

CAR Architecture Optimization Using Strategical Motif Combinations and Predictive Modeling.



By

Zill E Noor

(Registration No: 000000402170)

Department of Bioinformatics

School of Interdisciplinary Engineering and Sciences

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

CAR Architecture Optimization Using Strategical Motif Combinations and Predictive Modeling.



By

Zill E Noor

(Registration No: 00000402170)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

**Master of Science in
Bioinformatics**

Supervisor: Dr. Mehak Rafiq

School of Interdisciplinary Engineering & Sciences (SINES)


National University of Sciences & Technology (NUST)

Islamabad, Pakistan

August 2024

THESIS ACCEPTANCE CERTIFICATE


Certified that final copy of MS Thesis written by Mr / Ms Zill e Noor
(Registration No. 00000402170), of SINES, NUST (School/College/Institute)
has been vetted by undersigned, found complete in all respects as per NUST Statutes/
Regulations/ Masters Policy, is free of plagiarism, errors, and mistakes and is accepted as
partial fulfillment for award of Master's degree. It is further certified that necessary
amendments as pointed out by GEC members and evaluators of the scholar have also been
incorporated in the said thesis.

Signature: 


Name of Supervisor Dr Mehabe Kalia

Date: 29-8-24

for HOD sci

Signature (HOD): 

Date: 29/08/2024

Signature (Dean/ Principal) 

Date: 29/08/2024

AUTHOR'S DECLARATION

IZill E Noor hereby state that my MS thesis titled “**CAR Architecture Optimization using Strategical Motif Combinations and Predictive Modeling**” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Zill E Noor

Date: 19-08-2024

DEDICATION

I dedicate this thesis to my beloved parents and siblings whose prayers, love, encouragement, and endless support made my success possible, and to my teachers, whose help and guidance helped me throughout my academic journey.

ACKNOWLEDGEMENTS

First of all, I am deeply grateful to **Allah Almighty**, whose countless blessings and guidance have brought me to this point in my educational journey. None of this would have been possible without His grace and mercy. I would like to express my deepest gratitude to **my Parents**, whose unwavering support and endless sacrifices have been my pillars of strength. Their encouragement and belief in me have always been the driving force behind my achievements. I am also extremely grateful to all my siblings, especially my brother **Muhammad Umair Akbar** and my sister **Rafaila Sameen** for their constant encouragement and for always being with me.

I want to express my sincere gratitude to my supervisor **Dr. Mehak Rafiq**. Your guidance, patience, and invaluable guidance have been of great help in completing this thesis. Your encouragement and belief in my abilities have helped me through the challenges of this research.

I am deeply thankful to my GEC members, **Dr. Masood Ur Rahman Kayani** and **Dr. Salma Sherbaz**, for their insightful guidance and constructive feedback. Your expertise and advice have helped shape this thesis and guided me through the research process.

My friend **Iqra Asif**, **Muhammad Inam Rafique** and **Maria Hassan Khan** also deserve special recognition for their continuous support and company during this journey. Your support and positivity have given me lots of comfort and inspiration.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
TABLE OF CONTENTS	2
LIST OF TABLES	4
LIST OF FIGURES	5
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS	7
ABSTRACT	8
CHAPTER 1 : INTRODUCTION	9
1.1 Structural Components of Chimeric Antigen Receptors	10
1.2 Evolution of CAR Technology	12
1.3 Mechanism of Action of CAR T Cell Therapy	17
1.4 Advancements in CAR T Therapy and role of Machine Learning	19
1.5 Predictive Modeling in CAR T Cell Therapy	20
1.6 Scope of the Study	21
1.7 Problem Statement	21
1.8 Objectives	22
CHAPTER 2 : LITERATURE REVIEW	23
2.1 The Role of Machine Learning in Cancer Diagnosis and Treatment	23
2.1.1 Diagnosis and Classification	23
2.1.2 Treatment of Cancer	24
2.1.3 Manufacturing of CARs	26
2.1.4 Personalized care for cancer treatment	26
2.1.5 Predicting Remission	27
CHAPTER 3 : MATERIALS AND METHODS	29
3.1 Data Acquisition	29
3.2 Data Preprocessing	29

3.3 Python Libraries Used	30
3.4 Regression Models	31
3.5 Comparative Analysis	34
3.6 Transformer Model	35
CHAPTER 4 : RESULTS AND DISCUSSION	37
4.1 Data Acquisition and Quality Assessment	37
4.2 Preprocessing and Feature Engineering	37
4.3 Model Implementation and Results	38
4.3.1 Random Forest	38
4.3.2 Support Vector Regression	40
4.3.3 Neural Network Regression	43
4.3.4 Linear Regression	44
4.3.5 Decision Tree	47
4.3.6 Gradient Boosting	48
4.4 Comparative Analysis	51
4.5 Model Predicting Cytotoxicity	54
4.5.1 Model Training and Validation Performance	54
4.5.2 Predictive Accuracy on Sequences	56
CHAPTER 5 : CONCLUSION AND FUTURE DIRECTIONS	58
REFERENCES	63

LIST OF TABLES

Page No.

Table 4.1 Performance and resource comparison of various machine learning models for Cytotoxicity.	52
Table 4.2 Performance and resource comparison of various machine learning models for stemness.	53
Table 4.3: Training and Validation Losses Across Epochs for the Transformer Model Applied to Cytotoxicity Prediction.	55
Table 4.4: Comparison of predicted and actual cytotoxicity values for various peptide sequences.	56

LIST OF FIGURES

	Page No.
Figure 1.1 Layout of the CAR structure. CARs have four major domains: Antigen-Binding domain, Hinge region, transmembrane domain, and signaling domain.....	10
Figure 1.2 Diagram of a First Generation CAR highlighting its components and limitations	12
Figure 1.3 Diagram of a second-generation CAR, highlighting its enhanced characteristics.....	13
Figure 1.4 Diagram of a third-generation CAR, highlighting its stronger characteristics and additional domains for enhanced functionality.	14
Figure 1.5 Diagram of a fourth-generation CAR, illustrating its enhanced functionality with additional domains and the secretion of transgenic cytokines for improved immune response.	15
Figure 1.6: Diagram of a fifth-generation CAR, illustrating its enhanced functionality with additional IL-2 receptor and JAK-STAT signaling domains.....	16
Figure 1.7: Illustration of the mechanism of action of CAR-T cells therapy	17
Figure 4.1: Scatter plot comparing actual vs. predicted cytotoxicity using the Random Forest model. The red dashed line represents perfect predictions, with green points showing that the model generally performs well, closely following the ideal prediction line.....	39
Figure 4.2: Scatter plot comparing actual vs. predicted stemness using the Random Forest model. The red dashed line represents perfect predictions, with blue points indicating that the model's predictions closely follow the ideal line, demonstrating good performance.	40
Figure 4.3: Scatter plot comparing actual vs. predicted cytotoxicity using the Support Vector Model. The red dashed line represents perfect predictions, with green points showing that the model generally performs well, closely aligning with the idea.....	41

Figure 4.4: Scatter plot comparing actual vs. predicted stemness using the Support Vector Model, where each point represents an individual data sample, indicating how closely the model's predictions align with the actual stemness values.42

Figure 4.5: Scatter plot comparing actual vs. predicted cytotoxicity using the Neural Network Model, where each point represents a data sample, illustrating how well the model's predictions align with the actual cytotoxicity values.....43

Figure 4.6: Scatter plot comparing actual vs. predicted stemness using the Neural Network Model, where each point represents a data sample, indicating the alignment between the model's predictions and the actual stemness values.44

Figure 4.7: Scatter plot comparing actual vs. predicted cytotoxicity using the Linear Regression Model. Each point represents a data sample, with the red dashed line indicating perfect predictions; the model's accuracy is reflected in how closely the points align.....45

Figure 4.8: Scatter plot comparing actual vs. predicted stemness using the Linear Regression Model. Each point represents a data sample, with the red dashed line indicating perfect predictions; the model's accuracy is reflected in how closely the points align.....46

Figure 4.9: Scatter plot comparing actual vs. predicted cytotoxicity using the Neural Network Model. The green points represent individual data samples, with the red dashed line indicating perfect predictions.....47

Figure 4.10: Scatter plot comparing actual vs. predicted stemness using the Decision Tree Model. Each blue point represents a data sample, with the red dashed line indicating perfect predictions.....48

Figure 4.11: Scatter plot comparing actual vs. predicted cytotoxicity using the Gradient Boost Model, with each point representing a data sample. The model's predictions closely align with the ideal prediction line.49

Figure 4.12: Scatter plot comparing actual vs. predicted stemness using the Gradient Boost Model, with each point representing a data sample. The model's predictions closely align with the ideal prediction line.50

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

CAR	Chimeric Antigen Receptor
MHC	Major Histocompatibility Complex
CGS	Center for Graduate Studies
scFv	Single-chain Variable Fragment
CRS	Cytokine Release Syndrome
K	Kappa
IL	Interleukin
TCRs	T cell receptors
ML	Machine Learning
AI	Artificial Intelligence
HLA	Human Leukocyte Antigen
MSE	MSE
R ²	R-Squared value

ABSTRACT

CAR T cell therapy has emerged as a promising approach for treating various forms of cancer. Despite its success, the effectiveness of CAR T cell therapy can be influenced by the specific configurations of signaling motifs within the CAR architecture. This study focuses on optimizing CAR architecture using strategic motif combinations and predictive modeling to enhance therapeutic outcomes. The research utilized a dataset of CAR T cell configurations, each characterized by different motif combinations, to explore the impact on cytotoxicity and stemness. Various machine learning models, including Random Forest, Support Vector Machine, Neural Networks, Linear Regression, Decision Tree, and Gradient Boosting, were employed to predict cytotoxicity and stemness based on these configurations. The transformer-based model was also implemented to predict cytotoxicity from protein sequences, showcasing the potential for integrating generative AI models into the healthcare domain. The results demonstrated that traditional machine learning models, such as Decision Tree and Gradient Boosting, effectively captured key features of cytotoxicity and stemness, particularly at moderate levels, and did so with significantly less computational power and time compared to more complex models like CNNs and LSTMs. This highlights the first major objective of the research: to show that machine learning models can achieve comparable performance to CNNs and LSTMs in feature extraction, while being more efficient in terms of computational resources. This study contributes to the growing field of CAR T cell therapy by providing a detailed analysis of how different motif combinations influence therapeutic outcomes. The research offers key insights that can lead to the refinement of CAR designs, making them more effective and safer. These findings support the development of improved CAR T cell therapies and pave the way for personalized treatment strategies that use predictive modeling to tailor interventions to individual patients.

CHAPTER 1 : INTRODUCTION

Chimeric Antigen Receptors (CAR) are highly specialized, engineered receptor proteins designed to enhance the capability of immune cells, particularly T cells, to identify and eliminate cancer cells [1]. Unlike naturally occurring receptors, CARs are synthesized in laboratories by combining different protein components. The primary function of these engineered receptors is to enable T cells to detect specific markers, known as antigens, present on the surface of cancer cells [2]. This precise targeting mechanism is crucial, as it allows the immune system to differentiate between malignant and healthy cells, thereby directing a focused immune response exactly where it is needed. Upon encountering the target antigen on a cancer cell, CAR-modified T cells become activated and launch a potent attack, leading to the destruction of the cancerous cells [3].

CAR T cells represent a unique and innovative form of immune cell therapy that differs significantly from conventional effector T cells. One of the key distinctions is that CAR T cells can recognize and bind to antigens on cancer cells independently of Major Histocompatibility Complex (MHC) presentation [4]. This capability allows CAR T cells to target cancer cells based solely on specific surface antigens, significantly expanding their therapeutic applicability across various cancer types. This independence from MHC restriction makes CAR T cells versatile in targeting tumors with diverse antigen presentations, potentially increasing the effectiveness of treatments [5]. Furthermore, CAR T cell therapy is typically generated using an autologous approach, wherein the patient's own cells are harvested, genetically modified to express CARs, and expanded in large numbers before being reinfused [6].

This method not only enhances treatment specificity and reduces the risk of immune rejection but also personalizes therapy to the unique cancer profile of each patient, offering a promising and highly targeted approach to cancer treatment [7]. The traditional design of CARs has utilized signaling domains derived from native immune receptors such as CD28 and 4-1BB, which provide essential signals for T-cell activation and survival [8]. However, there is growing recognition that the design space for CAR

signaling domains is vast and largely unexplored. The challenge lies in engineering CARs that not only effectively recognize and kill cancer cells but also maintain a stem-like phenotype for long-term persistence and memory formation [9].

1.1 Structural Components of Chimeric Antigen Receptors

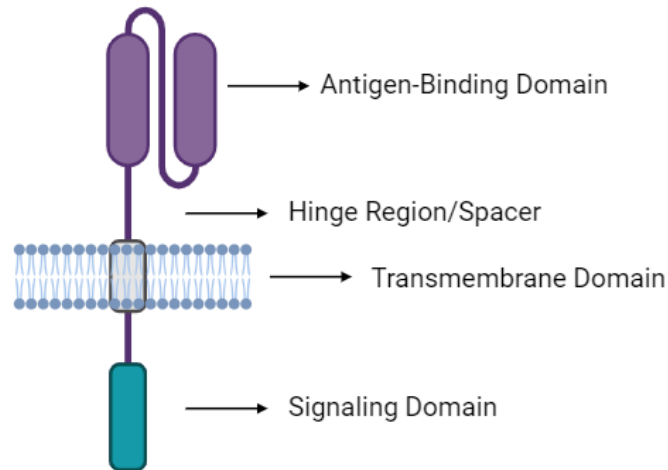


Figure 1.1 Layout of the CAR structure. CARs have four major domains: Antigen-Binding domain, Hinge region, transmembrane domain, and signaling domain.

- **Antigen-Binding Domain:** The antigen-binding domain of CARs, typically composed of a single-chain variable fragment (scFv), is derived from monoclonal antibodies. This domain is crucial for recognizing and binding to specific antigens present on the surface of cancer cells [10]. The flexibility in designing scFvs allows for the targeting of a diverse range of tumor-associated antigens, including well-known markers like CD19, HER2, and EGFRvIII. This adaptability is vital for customizing CAR-T cells to address various types of cancers [11].
- **Spacer/Linker Region:** The spacer, or hinge region, serves as a connector between the scFv and the transmembrane domain [12]. The design of this region, including its length and flexibility, plays a significant role in the CAR's ability to effectively engage with antigens and transmit signals. Optimizing the spacer is particularly important for CAR-T cells targeting solid tumors, where

antigen accessibility can be hindered by the tumor microenvironment [13]. An appropriately designed spacer enhances the efficacy of the CAR-T cells by facilitating better interaction with the target antigens [12].

