JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Data-Driven Approaches in Polymer Research

Dr. Ritesh Kumar

Assistant Professor, Department of Physics

Maulana Azad National Urdu University, Polytechnic, Bangalore – 560072

Headquarters: Maulana Azad National Urdu University (A Central University),

Hyderabad 500032

manuuphy2020@gmail.com

Abstract

Data-driven polymer research is reshaping how macromolecules are conceived, synthesized, and deployed by linking curated, FAIR data to physics-informed machine learning and autonomous experimentation. Computational approaches are now an important part of polymer chemistry. They let scientists guess a material's bulk properties based on its molecular structure. The goal is to figure out which structural elements control certain behaviours so that new materials can be designed in a logical way. But this procedure is made more difficult by a number of real-world problems, such as short datasets, changing molecular weights, and complicated polymer topologies. As a result, generic computational models are typically not very useful unless they are heavily customised. To get over these problems, chemists and data scientists need to work together in a way that helps both groups. The chemist's job is to help define the problem by using their knowledge of the field, and the data scientist's job is to make models work in the unique chemical context. This research seeks to bridge the multidisciplinary divide by tackling issues related to data quality and emphasising how recent scientific progress has enhanced data accessibility. We look at how materials are used in the field, from anticipating how well they will work to designing them for drug delivery. Finally, we talk about how important it is to share data (FAIR Data Principles) and how powerful it is to combine conventional polymer theory with new, data-driven insights.

Keywords: Data Scientists, FAIR Principles, Polymer Chemistry, Data, Drug Delivery.

Introduction

A significant transformation is underway in materials science, particularly in the polymer field, where the pace of discovery is accelerating. This change is driven by the adoption of data-driven informatics approaches [1,2], marking a shift towards research that is more predictive, efficient, and collaborative. The need for this shift is rooted in the immense complexity of polymers themselves. Despite being built from basic elements, polymers exhibit a vast range of structural variations at both microscopic and macroscopic levels[3]. These differences arise from atomic-level connectivity, the way chains pack together, and diverse morphological features like crystallinity, phase separation, porosity, and microstructure[4,5].

This enormous spectrum of chemical and structural possibilities poses a significant challenge. Identifying a polymer with properties tailored for a specific use requires exploring a vast design space[6]. Traditional experimental and computational methods are often inefficient for navigating such a complex, high-dimensional

domain. Moreover, the sheer volume of new knowledge in the polymer field makes it increasingly difficult to effectively harness existing information for future breakthroughs.

The application of data-driven methods to polymer design has led to the emergence of "polymer informatics" [7,8]. This interdisciplinary field merges polymer science with data science and machine learning, creating powerful new tools to both develop novel polymers and better understand the properties of existing materials.

The foundational topics of polymer informatics are already well-covered in existing literature. Numerous reviews explain the typical machine learning workflow [8-10,11,12-14].

2. Fundamentals of Machine Learning in Polymer Science

2.1. Data

2.1.1. Challenges in Data Collection and Cleaning in Polymer Informatics

Once a research problem in polymer science has been framed for a machine learning approach, the crucial next step in the pipeline is to gather and curate the necessary data. In today's era of big data, one might assume that high-quality datasets are readily available, but this is rarely the case in polymer informatics. The difficulty arises from the intrinsic nature of polymers themselves. They are stochastic and hierarchical materials, and their final morphology is highly sensitive to the specific conditions under which they were prepared, making standardized data collection exceptionally challenging.

A suite of polymer-focused databases—PoLyInfo [15], PI1M [16], Khazana [17], CROW [18], PubChem [19], and CAMPUS [20]—has pushed the field forward. Each resource is assembled through distinct curation workflows and therefore captures different kinds of information about monomers and their resulting polymers. Table 1 contrasts these databases along those dimensions. Efforts such as the Materials Project in the U.S. and Europe's Novel Materials Discovery Laboratory have also reshaped how new functional polymers are found, pivoting from conventional bench experiments to data-driven discovery that leverages these extensive repositories.

Even with numerous materials databases available, building a dataset that truly fits a specific research question remains difficult. Challenges include access restrictions and the relatively modest scale of polymer databases compared with major small-molecule repositories like ChemSpider [21], PubChem [22], and ChEMBL [23,24], which limits their utility in polymer informatics. The community's call for FAIR data [25]—findable, accessible, interoperable, and reusable—highlights the need for well-structured, shareable datasets, yet broadly adopted mechanisms to implement FAIR principles at scale are still lacking.

