022).

A study published by Gautam and colleagues focuses on the development of *in silico* methods for predicting highly effective CPPs. The researchers employed support vector machine (SVM)-based models to predict and design these peptides, aiming to facilitate rapid screening before synthesis. Various features were utilized as input for the SVM classifier, including amino acid composition, dipeptide composition, binary profile of patterns, and physicochemical properties. Specifically, the study found that specific residues such as Arg, Lys, Pro, Trp, Leu, and Ala were preferred at certain locations in CPPs, enabling discrimination between CPPs and non-CPPs based on amino acid composition. To illustrate the effect of amino acid composition on CPPs, the study compared the percent average amino acid composition of whole peptides, N-terminal residues, and C-terminal residues between CPPs and non-CPPs. This comparison revealed distinct differences in amino acid composition between the two groups, underscoring the significance of amino acid composition in distinguishing CPPs from non-CPPs. The study also showed that the hydrophobicity of CPPs plays a crucial role in their cell-penetrating ability, as certain hydrophobic residues are preferred at specific positions within the peptides. In summary, the study demonstrated that features like amino acid composition, binary profile of patterns, and motifs can effectively be used to train an SVM classifier for predicting CPPs with high accuracy. The developed hybrid model, which combines motif information and the binary profile of peptides, achieved a maximum accuracy of 97.40%, emphasizing the importance of amino acid composition in predicting CPPs (Gautam et al, 2013).

A study performed by Sanders and colleagues aimed to investigate the importance of a peptide's hydrophobicity value in its potential to act as a CPP. The use of support vector

machines (SVMs) was explored to predict novel CPPs and analyze how the composition of training datasets influences the classification of peptides as cell-penetrating. It was determined in the study that the cell-penetrating potential of peptides can be accurately predicted by SVMs based on primary biochemical properties, with a specific emphasis on the crucial role of hydrophobicity in determining CPP activity. Several peptides were synthesized and tested, revealing that those exhibiting higher hydrophobicity were more likely to exhibit cell-penetrating capabilities. In summary, insights into the biochemical properties associated with cell penetration are provided by the study, underscoring the potential of machine learning methods for the rapid screening of potential CPPs (Sanders et al., 2011).

In different study, the importance of secondary structure in cell-penetrating peptides (CPPs) was explored. It was observed that the contribution of stable secondary structures, such as alpha-helices, to the ability of CPPs to traverse membranes and undergo more efficient internalization into cells was investigated by the authors. Furthermore, modular CPPs were designed based on the environmental sensing ability of proline residues. The introduction of L-proline or modified guanidinyl L-proline was found to aid in controlling the secondary structure of the peptides, consequently enhancing their cell-penetrating properties. The overall emphasis of the study lies in highlighting the significance of considering secondary structure in the design of CPPs for achieving efficient intracellular drug delivery (Kalafatovic et al., 2017).

Basing on the literature reviews discussed above, physicochemical and structural features that could play a crucial role in a peptide's potential as a CPP have been identified. The features utilized in this study are deemed significant based on previous literature, indicating their importance. The selection of machine learning features took into account the requirements of this study, aiming for classification. Hence, a Random Forest model was employed.

Looking forward, several avenues for future research emerge. Firstly, the algorithm's predictive capacity could be enhanced by considering additional features, such as amino acid sequence order, dipeptide interaction, to further refine the discernment of CPPs. Exploring more advanced machine learning models or ensemble methods may also

contribute to the robustness of the algorithm. Practical applications of the developed model in drug design and biotechnology warrant thorough investigation, with potential implications for accelerating the discovery of novel CPPs with specific functionalities.

However, it is essential to acknowledge the limitations of this study. The study relies on a specific set of features, including amino acid composition, length, entropy, and other descriptors. Limiting the feature set to these may overlook potentially crucial features influencing the prediction of CPPs. Incorporating additional physicochemical or structural features might enhance the model's performance. Another limitation of this study is rooted in the relatively confined nature of the dataset. The dataset, while comprehensive for its purpose, is finite and may not fully capture the diversity inherent in CPP sequences. To address this limitation and enhance the robustness of the findings, future iterations of this study could greatly benefit from the continual expansion and updates of the dataset over time. By incorporating a more extensive and diverse set of peptide sequences, the model could be better generalized to various biological contexts, improving its predictive capabilities and bolstering the overall reliability of our algorithm. Continuous efforts to enrich the dataset would not only refine the discernment of CPPs but also ensure the adaptability of the model to emerging trends and variations in CPP sequences, thus advancing the study to a more sophisticated and encompassing level. Yet another constraint arises from the current approach, where each peptide sequence is treated independently, overlooking potential dependencies or correlations between sequences. Recognizing the interdependence of sequences could provide a more realistic representation of the predictive model. Acknowledging and incorporating such dependencies in future research may enhance the model's accuracy by capturing nuanced relationships between different peptide sequences, further refining the discernment of CPPs. In addition to these considerations, it is also possible to discuss limitations related to the model itself. The study utilizes a Random Forest classifier, known for its power and versatility. However, it is crucial to acknowledge that the model's performance may be influenced by the particular choice of the algorithm. Conducting comparative evaluations with other machine learning models could offer a more comprehensive understanding of the predictive capabilities. Recognizing these challenges lays the groundwork for future studies to address these gaps in knowledge.

