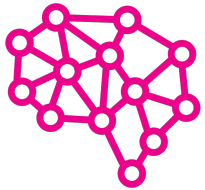


FourthBrain

Learning The Language of Proteins

Chrisogonas Odhiambo, Gilles Bouyer, Patrick Gebhard - April 2023

Outline

- Problem Definition
- Solution Approach
- Responsible AI Considerations
- Data + Model
- MLE Stack
- AWS Demo
- Conclusions
- Future Work

Developing Drug is hard

Amino acids can be represented as a character sequence: GMASKAGSVLGK ITKIALGAL

The 3D-shape of the protein determines its potential e.g. to treat cancer (ACP), to be antimicrobial (AMP) or to be DNA-Binding.

There are lots of potential amino acid sequences to choose from, but we want to know which ones are likely to have relevant properties.

Problem Definition

- Cancer drug development is costly in time and intensive in labor demands.
- It is improved by incorporating proteins bio properties to the drug
- The following bio properties also called tasks, are improving cancer drugs: ACP - targeting cancer cells, DNA replication, AMP - inhibitor of viruses

Solution Approach - Predict Protein Characteristics

1. Do not engineer proteins; predict protein characteristics
2. Represent protein as a sequence of tokens, each representing an amino acid e.g. GWKSVFRKAKKVGKTVGGLALDHYLG
3. Reduce protein sequences to a language comprised of amino acid alphabets
4. By (3) above, we therefore solve protein classification problem using linguistic machine learning techniques - NLPs, LSTM, XGBoost

Data + Model

- ACP: 344 (sequence, label) tuples in test_data.csv, 1,378 in train_data.csv
- AMP: 4042 (PDBs_Code, sequenceID, label) all_data tuples
- DNA binding: (code, sequence,label, origin) 2,272 test.csv 14,189 train.csv
- One letter of the alphabet for each of the 20 amino acids
- The letters B, J, O, U, X, Z are not used
- Chain Lengths: ACP [2,50] AMP [11 ,183] DNA binding [47 ,5184]

.