ML in Polymer Research

Machine learning is a groundbreaking computational approach, providing a range of algorithms designed to derive actionable insights from intricate, high-dimensional datasets. The main approaches involve supervised learning, which is highly effective at building predictive models by linking inputs (e.g., molecular descriptors) to established outputs (e.g., material properties) derived from labelled training data. Unsupervised learning works with unlabelled data to reveal underlying structures, allowing for the independent clustering of materials and the discovery of concealed correlations within extensive chemical landscapes. In addition to these, reinforcement learning offers a robust framework for enhancing sequential decision-making processes, demonstrating its worth in tasks like navigating synthetic pathways or managing autonomous experimental systems.

In the field of polymer research, the use of these ML frameworks is driving a transition from chance discoveries to systematic, purpose-driven design. The main focus is on creating surrogate models that swiftly and precisely forecast intricate polymer properties—like glass transition temperature, tensile strength, or ion conductivity—straight from their chemical structure, thus avoiding the laborious experimental synthesis and characterisation process. This advanced capability allows for extensive virtual screening of vast candidate

libraries. More profoundly, machine learning enables the concept of inverse design, allowing algorithms to explore the extensive chemical landscape to create innovative macromolecular structures specifically designed to satisfy distinct, frequently conflicting, performance requirements. This strategy is fast-tracking the creation of cutting-edge polymers for essential uses, such as eco-friendly packaging, high-performance composites, and innovative biomedical devices.

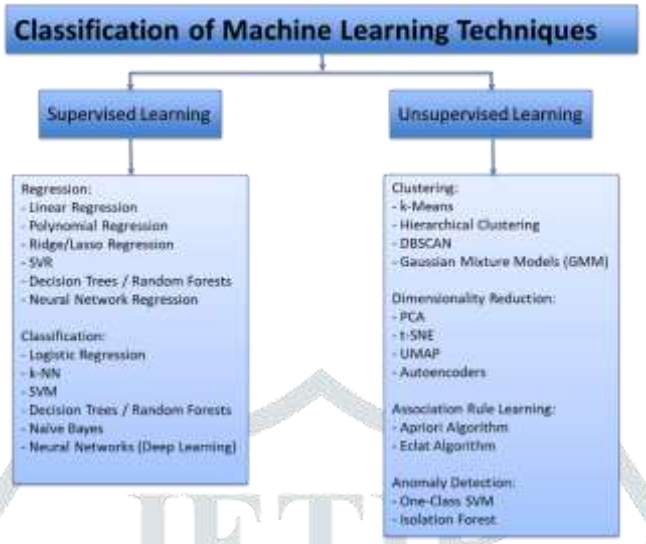

Figure 1. Classification of Machine learning

Table 1 Comparison of different databases containing information on polymers.

Database	Data Collection Method	Chemical Structures and Composition	Property Data
CAMPUS Plastics	Data supplied directly by raw material **manufacturers**. All testing follows strict, uniform ISO standards.	Describes specific, commercial grades of polymers, often including fillers and additives.	High-quality, comparable **experimental data** for engineering properties (mechanical, thermal, electrical).
CROW	Manual curation from scientific literature, handbooks, and historical sources.	A comprehensive roster of polymer names, trade names, and basic structural representations.	Contains basic physical properties (e.g., Tg, Tm); less focused on detailed engineering or performance data.
Khazana	Data generated from first-principles **computational methods** (like Density Functional Theory - DFT).	A large library of polymer repeating units represented in a format suitable for computation.	Repository of **computationally predicted** properties (e.g., band gap, dielectric constant) for training ML models.
PI1M	Algorithmic generation of polymer structures followed by high-throughput property **prediction** using ML models.	A vast library of one million hypothetical polymer structures, typically represented by SMILES.	**Predicted properties** for a massive number of structures, designed to screen a broad chemical space for discovery.
PoLyInfo	Manual curation and expert extraction from peer-reviewed scientific literature.	Detailed representations including repeating units, monomers, tacticity, and polymerization methods.	Extensive **experimental data**: mechanical, thermal, electrical, optical, and physical properties.
Polymer Genome	Primarily **computational** (e.g., DFT simulations) and high-	Focuses on representations of repeating units suitable for	Mainly **computationally predicted** properties,

