nanoHUB
UNDERGRADUATE RESEARCH EXPERIENCE

Chemical Autoencoder Tool User Guide

Bryan Arciniega, Mackinzie Farnell, Nicolae C. Iovanac, Brett Savoie

PURDUE UNIVERSITY
Network for Computational Nanotechnology
DISCOVERY PARK

NSF

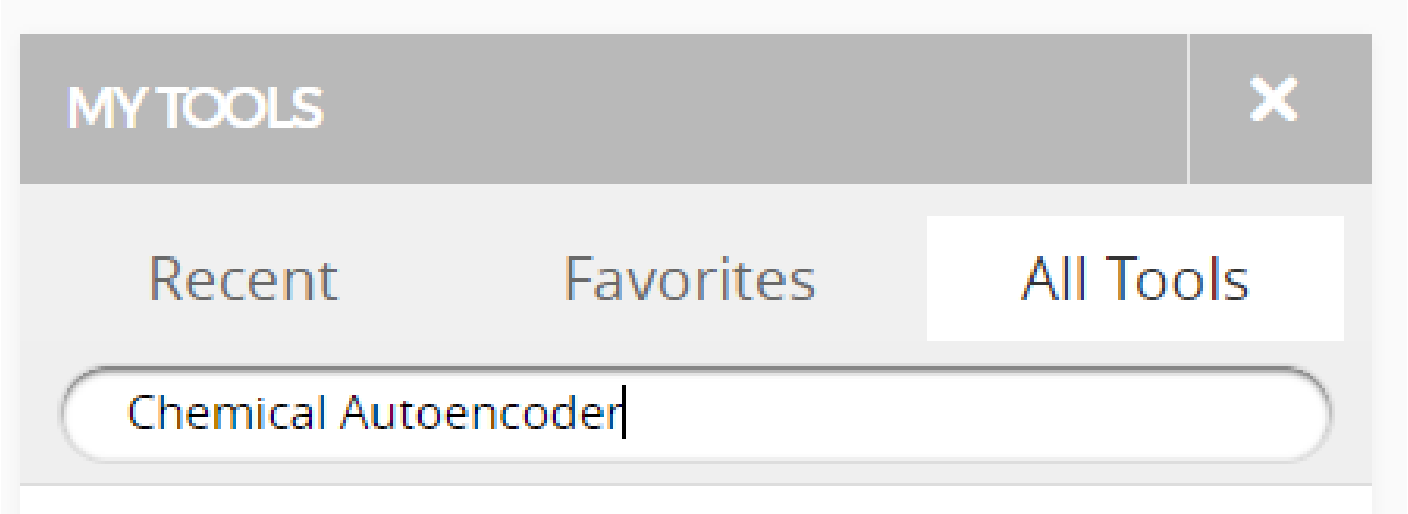
User Guide Contents

Background Information

- The Chemical Autoencoder tool for latent space enrichment is designed to allow users to joint train autoencoder models on property prediction and visualize the latent space of these models with principal component analysis plots
- The tool is based on the paper: Improved Chemical Prediction for Scarce Data Sets via Latent Space Enrichment by Nicolae C. Iovanac and Brett M. Savoie
  - Refer to this paper for more background information

