Universidade de São Paulo
Instituto de Física de São Carlos

Semana Integrada do Instituto de Física de São Carlos

13ª edição

Livro de Resumos

São Carlos
2023

Ficha catalográfica elaborada pelo Serviço de Informação do IFSC

Semana Integrada do Instituto de Física de São Carlos
(13: 21-25 ago.: 2023: São Carlos, SP.)
Livro de resumos da XIII Semana Integrada do Instituto de Física de São Carlos – Universidade de São Paulo / Organizado por Adonai Hilário da Silva [et al.]. São Carlos: IFSC, 2023.
358p.

Texto em português.
1.Física. I. Silva, Adonai Hilário da, org. II. Título.

ISSN: 2965-7679

PG84

Deep Variational Anomaly Generation: An Approach to Testing Molecular Representation Robustness

NOGUEIRA, Victor[1]; SHARMA, Rishabh[2]; KEISER, Michael[2]; GUIDO, Rafael Victorio Carvalho[1]

victor.nogueira@usp.br

[1]Instituto de Física de São Carlos - USP; [2]University of California - UC

Real-world datasets in various domains, ranging from telecommunications to healthcare, often contain anomalous or outlier data that deviate significantly from the norm. Before applying modeling techniques, it is essential to filter out these anomalies to ensure data quality. This requirement has led to the development of robust anomaly detection models that can be deployed for data cleaning or to raise alarms in dynamic information processing systems, such as browsing, spam detection, or credit card fraud detection. However, there are cases where anomalies are the focus of investigation, shifting the attention from anomaly detection to anomaly generation. In certain domains, anomaly detection models face limitations due to the scarcity of training data, which hinders their predictive potential. In such situations, generating anomalies to populate synthetic training datasets has emerged as a promising approach to address data scarcity. (1) To tackle this challenge, it is crucial to investigate advanced methods for testing the robustness of molecular representations. In this context, we highlight the need for exploring advanced representational robustness testing methods in conjunction with the progress made in maximizing molecular representation robustness. We propose leveraging deep learning techniques, specifically variational autoencoders (VAE), to generate anomalies in a recently developed molecular string representation called SELF-referencIng Embedded Strings (SELFIES). (2-3) The objective was to test the robustness of the SELFIES representation, which assumes 100% validity when converting to SMILES notation. Through the exploration of a hyper-spherical latent space, we demonstrated that a VAE trained on SELFIES representations can generate molecules that violate the assumption of validity, surpassing a set of null models in the same task. We propose the VAE and the associated anomaly generation methodology as an effective means of testing the robustness of molecular representations. Furthermore, we discuss potential sources of invalidity in the SELFIES representation (latest version 2.1.1) and suggest validating modifications to address these issues. This discussion aims to invite further discourse on SELFIES and molecular string representations, fostering continuous improvement and development in the field.

Palavras-chave: Variational Auto-Encoder. SELFIES. Anomaly generation.

Agência de fomento: CAPES (88887.357974/2019-00)

Referências:

1 LAPTEV, N. Anogen: deep anomaly generator. **Meta Research**. 2018. Disponível em: https://research.facebook.com/file/969101687155819/AnoGen-Deep-Anomaly-Generator.pdf. Acesso em: 2023.

2 GÓMEZ-BOMBARELLI, R. *et al*. Automatic chemical design using a data-driven continuous representation of molecules. **ACS Central Science**, v. 4, n.2, p. 268–276, Feb. 2018.

3 KRENN, M. *et al*. A self-referencIng embedded strings (SELFIES): a 100% robust molecular string representation. **Machine Learning: Science and Technology**, v.1, n.4, p.045024-1-045024-8, 2020.