Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 1, 2024

Biology-based AI Predicts T-cell Receptor Antigen Binding Specificity

Xinyu Shen[1, *], Baoming Wang[2], Zheng He[3], Huiming Zhou[4], Yanlin Zhou[5]

[1]Biostatistics Columbia University, USA
[2]Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA
[3]Applied Analytics, Columbia University, NY, USA
[4]Computer Science, Northeastern University, CA, USA
[5]Computer Science, Johns Hopkins University, Baltimore, MD, USA
[*]Corresponding author: Xinyu Shen (Email: xshen1007@gmail.com)

Abstract: Adoptive cell transfer (ACT) using T cells modified by the T cell receptor (TCR) gene is an exciting and rapidly developing field. Numerous preclinical and clinical studies have shown varying feasibility, safety, and efficacy of using TCR-engineered T cells to treat cancer and viral infections. Although there is evidence that their use is effective, to what extent and how these therapies can be improved is still a question of research. Since TCR affinity has been generally accepted as the primary role in defining T cell specificity and sensitivity, selecting and generating high-affinity TCRS remains a fundamental approach to designing more effective T cells. However, the traditional approach of increasing affinity by random mutagenesis can cause adverse cross-reactions that result in on-target and off-target adverse events, produce depleted effectors through overstimulation, and ignore other kinetic and cellular parameters that have been shown to affect antigen specificity. In this paper, we review the preclinical and clinical potential of TCR-modified T cells, summarize contributions that challenge the role of TCR affinity in antigen recognition, and explore how structure-guided design can be used to manipulate antigen specificity and TCR cross-reactivity to improve the safety and efficacy of TCR-modified T cells for ACT.

Keywords: Gene prediction; Artificial intelligence; T cell receptor; Antibody binding.

1. Introduction

T cells typically recognize antigens on members of the MHC protein family through highly diverse heterodimer T cell receptors (TCRs) expressed on the surface. These antigens are usually short peptide fragments of eight or more residues, and their presentation is largely dependent on the structural preferences of the MHC alleles. Lipids, metabolites and oligosaccharide T cell antigens have also been reported. [1]TCR usually binds to the antigen MHC complex via one or more of its six complementary determining rings (CDRS), with each chain of the TCR dimer contributing three.

The critical role of TCR in disease surveillance and response, as well as in the development of new vaccines and therapies, has driven a concerted effort to decode the rules by which T cells recognize homologous antigen-MHC complexes. [2-3]However, cost and experimental limitations limit the available databases to only a small fraction of the possible sample space for TCR-antigen-binding pairs. These datasets are also not well representative of the self and pathogenicity epitopes and the various MHC environments in which they may occur.

It is important to distinguish between TCR specificity and antigen immunogenicity. The former predicts the binding between the TCR set and the antigen MHC complex. The latter can be described as predicting whether a given antigen will induce a functional T cell immune response, i.e., a complex chain of events involving antigen expression, processing and presentation, TCR binding, T cell activation, amplification, and effector differentiation. Although great progress has been made in improving antigen processing and presentation prediction of common HLA alleles, the nature and extent to which presenting peptides trigger T cell responses remains to be elucidated. There are significant gaps in predicting [4-6]T cell activation for a given peptide, and the parameters that influence pathologic peptide or neoantigen immunogenicity are still being studied in depth. Only by integrating knowledge of antigen presentation, TCR recognition, dependent activation, and effector function at the cellular and tissue levels can the benefits of basic and translational science be fully realized. This paper addresses the problem of predicting TCR antigen specificity and elaborates on the general requirements for antigen-binding prediction models, highlights key challenges, and discusses how recent advances in digital biology such as single-cell technology and machine learning offer possible solutions. [7-8]Finally, they describe how predicting TCR specificity can help us understand the broader puzzle of antigen immunogenicity.