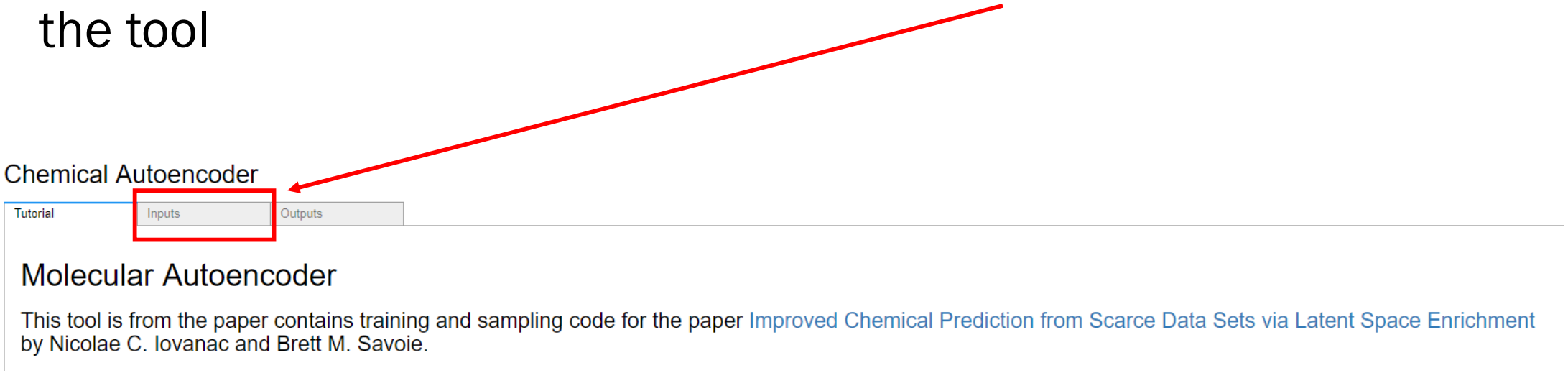
How to Access Chemical Autoencoder Tool

- Log on to your account on nanohub.org
- To find the tool, do one of the following:
  - Go to: https://nanohub.org/tools/chemae
  - Search for the Chemical Autoencoder for Latent Space Enrichment Tool

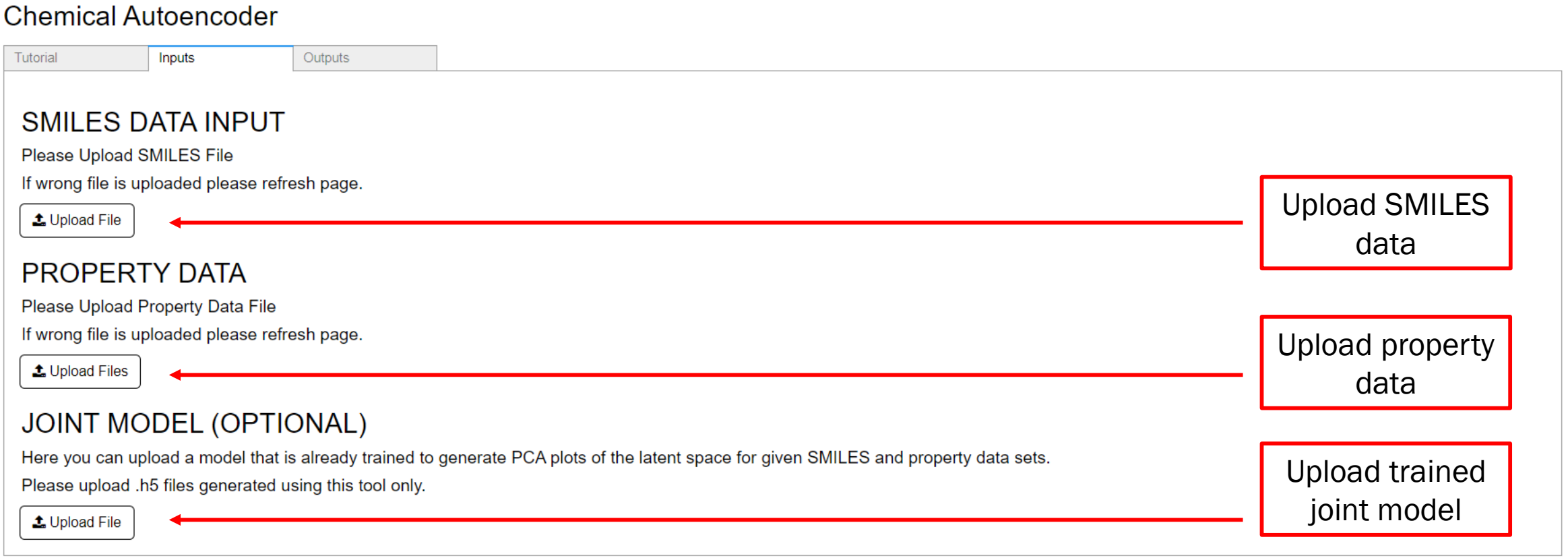


Front Page of Tool

- The front page of the tool is a short tutorial. Refer to this tutorial to find the minimum amount of information required to run the tool
- After reviewing the tutorial, move to the ‘Inputs’ tab to begin using the tool