- **Transmembrane Domain:** The transmembrane domain is responsible for anchoring the CAR to the T cell membrane [4]. It also affects the stability and expression levels of the receptor on the cell surface. Common transmembrane domains used in CAR construction are derived from molecules like CD8, CD28, and CD3 ζ . The choice of transmembrane domain can influence the overall functionality and durability of the CAR on the T cell surface [14].
- **Intracellular Signaling Domain:** The intracellular signaling domains are essential for initiating and sustaining T cell activation [15]. The CD3 ζ chain is a fundamental component, responsible for triggering the initial activation signal within the T cell [16]. In addition, co-stimulatory molecules such as CD28 and 4-1BB are often included to enhance various aspects of T cell functionality, including proliferation, survival, and cytokine production. The selection and combination of these intracellular domains are critical, as they can significantly impact the potency and therapeutic potential of CAR-T cells, influencing their effectiveness in eradicating cancer cells [17].

1.2 Evolution of CAR Technology

The development of CAR has undergone significant advancements, progressing through multiple generations, each aimed at overcoming the limitations of the previous iterations:

1. **First-Generation CARs:** The earliest CARs consisted of a scFv linked to the CD3 ζ signaling domain. While these CARs could activate T cells upon binding to their target antigen, they demonstrated limited clinical efficacy. This was primarily due to suboptimal T cell activation and insufficient persistence, which restricted their therapeutic potential [18].

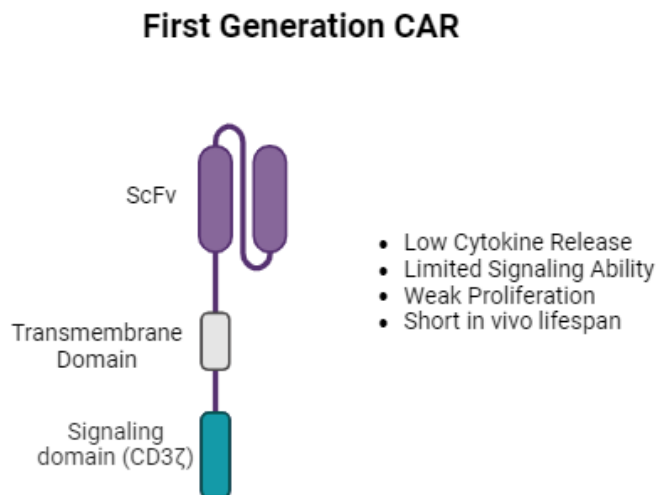


Figure 1.2 Diagram of a First-Generation CAR highlighting its components and limitations

2. **Second-Generation CARs:** To address the shortcomings of first-generation CARs, second-generation designs incorporated an additional co-stimulatory domain, such as CD28 or 4-1BB, alongside the CD3 ζ domain. This addition significantly enhanced T cell proliferation, survival, and cytotoxic activity, leading to improved clinical outcomes. The inclusion of these co-stimulatory molecules provided a more robust and sustained immune response against cancer cells [18].

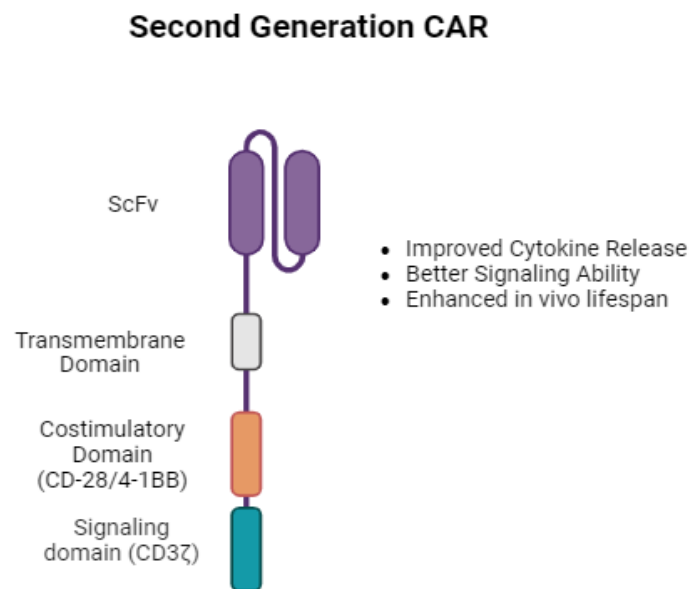


Figure 1.3 Diagram of a second-generation CAR, highlighting its enhanced characteristics

3. **Third-Generation CARs:** These CARs further evolved by integrating multiple co-stimulatory domains, typically combining CD28 and 4-1BB. The rationale behind this design was to harness the synergistic benefits of different co-stimulatory domains, thereby maximizing the anti-tumor efficacy of the CAR-T cells. This generation aimed to provide a more potent and durable immune response [19].

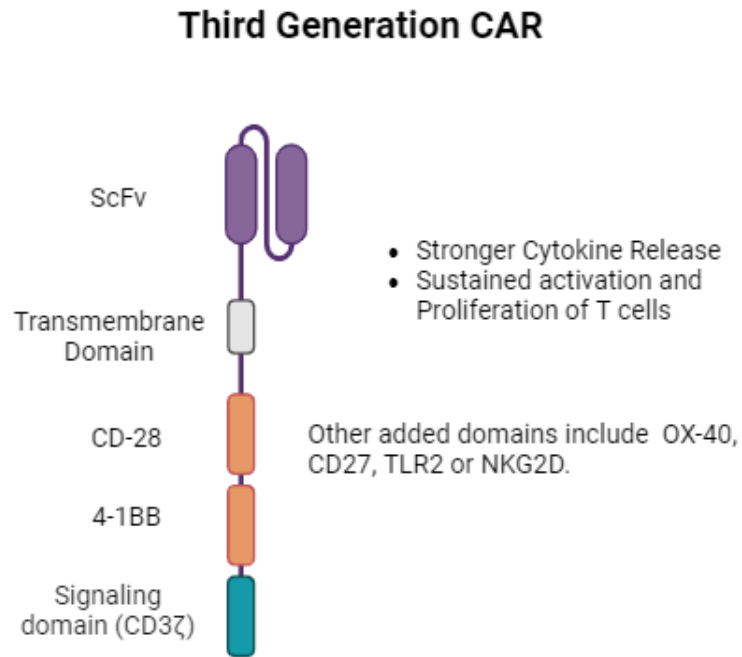


Figure 1.4 Diagram of a third-generation CAR, highlighting its stronger characteristics and additional domains for enhanced functionality.

4. **Fourth-Generation CARs:** Fourth-generation CARs also known as T cells Redirected for Universal Cytokine-mediated Killing (TRUCKs), introduced additional genetic modifications that enable the CAR-T cells to secrete pro-inflammatory cytokines, such as IL-12, upon activation. This innovation is intended to enhance the local immune response within the tumor microenvironment, potentially overcoming the immune-suppressive conditions that often hinder effective tumor eradication [20].

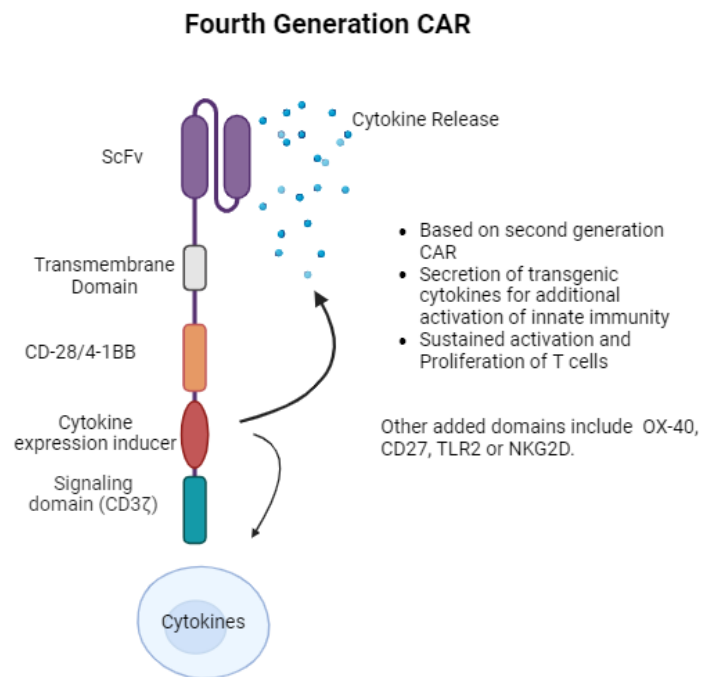


Figure 1.5 Diagram of a fourth-generation CAR, illustrating its enhanced functionality with additional domains and the secretion of transgenic cytokines for improved immune response.

5. **Fifth-Generation CARs:** The latest advancements in CAR technology involve the integration of cytokine signaling domains, such as IL-2R β , coupled with a STAT3/5 binding domain. These modifications aim to augment the persistence and functionality of CAR-T cells through both autocrine and paracrine signaling mechanisms. Fifth-generation CARs are designed to enhance the overall therapeutic durability and effectiveness of CAR-T cell therapy, offering promising potential for more robust and long-lasting cancer treatment responses [21].

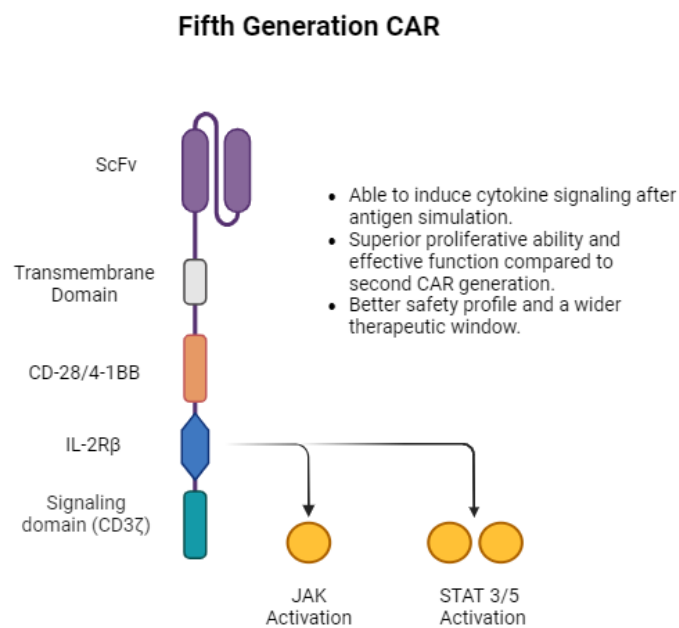


Figure 1.6: Diagram of a fifth-generation CAR, illustrating its enhanced functionality with additional IL-2 receptor and JAK-STAT signaling domains.

Each generation of CARs has contributed to improving the effectiveness, specificity, and safety of CAR-T cell therapies, bringing new hope for treating various forms of cancer. The ongoing evolution of CAR technology continues to push the boundaries of what is possible in the field of cancer immunotherapy.

1.3 Mechanism of Action of CAR T Cell Therapy

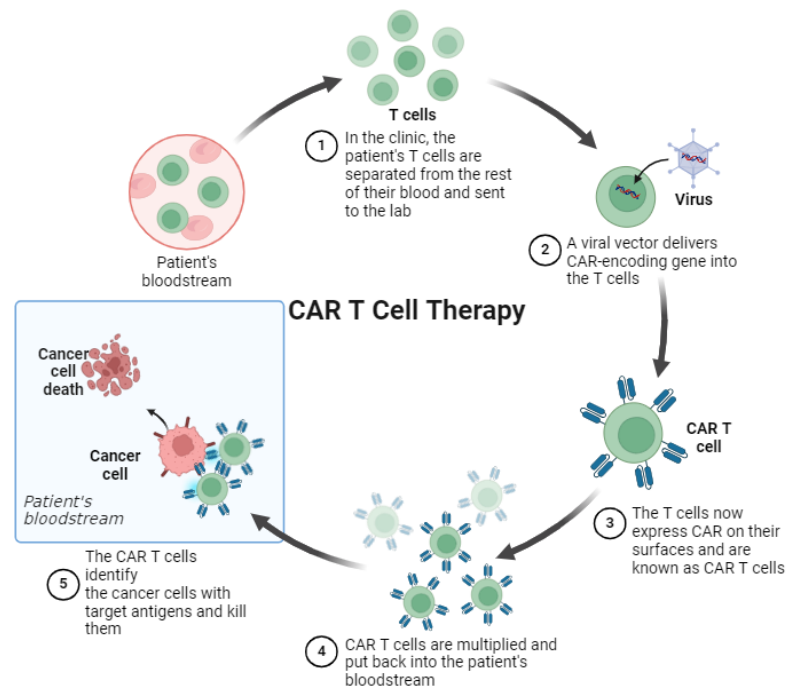


Figure 1.7: Illustration of the mechanism of action of CAR-T cells therapy

1. T Cell Collection: The process of CAR T cell therapy begins with the collection of T cells from the patient's blood. This is accomplished through a procedure called leukapheresis, where blood is drawn from the patient and passed through a machine that separates out the T cells from the rest of the blood components. The T cells, which are a crucial component of the immune system, are then collected for further processing [22].

2. Genetic Modification: The collected T cells are transported to the laboratory where they undergo genetic modification. In this step, scientists use a viral vector, often derived from a retrovirus or lentivirus, to introduce a gene into the T cells. This gene encodes a CAR a synthetic receptor designed to target a specific antigen found on the surface of cancer cells. The antigen-binding domain is engineered to recognize a specific protein expressed on cancer cells, such as CD19 in B-cell malignancies [23].

3. T Cell Expansion: After the T cells are successfully modified to express the CAR, they are expanded in culture. This involves growing the CAR T cells in a controlled

laboratory environment to produce a large number of cells. This expansion is necessary to generate enough CAR T cells to have a therapeutic effect when reintroduced into the patient. The expansion process can take several days to weeks, depending on the required dose [23].

4. Patient Conditioning: Prior to the infusion of the CAR T cells, the patient typically undergoes a conditioning regimen, which often includes chemotherapy. This treatment serves two primary purposes: it helps to reduce the patient's existing immune cells, thereby minimizing competition and creating space for the newly infused CAR T cells to proliferate, and it also helps to modulate the immune environment to enhance the efficacy of the therapy [24].

5. Infusion and Target Recognition: Once the CAR T cells are ready, they are infused back into the patient's bloodstream through an intravenous infusion. These CAR T cells then circulate throughout the body, searching for cells that express the target antigen [25]. The antigen-binding domain of the CAR enables the T cells to recognize and bind to the specific antigens present on the surface of cancer cells [26].

6. Activation and Tumor Cell Killing: Upon binding to the target antigen on a cancer cell, the CAR transmits an activation signal to the T cell through its intracellular signaling domains. This activation triggers the T cell to exert its cytotoxic functions [27]. The CAR T cells release cytotoxic molecules such as perforin and granzymes, which penetrate the cancer cell membrane and induce apoptosis, or programmed cell death, thereby killing the cancer cells [28].

7. Cytokine Release and Immune Modulation: In addition to direct killing of cancer cells, CAR T cells also release cytokines, which are signaling molecules that can enhance the immune response. Some CAR T cells are engineered to secrete specific cytokines like interleukin-12, which can recruit and activate other immune cells, further enhancing the anti-tumor response and modifying the tumor microenvironment to make it less conducive to cancer cell survival [29].