Comparatively, the methodology in this study demonstrates proficiency in maintaining a balanced class distribution, conducting a comprehensive feature analysis, performing an extensive hyperparameter tuning analysis, and avoiding overfitting. In this study, 50% of the dataset consists of CPPs, while the remaining 50% comprises non-CPPs. This results in a balanced dataset, mitigating the potential impact of an imbalanced dataset situation on the outcomes. Additionally, during the model development phase, the obtained results revealed a Training Accuracy of 100% and a Validation Accuracy of 84.75%. Recognizing the significant disparity between these two-accuracy metrics and the perfect Training Accuracy, indicative of overfitting, prompted further investigation. To address this issue, a thorough feature importance analysis was conducted. Subsequently, an extensive hyperparameter tuning analysis was performed to mitigate overfitting, ensuring a more robust and generalized model. This process was complemented by a comprehensive feature analysis, enhancing the model's overall performance and addressing potential overfitting concerns effectively. Considering the ever-evolving landscape of technology and experimental techniques, our model's effectiveness may be influenced by external factors. Future research should, therefore, remain attuned to advancements in these domains.

Moreover, ethical considerations surrounding the use of machine learning in peptide prediction should not be overlooked. Ensuring privacy, avoiding misuse, and considering the ethical implications of personalized medicine are vital aspects that warrant attention in subsequent investigations.

Opportunities for collaboration with researchers and industries interested in similar topics could foster interdisciplinary efforts. Such collaborations might lead to the refinement and application of the developed algorithm in diverse research and industrial settings. In conclusion, these future considerations underscore the dynamic nature of research in predicting CPPs and pave the way for continued advancements in this field.

6. CONCLUSION and SUGGESTIONS

In this study, a novel machine learning model has been developed to predict whether a peptide chain is a cell-penetrating peptide (CPP) or a non-CPP based on specific physicochemical properties of the peptide sequence. Approximately 1400 different peptide chains were utilized as data, incorporating features such as amino acid composition, length, entropy, and CTDC, along with peptide sequences as inputs. The output is binary, with a result of "1" indicating a CPP and "0" denoting a non-CPP. The Random Forest model was chosen as the machine learning model for classification.

The model achieved a training accuracy of 91.96% and a validation accuracy of 86.62%, demonstrating its effectiveness in predicting peptide classification. Furthermore, predictions were made for seven different peptides not present in the dataset, and the developed algorithm accurately predicted nearly all of them. Notably, protamine, a peptide not present in the dataset but known to be a CPP, was correctly predicted as a CPP by the algorithm. Experimental validation in the study confirmed the CPP property of protamine.

While the initial model has yielded promising results, there is potential for further enhancement in the future. The algorithm serves as a valuable tool for predicting CPP characteristics without the need for peptide synthesis. Looking ahead, the improved model could facilitate the easier development of new CPPs for use in gene therapy. Furthermore, this pre-manufacturing insight into peptide behavior, facilitated by machine learning, is poised to enhance the success of gene therapy. The ability to predict and understand the properties of peptides prior to synthesis not only streamlines the design process but also empowers researchers to tailor peptides specifically for improved efficacy in gene therapeutic applications. This integrated approach, coupling machine learning predictions with targeted peptide design, holds the potential to significantly advance the development of novel therapeutic interventions in gene therapy, ultimately contributing to more effective and personalized treatment strategies.

REFERENCES

A. Dobchev D, Mager I, Tulp I, Karelson G, Tamm T, Tamm K , et al. Prediction of cell-penetrating peptides using artificial neural networks. Current Computer Aided-Drug Design. 2010; 6: 79-89.

Al-Dosari MS, Gao X. (2009). Nonviral gene delivery: Principle, limitations, and recent progress. In *AAPS Journal* (Vol. 11, pp. 671-681).

Alhakamy NA, Nigatu AS, Berkland CJ, Ramsey JD. (2013). Noncovalently associated cell-penetrating peptides for gene delivery applications. In *Therapeutic Delivery* (Vol. 4, pp. 741-757): Newlands Press Ltd.

Alharbi F, Vakanski A. (2023). Machine learning methods for cancer classification using gene expression data: A review. In *Bioengineering* (Vol. 10): MDPI.

Alley EC, Turpin M, Liu AB, Kulp-McDowall T, Swett J, Edison R , et al. A machine learning toolkit for genetic engineering attribution to facilitate biosecurity. Nature Communications. 2020; 11.

Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: A corrected feature importance measure. Bioinformatics. 2010; 26: 1340-1347.

Anashkina AA, Petrushanko IY, Ziganshin RH, Orlov YL, Nekrasov AN. Entropy analysis of protein sequences reveals a hierarchical organization. Entropy. 2021; 23.

Arbab AS, Yocum GT, Kalish H, Jordan EK, Anderson SA, Khakoo AY , et al. Efficient magnetic cell labeling with protamine sulfate complexed to ferumoxides for cellular mri. Blood. 2004; 104: 1217-1223.

Arsenijevic Y, Berger A, Udry F, Kostic C. (2022). Lentiviral vectors for ocular gene therapy. In *Pharmaceutics* (Vol. 14): MDPI.

Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T , et al. An introduction to machine learning. Clinical Pharmacology and Therapeutics. 2020; 107: 871-885.

