CHALLENGES AND VISION FOR STANDARDIZATION OF BIOPOLYMER DATASETS FOR MACHINE LEARNING

Jessica N. Lalonde[1], **Defne Circi**[2], **Babetta L. Marrone**[1], **Stefan Zauscher**[2], **L. Catherine Brinson**[2]
[1]Bioscience Division, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM, USA
[2]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA

ABSTRACT

Machine learning (ML) is transforming materials research, yet potential for biopolymer discovery remains constrained by fragmented data and non-standardized reporting. Biopolymers differ significantly from synthetic polymers, requiring specialized approaches to represent their biosynthetic origins, hierarchical structures, and application-specific metrics. In this perspective, we identify three core challenges limiting biopolymer representation: information encoding, data quality, and data sharing. Unlike prior reviews on polymer informatics, this perspective explicitly focuses on biopolymer-specific challenges arising from biosynthetic variability, hierarchical structure, and environmental sensitivity, and outlines interoperable, ML-ready solutions tailored to these three key challenges. Recommendations include the design and adoption of biopolymer-specific fingerprinting frameworks, the development of hybrid data extraction strategies, and the expansion of Findable, Accessible, Interoperable, Reusable (FAIR)-compliant repositories. We propose a robust foundation to define interoperable, high-quality datasets that capture the full context of biopolymer materials. Standardized metadata, shared ontologies, and community-driven infrastructure will enable scalable, reproducible workflows and accelerate the ML-driven development of biopolymers.

1 INTRODUCTION

Machine learning (ML) is increasingly central to materials discovery, enabling scalable structure–property modeling, accelerated screening, and data-driven optimization across polymer systems Himanen et al. (2020); Mannodi-Kanakkithodi & Chan (2021); Lalonde et al. (2025); Artrith et al. (2021); Butler et al. (2018). While synthetic polymer informatics has advanced rapidly through standardized representations, curated repositories, and ML-ready datasets, Gurnani et al. (2023); Tran et al. (2024); Phan et al. (2024); Toland et al. (2023) analogous progress for biopolymers remains limited Kuenneth et al. (2022); Roumeli et al. (2025); Malashin et al. (2024). Biopolymers are a chemically diverse group of biomolecules ranging in size from 3 to 14+ linear carbon chains, to proteins with hundreds of amino acid residues and complex hierarchical structures, to polysaccharide chains that can be thousands of sugar units in length. They may be produced biologically or from biological sources, with examples such as polyhydroxyalkanoates (PHAs), polylactic acid (PLA), starch, silk, and chitin. These materials exhibit environmental sensitivity and process-dependent variability that are poorly captured by existing polymer data frameworks Fransen et al. (2023); Mateu-Sanz et al. (2024); Lin et al. (2024); Lim et al. (2017). As a result, biopolymer data are often fragmented, inconsistently encoded, and difficult to integrate into modern ML workflows.

In principle, ML-driven biopolymer discovery follows the same end-to-end pipeline used in other materials domains, from data assembly and encoding to model training and validation (Figure 1 A) McDonald et al. (2023); Kerner et al. (2021b); Gianti & Percec (2022). In practice, however, dataset preparation and standardization remain the dominant bottlenecks McDonald et al. (2023); Roumeli et al. (2025); Kerner et al. (2021b); Gianti & Percec (2022). In this perspective, we identify three core challenge areas that currently limit scalable ML for biopolymers from a materials perspective, which we list here as information encoding, data quality, and data sharing. We then outline practical, community-oriented recommendations for each (Figure 1 B). Rather than proposing a single

solution, we focus on articulating where current approaches break down and where specific interoperable advances could enable ML-ready biopolymer datasets. Further technical discussions are provided in the included Appendix.

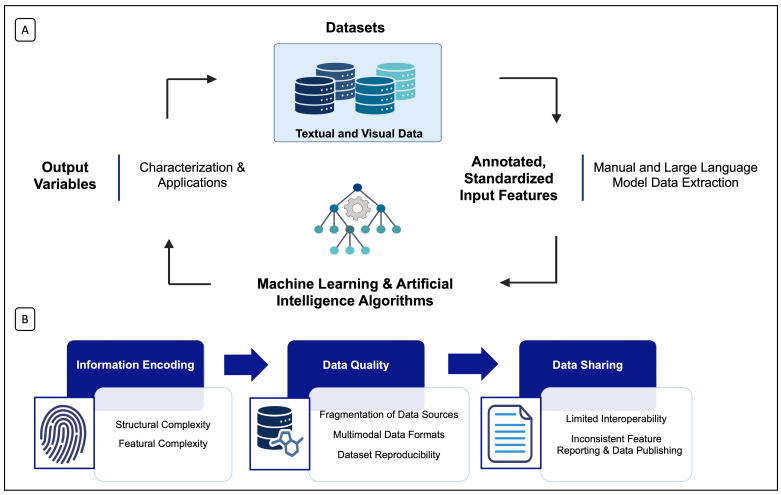

Figure 1: (A) An idealized ML/AI workflow for biopolymer materials, in which data from the literature and repositories are annotated and aggregated into structured datasets. Standardized preprocessing and information encoding creature input features for ML/AI models that in turn generate predictions, classifications, and property evaluations. The resulting outputs can then be re-integrated into future datasets, forming a cyclical data flow that promotes continuous improvement in data quality and reproducibility. (B) Key barriers to this workflow, highlighting biopolymer-specific challenges in information encoding, data quality, and data sharing.

