UDC 620.193

**Aydinsoy E.A., Aghamaliyev Z.Z.,
Aghamaliyeva D.B.**
(Academician Y.H. Mammadaliyev Institute of Petrochemical Processes
of the Ministry of Science and Education)

DEVELOPING AND EVALUATING BAG OF SMILES AND CHEMGRAMS FOR CHEMICAL DATA TOKENIZATION

Abstract. This thesis introduces two novel tokenisation algorithms, Bag of Smiles and ChemGrams, which are designed to revolutionize the field of molecular representation and predictive performance in the context of corrosion inhibitor datasets. These algorithms, developed based on data extracted from academic papers focusing on the efficiency of corrosion inhibitors, offer a fresh perspective by capturing critical chemical structures and contextual interactions within molecules that traditional SMILES-based tokenisation methods often overlook. Bag of Smiles tokenises molecules at the atomic and bond level, capturing positional and frequency information, while ChemGrams emphasises functional group-based tokenisation to identify chemically meaningful fragments. A comparative evaluation was conducted against established tokenisation techniques, such as ChemBERT and Byte-Pair Encoding, using machine learning models for predicting inhibitor efficiency. Results demonstrate that combining Bag of Smiles (with positional and count information) and ChemGrams improves the prediction of molecular efficiency, achieving a Test RMSE of 13.068 and Test R^2 of 0.768, outperforming other methods. These findings underscore the potential of these approaches in advancing machine-learning applications for corrosion inhibitor analysis.

Introduction. In recent years, advancements in natural language processing (NLP) and machine learning have opened new possibilities for analysing molecular structures, particularly in fields like corrosion inhibition. This research focuses on enhancing the predictability of corrosion inhibitor efficiency using data extracted from academic papers and CORDATA [1]. These papers detail various corrosion inhibitors and their respective properties, providing a valuable dataset for machine-learning modelling. Traditional SMILES-based approaches, such as ChemBERT and Byte-Pair Encoding (BPE), have demonstrated success in tokenising molecular structures but often need finer chemical details. Several additional techniques were evaluated in this work, including Smiles2Vec, Word-Level Tokenizer, and SentencePiece [2–4]. Each method was tested against the newly developed algorithms, Bag of Smiles and ChemGrams, which focus on capturing positional, count-based, and functional group information within molecules. By refining tokenisation strategies and focusing on chemically meaningful representations, the study aims to improve the accuracy of machine learning models in predicting corrosion inhibitor performance.

Methodology. In this study, we developed and tested two novel tokenisation algorithms, Bag of Smiles and ChemGrams, to improve the predictive performance of machine learning models on chemical datasets. The Bag of Smiles algorithm is a character-based tokenisation method that captures the atomic and bonding structure of molecules represented in the SMILES format. This method tokenises individual SMILES strings into atomic and bond-level tokens, allowing for the incorporation of both token counts and their positional information. We extended this by considering Bag of Smiles with location action and count, which captures not only the presence of specific atoms and bonds but also their relative positions and frequencies within the molecular structure.

The second algorithm, ChemGrams, is a fragment-based tokenisation method designed to capture chemically meaningful fragments, such as functional groups and ring structures. Using SMARTS patterns, ChemGrams extracts functional groups from molecules and generates n-grams of varying lengths to capture interactions between adjacent functional groups.

In addition to these algorithms, we incorporated molecular descriptors such as molecular weight, LogP (hydrophobicity), and the number of hydrogen bond donors/acceptors, as well as dynamic graph-based tokenisation techniques. These graph-based features, generated using RDKit, allow for the extraction of structural information related to atom connectivity and adjacency within the molecular graph. Together, these techniques were evaluated using XGBoost regression models to predict molecular efficiency, using root mean square error (RMSE) and R^2 as the primary evaluation metrics.

Results. The results of our experiments demonstrate the effectiveness of our tokenisation algorithms, Bag of Smiles and ChemGrams, as well as the added value of molecular descriptors and graph-based features. The results are displayed in Table 1.

The Bag of Smiles achieved a Train RMSE of 8.21 and a Test RMSE of 13.057, with a Test R^2 of 0.768. By extending the Bag of Smiles with location action and count information, we observed an improvement in performance, achieving a Train RMSE of 7.71 and a Test RMSE of 13.255, with a Test R^2 of 0.761. These results indicate that the addition of token positions and counts contributes to capturing more relevant structural information.

The ChemGrams algorithm, which tokenises based on chemically meaningful fragments, yielded a Train RMSE of 8.499 and a Test RMSE of 13.363, with a Test R^2 of 0.757. Although this method did not outperform the Bag of Smiles variants, it still provided valuable insights into the role of functional group interactions in molecular efficiency. The addition of molecular descriptors resulted in further improvements, with a Train RMSE of 7.85, a Test RMSE of 13.068, and a Test R^2 of 0.768, showing the benefit of combining fragment-based tokenisation with molecular property data.

