1. This data set includes two folders:
   1. ‘data’: include three items; namely,
      1. *“names\_rTypes\_classes\_superclasses\_training\_test\_set\_patent\_data.pkl”*: class, superclass names
      2. *“reactionTypes\_training\_test\_set\_patent\_data.pkl”*: reaction types (50 types)
      3. *“training\_set\_patent\_data.pkl.gz”*: reaction data (major file, 40,000 reactions; 800 for each reaction type)
   2. ‘Notebook’: it is a Jupyter Notebook; this file illustrates how to import ‘pkl’ file and showcase reaction SMILES; SMILES is The simplified molecular-input line-entry system (<https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system>)
2. This Jupyter Notebook can be executed under Anaconda software (<https://www.anaconda.com/>). The python language is Python3. One needs to install several packages to perform the notebook (pickle, gzip).
3. The purpose of this set to familiarize with reaction data structure.
4. The data is originally from patent data:
   1. See references: <https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873>
5. Point of inquiry: Shiyan Wang ([wang2502@purdue.edu](mailto:wang2502@purdue.edu))