2 Information Encoding

Biopolymers present encoding challenges distinct from synthetic polymers due to biosynthesis-dependent variability, hierarchical organization, and sensitivity to processing and environmental conditions Bejagam et al. (2022); Koller & Mukherjee (2022); Popa et al. (2022); McAdam et al. (2020). These factors complicate representation of structure, processing history, and application-relevant measurements for ML. As a result, structurally distinct samples are often reduced to simplified representations, and critical contextual information is lost during dataset construction Altamira-Algarra et al. (2025); Anjum et al. (2016); Sabapathy et al. (2020); Urtuvia et al. (2014); Wang et al. (2024); Surendran et al. (2020); Utsunomia et al. (2020); Knoop et al. (2010; 2013). A key limitation is the inconsistent treatment of context. Measured properties depend on biosynthesis conditions, extraction methods, processing history, and application environments, yet these variables are rarely encoded in standardized, machine-readable formats Pilania et al. (2019); Bejagam et al. (2021); Fatriansyah et al. (2024); Martí et al. (2024); Tao et al. (2021). Additionally, biopolymers are often represented as molecular structures rather than materials, leading to fragmented reporting of properties across text, tables, and figures. Without harmonized metadata, datasets from multiple sources remain difficult to merge, compare, and reuse.

2.1 Consequences for ML Workflows

Inadequate encoding limits model performance, interpretability, and generalization Talaei-Khoei & Motiwalla (2023). Missing structural hierarchy and environmental context lead to incomplete or noisy feature spaces, weakening structure-processing-property relationships and reducing transferability across datasets and applications. Specifically, biopolymer structural encoding is hindered by organism-dependent biosynthesis and environmental sensitivity, which produce batch-to-batch variability in monomer distribution, molecular-weight and dispersity profiles, stereochemistry/enantiomeric ratios, branching, crystallinity, and conformational states Kerner et al. (2021b);

Rickert & Lieleg (2022); Schuster & Stadler (1994); Caputo et al. (2022); Tang et al. (2020). Without standardized encodings, datasets remain siloed and difficult to integrate into automated ML pipelines. These inconsistencies also hinder data extraction. Although NLP methods are effective on structured tables, they still face limitations, and performance degrades when key information is implicit or embedded in figures or variable text formats Shetty et al. (2023); Gupta et al. (2024); Alampara et al. (2025); Circi et al. (2024). The resulting feature space is fragmented and noisy, limiting downstream learning. Finally, modern representation-learning approaches, including GNNs and self-supervised methods, depend on consistent schemas and identifiers. Without standardized encodings, these methods cannot reliably connect raw data to ML-ready features.

2.2 Recommendations for Standardizing Information Encoding

Biopolymer-specific structural representations are needed to move beyond linear notations optimized for synthetic polymers. Key priorities include:

- **Hierarchical representations:** Graph-based or multiscale encodings that capture monomer identity, stereochemistry, molecular-weight distributions, and higher-order structure Kunchapu & Jablonka (2025).
- **Minimal extensions to existing systems:** Add standardized tags to current polymer encodings rather than replacing them.
- **Explicit contextual metadata:** Standardize descriptors for biosynthesis, processing, and application environments using controlled vocabularies and consistent units.

These steps enable interoperable, metadata-rich representations that link structure and context in ML workflows (additional examples are provided in the Appendix). Furthermore, as a proof-of-concept, a minimal encoding pipeline can be constructed by combining automated extraction with standardized metadata mapping. For example, polyhydroxybutyrate (PHB) is a representative member of the family of PHA biopolymers Popa et al. (2022). Samples of PHB reported in the literature can be processed using an NLP pipeline to extract molecular weight, dispersity, and degradation conditions (e.g., temperature, enzyme concentration), followed by normalization into a controlled vocabulary and unit-consistent schema. These features can then be represented as (i) tabular fingerprints for classical ML and (ii) graph-based structures linking material identity, processing conditions, and measured outputs. In this way, heterogeneous text and figure data can be transformed into interoperable, ML-ready representations using existing tools with minimal extensions.

3 Data Quality

Even with improved information encoding, ML for biopolymers remains constrained by data quality. Here, data quality refers to the completeness, consistency, and reproducibility of encoded information as it moves from raw experimental outputs to ML-ready datasets Cencer et al. (2022); Meyer et al. (2022). The primary challenge is fragmentation: key descriptors such as structure, processing conditions, environmental exposure, and measured properties are rarely reported together and are instead distributed across text, tables, figures, supplementary materials, and external databases Ge et al. (2025); Kong et al. (2025). This fragmentation is further compounded by heterogeneous terminology, inconsistent units, and variability in experimental protocols across studies.

Biopolymer data are also inherently multimodal, appearing in diverse formats including structured tables, narrative text, and graphical representations Kerner et al. (2021a); Pengcheng Xu & Lu (2023). Assembling ML-ready datasets therefore requires aggregation, normalization, and semantic alignment across sources. However, standardized practices for capturing provenance, preprocessing steps, and uncertainty are often lacking, limiting dataset transparency and reuse Patel & Webb (2024); Corvi et al. (2023). In addition, many datasets omit critical documentation such as scaling and normalization procedures, missing-data handling, and versioning of both data and code. These gaps hinder reproducibility and reduce confidence in downstream applications.

