token\_regex = "(\[[^\]]+]|Br?|Cl?|N|O|S|P|F|I|b|c|n|o|s|p|\(|\)|

\.|=|#|-|\+|\\\\|\/|:|~|@|\?|>|\\*|\$|\%[0-9]{2}|[0-9])

**Team Name:** Chemistry AI

**Team members**: Naiji Jabin Gong, Siddharth Pillai, Terry Ma

**Project title:** Organic Chemical Reaction Prediction

## Project summary:

The prediction of chemical reaction products is an important part of the chemical synthesis of new compounds. Trained chemists have an in-depth understanding of the mechanisms by which compounds react to form new compounds, and use this knowledge to develop complex procedures by which to synthesize new compounds. Conventionally, new compounds must be synthesized and identified experimentally, which is resource-intensive and time-consuming. This project aims to explore deep learning approaches to chemical reaction prediction which could expedite the discovery of new compounds.

## Approach:

Chemical compounds and reactions can be represented as text strings using the Simplified Molecular Input Line Entry System (SMILES). Because of the large availability of reactions in text format (over 900,000 single-product reactions collected by the U.S. Patent Office), many studies have opted to frame the problem as a translation problem, where deep learning models are used to map reactant strings to product strings [1], [2]. We expect to reproduce the results from the LSTM approach in [2], and we would like to explore other approaches to this problem, such as using a Transformer model approach. Another direction we would like to explore is solving the retrosynthesis problem (predicting reactants from products, [3]) and using explainable AI to help understand how deep models are predicting reaction products through visualization.

## Resources:

[1] D. Fooshee *et al.*, “Deep learning for chemical reaction prediction,” *Mol. Syst. Des. Eng.*, vol. 3, no. 3, pp. 442–452, Jun. 2018, doi: 10.1039/C7ME00107J.

[2] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, and T. Laino, “‘Found in Translation’: Predicting Outcomes of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models.” arXiv, Nov. 15, 2017. doi: 10.48550/arXiv.1711.04810.

[3] L. Wang, C. Zhang, R. Bai, J. Li, and H. Duan, “Forward Reaction Prediction as Reverse Verification: A Novel Approach to Retrosynthesis.” ChemRxiv, Oct. 27, 2020. doi: 10.26434/chemrxiv.13138775.v1.

[4] L. Howes, “A toxic twist on AI for drug design,” *Chemical & Engineering News*. https://cen.acs.org/policy/chemical-weapons/toxic-twist-AI-drug-design/100/i17 (accessed Mar. 19, 2023).

[5] H. Wang *et al.*, “Chemical-Reaction-Aware Molecule Representation Learning,” presented at the International Conference on Learning Representations, Jan. 2022. Accessed: Mar. 19, 2023. [Online]. Available: https://openreview.net/forum?id=6sh3pIzKS-

[6] J. Jiménez-Luna, F. Grisoni, and G. Schneider, “Drug discovery with explainable artificial intelligence,” *Nat. Mach. Intell.*, vol. 2, no. 10, Art. no. 10, Oct. 2020, doi: 10.1038/s42256-020-00236-4.

## Related Work and Current SOTA method:

* In [2] the authors explore the application of neural sequence-to-sequence models to predict the outcomes of complex organic chemistry reactions. The authors employ an encoder-decoder model with Luongs’ attention mechanism, which has been previously successful in NLP applications.
* They use this model to translate reactants into products in a similar manner to how sentences are translated between different languages.
* The authors demonstrate that their deep learning approach can accurately (achieving 80.1% top 1 accuracy on SMILES test set) predict reaction outcomes and outperform traditional methods that are based on a set of rules and other machine learning techniques.
* The authors also show that the model can generalize well for previously unseen chemical reactions which could pave the way for predicting novel chemical reactions.

## Dataset:

* <https://www.kaggle.com/datasets/mmelahi/organic-chemistry-reactions?resource=download>
* Chemical reactions and product in SMILES (Simplified Molecular Input Line Entry System) format
* Over 900,000 single-product reactions

## Ethical considerations:

* Safety: Incorrect / inaccurate predictions from the model could result in unstable reactions, potentially resulting in safety hazards.
* Human impact: Models such as this could result in replacement of human labor to cut costs, which could put people trained in this area at a disadvantage.
* Environmental impact: Inaccurate/ inefficient results from the model could result in wastage of chemicals or inefficient disposal of chemicals.
* Misuse: use of AI to design harmful substances (dangerous drugs, chemical weapons) [4]

**Mitigation :** These effects can be mitigated by cautioning against direct use of the model to synthesize products, but instead for it to be used as a supplementary resource by chemists.

Representation learning for songs:

https://cs229.stanford.edu/proj2017/final-reports/5218770.pdf