BIOCHEMISTRY
LITERACY
for kids

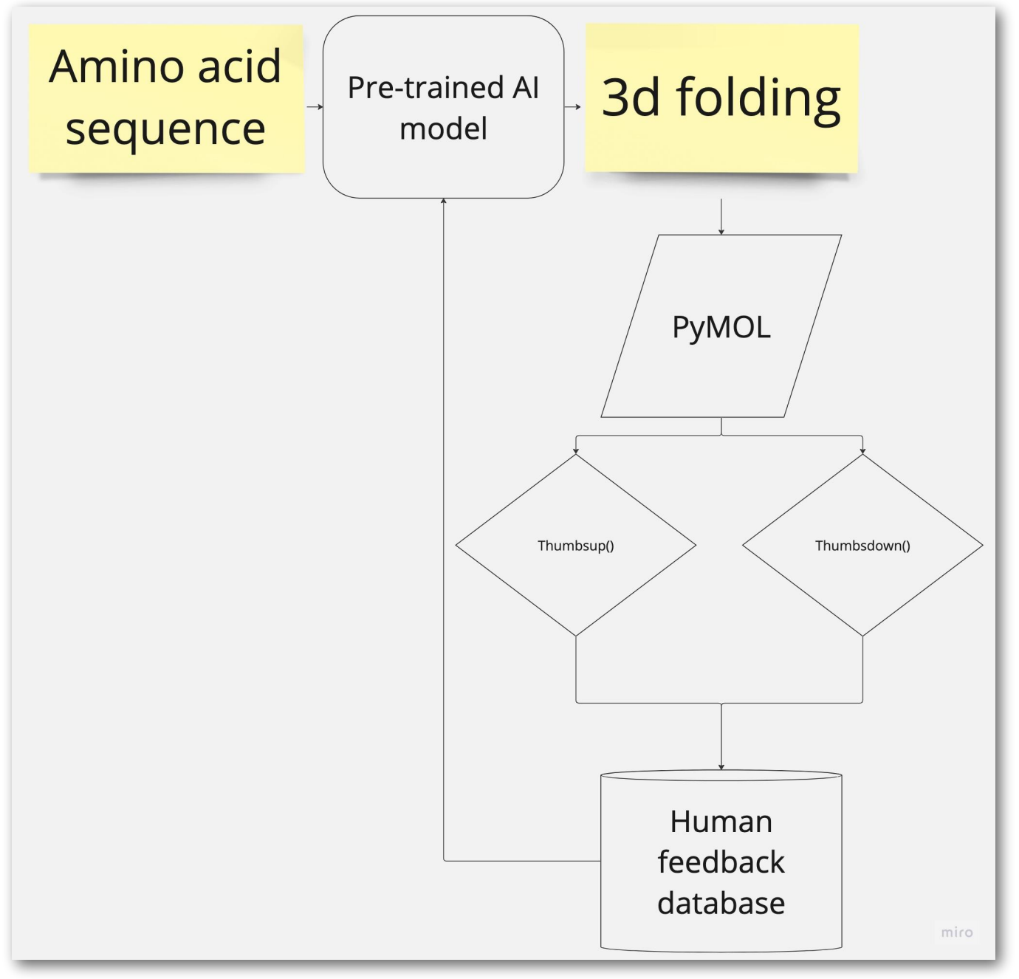
IMPROVING AI PROTEIN FOLDING PREDICTIONS WITH HUMAN FEEDBACK: INTRODUCING FOLDX

HANSON JIANG & ACE JIANG

BACKGROUND

- **Protein folding and its importance:** Protein folding is a vital biological process where a protein acquires its functional, three-dimensional structure from a linear amino acid chain. Misfolding can lead to various diseases, making accurate prediction crucial in health and medicine.
- **AI in protein folding prediction:** The ESM Metagenomic Atlas is a pioneer AI model in protein folding prediction, harnessing the power of machine learning to predict protein structures from their primary sequences.
- **The need for improved accuracy:** While effective, the ESM Atlas's predictions could benefit from enhanced accuracy, especially for complex and novel proteins.