Database	Data Collection Method	Chemical Structures and Composition	Property Data
	throughput virtual screening.	computational modeling.	with a strong focus on dielectric and electronic properties.
PubChem	Aggregation from various sources including literature, patents, and data contributors (e.g., vendors).	Stores polymers using standard chemical identifiers (SMILES, InChI). Polymer-specific details can be limited.	Broad range of data, including physical properties, bioactivity, and safety. Can be sparse for specific material properties.
SciFinder^{n} / CAS REGISTRY	Expert human curation and indexing of global scientific literature, patents, and technical reports.	Highly structured data including monomer composition, end-groups, and cross-linking via CAS Registry Numbers®.	Vast collection of **experimentally measured** properties linked directly to the literature source.

Application of Machine Learning in Polymer Science

Machine learning now spans virtually all areas of polymer science, from informing synthesis to improving end-use performance [26]. Here, we emphasize the prediction targets rather than the specific algorithms, concentrating on polymerization reaction rates and on polymer properties in both the solid state and in solution. In solids, common targets include the glass transition temperature (Tg), specific heat capacity (cp), decomposition temperature, mechanical performance, gas solubility, and electrical behavior.

Table 2. Comparison of different machine learning models.

Model	Learning Task	Common Applications	Pros	Cons
Linear / Logistic Regression	Regression / Classification	- Predicting house prices, sales forecasts.- Spam detection, disease diagnosis, churn prediction.	- Fast, simple, and highly interpretable.- Good baseline model.- Low computational cost.	- Assumes a linear relationship between features.- Prone to underfitting complex data.- Sensitive to outliers.
K-Nearest Neighbors (KNN)	Classification & Regression	- Recommendation systems.- Image recognition.- Anomaly detection.	- Very simple to understand and implement.- No training phase ("lazy learning").- Effective for data with non-linear boundaries.	- Computationally expensive during prediction.- Performance degrades with high-dimensional data.- Requires feature scaling and is sensitive to the choice of 'K'.
Support Vector Machine (SVM)	Classification & Regression	- Image classification (e.g., facial recognition).- Text classification.- Bioinformatics.	- Effective in high-dimensional spaces.- Memory efficient (uses a subset of training points).- "Kernel trick" allows for modeling non-linear data.	- Can be slow to train on very large datasets.- Choosing the right kernel and hyperparameters is difficult.- Less interpretable than other models.
Decision Tree	Classification & Regression	- Credit scoring.- Simple medical diagnosis.- Customer segmentation.	- Highly interpretable and easy to visualize.- Requires little data preprocessing.- Can handle both numerical	- Very prone to overfitting.- Can be unstable (small data changes can alter the tree).- Can create biased

Model	Learning Task	Common Applications	Pros	Cons
			and categorical data.	trees if some classes dominate.
Random Forest	Classification & Regression	- Fraud detection.- Predicting stock prices or patient outcomes.- E-commerce product recommendations.	- High accuracy and robust to overfitting.- Handles missing values and maintains accuracy.- Can provide feature importance scores.	- Less interpretable than a single decision tree.- Can be slow to train and predict with many trees.- Requires more memory.
Gradient Boosting (XGBoost, LightGBM)	Classification & Regression	- Dominant model for tabular data competitions (Kaggle).- Search engine ranking.- Click-through rate prediction.	- Often achieves the highest performance on structured data.- Highly flexible and handles missing values well.- Many tunable parameters for optimization.	- Complex to tune and prone to overfitting if not careful.- Can be slow to train as it builds trees sequentially.- Can be considered a "black box" model.
K-Means Clustering	Unsupervised (Clustering)	- Customer segmentation for marketing.- Document clustering for topic analysis.- Image compression.	- Simple, fast, and scalable for large datasets.- Easy to implement and understand.	- Must specify the number of clusters (K) beforehand.- Sensitive to initial centroid placement.- Assumes clusters are spherical and of equal size.
Neural Networks (Deep Learning)	Classification, Regression, and more	- Image and speech recognition.- Natural Language Processing (NLP).- Self-driving cars.	- State-of-the-art performance on unstructured data.- Can learn features automatically from raw data.- Can model extremely complex relationships.	- Requires very large amounts of data.- Computationally expensive to train (often needs GPUs).- Highly complex and difficult to interpret ("black box").

