Published in final edited form as:

ACS Macro Lett. 2017 October ; 6(10): 1078–1082. doi:10.1021/acsmacrolett.7b00228.

Polymer Informatics: Opportunities and Challenges

Debra J. Audus[†] and Juan J. de Pablo^{‡,¶}

[†]Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899 USA

[‡]The Institute for Molecular Engineering, The University of Chicago, Chicago, Illinois, 60637 USA

[¶]The Computation Institute, The University of Chicago, Chicago, Illinois, 60637 USA

Abstract

We are entering an era where large volumes of scientific data, coupled with algorithmic and computational advances, can reduce both the time and cost of developing new materials. This emerging field known as materials informatics has gained acceptance for a number of classes of materials, including metals and oxides. In the particular case of polymer science, however, there are important challenges that must be addressed before one can start to deploy advanced machine learning approaches for designing new materials. These challenges are primarily related to the manner in which polymeric systems and their properties are reported. In this viewpoint, we discuss the opportunities and challenges for making materials informatics as applied to polymers, or equivalently polymer informatics, a reality.

Graphical abstract

Development times for new materials can be a staggering 10–20 years,¹ despite the ever-growing materials literature. In order to reduce both the time-to-market and development cost of new products, ideally by a factor of two or more,¹ a new approach known as materials informatics has emerged.^{2,3} The idea is to train machine learning algorithms on large databases, in order to identify previously unrecognized trends or patterns, and even propose new candidate materials. If successful, materials informatics could considerably improve how new materials, including polymers, are developed. Materials informatics has become an important component of recent international efforts, including the Materials Genome Initiative (USA),⁴ the “Materials Research by Information Integration” Initiative (Japan)⁵ and the NOMAD Laboratory: a European Centre for Excellence (EU).⁶

Specifically, the new framework for materials discovery enabled by materials informatics, as illustrated in Fig. 1, is (1) scientists define specifications for the new material, (2) materials informatics, coupled with physics-based models, is used to propose potential candidates, (3) candidates are tested experimentally for viability using insights from machine learning, theory and simulation, (4) an industrial process is developed, and (5) the new material is released. In this viewpoint, we focus on the materials informatics in the larger framework of materials discovery.

What is unique about materials informatics is the focus on data and informatics, sometimes referred to as the fourth paradigm in materials discovery.⁷ In materials discovery, the original paradigm is the essential process of experimentation. That often time-consuming process can be assisted by theory, the second paradigm, in the form of physics-based models. More recently, the third paradigm of computer simulations, including integrated computational materials engineering (ICME), where multiple length scales are linked to predictively model materials such as metal alloys, has been able to further reduce development times.⁸ Using the fourth paradigm of materials informatics does not displace any of the prior paradigms; instead, it should be used in conjunction with them to reach the ultimate goal of accelerating materials discovery (see Fig. 1). In fact, informatics relies on data from experiments and simulations. This framework for materials discovery, along with its associated goals, can be applied to all types of materials, including polymers.

Promising examples of materials informatics for design can be found for different classes of materials. Efforts in the domain of inorganic materials, have shown that machine learning can be used to successfully predict new stable compounds based on a database generated by high-throughput density functional theory calculations.⁹ To date, analogous efforts in polymer science and engineering have been limited in scope and examples of use for polymers design, as opposed to property prediction, are rare. Some of the earlier efforts in polymer informatics, often under the umbrella of quantitative structure-property relationships^{10–13} including the use of neural nets to predict the glass transition temperature in 1994,¹⁴ are highlighted in the 2012 review article by Winkler and coworkers.¹³ More recently, there has been work on predicting the dielectric constant,^{15,16} refractive index,¹⁶ and tensile strength at break.¹⁷ Of these efforts, only one¹⁵ goes beyond property prediction to design—starting with specifications and using the predictions to identify candidates (steps 1 and 2 in Fig. 1), and then confirming candidates with experiment (step 3 in Fig. 1). Although this case study relies primarily computational data, rather than both experimental and computational data, as input, it gives a glimpse of what is possible for polymer informatics.