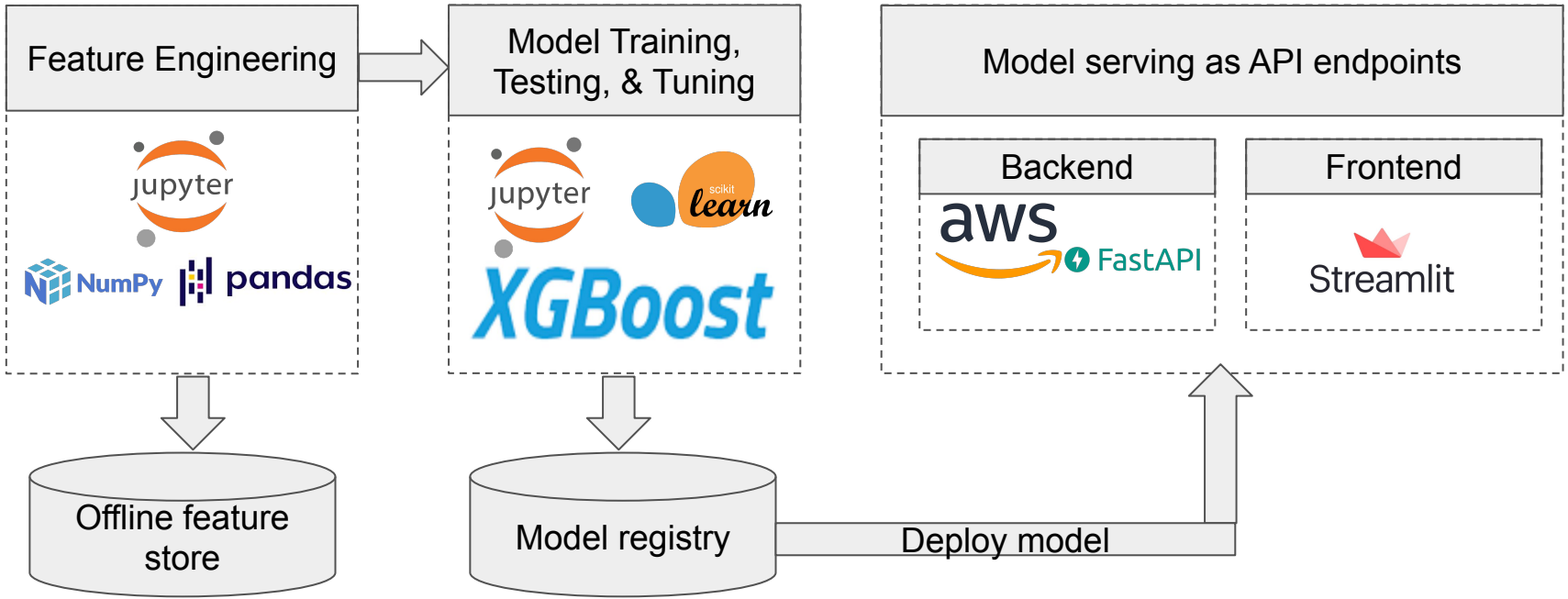
Potential Benefits

- Shortened life-cycle of drug development incorporating these protein to cure illnesses
- Efficient drug development process

Responsible AI Considerations

1. **Bias and Fairness:**
   - Use of available, diverse and representative datasets to avoid algorithmic bias.
   - Regular evaluation of the model for potential biases in its predictions.
2. **Transparency and Explainability:**
   - Develop clear and understandable documentation of the AI model's functionality and limitations.
   - Provide explanations for model predictions to facilitate trust and acceptance among medical professionals.
   - Ensure that the AI system is subject to third-party audits for transparency.
3. **Human-AI Collaboration:**
   - Design the protein analysis tool to augment, not replace, human expertise.
   - Encourage ongoing education on AI advancements for medical professionals.
4. **Accountability and Responsibility:**
   - Establish clear guidelines for the responsibility of AI developers, users, and regulators on new proteins
   - Implement a robust system for monitoring, reporting, and addressing AI-related ethical concerns.
5. **Environmental Sustainability:**
   - Optimize the AI model's energy consumption to reduce its environmental footprint.
   - Prioritize the use of sustainable energy sources and eco-friendly infrastructure.

MLE/MLOps Pipeline Infrastructure



Deployment

InstaDeep™ DeepChain by InstaDeep

Learning The Language of Proteins

Copy/Paste your amino acid sequence here:
'ATFCHCRRSCYSTEYSYGTCTVMGINWRFCCL(ACP)','FLSLIPHAINAVSTLVHHF(AMP)',
'GDVSVVGFDDSPLIAFTSPPLSTVRQPVQAMATAAVGALLEEIEGNPVQRTEFVFQPELVVRGSTAQPPGRVSQVLS(DNA)'