Table 1 – Performance Comparison of Tokenization Techniques in Predicting Corrosion Inhibitor Efficiency

Technique	Train RMSE	Train MAE	Train R2	Test RMSE	Test MAE	Test R2
ChemBert	7.241	4.41	0.924	13.522	8.837	0.752
Bag of Smiles (Ours)	8.21	5.324	0.903	13.057	8.966	0.768
Smiles2Vec	8.853	5.661	0.887	13.807	9.59	0.741
Byte-Pair Encoding (BPE)	7.241	4.41	0.924	13.522	8.837	0.752
Word-Level Tokenizer	8.377	5.427	0.899	13.347	9.23	0.758
SentencePiece	8.056	5.151	0.906	13.117	9.055	0.766
ChemGrams	8.499	5.456	0.896	13.363	9.297	0.757
Bag of Smiles (location action)	8.396	5.503	0.898	13.328	9.24	0.759
Bag Of Smiles (Location + count)	7.71	4.81	0.913	13.255	9.023	0.761
Bag Of Smiles (Location + count) added Molecular Descriptors	7.85	4.897	0.911	13.068	8.929	0.768

Overall, the highest performance was observed when both tokenisation methods were combined with molecular descriptors. The models using Bag of Smiles (location action and count) and molecular descriptors were able to achieve a Test R^2 of 0.768, indicating that a hybrid approach, incorporating multiple molecular representations, can lead to improved prediction accuracy. These results highlight the potential of our tokenisation methods for molecular efficiency prediction tasks and suggest that further improvements could be made by exploring additional tokenisation strategies.

Conclusion. This study introduced Bag of Smiles and ChemGrams, two novel tokenisation methods for predicting corrosion inhibitor efficiency, which were applied to data extracted from academic papers. These algorithms outperformed or matched traditional models like ChemBERT and Smiles2Vec by capturing both positional and count-based features alongside functional group data. Adding molecular descriptors further enhanced the prediction accuracy. While Bag of Smiles (Location + Count) and ChemGrams showed notable improvements, future work can explore graph-based representations and more advanced neural networks. This research contributes valuable insights to chemical informatics, accelerating the discovery of effective inhibitors.

REFERENCES

1. Galvão, T.L.P. CORDATA: an open data management web application to select corrosion inhibitors / T.L.P. Galvão, I. Ferreira, A. Kuznetsova, G. Novell-Leruth, C. Song, C. Feiler, S.V. Lamaka, C. Rocha, F. Maia,

M.L. Zheludkevich, J.R.B. Gomes, J. Tedim // Npj Materials Degradation. – 2022. – Vol. 6. – № 1. – Article number 48. https://doi.org/10.1038/s41529-022-00259-9.

2. Davronov, R. BERT-based drug structure presentation: A comparison of tokenisers / R. Davronov, F. Adilova // AIP Conference Proceedings. – 2023. – Vol. 2781. – № 1. Article number 020039. https://doi.org/10.1063/5.0144799.

3. Bostrom, K. Byte Pair Encoding is Suboptimal for Language Model Pretraining / K. Bostrom, G. Durrett // arXiv (Cornell University). – 2020. – https://doi.org/10.48550/arxiv.2004.03720.

4. Jo, J. The message passing neural networks for chemical property prediction on SMILES / J. Jo, B. Kwak, H.-S. Choi, S. Yoon // Methods. – 2020. – Vol. 179. – P. 65-72. – https://doi.org/10.1016/j.ymeth.2020.05.009.

УДК 665.6

Таранова Л.В.

(ФГБОУ ВО Тюменский индустриальный университет)

ОПТИМИЗАЦИЯ СИСТЕМ ТЕПЛООБМЕНА В ПРОЦЕССАХ ПЕРЕРАБОТКИ УГЛЕВОДОРОДНОГО СЫРЬЯ

Процессы переработки нефтегазового сырья, как известно, относятся к высокозатратным с точки зрения энергопотребления технологическим процессам. Это связано с необходимостью перемещения потоков в пределах производственной установки с созданием требуемых давлений и преодолением возникающих в аппаратах и трубопроводной сети гидравлических сопротивлений, а также с требованием создания соответствующих температурных режимов проведения процессов в единицах технологического оборудования.

Последний аспект связан с весьма широким диапазоном варьирования температур реализации отраслевых процессов и необходимостью применения разнообразного оборудования для тепловых процессов в любых процессах подготовки и различных этапов переработки углеводородного сырья на стадиях первичной и глубокой переработки и нефтегазохимии.

С учетом особенностей осуществления отраслевых процессов при их организации по непрерывной технологии и параметрами их проведения объективно возникает необходимость решения задач повышения энергоэффективности, реализации энерго- и ресурсосберегающих подходов, сокращения доли нерационально используемого тепла и его потерь.