CHEMICAL ENGINEERING

UW Data Science Graduate Capstone Project Proposal

The project should allow trainees to cement the acquisition of data science skills and develop proficiency in the conduct of team-based interdisciplinary research

PROJECT NAME: Achieving High Resolution at Low Cost via Machine Learning: The Case for Nuclear Magnetic Resonance (NMR)

SPONSOR NAME: Dr. Vikas Varshney (Air Force Research Laboratory); Dr. Davide Simone (Air Force Research Laboratory)
May we list you on our website as a partner DIRECT project partner? Yes
Will graduate students be asked to sign a non-disclosure agreement? ***This is something that I will connect back with you. Most likely the answer is no but I want to make sure.***
<< This information is for planning purposes only. However, there must be a final 'public-facing' version of the project for student portfolio and presentations. Please ask clarifying questions as needed >>

PROJECT DESCRIPTION:
Higher resolution data, irrespective of its origin (images, spectra, signals) is associated with larger acquisition, operational, & maintenance costs. Data resolution is often limited by the quality of acquisition media (detector, filter, camera resolution). In cases where the underlying physics of data generation is not affected by acquisition media, it should be possible to derive frameworks to upscale lower resolution signals via machine learning (ML). Our specific interest is in developing a framework to upscale low-resolution bench-top 60MHz NMR data to 400MHz, a 7x resolution improvement, for better identification of reactants & products in synthetic chemistry reactions. While high-resolution NMRs do provide better quality data, it comes with notably higher acquisition, operational, & maintenance costs. The developed framework will result in better return of investment associated with these costs and further enable the broader use of bench-top NMRs as an in-line characterization tool with upscaled high-resolution data. Such advancements can also accelerated closed-loop synthesis on variety of autonomous platforms.

It is expected that the framework will be developed using encoder-decoder networks ideas. Instead of providing explicit encoding & decoding functions, convolutional neural networks can be used to learn these functions to transform 60MHz spectra to 400MHz. While the long term vision is to incorporate the framework as an in-line characterization/policy guiding tool for autonomous experimentation, this seed project can test the hypothesis via generating synthetic NMR data @ 600MHz & 400 MHz using open-source tools (nmrsim), training encoder-decoder based ML model, & validate it using experimental datasets. If successful, not only this concept can be applied to further upscale the resolution (>400 MHz) using advanced ML (transfer learning) methods, its applicability can also be expanded to other domains where resolution plays a key role.

DESCRIPTION OF DATA TO BE USED: The synthetic data for training ML models will be generated using nmrsim tool. The framework for data generation is already exits. It will be composed of 1-D spectral data consisting of NMR peaks with random configurations (chemical shifts, coupling, multiplicity) for 60 MHz and 400 MHz.

PROJECT START DATE: 1/3/2023	**PROJECT END DATE:** 6/15/23

UNIVERSITY of WASHINGTON

Box 351750 105 Benson Hall Seattle, WA 98195-1750

206.543.2250 fax 206.543.3778 cheme.washington.edu

PROBLEM TO SOLVE/OBJECTIVE:
<< describe the problem in a bit more detail and what are the specific objectives >>
The background of the problem is discussed above. The specific problem would be to test the hypothesis that ML can learn the intrinsic physics that govern the resolution and accurately upscale lower resolution spectra

Objectives:

1. Train machine learning models of synthetic data. Explore hyperparameter tuning.
2. Explore how complex one can go with respect to NMR spectra randomization (in terms of complexity) with sufficient accuracy of upscaling
3. Utilize real experimental data to use further augment the model using transfer/active learning.

TIMELINES AND DELIVERABLES: << outline the expected work plan as well as what is expected to be delivered at the end of the project. The work plan should include the use or development of Data Science software/tools >>

As listed in the objectives, specific deliverables and their timelines can be as follows.

1. Generate synthetic data using nmrsim tool while developing software engineering skills. (~end of Jan)
2. Train Version 1 of ML models on the generated data, focus on hyper-parameter as well as model optimization. Validate against test synthetic data (~mid-March)
3. Augment ML model to incorporate more complex NMR spectra, also adding noise to synthetic data. Optimize the model (~end of April)
4. Further augment the model by incorporating real experimental data for small molecules NMR at both 60 and 400 MHz. (~mid-May)
5. Dissemination of the research as potential publication as a publication draft. (~mid-Jun)
6. Developing a github like repo of the tool, graphical user interface, etc. as applicable as part of the deliverable (~mid-Jun)

* our students are versed in Python and SKLearn environment but can quickly pick up other languages/environments

PROJECT MENTOR(S):
Dr. Vikas Varshney (Air Force Research Laboratory) vikas.varshney.2@us.af.mil
Dr. Davide Simone (Air Force Research Laboratory) davide.simone@us.af.mil
Dr. Patrick Hewitt (Air Force Research Laboratory, UES, Inc.) patrick.hewitt.ctr@us.af.mil

UW FACULTY CO-ADVISOR: David Beck

PROJECT TEAM MEMBERS: