# BMS Molecular Translation Project

## Domain Considerations

* Inchi is a structured representation with multiple *layers*. Each layer provides different information to reconstruct a portion of the chemical composition.
* There is a main layer with chemical formula and other optional layers, each with a prefix denoting the type of information.
* We can take advantage of this structure and break up the problem into multiple translation problems:
  1. Extract the chemical formula of molecule given image
     1. Extract (Model 1.a) atomic symbols that image consists of (multi-word tokens, usually single letter)
     2. Extract (Model 1.b) frequency of each atom
  2. Enumerate over each optional layer
     1. Predict (Model 2.a) if this optional layer is required for this image and chemical formula (the prediction of step 1).
     2. If this layer is required, then Extract (Model 2.b) its string representation given the image and the chemical formula (the prediction of step 1)
  3. Compose these intermediate outputs into a Inchi string representation (final output).
  4. For prediction we call the model pipeline in order of dependencies.
  5. For training
     1. We would need to parse the Inchi codes to split into layers
     2. Train each model separately while handling dependency in model pipeline.
     3. Teacher force outputs, for instance result of Model 1 with actual inputs to Model 2.
     4. Gradients will flow from each Inchi layer model to convolution model.
* The input image is a composition of structures as well. There are a finite set of alphabets and numbers albeit in different position, scale and rotation orientations in an image. The connection lines and glyphs are symbolic as well, for instance a double line is different than two one lines and need to be treated as a different symbol.
  1. Perhaps we can use pre trained convolutional layers for alphabets and digits data readily available on internet.
  2. Pre train convolution model with different connection glyphs and different rotation orientations of them. (need to manually prepare a small set of labelled data)

## High Level Design

Diagram

Description automatically generated with medium confidence