3.1 Consequences for ML Workflows

Poor data quality directly impacts ML model performance and reliability. Fragmented and inconsistently normalized datasets introduce noise that obscures structure–processing–property relationships, reducing predictive accuracy and limiting generalization across datasets. Multimodal inconsistencies further hinder automated data extraction and integration, often requiring manual intervention that is difficult to standardize or reproduce. As a result, ML pipelines become less scalable and more error-prone. Lack of provenance and preprocessing transparency introduces additional uncertainty. Without clear documentation of normalization, unit conversion, or missing-data handling, it becomes difficult to determine whether model performance reflects true material behavior or artifacts of data curation. Together, these limitations constrain reproducibility, impede cross-study validation, and prevent effective aggregation of datasets at scale.

3.2 Recommendations for Standardizing Data Quality

Key priorities include:

- **Hybrid human–LLM curation pipelines:** Use LLMs to extract candidate entities, values, and relationships from literature, while domain experts validate ambiguous cases, resolve conflicting metadata, and enforce schema alignment Ramprasad et al. (2025); Duval et al. (2024). This approach enables scalable yet accurate normalization of fragmented data sources.
- **Multimodal data integration:** Compile textual, visual, and computational data into unified, ML-compatible formats linked by shared sample identifiers, method descriptors, and controlled vocabularies Fan et al. (2024); Wei et al. (2024). This ensures that extracted features remain interpretable and interoperable across datasets.
- **Explicit separation of data types:** Distinguish measured values from inferred, estimated, or assumed quantities to prevent silent errors during dataset construction and downstream ML tasks.
- **Reproducibility and provenance tracking:** Standardize documentation of preprocessing steps, including unit conversions, normalization procedures, missing-data handling, and dataset versioning, to enable auditability and reuse.

Further examples of existing workflows, validation strategies, and illustrative examples are provided in the Appendix.

4 Data Sharing

Effective data sharing is the final link connecting standardized encoding and high-quality curation to reusable, ML-ready biopolymer datasets. In this context, data sharing refers not only to making data accessible, but to publishing structured datasets in formats that support interoperability, long-term reuse, and automated analysis. For biopolymers, two persistent challenges limit progress: fragmented repositories and inconsistent reporting practices. Relevant datasets are dispersed across journal supplements, institutional repositories, proprietary databases, and laboratory-specific archives, often with limited alignment in schema, identifiers, or metadata.

Although several mature platforms exist for synthetic polymers and general materials science, Lalonde et al. (2025); Kuenneth et al. (2022); Gurnani et al. (2023) biopolymer-relevant data—particularly application-specific metrics tied to environmental exposure or bioactivity—are rarely captured in interoperable forms. Proprietary databases may contain valuable measurements, but licensing restrictions and export limitations prevent their aggregation into open ML datasets. At the same time, many biopolymer datasets appear only as supplementary files in heterogeneous formats, lacking standardized metadata or persistent identifiers. Together, these factors impede discoverability, integration, and reuse across the broader research community.

4.1 CONSEQUENCES FOR ML WORKFLOWS

Disparate data sharing practices directly constrain the scale and diversity of datasets available for ML. Fragmented repositories prevent aggregation across studies, limiting sample sizes and increasing susceptibility to bias. Inconsistent reporting further complicates reuse: identical measurements may be labeled differently, expressed in incompatible units, or stripped of critical context describing processing or testing conditions. As a result, datasets that are technically "available" often require extensive re-curation before they can be incorporated into ML pipelines, reducing efficiency and increasing the risk of error. These limitations also undermine reproducibility and model transferability. Without persistent identifiers, versioning, and clear licensing, datasets cannot be reliably cited, updated, or extended. ML models trained on such data may perform well in narrowly defined settings but fail to generalize when applied to independently curated datasets.

4.2 RECOMMENDATIONS FOR STANDARDIZING DATA SHARING PRACTICES

Key priorities include:

- **FAIR-compliant, extensible repositories:** Expand or develop centralized data platforms that support biopolymer-specific schemas while leveraging existing materials informatics infrastructure Zhu et al. (2023). Extensions should include application-relevant metadata and standardized identifiers.
- **Federated data architectures:** Enable interoperability across distributed datasets through shared identifiers and metadata standards, allowing integration without requiring centralized data ownership.
- **Incentives for data sharing:** Promote adoption through dataset DOIs, citation credit, and increased visibility for reusable datasets, ML-ready data, and aligning academic incentives with open data practices.
- **Inclusion of negative and partial results:** Broaden reporting standards to include incomplete or negative outcomes, which are critical for reducing bias and improving generalization in ML models.

In conclusion, to support near-term adoption, we propose a phased approach to standardization:

- **Phase I (near-term):** Define minimal metadata schemas and controlled vocabularies for biosynthesis, processing, and testing conditions, along with unit normalization standards. These extensions should be compatible with existing polymer representations and require minimal overhead for data contributors.
- **Phase II (mid-term):** Develop and validate biopolymer-specific encoding frameworks, including hierarchical or graph-based fingerprints, and establish hybrid human–LLM pipelines for scalable data extraction and curation.
- **Phase III (long-term):** Integrate standardized encodings into FAIR-compliant, federated repositories with persistent identifiers, enabling cross-dataset interoperability, benchmarking, and automated ML workflows at scale.

This phased strategy prioritizes immediate interoperability while enabling progressive incorporation of more expressive representations and infrastructure.

5 OUTLOOK

The effective application of ML to biopolymer materials depends less on algorithmic advances than on the availability of interoperable, high-quality, and shareable data. Although biopolymers present challenges due to biosynthetic variability, hierarchical structure, and environmental coupling, these can be addressed through standardized encoding, structured metadata, and improved curation workflows. Establishing hybrid human-LLM pipelines and FAIR-aligned infrastructure will enable the transformation of fragmented data into coherent, ML-ready datasets. Long-term success in this area will require sustained interdisciplinary collaboration to develop shared standards bridge biology and materials science through an enduring data infrastructure.

ACKNOWLEDGEMENTS

JNL gratefully recognizes the National Research Traineeship Program at Duke University (aiM) and the Laboratory Directed Research and Development (LDRD) office at the Los Alamos National Laboratory for research and resource support.

REFERENCES

Rdkit: Open-source cheminformatics, 2024. URL https://www.rdkit.org/.

Nawaf Alampara, Mara Schilling-Wilhelmi, and Kevin Maik Jablonka. Lessons from the trenches on evaluating machine-learning systems in materials science. *arXiv preprint arXiv:2503.10837*, 2025.

Beatriz Altamira-Algarra, Joan Garcia, and Eva Gonzalez-Flo. Cyanobacteria microbiomes for bioplastic production: Critical review of key factors and challenges in scaling from laboratory to industry set-ups. *Bioresource technology*, pp. 132231, 2025.

Anbreen Anjum, Mohammad Zuber, Khalid Mahmood Zia, Aqdas Noreen, Muhammad Naveed Anjum, and Shazia Tabasum. Microbial production of polyhydroxyalkanoates (phas) and its copolymers: A review of recent advancements. *International journal of biological macromolecules*, 89: 161–174, 2016.

Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature chemistry*, 13(6):505–508, 2021.

Karteek K Bejagam, Carl N Iverson, Babetta L Marrone, and Ghanshyam Pilania. Composition and configuration dependence of glass-transition temperature in binary copolymers and blends of polyhydroxyalkanoate biopolymers. *Macromolecules*, 54(12):5618–5628, 2021.

Karteek K Bejagam, Jessica Lalonde, Carl N Iverson, Babetta L Marrone, and Ghanshyam Pilania. Machine learning for melting temperature predictions and design in polyhydroxyalkanoate-based biopolymers. *The Journal of Physical Chemistry B*, 126(4):934–945, 2022.

Johannes Brandrup, Edmund H Immergut, Eric A Grulke, Akihiro Abe, and Daniel R Bloch. *Polymer handbook*, volume 89. Wiley New York, 1999.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

Maria Rosaria Caputo, Xiaoyan Tang, Andrea H Westlie, Haritz Sardon, Eugene Y-X Chen, and Alejandro J Muller. Effect of chain stereoconfiguration on poly (3-hydroxybutyrate) crystallization kinetics. *Biomacromolecules*, 23(9):3847–3859, 2022.

Morgan M. Cencer, Jeffrey S. Moore, and Rajeev S. Assary. Machine learning for polymeric materials: an introduction. *Polymer International*, 71(5):537–542, 2022. doi: https://doi.org/10.1002/pi.6345.

Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, and L Catherine Brinson. How well do large language models understand tables in materials science? *Integrating Materials and Manufacturing Innovation*, 13(3):669–687, 2024.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

J. O. Corvi, A. McKitrick, J. M. Fernández, C. V. Fuenteslópez, J. L. Gelpí, M. P. Ginebra, and O. Hakimi. Debbie: the open access database of experimental scaffolds and biomaterials built using an automated text mining pipeline. *Advanced Healthcare Materials*, 12(25):2300150, 2023. doi: https://doi.org/10.1002/adhm.202300150.

Alexandre Duval, Lucile Ritchie, Martin Siron, Inel Djafar, Etienne du Fayet, Amandine Rosello, Ali Ramlaoui, Leandro von Werra, and Thomas Wolf. Lematerial: an open source initiative to accelerate materials discovery and research, 2024. URL https://huggingface.co/blog/lematerial.

V. Fan, Y. Qian, A. Wang, A. Wang, C. W. Coley, and R. Barzilay. Openchemie: An information extraction toolkit for chemistry literature. *J Chem Inf Model*, 64(14):5521–5534, 2024. ISSN 1549-960X (Electronic) 1549-9596 (Linking). doi: 10.1021/acs.jcim.4c00572. URL https://www.ncbi.nlm.nih.gov/pubmed/38950894. Fan, Vincent Qian, Yujie Wang, Alex Wang, Amber Coley, Connor W Barzilay, Regina eng 2024/07/02 J Chem Inf Model. 2024 Jul 22;64(14):5521-5534. doi: 10.1021/acs.jcim.4c00572. Epub 2024 Jul 1.

