УДК 004.942

Analysis Of Multimodal Data For Machine Learning Based Prediction Of Film Surface Characteristics For Calixarene Compound Groups

Ireddy Ashish. T.S. (ITMO)
Scientific Supervisor – PhD, ведущий инженер Zun Pavel.S. (ITMO)

Introduction. Drug delivery systems are a crucial research agenda in the medical domain. With immense progress being made in bio-medicinal research, there still exist areas where drug effectiveness falls short namely due to secondary factors such as general health, method of drug administration, environmental, etc. Targeted delivery systems can effectively transport drugs that may otherwise take longer to induce. Calixarene group of compounds provide a well-suited platform to support and carry biomedical compounds [1]. Despite its high binding capacity, there still exists ambiguity on its external surface characteristics that are crucial and on par with its molecular structure. By analysing Atomic force microscopy (AFM) data and molecular descriptors via machine learning approaches we investigate correlations to predict surface characteristics using theoretical and experimental conditions of Calixarene compounds.

Main Part. A two-stage data analysis is carried out on theoretical data (i.e. packing factor, molecular sphericity, molecular descriptors from RDKIT toolbox [2] and the synthesis conditions) and experimental data (atomic force microscopy data) of Calixarene compounds generated under various conditions. The obtained AFM scans reveal patterns of aggregates in the shape of spheres. Based on a total of twelve combinations of compound type, solvent, temperature and concentration, we analyse the relation of the formation of spheres on the surface of AFM samples. We introduce three metrics (mean particle size, polydispersity and ratio of spheres) to describe the individuality of each surface extracted using a combination of custom scripts and the Cellpose Plus toolkit [3]. With Correlation analysis and mutual information metrics, we obtain results of underlying correlations. We extend the analysis by modelling the data using linear and random forest regression to classify the behaviour as per their descriptive theoretical properties. The modelling results aim to extract and highlight specific descriptors, and characteristics that play a key role in the prediction of spheres on the surface of Calixarene compounds.

Conclusions. The results showcase the impact of certain theoretical descriptors such as packing factor, electron state, temperature of solvent and compound etc, as most important and influential when predicting surface sphere aggregates. The practical implications allow us to select and generate a regression model that predicts the surface characteristics given the above-mentioned descriptors. Therefore massively minimizing the requirement of physical development of compounds and selection based on requirements of surface properties. Further, we extend our current analysis by selecting a larger set of twenty Calixarene compounds with narrower synthesis conditions, therefore benchmarking the accuracy of results given the prior selected descriptors.

Список использованных источников:

1. Muravev, Anton A., et al. "Calix [4] arene–pyrazole conjugates as potential cancer therapeutics." Bioorganic Chemistry 139 (2023): 106742
2. Huaman, Israel A., et al. "Cellpose+, a morphological analysis tool for feature extraction of stained cell images." arXiv preprint arXiv:2410.18738 (2024).
3. Greg Landrum, et al. Rdkit: 2024_09_5 Release. doi:10.5281/zenodo.14779836.