Materials Informatics and Polymer Science: Pushing the Frontiers of our Understanding

Taylor D. Sparks[1] and Debanshu Banerjee[2]

[1]Materials Science & Engineering Department, University of Utah, Salt Lake City, UT, 84112, USA

[2]Metallurgical and Material Engineering Department, Jadavpur University, Kolkata, West Bengal, 700032, India

Summary: Humankind's unparalleled access to computing, data, and materials resources has resulted in the discovery of new, technology enabling materials. This work is a preview of a recent publication wherein predictive and generative machine learning models are applied on the largest polymer dataset to correlate chemical structure with glass transition temperature thereby leading to the discovery of new high-temperature polymers.

In the introduction to Stephen L. Sass' classic book "The Substance of Civilization" we read that "Materials not only affect the destinies of nations but define the periods within which they rise and fall. Materials and the story of human civilization are intertwined as the naming of eras after materials – the Stone Age, the Bronze Age, the Iron Age – reminds us." [1] This statement begs the obvious question of "What age do we currently live in?" Sass himself argues that we are living in the Silicon Age impacted pre-eminently by computer hardware. Others have suggested that we live in the Information Age with data itself as the key innovation of the 21st century. The Economist went so far as to publish a story entitled "The world's most valuable resource is no longer oil, but data". [2] Still others contend that we are currently living in the Plastic Age. [3]

It's certainly easy to argue for the merits of each of these three possibilities! Transistors are now fabricated at the nm length scale enabling radical computer hardware miniaturization. The internet of things means that more and more of the devices that we interact with are in a constant state of data connectivity and data collection. Meanwhile, synthetic polymers, the youngest of the traditional materials categories has burst on the scene in only the last century but has managed to infiltrate every aspect of our lives. With their nearly infinite range of structures and compositions, polymers are certainly among the most versatile of materials.

After reading the recent paper by Tao, Chen, and Li entitled "Machine learning discovery of high-temperature polymers" [4] we are of the opinion that we needn't choose between silicon, information, or polymers when defining our age since the most exciting innovations come at the intersection of these three! In the last decade we have seen materials informatics, or the application of data science to materials research, grow from relative obscurity to a formidable and well-respected technique driving the discovery of new materials. The field has had particular success among metal alloy and inorganic compound development. The first 3D printable aluminium alloy and new superhard materials are great examples. [5,6] However, machine learning development of new polymers has been slower in comparison.

This is not for lack of utility in discovering new polymers but the lack of data! Polymers having high strength to weight ratio – particularly those that exhibit high-temperature durability, high thermal decomposition temperatures or high glass transition temperatures – would have ample applications in industry. The development of such polymers began in the late 1950s primarily to satisfy the needs of aerospace and electronics industry. The following decades have witnessed the discovery and commercialization these polymers due to their low cost, excellent processability and moderate mechanical properties. Polymers like polytetrafluoroethylene (PTFE), perfluoro alkoxy alkanes (PFA), polyether ether ketone (PEEK) and fluorinated ethylene propylene (FEP) fall under the aforesaid category and find a wide range of applications due to their low density and high specific strength. However, the exploration of the molecular engineering of such polymers were limited to experimental trial and error strategies and in some cases, such as in the unintentional discovery of PTFE, serendipity! These Edisonian approaches are time and cost consuming, biased towards known chemical groups, and all too often fail to result in promising new materials. The time is right for a data-driven revolution in high-temperature polymer discovery.

The glass transition temperature, T_g is defined as the range of temperatures over which an amorphous material exhibits a gradual and reversible transformation from a hard and brittle (glassy) state to a rubbery (viscous) state. To develop robust and high-throughput screening methods for designing high-temperature polymers, researchers have established empirical relations where T_g is be expressed as functions of relative rigidities of chain backbone and side groups, repeating units of polymers chains, atomic mobilities etc. But most of these relations are applicable primarily to previously investigated polymer structures and tend to fail when extrapolated into new materials. To date there is not yet a universal model that can connect a polymer's T_g with its repeating units and molecular structure. This has led to the development of molecular dynamics (MD), i.e., molecular simulation and high-performance computing to simulate this important property. However, this too has limitations due to the high computational cost involved.