Jaka Fajar Fatriansyah, Baiq Diffa Pakarti Linuwih, Yossi Andreano, Intan Septia Sari, Andreas Federico, Muhammad Anis, Siti Norasmah Surip, and Mariatti Jaafar. Prediction of glass transition temperature of polymers using simple machine learning. *Polymers*, 16(17):2464, 2024.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Katharina A Fransen, Sarah HM Av-Ron, Tess R Buchanan, Dylan J Walsh, Dechen T Rota, Lana Van Note, and Bradley D Olsen. High-throughput experimentation for discovery of biodegradable polyesters. *Proceedings of the National Academy of Sciences*, 120(23):e2220021120, 2023.

Wei Ge, Ramindu De Silva, Yanan Fan, Scott Sisson, and Martina H. Stenzel. Machine learning in polymer research. *Advanced Materials*, 37(11):2413695, 2025. doi: https://doi.org/10.1002/adma.202413695.

Eleonora Gianti and Simona Percec. Machine learning at the interface of polymer science and biology: How far can we go? *Biomacromolecules*, 23(3):576–591, 2022.

Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. Data extraction from polymer literature using large language models. *Communications materials*, 5 (1):269, 2024.

Rishi Gurnani, Christopher Kuenneth, Aubrey Toland, and Rampi Ramprasad. Polymer informatics at scale with multitask graph neural networks. *Chemistry of Materials*, 35(4):1560–1567, 2023.

Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Erratum: Data-driven materials science: Status, challenges, and perspectives. *Advanced Science (Weinheim, Baden-wurttemberg, Germany)*, 7(2):1903667–1903667, 2020.

Anubhav Jain. Matminer, 2015. URL https://hackingmaterials.lbl.gov/matminer/.

J. Kerner, A. Dogan, and H. von Recum. Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomater*, 130:54–65, 2021a. ISSN 1878-7568 (Electronic) 1742-7061 (Linking). doi: 10.1016/j.actbio.2021.05.053. URL https://www.ncbi.nlm.nih.gov/pubmed/34087445. Kerner, Jacob Dogan, Alan von Recum, Horst eng Review England 2021/06/05 Acta Biomater. 2021 Aug;130:54-65. doi: 10.1016/j.actbio.2021.05.053. Epub 2021 Jun 1.

Jacob Kerner, Alan Dogan, and Horst von Recum. Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomaterialia*, 130:54–65, 2021b.

Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C*, 122(31):17575–17585, 2018.

Henning Knoop, Yvonne Zilliges, Wolfgang Lockau, and Ralf Steuer. The metabolic network of synechocystis sp. pcc 6803: systemic properties of autotrophic growth. *Plant physiology*, 154(1): 410–422, 2010.

Henning Knoop, Marianne Gründel, Yvonne Zilliges, Robert Lehmann, Sabrina Hoffmann, Wolfgang Lockau, and Ralf Steuer. Flux balance analysis of cyanobacterial metabolism: the metabolic network of synechocystis sp. pcc 6803. *PLoS computational biology*, 9(6):e1003081, 2013.

Martin Koller and Anindya Mukherjee. A new wave of industrialization of pha biopolyesters. *Bioengineering*, 9(2):74, 2022.

Jessica Kong, Gihan Panapitiya, and Emily Saldanha. Extracting material property measurements from scientific literature with limited annotations. *Journal of Chemical Information and Modeling*, 65:4906–4917, 2025. doi: https://doi.org/10.1021/acs.jcim.4c01352.

Christopher Kuenneth, Jessica Lalonde, Babetta L Marrone, Carl N Iverson, Rampi Ramprasad, and Ghanshyam Pilania. Bioplastic design using multitask deep neural networks. *Communications materials*, 3(1):96, 2022.

Sreekanth Kunchapu and Kevin Maik Jablonka. Polymetrix: an ecosystem for digital polymer chemistry. *npj Computational Materials*, 11(1):312, 2025.

Jessica N Lalonde, Ghanshyam Pilania, and Babetta L Marrone. Materials designed to degrade: structure, properties, processing, and performance relationships in polyhydroxyalkanoate biopolymers. *Polymer Chemistry*, 16(3):235–265, 2025.

Janice Lim, Mingliang You, Jian Li, and Zibiao Li. Emerging bone tissue engineering via polyhydroxyalkanoate (pha)-based scaffolds. *Materials Science and Engineering: C*, 79:917–929, 2017.

Sicong Lin, Yan Zhuang, Ke Chen, Jian Lu, Kefeng Wang, Lin Han, Mufei Li, Xiangfeng Li, Xiangdong Zhu, Mingli Yang, et al. Osteoinductive biomaterials: Machine learning for prediction and interpretation. *Acta Biomaterialia*, 187:422–433, 2024.

Tzyy-Shyang Lin, Connor W Coley, Hidenobu Mochigase, Haley K Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L Craig, Jeremiah A Johnson, Julia A Kalow, et al. Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9):1523–1531, 2019.

Tzyy-Shyang Lin, Nathan J Rebello, Haley K Beech, Zi Wang, Bassil El-Zaatari, David J Lundberg, Jeremiah A Johnson, Julia A Kalow, Stephen L Craig, and Bradley D Olsen. Polydat: a generic data schema for polymer characterization. *Journal of chemical information and modeling*, 61(3): 1150–1163, 2021.

Ivan Malashin, Dmitriy Martysyuk, Vadim Tynchenko, Andrei Gantimurov, Andrey Semikolenov, Vladimir Nelyub, and Aleksei Borodulin. Machine learning-based process optimization in biopolymer manufacturing: A review. *Polymers*, 16(23):3368, 2024.

