Spherical CNN for QSAR prediction*

Dorokhin S.[1]***, Popova M.***[2]
[1]Moscow Institute of Physics and Technology, [2]University of North Carolina

Abstract: The task of predicting molecular properties e.g. biological activity or solubility based on the atomic structure is called QSAR (quantitative structure-activity relationship) prediction. It is a classical problem in drug design. Despite various algorithms (quantile regression, radial basis function neural networks) are an acceptable solution, there is still a need for more precise models. A model originally developed for 3D shapes recognition was chosen and put under carefull examination in context of QSAR forecasting. This model is Spherical CNN, first suggested by Taco S. Cohen et. al., who managed to demonstrate that this NN performs well in several applications, including atomization energy prediction. The implemented model is compared with common CNN, RNN, graph CNN and Random Forest.

Keywords: *QSAR prediction, Spherical CNN, drug design*

Introduction

The idea of QSAR (Qualitive Structure Activity Realtionships) is to associate 2D or 3D structural representaion of a molecule with its biological or chemical properties. This research is aimed at building an accurate QSAR prediction tool. There were several attempts to solve the problem. Meryam Zeryouh et. al. [6] used graph representaion and suggested formulas for calculation of Wiener indices of complicated graphs. Wiener indices do correlate with such properties as critical point (Stiel and Thodos [5]) or viscosity (Rouvray and Crafford [4]), but there is still no distinct relation to solubility or target activity, which are of extreme importance in drug design. Nupur S Munjal et. al. [3] performed a non-linear multi-colinearity regression analysis to build model which predicts paclitaxel solubility. However, their model is designed for a specific compound and thus lacks universality. Fatima Adilova and Alisher Ikramov [1] analyzed Matched Molecular Pairs (MMP) method in context of QSAR prediction. They managed to demonstrate that such an approach is inappropiate for QSAR modelling.
The method suggested in this article is based on Spherical Convolution Neural Networks. This concept was introduced by Taco S. Cohen et. al. [2], who defined the correlation of two signals in SO(3) rotation group and the generalized convolution. The unique feature of the CNN suggested in their article is that the abovementioned convolution allows to create a distortion-free projection of a spherical signal. Taco et. al. tested the CNN in various tasks, including prediction of atomization energies from molecular geometry. The model yielded excellent results and this makes applying it to QSAR prediction an interesting challenge.
The main drawback of the suggested model is its complexity resulting in significant number of parameters (around 1M) and huge memory and time resources required. However, the resulting model is expcted to be a universal solution. It is compared with conventional CNN, RNN, graph CNN and Random Forest. (The articles in the intro will soon be replaced)

Problem statement

Let $\mathbb{M} = \{ m_i \mid i = \overline{1,n}\}$ be a set of molecules m_i, each described by 3D cartesian coordinates of all atoms it contains: $m_i = \{ \mathbf{x}_j \in \mathbb{R}^3 \mid j = \overline{1,k_i}\}$, where k_i is the number of atoms in the molecule m_i. Every molecule $m_i \in \mathbb{M}$ has a certain property $y_i \in Y \subset \mathbb{R}$ associated with it,

Scientific advisor: Strijov V. V.

where $Y = \{ y_1, y_2, ... y_n\}$ are molecular properties. Their nature is not of great importance: it may be solubility, toxicity, bioactivity e. g.
Let us consider a set of parametric models $\mathfrak{F}$ derived from convolutional neural networks class: $\mathfrak{F} = \{ f_i \colon (\mathbf{w}, m) \to \hat{y}) \mid i \in \mathfrak{I}, m \in \mathbf{M}\}$, where $\mathbf{w} \in W$ are parameters of a model and $\hat{y} \in \mathbb{R}$ is an estimated property. The task is to predict the property y_i of a molecule m_i based on its spacial structure only. It is considered to be a regression problem assuming $y_i \in N(\overline{y}, \sigma_y)$. Denoting the merit function as

$$E(y, m, \mathbf{w}) = (y - f(m, \mathbf{w}))^2 \tag{1}$$

the problem of training coefficients $\mathbf{w}$ could be represented by the following equation:

$$\hat{w} = \arg\min_{w \in W} \sum_{(y,m) \in (Y,M)} E(y, m, \mathbf{w}) \tag{2}$$

Литература

[1] Fatima Adilova and Alisher Ikramov. Case study: Matched molecular pairs approach in qsar modelling. In *ICISC2017*, November 2017.

[2] Taco Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, March 2018.

[3] Nupur S. Munjal, Narendra Kumar, Manu Sharma, and Chittaranjan Rout. Qsar model development for solubility prediction of paclitaxel. In *INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND SYSTEMS BIOLOGY*, March 2016.

[4] D. H. Rouvray and B. C. Crafford. The dependence of physical-chemical properties on topological factors. *South African Journal of Science*, 72:47, September 1976.

[5] Leonard I. Stiel and George Thodos. The normal boiling points and critical constants of saturated aliphatic hydrocarbons. *AIChE Journal*, 8:527–529, September 1962.

[6] Meryam Zeryouh, Mohamed El Marraki, and Mohamed Essalih. Some tools of qsar/qspr and drug development: Wiener and terminal wiener indices. In *Proceedings of 2015 International Conference on Cloud Computing Technologies and Applications (CloudTechбЂ™15)*, March 2015.