The largest hurdle for the widespread use of polymer informatics, especially for use in design, is the lack of databases^{13,18,19} not a lack of machine learning algorithms. Currently, there are numerous free resources ranging from textbooks^{20,21} and articles^{22,23} to software including scikit-learn and TensorFlow, among others. Nor is the problem a lack in the quantity of data, as there has been a near doubling of the number of polymer related articles published in the last 20 years.²⁴ The traditional solution of relying on domain experts to find, read and extract relevant information manually from journal articles is not viable long term as the literature is growing exponentially, while resources (people, time, funding) are

unlikely to keep pace. The continued improvement of machine learning opens up the intriguing prospect of generating databases automatically, thereby eliminating or, more likely, reducing human intervention in database creation. Machine learning, with proper training sets, can not only identify which journal articles are the most likely to contain desired data,²⁵ but can also be used to read and interpret such articles. Recent successes include IBM's Watson project, which parsed prose-based sources, including Wikipedia, to win Jeopardy!²⁶ Additionally, M.C. Swain and J.M. Cole at Cambridge University applied these concepts and developed a toolkit for the automated extraction of thermodynamic properties, including the melting temperature and measurements related to NMR spectra (e.g., peak values) for small molecules.²⁷ To yield improved accuracy, such efforts can be coupled with review of lower certainty data by human curators, reducing, but not eliminating human effort. In addition to generating the necessary databases for polymer informatics, the automatic creation of databases also has the potential to reduce the burden on individual scientists by making it easier to find data and validate new models. Perhaps, these efforts can one day be expanded to also capture ideas—further reducing the burden. In the meantime, they are a promising route for database creation.

Even with these state-of-the-art algorithmic advances, the creation of databases, and in particular polymeric databases still suffers from challenges that predate the invention of the computer. These include a proper description of the material of interest, subtleties associated with the property, the context of the underlying measurement and the reporting of all the necessary information. The limited scope of existing polymer data resources^{28–36} is partly due to these challenges; addressing them will require both fundamental research and, importantly, that consensus be reached by the polymer science community. Barriers to database creation must be reduced, reproducibility must be encouraged, and discussion must be fostered in order to make polymer informatics possible. The importance of such efforts, which will also benefit individual scientists, was highlighted in a recent National Science Foundation Workshop entitled Frontiers in Polymer Science and Engineering, which took place on August 17–18, 2016.³⁷ As in research, the process of tackling these challenges will be iterative—adopt available solutions while developing new solutions.

An inherent challenge to polymeric database creation stems from the fact that a synthetic polymer is rarely a single entity. Unlike small molecules or proteins, even individual polymer samples are generally described by distributions. In the simplest case, a sample synthesized from a single achiral monomer has only one distribution, the molecular mass distribution. However, the number of relevant distributions can rapidly become intractable when chiral monomers, multiple monomers, or chain branching are considered. For now, both the molecular weight and the dispersity should be captured in databases. In the future and as polymer synthesis continues to advance, additional data such as the degree of branching or even the raw output of characterization techniques could be reported in a standardized form.

The need to describe these distributions, combined with often complicated monomeric structures and sequences leads to nonstandard naming conventions. The use of commercial trade names further complicates matters. For example, polystyrene can be described by at least 1,800 different names.³⁸ Even International Union of Pure and Applied Chemistry

(IUPAC) naming conventions and Chemical Abstracts Service Registry Numbers, or simply, CAS numbers, both of which identify polymers based on their monomeric structure, have shortcomings—this helps explain why PolyInfo,²⁸ the largest online polymers database, created its own numbering system. The problem of nomenclature is exacerbated in sequenced defined polymers. The viewpoint by J.-F. Lutz³⁹ and the ensuing discussion⁴⁰ in this journal serve to emphasize this point. One potential, partial solution, is provided by the IUPAC international chemical identifier (InChI).⁴¹ This system is derived from chemical structures (unlike CAS numbers), and is relatively compact (unlike some IUPAC names). For these reasons, it is already used in databases such as PubChem, which reports the biological activity of small molecules.⁴² However, InChI only recently added experimental support for linear polymers, and does not support branched polymers. For now, it should be used whenever possible, and efforts should be made to convince the InChI trust to continue development for more complicated polymeric structures (branched, organometallic, Markush) in the future.