Balogh OM, Benczik B, Horváth A, Pétervári M, Csermely P, Ferdinandy P , et al. Efficient link prediction in the protein–protein interaction network using topological information in a generative adversarial network machine learning model. BMC Bioinformatics. 2022; 23.

Belle V, Papantonis I. (2021). Principles and practice of explainable machine learning. In *Frontiers in Big Data* (Vol. 4): Frontiers Media S.A.

Bulaklak K, Gersbach CA. (2020). The once and future gene therapy. In *Nature Communications* (Vol. 11): Nature Research.

Burny A, Bex F, Dequiedt F, Adam E, Gatot JS, Kerkhofs P , et al. Leukemia viruses. In PJ Delves (Ed.), Encyclopedia of immunology (second edition). Oxford: Elsevier; 1998, p. 1562-1565.

Byun Y, Singh VK, Yang VC. Low molecular weight protamine: A potential nontoxic heparin antagonist. Thrombosis research. 1999; 94: 53-61.

Carrasco-Padilla C, Aguilar-Sopeña O, Gómez-Morón A, Alegre-Gómez S, Sánchez-Madrid F, Martín-Cófreces NB , et al. Chapter 3 - t cell activation and effector function in the human jurkat t cell model. In C Thomas L Galluzzi (Eds.), Methods in cell biology. Academic Press; 2023, 178, p. 25-41.

Chang ST, Ghosh D, Kirschner DE, Linderman JJ. Peptide length-based prediction of peptide - mhc class ii binding. Bioinformatics. 2006; 22: 2761-2767.

Chen L, Chu C, Huang T, Kong X, Cai YD. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. Amino Acids. 2015; 47: 1485-1493.

Chen N, Wang HB, Wu BQ, Jiang JH, Yang JT, Tang LJ , et al. Using random forest to detect multiple inherited metabolic diseases simultaneously based on gc-ms urinary metabolomics. Talanta. 2021; 235.

Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Peter Campbell J. Introduction to machine learning, neural networks, and deep learning. Translational Vision Science and Technology. 2020; 9.

Clark WR, Gao D. (2001). *Low-molecular weight proteins in end-stage renal disease: Potential toxicity and dialytic removal mechanisms renal handling of lmwp normal renal elimination*. Retrieved from http://journals.lww.com/jasn

Cring MR, Sheffield VC. (2022). Gene therapy and gene correction: Targets, progress, and challenges for treating human diseases. In *Gene Therapy* (Vol. 29, pp. 3-12): Springer Nature.

De Jong H, Bonger KM, Löwik DWPM. (2020). Activatable cell-penetrating peptides: 15 years of research. In *RSC Chemical Biology* (Vol. 1, pp. 192-203): Royal Society of Chemistry.

de Oliveira ECL, Santana K, Josino L, Lima e Lima AH, de Souza de Sales Júnior C. Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space. Scientific Reports. 2021; 11: 7628.

Dixon JE, Osman G, Morris GE, Markides H, Rotherham M, Bayoussef Z , et al. Highly efficient delivery of functional cargoes by the synergistic effect of gag binding motifs and cell-penetrating peptides. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113: E291-E299.

Duan D. (2018). Micro-dystrophin gene therapy goes systemic in duchenne muscular dystrophy patients. In *Human Gene Therapy* (Vol. 29, pp. 733-736): Mary Ann Liebert Inc.

Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH. Recognition of a protein fold in the context of the scop classification. Proteins: Structure, Function and Genetics. 1999; 35: 401-407.

Dubchak I, Muchnikt I, Holbrook SR, Kim S-H. (1995). *Prediction of protein folding class using global description of amino acid sequence*. Retrieved from https://www.pnas.org

Durand S, Cimarelli A. (2011). The inside out of lentiviral vectors. In *Viruses* (Vol. 3, pp. 132-159).

Frangoul H, Altshuler D, Cappellini MD, Chen Y-S, Domm J, Eustace BK , et al. Crispr-cas9 gene editing for sickle cell disease and β-thalassemia. New England Journal of Medicine. 2021; 384: 252-260.

Friedmann T. A brief history of gene therapy. Nature Genetics. 1992; 2: 93-98.

Fu X, Cai L, Zeng X, Zou Q. Stackcppred: A stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics. 2020; 36: 3028-3034.

Gäding J, Tocci G, Busch M, Huber P, Meißner RH. Impact of confinement and polarizability on dynamics of ionic liquids. Journal of Chemical Physics. 2022; 156.

Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A , et al. In silico approaches for designing highly effective cell penetrating peptides. Journal of Translational Medicine. 2013; 11.

Gautam A, Nanda JS, Samuel JS, Kumari M, Priyanka P, Bedi G , et al. Topical delivery of protein and peptide using novel cell penetrating peptide imt-p8. Scientific Reports. 2016; 6.

Gessner I, Neundorf I. Nanoparticles modified with cell-penetrating peptides: Conjugation mechanisms, physicochemical properties, and application in cancer diagnosis and therapy. International Journal of Molecular Sciences. 2020; 21.

Ghazal TM, Al Hamadi H, Umar Nasir M, Atta Ur R, Gollapalli M, Zubair M , et al. Supervised machine learning empowered multifactorial genetic inheritance disorder prediction. Computational Intelligence and Neuroscience. 2022; 2022.