Tool Inputs


Chemical Autoencoder
Tutorial
Inputs
Outputs
SMILES DATA INPUT
Please Upload SMILES File
If wrong file is uploaded please refresh page.
Upload File
PROPERTY DATA
Please Upload Property Data File
If wrong file is uploaded please refresh page.
Upload Files
JOINT MODEL (OPTIONAL)
Here you can upload a model that is already trained to generate PCA plots of the latent space for given SMILES and property data sets.
Please upload .h5 files generated using this tool only.
Upload File
Upload SMILES data
Upload property data
Upload trained joint model

Required File Format – SMILES Data

- Either a .txt or .csv file can be uploaded
- For the .txt files, the SMILES should be separated by newline characters
- SMILES strings should be no longer than 80 characters
- SMILES strings may only contain the following unique characters:
  - " ", "B", "b", "C", "c", "N", "n", "O", "o", "F", "Si", "P", "p", "S", "s", "Cl", "As", "Se", "Br", "I", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "(", ")", "[", "]", ".", "#", "@", "=", "-", "\\", "/", "+", "%", "H“
- For more information on SMILES strings, please refer to: https://archive.epa.gov/med/med_archive_03/web/html/smiles.html

SMILES validity check

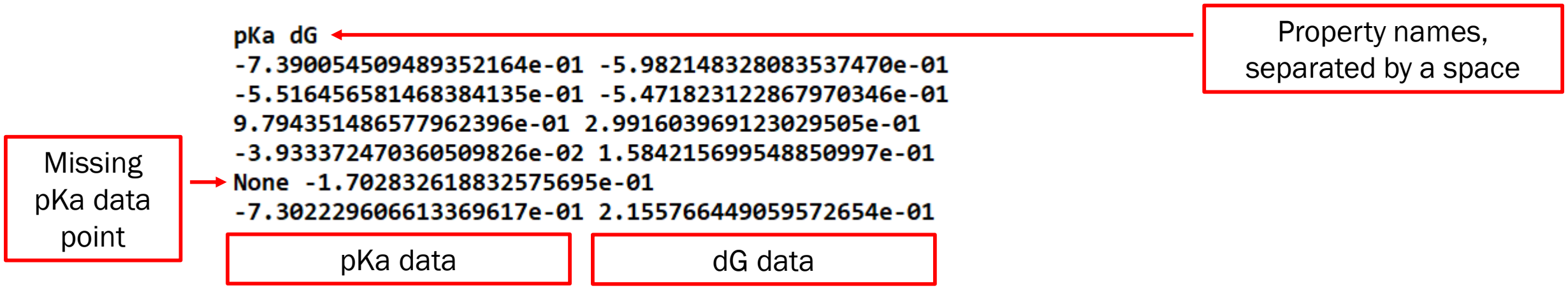
- Upon uploading SMILES data, the validity of each SMILES string is evaluated using the following command:
  - Chem.MolFromSmiles(smile) from RDkit library
- If there are invalid SMILES, the tool will print the invalid string and the index of that string in the list of SMILES (where the indices go from 1 to number of SMILES uploaded)
- If you know these SMILES to be valid, continue running the tool
- If you decide to remove these SMILES from the list, you will have to refresh the tool page and re-upload the new SMILES file