On the other hand, the barriers for data-driven techniques are being lowered continually with the growth of polymer databases. Indeed, data-driven approaches like quantitative structure property relationship (QSPR) and machine learning (ML) have emerged as effective methods in correlating molecular structure with a polymer's T_g. In these approaches a large array of molecular descriptors is extracted from the polymer's repeating unit. These are then trained using multi-step linear regression, neural networks, support vector regression (SVR) etc. resulting in a good match between predicted and experimental values of T_g. QSPR however, becomes time consuming where density functional theory (DFT) calculations are involved and it is difficult to give physical interpretation to the parameters generated through this process. While ML recent models have shown promise, they have also been very limited in their generalizability due to small and relatively confined ranges in chemistry in the training data sets. Furthermore, it's not clear which of the many descriptor tools should be implemented to be represent the structure for the models.

The new editorial by Tao et al. focuses in overcoming the above challenges. The largest polymer dataset, PoLyInfo, has been considered in this work which comprises about 13,000 homopolymers. This dataset is divided into two parts, dataset-1 and dataset-2. Dataset-1 consists of labelled data (6923 polymers with their experimentally measured T_g values). The authors compare three types of feature representations that have been considered based on the SMILES notation of each polymer: molecular descriptors, Morgan fingerprints and images. The authors then use these representations to construct four different machine learning models: least absolute shrinkage and selection operator (Lasso), deep neural networks (DNN) and convolutional neural networks (CNN). In addition to discussing the advantages and disadvantages of these models and descriptors on the labelled dataset, they then carry out an important analysis of these models on unlabelled data to probe generalizability to new data outside of the training data. This is done not only on 5690 real polymers (dataset-2) from the PoLyInfo database, but also on an additional 1 million hypothetical polymers generated by a generative machine learning model.

A particularly captivating result from this paper is that the authors use their machine models to predict new high-temperature polymers. Across the various datasets, the authors demonstrate an ability to discover thousands of new candidate materials for future evaluation. This approach shows promise for identifying polymers with $T_g \geq 400$ °C particularly since subsequent MD simulations showed a good agreement with the T_g values predicted by the ML models.

It's noteworthy to point out that the authors did not simply build a black-box tool for recommending new polymer chemistries. Instead, they also focused on interpretability by analysing feature weighting and posterior analysis of chemical descriptors. Their findings reinforce and supplement existing models to offer a better understanding of chemical insight related to glass transition temperature.

This paper demonstrates the ability for machine learning models to generate new compounds and new understanding in the traditionally challenging field of high-temperature polymers by leveraging growing databases of experimental and computational data alongside predictive and generative machine learning models. This latest example of materials informatics portends a bright future for humankind as we leverage silicon, information, and polymers to address the technological challenges of the 21st century.

Declaration of Interests: The authors declare no competing interests.

[1] Sass, Stephen L. The substance of civilization: Materials and human history from the stone age to the age of silicon. Arcade Publishing, 1998.

[2] David Parkins "The world's most valuable resource is no longer oil, but data" The Economist, May 6th 2017, retrieved at https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data on March 22 2021.

[3] Danny Lewis "Are We Living in the Plastic Age?" Smithsonian Magazine, January 22nd 2016, retrieved at https://www.smithsonianmag.com/smart-news/are-we-living-plastic-age-180957817/ on March 22 2021.

[4] Lei Tao, Guang Chen, and Ying Li "Machine learning discovery of high-temperature polymers" Patterns (2021): 100225.

[5] Martin, John H., et al. "3D printing of high-strength aluminium alloys." Nature 549.7672 (2017): 365-369.

[6] Mansouri Tehrani, Aria, et al. "Machine learning directed search for ultraincompressible, superhard materials." Journal of the American Chemical Society 140.31 (2018): 9844-9853.