Gill P, Stewart K, Treasure E, Chadwick B. Methods of data collection in qualitative research: Interviews and focus groups. British Dental Journal. 2008; 204: 291-295.

Goldstein BA, Polley EC, Briggs FBS. (2011). Random forests for genetic association studies. In *Statistical Applications in Genetics and Molecular Biology* (Vol. 10).

Gonçalves GAR, Paiva RdMA. (2017). Gene therapy: Advances, challenges and perspectives. In *Einstein (Sao Paulo, Brazil)* (Vol. 15, pp. 369-375).

González-Aramundiz JV, Peleteiro M, González-Fernández Á, Alonso MJ, Csaba NS. Protamine nanocapsules for the development of thermostable adjuvanted nanovaccines. Molecular Pharmaceutics. 2018; 15: 5653-5664.

Grotjohann T, Testa I, Leutenegger M, Bock H, Urban NT, Lavoie-Cardinal F , et al. Diffraction-unlimited all-optical imaging and writing with a photochromic gfp. Nature. 2011; 478: 204-208.

Gu X, Chen Z, Wang D. Prediction of g protein-coupled receptors with ctdc extraction and mrmd2.0 dimension-reduction methods. Frontiers in Bioengineering and Biotechnology. 2020; 8.

Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M , et al. (2019). Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. In *American Journal of Roentgenology* (Vol. 212, pp. 38-43): American Roentgen Ray Society.

He H, Ye J, Liu E, Liang Q, Liu Q, Yang VC. (2014). Low molecular weight protamine (lmwp): A nontoxic protamine substitute and an effective cell-penetrating peptide. In *Journal of Controlled Release* (Vol. 193, pp. 63-73): Elsevier B.V.

Hofmann B, Bushnaq H, Kraus FB, Raspé C, Simm A, Silber RE , et al. Immediate effects of individualized heparin and protamine management on hemostatic activation and platelet function in adult patients undergoing cardiac surgery with tranexamic acid antifibrinolytic therapy. Perfusion (United Kingdom). 2013; 28: 412-418.

Hu J, Szymczak S. (2023). A review on longitudinal data analysis with random forest. In *Briefings in Bioinformatics* (Vol. 24): Oxford University Press.

Huda SA, Ila IJ, Sarder S, Shamsujjoha M, Ali MNY. (2019). *An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms.* Paper presented at the Institute of Electrical and Electronics Engineers.

Hymel HC, Rahnama A, Sanchez OM, Liu D, Gauthier TJ, Melvin AT. How cargo identity alters the uptake of cell-penetrating peptide (cpp)/cargo complexes: A study on the effect of net cargo charge and length. Cells. 2022; 11.

Jafarlou M, Baradaran B, Saedi TA, Jafarlou V, Shanehbandi D, Maralani M , et al. An overview of the history, applications, advantages, disadvantages and prospects of gene therapy. J Biol Regul Homeost Agents. 2016; 30: 315-321.

Jarzebska NT, Mellett M, Frei J, Kündig TM, Pascolo S. (2021). Protamine-based strategies for rna transfection. In *Pharmaceutics* (Vol. 13): MDPI.

Jayanthi K, Mahesh C. A study on machine learning methods and applications in genetics and genomics. International Journal of Engineering & Technology. 2018; 7: 201-201.

Jeong C, Yoo J, Lee D, Kim Y-C. A branched tat cell-penetrating peptide as a novel delivery carrier for the efficient gene transfection. Biomaterials research. 2016; 20: 1-8.

Kabir M, Nantasenamat C, Kanthawong S, Charoenkwan P, Shoombuatong W. Large-scale comparative review and assessment of computational methods for phage virion proteins identification. EXCLI Journal. 2022; 21: 11-29.

Kalafatovic D, Giralt E. (2017). Cell-penetrating peptides: Design strategies beyond primary structure and amphipathicity. In *Molecules* (Vol. 22): MDPI AG.

Kato S, Kobayashi K, Inoue KI, Kuramochi M, Okada T, Yaginuma H , et al. A lentiviral strategy for highly efficient retrograde gene transfer by pseudotyping with fusion envelope glycoprotein. Human Gene Therapy. 2011; 22: 197-206.

Katz MG, Fargnoli AS, Williams RD, Bridges CR. (2013). Gene therapy delivery systems for enhancing viral and nonviral vectors for cardiac diseases: Current concepts and future applications. In *Human Gene Therapy* (Vol. 24, pp. 914-927).

Kauffman WB, Guha S, Wimley WC. Synthetic molecular evolution of hybrid cell penetrating peptides. Nature Communications. 2018; 9.

Kawashima S, Kanehisa M. (2000). *Aaindex: Amino acid index database*. Retrieved from www.genome.ad.jp/aaindex/

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. Aaindex: Amino acid index database, progress report 2008. Nucleic Acids Research. 2008; 36.

Khalaf Jabbar Rafiqul Zaman Khan HDR. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). 2015.

Khan MKH, Guo W, Liu J, Dong F, Li Z, Patterson TA , et al. (2023). Machine learning and deep learning for brain tumor mri image segmentation. In *Experimental Biology and Medicine*: SAGE Publications Inc.

