Llama-TCR: Generate De Novo TCR with Large Language Model

Jun Zhou, Zhenzhou Wu, Mengren Man, Yonghui Wu, Le Dinh Linh, Manna Dai, Yong Liu, Rick Goh
Institute of High Performance Computing
Agency for Science, Technology and Research (A*STAR)
Singapore
Email: {zhou_jun, wu_zhenzhou, mengren_man, wu_yonghui, linh_le_dinh, manna_dai, liuyong, gohsm}@ihpc.a-star.edu.sg

***Abstract*—T cell receptor-engineered T cell (TCR-T), where T cells are equipped with engineered antigen-specific receptors, is a promising approach of cancer immunotherapy. However, it is known that naturally occurred TCR is not able to detect some cancer antigens due to negative selection and resulted in the proliferation of cancer cells. This study introduces a novel application of large language models (LLM) in the de novo generation of T cell receptors (TCR). We propose Llama-TCR, a generative model trained on database of antigen and TCR pairs, for diverse and functional candidate TCR sequence generation. Using both sequence based benchmarking and 3D structure inspection, we show that the model is able to generate TCR sequences targeting to given antigens and has great potential in the development of novel TCR-T immunotherapy.**

***Index Terms*—Large Language Model, Generative AI, TCR, Immunotherapy**

I. Introduction

T cell receptor-engineered T cell (TCR-T) [1] is a promising approach of cancer immunotherapy, where T cells are equipped with engineered antigen-specific receptors to recognize and attack cancer cells. However, it is known that naturally occurred TCR cannot detect some cancer antigens due to negative selection or immuno-suppression of naturally positive TCR and resulted in the proliferation of cancer cells [2]. To tackle this issue, T cell Chimeric Antigen Receptor (CAR-T) [2, 3] and TCR-T was invented by finding a variable fragment from a large database that has a close affinity with the neoantigen, thus equipped the T cell with the recognition capability for the target antigen. However, due to the proximity of the cancer peptide to some healthy peptides, and the low specificity of the transplanted receptor, there is often strong autoimmune response from CAR-T or TCR-T treatment [4].

To overcome this issue, we propose a solution whereby we can make use of large language models (LLM) to generate de novo receptors that have close to perfect specificity for a target antigen. We hypothesize that for naturally occurring receptors, it will be exceedingly difficult to evolve into super specific receptors (SSR), although there is research that indicates such receptors may actually exist [5]. Our goal is to build a library of SSR using LLM such that, we can administer human specific combinations of SSR based on the cancer antigens profile of the patient and ultimately to test whether such a framework of precise and targeted treatment is effective and safe clinically in future.

With advent of Alphafold [6], for the first time, we can translate DNA codes into protein structure directly and instantly with extremely high accuracy and render this problem solved by the organizer of the CASP14 protein folding prediction competition [7]. This breakthrough leads to translation of millions of proteins within hours, which if by conventional route of cryoelelctron microscopy, will take years. The most exhilarating outcome from this breakthrough is the possibility of designing and engineering all kinds of biological functions from scratch into cells, which will open a new era of de novo biological designs. A lot of interesting works have already started in this direction. From designing de novo luciferase [8] to designing general binding protein for target structure [9, 10, 11, 12]. We see the great potential of using neural network to create new recognition receptors for immune cells to give them much more powerful recognition capability.

II. Related Works

A. Protein Language Model

Based on the nature that proteins are encoded by sequence of amino acid, researchers have been applying advancement of transformer based large language model to protein domain. Protein language models has emerged as a transformative tool in computational biology to understand and predict protein structures and functions. The development of protein language model focuses on both protein structures and sequences.

In the area of protein structure, multiple sequence alignment (MSA) plays a crucial role. This technique aligns protein/DNA sequences from related organisms and highlights conserved regions and evolutionary relationships. Rao et al. developed MSA Transformer [13] in 2021, discovered that training across multiple sequence alignments (MSA) significantly improves unsupervised structure learning methods for proteins. AlphaFold [6] represents a significant breakthrough in protein structure prediction, utilizing deep learning to accurately predict protein structures. It predicted over 200 million protein structures and built AlphaFold DB, an open access database. Meier et al. developed Evolutionary Scale Modeling (ESM) [14], using transformer-based models to enable zero-shot prediction of the effects of mutations on protein function.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00158

In 2022, an open atlas of 772 million predicted metagenomic protein structures is released based on ESM [15].