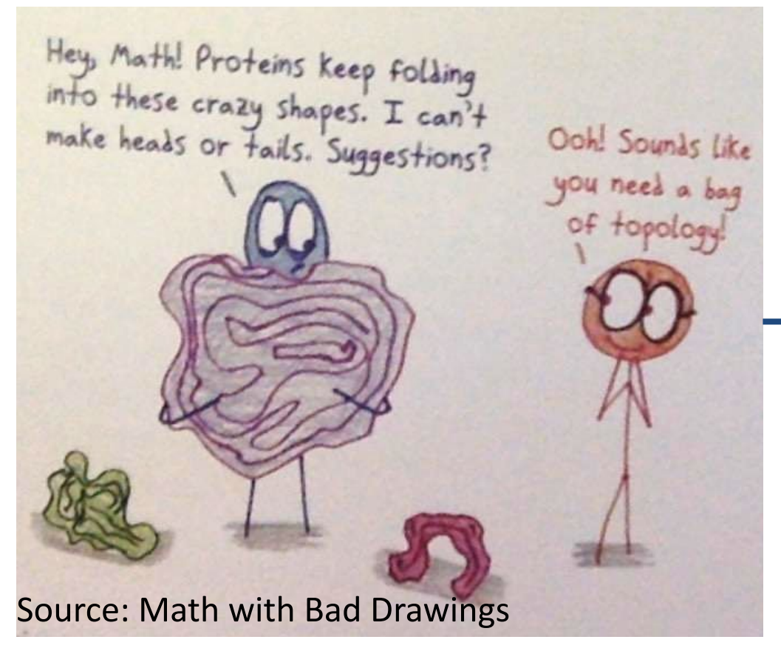
ARCHITECTURE



1. **Input:** The user inputs an amino acid sequence in PyMOL.
2. **Generation:** ESM-2 model processes the sequence to predict its 3D structure.
3. **Visualization:** PyMOL visualizes the predicted structure for examination and modification by the user.
4. **Feedback Collection:** User feedback and modifications of the structure are gathered.
5. **Analysis:** Feedback data is analyzed to identify areas of improvement in model predictions.
6. **Re-training:** The model is trained again using user modifications as new targets, enhancing its accuracy.
7. **Continuous Improvement:** The process iterates, steadily refining the model through user-driven, real-world feedback.

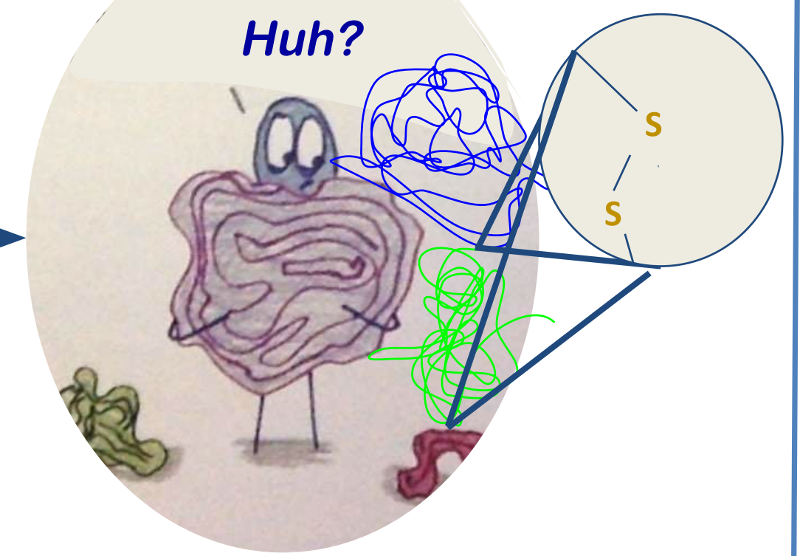
THE FOLDX SOLUTION

- **Utilization of the Advanced ESM Model**: FoldX takes advantage of the ESM model, an AI model pre-trained on protein structure data, harnessing its ability to predict protein folding structures from primary amino acid sequences.
- **Facilitating Human Interaction**: One of FoldX's significant innovations is its user-friendly interface integrated with the PyMOL molecular visualization system. This allows users to easily examine and give feedback on the AI-predicted protein structures, ensuring that the system is not only approachable but interactive.
- **Collection and Storage of Human Feedback:** FoldX employs a sophisticated system for gathering and storing human feedback, capitalizing on the knowledge and insights of users to enhance the model's learning.
- **Reinforced Learning via Human Feedback:** Similar to the approach used in InstructGPT, FoldX uses the collected human feedback to refine the ESM model's predictions, driving improvements in its accuracy and performance. This continuous learning process leverages human intelligence to teach the AI model, leading to better and more reliable predictions over time.



