The Polymer Chemical Linguist: polyBERT's Role in Next-Generation Polymer Informatics

Christopher Kuenneth[1]
[1]Faculty of Engineering Science, University of Bayreuth, 94557 Bayreuth, Germany

christopher.kuenneth@uni-bayreuth.de

Polymer informatics has emerged as a critical field in materials science and chemistry, offering unprecedented opportunities for rapid identification and design of polymers tailored to specific applications. This study presents an end-to-end polymer informatics pipeline that advances the search for suitable polymer candidates with unparalleled speed and accuracy.

At the heart of our pipeline lies polyBERT, a large language model-based fingerprinting capability that functions as a "chemical linguist," interpreting polymer structures as a form of chemical language. This innovative approach demonstrates a remarkable two-order-of-magnitude speed improvement over traditional manually designed fingerprinting schemes while maintaining high accuracy. Such performance enhancements position polyBERT as an ideal candidate for deployment in scalable architectures, including cloud infrastructures.

Our pipeline also incorporates a data fusion learning approach to map polyBERT fingerprints to a diverse array of polymer properties. By employing an inverse design strategy, we successfully identify polymer candidates with specific desired properties. We demonstrate the practical applications of this approach through two case studies: the search for biodegradable alternatives to commodity plastics and the identification of redox-active polymers for battery applications.

The significance of this research extends beyond its immediate applications. By drastically accelerating the polymer discovery process, our pipeline has the potential to catalyze innovation across multiple industries, from sustainable materials to energy storage. Moreover, the adaptability of our approach suggests its potential applicability to other areas of cheminformatics.

References

[1] Chen, L., Pilania, G., Batra, R., Huan, T.D., Kim, C., Kuenneth, C., Ramprasad, R., Materials Science and Engineering: R: Reports, Volume 144, 100595 (2021)
[2] Audus, D.J., de Pablo, J.J, ACS Macro Lett. 6,10,1078-1082 (2017)
[3] Kuenneth, C., Lalonde, J., Marrone, B.L., Iverson, C.N., Ramprasad, R., Pilania, G., Commun Mater 3, 96 (2022)
[4] Kuenneth, C., Ramprasad, R., Nat Commun 14, 4099 (2023).

Figures

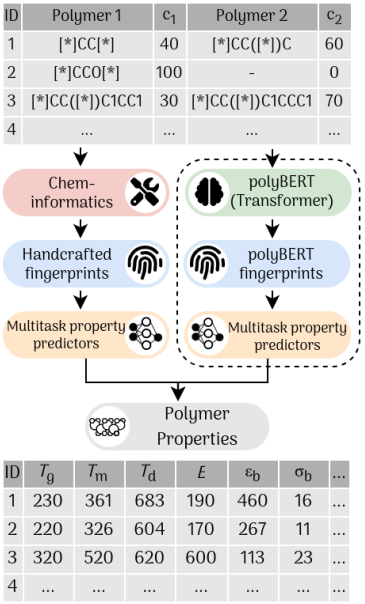

Figure 1. The traditional method (left) relies on handcrafted fingerprints, while the new approach (right) uses polyBERT for a fully end-to-end machine-driven prediction system.

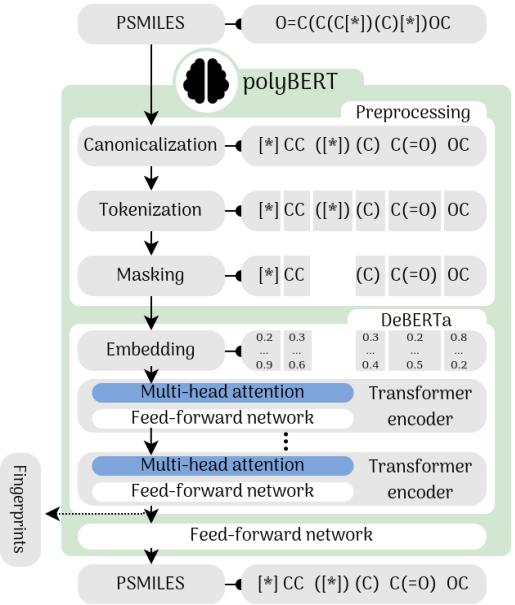

Figure 2. polyBERT acts as a "chemical linguist" for polymers. polyBERT canonicalizes, tokenizes, and masks Polymer Simplified Molecular-Input Line-Entry System (PSMILES) strings. This prepared data is then fed into the DeBERTa model, a powerful language processing system. The model utilizes 12 Transformer encoders, each equipped with 12 attention heads, to analyze the masked PSMILES strings. Finally, a dense layer with a softmax activation function uncovers the masked tokens. The key output of polyBERT, referred to as "fingerprints," are calculated by averaging the values (over the token dimension) from the final Transformer encoder.