In the area of protein sequences, efforts are made to adapt large language models from natural language domain to protein domain, especially in protein sequence completion and de novo protein generation. ProteinBERT [16] is a deep language model specifically for proteins. It achieves state-of-the-art performance on diverse protein properties using a smaller model than competing methods. ProtGPT2 [17], a deep unsupervised language model for protein design is capable of generating de novo protein sequences. Madani et al. introduced ProGen [18], a 1.2B-parameter language model, effective in generating protein sequences with diversity.

B. Structure Based Methods

1) Diffusion models: Diffusion models are a class of generative neural network models capable of handling high dimensional data trained by adding and removing noises to input data [19, 20]. Owing to their unique capacities, diffusion models have found tremendous success in the generative modeling of images and languages, producing high-quality images while requiring fewer computational resources compared to traditional generative methods such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). In bioinformatics and computational biology, diffusion models are finding increasing applications in protein design and generation due to their ability to produce diverse outputs.

2) Protein backbone generation using diffusion models: The structure-first approach to protein design was poineered by the David lab at the University of Washington. In this approach, the process for designing a protein starts with generating the protein's structure as a set of coordinates known as the backbone. More specifically, a protein's backbone is a continuous chain of four atoms that runs throughout the length of a protein. These four atoms are nitrogen, alpha-carbon, carbon, and oxygen. Alpha-carbon is the central point for each amino acid residue within the protein, while the coordinates of the other three atoms relative to alpha-carbon uniquely determines the orientation of the residue.

The starting point of structure-first approach to protein design is backbone generation [11, 12]. The workflow takes advantage of the deep understanding of protein structures and sequence-structure relationships already implicit in powerful structure prediction models such as AlphaFold2 and RoseTTAFold, and inverse-folding models like ProteinMPNN and ESM-IF1. Following backbone generation, viable protein sequences can be readily obtained by solving the inverse folding problems, with models such as ProteinMPNN [21, 22] and ESM-IF1 [23]. Subsequent applications of structure prediction models such as AlphaFold2 or RoseTTAFold solves the forward folding problem fully determines the proteins.

Backbone generation is both the starting point and currently the bottleneck of the entire workflow. The groundbreaking work demonstrated through the development of RFdiffusion is aimed directly at tackling this bottleneck [11, 12]. With the structure prediction model RoseTTAFold taken as the basis, the generative model RFdiffusion was created through fine tuning the RoseTTAFold network on protein structure denoising tasks. RoseTTAFold has a number of desirable characteristics built-in that make it particularly well-suited as a basis for fine tuning into a diffusion model. These include the ability to carry out conditioning on protein design specifications. However, a nearly indispensable property of RoseTTAFold as a structure prediction network is its equivariance under the action of the Lie group SE(3), defined as the semi-direct product of the group of translations and rotations in three-dimensional Euclidean space. SE(3)-equivariance ensures that the same protein structure appearing in different orientations and positions can be treated as the exact same object, greatly improving data efficiency of model training [24, 25, 26, 27, 28]. Other SE(3)-equivariant structure prediction networks, such as AlphaFold2, OmegaFold and ESMFold can in principle be substituted for the basis of the diffusion model [12].

While RFdiffusion has enjoyed huge success as a viable solution for protein backbone generation, efforts have been carried out to construct diffusion models without the use of a pre-trained protein structure prediction as the basis. This new approach has been demonstrated through the model FrameDiff [25]. Here, a principled SE(3) diffusion model is shown to better formulate the protein backbone generation problem, achieving comparable performance with four-fold fewer network weights and without the need to leverage a protein structure prediction network, compared to RFdiffusion.