8. Persistence and Memory Formation: After the initial therapeutic effect, a subset of the infused CAR T cells can persist in the patient's body and develop into memory T cells. These memory CAR T cells provide long-term immunological surveillance,

potentially preventing cancer recurrence by recognizing and eliminating any residual or emerging cancer cells. This persistent immune response is a key feature that contributes to the potential for durable remissions in patients treated with CAR T cell therapy [30].

1.4 Advancements in CAR T Therapy and role of Machine Learning

CAR T cell therapy is undergoing significant advancements aimed at improving its efficacy, safety, and applicability. One major area of focus is the treatment of solid tumors. Unlike blood cancers, solid tumors present challenges such as heterogeneous antigen expression, a dense and suppressive tumor microenvironment, and immune evasion mechanisms. Researchers are working to identify novel target antigens specific to solid tumors, develop CARs capable of better penetration and persistence within these environments, and engineer CAR T cells that can resist immunosuppressive signals from the tumor microenvironment [31].

Enhancing the persistence and durability of CAR T cells is another critical area of development. The long-term efficacy of CAR T cell therapy depends on the sustained presence and functionality of these cells within the patient [32]. To address this, scientists are optimizing the co-stimulatory domains within CAR constructs, employing gene editing technologies like CRISPR to improve T cell fitness, and incorporating cytokine signaling domains, such as IL-2R β , to support the survival and proliferation of CAR T cells [33].

Reducing the toxicities associated with CAR T cell therapy, such as CRS and neurotoxicity, is also a priority. Innovations include designing safer CAR constructs with "safety switches" that can deactivate or eliminate the CAR T cells if severe adverse effects occur [34]. Additionally, by optimizing the affinity of CARs for their target antigens and exploring dual-targeting CARs that require the recognition of two different antigens, researchers aim to minimize off-target effects and enhance safety [35].

The development of universal or allogeneic CAR T cells is a promising area aimed at making the therapy more accessible and cost-effective. Unlike autologous CAR T cell therapy, which uses the patient's own cells, allogeneic CAR T cells are derived from

healthy donors and are engineered to be universally applicable, reducing production time and costs. This approach involves modifications to prevent rejection by the patient's immune system and to ensure a consistent therapeutic product [36].

Machine Learning (ML) and Artificial Intelligence (AI) are increasingly instrumental in advancing CAR T cell therapy. ML algorithms help in identifying new antigen targets by analyzing extensive datasets, including genetic and proteomic information, to prioritize antigens that are highly expressed on cancer cells but minimally on healthy tissues [37]. ML models, including techniques like particle swarm optimization, principal component analysis, and gene set enrichment analysis, have been incorporated into CAR-T cell research to identify key cell-intrinsic factors such as proliferative capacity, death rates, and cytotoxicity potential that outperform traditional phenotypic markers in predictive accuracy [38]. Furthermore, machine learning and artificial intelligence are being utilized to optimize CAR T cell therapies by predicting cytotoxicity and stemness, thereby enhancing the precision and effectiveness of these treatments. Neural networks, specifically trained to interpret the combinatorial patterns of CAR signaling motifs, have enabled the identification of crucial design principles, further advancing the development of more effective CAR T cell therapies [39].

1.5 Predictive Modeling in CAR T Cell Therapy

Predictive modeling, a sophisticated application of machine learning, is significantly enhancing the development and application of CAR T cell therapy. This technology utilizes statistical algorithms to predict outcomes based on extensive datasets, providing critical insights that improve various aspects of CAR T therapy [40].

Moreover, predictive modeling plays a crucial role in personalizing CAR T cell therapy. By incorporating patient-specific data such as genetic markers, tumor characteristics, and past treatment responses, these models can predict how individual patients are likely to respond to the therapy. This personalized approach enables clinicians to tailor treatment plans, adjust dosages, and select the most appropriate CAR constructs, significantly enhancing therapeutic outcomes and reducing the likelihood of adverse effects [41].

The management of potential toxicities, such as CRS and neurotoxicity, is another critical area where predictive modeling proves invaluable. By assessing pre-treatment patient data, predictive models can estimate the risk of these side effects, allowing for proactive management strategies and rapid intervention if needed. This capability enhances patient safety and improves the overall therapeutic experience [41].

Furthermore, predictive modeling enhances the ability to monitor and forecast treatment outcomes after infusion, providing critical insights that refine risk stratification and guide personalized treatment decisions, thereby optimizing the overall therapeutic strategy [42].

Beyond clinical applications, predictive modeling is instrumental in research and development. By analyzing complex data from preclinical and clinical studies, these models help identify novel therapeutic targets, optimize CAR T cell constructs, and forecast the potential success of new therapies. This accelerates the development pipeline, bringing innovative treatments to patients more efficiently [38].

1.6 Scope of the Study

The scope of this study is to optimize the architecture of CARs through strategic motif combinations and predictive modeling, with the aim of enhancing the efficacy of CAR T-cell therapies in cancer treatment. This research employs various machine learning models, including Random Forest, Support Vector Machine, Neural Networks, Decision Tree and Gradient Boosting, to predict cytotoxicity and stemness based on different CAR configurations. Moreover, the study delves into predicting cytotoxicity based on motif sequence using a transformer model, highlighting the integration of advanced AI techniques in the field of cancer immunotherapy.

1.7 Problem Statement

In bioinformatics, deep learning models such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) have shown the potential to achieve high predictive accuracy. However, these models are computationally intensive and time-consuming which poses a significant challenge for predictive tasks where

efficient processing is crucial. Therefore, there is a critical need to explore alternative machine learning approaches that balance accuracy with computational efficiency, enabling the effective prediction of key therapeutic outcomes.

1.8 Objectives

The objectives of the study are as follows:

- To develop a model to observe the recombination of motifs on phenotypes of CAR-T cells.
- To conduct comparative analysis to find the best model to predict cytotoxicity by motif combinations.
- To develop a model using generative AI to predict the cytotoxicity based on motif sequences.

CHAPTER 2 : LITERATURE REVIEW

The previous few decades have seen significant advancements in cancer treatment, which are mostly attributable to technological advancements. One of the most significant developments in recent years has been the application of ML and AI in cancer care. These technologies have improved diagnosis, treatment, and drug discovery by enhancing precision, personalization, and efficiency. Moreover, AI and ML are accelerating the discovery of new therapeutics, offering hope for more effective and less invasive cancer treatments in the future.

2.1 The Role of Machine Learning in Cancer Diagnosis and Treatment

2.1.1 *Diagnosis and Classification*

Digital pathology has been one of the most impacted areas by AI and ML. Traditionally, cancer diagnosis through histopathology has relied on the expertise of pathologists, who manually examine tissue samples under a microscope. However, this process is time consuming and prone to human error, especially in complex cases. AI and ML have introduced a new paradigm in digital pathology by automating and enhancing the accuracy of image analysis [43].

Deep learning models have been trained on vast datasets of digitized histopathological images, enabling them to identify and quantify morphological features of tissues, such as tumor cells, with remarkable accuracy, often outperforming human experts. For instance, AI algorithms have successfully detected invasive carcinoma in breast cancer slides, even in cases with complex morphological patterns. Moreover, AI can help grade tumors by assessing histological features, such as nuclear atypia and mitotic count, which are important for determining cancer aggressiveness. These advancements not only improve diagnostic accuracy but also significantly reduce the time required for pathology reports, allowing for faster initiation of treatment [43]. Additionally, this image analysis approach is being applied to diagnose several other cancers, including lung cancer [44], head and neck cancer [45], and more, showcasing its versatility and effectiveness across various cancer types.

Furthermore, ML was able to classify genetic mutations, a cornerstone for developing targeted cancer therapies. By analyzing complex genomic datasets, ML models have greatly improved the identification and classification of mutations that drive cancer progression. This enhanced accuracy accelerates the development process and lessens the manual work required to interpret genetic data, resulting in more efficient and personalized treatment regimens. Thus, patients receive more focused, effective, and efficient cancer therapy, highlighting the revolutionary role that AI and ML have played in modern oncology [46].

2.1.2 Treatment of Cancer

AI and ML are transforming the landscape of cancer treatment by enhancing immunotherapy [47]. It plays a pivotal role in predicting neoantigens. Large multi-omics datasets have been evaluated using ML algorithms to improve the discovery of tumor-specific neoantigens, which is critical for creating tailored cancer treatments. These methods enable accurate predictions of neoantigen binding to MHC molecules, a vital step in the immune response. Also, ML algorithms are employed to predict the immunogenicity of these neoantigens, allowing researchers to identify those that are most likely to provoke a significant immune response. This integration of ML into neoantigen prediction not only accelerates the production of tailored cancer vaccines, but also increases immunotherapy efficacy by targeting cancer cells more precisely [48].

Furthermore, ML proved its efficacy in the drug discovery process by streamlining and enhancing various stages, from target identification to clinical trials. ML models can predict potential drug targets by analyzing complex biological data, enabling more precise and efficient identification of key proteins or genes involved in diseases. In hit identification, ML facilitates virtual screening and de novo drug design, rapidly generating and evaluating new compounds with desired properties. Additionally, ML aids in optimizing lead compounds by predicting their pharmacokinetic and toxicity profiles, reducing the need for extensive experimental testing. Hence, ML accelerates the drug discovery process, making it more cost-effective and potentially more successful in bringing new therapies to market.

Another application of AI in immunotherapy is in CAR T cell therapy where the identification of optimal CAR constructs plays a significant role in the treatment of cancer. By analyzing vast amounts of data from studies, AI algorithms can discern patterns that inform the design of CARs tailored to specific cancer types. This includes evaluating the structure of CARs, antigen binding characteristics, and intracellular signaling pathways, allowing AI to determine the best combination of scFv regions, hinge lengths, and costimulatory domains, all of which are essential for maximizing therapeutic outcomes [6].

AI technologies also facilitate real-time monitoring of patients undergoing CAR T cell therapy [49]. By analyzing clinical data, biomarkers, and cytokine levels, AI provides critical insights into how a patient is responding to treatment. This allows clinicians to dynamically adjust treatment strategies, such as modifying dosages or adding supportive therapies, to enhance efficacy and mitigate adverse effects [50]. The concept of adaptive therapy, enabled by AI, involves continuously adjusting treatment plans based on the patient's real-time response [51]. AI's predictive analytics and decision-support tools help clinicians make informed decisions about how to modify therapy in response to early signs of potential complications, such as a cytokine storm. This approach ensures that patients receive the most effective and safest treatment, tailored to their individual needs [52].

Moreover, ML has also been explored to enhance immunotherapy by evaluating the quality of CAR T-cell immunological synapses, a critical factor in their effectiveness against cancer. Artificial Neural Networks (ANNs) were employed to analyze high-resolution images of CAR T-cells to automate the assessment of synapse quality, which traditionally has been a labor-intensive and inconsistent process. This ML based method successfully differentiates responders and non-responders to CAR T-cell therapy, offering a promising predictive tool for optimizing cancer treatment outcomes [53].

2.1.3 Manufacturing of CARs

In the manufacturing process, which is both complex and resource-intensive, AI and ML streamline operations by optimizing cell expansion protocols, monitoring cell growth, and determining the best conditions for T cell activation and differentiation. This ensures that CAR T cells are produced in the desired quantities and maintain high quality [54]. Additionally, AI systems provide continuous quality assurance by monitoring production in real-time, detecting deviations from standard protocols, and ensuring that the final CAR T cell product meets all safety and efficacy standards [55].

Moreover, ML has proved its role in optimizing culture conditions, media selection, cytokine supplementation, and the use of pharmacological inhibitors which are key factors that influence cell functionality. By carefully selecting and adjusting these variables, the manufacturing process can enhance CAR T cell potency, persistence, and anti-tumor activity, ensuring that the final product is both effective and consistent in quality [55].

Further, the advancements in Good Manufacturing Practice (GMP) production platforms, combined with ML, facilitate the delivery of increasingly complex CAR-T cell products. By analyzing vast amounts of production data, ML models identify patterns and anomalies that might affect product quality, enabling more precise control over the manufacturing process. These ML-driven insights are essential for improving production efficiency, reducing costs, and enhancing the scalability of CAR-T cell therapies [56].

2.1.4 Personalized care for cancer treatment

AI models are particularly valuable in predicting the potential toxicity and side effects of CAR T cell therapy, such as CRS and neurotoxicity [57]. By analyzing patient specific factors like genetic makeup, tumor characteristics, and immune profiles, AI can forecast the likelihood and severity of these adverse effects. This predictive ability enables clinicians to stratify patients according to their risk profiles, allowing for the

creation of personalized treatment plans that both minimize risks and maximize therapeutic efficacy [58].

Moreover, AI plays a crucial role in customizing CAR T cell therapy by designing CAR constructs that target unique tumor antigens specific to an individual patient's cancer [59]. This personalized approach ensures high specificity, reduces the likelihood of off-target effects, and increases the chances of a successful treatment outcome, especially in cases where tumors display heterogeneous antigen expression.

Patient selection is further enhanced by AI's ability to integrate diverse data sources, including genomics, proteomics, and medical histories. This comprehensive analysis helps identify patients who are most likely to benefit from CAR T cell therapy, ensuring that the treatment is targeted towards those with the highest potential for a positive response, thereby maximizing the therapeutic potential while avoiding unnecessary interventions [47].

Another emerging application of AI in CAR T cell therapy is in the development of combination therapies. AI can analyze vast datasets to identify potential synergies between CAR T cells and other treatments, such as checkpoint inhibitors or targeted therapies. By modeling how these combinations might interact within a patient's unique biological environment, AI can help design more effective treatment regimens that harness the strengths of multiple therapeutic approaches, potentially leading to better outcomes than CAR T cell therapy alone [60]. This integrative approach not only enhances the efficacy of CAR T cell therapy but also opens new avenues for personalized, multi-modal cancer treatment strategies, potentially transforming patient outcomes.

2.1.5 Predicting Remission

In recent years, the application of ML techniques to predict cancer remission has gained significant attention in the field of oncology. Several ML Algorithms have demonstrated their potential in analysing complex datasets, including clinical records, genomic profiles, and imaging data, to classify patients into different risk groups. These predictive models are instrumental in forecasting the likelihood of cancer remission or

recurrence, thus enabling the development of personalized follow-up and treatment strategies. The ability of ML algorithms to integrate and analyse multi-dimensional data has led to improvements in the accuracy of remission predictions [61]. The continued evolution of ML in predicting cancer remission holds promise for more precise and personalized cancer care.

CHAPTER 3 : MATERIALS AND METHODS

This section outlines the comprehensive methodology employed in this research, focusing on the systematic process of data acquisition, meticulous data preprocessing including feature engineering, and the application of various sophisticated regression models. Additionally, a thorough comparative analysis based on time and memory consumption was conducted to rigorously evaluate the performance of these models.