FLSLIPHAINAVSTLVHHF

The sequence you entered is: FLSLIPHAINAVSTLVHHF

Not an Anticancer Peptide

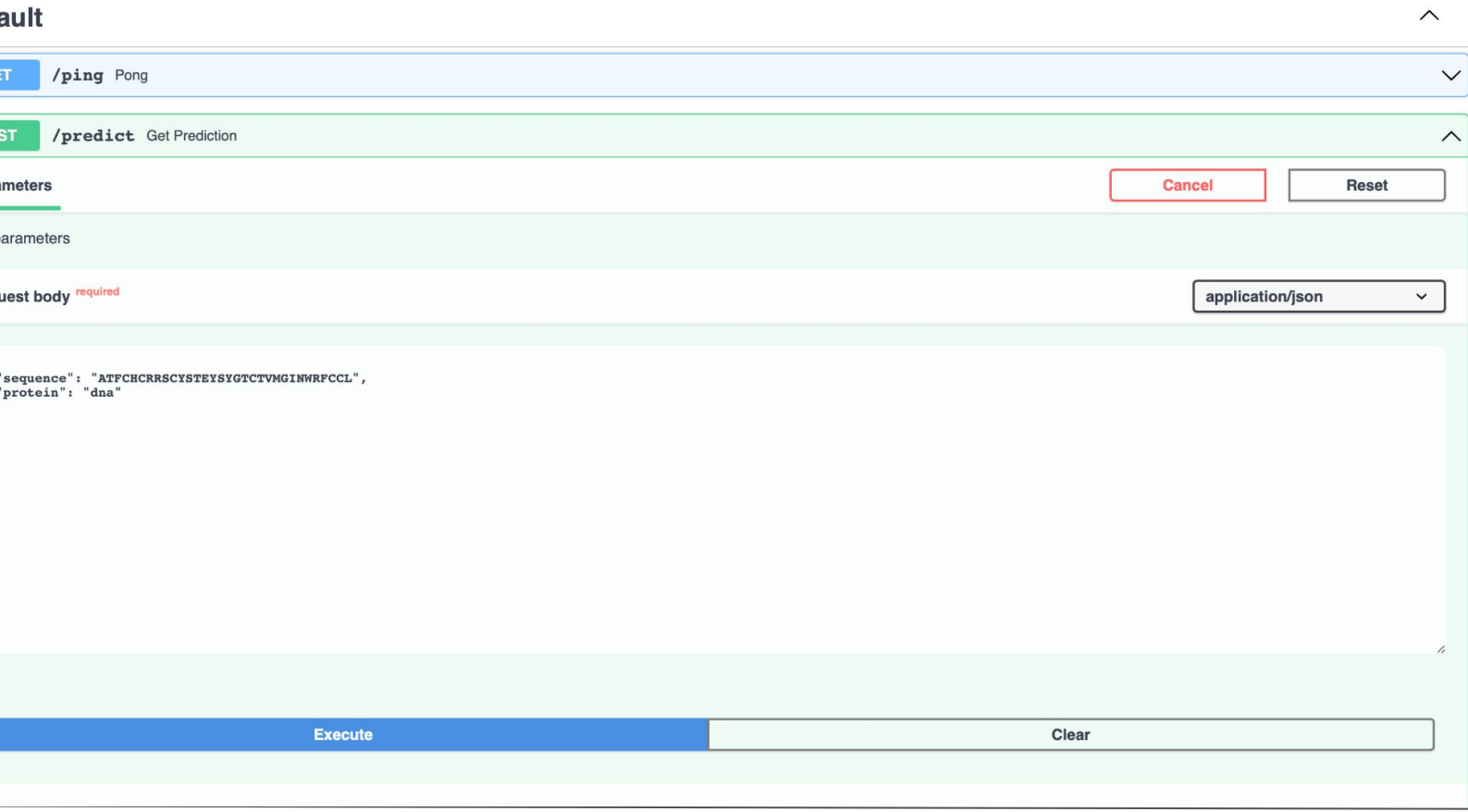
It is an Antimicrobial Peptide

Not a DNA Binding Protein

AWS Deployment in backup slides section

https://gillesbouyer-lop-language-of-proteins-streamlitv7-m20ksg.streamlit.app

AWS Demo



Conclusions

- We created an end-to-end prototype to protein classification model.
- Modern ML algorithms hold much promise to drug development and cancer research.

- This is a sensitive research area with implications on ethics. It is crucial to be aware of unintended consequences such as the creation of dangerous proteins, copyrighting of nature, deviation off the moral compass, etc.