Kim A, Shin TH, Shin SM, Pham CD, Choi DK, Kwon MH , et al. Cellular internalization mechanism and intracellular trafficking of filamentous m13 phages displaying a cell-penetrating transbody and tat peptide. PLoS ONE. 2012; 7.

Kim GC, Cheon DH, Lee Y. Challenge to overcome current limitations of cell-penetrating peptides. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2021; 1869: 140604.

Kohn DB, Chen YY, Spencer MJ. Successes and challenges in clinical gene therapy. Gene Therapy. 2023; 30: 738-746.

Kotterman MA, Schaffer DV. Engineering adeno-associated viruses for clinical gene therapy. Nature Reviews Genetics. 2014; 15: 445-451.

Kremer J, Steenstrup Pedersen K, Igel C. Active learning with support vector machines. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014; 4: 313-326.

Kristensen M, Birch D, Nielsen HM. (2016). Applications and challenges for use of cell-penetrating peptides as delivery vectors for peptide and protein cargos. In *International Journal of Molecular Sciences* (Vol. 17): MDPI AG.

Kumar SRP, Markusic DM, Biswas M, High KA, Herzog RW. (2016). Clinical development of gene therapy: Results and lessons from recent successes. In *Molecular Therapy Methods and Clinical Development* (Vol. 3, pp. 16034-16034): Elsevier Inc.

Kunz SA, Miles LF, Ianno DJ, Mirowska-Allen KL, Matalanis G, Bellomo R , et al. The effect of protamine dosing variation on bleeding and transfusion after heparinisation for cardiopulmonary bypass. Perfusion (United Kingdom). 2018; 33: 445-452.

Kurrikoff K, Gestin M, Langel Ü. (2016). Recent in vivo advances in cell-penetrating peptide-assisted drug delivery. In *Expert Opinion on Drug Delivery* (Vol. 13, pp. 373-387): Taylor and Francis Ltd.

Lee H, Lim SI, Shin SH, Lim Y, Koh JW, Yang S. Conjugation of cell-penetrating peptides to antimicrobial peptides enhances antibacterial activity. ACS Omega. 2019; 4: 15694-15701.

Libbrecht MW, Noble WS. (2015). Machine learning applications in genetics and genomics. In *Nature Reviews Genetics* (Vol. 16, pp. 321-332): Nature Publishing Group.

Liu B, Li S, Wang Y, Lu L, Li Y, Cai Y. Predicting the protein sumo modification sites based on properties sequential forward selection (psfs). Biochemical and Biophysical Research Communications. 2007; 358: 136-139.

Liu BR, Chen HH, Chan MH, Huang YW, Aronstam RS, Lee HJ. Three arginine-rich cell-penetrating peptides facilitate cellular internalization of red-emitting quantum dots. Journal of Nanoscience and Nanotechnology. 2015; 15: 2067-2078.

Liu E, Vega S, Treiser MD, Sung HJ, Moghe PV. 3.317 - fluorescence imaging of cell–biomaterial interactions. In P Ducheyne (Ed.), Comprehensive biomaterials. Oxford: Elsevier; 2011, p. 291-303.

Liu Y, Wan H-H, Tian D-M, Xu X-J, Bi C-L, Zhan X-Y , et al. Development and characterization of high efficacy cell-penetrating peptide via modulation of the histidine and arginine ratio for gene therapy. Materials. 2021; 14: 4674.

Lo Piano S. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. Humanities and Social Sciences Communications. 2020; 7.

Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C , et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. Journal of medical Internet research. 2016; 18: e323-e323.

Lv Z, Jin S, Ding H, Zou Q. A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. Frontiers in Bioengineering and Biotechnology. 2019; 7.

M H, M.N S. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process. 2015; 5: 01-11.

Ma L, Zhang G, Doyle MP. Green fluorescent protein labeling of listeria, salmonella, and escherichia coli o157:H7 for safety-related studies. PLoS ONE. 2011; 6.

Mairhofer J, Grabherr R. (2008, 2008/6//). *Rational vector design for efficient non-viral gene delivery: Challenges facing the use of plasmid DNA.* Paper presented at the Molecular Biotechnology.

Martín I, Teixidó M, Giralt E. (2010). Building cell selectivity into cpp-mediated strategies. In *Pharmaceuticals* (Vol. 3, pp. 1456-1490): MDPI AG.

McCall AL, Farhy LS. Treating type 1 diabetes: From strategies for insulin delivery to dual hormonal control. Minerva endocrinologica. 2013; 38: 145-163.

Merten OW, Hebben M, Bovolenta C. (2016). Production of lentiviral vectors. In *Molecular Therapy Methods and Clinical Development* (Vol. 3, pp. 16017-16017): Elsevier Inc.

Milone MC, O'Doherty U. (2018). Clinical use of lentiviral vectors. In *Leukemia* (Vol. 32, pp. 1529-1541): Nature Publishing Group.

Nakai K, Kidera A, Kanehisa M. (1988). *Cluster analysis of amino acid indices for prediction of protein structure and function*. Retrieved from https://academic.oup.com/peds/article-abstract/2/2/93/1485093

Nam SH, Park J, Koo H. Recent advances in selective and targeted drug/gene delivery systems using cell-penetrating peptides. Archives of Pharmacal Research. 2023; 46: 18-34.