Arun Mannodi-Kanakkithodi and Maria KY Chan. Computational data-driven materials discovery. *Trends in Chemistry*, 3(2):79–82, 2021.

Didac Martí, Rémi Pétuya, Emanuele Bosoni, Anne-Claude Dublanchet, Stephan Mohr, and Fabien Léonforte. Predicting the glass transition temperature of biopolymers via high-throughput molecular dynamics simulations and machine learning. *ACS Applied Polymer Materials*, 6(8): 4449–4461, 2024.

Miguel Mateu-Sanz, Carla V Fuenteslópez, Juan Uribe-Gomez, Håvard Jostein Haugen, Abhay Pandit, Maria-Pau Ginebra, Osnat Hakimi, Martin Krallinger, and Athina Samara. Redefining biomaterial biocompatibility: challenges for artificial intelligence and text mining. *Trends in biotechnology*, 42(4):402–417, 2024.

Blaithín McAdam, Margaret Brennan Fournet, Paul McDonald, and Marija Mojicevic. Production of polyhydroxybutyrate (phb) and factors impacting its chemical and mechanical characteristics. *Polymers*, 12(12):2908, 2020.

Samantha M McDonald, Emily K Augustine, Quinn Lanners, Cynthia Rudin, L Catherine Brinson, and Matthew L Becker. Applied machine learning as a driver for polymeric biomaterials design. *Nature Communications*, 14(1):4838, 2023.

Travis A. Meyer, Cesar Ramirez, Matthew J. Tamasi, and Adam J. Gormley. A user's guide to machine learning for polymeric biomaterials. *ACS Polym Au*, 3(2):141–157, 2022. doi: https://doi.org/10.1021/acspolymersau.2c00037.

R. A. Patel and M. A. Webb. Data-driven design of polymer-based biomaterials: High-throughput simulation, experimentation, and machine learning. *ACS Appl Bio Mater*, 7(2):510–527, 2024. ISSN 2576-6422 (Electronic) 2576-6422 (Linking). doi: 10.1021/acsabm.2c00962. URL https://www.ncbi.nlm.nih.gov/pubmed/36701125. Patel, Roshan A Webb, Michael A eng Research Support, U.S. Gov't, Non-P.H.S. Review 2023/01/27 ACS Appl Bio Mater. 2024 Feb 19;7(2):510-527. doi: 10.1021/acsabm.2c00962. Epub 2023 Jan 26.

Roshan A Patel and Michael A Webb. Data-driven design of polymer-based biomaterials: high-throughput simulation, experimentation, and machine learning. *ACS Applied Bio Materials*, 7(2): 510–527, 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Minjie Li Pengcheng Xu, Xiaobo Ji and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(42):1–15, 2023. doi: 10.1038/s41524-023-01000-z.

Brandon K Phan, Kuan-Hsuan Shen, Rishi Gurnani, Huan Tran, Ryan Lively, and Rampi Ramprasad. Gas permeability, diffusivity, and solubility in polymers: Simulation-experiment data fusion and multi-task machine learning. *npj Computational Materials*, 10(1):186, 2024.

Ghanshyam Pilania, Carl N Iverson, Turab Lookman, and Babetta L Marrone. Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. *Journal of Chemical Information and Modeling*, 59(12):5013–5025, 2019.

Marius Stelian Popa, Adriana Nicoleta Frone, and Denis Mihaela Panaitescu. Polyhydroxybutyrate blends: A solution for biodegradable packaging? *International journal of biological macromolecules*, 207:263–277, 2022.

R. Ramprasad, C. Kim, J. Goldman, and B. Midgette. Matmerize: Accelerated materials design, 2025. URL https://www.matmerize.com.

Bharath Ramsundar. *Molecular machine learning with DeepChem*. PhD thesis, Stanford University, 2018.

Bharath Ramsundar. Deepchem, 2024. URL https://github.com/deepchem/.

Carolin A Rickert and Oliver Lieleg. Machine learning approaches for biomolecular, biophysical, and biomaterials research. *Biophysics Reviews*, 3(2), 2022.

Eleftheria Roumeli, Sourena Azidhak, Ana F Costa, Anlan Chen, Io Saito, Yiyang Sun, L Cate Brinson, Cynthia Rudin, Linda S Schadler, and Kayla Sprenger. From biomatter to bioplastics: A perspective on modeling, structure, and data-driven design: E. roumeli et al. *MRS Bulletin*, pp. 1–15, 2025.

Poorna Chandrika Sabapathy, Sabarinathan Devaraj, Katharina Meixner, Parthiban Anburajan, Preethi Kathirvel, Yuvaraj Ravikumar, Hossain M Zabed, and Xianghui Qi. Recent developments in polyhydroxyalkanoates (phas) production–a review. *Bioresource technology*, 306:123132, 2020.

Harikrishna Sahu, Kuan-Hsuan Shen, Joseph H Montoya, Huan Tran, and Rampi Ramprasad. Polymer structure predictor (psp): a python toolkit for predicting atomic-level structural models for a range of polymer geometries. *Journal of Chemical Theory and Computation*, 18(4):2737–2748, 2022.

Peter Schuster and Peter F Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers & chemistry*, 18(3):295–324, 1994.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52, 2023.