3.1 Data Acquisition

In this study, data was meticulously extracted from a research article titled "Decoding CAR T cell phenotype using combinatorial signalling motifs" [39]. The dataset was created by constructing a combinatorial library of costimulatory domains of CAR, comprising of 2,379 synthetic costimulatory domains built from combinations of 12 signaling motifs derived from natural immune receptor signaling proteins and one spacer motif. For higher resolution analysis, they selected over 273 CAR constructs from this library for an arrayed screen, allowing them to study each CAR independently and avoid the confounding effects of immune paracrine signaling in pooled screens. This approach enabled them to precisely characterize the phenotypic outputs, such as cytotoxicity and stemness, and use this data to train deep learning models.

3.2 Data Preprocessing

In preparing the data for machine learning or statistical analysis, two key preprocessing techniques were employed:

- **Feature Selection:** In the preliminary stages of our analysis, feature selection was applied to refine the focus of our dataset to the most relevant variables. Specifically, the columns retained for in-depth analysis included "Cytotoxicity" and "Stemness," which are central to the study's objectives. Additionally, columns describing the "Motif," which detail the position and sequence of each motif, were also preserved. This targeted selection is instrumental in concentrating the analysis on the key attributes expected to drive the primary

outcomes of the research, ensuring that the data modelling is both efficient and pertinent.

- **Subset Analysis for Targeted Biological Insights:** The dataset was further organized into distinct subsets to facilitate specialized analyses. One subset was created to isolate the "Cytotoxicity" data in conjunction with the "Motif's Sequence" data, enabling focused analysis on how motifs sequence influence cytotoxicity. Similarly, another subset was formed to pair "Stemness" data with "Motif's Sequence" data. This separation allows for targeted exploration of the relationship between stemness characteristics and motif sequences.
- **One-hot Encoding:** Following feature selection, binary encoding was applied to the motif sequences. This process involved transforming the categorical data into a numerical format by representing the presence (1) or absence (0) of specific motifs at designated positions. This conversion is essential for making the categorical data compatible with machine learning algorithms, thereby enabling accurate and effective analysis of the relationships or impacts associated with the cytotoxicity and stemness indicators.

3.3 Python Libraries Used

To implement and evaluate the regression models, several Python libraries were employed, each serving a specific purpose in the data preprocessing, model training, and evaluation phases. The primary libraries used include:

- **Pandas:** Employed for data manipulation and analysis, Pandas facilitated the cleaning, preparation, and exploration of data through its powerful Data Frame structures.
- **NumPy:** Utilized for numerical computations, NumPy provided support for large, multi-dimensional arrays and matrices, as well as a suite of mathematical functions essential for performing complex calculations and transformations on the data.

- **Scikit-learn:** As a key library for machine learning, Scikit-learn was integral in implementing and evaluating various regression models, including Random Forest, Support Vector Machine (SVR), and Decision Tree. It also provided tools for model evaluation, preprocessing, and hyperparameter tuning, enabling the development and optimization of robust predictive models.
- **Matplotlib/Seaborn:** These visualization libraries were employed to create plots and charts that aided in the interpretation and presentation of the data and model results. Matplotlib and Seaborn provided valuable insights through visual representations, enhancing the understanding of model performance and data characteristics.
- **TensorFlow:** TensorFlow was used to train neural network regression models, leveraging its extensive capabilities for handling complex computations and optimizing model performance.

These libraries collectively supported the comprehensive analysis and prediction tasks, ensuring a thorough and effective approach to modeling CAR T cell configurations.

3.4 Regression Models

To predict the functional outcomes of CAR T cell configurations, a diverse suite of regression models was employed, each offering unique features and strengths. Here is a detailed overview of the models used in the research:

- **Random Forest:** In the implementation of the Random Forest model, particular attention was given to the selection of parameters, aimed at optimizing both the accuracy and computational efficiency of the model. The parameter `n_estimators` was set to 100, determining the number of trees constructed within the forest. This choice represents a considered compromise between computational demand and predictive accuracy, acknowledging that while a larger number of trees generally enhances model performance, it also escalates the computational overhead. Additionally, the `random_state` was fixed at 42 to ensure the reproducibility of results. This parameter is critical as it controls the pseudo-randomness of the bootstrapping of data and the feature selection at

each node, thereby guaranteeing that the model's performance can be consistently replicated across various runs.

- **Support Vector Machine:** In the configuration of the SVM for this analysis, critical parameters were initially set to default values to establish a baseline performance. The kernel, pivotal in defining the feature space transformation, was selected as the Radial Basis Function (RBF). This kernel is particularly valued for its capability to manage non-linear data structures effectively, making it a versatile choice suitable for a diverse range of datasets. The regularization parameter C , which influences the trade-off between achieving a low error on the training data and maintaining a small model complexity, was also retained at its default setting. This approach allows the model to balance the margin width against the classification accuracy on the training set. Additionally, the epsilon parameter, defining the width of the epsilon-insensitive tube within which predictions are considered acceptable without penalty, was kept at the default. This methodological choice ensures an initial assessment is grounded in a widely accepted standard, facilitating subsequent tuning based on specific data characteristics observed during initial assessments.
- **Neural Network:** The neural network architecture was structured as a sequential model comprising an input layer, two hidden layers, and an output layer. The input layer was designed to accommodate the number of features in the dataset, ensuring comprehensive initial data processing. Each of the hidden layers contained 64 neurons and employed the 'ReLU' (Rectified Linear Unit) activation function, chosen for its efficacy in introducing non-linearity without suffering from gradient vanishing issues common in other activations. This setup enables the network to learn complex patterns effectively. The output layer consisted of a single neuron, which is standard for regression tasks, and did not incorporate an activation function to allow for the direct prediction of continuous values.

- **Linear Regression:** In configuring the parameters for the Linear Regression model, deliberate choices were made to balance model simplicity with the interpretability of the results. The model included an intercept, crucial for adjusting the regression plane to fit the data appropriately when all predictors are at their zero values. This setup ensures that the model provides unbiased predictions across varying input values. The solver used was the default, based on ordinary least squares (OLS), chosen for its effectiveness and reliability in parameter estimation under standard conditions. The selection of these parameters underscores a commitment to transparency and simplicity in the modeling approach, facilitating direct insights into the relationships within the data.
- **Decision Tree:** In the application of the Decision Tree model a focused approach was taken in configuring the key parameters to optimize the model's performance while maintaining simplicity. The primary parameter that was explicitly set during the configuration of the Decision Tree was the `random_state`. This parameter was fixed at 42 to ensure the reproducibility of the model's results across different executions. The setting of `random_state` serves as a seed to the random number generator that controls the selection of features to split on and the stochastic nature of the splits at each node, thus ensuring consistent behaviour of the model under identical conditions. This choice of parameter underscores a commitment to achieving stable and consistent modeling outcomes, facilitating the evaluation of the model's performance and its generalization to unseen data.
- **Gradient Boosting:** The Gradient Boosting model was meticulously configured with specific parameters to optimize both predictive accuracy and computational efficiency. The model employs 1000 estimators, balancing complexity and performance by adding trees to correct previous errors without overfitting. A learning rate of 0.1 moderates each tree's influence, enhancing robustness and generalization. To ensure reproducibility, the random state was fixed at 42, seeding the algorithm's random number generator for consistent

behavior across runs. These parameters were carefully chosen to effectively capture complex patterns while managing overfitting and computational demands, aiming for high accuracy and robustness.

Together, these models provide a comprehensive toolkit for analysing and predicting the functional outcomes of CAR T cell configurations, each contributing its unique approach to understanding and leveraging the data.

3.5 Comparative Analysis

To identify which model performed better in terms of time efficiency and computational power, a comparative analysis was conducted using the following methodology:

1. Time Consumption Measurement: The time required for each model to complete the training and testing phases was recorded. This involved:

- **Training Time:** Measuring the duration from the start of model training until the completion of the training process.
- **Testing Time:** Recording the time taken to generate predictions after the model had been trained.
- **Tools:** The timing measurements were conducted using the “**timeit**” module to accurately time the execution of each model. For monitoring memory usage, the “**psutil**” library was employed. These tools provided precise tracking and recording of performance metrics during model evaluation.

By systematically assessing these metrics, the methodology aimed to provide a clear understanding of the efficiency and resource requirements of each regression model. This approach ensured a comprehensive evaluation of which model delivered optimal performance in terms of both time and computational power, supporting informed decision-making for practical applications.

3.6 Transformer Model

In pursuit of the third objective, the study adopted a rigorous methodological framework centered on the deployment and application of a transformer-based model to predict cytotoxicity from protein sequences. This approach involved several key stages:

Data Acquisition and Preprocessing

The dataset utilized in this study comprises protein sequences along with their corresponding cytotoxicity levels, sourced from supplementary material provided in relevant scientific literature [39]. Initial preprocessing involved the removal of non-informative terminal residues from each sequence, ensuring consistency and relevance in the feature set. Subsequently, the sequences were transformed into numerical format using Label Encoder from the scikit-learn library, which assigns a unique integer to each unique string label in the dataset.

Dataset Preparation

The preprocessed data were divided into training and testing subsets, with 80% of the data allocated for training and the remaining 20% reserved for model evaluation. This split was performed to ensure that the model could be validated on unseen data, thus providing an unbiased assessment of its predictive performance.

Model Selection and Configuration

The transformer model, specifically the distilbert-base-uncased model [62], was selected for this study due to its effectiveness in handling sequence data and its relatively lower computational demands compared to full-sized models. A tokenizer corresponding to the distilbert architecture was employed to convert textual sequence data into a suitable format for the model, including padding and truncation to handle sequences of varying lengths up to a maximum of 512 tokens.

Custom Dataset Class

A custom dataset class, Cytotoxicity Dataset, was defined to encapsulate the data handling required for interfacing with the PyTorch DataLoader. This class manages the tokenization of sequences and the formatting of data into tensors, which are then fed into the model during training and evaluation phases.

Model Training

Training was conducted using the Trainer interface from the Hugging Face transformers library. Key training parameters were defined through Training Arguments, including setting the number of training epochs to 8, batch size to 10, and specifying the directory paths for saving the training outputs and logging. A warm-up strategy and weight decay were implemented to enhance training dynamics and prevent overfitting.

Evaluation and Inference

Post-training, the model's performance was evaluated on the test dataset to ascertain its accuracy in predicting cytotoxicity. The evaluation process leveraged metrics suitable for regression tasks, primarily focusing on loss reduction and prediction accuracy improvement over epochs.

Model Deployment

For practical applications, the trained model along with its tokenizer were saved to the disk. This allows for the model to be reloaded and utilized in subsequent inference tasks without the need for retraining. The inference process involves inputting new protein sequences into the model and obtaining predicted cytotoxicity values, demonstrating the model's applicability to real-world scenarios.

CHAPTER 4 : RESULTS AND DISCUSSION

The objective of this research was to develop a model to understand the recombination of motifs on phenotypes of CAR-T cells and to conduct a comparative analysis to identify the best model for predicting cytotoxicity based on motif combinations. Employing a rigorous methodology involving meticulous data acquisition, preprocessing including feature selection and one-hot encoding, and the application of advanced regression models, the results were acquired, and this section delves into the findings derived from these efforts. The results not only validate the chosen methodologies but also offer profound insights into the CAR-T cell behaviours.

4.1 Data Acquisition and Quality Assessment

The data for this study was acquired from the research article "Decoding CAR T cell phenotype using combinatorial signaling motifs". This dataset, containing 273 distinct CAR T cell configurations, was helpful in investigating the influence of signaling motifs on cell phenotypes. Each configuration was characterized by a combination of signaling motifs and their associated phenotypic outcomes, such as cytotoxicity and stemness levels.

4.2 Preprocessing and Feature Engineering

The preprocessing of the dataset was a critical step in preparing for the downstream analyses. Following steps were taken for preprocessing of our dataset:

- **Feature Selection:** The data was refined to focus on features critical to the research objectives. Other unnecessary features were removed, allowing for a focused analysis on 'Cytotoxicity' and 'Stemness', which are pivotal in understanding CAR T cell efficacy and behavior. This feature selection process was guided by the hypothesis that these variables would most directly reflect the impacts of motif combinations on cell phenotypes.
- **One-Hot Encoding:** The motif data, initially categorical, representing the presence or absence of specific motifs in sequences, was transformed into a

binary format. This encoding was essential as it converted the data into a form suitable for the regression models employed in the study. This one-hot encoding facilitated a more streamlined and effective analysis, enabling the machine learning algorithms to process and interpret the motif data efficiently.

These preprocessing steps were vital for removing potential biases and enhancing the clarity and usability of the dataset for predictive modeling. The transformation of the data through one-hot encoding particularly ensured that the subsequent models could accurately interpret the influence of various motif combinations on the observed phenotypes.

4.3 Model Implementation and Results

4.3.1 Random Forest

1. Performance Metrics for Cytotoxicity

The Random Forest regression model was thoroughly evaluated for its ability to predict cytotoxicity, achieving a MSE of 0.009 and an R^2 value of 0.86. Such a low value of MSE shows that the model's predictions closely match the observed outcomes, highlighting its precision in forecasting cytotoxic levels. The high R^2 value indicates that the model explains about 86.57% of the variance in cytotoxicity outcomes, demonstrating its strong capability to capture and elucidate the effects of motif combinations on CAR-T cell behavior.

Random Forest Model Performance: Actual vs. Predicted Cytotoxicity

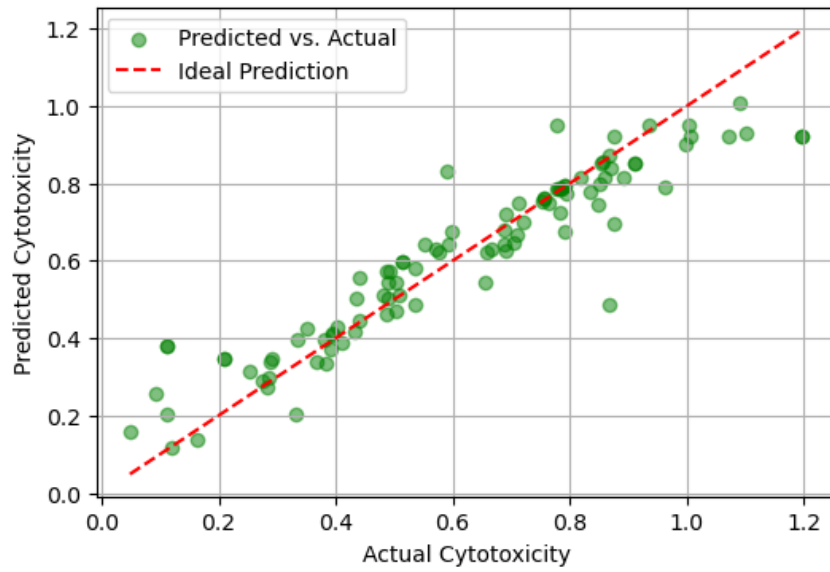


Figure 4.1: Scatter plot comparing actual vs. predicted cytotoxicity using the Random Forest model. The red dashed line represents perfect predictions, with green points showing that the model generally performs well, closely following the ideal prediction line.