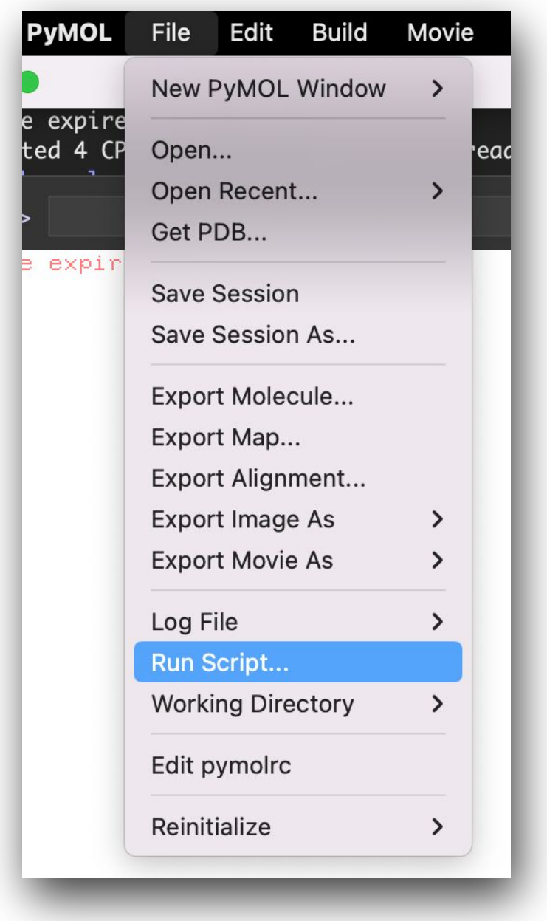
Source: Math with Bad Drawings

And then with FoldX...



