**National Cancer Institute (NCI)-US Department of Energy (DOE) Collaboration Use Case for AMPL on Azure Machine Learning**

**Introduction**

This Use Case Template is intended to describe and interpret real-world examples of computational resources developed by the NCI-DOE Collaboration—and others—that can be used to explore solutions to complex cancer-risk factors, cancer diagnosis and treatments, as well as related biomedical research challenges. This template was developed with input from external subject matter experts. We ask you to share your experience in using these computational resources by completing this template in accordance with the instructions below. To provide feedback please contact *computational-cancer@nih.gov* Thank you!

**Instructions**

Please fill in the information requested. Each section includes the level of expected length or level of detail: **brief** (one-two sentences), **informative** (two-four sentences), and **detailed** (as much detail as required to properly inform the user). As each use case is different, please provide additional information as necessary.

**Use Case Overview**

This use case aims to provide users of the ATOM Modeling PipeLine (AMPL) with a template for application on Azure Machine Learning Studio (AML). For this example, we are using the QMugs datasets, a quantum mechanical (QM) property dataset which consists of over 600K unique SMILES strings and their respective QM data. This example use case will provide the framework to perform a graph convolution neural network (NN) model to make predictions on the HOMO, LUMO, and HOMO-LUMO gap values of this dataset. This use case can be used on CPUs or GPUs provided by AML, depending on how the users AMPL environment has been set up. After completion of this use case, the user can expect to have a NN predictive model that is generalizable as indicative of the train, validation, and test R2 scores. You can find all necessary files in the GitHub repo at <https://github.com/joverhul/AMPL_Azure>. For instructions on how to run AMPL not on AML, please go to <https://github.com/ATOMScience-org/AMPL>.

**Use Case Details**

Scientific Background and Goals

* **Problem**: This use case uses the ATOM Modeling PipeLine (AMPL) to train models and make predictions on molecules. Users can use Azure Machine Learning (AML) resources to perform the AMPL pipeline. In this example, a NN model was trained to predict HOMO-LUMO gap values using the QMugs dataset. This quantum mechanical property is important in determining the stability of compounds.
* **Background**:AMPL is an end-to-end modeling pipeline that allows users to perform data cleaning, curation, training, and visualizations all within the pipeline. There are several tutorials available online to walk by training models using AMPL. This use case is specific for HOMO LUMO gap values but can be adjusted to run predictions on any property of interest.
* **Scientific purpose**: This use case provides a working example for how to use AMPL both as a submitted job but also as a Sleep Job using VSCode or Jupyter notebook on Azure Machine Learning to train a predictive model for quantum mechanical properties.
* **Goal**: This use case will produce a machine learning predictive model for quantum mechanical properties, specifically the homo and lumo gap values, using chemical structures and QM values as input. The model can give insight into the stability of drug like compounds. This model could also be used with transfer learning to create additional models that are finetuned for the users problem.
* **Objective(s)**: This use case provides a specific example of creating a machine learning predictive model using smiles strings and homo lumo gap value data as the input to the model. This serves as a general example of how to use AMPL within AML, giving users an option to use cloud based computing if they do not have access to an HPC or other computing resources. This use case provides a basic example for how to use AMPL on AML.
* **Constraints**:This example is meant to develop a tool that helps medicinal chemists filter out problematic, reactive compounds from the predicted, generated compounds. As with all machine learning models, users should be cautious that this model corresponds to their dataset, and is not too far from the molecular domain.

User Background and Skills

* **Education**: This use case is a specificexample to train a model on quantum mechanical properties using AML. It is recommended that users have a working knowledge of chemistry, machine learning approaches, and AML. Tutorials for AMPL can be found on the GitHub.
* **Skills required**: A working knowledge of Python is required and familiarity with AML, chemistry, quantum mechanical properties, and SMILES representation of chemical structures is suggested. AML provides great support in case of issues with running jobs that are specific to AML.