2. Related Work

2.1. Cytodynamics of CAR-Ts

The cytodynamics of circulating CAR Ts (which has properties of pharmacokinetics) are divided into three distinct phases: initial expansion, followed by rapid contraction, followed by slow long-term decay 3. Degree of cell expansion (Cmax) and long-term exposure (area under the curve (AUC)) vary widely among patients (about three orders of magnitude) and can predict efficacy (tumor volume reduction) and toxic4. However, the product intrinsic factors and host intrinsic factors that mediate this pharmacology are still not clearly defined.[9] An empirical nonlinear mixed-effect model was developed to quantify the pharmacokinetics of Kymriah (tisagenlecleul, CTL019)5 and provided as part of the Biologics License Application (BLA)4. This formulation has

been shown to be suitable for a variety of other CAR-T therapies for a variety of indications and has been adopted by the FDA as a benchmark. [10-13]Although the empirical equation is a useful tool for quantifying clinical data, it does not explain the underlying biology and is therefore of limited value in modeling the effects of alternate CAR-T designs, cell origins, or treatment regimens. A mathematical model that can quantitatively describe clinical data, also based on sound biological mechanisms, is useful for the development of novel CAR-T products, just as systematic pharmacological models have been shown for other therapeutic modalities.

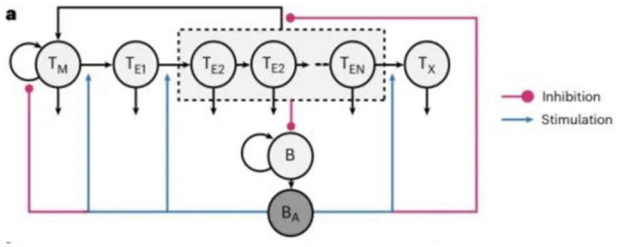

Figure 1. CAR-Ts model architecture

We think that T cells (and CAR-T products) are composed of three functionally distinct cell populations: T memory cells (TM), capable of long-term self-renewal and immune memory; T-effector (TE), which is responsible for target-mediated cell death; As well as depleted T cells (TX), which lack killing potential and proliferation capacity. The antigen-sensing switch coordinates the decision of self-renewal and differentiation of memory cells, the effector proliferation rate, depletion, and the rate of regeneration of memory cells from effectors (method). [14]This represents a conceptually simple but biologically sound description of T cell function and regulatory control in response to immune demands determined by the burden of systemic antigens (Figure 1).

The first attempt was to determine whether mathematical descriptions of T cell regulatory control could quantitatively capture CAR-T pharmacokinetics and tumor dynamics profiles, and whether parameter estimates revealed the biological basis of clinical variability. Fraietta et al. 18 [15]report average pharmacokinetics and tumor dynamics profiles of patients with chronic lymphoma (CLL) treated with Kymriah (CTL019, a CD19-targeted CAR T), divided by complete responders (CR), partial responders (PR), and non-responders (NR). We digitized the data (mean ±s.d.) and used particle swarm optimization (PSO) to estimate the model parameters characterizing the three population prototypes (Figure 2).

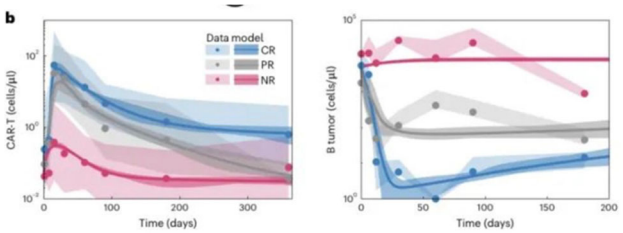

Figure 2. Particle swarm optimization (PSO) parameter model

Single sample gene set enrichment analysis (ssGSEA) is used to examine the distribution of pathway and cell characteristics in a single sample. The CR population was significantly abundant in the "non-depleted T cell" feature (Figure 1), consistent with simulations, in which the CR group had a significantly higher proportion of non-depleted cells on day 60 (peak anti-tumor effect) (Figure 2), while the cells of patients with NR rapidly progressed to exhaustion. These simulations are also consistent with clinical reports that CAR-T products that fail to expand in vivo exhibit enhanced expression of fatigue markers LAG3 and PD1[16-18].

2.2. pMT training and prediction

At present, with the development of genome sequencing technology, major histocompatibility complex (MHC) epitope database and prediction algorithm, it has become possible to identify and screen tumor neoantigens of individual patients. However, the complexity of T cells combined with the uncertainty of tumor variation makes it impossible to predict the specific binding of tumor neoplasm antigen to T cells, which is one of the urgent problems in modern immunology.