Polymer Materials

Smart Polymer

Smart polymers are a type of sophisticated material that can change their properties when they come into contact with outside forces [27]. Shape-memory polymers are one of the most interesting types of polymers because they can return to a shape that was programmed into them when they are activated [28]. This makes them useful in biomedical devices and sensors [29]. Another big step forward is self-healing polymers, which can fix damage on their own. This is an important feature for making materials last longer [30]. Advanced sensing polymers can also detect and respond to changes in the environment, such strain. This makes them useful for monitoring the health of structures and wearable technology [31]. To make things safer, new flame-retardant polymers are being made with complex molecular structures that stop fires. These polymers are very important for the construction, electronics, and transportation industries [32]. Machine learning has become a significant tool for developing these sectors in the last several years. It is a more efficient way to make these functional materials than traditional lab-based methods [33,34-36].

Polymers as Biological Imaging Agents

^{19}F MRI is a valuable biomedical imaging technique that produces clear, high-contrast images, making it possible to monitor cellular transport and quantify oxygenation over time. For this purpose, synthetic polymers

are appealing as contrast agents due to their flexible design and their capacity for holding many ^{19}F atoms, which enhances the signal [37-40]. However, the field faces a significant hurdle. Even with numerous studies on copolymer-based agents, it remains difficult to create a material that successfully combines two essential properties: dissolving in water and carrying enough fluorine to be visible on standard hospital MRI equipment [41].

The study proposed using an automated machine learning (AutoML) approach to simplify model creation, allowing researchers who are not machine learning specialists to find high-performing models on their own [42]. This AutoML system was designed as a self-contained optimization process that operates in a continuous loop, automatically testing different methods and selecting the best parameters. It refines its models with each cycle, incorporating new data from both computer simulations and lab experiments to balance multiple performance goals at once. The system tested a range of established machine learning methods—including XGBoost [43], Neutral Gradient Boosting [44], GPR, and Random Forests from the scikit-learn library [45]—in a supervised learning framework. The inputs for these models were standardized numerical data representing the polymer's chemical composition and analytical measurements.

Fair Data Principles

The FAIR principles—Findable, Accessible, Interoperable, and Reusable—offer a vital set of standards for managing scientific data. Their real value comes from turning research data from forgotten, one-off files into a lasting resource for the entire scientific community. When data is easy to locate, use, and integrate, it strengthens both transparency and the ability to reproduce results. This approach not only prevents crucial findings from disappearing over time but also drives new discoveries by giving scientists the foundation to build upon previous work, saving time and encouraging teamwork between different fields.

The FAIR principles provide a framework for managing research data to maximize its value. The goal is to make data:

- **Findable:** Easy to discover with unique identifiers and clear descriptions.
- **Accessible:** Simple to retrieve using standard, well-defined procedures.
- **Interoperable:** Able to be combined and used with other data and systems.
- **Reusable:** Well-described and licensed so it can be used for future studies.

For research data to have real value, it must be stored in a way that other scientists can easily access and use it. The FAIR data principles formalise this idea by calling for adequate data stewardship to make it possible to reuse data [46]. In reality, FAIR says that data must have precise descriptions (Findable), clear rules for who may access it (Accessible), be kept in a standard format (Interoperable), and have clear usage licenses (Reusable).
There is a rising push to use this standard method. The best method to do this is to get research groups to send their results to centralised platforms where the data can be organised in the same way.

Conclusion

The present study provides a useful insight on machine learning with a specific emphasis on its applications in polymer science. It aims to demystify the underlying mathematics and detail the end-to-end machine learning process. In order to convert large amounts of diverse experimental data into the reliable prediction models required to speed up materials discovery, we stress the importance of a quantitative framework. In order to successfully work on interdisciplinary research at the cutting edge of materials science, experimentalists must be given the conceptual tools necessary to traverse this data-driven paradigm. Due to the inherent complexity of polymers—whose stochastic topologies and polydispersity make them difficult to represent digitally—the area

lacks a common representational standard. Additionally, whereas high-throughput experimentation is accelerating the velocity of data, a corresponding cultural shift towards the acceptance of large-scale data analysis has not yet fully taken shape. In order to fully realise the transformative potential of machine learning, the community must work together to embrace FAIR data principles and, more importantly, to integrate core polymer theory with contemporary data science techniques.