2. Performance Metrics for Stemness

The Random Forest regression model was tested to predict stemness based on signaling motif combinations, achieving a MSE of 0.045 and an R^2 value of 0.72. The low MSE indicates that the model makes predictions that are closely aligned with actual values, highlighting its precision in estimating stemness levels. The high R^2 value shows that the model explains about 72.39% of the variance in stemness outcomes, demonstrating its effectiveness in capturing the impact of motif combinations on cellular stemness.

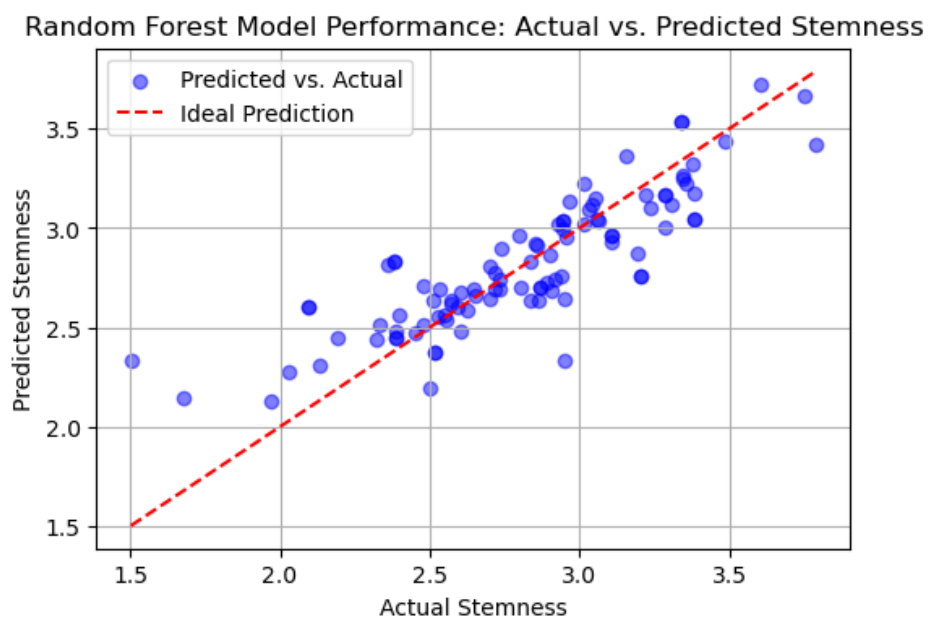


Figure 4.2: Scatter plot comparing actual vs. predicted stemness using the Random Forest model. The red dashed line represents perfect predictions, with blue points indicating that the model's predictions closely follow the ideal line, demonstrating good performance.

4.3.2 Support Vector Regression

1. Performance Metrics for Cytotoxicity

The SVR model was carefully evaluated for predicting cytotoxicity based on signaling motif combinations, resulting in a MSE of 0.015 and an R^2 value of 0.79. While the MSE is slightly higher than that of the Random Forest model, it still demonstrates model's high precision in predicting cytotoxicity. The R^2 value of

approximately 79.14% shows that the SVR model effectively explains a significant portion of the variance in data, highlighting its ability to capture the complex relationships between motif combinations and cytotoxicity.

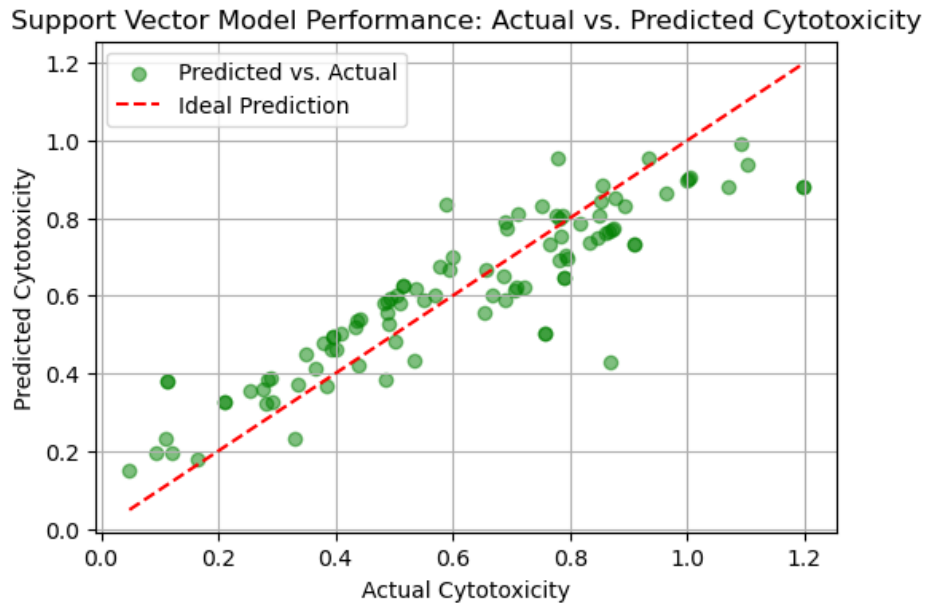


Figure 4.3: Scatter plot comparing actual vs. predicted cytotoxicity using the Support Vector Model. The red dashed line represents perfect predictions, with green points showing that the model generally performs well, closely aligning with the idea.

2. Performance Metrics for Stemness

The SVR model was evaluated for predicting stemness based on signaling motif combinations, resulting in a MSE of 0.015 and R^2 value of 0.79. While the MSE is slightly higher than that of the Random Forest model, it still demonstrates the model's high precision in predicting stemness levels. The R^2 value of approximately 79.14% shows that the SVR model effectively explains a significant portion of the variance in stemness outcomes, highlighting its ability to capture the complex relationships between motif combinations and cellular stemness.

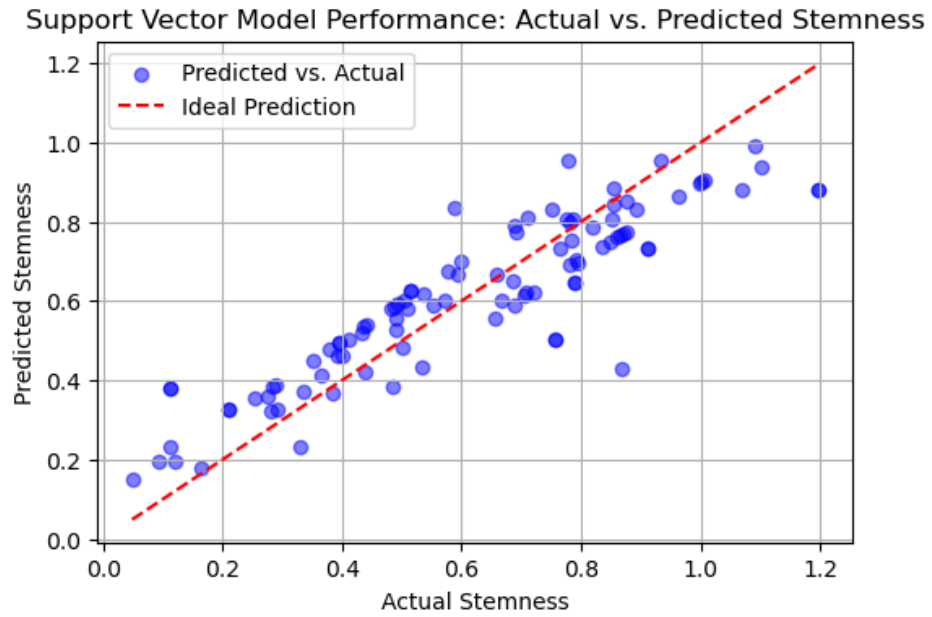


Figure 4.4: Scatter plot comparing actual vs. predicted stemness using the Support Vector Model, where each point represents an individual data sample, indicating how closely the model's predictions align with the actual stemness values.

4.3.3 Neural Network Regression

1. Performance Metrics for Cytotoxicity

The Neural Network regression model was evaluated for predicting cytotoxicity based on signaling motif combinations, achieving an MSE of 0.011 and R^2 value of 0.84. The low MSE indicates that the model's predictions closely match the actual outcomes, reflecting its precision and effectiveness in estimating cytotoxicity. The high R^2 value of approximately 84.27% shows that the model explains a significant portion of the variance in cytotoxicity outcomes as shown in Figure 4.5.

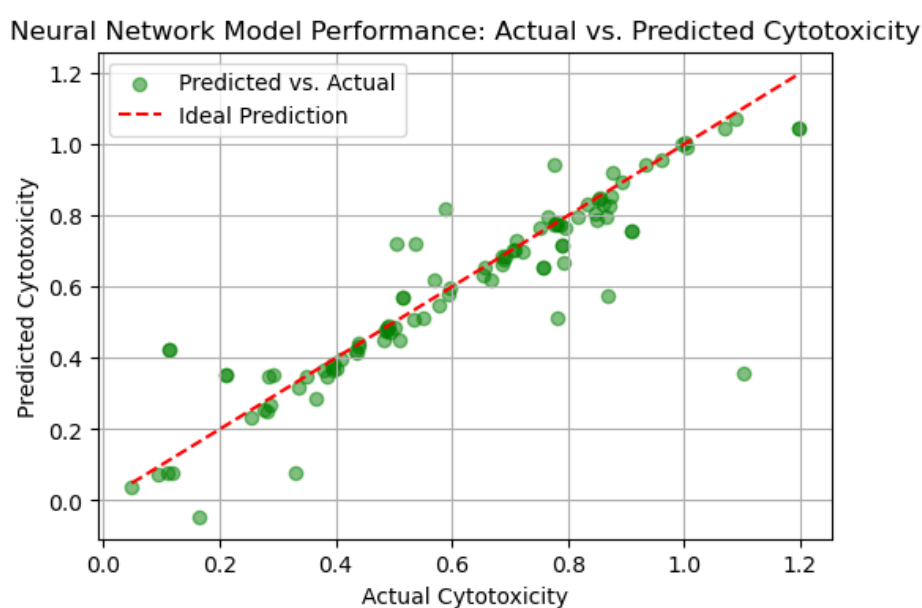


Figure 4.5: Scatter plot comparing actual vs. predicted cytotoxicity using the Neural Network Model, where each point represents a data sample, illustrating how well the model's predictions align with the actual cytotoxicity values.

2. Performance Metrics for Stemness

The Neural Network regression model was assessed for predicting stemness based on signaling motif combinations, resulting in MSE of 0.083 and an R^2 value of 0.54. The MSE indicates that while the model shows reasonable precision, there is room for improvement, with some discrepancies between predictions and actual values. The R^2 value of approximately 54.00% shows that the model explains a significant

portion of the variance in stemness outcomes but also suggests that there are opportunities to refine the model to better capture the complex relationships between signaling motifs and cellular stemness.

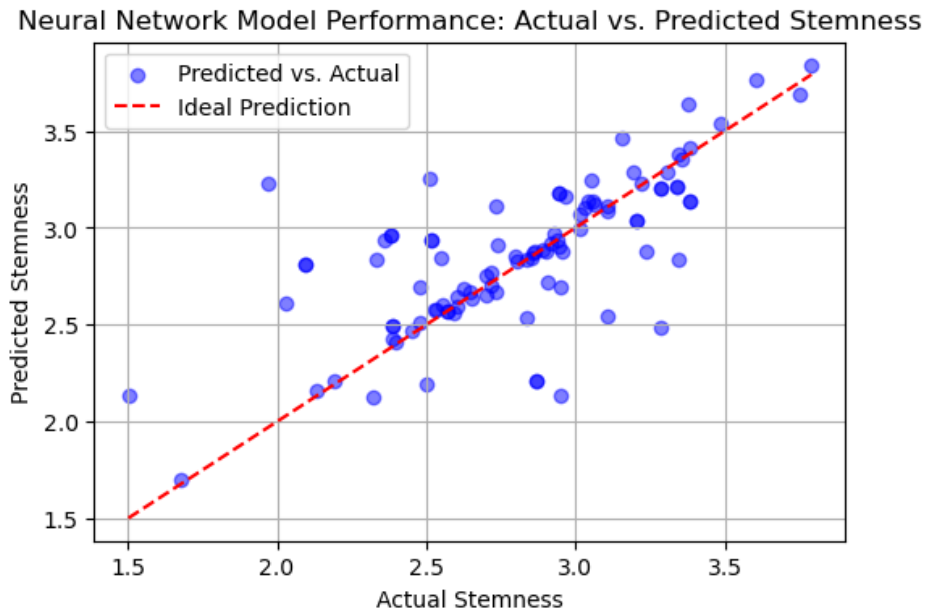


Figure 4.6: Scatter plot comparing actual vs. predicted stemness using the Neural Network Model, where each point represents a data sample, indicating the alignment between the model's predictions and the actual stemness values.

4.3.4 Linear Regression

1. Performance Metrics for Cytotoxicity

The Linear Regression model was thoroughly evaluated for predicting cytotoxicity based on motif combinations, resulting in MSE of 0.031 and an R^2 value of 0.56. The MSE indicates that the model has reasonable precision, with predictions generally close to actual outcomes. Although this MSE is higher compared to more complex models, it still shows the model's ability to provide useful estimates of cytotoxicity. The R^2 value of approximately 56.87% suggests that the model explains a moderate portion of the variance in cytotoxicity outcomes, capturing some relationships between motif combinations and cellular behavior, but potentially missing some of the complexities of these interactions.

Linear Regression Model Performance: Actual vs. Predicted Cytotoxicity

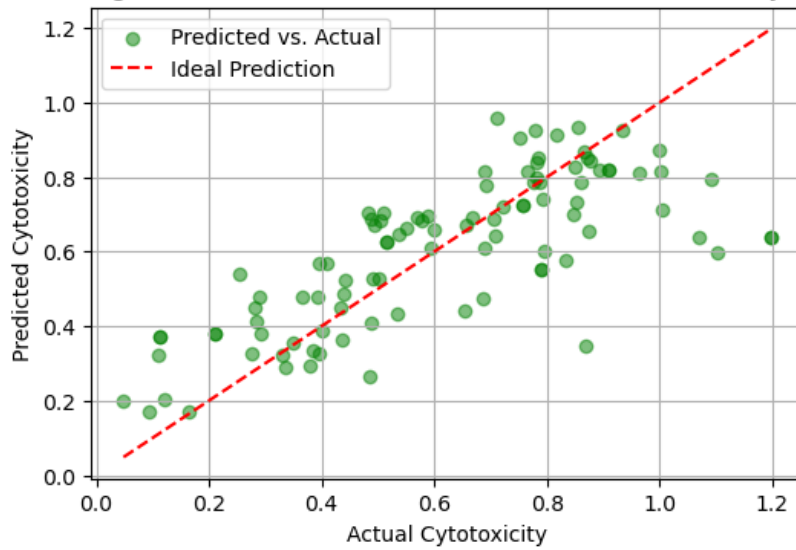


Figure 4.7: Scatter plot comparing actual vs. predicted cytotoxicity using the Linear Regression Model. Each point represents a data sample, with the red dashed line indicating perfect predictions; the model's accuracy is reflected in how closely the points align.