Conventional methods (such as MHC tetramer assay and tetramer assay) are technically demanding, time-consuming and costly. [19][20]The deep learning model pMTnet established by Wang Tao's research group, through the high-speed computing advantages of artificial intelligence machine learning, saves researchers cumbersome experiment time and expensive experiment costs, and provides a platform prediction method.

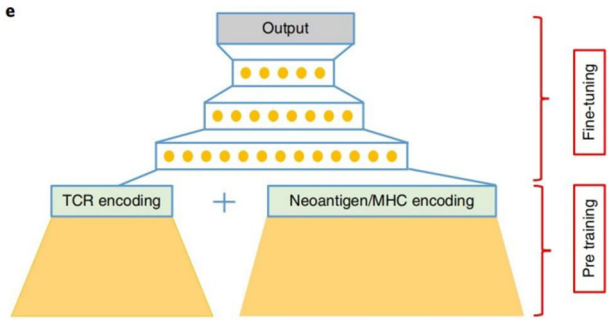

Figure 3. Final structure of the pMTnet model

The training and prediction of pMT is based on three main data points: the antigen sequence of T-cell-tumor cell binding, the sequence of the major tissue-compatible complex (MHC), and the sequence of the T-cell receptor. Through transfer-learning, pMTnet completes the training of deep learning models.

pMT is a microrNA that is hypothesized to play an important role in the nervous system. The training and prediction of pMT is important for the prediction of biological cells for the following reasons:

1. Understanding cellular regulatory mechanisms: [21]pMT may be involved in regulating gene expression of cells, affecting physiological functions and metabolic processes of cells. Through the training and prediction of pMT, we can better understand the mechanism of cell regulation and reveal the complex signal transduction network inside the cell.

2. Diagnose and treat diseases: Many diseases, such as cancer and neurodegenerative diseases, are closely related to abnormalities in intracellular signaling. By analyzing the role of pMT in the occurrence and development of diseases, we can provide new targets and strategies for the diagnosis and treatment of related diseases.

3. Frontier of biomedical research: pMT, as a new type of regulatory factor, is still in its infancy in the field of cell

biology and neuroscience. The training and prediction of pMT help to promote cutting-edge research in this field and provide new ideas and approaches for the development of new drugs and the discovery of therapeutic methods.

4. Personalized medicine: [22-23]By analyzing the expression and regulation of pMT in individual cells, important information can be provided for personalized medicine. Training and prediction based on pMT can achieve accurate diagnosis and treatment for individuals and improve medical outcomes and prognosis.

3. Methodology

In order to deepen the important advantages of artificial intelligence in biomedical detection and cell prediction, our understanding of the mechanism of disease provides a way to accelerate the development of safer and personalized vaccines and therapies, construct a complete map of TCR-antigen interaction as an experimental source, and explore the core content of why TCR binding specificity is universal prediction.

3.1. Experimental method

Antigen-mhc polymers can be used to determine TCR specificity using large (combined) T cell populations or newer single-cell methods. Most methods are widely used and relatively inexpensive, but do not provide information about αβTCR chain pairing or function. [24]As a result, single-strand TCR sequences dominate public datasets. However, both the alpha and beta chains contribute to antigen recognition and specificity. Multimodal single-cell techniques provide insights into chain pairing, transcriptome, and phenotypic characteristics at cellular resolution, but are still very expensive, return fewer TCR sequences per run compared to batch experiments, and show a significant bias against highly specific TCRS.[25] Appropriate experimental protocols for reducing the binding of nonspecific polymers, verification of correct folding, and computational improvements in the signal-to-noise ratio remain areas of active debate.

Due to these scalability barriers, only a small fraction of the total sample space of TCR antigen pairs has been experimentally validated. Fewer than 1 million unique TCR epitope pairs are available from VDJdb, McPasTCR, immune epitope databases, and MIRA datasets. Only 4% of these instances contained complete chain pairing information.