Alert Image:

```
ALERT! SMILES:O[C@H]1[CH]23C[C]12(CC#C)O3
 in index: 10 is not a valid molecular structure.
ALERT! SMILES:C[C]123O[CH]41[C@@H]2CC[C@@H]34
 in index: 198 is not a valid molecular structure.
ALERT! SMILES:CC[C]123C[CH]1(O2)[C@H]3CO
 in index: 502 is not a valid molecular structure.
ALERT! SMILES:C#C[C]123C[CH]1(C2)O3
 in index: 523 is not a valid molecular structure.
```

Required File Format – Property Data

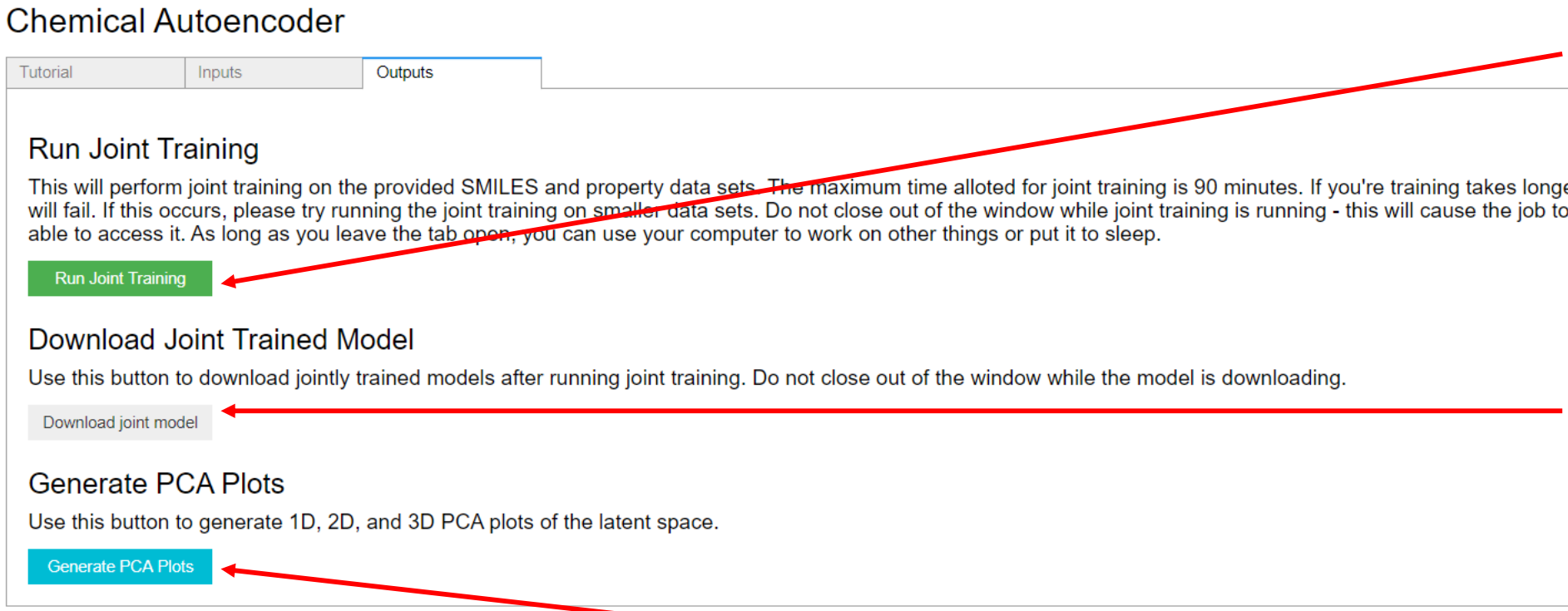
- Property data file can be uploaded in .txt or .csv format
- For .txt files, there should only be one space between the properties in each row
- Please ensure that each property is indexed the same way the SMILES strings are indexed (i.e. the first row of properties corresponds to the first SMILE)
- If a property data value is missing, write 'None' in place of the value
  - Failure to do so will result prevent the tool from running or produce invalid results
- You may upload data to train on as many properties as you would like
- Please include the names of the properties in the first line of the file, corresponding with each column of properties



