XenonPy: a Python library for materials informatics

Qi Zhang　合研究大学院大学　統計科学専攻　博士後期課程 3 年

1 INTRODUCTION

XenonPy is a Python library that implements a comprehensive set of machine learning tools for materials informatics. Its functionalities partially depend on Python (PyTorch) and R (MXNet). This package still under hard working.

2 FEATURES

The current release provides some limited features:

- Interface to the public materials database
- Library of materials descriptors (compositional/structural descriptors)
- pre-trained model library XenonPy.MDL (v0.1.0.beta, 2019/8/9: more than 140,000 models (include private models) in 35 properties of small molecules, polymers, and inorganic compounds)
- Machine learning tools.
- Transfer learning using the pre-trained models in XenonPy.MDL

3 ISMD

This chapter is a step by step guidance of ISMD(Inversed Synthesizable Molecular Design), by which we are trying to cover the representative features of XenonPy.
The features included are as follow:

- Descriptor calculation
- Forward prediction
- Molecular transformer
- Sequential Monte Carlo

3.1 Problem setting

The task of ISMD is the identification of new molecules which have the desired properties, this algorithm also explores their synthetic path simultaneously.
Given a set of molecules called reactant pool,

$$R=\{r_1,r_2,\ldots,r_N\}$$

the amount N is very large(e.g. 10^{60}), we are interested in finding some subsets

$$X=\{x_1,x_2,\ldots,x_n\},n\leqslant 4;x\in R$$

as the reactants for a chemical reaction, such that, the properties Y of the product have a high probability of falling into a target region U.

3.1Method

Our goal is to sample from this posterior probability

$$P(X|Y\in U)$$

that is proportional to

$$P(Y\in U|S)P(S|X)P(X)$$

by the Bayes' theorem, $P(Y\in U|S)$ is the likelihood function that is derived by the descriptor and the forward prediction features in XenonPy, $P(S|X)$ is the molecular transformer used for chemical reaction prediction, this is a new function of XenonPy updated with ISMD tutorial. $P(X)$ is the prior that represents all possible combinations of reactants from the reactant pool.

3.3 Descriptor calculation

Descriptors are a numerical representation of complex chemical structures such as molecule and crystal. XenonPy comes with a general interface for descriptor calculation, by using this interface, users can implement their descriptor calculator with only a few lines of codes and run it smoothly.

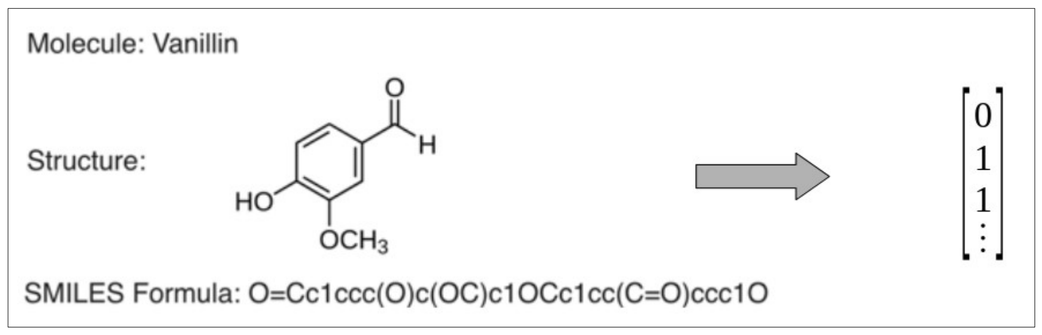

Fig.1 Descript graph structure as a binary vector

3.4 Forward prediction

The prepared descriptor class will be added to the forward model class used in iQSPR. The forward model calculates the likelihood value for a given molecule. iQSPR provides a Gaussian likelihood template, but users can also write their own BaseLogLikelihood class. To prepare the Gaussian likelihood, iQSPR provides two options: 1. use the default setting - Bayesian linear model 2. use your own pre-trained model that output mean and standard deviation.

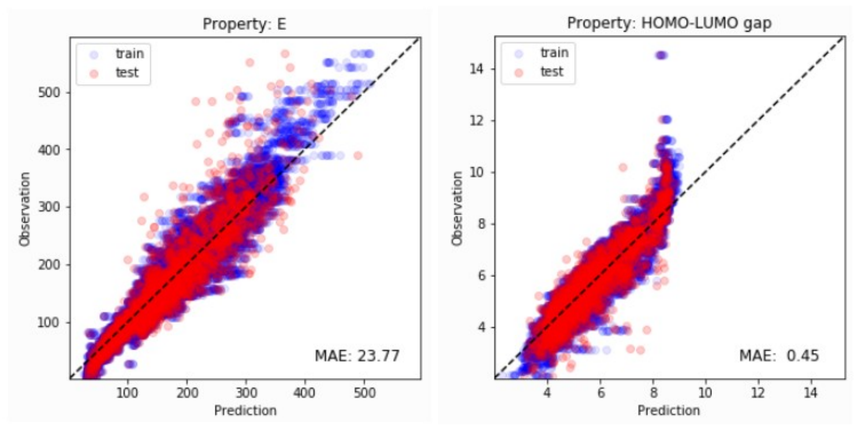

Fig.2 Probability predictions

3.5 Molecular transformer

The Molecular transformer is a chemical reaction prediction module based on attention network(Schwaller et al. 2019), treat reaction prediction as a machine translation problem between SMILES strings of reactants-reagents and the products, possible products can be predicted given the reactants.

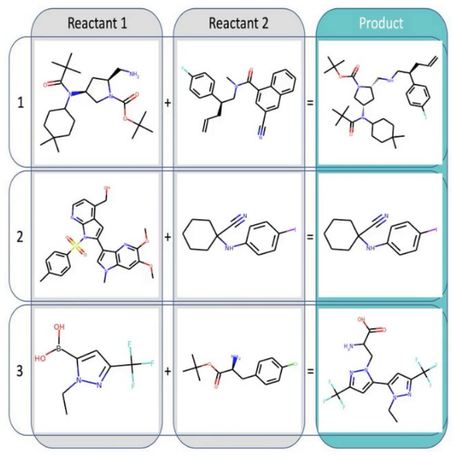

Fig.3 Reaction prediction

3.6 Sequential Monte Carlo

After the preparation of the forward model, molecular transformer, and reactant pool, we are now ready to perform the actual iteration of ISMD to generate reactants in our target property region. The following image shows the likelihood of the products(vertical) at each step(horizontal).

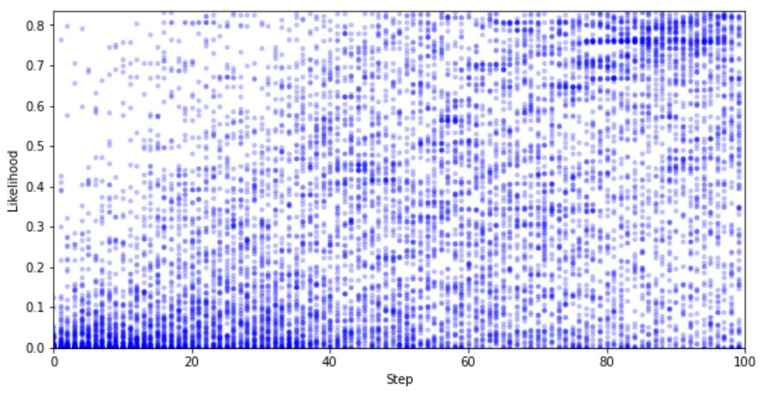

Fig.4 Likelihood of particles in each SMC step

Project homepage:
https://xenonpy.readthedocs.io