3.2. Function calculation method

Computational models at this stage can be divided into two categories, the supervised predictive model (SPM) and the unsupervised clustering model (UCM), because they use supervised and unsupervised learning respectively. The first is supervised prediction models, SPMS are those that try to learn a function that will correctly predict the homologous epitopes of a given input TCR with unknown specificity, given some training data set of known TCR-peptide pairs. Over the past two years, there has been an increase in the number of publications aimed at solving this challenge through deep neural networks (DNNS). [26]Although there are many possible ways to compare SPM performance, the most common method is the area under the receiver operating Characteristic Curve (ROC-AUC). Assuming a proportional equilibrium of negative and positive pairs, one would expect 50% ROC-AUC to be observed from random guesses in binary (combined or uncombined) tasks.

Together, these results highlight the urgent need for comprehensive, independent benchmarking studies of data set models that are compiled and analyzed in a consistent manner. Until then, newer models could reasonably be used to predict the binding of common HLA alleles to immunodominant viral epitopes. However, SPM should be used with caution when generalizing to the prediction of any epitope, as performance may degrade the further an epitope is from the sequence of epitopes in the training set.

3.3. Major challenges in T cell antigen-specific prediction of TCR

Despite the exponential growth of unlabeled immune library data and recent unprecedented breakthroughs in data science and artificial intelligence, quantitative immunology still lacks a framework for systematic and universal inferences about TCR T cell antigen specificity. The most plausible explanations are data limitations, methodological gaps, and incomplete basic immunological models.

(1) Data

The single most important limitation in model development is the availability of high-quality TCR and antigen MHC pairs. The need is most urgent for underrepresented antigens, antigens with low HLA allele frequency, and the association between epitope specificity and T cell function. [27]Meanwhile, single-cell multimodal techniques have generated hundreds of millions of unlabeled TCR sequences that correlate with transcriptomic, phenotypic, and functional information. However, this unlabeled data is not without significant limitations. It is important to note that biological factors such as age, gender, ethnicity, and disease background vary across studies and may influence immune responses. Differences in experimental protocols, sequence preprocessing, total variation filtering (denoising) and standardization between laboratory groups may also have an impact: it is likely that batch correction will need to be applied. As a result, a thoughtful approach to data integration, noise correction, processing, and annotation may be critical to advancing state-of-the-art predictive models.

(2) Modeling

The exponential growth of TCR data from single-cell technologies, as well as cutting-edge advances in artificial intelligence and machine learning, have brought TCR antigene-specific inference into focus. However, several key gaps must be addressed before a solution for generalized epitope-specific inference can be implemented.

First, models with TCR sequence input limited to coding using β-chain CDR3 rings and VDJ genes are only likely to tell part of the story of antigen recognition, and the extent to which single-strand pairing is sufficient to describe TCR antigen specificity remains an open question. Structural and statistical analyses showed that the α-chain and β-chain contribute equally to specificity, and combining the two chains can improve prediction performance.

However, chain pairing information is largely missing, and many of the most advanced SPMS and UCMs rely on single-strand information. Although the CDR3 ring may be primarily responsible for antigen recognition, residues of CDR1, CDR2, and even the framework regions of the alpha and beta chains may also be involved. Subtle compensatory changes in the peptide-MHC and TCR interaction network, changes in the binding pattern and conformational flexibility of TCR and MHC may support TCR cross-reactivity. Until recently, explicit encoding of structural information for specific

inference was limited to the study of the structure of a limited set of crystals.

(3) Immunology

It is now clear that the potential immunological relevance of T cell interactions with their homologous ligands is highly variable and only partially understood, which has important implications for model design. Importantly, TCR antigen-specific inference is only one piece of the larger puzzle of predicting antigen immunogenicity, which the authors divide into three stages: antigen processing and MHC presentation, TCR recognition, and T cell response.

Antigen processing and presentation pathways have been extensively studied, and computational models for predicting peptide binding affinity to certain MHC alleles, especially Class I HLA, have achieved near-perfect ROC AUC for common alleles. However, this problem is far from solved, especially for the lower frequency MHC Class I alleles and MHC Class II alleles.