C. TCR Binding Evaluation

Beside TCR generation, binding evaluation is another crucial topic in TCR-T immunotherapy development. Traditionally, wet labs need to manufacture the TCR engineered T cells and measure their binding affinity with targeted cells, which is expensive in both time and cost. In recent years, using AI to predict the binding effects becomes a rapidly evolving field. Although this is not of the focus of this paper and is orthogonal, we give a brief review of the recent developments. AI based structure modelling and docking is an important direction for binding prediction. Pierce and Weng developed TCRFlexDock, a flexible docking approach for TCR/pMHC complex prediction, leading to near-native predictions for a significant proportion of models [29]. Myronov et al. developed an AI platform that combines molecular dynamics simulations with deep learning to predict binding probabilities for potential adoptive T-cell receptor cancer therapies [30]. Bradley explored specialized AlphaFold to generate models of TCR:peptide-MHC interactions for binding prediction. Yin et al. developed TCRModel2 for high-resolution modeling of T cell receptor recognition. Another important direction is to build AI models based on the sequence databases. Springer et al. developed ERGO, a highly specific and generic TCR-peptide binding predictor, demonstrating its accuracy in detecting TCR binding to peptides and peptide-TCR binding [33]. Montemurro et al. developed NetTCR-2.0 [34], which accurately predicts TCR-peptide binding using TCR sequence data, improving T-cell specificity prediction. Pham et al. developed

epiTCR [35], a highly sensitive predictor for TCR–peptide binding using over 3 million of TCR data.

III. Background

A. Antigen Presentation and Structure

The Major Histocompatibility Complex (MHC) pathway plays an important role in immune system by presenting the cell-specific (e.g. cancer cell or virus infected cell) peptide fragments – antigen – to the membrane, so that the immune system can identify and attack these cells. The pathway synthesizes and presents the MHC molecule with α and β chains enclosing the antigen peptide, as shown in Figure 1.

B. T Cell and T Cell Receptor (TCR)

In the immune system, T cells recognize the target cells (e.g. cancer cell or infected cell) by the T cell receptors, which binds to the antigen peptides presented by the MHC molecules on the surface of target cells.

As shown in Figure 1, each TCR consists of two different polypeptide chains, known as the α and β chains. Each chain contains both the constant (C) and variable (V) regions. The variable regions are the most crucial for the antigen-binding specificity of TCRs.

The variable regions consists of 3 pairs of complementarity-determining regions (CDR). CDR1 and CDR2 bind to the MHC molecules of the target cell and are more or less fixed by the limited varieties of human cell MHC. CDR3 binds to the antigen peptides presented by the MHC and is the highly variable region. Hence it is the focus of the TCR synthesis.

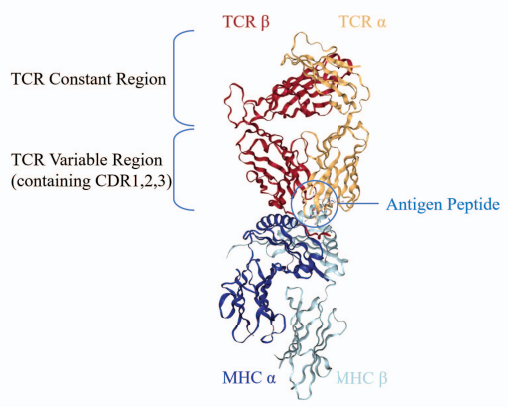

Fig. 1. Structure of TCR binding with MHC(Structure is downloaded from https://www.rcsb.org/3d-view/ngl/4may)

C. V(D)J Recombination

As illustrated in Figure 2, the variable region of TCR is produced by rearrangement of three gene segments, namely variable (V), diversity (D), and joining (J), known as V(D)J recombination. The α chain is composed of V and J segments, while the β chain is composed of V, D, and J segments. Human T cells contain 44 possible V segments, 27 D segments and 6 J segments. In DNA, V(D)J regions are coded separately. During V(D)J recombination, the cell selects from these segments and join them together to produce the TCR[36]. This recombination generates vast diversity of TCR. Additionally, junctions between these segments undergo random insertions and deletions of nucleotides (Figure 2b), further increasing diversity (junctional diversity). It is estimated that human immune system is able to produce about 10^{15} unique TCRs [37, 38], far greater than the number of T cells in an person.