Nayerossadat N, Maedeh T, Ali P. Viral and nonviral delivery systems for gene delivery. Advanced Biomedical Research. 2012; 1: 27-27.

Nguyen JM, Jézéquel P, Gillois P, Silva L, Azzouz FB, Lambert-Lacroix S , et al. Random forest of perfect trees: Concept, performance, applications and perspectives. Bioinformatics. 2021; 37: 2165-2174.

Park H, Park JH, Kim MS, Cho K, Shin JM. In silico screening and optimization of cell-penetrating peptides using deep learning methods. Biomolecules. 2023; 13.

Park SY, Kim KH, Kim S, Lee YM, Seol YJ. (2019). Bmp-2 gene delivery-based bone regeneration in dentistry. In *Pharmaceutics* (Vol. 11): MDPI AG.

Pawelczyk E, Arbab AS, Pandit S, Hu E, Frank JA. Expression of transferrin receptor and ferritin following ferumoxides-protamine sulfate labeling of cells: Implications for cellular magnetic resonance imaging. NMR in Biomedicine. 2006; 19: 581-592.

Porosk L, Arukuusk P, Põhako K, Kurrikoff K, Kiisholts K, Padari K , et al. Enhancement of sirna transfection by the optimization of fatty acid length and histidine content in the cpp. Biomaterials science. 2019; 7: 4363-4374.

Porosk L, Põhako K, Arukuusk P, Langel Ü. Cell-penetrating peptides predicted from casc3, akip1, and ahrr proteins. Frontiers in Pharmacology. 2021; 12.

Prabhakaran M. The distribution of physical, chemical and conformational properties in signal and nascent peptides. The Biochemical journal. 1990; 269: 691-696.

Pranatharthiharan S, Patel MD, D'Souza AA, Devarajan PV. Inorganic nanovectors for nucleic acid delivery. Drug Delivery and Translational Research. 2013; 3: 446-470.

Pritišanac I, Vernon RM, Moses AM, Forman Kay JD. Entropy and information within intrinsically disordered protein regions. Entropy. 2019; 21.

Pujals S, Giralt E. (2008). Proline-rich, amphipathic cell-penetrating peptides. In *Advanced Drug Delivery Reviews* (Vol. 60, pp. 473-484).

Quist J, Taylor L, Staaf J, Grigoriadis A. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. 2021.

Răcuciu M, Barbu-Tudoran L, Oancea S, Drăghici O, Morosanu C, Grigoras M , et al. Aspartic acid stabilized iron oxide nanoparticles for biomedical applications. Nanomaterials. 2022; 12.

Ramaker K, Henkel M, Krause T, Röckendorf N, Frey A. Cell penetrating peptides: A comparative transport analysis for 474 sequence motifs. Drug Delivery. 2018; 25: 928-937.

Ramamoorth M, Narvekar A. (2015). Non viral vectors in gene therapy - an overview. In *Journal of Clinical and Diagnostic Research* (Vol. 9, pp. GE01-GE06): Journal of Clinical and Diagnostic Research.

Ranasinghe T, Mays T, Quedado J, Adcock A. Thrombolysis following heparin reversal with protamine sulfate in acute ischemic stroke: Case series and literature review.

Renner R, Vocke K, Hepp KD. Comparison of the effects of protamine-bound nph porcine insulin and nph human insulin in type 1 and 2 diabetes. Deutsche Medizinische Wochenschrift (1946). 1986; 111: 1316-1320.

Rigatti SJ. Random forest. Journal of Insurance Medicine. 2017; 47: 31-39.

Robla S, Varela Calviño R, Ambrus R, Csaba N. A ready-to-use dry powder formulation based on protamine nanocarriers for pulmonary drug delivery. European Journal of Pharmaceutical Sciences. 2023; 185.

Roman-Naranjo P, Parra-Perez AM, Lopez-Escamez JA. (2023). A systematic review on machine learning approaches in the diagnosis and prognosis of rare genetic diseases. In *Journal of Biomedical Informatics* (Vol. 143): Academic Press Inc.

Roy S, Martinez D, Platero H, Lane T, Werner-Washburne M. Exploiting amino acid composition for predicting protein-protein interactions. PLoS ONE. 2009; 4.

Ruseska I, Fresacher K, Petschacher C, Zimmer A. (2021). Use of protamine in nanopharmaceuticals—a review. In *Nanomaterials* (Vol. 11): MDPI AG.

Saczynski JS, McManus DD, Goldberg RJ. (2013). Commonly used data-collection approaches in clinical research. In *American Journal of Medicine* (Vol. 126, pp. 946-950).

Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO. Prediction of cell penetrating peptides by support vector machines. PLoS Computational Biology. 2011; 7.

Saravanan V, Gautham N. Harnessing computational biology for exact linear b-cell epitope prediction: A novel amino acid composition-based feature descriptor. OMICS A Journal of Integrative Biology. 2015; 19: 648-658.

Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I , et al. (2023). Artificial intelligence and machine learning technology driven modern drug discovery and development. In *International Journal of Molecular Sciences* (Vol. 24): MDPI.

Shahryari A, Burtscher I, Nazari Z, Lickert H. (2021). Engineering gene therapy: Advances and barriers. In *Advanced Therapeutics* (Vol. 4): John Wiley and Sons Inc.