Partial source: Math with Bad Drawings

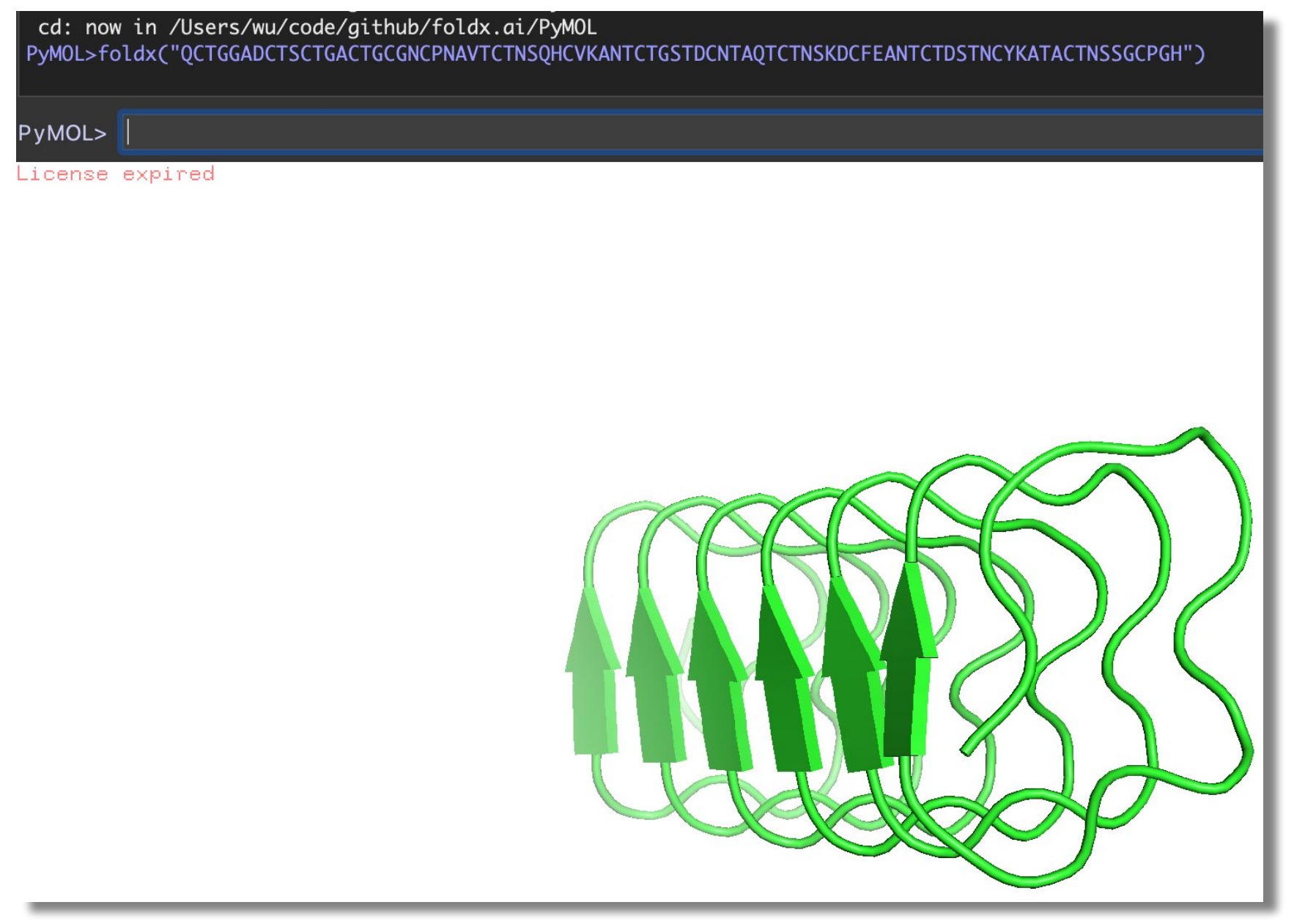
HUMAN FEEDBACK IN FOLDX



Step I: Run the FoldX script in PyMOL

```
PyMOL> foldx("QCTGGADCTSCTGACTGCGNCPNAVTCTNSQHCVKANTCTGSTDCNTAQTCTNSKDCFEANTCTDSTNCYKATACTNSSGCPGH")
```