As the result, the variable region of TCR is formed, including CDR1 and CDR2 encoded by V segment, and CDR3 spans the V(D)J segments, as shown in Figure 2c.

D. TCR Encoding with V, CDR3, J Segments

Based on the result of V(D)J recombination illustrated in Figure 2c, TCR sequences are commonly recorded as the V, CDR3 and J in computational methods. As V and J segments are selected from a pool of possible ones, they are recorded as gene identifiers, e.g. TRAJ43 and TRBV24-1 where TRAJ refers to "TCR α chain J segment" and TRBV refers to "TCR β chain V segment". CDR3 is the most variable region, therefore, it is recorded directly as the amio acid sequence, such as "CASSYLPGQGDHYSNQPQHF".

IV. Large Language Model for TCR Generation

Antigen and TCR are encoded by gene and protein sequences, which can be considered as a language. The recent rise of large language model (LLM), especially the foundation LLM, inspired us to adapt large language model to generate de novo TCR specific to the given antigen. We propose Llama-TCR, a foundation TCR language model for antigen-specific TCR generation. As shown in Figure 3, the language based TCR generation model will receive the antigen sequence as the input and produce the TCR sequence as V, CDR3, J encoding.

A. VDJdb Instruction

VDJdb [39] is a curated database of TCR sequences with known antigen specificities. Based on this database, We filtered the incomplete records and obtained 28452 antigen-TCR pairs. We then built the instruction dataset from these pairs. The input is the antigen consisting of the MHC α chain, Epitope sequence, and MHC β chain, separated by "∼", e.g. "HLA-A*02:01∼KVAELVHFL∼B2M". The output is the TCR α and β, each consisting of the V, CDR3 and J segments, also separated by "∼", e.g. "TRAV8-3*01∼CAVVPMYGGSQGNLIF∼TRAJ42*01∼TRBV5-6*01∼CASSPLRGGVTYNEQFF∼TRBJ2-1*01".

B. Transformer based Large Language Model

Transformers [40] have revolutionized LLMs by relying on self-attention, a mechanism that captures long-range dependencies in text data. It has become the foundation of the large language models. We first briefly introduce the key concepts of Transformer with the context of TCR generation.

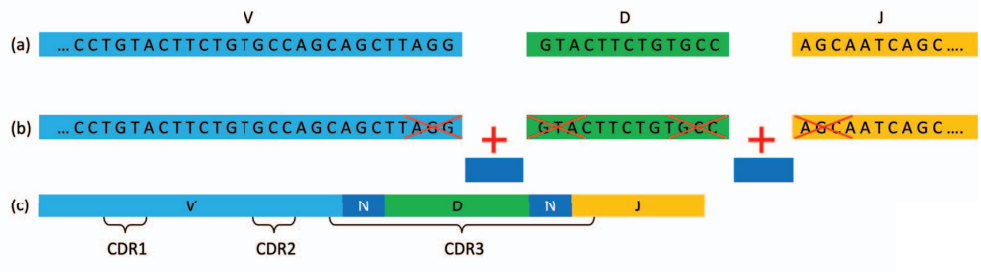

Fig. 2. V(D)J Recombination

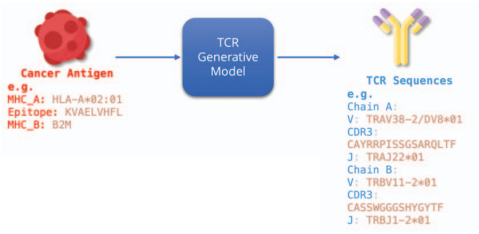

Fig. 3. Framework of TCR Generation

*1) **Input Embedding**:* The each word of the input sequence (e.g. text and protein sequences) to transformers is encoded as a vector of a predefined length, aka. embedding dimension. As a result, the whole sequence is encoded as a matrix of sequence length $\times$ embedding dimension. For example, using embedding dimension 100, an $alpha$ chain CDR3 sequence "CIVRAPGRADMRF" of 13 amino acid can be encoded as a matrix of 13×100.

