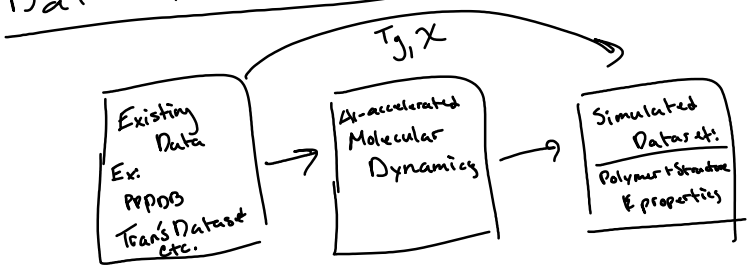
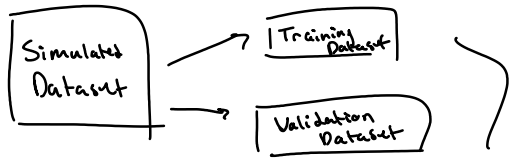
Pipeline Ideation

Sunday, June 26, 2022 2:01 PM

Dataset Creation:



ML Models for Property Prediction



Cross-Validation by k-Fold.

ML Models to Consider

- Gaussian Process Regression
  - + works well on smaller datasets (less MD simulation)
  - ↳ utilized in Polymer Genome (existing ML-based PPP platform)
  - + uncertainty measurements
  - − not sparse (use all features), lose efficiency when features exceed few dozen
- Deep Reinforcement Learning
  - + Molecular Generation model
  - Ø Not as transferable to property prediction (CNN generally don't perform as well for PPP?)
- Random Forest
  - + Concatenating features and select for most important features after training (i.e. MF + QSPR descriptors + ME)
- CNN
  - + identified as possible competitive model (CNN + Representation ME)
  - • 1D CNN exists
- Molecular Generation by Inverse Design (MGenID; made by IBM)
  - • Not necessarily ML, but whatever's
  - • Feature search by Particle Swarm Organization for most influential substructures from existing molecules.
  - • Mainly algorithmic.
- Retrosynthesis by Molecular Transformer (IBM)
  - • Utilized in IBM RXN for Chemistry
  - − requires significant chemical reaction dataset
  - − SMILES ?>:((I guess)
- GeoMol
  - + Provides valid conformers
  - ↳ ties in with RDKit
  - ? affects descriptors & vectorization

Guiding Objectives:

Unity / Low-Budget:
Tool for gaining insight through Machine
Autonomous / High-Budget:
Autonomous tool for
While optimizin

Pipeline Necessities:

- Polymer Property Predi
- Polymer Generation op
- Way to screen generated

Molecular Dynamics:

- Utilize TorchMD for
  - ↳ Uses PyTorch (i
  - ↳ Designed for prot

ML Backends:

- Scikit-Learn for ati
- PyTorch for TorchM

* utilize fo
building b
for DeepRL

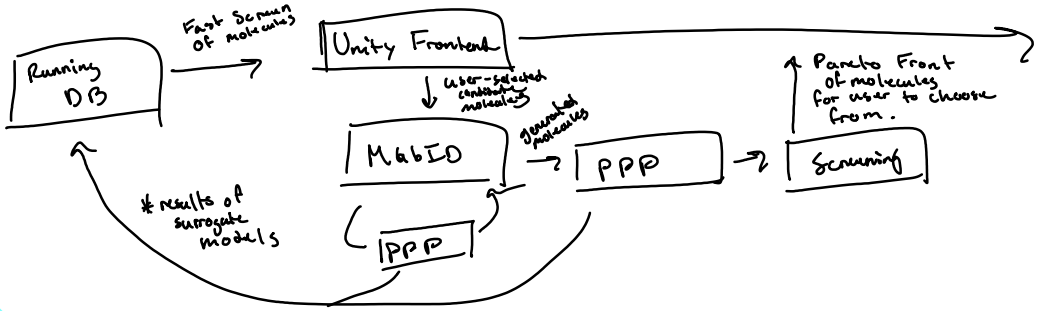
Efficient multi-objective molecular optimizati continuous latent space†

Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé and Dj

UNITY PIPELINE

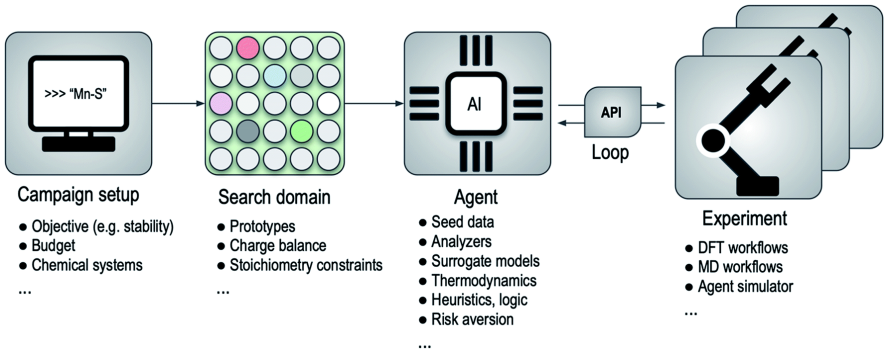
Training Dataset:

- DO NOT generate new molecules for training
- Minimum available dataset, i.e. implement w/ least amount of simulations



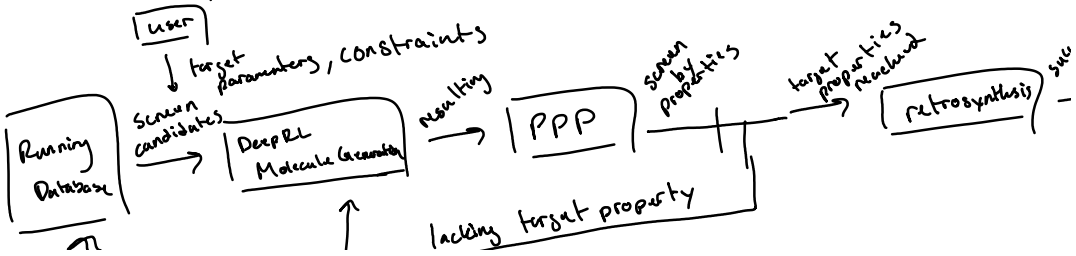
AUTOMATED PIPELINE

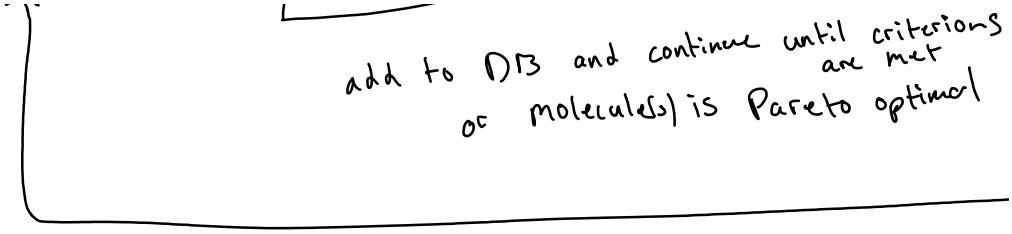
1) Computational Autonomy for Materials Discovery
   - see Montoya et al
   - code on github; open-source framework



2) My "Brash" Pipeline

Training Dataset
- "Randomly" generate new molecules for training to fill in the gaps
- Maximally available dataset
- Utilize GeoMol to enhance featurization.




add to DB and continue until criterions are met
or molecule(s) is Pareto optimal