File Requirements – Joint Model File

- Uploading a joint model is optional
- The uploaded joint model will only be used to generate PCA plots and will not affect joint training
- If a joint model is not uploaded, PCA plots will be generated using a model trained only on reconstructing 128,000 SMILES with no property prediction task
- All joint models uploaded must have been trained using the Chemical Autoencoder for Latent Space Enrichment tool so that they have the correct architecture
  - Upload the .h5 file generated by joint training a model

Tool Outputs



Run a joint training on uploaded SMILES and property data

Download a trained joint model – must be pressed after joint training runs

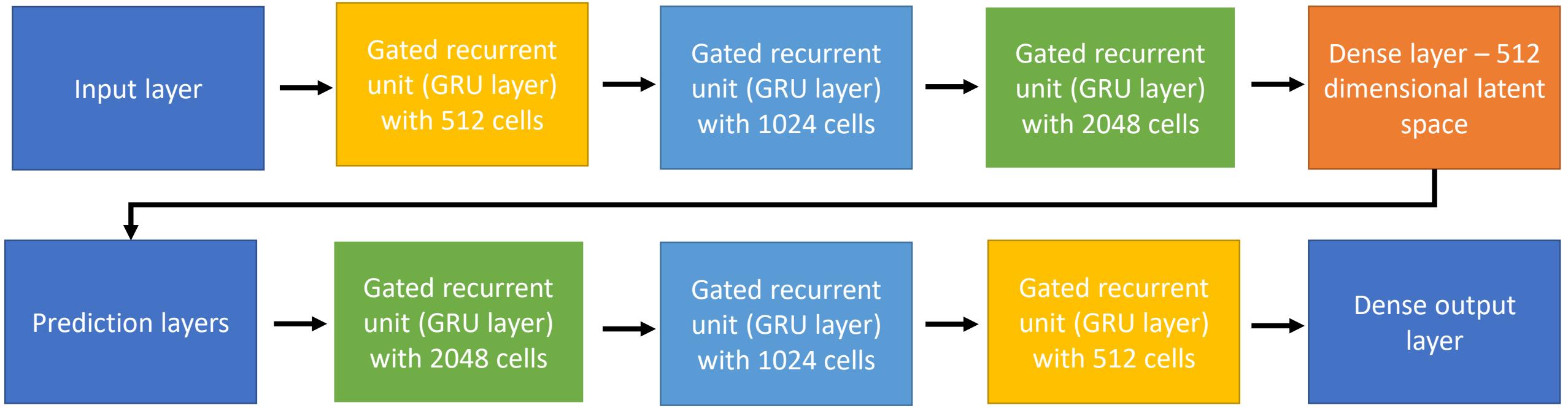
Produce PCA plots based on uploaded SMILES and property data and a previously joint trained model

Run Joint Training Button - Details

- Joint training will be run on Gilbreth GPUs
- Before running joint training, make sure you have at least 2 GB of storage available on your nanoHUB account - the tool will use this space to save the jointly trained model.
- Hyperparameters used for joint training:
  - Batch size: 64
  - Dimensions of Latent Space: 512
  - Epochs: 300
  - Learning Rate: 2×10^{-4}
  - Predictor Dropout: 0.50
  - Predictor Loss: Mean squared error
  - Predictor Metric: Mean absolute error

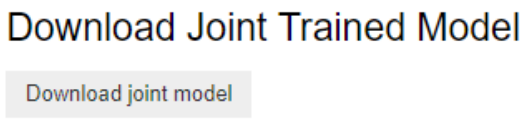
Run Joint Training Button - Details

- Model architecture is built in Keras with a tensorflow backend:



Accessing Joint Model After Training

- Primary way to access model: download button



- Other option for accessing model: go to the nanoHUB Workspace tool and use the following command to find the joint model file:
  - find ${HOME} -name joint_model.h5
    - This will find all jointly trained models you have produced using the tool. We recommend deleting old joint_model.h5 files to allow you to more easily find newly generated files.
  - After finding the file, you can download it through the Workspace tool
  - This option requires you to have access to the Workspace tool and some experience using UNIX but downloading through the Workspace may be faster than downloading directly through the tool.