In addition to the list of components that make up a sample, a database also requires description of the property of interest. Such properties can be further categorized into well-defined groups such as thermodynamic, mechanical, transport, electromagnetic, optical. While properties may initially be thought of as well-defined entities, this can often not be true. Properties can be split up into three distinct categories: fundamental, application, and phenomenological. In the first case, no additional information needs to be specified about the property itself, for example, density, viscosity, and heat capacity. In the second case, the method of measurement must be specified, but a well-defined framework for performing and interpreting the measurement exists, for example the ASTM standard for determination of tensile elongation. In the last case, even an exact physical meaning is lacking. Perhaps the most notable example is the effective Flory-Huggins χ parameter. This simple, single parameter has been used to map out complicated phase diagrams of homopolymers and copolymers with considerable success.^{43,44} Yet, due to its phenomenological nature, its value depends on choice of system, measurement method and analysis. For example, choosing to report a mean-field or fluctuation corrected χ can significantly alter its value⁴⁵ and issues associated with its applicability still remain.⁴⁶ For these cases, additional research is warranted.

The contextual information associated with the property can also play an important role. Even fundamental properties cannot always escape the need for detailed contextual information. For example, polymeric density, naively a state variable, may vary significantly depending on the processing history.⁴⁷ The issue in this case is not the description of the property, but rather the accurate characterization of the sample being measured. As one moves toward application properties, where properties now become a function of both the sample and the method of measurement, the amount of contextual information required is even greater. In the most extreme case, phenomenological properties require detailed description of the sample, method of measurement and analysis. For these cases, raw data and analysis routines should be provided, as they will improve polymer informatics predictions and improve reproducibility. The need to describe such key contextual information, including the processing history, will only become more important as polymer science continues toward new, often out-of-equilibrium, frontiers. However, each additional

piece of contextual information will make it more challenging for algorithms to automatically capture the information. For now, the decisions regarding contextual information are made by database creators when they develop the schema for properties of interest, but as time progresses, feedback from both human and computer database users can inform these choices.

In addition to all of the aforementioned aspects that need to be captured, there is also the issue of accuracy, precision, veracity and non-reporting. For the former two, error bars along with a discussion of the error bars should be reported, and for veracity, erroneous published results should be challenged. Specifically, a resource needs to be put in place for challenging erroneous data while minimizing negative side effects. Finally, for non-reporting, a resource needs to exist to collect these “non-results.” This resource could take the form of a database or a set of journals. Such a resource would be particularly useful in the case of well-defined situations such as unexpected results from chemical reactions with known reactants and conditions. Long-term, this requires a push from both the community and publishers, while in the short-term, outliers in databases can be identified and flagged, allowing such data to be ignored during informatics but potentially added back later, if warranted.

We have faced many of the aforementioned challenges in our own efforts to build a polymer χ parameter database, where we have relied on a combination of automation and crowd-sourcing.³¹ Structured quantities such as the article title were entered automatically, while unstructured quantities, appearing in different forms and locations, such as the method of measurement, were entered by humans.⁴⁸ Based on our results, we demonstrated that machine learning can be used to identify the articles that contain χ , thereby reducing the human effort, but not eliminating it.²⁵ To further complicate the problem, we find that not all of the necessary contextual information is always specified. We are currently working on the next steps: reducing the need for crowd-sourcing, expanding the properties included in the database, exploring ways to ensure the accurate capture of data, and continuing to lay the groundwork for polymer informatics. We suspect that such efforts will still be lacking in part due to an inability to capture the information that did not make it in the publication.