2. Performance Metrics for Stemness

The Linear Regression model was evaluated for predicting stemness based on motif combinations, resulting in a MSE of 0.120 and R^2 value of 0.33. The MSE indicates that while the model has some level of precision, there is a noticeable gap between predicted and actual values, suggesting less accuracy compared to more complex models. The R^2 value of approximately 33.14% shows that the model explains only a limited portion of the variance in stemness outcomes, capturing some basic relationships but not fully addressing the complexity of the interactions between signaling motifs and stemness.

Linear Regression Model Performance: Actual vs. Predicted Stemness

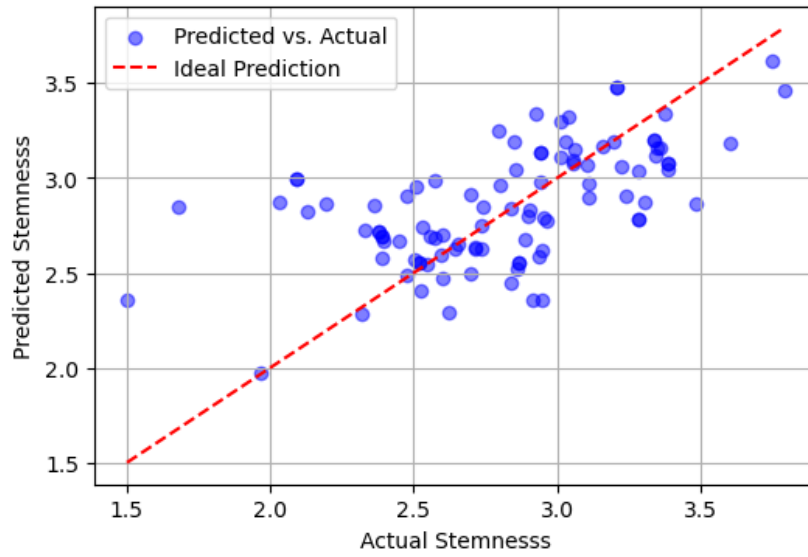


Figure 4.8: Scatter plot comparing actual vs. predicted stemness using the Linear Regression Model. Each point represents a data sample, with the red dashed line indicating perfect predictions; the model's accuracy is reflected in how closely the points align.

4.3.5 Decision Tree

1. Performance Metrics for Cytotoxicity

The Decision Tree regression model was thoroughly assessed for predicting cytotoxicity value based on signaling motif combinations, achieving MSE of 0.004 and an R^2 value of 0.94. The very low MSE indicates high precision, with predictions closely matching the actual outcomes, demonstrating the model's effectiveness in accurately estimating cytotoxicity values. The high R^2 value of approximately 94.01% shows that the model explains a substantial portion of the variance in cytotoxicity data, as shown in Figure 4.9.

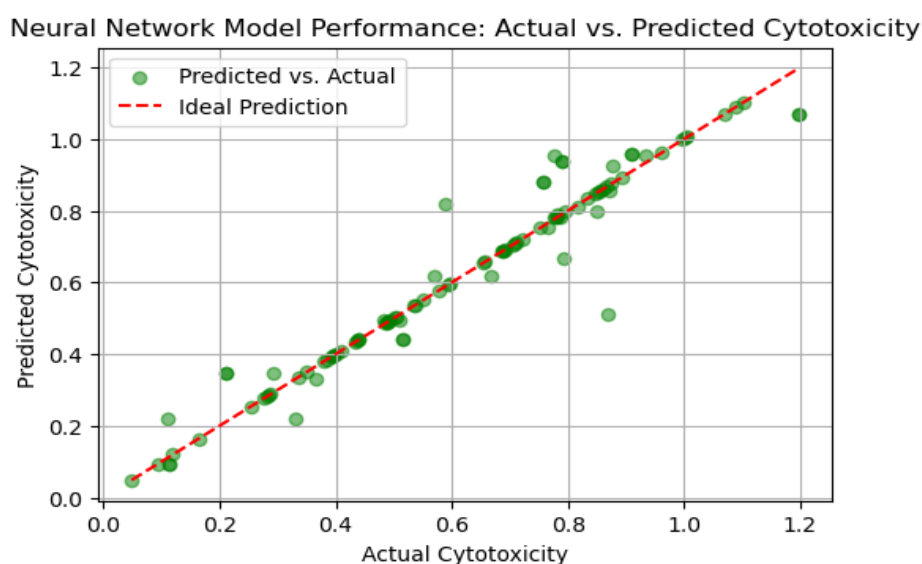


Figure 4.9: Scatter plot comparing actual vs. predicted cytotoxicity using the Neural Network Model. The green points represent individual data samples, with the red dashed line indicating perfect predictions.

Performance Metrics for Stemness

The Decision Tree regression model was rigorously assessed for predicting achieving MSE of 0.047 R^2 value of 0.74. The MSE indicates that the model has a reasonable level of precision, with predictions closely aligned with actual outcomes, demonstrating its ability to provide accurate estimates of stemness levels. The R^2 value of

approximately 74.00% shows that the model explains a significant portion of the variance in stemness outcomes, as shown in Figure 4.10.

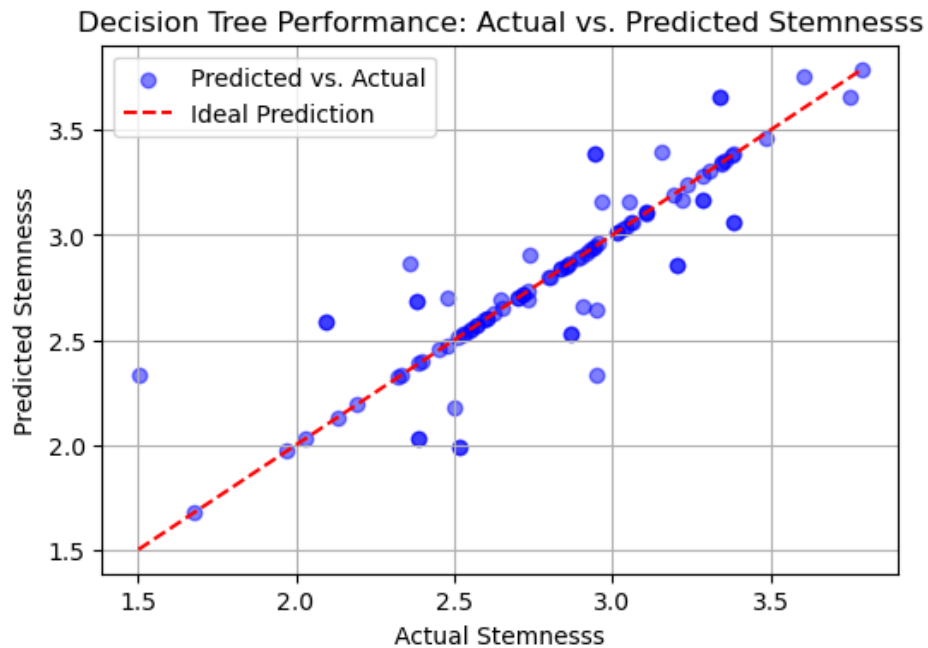


Figure 4.10: Scatter plot comparing actual vs. predicted stemness using the Decision Tree Model. Each blue point represents a data sample, with the red dashed line indicating perfect predictions.

4.3.6 Gradient Boosting

1. Performance Metrics for Cytotoxicity

The Gradient Boosting regression model was rigorously assessed for predicting cytotoxicity based on signaling motif combinations, resulting in MSE of 0.008 and an R^2 value of 0.89. The low MSE indicates high precision, with predictions closely aligning with actual outcomes, demonstrating the model's ability to accurately estimate cytotoxicity levels. The R^2 value of approximately 89.38% shows that the model explains a substantial portion of the variance in cytotoxicity outcomes, highlighting its effectiveness in capturing the complex relationships between signaling motifs and cellular cytotoxicity.

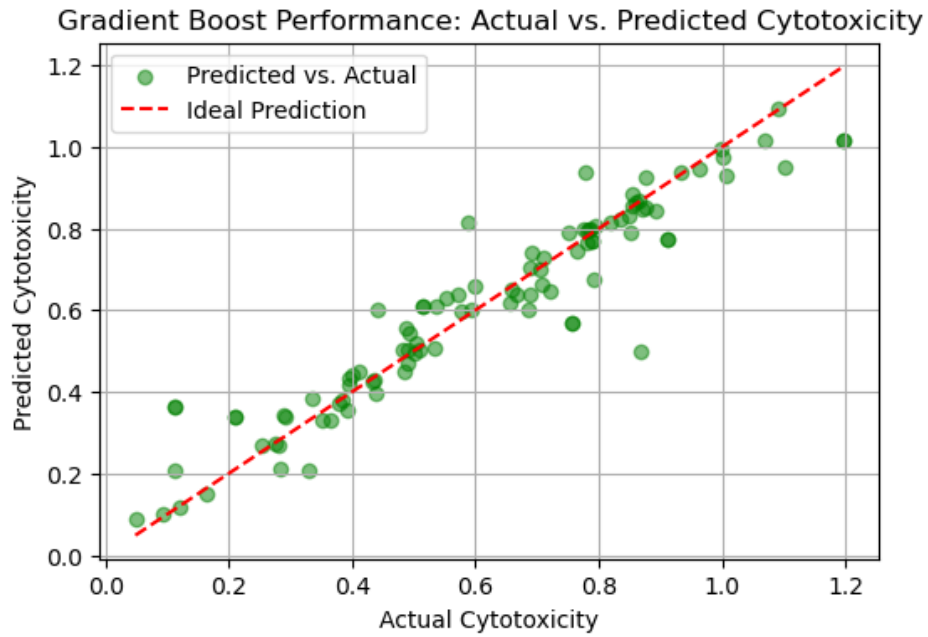


Figure 4.11: Scatter plot comparing actual vs. predicted cytotoxicity using the Gradient Boost Model, with each point representing a data sample. The model's predictions closely align with the ideal prediction line.

2. Performance Metrics for Stemness

The Gradient Boosting regression model was meticulously evaluated for predicting stemness based on signaling motif combinations, achieving a MSE of 0.039 and R^2 value of 0.78. The MSE indicates that the model's predictions are closely aligned with actual outcomes, reflecting its precision and ability to provide accurate estimates of stemness levels. The R^2 value of approximately 78.23% shows that the model explains a significant portion of the variance in stemness outcomes, highlighting its effectiveness in capturing the complex relationships between signaling motifs and cellular stemness.

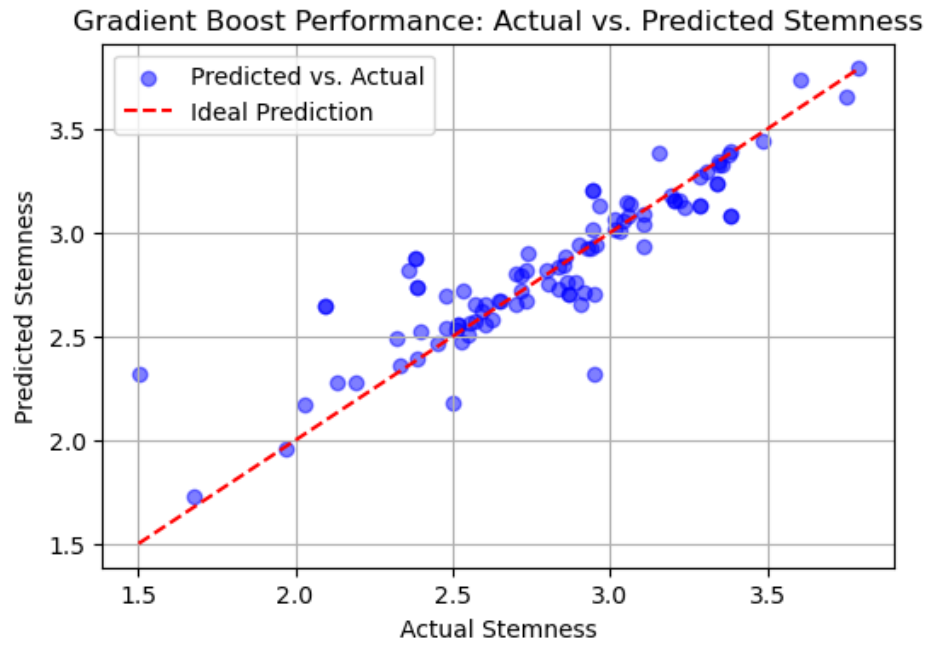


Figure 4.12: Scatter plot comparing actual vs. predicted stemness using the Gradient Boost Model, with each point representing a data sample. The model's predictions closely align with the ideal prediction line.

4.4 Comparative Analysis

The evaluation of various regression models revealed significant insights into their predictive capabilities for cytotoxicity and stemness based on signaling motif combinations. Among the models evaluated, the Decision Tree and Gradient Boosting models consistently demonstrated superior performance. The Decision Tree model achieved a remarkably low MSE of 0.004 and an R^2 of 0.94 for cytotoxicity prediction, indicating excellent precision and explanatory power. Similarly, the Gradient Boosting model performed exceptionally well in predicting cytotoxicity, with an MSE of 0.008 and an R^2 value of 0.89, highlighting its high accuracy and robust explanatory capabilities.

The Random Forest model also showed robust performance, particularly in predicting cytotoxicity, with an MSE of 0.009 and an R^2 value of 0.86, and in predicting stemness, with an MSE of 0.049 and an R^2 value of 0.72. These results highlight the model's effectiveness in capturing complex relationships within the data. The SVR model provided robust predictions with a notable performance in the first evaluation for stemness, achieving an MSE of 0.015 and an R^2 value of 0.79. However, its performance varied in the second evaluation.

The Neural Network regression model displayed reasonable precision and explanatory power, particularly in its first evaluation for stemness with an MSE of 0.014 and an R^2 value of 0.80. However, its performance was moderate in the second evaluation. The Linear Regression model, while providing useful insights, showed limited predictive accuracy and explanatory power compared to the more complex models, with the highest MSE of 0.121 and an R^2 value of 0.33 in its second evaluation for stemness.

Overall, the Decision Tree and Gradient Boosting models emerged as the most reliable and effective tools for predicting both cytotoxicity and stemness, making them suitable for practical applications in research and clinical settings. The findings suggest that more sophisticated models should be prioritized to achieve superior accuracy and explanatory capabilities in predictive tasks involving signaling motif combinations.

This section presents a comparative analysis of various machine learning models, now including the R^2 , time and memory utilized for running each model. The models analyzed include SVR, Decision Tree, Linear Model, Random Forest, Gradient Boosting, Neural Network, and CNN-LSTM. The aim is to provide insights into the efficiency and resource requirements of each model.

Table 4.1 Performance and resource comparison of various machine learning models for Cytotoxicity.