Generate PCA Plots Button

- This button will generate 1, 2, and 3D PCA plots that condense 512-D latent space
- Latent space is generated based off a previously trained joint model
  - The joint model uploaded by the user
  - Default model is an autoencoder trained only on SMILES translation
- The code will also calculate R2 and p for the PCA plots
  - Refer to Iovanac and Savoie 2019 Supporting Information to see how R2 and p are calculated for 2D PCA plots

PCA plots

Results of statistical analysis of PCA plot

Dropdown menu to display PCA plots of each property

Plotly menu bar

SMILES string and image displayed on hovering over plot

Plotly Menu Bar

- Descriptions of the menu bar icons pop up when you scroll over the menu bar, as shown below:

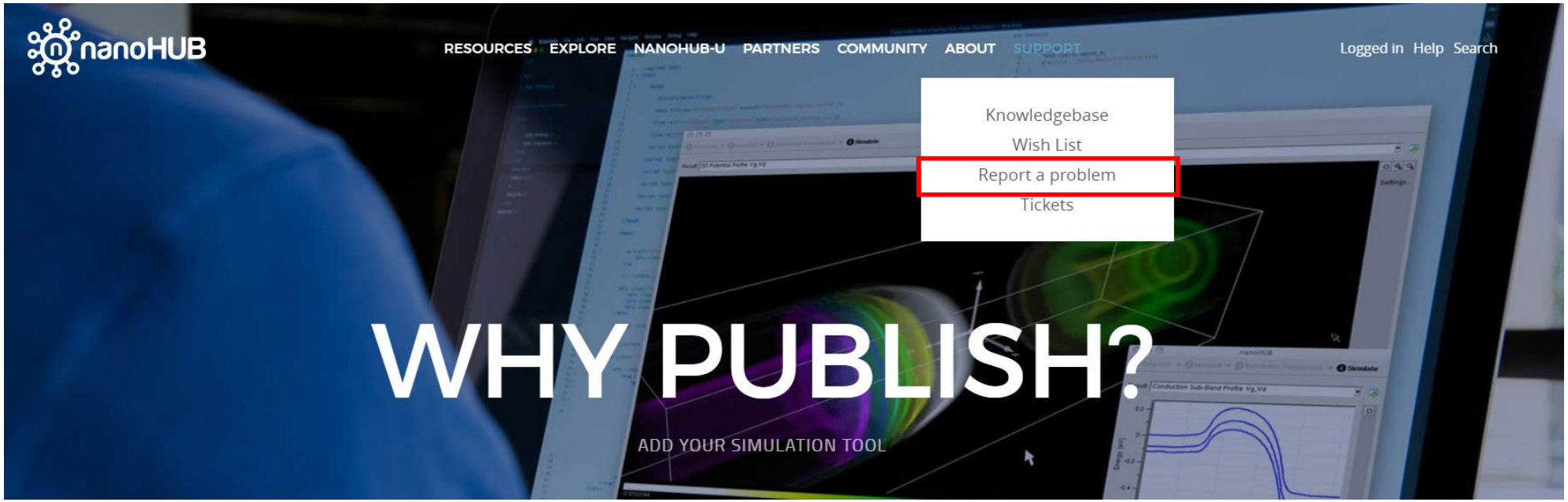
- Some icons we found particularly useful are:
  - Download plot as png
  - Zoom
  - Pan
  - Reset Axes