References

1. Lookman, T., Alexander, F. J., & Rajan, K. (2016). *Information science for materials discovery and design* (Vol. 225). Springer.
2. Basu, B., Gowtham, N. H., Xiao, Y., Kalidindi, S. R., & Leong, K. W. (2022). Data-driven approaches for the discovery and design of biomaterials. *Acta Biomaterialia, 143*, 1–27.
3. Sha, W., Li, Y., Tang, S., Tian, J., Zhao, Y., Guo, Y., Zhang, W., Zhang, X., Lu, S., Cao, Y. C., & Cheng, S. (2021). Machine learning in polymer informatics: Recent developments and future directions. *InfoMat, 3*(3), 353–377.
4. Ehrenstein, G. W. (2012). *Polymeric materials: Structure, properties, applications*. Carl Hanser Verlag.
5. Huan, T. D., Boggs, S., Teyssedre, G., Laurent, C., Cakmak, M., Kumar, S., & Ramprasad, R. (2016). Advanced polymeric dielectrics for high energy density applications. *Progress in Materials Science, 83*, 236–269.
6. Gartner, T. E., & Jayaraman, A. (2019). Machine learning in polymer science. *Macromolecules, 52*(3), 755–786.
7. Tran, H. D., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J. P., Gurnani, R., Shetty, P., Ramprasad, M., Laws, J., Shelton, M., & Ramprasad, R. (2020). The polymer genome project: A materials informatics platform for polymer dielectrics. *Journal of Applied Physics, 128*(17), 171104.
8. Chen, L., Pilania, G., Batra, R., Huan, T. D., Kim, C., Kuenneth, C., & Ramprasad, R. (2021). The materials science of machine learning. *Materials Science and Engineering: R: Reports, 144*, 100595.
9. Yan, C., & Li, G. (2023). Polymer informatics with multi-task learning. *Advanced Intelligent Systems, 5*(1), 2200243.
10. Xu, P., Chen, H., Li, M., & Lu, W. (2022). Inverse design of polymer dielectrics by combining machine learning and coarse-grained molecular dynamics simulation. *Advanced Theory and Simulations, 5*(2), 2100565.
11. Martin, T. B., & Audus, D. J. (2023). Polymer data and databases. *ACS Polymers Au, 3*(3), 239–252.
12. Hatakeyama-Sato, K. (2023). Materials informatics for polymer design: Present and future. *Polymer Journal, 55*(2), 117–128.
13. Meyer, T. A., Ramirez, C., Tamasi, M. J., & Gormley, A. J. (2023). Active learning in polymer science and engineering. *ACS Polymers Au, 3*(2), 141–164.
14. Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials, 5*, Article 83.
15. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011). Development of a data-driven materials design system for polymers. In *2011 International Conference on Emerging Intelligent Data and Web Technologies* (pp. 22–29). IEEE Computer Society.
16. Ma, R., & Luo, T. (2020). Machine learning for polymer design and property prediction. *Journal of Chemical Information and Modeling, 60*(10), 4684–4693.

17. Huan, T. D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G., & Ramprasad, R. (2016). A data-driven approach for the predictive, design and discovery of polymer dielectrics. *Scientific Data, 3*, Article 160012.

18. Chi, M., Gargouri, R., Schrader, T., Damak, K., Maâlej, R., & Sierka, M. (2022). Machine learning in polymer science: A case study of mechanical properties of polyamide 6 and its composites. *Polymers, 14*(1), Article 26.

19. Hu, B., Lin, A., & Brinson, L. C. (2021). Chemform, a chemical formula parsing tool for polymer informatics. *Journal of Cheminformatics, 13*, Article 22.

20. Lobato, H., Cernuda, C., Arriaga, A., Zulueta, K., Burgoa, A., & Coop, L. S. (2023). *Prediction of the mechanical behavior of a short fiber reinforced polymer based on a data-driven RVE* [Conference paper]. VII ECCOMAS Young Investigators Conference, Porto, Portugal.

21. Pence, H. E., & Williams, A. (2010). ChemSpider: An online chemical information resource. *Journal of Chemical Education, 87*(11), 1123–1124.

22. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021). PubChem 2021: New data content and improved web interfaces. *Nucleic Acids Research, 49*(D1), D1388–D1395.

23. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research, 43*(W1), W612–W620.

24. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research, 43*(W1), W612–W620.

25. Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Felix, E., Magarinos, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Maranon, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., ... Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research, 47*(D1), D930–D940.

26. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018.

27. McDonald, S. M., Augustine, E. K., Lanners, Q., Rudin, C., Brinson, L. C., & Becker, M. L. (2023). An interpretable machine learning model for the rational design of biodegradable polymers. *Nature Communications, 14*, Article 4838.

28. Bratek-Skicki, A. (2021). Protein adsorption on shape-memory polymers. *Applied Surface Science Advances, 4*, 100068.

29. Lendlein, A., & Kelch, S. (2002). Shape-memory polymers. *Angewandte Chemie International Edition, 41*(12), 2034–2057.

30. Pilate, F., Toncheva, A., Dubois, P., & Raquez, J.-M. (2016). Shape-memory polymers for multiple applications in the materials world. *European Polymer Journal, 80*, 268–294.

31. Wang, S., & Urban, M. W. (2020). Self-healing polymers. *Nature Reviews Materials, 5*(8), 562–582.

32. Yan, J., Chen, A., & Liu, S. (2024). Advances in self-healing polymers and composites: A review. *Alexandria Engineering Journal, 86*, 405–416.

33. Shen, J., Liang, J., Lin, X., Lin, H., Yu, J., & Wang, S. (2022). Progress of intrinsic self-healing polymers based on dynamic covalent bonds. *Polymers, 14*(1), Article 82.

34. Yan, C., Feng, X., Wick, C., Peters, A., & Li, G. (2021). Predicting glass transition temperature of polymers with the graph neural network. *Polymer, 214*, 123351.

35. Anwar Ali, H. P., Zhao, Z., Tan, Y. J., Yao, W., Li, Q., & Tee, B. C. K. (2022). Computationally guided design of mechanically robust and multistimuli-responsive self-healing iminoboronate-based elastomers. *ACS Applied Materials & Interfaces, 14*(46), 52486–52495.

36. Chen, F., Weng, L., Wang, J., Wu, P., Ma, D., Pan, F., & Ding, P. (2023). A self-healing and recyclable shape memory polymer based on dynamic acylhydrazone bond and disulfide bond. *Composites Science and Technology, 231*, 109818.

37. Srinivas, M., Heerschap, A., Ahrens, E. T., Figdor, C. G., & De Vries, I. J. M. (2010). ^{19}F MRI for quantitative in vivo cell tracking. *Trends in Biotechnology, 28*(7), 363–372.

38. Janjic, J. M., & Ahrens, E. T. (2009). Fluorine-containing nanoemulsions for quantitative ^{19}F MRI/MRS of cell migration in vivo. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology, 1*(5), 492–501.

39. Chapelin, F., Capitini, C. M., & Ahrens, E. T. (2018). Fluorine-19 magnetic resonance imaging for detection and quantification of CAR T cells in vivo. *Journal for ImmunoTherapy of Cancer, 6*, Article 105.

40. Taylor, N. G., Chung, S. H., Kwansa, A. L., Johnson III, R. R., Teator, A. J., Milliken, N. J., Koshlap, K. M., Yingling, Y. G., Lee, Y. Z., & Leibfarth, F. A. (2020). A data-driven approach to creating a library of fluorinated polymers for use as ^{19}F MRI agents. *Chemistry–A European Journal, 26*(44), 9982–9988.

41. Tirotta, I., Dichiarante, V., Pigliacelli, C., Cavallo, G., Terraneo, G., Bombelli, F. B., Metrangolo, P., & Resnati, G. (2015). Halogen-bonded supramolecular materials. *Chemical Reviews, 115*(3), 1106–1129.

42. Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

43. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

44. Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020). *AutoGluon-tabular: Robust and accurate AutoML for structured data*. ArXiv. https://arxiv.org/abs/2003.06505

45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.