Arthy Surendran, Manoj Lakshmanan, Jiun Yee Chee, Azlinah Mohd Sulaiman, Doan Van Thuoc, and Kumar Sudesh. Can polyhydroxyalkanoates be produced efficiently from waste plant and animal oils? *Frontiers in Bioengineering and Biotechnology*, 8:169, 2020.

A. Talaei-Khoei and L. Motiwalla. A new method for improving prediction performance in neural networks with insufficient data. *Decision Analytics Journal*, 6(100172), 2023. doi: https://doi.org/10.1016/j.dajour.2023.100172.

Xiaoyan Tang, Andrea H Westlie, Lucia Caporaso, Luigi Cavallo, Laura Falivene, and Eugene Y-X Chen. Biodegradable polyhydroxyalkanoates by stereoselective copolymerization of racemic diolides: stereocontrol and polyolefin-like properties. *Angewandte Chemie*, 132(20):7955–7964, 2020.

Lei Tao, Vikas Varshney, and Ying Li. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *Journal of Chemical Information and Modeling*, 61(11):5395–5413, 2021.

Aubrey Toland, Huan Tran, Lihua Chen, Yinghao Li, Chao Zhang, Will Gutekunst, and Rampi Ramprasad. Accelerated scheme to predict ring-opening polymerization enthalpy: simulation-experimental data fusion and multitask machine learning. *The Journal of Physical Chemistry A*, 127(50):10709–10716, 2023.

Huan Tran, Rishi Gurnani, Chiho Kim, Ghanshyam Pilania, Ha-Kyung Kwon, Ryan P Lively, and Rampi Ramprasad. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials*, 9(12):866–886, 2024.

Viviana Urtuvia, Pamela Villegas, Myriam González, and Michael Seeger. Bacterial production of the biodegradable plastics polyhydroxyalkanoates. *International journal of biological macromolecules*, 70:208–213, 2014.

Camila Utsunomia, Qun Ren, and Manfred Zinn. Poly (4-hydroxybutyrate): current state and perspectives. *Frontiers in Bioengineering and Biotechnology*, 8:257, 2020.

Jianfei Wang, Jiaqi Huang, and Shijie Liu. The production, recovery, and valorization of polyhydroxybutyrate (phb) based on circular bioeconomy. *Biotechnology advances*, 72:108340, 2024.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res*, 52:W540–W546, 2024. doi: https://doi.org/10.1093/nar/gkae235.

Kun Yao, John E Herr, David W Toth, Ryker Mckintyre, and John Parkhill. The tensormol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical science*, 9(8): 2261–2269, 2018.

He Zhao, Yixing Wang, Anqi Lin, Bingyin Hu, Rui Yan, James McCusker, Wei Chen, Deborah L McGuinness, Linda Schadler, and L Catherine Brinson. Nanomine schema: An extensible data representation for polymer nanocomposites. *APL Materials*, 6(11), 2018.

Jun-Jie Zhu, Meiqi Yang, and Zhiyong J. Ren. Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science Technology*, 57(16):17671–17689, 2023. doi: https://doi.org/10.1021/acs.est.3c00026.

A Appendix

A.1 Action Items for Standardizing Information Encoding

A.1.1 Expand Biopolymer Specific Structural Encoding Systems

1. Launch a structural ontology that links to, and does not duplicate, PRO, PDB, ChEBI, and digitized resources such as the Polymer Handbook, which includes comprehensive polymer physical property data Brandrup et al. (1999). Such digital resources could be further curated as an ontology foundation for biopolymers with machine-readable fields for stereochemistry, molecular-weight distributions, crystallinity, branching, and hierarchical motifs.
2. Develop and prototype representation frameworks that extend on SMILES. Publish a minimal specification with examples of commonly referenced biopolymers, such as standard compositions of PLA, polyhydroxybutyrate (PHB) and other common PHAs, and cellulose acetate.
3. Provide open reference implementations (RDKit extensions + converters to graph objects) so datasets can export unified structural fingerprints.

A.1.2 Standardize Metadata

1. Build a biopolymer processing and environmental (application-specific) metadata library. This library would be analogous to the MaterialsMine framework, or perhaps a more specific application of PolyDAT Lin et al. (2021), in which every characterization on a relevant species is annotated, but tailored to biological/environmental interactions. Additionally, publish JSON/YAML schemas and controlled vocabularies Lin et al. (2021); Zhao et al. (2018).
2. Provide standard extract/validate tooling that maps literature terms to the vocabulary. For example, mapping "industrially compostable" to specific temperature/humidity/time profiles; and exports a useable format. As MaterialsMine does not currently have these entities, expanding the capabilities of existing resources will be useful.
3. Encode the standards above as a digital-twin schema to enable consistent collection, training, and cross-study validation; and align with FAIR standards.

A.1.3 Apply Node-Based Frameworks & NLP Methods for Uniform Feature Extraction and Expand Fingerprinting Libraries

1. Develop an NLP pipeline that detects entities/values in text/tables/figure captions ("ASTM D5338," "55 °C," "proteinase K, 2 U/mL") and normalizes units via the metadata library.
2. Train SSL/GNN models to learn embeddings, by which we mean vectorized, numerical representations of structural and contextual information such as molecular graphs, processing conditions, or environmental parameters, that capture relationships between similar samples Roumeli et al. (2025); Patel & Webb (2023). These embeddings allow the model to recognize patterns and semantic similarity even when explicit metadata are missing, improving extraction recall by identifying related entities or conditions across partially labeled datasets and making automated data extraction more complete and more accurate.
3. Implement a human-in-the-loop pass to validate uncertain fields. Incorporate explicit notes to separate limitations and highly uncertain values, such as ambiguous labels in visual data. These fields will then be further annotated with a hybrid process incorporating a domain expert.

