Weekly Report (May 5, 2025 – May 11, 2025)

1. Summary of Work

1. Reviewed and organized the overall project proposal “Integrating ESM3 Structural Representation with Self-Play Reinforcement Learning” and its planned adaptation of the EvoPlay method based on the token space of ESM3.
2. Dive deep into Wang et al. (2023) “Self-play reinforcement learning guides protein engineering” (EvoPlay), tried to understand its algorithmic framework, key formulas, and experimental workflow.
3. Created and finalized a *Journal Club* PPT for next week’s group meeting, covering EvoPlay’s background, MCTS simulation details, policy-value network training, and its applications to GFP and PAB1 protein design. You can find the PPT at here: EvoPlay_Journal_Club.pptx
4. However, the paper focuses more on the experimental results and the application of EvoPlay to GFP and PAB1 protein design, rather than the algorithmic details. We have tried to understand the algorithmic details of EvoPlay, but for a full understanding of the MCTS algorithm, we might need to read the original paper of AlphaZero. We plan to have a quick overview the AlphaZero paper next week.

---

2. Detailed Work

2.1 EvoPlay Paper Deep-Dive

- **Algorithm Overview**
  - Transforms AlphaZero’s self-play into a single-agent optimization loop (EvoPlay).
  - Uses MCTS to simulate ~400 trajectories per update; the policy-value network outputs mutation probabilities and scores.
- **MCTS Simulation Details**
  - Node selection:

$$U_{s,a} \leftarrow c \cdot P_{s,a} \frac{\sqrt{N_{\text{parent}}}}{1 + N_{s,a}}, \quad Q_{s,a} \leftarrow \frac{Q_{s,a}\, N_{s,a} + v}{N_{s,a} + 1}$$

    where c is a constant controlling the balance between exploration and exploitation. $P_{s,a}$ is the prior probability of action a in state s, and $N_{s,a}$ is the visit count. v is the value of the state which is the output of the policy-value network.
  - The terminal state is reached when:
    - The mutation sequence is repeated.
    - The score drops below a threshold.
  - The loss function:

$$\mathcal{L} = (r - v)^2 - \boldsymbol{\pi}^T \log P + \gamma \|\theta\|^2$$

where r is the reward which is given by the surrogate model, $\boldsymbol{\pi}$ is the selected action probability, and θ is the network parameters.

- **Real Play**
  - Sample mutation sequences based on visit counts.
- **Experimental Setup**
  - Test on GFP (20 $\times$ 237 action space) and PAB1 (20×44).
  - Surrogate CNN predicts fluorescence/enrichment scores.
  - Initiate from 5 starting sequences per protein.

2.3 PPT Preparation

- **Structure**
  1. Background & motivation
  2. EvoPlay algorithm workflow
  3. MCTS simulation & real play
  4. Loss function & network training
  5. Experimental design & results
  6. Summary & open questions
- **Figures & Charts**
  - Recreated key flow diagrams, formulas, and GFP/PAB1 performance bar plots.

---

3. Challenges & Reflections

1. **Large Action Space**
   - EvoPlay's 20 $\times$ L sequence space is already vast; moving to ~4,096 structural tokens risks search inefficiency. Consider hierarchical or constrained search strategies.
2. **Reward Function Design**
   - Current surrogate scores are simplistic. When we turns to the structural token space and backbone generation, we need to design a more sophisticated reward function that captures the nuances of protein folding and stability. Directly utilizing ESM3 for the transfer from the token space to the structural token space may not be feasible as it might takes too long times during simulation.
3. **Compute Resources**
   - MCTS simulations, network training, and ESM3 for truning tokens back to protein backbone demand high GPU/TPU resources. Might need to coordinate for resource allocation.

---

4. Plan for Next Week

1. **Design Structural Token Action Space**
   - Define token mutation rules and impose constraints on mutation count/locations.
2. **Prototype MCTS in Sequence Space**

- Reproduce EvoPlay MCTS workflow on sequence data to validate code framework and network interfaces.

3. **Draft Reward Function Prototype**
    - Draft a composite scoring function using packing density or other backbone generation metrics, like *designablity*, *diversity*, and *novelty*.

4. **Team & Resource Coordination**
    - Dsicuss in team for hierarchical MCTS feasibility.