Artificial Intelligence for Fast Discovery of Novel Materials for Healthcare

Jinhao Gu, Supervised by Ángel Garcia-Fernandez, Robert Firth (STFC), Rasmita Raval, Joanne O'Keeffe (Unilever)

EPSRC Centre for Doctoral Training in Distributed Algorithms, University of Liverpool, Liverpool, UK

Background & Aims

Innovation in healthcare and personal care both depend on the creation of novel materials for public hygiene and infection treatment. In molecule synthesis process, a chemist would prefer to know the characteristics of the molecules before deciding which ones to synthesis.

In order to perform effective analysis, molecules need to be represented in suitable ways, as different representation may affect the methods for molecular property investigation.

As machine learning has achieved a great success in various areas, it can also, and has already been applied to chemistry. In this project, we target on Gaussian Process models for molecules property prediction for the discovery of novel materials.

Gaussian Processes(GPs): Non-parametric models for Bayesian inference with kernel methods [1].

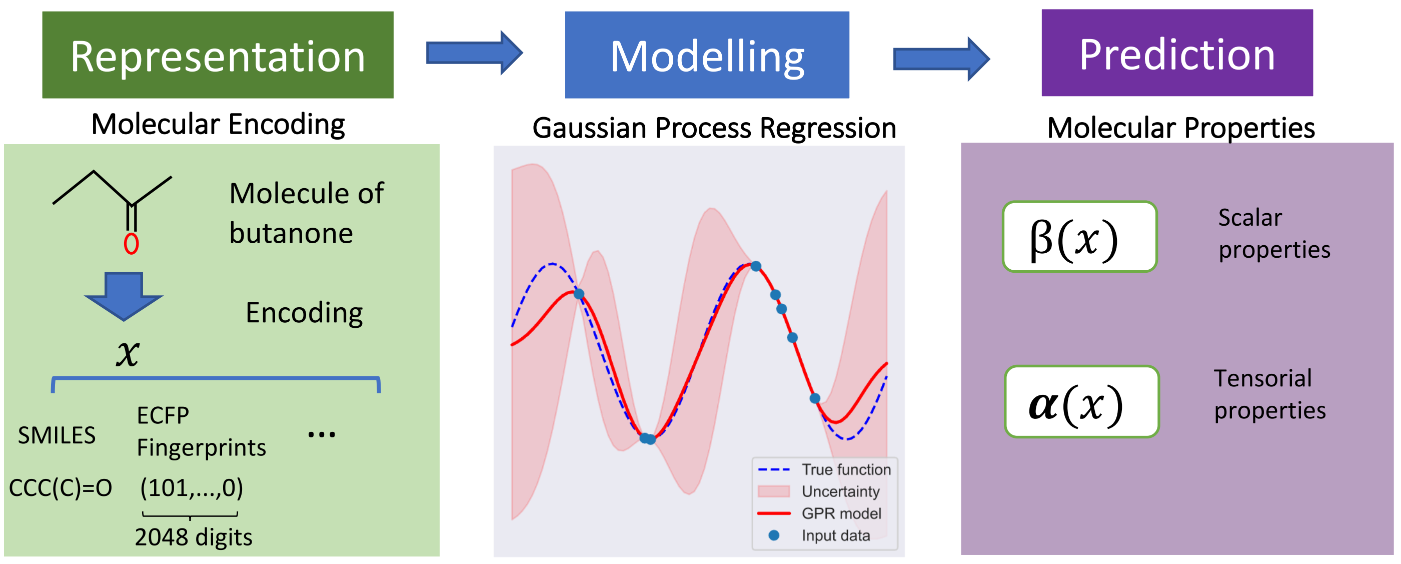
Aims:

Bayesian Optimisation
Machine Learning
High-performance Computing

Accelerate

- Discovery
- Optimisation
- Development

Advanced Materials



Representation → Modelling → Prediction

Molecular Encoding

Molecule of butanone

Encoding

x

SMILES: CCC(C)=O

ECFP Fingerprints: (101,...,0) 2048 digits

...

Gaussian Process Regression

Molecular Properties

$\beta(x)$ Scalar properties

$\boldsymbol{\alpha}(x)$ Tensorial properties

A typical process for molecular property prediction involves three stages:

1) Representation: encoding molecules into forms suitable for specific tasks (left).
2) Modelling: building models from observed data (middle).
3) Prediction: predicting molecular properties with model from previous stage (right).

Ongoing work

Currently we are focusing on the Simplified molecular-input line-entry system(SMILES) based molecule representation. As there are different encoding algorithms for SMILES string generation, a single molecule can have multiple forms of SMILES, and training with only canonical SMILES can lead to biased models [3].

Sub-sequence String Kernels: String kernels measure the similarity of strings through the number of shared strings [4].

Ideally, all SMILES belong to the same molecule should provide similarity values close to 1.

Set String Kernels: To avoid problems of various SMILES representation of a molecule, we treat each molecule as a set, and compute similarities between sets instead of single SMILES strings.