4. Conclusion

First, a merge and validation library of labeled and unlabeled TCR data should be provided to facilitate model pre-training and system comparison. Second, efforts should be coordinated to improve coverage of TCR antigen pairs presented by less common [28-29]HLA alleles and non-viral epitopes. Continued publication of negative and positive TCR epitope-binding data is encouraged to produce balanced datasets. Third, an independent, unbiased, and systematic evaluation of model performance for SPM, UCM, and a combination of both is of great help. This comparison should use consistent evaluation measures, including but not limited to the ROC-AUC and the area under the precise recall curve, that describe the performance of common and rare HLA subtypes, visible and invisible TCRS, and epitopes. The validation strategy used in the ImRex and TITAN evaluations is encouraged to confirm model performance comparisons. In the future, TCR-specific inference data should be expanded to include multimodal information as a means of bridging from TCR binding to immunogenicity prediction.

The scale and complexity of this task means that an interdisciplinary consortium approach is needed to systematically combine the latest immunological understanding of cellular immunity with cutting-edge developments in artificial intelligence and data science at the organizational level. This should include experimental and computational immunologists, machine learning experts, and translation and industrial partners. Given the success of the critical evaluation of the protein structure prediction series, a similar approach is encouraged to address the enormous challenge of TCR-specific inferibility in the short term and ultimately to predict integrated T and B cell immunogenicity. Competing models should be made available for research based on commendable cases in protein structure prediction.

Acknowledgment

At the end of the article, I want to thank him from the bottom of my heart, Zheng et al., Application of K-means clustering based on artificial intelligence in gene statistics of biological information The outstanding work presented in the article "engineering". Their research provides profound inspiration and important reference basis for this paper, and deepens our understanding of artificial intelligence, biological monitoring, cell prediction and other fields.

In particular, the study has made significant progress in the field of genetic statistics and bioinformation engineering using the K-means clustering algorithm, which provides valuable empirical support for the related topics we explore in the paper. Their work not only enriches the academic community's understanding of the application of artificial intelligence in the biomedical field, but also provides us with new ideas and methods for interdisciplinary research.

In addition to thanking He, Zheng and others for their research, I would like to especially thank the author of this paper, whose professional guidance and support have played a key role in our research work. Throughout the research process, their advice and insights have provided us with valuable guidance and helped us overcome many difficulties

References

[1] “Unveiling the Future Navigating Next-Generation AI Frontiers and Innovations in Application”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 147-56, https://doi.org/10.62051/ijcsit.v1n1.20.

[2] Chen, Wangmei, et al. “Applying Machine Learning Algorithm to Optimize Personalized Education Recommendation System”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Feb. 2024, pp. 101-8, doi:10.53469/jtpes.2024.04(01).14.

[3] H. Zhu and B. Wang, "Negative Siamese Network for Classifying Semantically Similar Sentences," 2021 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2021, pp. 170-173, doi: 10.1109/IALP54817.2021.9675278.

[4] “The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier”. Academic Journal of Science and Technology, vol. 8, no. 2, Dec. 2023, pp. 57-61, https://doi.org/10.54097/ajst.v8i2.14945.

[5] Wei, Kuo, et al. “Strategic Application of AI Intelligent Algorithm in Network Threat Detection and Defense”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 49-57, doi:10.53469/jtpes.2024.04(01).07.

[6] He, Zheng & Shen, Xinyu & Zhou, Yanlin & Wang, Yong. (2024). Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering. 10.13140/RG.2.2.11207.47527.

[7] Pan, Linying & Xu, Jingyu & Wan, Weixiang & Zeng, Qiang. (2024). Combine deep learning and artificial intelligence to optimize the application path of digital image processing technology.

[8] Duan, Shiheng, et al. "THE INNOVATIVE MODEL OF ARTIFICIAL INTELLIGENCE COMPUTER EDUCATION UNDER THE BACKGROUND OF EDUCATIONAL INNOVATION." The 2nd International scientific and practical conference “Innovations in education: prospects and challenges of today”(January 16-19, 2024) Sofia, Bulgaria. International Science Group. 2024. 389 p.. 2024.

[9] Wan, Weixiang & Sun, Wenjian & Zeng, Qiang & Pan, Linying & Xu, Jingyu. (2024). Progress in artificial intelligence applications based on the combination of self-driven sensors and deep learning.