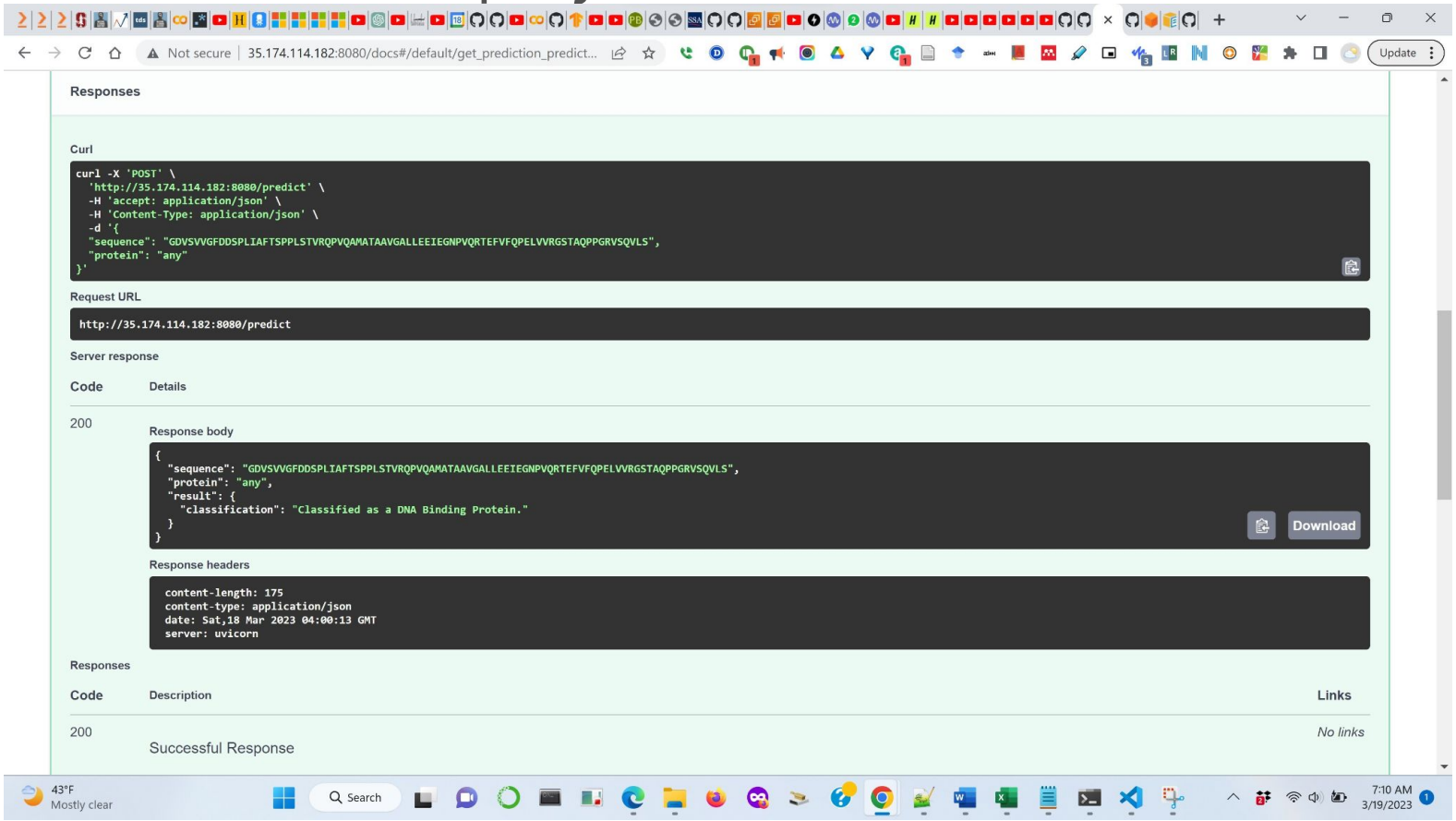
Future Work

- Use of transformers to classify proteins, trained on even larger datasets
- Training of a unified model instead of separate models for different protein classes
- Improving the model performance through robust drift monitoring and continuous experimentation
- Suggest chains of proteins

Thank You! Questions?

Backup Slides

AWS Deployment API Tests - 2



AWS Deployment API Tests - 3