*2) **Self-Attention Mechanism**:* The core of transformer is the self-attention mechanism. The input matrix is multiplied with three trainable weights, namely W_Q, W_K, W_V to produce the three vectors Q (Query), K (Key) and V (Value). The attention score is then computed by the following equation:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k}) * V$$

where

- Q (Query): A vector representing the current word or token being analyzed.
- K (Key): A vector representing all other words or tokens in the sequence.
- V (Value): A vector containing the information associated with each word or token.
- d_k: Dimensionality of the query and key vectors.

Softmax is applied to normalize the attention score across all words, ensuring it sums to 1.

This equation calculates the attention score for each word in the sequence relative to the current word. The score indicates how relevant each word is to the current one, capturing long-range dependencies crucial for tasks like machine translation and text generation.

	#Parameters	MAE ↓	BLEU↑	ROUGE-1↑
ProtGPT2	738M	0.266	0.785	0.415
Molinst-Protein	7B	10.36	0	0
Llama-TCR	7B	**0.264**	**0.858**	**0.912**

*3) **Encoder-Decoder vs. Decoder Only**:* When first proposed [40], transformers utilize an encoder-decoder architecture where the encoder processes the input sequence into context vector and the decoder generates the output sequence based on the context vector. It excels in sequence-to-sequence tasks like translation and summarization.

Decoder-only models, on the other hand, break free from the rigid encoder-decoder paradigm. They directly process the input prompts and generate the output from them. This allows for greater flexibility and creativity, particularly in open-ended tasks like dialogue and creative writing. Also, the decoder-only models do not reply on data labelling, which enables them to make use of the extremely large scale of text data in literature and internet. With such advantages, decoder-only models is leading the storm of generative AI with famous models like GPT [41], Llama [42] etc. As we aim to design de novo TCR with large language model, creativity is important. Therefore, we decide to adopt decoder-only transformers for TCR design.

C. Finetune LLM for De Novo TCR Design

In this study, we adopted and finetuned the Llama-2[42] model into Llama-TCR to generated diverse and functional T cell receptors (TCR) with instructions derived from VDJdb. By feeding the model with instructions based on real-world functional TCR and antigens, we enable the model to generate TCR with antigen specificity.

V. Results and Discussion

A. Sequence Based Performance

We adopated and finetuned the Llama-2-7b-hf model for TCR generation using the VDJdb Instruction database and kepted 10% of dataset for testing. We also finetuned two protein specific language model – ProtGPT2 [17] and Molinst-Protein [43] for comparison. We compare the performance of models using three common metrics in natural language processing - MAE, BLEU and ROUGE.

From the comparison, we see that the performance of Llama-TCR performs the best with marginal advantage over

ProtGPT2. VDJdb Instruction is a small dataset compared with common NLP datasets and therefore the advantage of larger model is not obvious. In contrast, Molinst-Protein gives very low performance in TCR generation.

*1) **Specialization vs. Generalization**:* We further investigated the performance of Molinst-Protein. Molinst-Protein is an fully instruction finetuned Llama model using Mol-Instruction protein dataset with tasks like protein design, protein function prediction etc. Further investigation shows that the model is over specialized by the initial finetune and its generalizability is affected

*2) **Precision v.s. Creativity**:* The language metrics shows that the LLM is able to generate TCR sequences close to the training data. However, such metrics is not able to reflect its ability to create de novo TCR design. At the state of the art, functional evaluation of de novo TCR sequences largely relies on wet lab experiment. We would like to conduct further studies on AI based evaluation in future works.

B. 3D Structure Visualization

As TCR functions largely depend on its folded structure, it is difficult to evaluate the generated design solely by the sequence. Therefore, we generate the de novo TCR targeting to MAGE-A3, a Melanoma-associated antigen and fold it to its 3D structure for visualization.