Methods

* **Computational resource selection**: This use case was selected as AMPL has been used extensively and successfully by more than 100 students in multiple projects, and not all students and users have access to the necessary compute resources to train models and run predcitions.
* **Inputs**: This use case employs tabular text data, which includes the HOMO, LUMO, and HOMO-LUMO gap value data along with SMILES string representation in four separate columns. The dataset used can be found on the GitHub. Before training the large model, it is recommended to only take a subsection of the dataset.
* **Methods**: This use case is designed to be use on AML. Users can submit jobs directly to the queue, where the training will run on the requested resources, or users can run the jobs using a sleep job, where users can interact with AMPL in either Jupyter or VSCode. Users should first create an AMPL environment as highlighted in the GitHub. Users should also clone the AMPL environment into their workspace. Input files should be added as assets within AML. Users will then curate their datasets either by submitting a job, or using the Sleep Job for interactivity. Once a curated dataset is prepared, this is the input file for the dataset splitting or model training. The predictive model will be generated as an output .tar.gz file which can be used to make predictions on external datasets. This use case can be run using either CPU or GPU provided by AML.
* **Workflow figure**:

**

* **Methods customizability**: The methods for this use case can be customized in several ways. AMPL as a pipeline is highly customizable, providing the ability customize the featurization, the splitting, the model type*.* The input dataset in this example is also of HOMO LUMO properties, but users can use any molecular data such as solubility, logP, hERG, IC50, or anything else of interest.
* **Common pitfalls**:Some common issues with this computational resource is not correctly adding the assets to the yaml file and within the job submission. Other pitfalls could include issues with the computational resources on Azure, which may crash occasionally. Additionally, if users do not properly clean and curate their datasets, they may run into issues of duplicates, or NaN values.
* **Requirements**:This use case requires dedicated cpu or gpu resources. The jobs can run on non-dedicated resources but could result in jobs failing due to time depending on parameters set. This use case also requires an AML subscription.

Results

* **Outputs**: The outputs from this use case provide insights into the machine learning approach on AML as well as provides a reusable model for predicting the HOMO, LUMO, and HOMO-LUMO gap values*.* Users are able to use the AMPL tutorials to prepare any visualizations available based on the model training results. For this specific NN use case, the output model had a training R2 score of 0.832, a validation R2 score of 0.788, and a test R2 score of 0.781. The output of this use case is only the model .tar.gz, but users can modify this to print plots plotting the R2 score as seen in other use cases for AMPL.
* **Results**: The final model prepared was a graph convolution NN model, using a scaffold splitter (0.7, 0.15, 0.15). The batch size was set to 50, and the best epoch for training was 29. Additional meta data results can be found in the GitHub.

Discussion

* **Results interpretation:** The generated model provided as output suggests that this model works well as a generalizable model but does have some limitations in the overall generality with R2 scores below 0.8. The model chosen for this use case was NN, which provided promising initial results. Further hyperparameter optimization should be performed as a follow up. Overall, this model could be used as a general model to predict the stability of drug-like compounds.
* **Objective(s) met**:The original objective was met, and users can use AMPL on AML, training their own predictive models. The model is generalizable, but further hyperparameter optimization could be performed to further increase the test and valid R2 scores.

Conclusion

* **Limitations:** Although the dataset consists of over 600K compounds, users should be wary of the chemical space being explored in the calculations. There was not hyperparameter optimization performed, which could lead to better models. A similar workflow could be applied, but be careful with memory usage on AML.
* **Goal**: This work highlights an example use of the ATOM modeling pipeline (AMPL) on cloud services, specifically AML. This will give researchers the ability to use this pipeline on services if they do not have access to their own HPC as well as if they are using a larger dataset.

**Key Links to Related Portal Pages:**

*Related resources or activities relevant to this use case may be found at the links below.*

*NOTE: these are intentionally not hyperlinked because the site is under development.*

* Publications

AMPL: A Data-driven Modeling Pipeline for Drug Discovery. <https://pubs.acs.org/doi/10.1021/acs.jcim.9b01053>

Isert, C., Atz, K., Jiménez-Luna, J. *et al.* QMugs, quantum mechanical properties of drug-like molecules. *Sci Data* **9**, 273 (2022). https://doi.org/10.1038/s41597-022-01390-7

* Computational Resources: Models, Software, Datasets

<https://github.com/ATOMScience-org/AMPL>

<https://modac.cancer.gov/assetDetails?dme_data_id=NCI-DME-MS01-106757106> (curated QMugs dataset)

* Educational Materials

https://github.com/ATOMScience-org/AMPL/tree/master/atomsci/ddm/examples/tutorials

<https://computational.cancer.gov/use-cases/creating-predictive-model-using-atom-modeling-pipeline-ampl>