A.2 Action Items for Standardizing Data Quality

A.2.1 Implement Hybrid LLM and Manual Curation to Normalize Fragmented Data

1. Guidelines and benchmarks. Release small "gold-standard" corpora with full-text annotations and clear instructions to calibrate tools and measure progress. Examples of these standard datasets should be provided for several different biopolymers.

2. Open curation playbooks. Provide templates for data conversion tables, vocabulary mappings, and normalization logs so different groups produce interoperable outputs.

A.2.2 Compile Text, Visual Contextual Data into ML-Compatible Formats

We also provide the following examples for data templates:

- From text. Extract, with a combination of manual and LLM/NLP workflows: polymer identity; molecular-weight stats; biosynthetic source, extraction conditions; processing parameters; application testing environment; and measured outputs. Attach method IDs and store with sample-level keys.
- From visuals. For example, from a degradation curve, deposit a table with sample_id, method_id, time_unit, time_value, mass_unit, mass_value, plus degradation environment tags. For a Fourier Transform Infrared spectroscopy (FT-IR) peak table or differential scanning calorimetry (DSC) trace, export peak positions/areas or tabulated arrays with instrument and calibration fields.
- From computation. Molecular dynamics (MD) and density functional theory (DFT) outputs should be stored with simulation provenance (force field or functional, timestep, temperature, box size) mapped to controlled fields so they can be compared with experimental analogs.

Existing resources such as Polymer Genome illustrate this practice for conventional polymers by integrating structure, synthesis, and testing metadata under unified schemas Kim et al. (2018). In particular, Polymer Genome encodes polymers using chemically informed descriptors—ranging from atomistic and topological to electronic features—and maps these to application-specific metrics through supervised ML models Kim et al. (2018).

Building on these frameworks, biopolymer data systems should be designed to capture not only structural, synthesis, and property information, but also the diverse application-specific metrics by which biopolymers are evaluated.

This paradigm can be further supported by advances in broader materials data infrastructure. For example, LeMaterial organizes inorganic materials data into curated datasets by standardizing representations across sources and linking entries through unique material identifiers Duval et al. (2024). Complementary cheminformatics tools such as RDKit, listed in Table 1, provide extensible foundations that can be adapted to support biopolymer-specific representations.

We further suggest the following next steps to demonstrate effective application of these existing resources:

1. Provide two export forms: (i) normalized tabular fingerprints, ready for classical ML; and (ii) graph objects for GNNs where nodes represent materials/process/measurement events and edges encode relationships. Both must reference the same ontology IDs to ensure one-to-one mapping between table columns and graph attributes.
2. Integrate with open cheminformatics libraries, such as an extended RDKit, for featurization and with repositories that already implement FAIR principles.

Table 1: An Overview of Existing Cheminformatics Packages. These packages for encoding biopolymers are available as resources to improve data quality.

PACKAGE	DESCRIPTION
Scikit-learn	**General ML Toolkit**: All-purpose feature extraction, dimensionality reduction, extensive analysis of descriptors from other libraries Pedregosa et al. (2011)
RDKit: Open-source cheminformatics	**Small-Molecule Focus** **2D Structural:** Molecular weight, LogP, polar surface area, number of rotatable bonds **3D Structural:** Molecular shape, conformation-dependent properties, connectivity indices Fingerprinting: Monomer analysis, Morgan fingerprints, MACCS keys, path-based identifiers RN3 (2024)
Matminer	**General Materials Science** **Elemental properties:** Atomic radii, electronegativity **3D Structural:** Crystal structure embedding and crystallinity **Fingerprinting:** Chemical formulas, atomic mass, valence, oxidization states Jain (2015)
DeepChem	**Deep Learning for Molecules** **Graph-based representations:** Convolutional network modeling of polymer chains Fingerprinting: Extended-connectivity, SMILES, monomer analysis Ramsundar (2018; 2024)
TensorMol-0.1	**Quantum Chemistry ML** **Graph-based representations:** Neural network modeling of polymer chains and molecules **Interatomic:** atomic environment vectors, electronic structure, quantum chemistry Yao et al. (2018)
Polymer Structure Predictor (PSP)	**Polymer Informatics** **Property Data:** Glass transition temperature, melting temperature **3D Structural:** Degree of polymerization, branching factor Fingerprinting: Monomer analysis, SMILES, InCHI Sahu et al. (2022)
Biopython	**Other Biopolymers (DNA, proteins, peptides)** Sequence: Nucleotides, codons 2D and 3D Structural: Molecular weight, secondary structure, isoelectric point, hydrophobicity, polarity Cock et al. (2009)
BIGSMILES	**Polymer Representation** **Line notation:** Extension of SMILES designed to represent stochastic polymer structures, repeating units, and macromolecular architectures for polymer informatics Lin et al. (2019)
PyTorch Geometric	**Graph-based Deep Learning** **Graph neural networks:** PyTorch-based library for building and training GNN models on molecular and materials graphs for structure–property prediction Fey & Lenssen (2019)