Shen X, Jiang C, Wen Y, Li C, Lu Q. A brief review on deep learning applications in genomic studies. Frontiers in Systems Biology. 2022; 2.

Shields RC, Kaspar JR, Lee K, Underhill SAM, Burne RA. Fluorescence tools adapted for real-time monitoring of the behaviors of streptococcus species. Applied and Environmental Microbiology. 2019; 85.

Slade N. (2001). *Viral vectors in gene therapy*. Paper presented at the Periodicum Biologorum.

Soboleski MR, Oaks J, Halford WP. Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells. The FASEB Journal. 2005; 19: 1-20.

Soofiyani SR, Baradaran B, Lotfipour F, Kazemi T, Mohammadnejad L. Gene therapy, early promises, subsequent problems, and recent breakthroughs. Advanced Pharmaceutical Bulletin. 2013; 3: 249-255.

Stokes JV, Moraru GM, McIntosh C, Kummari E, Rausch K, Andrea SVS. Differentiated thp-1 cells exposed to pathogenic and nonpathogenic borrelia species demonstrate minimal differences in production of four inflammatory cytokines. Vector-Borne and Zoonotic Diseases. 2016; 16: 691-695.

Sun Y. Machine learning applied in the financial industry. Financial Forum. 2021; 9: 239-239.

Sung YK, Kim SW. (2019). Recent advances in the development of gene delivery systems. In *Biomaterials Research* (Vol. 23): BioMed Central Ltd.

Szerlauth A, Muráth S, Szilagyi I. Layered double hydroxide-based antioxidant dispersions with high colloidal and functional stability. Soft Matter. 2020; 16: 10518-10527.

Tantoso E, Li KB. Aaindexloc: Predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids. 2008; 35: 345-353.

Tashima T. (2017). Intelligent substance delivery into cells using cell-penetrating peptides. In *Bioorganic and Medicinal Chemistry Letters* (Vol. 27, pp. 121-130): Elsevier Ltd.

Taylor RE, Zahid M. (2020). Cell penetrating peptides, novel vectors for gene therapy. In *Pharmaceutics* (Vol. 12): MDPI AG.

Thomas CE, Ehrhardt A, Kay MA. (2003). Progress and problems with the use of viral vectors for gene therapy. In *Nature Reviews Genetics* (Vol. 4, pp. 346-358).

Timotievich ED, Shilovskiy IP, Khaitov MR. (2023). Cell-penetrating peptides as vehicles for delivery of therapeutic nucleic acids. Mechanisms and application in medicine. In *Biochemistry (Moscow)* (Vol. 88, pp. 1800-1817).

Tomii K, Kanehisa M. (1996). *Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins*. Retrieved from http://www.genome.ad.jp/

Tsuchiya K, Horikoshi K, Fujita M, Hirano M, Miyamoto M, Yokoo H , et al. Development of hydrophobic cell-penetrating stapled peptides as drug carriers. International Journal of Molecular Sciences. 2023; 24.

Vujović Ž. Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications. 2021; 12: 599-606.

Wahlfors J, Loimas S, Pasanen T, Hakkarainen T. Green fluorescent protein (gfp) fusion constructs in gene therapy research. Histochemistry and Cell Biology. 2001; 115: 59-65.

Wang M, Yang T, Flechas MA, Harris P, Hawks B, Holzman B , et al. Gpu-accelerated machine learning inference as a service for computing in neutrino experiments. Frontiers in Big Data. 2021a; 3.

Wang Y, Zhang Z, Luo J, Han X, Wei Y, Wei X. (2021b). Mrna vaccine: A potential therapeutic strategy. In *Molecular Cancer* (Vol. 20): BioMed Central Ltd.

Wei L, Tang J, Zou Q. Skipcpp-pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics. 2017a; 18.

Wei L, Xing P, Su R, Shi G, Sam Ma Z, Zou Q , et al. (2017b). *Cppred-rf: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency*. Retrieved from http://server.malab.cn/CPPred-RF.

Weide B, Pascolo S, Scheel B, Derhovanessian E, Pflugfelder A, Eigentler TK , et al. (2009). *Direct injection of protamine-protected mrna: Results of a phase 1/2 vaccination trial in metastatic melanoma patients*. Retrieved from www.immunotherapy-journal.com

Wu J, Jones N, Fayez NAL, Chao PH, Wu A, de Araujo DR , et al. Protamine-mediated efficient transcellular and transmucosal delivery of proteins. Journal of Controlled Release. 2023; 356: 373-385.

Xu J, Khan AR, Fu M, Wang R, Ji J, Zhai G. (2019). Cell-penetrating peptide: A means of breaking through the physiological barriers of different tissues and organs. In *Journal of Controlled Release* (Vol. 309, pp. 106-124): Elsevier B.V.

Yang J, Gong J, Tang W, Shen Y, Liu C, Gao J. Delineation of urban growth boundaries using a patch-based cellular automata model under multiple spatial and socio-economic scenarios. Sustainability (Switzerland). 2019; 11.

Ye J, Liu E, Gong J, Wang J, Huang Y, He H , et al. High-yield synthesis of monomeric lmwp (cpp)-sirna covalent conjugate for effective cytosolic delivery of sirna. Theranostics. 2017; 7: 2495.