[10] Sun, Wenjian & Xu, Jingyu & Pan, Linying & Wan, Weixiang & Wang, Yong. (2024). Automatic driving lane change safety prediction model based on LSTM.

[11] Qian, Wenpin, et al. "NEXT-GENERATION ARTIFICIAL INTELLIGENCE INNOVATIVE APPLICATIONS OF LARGE LANGUAGE MODELS AND NEW METHODS."

OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 262.

[12] Liang, Penghao, et al. "Enhancing Security in DevOps by Integrating Artificial Intelligence and Machine Learning." Journal of Theory and Practice of Engineering Science 4.02 (2024): 31-37.

[13] Zhang, Chenwei, et al. "SegNet Network Architecture for Deep Learning Image Segmentation and Its Integrated Applications and Prospects." Academic Journal of Science and Technology 9.2 (2024): 224-229.

[14] Gong, Yulu, et al. "RESEARCH ON A MULTILEVEL PRACTICAL TEACHING SYSTEM FOR THE COURSE'DIGITAL IMAGE PROCESSING." OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 272.

[15] Wang, Yong, et al. "Autonomous Driving System Driven by Artificial Intelligence Perception Fusion." Academic Journal of Science and Technology 9.2 (2024): 193-198.

[16] Zhenghua Hu, Xianmei Wang, Kangming Xu, and Pu Dong. 2020. Real-time Target Tracking Based on PCANet-CSK Algorithm. In Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence (CSAI '19). Association for Computing Machinery, New York, NY, USA, 343–346. https://doi.org/10.1145/3374587.3374607.

[17] Chen, Jianhang, et al. "One-stage object referring with gaze estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[18] Zhang, Quan, et al. "Application of the AlphaFold2 Protein Prediction Algorithm Based on Artificial Intelligence." Journal of Theory and Practice of Engineering Science 4.02 (2024): 58-65.

[19] Wu, C., Cui, J., Xu, X. et al. The influence of virtual environment on thermal perception: physical reaction and subjective thermal perception on outdoor scenarios in virtual reality. Int J Biometeorol 67, 1291–1301 (2023). https://doi.org/10.1007/s00484-023-02495-3

[20] Shen, Zepeng, et al. "EDUCATIONAL INNOVATION IN THE DIGITAL AGE: THE ROLE AND IMPACT OF NLP TECHNOLOGY." OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 281.

[21] Wang, Yong & Ji, Huan & Zhou, Yanlin & He, Zheng & Shen, Xinyu. (2024). Construction and application of artificial intelligence crowdsourcing map based on multi-track GPS data. 10.13140/RG.2.2.24419.53288.

[22] Zheng, Jiajian & Xin, Duan & Cheng, Qishuo & Tian, Miao & Yang, Le. (2024). The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance.

[23] K. Xu, X. Wang, Z. Hu and Z. Zhang, "3D Face Recognition Based on Twin Neural Network Combining Deep Map and Texture," 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 2019, pp. 1665-1668, doi: 10.1109/ICCT46805.2019.8947113.

[24] “Exploring New Frontiers of Deep Learning in Legal Practice: A Case Study of Large Language Models”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 131-8, https://doi.org/10.62051/ijcsit.v1n1.18.

[25] Yang, Le & Tian, Miao & Xin, Duan & Cheng, Qishuo & Zheng, Jiajian. (2024). AI-Driven Anonymization: Protecting Personal Data Privacy While Leveraging Machine Learning.

[26] Cheng, Qishuo & Yang, Le & Zheng, Jiajian & Tian, Miao & Xin, Duan. (2024). Optimizing Portfolio Management and Risk Assessment in Digital Assets Using Deep Learning for Predictive Analysis.

[27] Duan, Shiheng, et al. “Prediction of Atmospheric Carbon Dioxide Radiative Transfer Model Based on Machine Learning”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 132-6, https://doi.org/10.54097/ObMPjw5n.

[28] Xiao, J., Chen, Y., Ou, Y., Yu, H., & Xiao, Y. (2024). Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for Dialogue Summarization. arXiv preprint arXiv:2401.15496.

[29] Sun, Guolin, et al. "Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs." Future Generation Computer Systems 104 (2020): 60-73.