The 3D folding is conducted by the pipeline of Stitchr [44] and Alphafold 2 [6]. Llama-TCR outputs the TCR α and β chain in V, CDR3, J encoding. As AlphaFold 2 makes single chain structure prediction on amino acid sequences, We need to process the two chains separately. For each chain, we first use Stitchr [44] to convert the V, CDR3, J segments into full amino acid sequence, and then use AlphaFold 2 to predict the structure. Figure 4 shows the AlphaFold predicted 3D structure of the TCR generated by Llama-TCR. Comparing with Figure 1, we can see the generated TCR contains the key structure of a functional TCR.

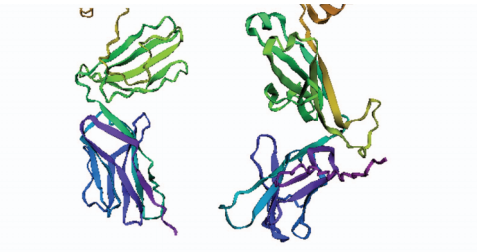

Fig. 4. 3D structure of generated TCR targeting at MAGE-A3 antigen

VI. Conclusion

In conclusion, this study developed Llama-TCR, a generative AI model for diverse and functional TCR candidate, leveraging the advanced capabilities of large language models. By both sequence based benchmark and 3D structure visualization, Llama-TCR has successfully demonstrated the potential in generating de novo TCR sequences for immunotherapy, particularly in the realm of cancer treatment and vaccine development. At the same time, it is crucial to acknowledge the need for further validation and refinement of the model through experimental studies, which is a focus area of our future works.

References

[1] Z. Pang, M.-m. Lu, Y. Zhang, Y. Gao, J.-j. Bai, J.-y. Gu, L. Xie, and W.-z. Wu, "Neoantigen-targeted tcr-engineered t cell immunotherapy: current advances and challenges," *Biomarker Research*, vol. 11, no. 1, 2023.

[2] K. L. M. . H. V. Shafer, P., "Cancer therapy with tcr-engineered t cells: Current strategies, challenges, and prospects," *Frontiers in Immunology*, 2022.

[3] . S. R. M. Sterner, R. C., "Car-t cell therapy: current limitations and potential strategies," *Blood cancer journal*, vol. 11, no. 4, 2021.

[4] G. C. T. B. P. S. R. . A. K. Burns, E. A., "Uniprot: the universal protein knowledgebase in 2023," *Nucleic Acids Research*, vol. 9, no. 1, 2021.

[5] . M. M. Lee, M. N., "Antigen identification for hla class i–and hla class ii–restricted t cell receptors using cytokine-capturing antigen-presenting cells," *Science immunology*, vol. 6, no. 55, 2021.

[6] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, 2021.

[7] A. Al-Janabi, "Has deepmind's alphafold solved the protein folding problem?" 2022.

[8] A. H.-W. Yeh, C. Norn, Y. Kipnis, D. Tischer, S. J. Pellock, D. Evans, P. Ma, G. R. Lee, J. Z. Zhang, I. Anishchenko *et al.*, "De novo design of luciferases using deep learning," *Nature*, vol. 614, no. 7949, pp. 774–780, 2023.

[9] L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. Verschueren *et al.*, "Design of protein-binding proteins from the target structure alone," *Nature*, vol. 605, no. 7910, pp. 551–560, 2022.

[10] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera *et al.*, "De novo protein design by deep network hallucination," *Nature*, vol. 600, no. 7889, pp. 547–552, 2021.

[11] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles *et al.*, "De novo design of protein structure and function with rfdiffusion," *Nature*, vol. 620, no. 7976, 2023.

[12] ——, "Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models," *BioRxiv*, 2022.

[13] R. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "Msa transformer," *bioRxiv*, 2021.

[14] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *bioRxiv*, 2021.

[15] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, 2023.

[16] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "Proteinbert: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, 2021.

[17] N. Ferruz, S. Schmidt, and B. Höcker, "Protgpt2 is a deep unsupervised language model for protein design," *Nature Communications*, vol. 13, 2022.

[18] A. Madani, B. McCann, N. Naik, N. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher, "Progen: Language modeling for protein generation," *bioRxiv*, 2020.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, 2020.