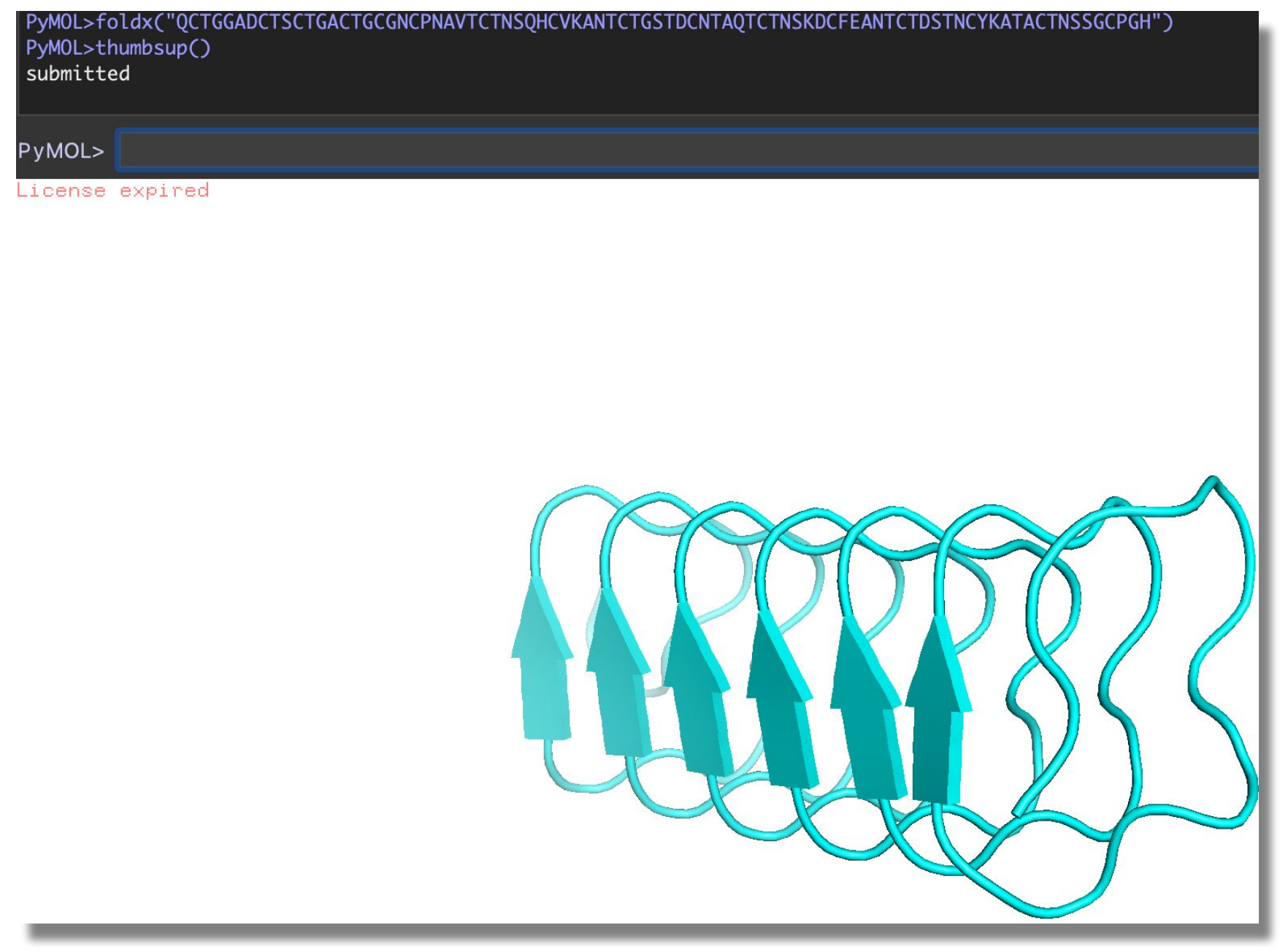
Step II: Call the FoldX Function to generate the 3D folding structure



```
 cd: now in /Users/wu/code/github/foldx.ai/PyMOL
PyMOL>foldx("QCTGGADCTSCTGACTGCGNCPNAVTCTNSQHCVKANTCTGSTDCNTAQTCTNSKDCFEANTCTDSTNCYKATACTNSSGCPGH")

PyMOL>
License expired
```

Step III: Examine and edit the structure



```
PyMOL>foldx("QCTGGADCTSCTGACTGCGNCPNAVTCTNSQHCVKANTCTGSTDCNTAQTCTNSKDCFEANTCTDSTNCYKATACTNSSGCPGH")
PyMOL>thumbsup()
submitted

PyMOL>
License expired
```

Step IV: Provide human input and feedback (thumbs-up or thumbs-down)

References

Lin, Z., et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. Cold Spring Harbor Laboratory.
Long, O., et al. (2022). Training language models to follow instructions with human feedback. arXiv.

FUTURE WORK

Continuous Reinforced Learning: A key aspect of our future work lies in the continuous reinforcement learning via human feedback. FoldX will continually collect human insights via PyMOL, enhancing the accuracy and performance of the model. As the feedback pool grows, so does our model's ability to deliver accurate protein folding predictions. It's a cycle of improvement that grows more powerful over time.

Expansive Potential Applications: The possibilities unlocked by FoldX's advanced prediction capabilities are vast. Engineered bacteria, optimized with our accurate protein structure predictions, could lead to cleaner and more efficient energy production methods. Furthermore, FoldX's potential in the realm of drug discovery is monumental. By accurately predicting protein-drug interactions, we could significantly speed up the drug discovery process and possibly contribute to the development of treatments for various diseases.

Scan QR code to learn more about FoldX

BIOCHEMISTRY AND PHYSICS LITERACY SUMMER INSTITUTE 2023: POTTERSVILLE, NJ