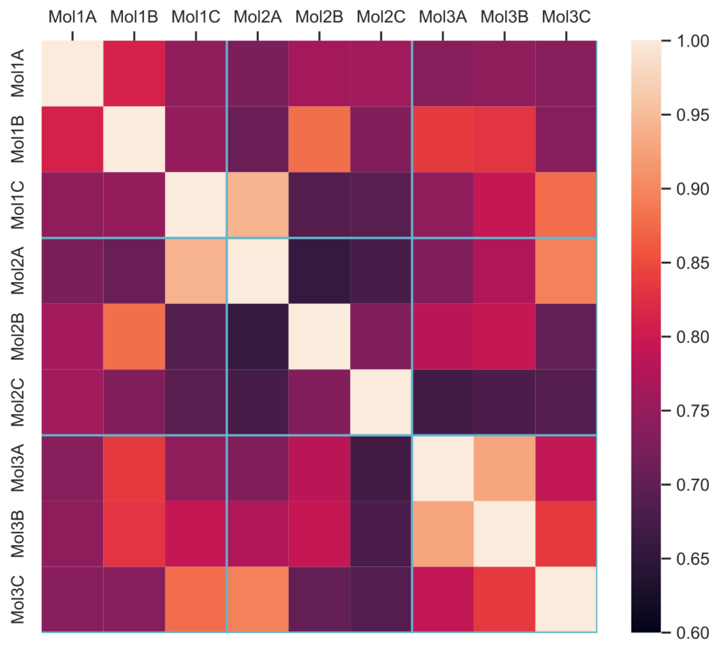


FIG 1: Similarity matrix of different SMILES strings of three molecules with Sub-sequence String Kernel.

Preliminary Results

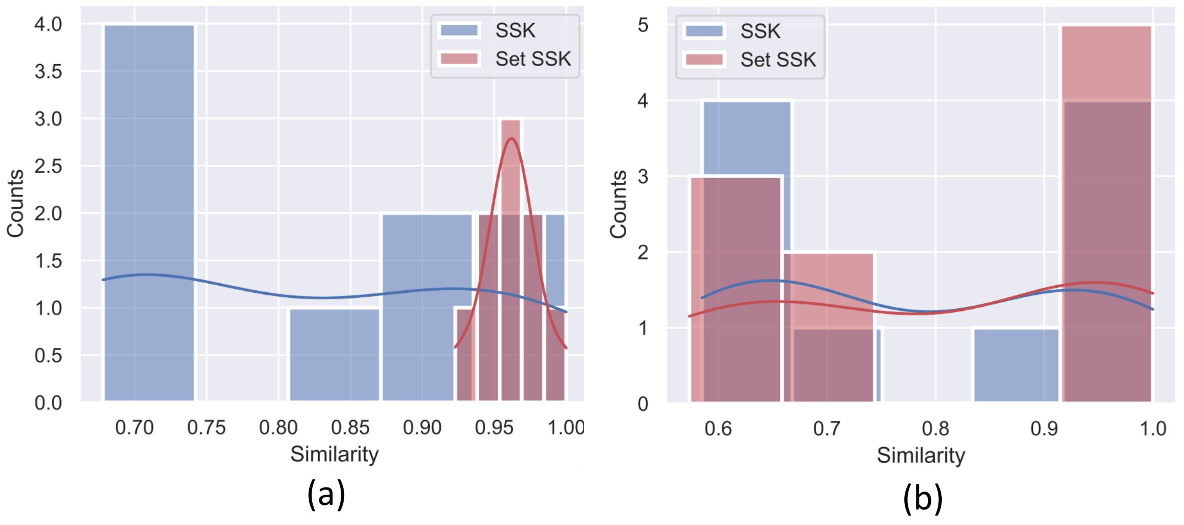


FIG 2: Distribution of similarities (a) within a 10-sized random SMILES set of a molecule; (b) between 10 different molecules with Subsequence String Kernels(SSK) and Set String Kernels(Set SSK).

Set SSK provide generally better performance in both similar and distinct molecules.

However, the computation complexity and no guaranteed positive semidefinite covariance matrix of Set SSK prevent its application in various fields.

Future Work

- **Set SSK:** Implementation with reduced computation complexity and guaranteed positive semidefinite covariance matrix.
- **Molecule Representation:** Exploration for more encoding methods appropriate for different tasks.
- **Gaussian Processes:** Looking for methods that enable Gaussian Processes scalable for large datasets and faster realisations.
- **Application:** Developing a system exploit Gaussian Processes and High- Performance Computing for fast discovery of novel materials for healthcare.

Conclusion

This project will explore the possibility of developing a system incorporating machine learning, Bayesian optimisation and high-performance computing for the development of new generation of novel materials and products to match the need for public healthcare.

References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press, 2006.

[2] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, 'Gaussian Process Regression for Materials and Molecules', *Chem. Rev.*, vol. 121, no. 16, pp. 10073–10141, Aug. 2021, doi: 10.1021/acs.chemrev.1c00022.

[3] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, and O. Engkvist, 'Exploring the GDB-13 chemical space using deep generative models', *Journal of Cheminformatics*, vol. 11, no. 1, p. 20, Mar. 2019, doi: 10.1186/s13321-019-0341-z.

[4] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, 'Text classification using string kernels', *J. Mach. Learn. Res.*, vol. 2, pp. 419–444, 2002, doi: 10.1162/153244302760200687.

Engineering and Physical Sciences Research Council

UNIVERSITY OF LIVERPOOL

Unilever