[20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015.

[21] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, "Generative models for graph-based protein design," *Advances in neural information processing systems*, vol. 32, 2019.

[22] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel *et al.*, "Robust deep learning–based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, 2022.

[23] H. Khakzad, I. Igashov, A. Schneuing, C. Goverde, M. Bronstein, and B. Correia, "A new age in protein design empowered by deep learning," *Cell Systems*, vol. 14, no. 11, 2023.

[24] Z. Guo, J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu, and J. Cheng, "Diffusion models in bioinformatics and computational biology," *Nature Reviews Bioengineering*, 2023.

[25] J. Yim, B. L. Trippe, V. De Bortoli, E. Mathieu, A. Doucet, R. Barzilay, and T. Jaakkola, "SE(3) diffusion model with application to protein backbone generation," *arXiv preprint arXiv:2302.02277*, 2023.

[26] N. Anand and T. Achim, "Protein structure and sequence generation with equivariant denoising diffusion probabilistic models," *arXiv preprint arXiv:2205.15019*, 2022.

[27] B. Elesedy and S. Zaidi, "Provably strict generalisation benefit for equivariant models," in *International Conference on Machine Learning*. PMLR, 2021.

[28] T. Cohen *et al.*, "Equivariant convolutional networks," Ph.D. dissertation, Taco Cohen, 2021.

[29] B. Pierce and Z. Weng, "A flexible docking approach for prediction of t cell receptor–peptide–mhc complexes," *Protein Science*, 2013.

[30] A. Myronov *et al.*, "Modeling phla:tcr interactions for effective tcr therapies: Leveraging ai and molecular dynamics," *Cancer Research*, 2022.

[31] P. Bradley, "Structure-based prediction of t cell receptor: peptide-mhc interactions," *Elife*, vol. 12, 2023.

[32] R. Yin, H. V. Ribeiro-Filho, V. Lin, R. Gowthaman, M. Cheung, and B. G. Pierce, "Tcrmodel2: high-resolution modeling of t cell receptor recognition using deep learning," *Nucleic Acids Research*, 2023.

[33] I. Springer *et al.*, "Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs," *Frontiers in Immunology*, 2020.

[34] A. Montemurro *et al.*, "Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcr and sequence data," *Communications Biology*, 2021.

[35] M.-D. N. Pham, T.-N. Nguyen, L. S. Tran, Q.-T. B. Nguyen, T.-P. H. Nguyen, T. M. Q. Pham, H.-N. Nguyen, H. Giang, M.-D. Phan, and V. Nguyen, "epitcr: a highly sensitive predictor for tcr–peptide binding," *Bioinformatics*, vol. 39, no. 5, 2023.

[36] F. W. Alt, E. M. Oltz, F. Young, J. Gorman, G. Taccioli, and J. Chen, "Vdj recombination," *Immunology today*, vol. 13, no. 8, 1992.

[37] M. M. Davis and P. J. Bjorkman, "T-cell antigen receptor genes and t-cell recognition," *Nature*, vol. 334, no. 6181, 1988.

[38] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan Jr, "Statistical inference of the generation probability of t-cell receptors from sequence repertoires," *Proceedings of the National Academy of Sciences*, vol. 109, no. 40, 2012.

[39] M. Goncharov, D. Bagaev, D. Shcherbinin, I. Zvyagin, D. Bolotin, P. G. Thomas, A. A. Minervina, M. V. Pogorelyy, K. Ladell, J. E. McLaren *et al.*, "Vdjdb in the pandemic era: a compendium of t cell receptors specific for sars-cov-2," *Nature Methods*, vol. 19, no. 9, 2022.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] OpenAI, "Gpt-4 technical report," 2023.

[42] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[43] Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen, "Mol-instructions: A large-scale biomolecular instruction dataset for large language models," *arXiv preprint arXiv:2306.08018*, 2023.

[44] J. M. Heather, M. J. Spindler, M. Alonso, Y. Shui, D. G. Millar, D. Johnson, M. Cobbold, and A. Hata, "Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information," *Nucleic Acids Research*, 03 2022.