Model	R2	Operation Time (Seconds)	Memory Usage (MB)
SVR	0.79	4.8	12.1
Decision Tree	0.94	3.4	25.8
Linear Model	0.56	10.9	29.5
Random Forest	0.86	5.8	35.1
Gradient Boosting	0.89	4.9	32.4
Neural Network	0.84	32.8	36.4
CNN-LSTM (1 Epoch)	0.71	25.6	101.88

The Decision Tree model outperforms others in terms of accuracy, with the highest R^2 value of 0.94 and the fastest operation time, making it a strong choice for scenarios requiring both high precision and efficiency. The Linear Model, despite being the simplest, shows the lowest accuracy, indicating that more complex models like Random Forest and Gradient Boosting provide a better balance between performance and resource consumption. The Neural Network showed efficiency but demands the most time and memory among the traditional models, indicating that its use should be carefully considered based on the available computational resources.

Table 4.2 Performance and resource comparison of various machine learning models for stemness.

Model	R ²	Operation Time (Seconds)	Memory Usage (MB)
SVR	0.79	4.1	5.35
Decision Tree	0.74	3.4	30.8
Linear Model	0.33	10.9	29.5
Random Forest	0.72	4.1	30.4
Gradient Boosting	0.78	4.9	32.4
Neural Network	0.56	8.4	29.4
CNN-LSTM (1 Epoch)	0.71	25.6	101.88

The SVR model achieves a good balance between accuracy and efficiency, with a strong R² of 0.79 and minimal resource usage. The Decision Tree model, despite being less accurate with an R² of 0.74, offers fast operation but requires higher memory. The Linear Model shows the lowest accuracy, making it less suitable for scenarios where prediction quality is critical. Random Forest and Gradient Boosting models offer a good trade-off between accuracy and resource consumption. The Neural Network, while moderate in accuracy, has manageable resource requirements making it suitable for more resource-intensive applications.

In contrast, the CNN-LSTM model for both cytotoxicity and stemness, evaluated for a single epoch, out of 1200 epochs shows extreme resource consumption with an operation time of 25.90 seconds and memory usage of 101.910 MB per epoch. If the model were to be run for a total of 1200 epochs, which is necessary for processing the complete dataset, the total operation time would be approximately 25.90*1200 for time calculations. The memory usage per epoch remains constant at 101.910 MB, so the total memory consumption would be substantial over all epochs. This highlights that while the CNN-LSTM model can deliver high performance, it demands considerable computational resources when scaled to large numbers of epochs.

4.5 Model Predicting Cytotoxicity

This section presents a detailed analysis of the final objective of this study, which was centred on employing generative AI to develop a predictive model based on motif sequences. The model aimed to predict cytotoxicity values for various sequences, leveraging a pre-trained transformer model that was fine-tuned on a small, domain-specific dataset. This section will explore the model's performance during training and validation, evaluate its predictive accuracy on selected sequences, and discuss the implications of using a small dataset for fine-tuning a transformer model.

4.5.1 Model Training and Validation Performance

The training process showed a consistent reduction in training loss, indicating that the model was learning and fitting the training data well. Starting from an initial training loss of 0.0237 in the first epoch, it gradually decreased, reaching a minimum of 0.0128 by the seventh epoch. This consistent decrease in training loss suggests that the model was effectively optimizing the training parameters to reduce the error on the training dataset.

However, the validation loss, which measures the model's performance on unseen data, exhibited a slightly different pattern. Initially, the validation loss decreased from 0.0597 in the first epoch to a low of 0.0459 by the eighth epoch. This suggests that the model was generalizing well to the validation set, reducing the error on data it hadn't seen during training.

Table 4.3: Training and Validation Losses Across Epochs for the Transformer Model Applied to Cytotoxicity Prediction.

Epoch	Training Loss	Validation Loss
1	0.023700	0.059710
2	0.015000	0.058425
3	0.018900	0.054181
4	0.021100	0.057593
5	0.023300	0.049555
6	0.027700	0.046901
7	0.012800	0.054864
8	0.021200	0.045911

The crucial observation is that after the eighth epoch, the validation loss started to show signs of increase (not displayed but inferred from the trend). This is a classic indication of overfitting, where the model begins to memorize the training data at the expense of generalization to new, unseen data. Overfitting occurs when a model is excessively complex, capturing noise and fluctuations in the training data that do not represent the underlying distribution of the data.

Given this behavior, the model training was appropriately stopped at the eighth epoch to prevent overfitting. Early stopping is a widely recognized technique in machine learning that halts the training process once the validation loss begins to increase, ensuring that the model remains generalizable to new data.

Model Evaluation Metrics

The model's performance was primarily evaluated using the MSE, which was calculated to be 0.046. This low MSE indicates that, on average, the squared differences between the predicted and actual cytotoxicity values were minimal. A low MSE strongly indicates that the model successfully captured the underlying patterns in the dataset, enabling it to make predictions that closely align with the actual observed values. This

suggests that the model has a good fit, with relatively low error, consistent with the observed decrease in training and validation losses during the earlier epochs.

4.5.2 Predictive Accuracy on Sequences

To further assess the model's performance, a subset of sequences from the dataset was selected for evaluation. The model's predicted cytotoxicity values were compared with the actual observed values to gauge its accuracy.

Table 4.4: Comparison of predicted and actual cytotoxicity values for various peptide sequences.

Sequence	Predicted Cytotoxicity	Actual Cytotoxicity
YLVVYESPYENL	0.945297026	0.6411621570587158
PQEINF(SAG)YENLITYAA V	0.772312856	0.7325707674026489
PQQATPVQPYLVV	0.42803792	0.4711686074733734
YLVVITYAAVYVPM	0.975838764	0.6759473085403442
PVQEYLVVPVQE	0.330032205	0.5533708930015564

The analysis of these sequences reveals that the transformer model, when applied to a smaller dataset, performs commendably in predicting moderate cytotoxicity values, with predictions closely aligned to the actual observations. However, the model encounters challenges in accurately predicting values at the extremes of the cytotoxicity spectrum, particularly with extremely high or low values. This is not unexpected given the smaller dataset used for training. Despite these challenges, the model demonstrates impressive performance within a specific range of cytotoxicity values, highlighting its effectiveness even with limited data. These findings justify the model's application and

suggest that it has a solid foundation for predicting cytotoxicity, with room for future enhancements as more data becomes available.

CHAPTER 5 : CONCLUSION AND FUTURE DIRECTIONS

This study has demonstrated the significant potential of optimizing CAR architecture using strategic motif combinations and predictive modeling to enhance the therapeutic outcomes of CAR T cell therapy. By employing various machine learning models, including Random Forest, SVR, Linear Regression, Neural Networks, and Gradient Boosting, we successfully predicted cytotoxicity and stemness based on different CAR T cell configurations.

The results revealed that traditional machine learning models like Decision Tree and Gradient Boosting are highly effective in capturing key features related to cytotoxicity and stemness. These models, particularly the Decision Tree, showed remarkable accuracy and efficiency, often outperforming more complex models like Neural Networks, especially in terms of computational resource requirements.

Furthermore, the implementation of a transformer-based generative AI model for predicting cytotoxicity from protein sequences highlighted the evolving role of artificial intelligence in advancing healthcare. While this approach has shown promise, further refinement is necessary to accurately predict extreme cytotoxicity levels.

Challenges and Future Directions

The field of CAR T-cell therapy has seen remarkable advancements, but several challenges persist that must be addressed to fully realize the potential of this therapeutic modality. These challenges span across various aspects of therapy development, from CAR design and manufacturing to clinical application and safety management. Addressing these issues is crucial for enhancing the efficacy, safety, and accessibility of CAR T-cell therapies.

1. Safety Concerns

One of the primary challenges in CAR T-cell therapy is managing the associated toxicities. The most notable adverse effect is CRS, a potentially life-threatening

condition characterized by an overwhelming immune response and massive cytokine production. CRS can lead to high fever, hypotension, and organ dysfunction, necessitating intensive medical intervention. Neurotoxicity, another significant concern, can manifest as confusion, seizures, or cerebral edema, further complicating patient management.

To mitigate these risks, researchers are exploring various strategies, including the incorporation of safety switches into CAR constructs. These molecular switches allow for the controlled activation or deactivation of CAR T cells, providing a mechanism to halt the therapy in case of severe adverse reactions. For example, inducible caspase-9 (iCasp9) is a commonly used safety switch that can trigger apoptosis in CAR T cells upon administration of a small molecule, thereby controlling the therapy's intensity and duration.

2. Target Antigen Selection and Tumor Heterogeneity:

Another significant challenge is the identification of suitable target antigens that are expressed exclusively or predominantly on cancer cells but not on healthy tissues. This selectivity is crucial to minimize off-tumor, off-target effects that can result in damage to normal tissues. However, many tumor-associated antigens are also expressed at low levels on normal cells, leading to potential off-target toxicity.

Moreover, tumor heterogeneity poses a substantial challenge to the efficacy of CAR T-cell therapy. Tumors often consist of diverse cell populations with varying antigen expression profiles. This heterogeneity can lead to the emergence of antigen-negative tumor escape variants, which can evade CAR T-cell recognition and cause disease relapse. To address this, researchers are developing multi-specific CARs that target multiple antigens simultaneously, thereby reducing the likelihood of tumor escape. Additionally, strategies such as tandem CARs and bispecific T-cell engagers (BiTEs) are being explored to enhance targeting precision and therapeutic efficacy.

3. Manufacturing Challenges:

The production of CAR T cells is a complex and resource-intensive process. It involves the collection of T cells from the patient, genetic modification to express the CAR, expansion to achieve therapeutic doses, and quality control to ensure the product's safety and efficacy. This process can take several weeks, during which time the patient's condition may deteriorate.

To streamline this process, advancements are being made in automation and standardization of CAR T-cell manufacturing. Closed-system bioreactors, automated cell processing technologies, and gene-editing tools like CRISPR/Cas9 are being employed to improve efficiency and scalability. Additionally, the development of allogeneic CAR T cells, derived from healthy donors rather than the patient, holds promise for reducing manufacturing time and costs, potentially making the therapy more accessible.

4. Overcoming the Tumor Microenvironment

The immunosuppressive tumor microenvironment (TME) is another formidable barrier to the success of CAR T-cell therapy, particularly in solid tumors. The TME can inhibit T-cell infiltration, suppress T-cell activation, and promote immune evasion through various mechanisms, including the expression of inhibitory ligands, secretion of immunosuppressive cytokines, and recruitment of regulatory immune cells.

To overcome these challenges, researchers are exploring several strategies. These include the use of checkpoint blockade therapies, such as PD-1/PD-L1 inhibitors, in combination with CAR T-cell therapy to enhance T-cell function and persistence. Additionally, engineering CAR T cells to secrete cytokines like IL-12 or to express dominant-negative receptors can help modulate the TME and improve anti-tumor responses. Strategies to enhance T-cell trafficking to the tumor site, such as the expression of chemokine receptors on CAR T cells, are also being investigated.

5. Regulatory and Ethical Considerations

The regulatory landscape for CAR T-cell therapy is complex, involving stringent safety and efficacy requirements. Regulatory agencies such as the FDA and EMA closely monitor the development and clinical application of CAR T-cell therapies to ensure patient safety. This regulatory scrutiny can sometimes delay the approval and availability of new therapies.

Moreover, ethical considerations around the use of gene-editing technologies, patient selection, and access to treatment are significant. The high cost of CAR T-cell therapies limits accessibility, raising concerns about equity and fairness in healthcare. Efforts are underway to reduce costs through technological innovations and policy initiatives, but these issues remain a significant challenge.

Future Directions: Looking forward, the future of CAR T-cell therapy is likely to involve several key advancements:

1. **Next-Generation CARs:** Researchers are developing "armored" CARs with additional functionalities, such as resistance to immunosuppressive signals or the ability to secrete pro-inflammatory cytokines. These next-generation CARs aim to enhance efficacy, especially in solid tumors.
2. **Universal CAR T Cells:** The development of allogeneic or "off-the-shelf" CAR T cells, which can be produced in bulk and administered to multiple patients, could revolutionize the field by improving accessibility and reducing costs.
3. **Combination Therapies:** Combining CAR T-cell therapy with other treatment modalities, such as immune checkpoint inhibitors, chemotherapy, or radiotherapy, may enhance therapeutic outcomes and overcome resistance mechanisms.
4. **Personalized Medicine:** Advances in genomics and bioinformatics will likely lead to more personalized CAR T-cell therapies, tailored to the genetic and molecular profile of individual patients' tumors.

5. **Clinical Trials and Real-World Evidence:** Continued clinical trials and the collection of real-world data will be crucial for refining CAR T-cell therapies, understanding long-term outcomes, and expanding indications.

In conclusion, while CAR T-cell therapy holds tremendous promise for treating a variety of cancers, overcoming the current challenges will require a multifaceted approach involving innovative CAR designs, improved manufacturing processes, and strategies to mitigate toxicity and enhance efficacy. As research progresses, the integration of new technologies and approaches will likely expand the scope and impact of CAR T-cell therapies, making them a cornerstone of cancer treatment.

REFERENCES

- [1] G. Dotti, S. Gottschalk, B. Savoldo, and M. K. Brenner, “Design and development of therapies using chimeric antigen receptor-expressing T cells,” *Immunol. Rev.*, vol. 257, no. 1, pp. 107–126, Jan. 2014, doi: 10.1111/imr.12131.
- [2] A. D. Fesnak, C. H. June, and B. L. Levine, “Engineered T cells: the promise and challenges of cancer immunotherapy,” *Nat. Rev. Cancer*, vol. 16, no. 9, pp. 566–581, Sep. 2016, doi: 10.1038/nrc.2016.97.
- [3] R. Mohanty, C. Chowdhury, S. Arega, P. Sen, P. Ganguly, and N. Ganguly, “CAR T cell therapy: A new era for cancer treatment (Review),” *Oncol. Rep.*, Sep. 2019, doi: 10.3892/or.2019.7335.
- [4] M.-R. Benmebarek, C. H. Karches, B. L. Cadilha, S. Lesch, S. Endres, and S. Kobold, “Killing Mechanisms of Chimeric Antigen Receptor (CAR) T Cells,” *Int. J. Mol. Sci.*, vol. 20, no. 6, p. 1283, Mar. 2019, doi: 10.3390/ijms20061283.
- [5] L. Labanieh, R. G. Majzner, and C. L. Mackall, “Programming CAR-T cells to kill cancer,” *Nat. Biomed. Eng.*, vol. 2, no. 6, pp. 377–391, Jun. 2018, doi: 10.1038/s41551-018-0235-9.
- [6] R. C. De Marco, H. J. Monzo, and P. M. Ojala, “CAR T Cell Therapy: A Versatile Living Drug,” *Int. J. Mol. Sci.*, vol. 24, no. 7, p. 6300, Mar. 2023, doi: 10.3390/ijms24076300.
- [7] K. K. Jain, “Personalized Immuno-Oncology,” in *Textbook of Personalized Medicine*, Cham: Springer International Publishing, 2021, pp. 479–508. doi: 10.1007/978-3-030-62080-6_20.
- [8] E. Drent *et al.*, “Combined CD28 and 4-1BB Costimulation Potentiates Affinity-tuned Chimeric Antigen Receptor-engineered T Cells,” *Clin. Cancer Res.*, vol. 25, no. 13, pp. 4014–4025, Jul. 2019, doi: 10.1158/1078-0432.CCR-18-2559.
- [9] L. Jafarzadeh, E. Masoumi, K. Fallah-Mehrjardi, H. R. Mirzaei, and J. Hadjati, “Prolonged Persistence of Chimeric Antigen Receptor (CAR) T Cell in Adoptive Cancer Immunotherapy: Challenges and Ways Forward,” *Front. Immunol.*, vol. 11, p. 702, Apr. 2020, doi: 10.3389/fimmu.2020.00702.
- [10] R. C. Sterner and R. M. Sterner, “CAR-T cell therapy: current limitations and potential strategies,” *Blood Cancer J.*, vol. 11, no. 4, p. 69, Apr. 2021, doi: 10.1038/s41408-021-00459-7.
- [11] A. Feldmann *et al.*, “Versatile chimeric antigen receptor platform for controllable and combinatorial T cell therapy,” *OncImmunity*, vol. 9, no. 1, p. 1785608, Jan. 2020, doi: 10.1080/2162402X.2020.1785608.

