Pond of Ideas

Some ideas for the next big project to do, some of them are:

1. GAN/Generative element on top of Alchemite [Potential project for Alex to help with]. We have Alchemite/Other methods train a great model for Endpoints and fill in the gaps. We then select desired properties. It seemed that Alchemite could then propose descriptors that will make those properties (i.e. from the alloys dataset example). Then we can fill in the StarDrop descriptors (and others). From here we can trigger a generative process to find compounds which match the StarDrop descriptors (i.e. a simple inverse function on the descriptors). This way Alchemite is doing the hard work for the activity/ADME model, and the GAN is doing only one job rather than both structure and target as in the regular implementations in the literature. [The only part we need is a version of Alchemite that proposes the values for the StarDrop descriptors]…  
   GuacaMol might be related.
2. Text association: Some of the most useful (sparse) information we might want to produce is based on text. Some compounds have names/Wikipedia pages/Database entries etc. New compounds don’t. So we have a column which is populated for some entries, we might have molecular descriptors for all of them. Can we ‘impute’ the words that would be associated with a compound if it did have a Wikipedia/Database entry. We can compile (download) bundles of textual information associated with bags of compounds. We can turn bags of compounds into bags of descriptors. Can we derive pharmacophores associated with certain diseases etc from descriptors?
3. Some kind of wider AI that regulates all of the processes.
4. A machine which writes literature reviews given a set of keywords.
5. Random matrix theory? Descriptor generation. Inverse problem. Does it work with Alchemite? If we take a random chemical compound as a baseline, this may be much more effective than fully random. Certain descriptors will have certain distributions which are not quite normal.
6. Apply Alchemite to UGT’s data, to try and fill in the missing activity parts. I.e. we have sparse data with respect to isoforms
7. Extension to Alchemite where we filter distribution samples based on the inequalities, or some other method to deal with them. [Have run through a few ideas with Gareth]

**Classification/Database of weak models:**

* **For example** <https://en.wikipedia.org/wiki/Melatonin_receptor_agonist>
* **This shows a scaffold, a simple Boolean of whether a given compound matches that scaffold would be a descriptor**
* **Could use other things to infer a weak model, I.e. if the model is + ve there is a chance of activity, if the model is - ve it is unclear.**
* **Characterisation of many of these weak models (through small lists of agonists etc. for proteins) might be useful.**
* **Each model has:**
  + **A name, i.e. “general\_kinase\_inhibition” or “adenosine\_receptor\_agonist”**
  + **A serial**
  + **A set of known compounds:**
    - **Primary Compounds**
    - **Synthesis Compounds (steps in the synthesis)**
    - **Breakdown Compounds (steps in the metabolism)**
  + **Methods to generate descriptors**
  + **A formula to combine the descriptors**
* **Examples of models might be:**
  + **Drug metabolism, i.e. UGT’s FMO’s**
  + **GPCR’s**
  + **LGIC’s**
  + **Fluorescence**
  + **Antibiotic**
  + **Anti-Malarial, other diseases**
  + **Anti-Helminthic**
  + **Toxic Compounds**
  + **Hep-tox comp**
  + **Brain-tox**
  + **Massively Inactive Compounds**
  + **Drugs (in the market) which are taken orally**
  + **Drugs (in the market) which are taken intravenously**
  + **Drugs which are documented to cause Gastro-intestinal-irritation (can we measure this from regulatory databases?)**
  + **Side effects, and prediction of side effects (cases per million) based on descriptors?**
  + **Drugs which cause pain relieving symptoms etc.**
* **Libraries of Meta-Descriptors:**
  + **Store for example combinations of StarDrop descriptors (as weight vectors)**
  + **Store NN-based matrix iterative vectors (non-linear)**
  + **Also store custom descriptors if needed**
  + **Use a canonical StarDrop ordering**
  + **Store these meta descriptors with tags**
  + **Models can have lists of tags associated with them**
  + **Vectors which best discriminate two families of compounds**
* For a given (anonymised) data set, we will be able to know roughly the kind of compounds in the data set without knowing what is in the dataset
* This will allow us to partition the models and train the right kind of model and suggest compounds to use (from MDC data and CHEMBL etc.)
* This will in turn allow a Nova like generation step to kick off a larger pool of compounds that might be related and synthetically tractable.
* From this a bespoke GAN model could be triggered.
* Everything will feed into Alchemite.
* Predicted compounds can be screened again against the database to see if any partial hits, or new hits have changed/appeared. This will revise the data source.
* If any compounds are found to be linked to a model, then a note can be made (taking care of sensitive information). I.e. generate SMILES, convert SMILES to names, search the web for the names in pairs, if articles come up, find the keywords in the articles, if a keyword is dense, consider a model with the compound names, search for each compound name, keyword pair, try to find which compounds are the most important.