

Medium

In today's world, the advent of certain technologies has made problem-solving easier. For instance, computational chemistry has enabled rapid resolution of chemical issues, while mathematics and computer tasks have been streamlined. Chemists can simulate experimental outcomes and determine material properties. Additionally, there's a growing integration of Machine Learning (ML) concepts and algorithms in various fields.

The paper [‘Open-Source Machine Learning in Computational Chemistry’](#) reviewed 179 open-source software projects with corresponding peer-reviewed papers published within the last 5 years, to better understand what topics in this area are being explored by Machine Learning methods. For each project, a description was made, the code was assigned, the type of accompanying license was made, and whether training data and resulting models were available in the operation. Based on revenue data hosted in GitHub repositories, the most popular Python libraries have been identified.

The essence of this article lies in its exploration of machine learning algorithms within computational chemistry. Through this survey, the authors aimed to provide a comprehensive understanding of the current state of the field. Studies have shown that out of the 179 repositories tested, 94% of the cases identified the presence of data for training, but 54% of the cases identified the presence of data only for the use of models. This means that there is an obvious initiative that stimulates the development of their repository of the developed model, which includes the code, a list of necessary libraries, and the model parameters. One such example is listed in OpenML (<https://www.openml.org>; BSD-3), which was created by the Open Machine Learning Foundation.

Regarding licenses, the article showed that 78% of the projects surveyed included a license in their release, while 22% did not. The paper emphasized the connection between the model and the licensing of the associated code. Common licenses included GPL 2.0 (11%) and BSD-3 clause (8%), both adhering to Free and Open Source Software (FOSS) principles.

When coding is an important research component, the paper's processing, information acquisition, and robustness of its knowledge can be greatly enhanced by access to algorithms. Therefore, it is important to release code that is written concisely, because it is built and properly commented - in fact, to follow scientific practice.

The value of the findings is that the project has drawn attention to important details and strongly encourages researchers to go one step further and publish their optimized models. This will allow non-ML experts to use these models in their research and increase the reproducibility of the published data. In addition, the idea of creating a centralized online platform dedicated to the development of algorithms and datasets focused on computational chemistry can foster collaboration and accelerate progress in the field, which would benefit the entire scientific community.

If you are looking for Machine Learning solutions for your discovery research projects, [Chemspace](#) is the ideal partner. Our catalog provides access to ultra-large spaces of small molecules, allowing you to quickly find powerful hits already at the initial screening stage. This capability greatly simplifies the process and saves valuable time for the researcher.