ascenion Technology transfer for academic research
A company of the LifeScience Foundation

REFERENCE NUMBER TO 02-00412

ChemLM: a chemical LLM for molecular property prediction of experimental compounds

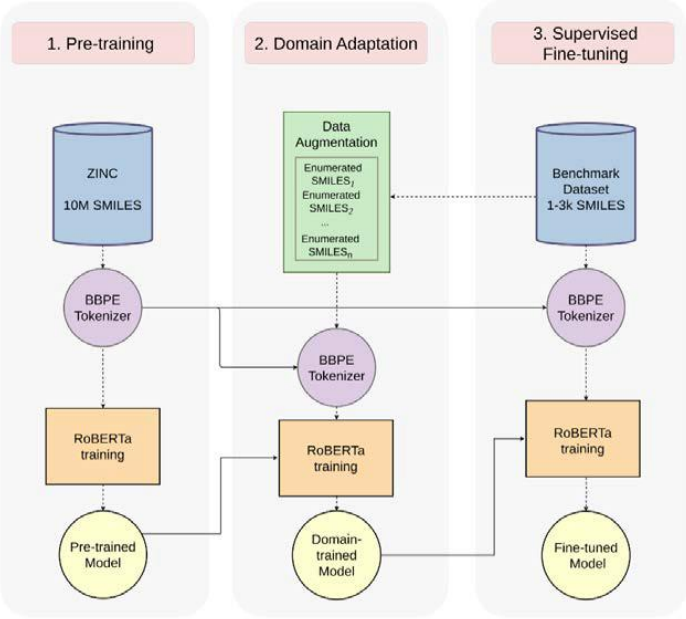
Computational chemistry, large chemical language models, molecular property prediction, deep learning

INVENTION NOVELTY

Large language models (LLMs) have the potential to revolutionize drug development. Building on the achievements of models such as ChatGPT, LLMs are now making strides in life sciences. Here, we introduce ChemLM, a powerful large language chemical model for accurate molecular property prediction empowering chemists by allowing them to efficiently prioritize or eliminate chemical structures in drug development. With our approach, chemical compounds are considered as sentences consisting of chemical 'words' of the molecules and processed using deep language methods. The model outperforms state-of-the-art methods on multiple benchmark datasets and on a real-world problem in molecular property prediction. The advantages of our approach stem from an additional training stage and data augmentation, wherein task-specific domain data is utilized to tailor the model to the specific task. This sets our methodology apart from conventional approaches that directly fine-tune models for the final task.

VALUE PROPOSITION

ChemLM will serve as a valuable tool in medicinal chemistry in the future. Particularly for datasets characterized by highly imbalanced data with a small percentage of positive samples, as common in drug development settings, it performs substantially better than other state-of-the-art models. For example, given a few hundred pathoblocker compounds designed for *Pseudomonas aeruginosa*, it accurately distinguished highly potent compounds (IC50 less than 500nM) from less potent ones. In this challenging scenario, ChemLM outperformed competing models by almost 30%. This underscores its capability to thrive in complex and practical applications within the field of molecular property prediction.



The training stages of ChemLM model. All the trained models are represented by circular shapes. Procedures like training, augmentation and prediction are indicated with rectangular shapes. Source: Kallergis G et al. doi:10.26434/chemrxiv-2023-cpkfk.

TECHNOLOGY DESCRIPTION

ChemLM uses SMILES, a string representation of molecules akin to a language as well as transformers, a deep neural architecture, highly successful in nature language processing. The method involves a series of steps that begin with pre-training the language model using a large number of chemical compounds teaching the model the syntax and grammar of the „SMILES language". In the next step, the trained language model is adapted to a specific target domain through self-supervised training by using domain-specific training data of the chemical compounds. Data augmentation is also used in this stage to boost model's performance. Finally, the domain-adapted model is trained using supervised training for a specific task, such as predicting the IC50 value of a chemical compound.

COMMERCIAL OPPORTUNITY

ChemLM is offered for licensing, creation of specialized models or customized predictions.

DEVELOPMENT STATUS

ChemLM is ready to perform molecular property prediction for candidate compounds with unknown characteristics.

PATENT SITUATION

European Patent Application was filed in August 2023.

FURTHER READING

Kallergis G, Asgari E, Azarkhalili B, Empting M, Hirsch A, McHardy A. Domain adaptable language modeling of chemical compounds identifies potent pathoblockers for Pseudomonas aeruginosa. ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-cpkfk

HZI HELMHOLTZ Zentrum für Infektionsforschung

Ascenion GmbH
Herzogstraße 64
D-80803 München
info@ascenion.de
www.ascenion.de

Licensing Contact
Dr. Petra Köhler
Technology Manager
T: +49 531 6181-2091
koehler@ascenion.de

Last updated February 2024