Ying X. (2019, 2019/3//). *An overview of overfitting and its solutions.* Paper presented at the Journal of Physics: Conference Series.

Yu S, Yang H, Li T, Pan H, Ren S, Luo G , et al. Efficient intracellular delivery of proteins by a multifunctional chimaeric peptide in vitro and in vivo. Nature Communications. 2021; 12.

Yu X, Hou J, Shi Y, Su C, Zhao L. Preparation and characterization of novel chitosan–protamine nanoparticles for nucleus-targeted anticancer drug delivery. International Journal of Nanomedicine. 2016; 11: 6035-6046.

Zahid M, Robbins PD. (2015). Cell-type specific penetrating peptides: Therapeutic promises and challenges. In *Molecules* (Vol. 20, pp. 13055-13070): MDPI AG.

Zarei H, Malaekeh-Nikouei B, Ramezani M, Soltani F. Multifunctional peptides based on low molecular weight protamine (lmwp) in the structure of polyplexes and lipopolyplexes: Design, preparation and gene delivery characterization. Journal of Drug Delivery Science and Technology. 2021; 62.

Zhang H, Qin C, An C, Zheng X, Wen S, Chen W , et al. (2021). Application of the crispr/cas9-based gene editing technique in basic research, diagnosis, and therapy of cancer. In *Molecular Cancer* (Vol. 20): BioMed Central Ltd.

Zhou J, Gandomi AH, Chen F, Holzinger A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. In *Electronics (Switzerland)* (Vol. 10, pp. 1-19): MDPI AG.

Zhu X, Shan W, Zhang P, Jin Y, Guan S, Fan T , et al. Penetratin derivative-based nanocomplexes for enhanced intestinal insulin delivery. Molecular Pharmaceutics. 2014; 11: 317-328.

Zhu Z, Tian D, Gao P, Wang K, Li Y, Shu X , et al. Cell-penetrating peptides transport noncovalently linked thermally activated delayed fluorescence nanoparticles for time-resolved luminescence imaging. Journal of the American Chemical Society. 2018; 140: 17484-17491.

Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V , et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nature Communications. 2020; 11.

RESUME

Personal Information

Name	Elif Çisem	**Race**	
Surname	ÇÖLDÜR	**Tel no**	
Date of Birth		**e-mail**	

Educational Information

Institution		**Graduation Date**
High School	Uşak Fen Lisesi	2014
University	Yeditepe University, Department of Pharmacy (Major)	2020
University	Yeditepe University, Department of Genetics and Bioengineering (Minor)	2020

Work Experience

Mission	**Organisation**	**Duration**
Intern	Lithuanian University of Health of Sciences	2020
Intern	Yeditepe University Toxicology Department	2019-2020

Foreign Languages	**Name of the Exam**	**Score**
English	IELTS	6.5
English	YÖKDİL	85

Achievements & Poster/Oral Presentations

- Championship award with the project “Developing and Optimizing Non-viral Gene Carrier System Formulations That Can Be Used in Breast Cancer Gene Therapy Using Artificial Intelligence” in the IAMGENIUS competition organized by Amgen in Turkey (2021)
- Oral Presentation in BIO Turkey – International Biotechnology Congress with the project “Single Guide RNA (sgRNA) Design and Off-target Determination for the Multidrug Resistance Gene-PDR1 from Saccharomyces cerevisiae with crispRdesignR Bioinformatic Application (2020)
- Poster presentation about "Formulation of Progestogen Only Vaginal Tampon with 3D Printer" project in congress of ACUGEN (2018)

- Championship award with "Formulation of Progestogen Only Vaginal Tampon with 3D Printer" in Alfa Rx 2018 Competition
- Championship award with "Pharmarea" project in Farmafikir 2018 Competition

List of Publications

- R. REİS, E. ÇÖLDÜR, Public Health Impacts of Exposure to Disinfectants, Therapeutics, and Illicit Substances During the COVID-19 Pandemic, Hacettepe Univ. J. Fac. Pharm. 42 (2022) 60–71. https://doi.org/10.52794/hujpharm.978727.
- Çöldür, E.Ç., and Dora, D.D. (2021, September) Single Guide RNA (sgRNA) Design and Off-target Determination for the Multidrug Resistance Gene-PDR1 from Saccharomyces cerevisiae with crispRdesignR Bioinformatic Application. Paper presented at the International Biotechnology Congress. Available from: https://bioturkiye.org/bioData/Uploads/files/bildiri_kitabi.pdf
- Çöldür, E. Ç., The Role of CRISPR-Cas9 in Cancer Treatment. Bezelye Journal, 12th issue,58-59. Available from: https://www.bezelyedergi.net/12-say%C4%B1
- Soyer Can, O., Ünlü, S., Ocak, H., Çöldür, E. Ç., Sipahi, H., & Bilgin Eran, B. (2022). Synthesis of new imidazothiazole derivatives and investigation of their anti-inflammatory and analgesic activities. Journal of the Iranian Chemical Society, 19(2), 579-587.
- Soyer Can, O., Ünlü, S., Ocak, H., Çöldür, E. Ç., Sipahi, H., & Bilgin Eran, B. (2022). Synthesis of new imidazothiazole derivatives and investigation of their anti-inflammatory and analgesic activities. Journal of the Iranian Chemical Society, 19(2), 579-587.