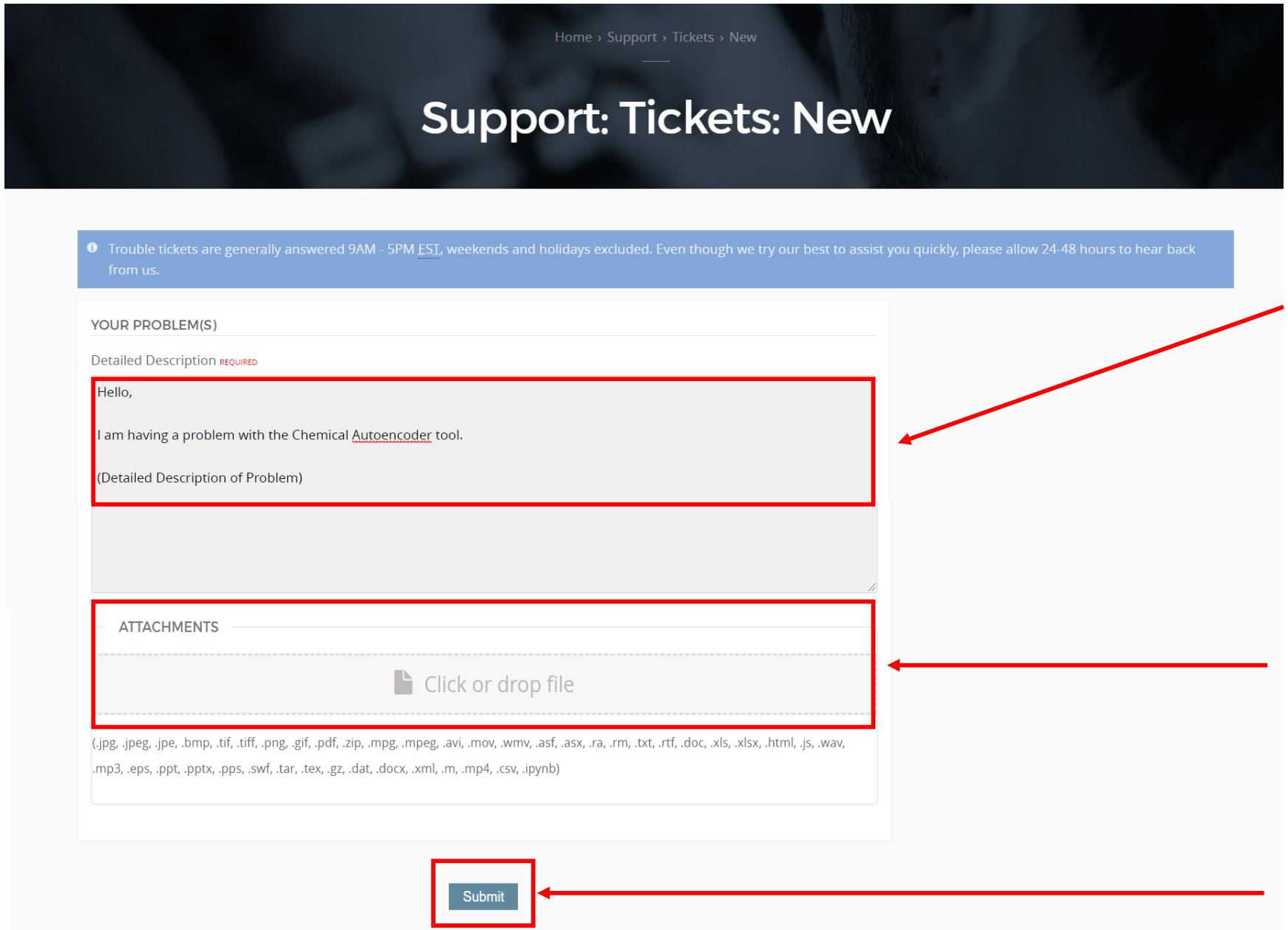
Potential Errors – File Upload

If the progress bar remains grey for more than 5 seconds after uploading a file, then there has been a problem with the upload. Please refresh the page and try again. The progress bar will be green when the file is uploaded.

What to do when there are errors:

- When you encounter an error, please read the statements printed within the tool to determine if all files were correctly uploaded.
- If the error persists, submit a support ticket on nanoHUB
  - Go to Support tab and click ‘Report a problem’





Include that you are having problems with the Chemical Autoencoder tool and provide a detailed description of the problem

Attach pictures to further explain the problem and any files required to reproduce the problem

Hit the Submit button to submit your ticket