- [12] R. Abrantes, H. O. Duarte, C. Gomes, S. Wälchli, and C. A. Reis, “CAR-Ts: new perspectives in cancer therapy,” *FEBS Lett.*, vol. 596, no. 4, pp. 403–416, Feb. 2022, doi: 10.1002/1873-3468.14270.
- [13] J. A. Kyte, “Strategies for Improving the Efficacy of CAR T Cells in Solid Cancers,” *Cancers*, vol. 14, no. 3, p. 571, Jan. 2022, doi: 10.3390/cancers14030571.
- [14] M. Cartellieri *et al.*, “Chimeric Antigen Receptor-Engineered T Cells for Immunotherapy of Cancer,” *J. Biomed. Biotechnol.*, vol. 2010, pp. 1–13, 2010, doi: 10.1155/2010/956304.
- [15] H. M. Finney, A. D. G. Lawson, C. R. Bebbington, and A. N. C. Weir, “Chimeric Receptors Providing Both Primary and Costimulatory Signaling in T Cells from a Single Gene Product,” *J. Immunol.*, vol. 161, no. 6, pp. 2791–2797, Sep. 1998, doi: 10.4049/jimmunol.161.6.2791.
- [16] X. Xu, H. Li, and C. Xu, “Structural understanding of T cell receptor triggering,” *Cell. Mol. Immunol.*, vol. 17, no. 3, pp. 193–202, Mar. 2020, doi: 10.1038/s41423-020-0367-1.
- [17] E. Roselli *et al.*, “4-1BB and optimized CD28 co-stimulation enhances function of human mono-specific and bi-specific third-generation CAR T cells,” *J. Immunother. Cancer*, vol. 9, no. 10, p. e003354, Oct. 2021, doi: 10.1136/jitc-2021-003354.
- [18] K. J. Curran, H. J. Pegram, and R. J. Brentjens, “Chimeric antigen receptors for T cell immunotherapy: current understanding and future directions,” *J. Gene Med.*, vol. 14, no. 6, pp. 405–415, Jun. 2012, doi: 10.1002/jgm.2604.
- [19] P. George *et al.*, “Third-generation anti-CD19 chimeric antigen receptor T-cells incorporating a TLR2 domain for relapsed or refractory B-cell lymphoma: a phase I clinical trial protocol (ENABLE),” *BMJ Open*, vol. 10, no. 2, p. e034629, Feb. 2020, doi: 10.1136/bmjopen-2019-034629.
- [20] M. Chmielewski and H. Abken, “TRUCKs: the fourth generation of CARs,” *Expert Opin. Biol. Ther.*, vol. 15, no. 8, pp. 1145–1154, Aug. 2015, doi: 10.1517/14712598.2015.1046430.
- [21] J.-Y. Wang and L. Wang, “CAR-T cell therapy: Where are we now, and where are we heading?,” *Blood Sci.*, vol. 5, no. 4, pp. 237–248, Nov. 2023, doi: 10.1097/BS9.0000000000000173.
- [22] M. Qayed *et al.*, “Leukapheresis guidance and best practices for optimal chimeric antigen receptor T-cell manufacturing,” *Cytotherapy*, vol. 24, no. 9, pp. 869–878, Sep. 2022, doi: 10.1016/j.jcyt.2022.05.003.

- [23] B. L. Levine, J. Miskin, K. Wonnacott, and C. Keir, “Global Manufacturing of CAR T Cell Therapy,” *Mol. Ther. - Methods Clin. Dev.*, vol. 4, pp. 92–101, Mar. 2017, doi: 10.1016/j.omtm.2016.12.006.
- [24] A. X. Wang, X. J. Ong, C. D’Souza, P. J. Neeson, and J. J. Zhu, “Combining chemotherapy with CAR-T cell therapy in treating solid tumors,” *Front. Immunol.*, vol. 14, p. 1140541, Mar. 2023, doi: 10.3389/fimmu.2023.1140541.
- [25] S. Singh, S. Khasbage, R. Kaur, J. Sidhu, and B. Bhandari, “Chimeric antigen receptor T cell: A cancer immunotherapy,” *Indian J. Pharmacol.*, vol. 54, no. 3, p. 226, 2022, doi: 10.4103/ijp.ijp_531_20.
- [26] R. Smith, “Bringing cell therapy to tumors: considerations for optimal CAR binder design,” *Antib. Ther.*, vol. 6, no. 4, pp. 225–239, Oct. 2023, doi: 10.1093/abt/tbad019.
- [27] S. S. Kenderian, D. L. Porter, and S. Gill, “Chimeric Antigen Receptor T Cells and Hematopoietic Cell Transplantation: How Not to Put the CART Before the Horse,” *Biol. Blood Marrow Transplant.*, vol. 23, no. 2, pp. 235–246, Feb. 2017, doi: 10.1016/j.bbmt.2016.09.002.
- [28] R. Pandey, C.-C. Chiu, and L.-F. Wang, “Immunotherapy Study on Non-small-Cell Lung Cancer (NSCLC) Combined with Cytotoxic T Cells and miRNA34a,” *Mol. Pharm.*, vol. 21, no. 3, pp. 1364–1381, Mar. 2024, doi: 10.1021/acs.molpharmaceut.3c01040.
- [29] M. Koneru, T. J. Purdon, D. Spriggs, S. Koneru, and R. J. Brentjens, “IL-12 secreting tumor-targeted chimeric antigen receptor T cells eradicate ovarian tumors *in vivo*,” *OncoImmunology*, vol. 4, no. 3, p. e994446, Mar. 2015, doi: 10.4161/2162402X.2014.994446.
- [30] G. López-Cantillo, C. Urueña, B. A. Camacho, and C. Ramírez-Segura, “CAR-T Cell Performance: How to Improve Their Persistence?,” *Front. Immunol.*, vol. 13, p. 878209, Apr. 2022, doi: 10.3389/fimmu.2022.878209.
- [31] G. Guzman, M. R. Reed, K. Bielamowicz, B. Koss, and A. Rodriguez, “CAR-T Therapies in Solid Tumors: Opportunities and Challenges,” *Curr. Oncol. Rep.*, vol. 25, no. 5, pp. 479–489, May 2023, doi: 10.1007/s11912-023-01380-x.
- [32] S. Srivastava and S. R. Riddell, “Chimeric Antigen Receptor T Cell Therapy: Challenges to Bench-to-Bedside Efficacy,” *J. Immunol.*, vol. 200, no. 2, pp. 459–468, Jan. 2018, doi: 10.4049/jimmunol.1701155.
- [33] M. S. Choudhery, T. Arif, R. Mahmood, and D. T. Harris, “CAR-T-Cell-Based Cancer Immunotherapies: Potentials, Limitations, and Future Prospects,” *J. Clin. Med.*, vol. 13, no. 11, p. 3202, May 2024, doi: 10.3390/jcm13113202.

- [34] M. D. Jain, M. Smith, and N. N. Shah, “How I Treat Refractory CRS and ICANS Following CAR T-cell Therapy,” *Blood*, p. blood.2022017414, Mar. 2023, doi: 10.1182/blood.2022017414.
- [35] P. Celichowski *et al.*, “Tuning CARs: recent advances in modulating chimeric antigen receptor (CAR) T cell activity for improved safety, efficacy, and flexibility,” *J. Transl. Med.*, vol. 21, no. 1, p. 197, Mar. 2023, doi: 10.1186/s12967-023-04041-6.
- [36] C. Lonez and E. Breman, “Allogeneic CAR-T Therapy Technologies: Has the Promise Been Met?,” *Cells*, vol. 13, no. 2, p. 146, Jan. 2024, doi: 10.3390/cells13020146.
- [37] B. Bravi, “Development and use of machine learning algorithms in vaccine target selection,” *Npj Vaccines*, vol. 9, no. 1, p. 15, Jan. 2024, doi: 10.1038/s41541-023-00795-8.
- [38] A. M. Mc Laughlin, P. A. Milligan, C. Yee, and M. Bergstrand, “Model-informed drug development of autologous CAR-T cell therapy: Strategies to optimize CAR-T cell exposure leveraging cell kinetic/dynamic modeling,” *CPT Pharmacomet. Syst. Pharmacol.*, vol. 12, no. 11, pp. 1577–1590, Nov. 2023, doi: 10.1002/psp4.13011.
- [39] K. G. Daniels *et al.*, “Decoding CAR T cell phenotype using combinatorial signaling motif libraries and machine learning,” *Science*, vol. 378, no. 6625, pp. 1194–1200, Dec. 2022, doi: 10.1126/science.abq0225.
- [40] Y. Li, X. Wu, D. Fang, and Y. Luo, “Informing immunotherapy with multi-omics driven machine learning,” *Npj Digit. Med.*, vol. 7, no. 1, p. 67, Mar. 2024, doi: 10.1038/s41746-024-01043-6.
- [41] S. Das, M. K. Dey, R. Devireddy, and M. R. Gartia, “Biomarkers in Cancer Detection, Diagnosis, and Prognosis,” *Sensors*, vol. 24, no. 1, p. 37, Dec. 2023, doi: 10.3390/s24010037.
- [42] J. Chen *et al.*, “A promising prognostic model for predicting survival of patients with HIV -related diffuse large B-cell lymphoma in the CART era,” *Cancer Med.*, vol. 12, no. 11, pp. 12470–12481, Jun. 2023, doi: 10.1002/cam4.5957.
- [43] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, “Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology,” *Nat. Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019, doi: 10.1038/s41571-019-0252-y.
- [44] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, “Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis,” *Genomics Proteomics Bioinformatics*, vol. 20, no. 5, pp. 850–866, Oct. 2022, doi: 10.1016/j.gpb.2022.11.003.

- [45] F. M. Howard, S. Kochanny, M. Koshy, M. Spiotto, and A. T. Pearson, “Machine Learning–Guided Adjuvant Treatment of Head and Neck Cancer,” *JAMA Netw. Open*, vol. 3, no. 11, p. e2025881, Nov. 2020, doi: 10.1001/jamanetworkopen.2020.25881.
- [46] G. Li and B. Yao, “Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches,” *Int. J. Des.*, vol. 7, no. 1, 2018.
- [47] N. J. Schork, “Artificial Intelligence and Personalized Medicine,” in *Precision Medicine in Cancer Therapy*, vol. 178, D. D. Von Hoff and H. Han, Eds., in *Cancer Treatment and Research*, vol. 178. , Cham: Springer International Publishing, 2019, pp. 265–283. doi: 10.1007/978-3-030-16391-4_11.
- [48] Y. Cai *et al.*, “Artificial intelligence applied in neoantigen identification facilitates personalized cancer immunotherapy,” *Front. Oncol.*, vol. 12, p. 1054231, Jan. 2023, doi: 10.3389/fonc.2022.1054231.
- [49] S. A. Alowais *et al.*, “Revolutionizing healthcare: the role of artificial intelligence in clinical practice,” *BMC Med. Educ.*, vol. 23, no. 1, p. 689, Sep. 2023, doi: 10.1186/s12909-023-04698-z.
- [50] G. Krishnan *et al.*, “Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm,” *Front. Artif. Intell.*, vol. 6, p. 1227091, Aug. 2023, doi: 10.3389/frai.2023.1227091.
- [51] Y. Derbal, “Adaptive Cancer Therapy in the Age of Generative Artificial Intelligence,” *Cancer Control*, vol. 31, p. 10732748241264704, Jan. 2024, doi: 10.1177/10732748241264704.
- [52] M. Wu, X. Du, R. Gu, and J. Wei, “Artificial Intelligence for Clinical Decision Support in Sepsis,” *Front. Med.*, vol. 8, p. 665464, May 2021, doi: 10.3389/fmed.2021.665464.
- [53] A. Naghizadeh *et al.*, “In vitro machine learning-based CAR T immunological synapse quality measurements correlate with patient clinical outcomes,” *PLOS Comput. Biol.*, vol. 18, no. 3, p. e1009883, Mar. 2022, doi: 10.1371/journal.pcbi.1009883.
- [54] V. Y. Odeh-Couvertier *et al.*, “Predicting T-cell quality during manufacturing through an artificial intelligence-based integrative multiomics analytical platform,” *Bioeng. Transl. Med.*, vol. 7, no. 2, p. e10282, May 2022, doi: 10.1002/btm2.10282.
- [55] N. Bäckel *et al.*, “Elaborating the potential of Artificial Intelligence in automated CAR-T cell manufacturing,” *Front. Mol. Med.*, vol. 3, p. 1250508, Sep. 2023, doi: 10.3389/fmmed.2023.1250508.
- [56] J. Dias, J. Garcia, G. Agliardi, and C. Roddie, “CAR-T cell manufacturing landscape—Lessons from the past decade and considerations for early clinical

development,” *Mol. Ther. - Methods Clin. Dev.*, vol. 32, no. 2, p. 101250, Jun. 2024, doi: 10.1016/j.omtm.2024.101250.

- [57] C. K. Chou and C. J. Turtle, “Assessment and management of cytokine release syndrome and neurotoxicity following CD19 CAR-T cell therapy,” *Expert Opin. Biol. Ther.*, vol. 20, no. 6, pp. 653–664, Jun. 2020, doi: 10.1080/14712598.2020.1729735.
- [58] M. Boettcher, A. Joechner, Z. Li, S. F. Yang, and P. Schlegel, “Development of CAR T Cell Therapy in Children—A Comprehensive Overview,” *J. Clin. Med.*, vol. 11, no. 8, p. 2158, Apr. 2022, doi: 10.3390/jcm11082158.
- [59] A. C. Uscanga-Palomeque *et al.*, “CAR-T Cell Therapy: From the Shop to Cancer Therapy,” *Int. J. Mol. Sci.*, vol. 24, no. 21, p. 15688, Oct. 2023, doi: 10.3390/ijms242115688.
- [60] “Budryte - Therapy selection gets AI assistance.pdf.”
- [61] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
- [62] *distilbert/distilbert-base-uncased*.