**Introduction**

Cancer is the leading cause of death worldwide. In 2022, there were an estimated 20 million new cancer cases and 9.7 million deaths **(World Health Organization: WHO, 2024)**. Although traditional methods such as surgery and chemotherapy are effective in treating cancer to a certain extent, they are often not effective in preventing the metastatic spread of the disease through disseminated tumor cells **(Schuster et al., 2006)**. With the deepening research on the immune system, immunotherapy that can harness the immune system to fight cancer has now firmly established itself as a novel pillar of cancer care **(Esfahani et al., 2020)**.

The specificity of T cell receptor (TCR) plays a crucial role in immunotherapy. As an important class of lymphocytes in the immune system, T cells can recognize and bind to antigen peptides from the major histocompatibility complex (MHC) through the TCR on the surface to trigger an immune response to recognize and attack pathogens and abnormal cells. Researching the specificity of TCR contributes to the development of immunotherapy, but the various experimental methods used to identify the interactions between TCRs and peptides presented by MHC molecules (pMHCs) have limitations such as time-consuming, costly, or technically demanding **(Zhao et al., 2023)**. Therefore, algorithms that can accurately identify and predict TCR specificity need to be developed.

We establish a TCR generative specificity detection framework, which enables efficient screening of TCRs and specific antigens using distance and machine learning algorithms. Furthermore, the performance of this model is compared with some baseline models, which demonstrates a further improvement in the accuracy and confidence of the model in predicting TCR specificity.

**literature review**

We researched various computational models developed for TCR specificity prediction, which can be broadly divided into three categories according to the working principle.

The first class of models predicts the specificity of TCR based on three-dimensional structural information of TCR and pMHC. These models perform well when there is a need for detailed knowledge and high-resolution prediction of binding patterns and specificity. However, the structural basis of TCR activation is poorly understood **(Mariuzza et al., 2020)**, which reduces the accuracy and feasibility of the model.

The second class of models is based on the sequence information of the TCR and pMHC. The development of high-throughput sequencing methods and single-cell RNA sequencing technologies has enabled efficient and rapid derivation of large numbers of TCR sequences from donor samples, which are collated into rich datasets **(Shugay et al., 2017)**. It promotes the application of deep learning methods in the modeling of TCR-pMHC interactions. However, these models only utilize the sequence information of the TCR and ignore structural information. Therefore, the accuracy of this class of model may not be ideal if structural features need to be considered in the prediction.

The third class of models combines the characteristics of the first two classes of models, which can make full use of structure and sequence information to improve accuracy and reliability. In other words, it also has higher requirements for the variety and quantity of data. Therefore, these models may cause high computational costs and training difficulties if computing resources and training data are limited.

**Reference:**

Esfahani, K., Roudaia, L., Buhlaiga, N., Del Rincón, S. V., Papneja, N., & Miller, W. H. (2020). A review of cancer immunotherapy: From the past, to the present, to the future. *Current Oncology*, *27*(12), 87–97. <https://doi.org/10.3747/co.27.5223>

Mariuzza, R. A., Agnihotri, P., & Orban, J. (2020). The structural basis of T-cell receptor (TCR) activation: An enduring enigma. Journal of Biological Chemistry/the Journal of Biological Chemistry, 295(4), 914–925. [https://doi.org/10.1016/s0021-9258(17)49904-2](Mariuzza, R. A., Agnihotri, P., & Orban, J. (2020). The structural basis of T-cell receptor (TCR) activation: An enduring enigma. Journal of Biological Chemistry/the Journal of Biological Chemistry, 295(4), 914–925. https:/doi.org/10.1016/s0021-9258(17))

Schuster, M., Nechansky, A., & Kircheis, R. (2006). Cancer immunotherapy. *Biotechnology Journal*, *1*(2), 138–147. <https://doi.org/10.1002/biot.200500044>

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M. A., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Елисеев, А. В., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K., Clemens, E. B., . . . Chudakov, D. M. (2017). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, *46*(D1), D419–D427. [https://doi.org/10.1093/nar/gkx760](https://doi.org/10.3389/fimmu.2022.1055151)

World Health Organization: WHO. (2024, February 1). Global cancer burden growing, amidst mounting need for services. WHO. [https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for- services](https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services)

Zhao, M., Xu, X. S., Yang, Y., & Yuan, M. (2023). GGNPTCR: a generative graph structure neural network for predicting immunogenic peptides for T-cell immune response. Journal of Chemical Information and Modeling, 63(23), 7557–7567. <https://doi.org/10.1021/acs.jcim.3c01293>

**中文版：**

癌症是全球主要死因。2022 年，估计有 2000 万新发癌症病例和 970 万人死亡【】。手术、化疗等传统方式虽然一定程度上能有效治疗癌症，但这些方式在防止疾病通过播散性肿瘤细胞进行转移扩散通常是无效的【】。随着对免疫系统的研究不断深入，能够利用免疫系统来对抗癌症的免疫疗法现已成为癌症治疗的新支柱【】。

TCR的特异性在免疫疗法中起着至关重要的作用。T细胞作为免疫系统中一类重要的淋巴细胞，可以通过表面的T细胞受体（TCR）识别和结合来自主要组织相容性复合物 (MHC)中的抗原片段来触发免疫反应，识别并攻击病原体和异常细胞。研究TCR的特异性有助于有助于免疫疗法的发展，但用来识别T细胞受体(TCR)和MHC分子(pMHC)呈递的肽之间的相互作用的各种实验方法具有各方面的局限性，例如耗时长、成本高或技术要求高【】。因此，我们需要开发出能够精准识别和预测TCR特异性的算法。

我们提出了一种TCR生成式特异性检测框架（Specificity Detection Framework），旨在实现利用距离和机器学习算法高效地筛选出TCR与对应抗原。此外，我们将该模型的性能与一些基线模型进行了比较，证明了在预测TCR特异性时该模型的准确率和可信度有进一步的提高。

**文章综述**

我们研究了为实现TCR特异性预测而开发的多种计算模型。根据工作原理，这些模型大致可分为三类。

第一类模型基于TCR和MHC-抗原复合物的三维结构信息来预测TCR的特异性。当对结合模式和特异性有详细了解和高分辨率的预测要求时，这类模型表现良好。但目前人们对TCR激活的结构基础还知之甚少(Mariuzza et al., 2020)，这降低了模型的准确度和可行度。

第二类模型基于TCR和MHC-抗原复合物的序列信息进行预测。高通量测序方法和单细胞 RNA 测序技术的发展实现了从供体样本中高效、快速地获得大量 TCR 序列，这些数据被整理成丰富的数据集(Shugay et al., 2017)。 这促进了深度学习方法在TCR-pMHC 相互作用建模中的运用。然而，此类模型只利用了TCR的序列信息，而忽略了其结构信息。因此，纯序列基于模型的准确率可能不够理想如果需要在预测时考虑结构特征。

第三类模型结合了前两类模型的特点，利用TCR和MHC-抗原复合物的结构和序列信息进行综合预测。这类模型可以充分利用结构和序列信息来提高预测的准确性和可靠性。但这对数据的种类和数量也有更高的要求。也就是说，当计算资源和训练数据有限时，复杂的组合模型可能会造成过高的计算成本和训练难度。