Many of the challenges associated with automatic data retrieval, including those related to a lack of reported data, could be avoided by adopting a conceptually simpler solution to populate databases—when scientists publish their results, they “deposit” their own data. Such data deposits could take three different forms. In the first, raw data are made available via a link in the journal article; this will improve reproducibility, but is lacking in regard to polymer informatics. In the second, better, form data are stored in a repository where they are indexed and have a persistent identifier, improving findability. In the third, and best form, data are directly deposited into a database with a common schema such that they can be immediately used for polymer informatics. This approach has been embraced by the biological community, who store their data in the Research Collaboratory for Structural Bioinformatics (RCSB) protein databank; the RCSB protein databank now contains a staggering 38,000 distinct protein structures.^{49,50} While this is the end goal, it is not a feasible place to start for data deposits if one wants to collect all polymeric data given that schemas would have to be developed for each type of data. A more practical solution for capturing polymeric data as a whole is to allow the authors of journal articles to create their

own schema or borrow existing schema from a central library of schema that other users created. This means that a framework needs to be created where the researchers can specify the values and the contextual information, possibly pulling from a “dictionary” of common terms that are often specified (e.g., number average molecular mass). Such efforts are already in progress by the creators of three resources: the Schema Repository and Registry,⁵¹ which provides a platform for registering schemas, the Materials Data Curation System,^{52,53} which already includes a schema creator to help transfer data into a structured format, and the Materials Data Facility,^{54,55} which is a site for publication of materials datasets of all sizes, as well as for materials data discovery. Dictionaries, schemas and data resources are also being developed in conjunction with the aforementioned efforts by the Center for Hierarchical Materials Design.⁵⁶

Encouraging researchers to deposit their data is also a challenge unto itself. To overcome this barrier, possible incentives include: (1) requirement by funding agencies, (2) requirement by journals, (3) recognition or gain directly from the activity. The first two incentives represent a change in policies toward data sharing—a direction in which they are slowly moving. For example, some grants require that journal articles be made available in PubMed. Additionally, the journal *Science* requires certain data, such as molecular structure data, be deposited in databases, and most major publishers have at least some data policies in place. The third incentive could take two forms. One is voluntarily data deposits for the good of the community, much like the driving force behind Wikipedia. Another is to provide resources to go along with data deposits. This could take the form as a platform, similar to the Galaxy Project⁵⁷ where analysis routines are shared, workflows are developed and data deposits are made trivial—not only eliminating the burden of depositing data but also making the researcher more efficient. Given that implementing such procedures and software could be slow, a two-pronged approach consisting of voluntary data entry and computer automated data extraction should be immediately supported to create databases and enable polymer informatics.

The polymer science community is currently presented with an extraordinary opportunity. Advances in computer science have shown that the immense amount of valuable scientific data trapped in the literature may be harnessed by automating database creation, which in turn could be greatly assisted by data deposits. The resulting resources, coupled with emerging machine learning algorithms, could be invaluable for discovery of promising new materials. Additionally, the process of developing these resources will bolster day-to-day research via improved access to data through accessible resources including previously unreported data, improved reliability through identification of errors, improved understanding through discussions around these ideas, and improved efficiency through analysis platforms. However, none of this will be possible unless polymer scientists finally come to terms with the decades long conundrum of how to describe the material of interest, interpret the property, determine what context is necessary and how to share their data. Polymers currently represent over five hundred billion dollars in the shipment of goods. That amount is expected to grow at twice the rate of the US gross domestic product.⁵⁸ That industry, the polymer community, and the world’s economy would be well served by spearheading efforts to organize, curate, and exploit experimentally and computationally generated data.

Acknowledgments

The authors thank Dr. Kathryn Beers for insightful discussions, as well as the reviewers for their thought provoking comments. This work was supported in part by NIST contract 60NANB15D077, the Center for Hierarchical Materials Design. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial equipment and/or materials are identified in this report. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment and/or materials used are necessarily the best available for the purpose.

References

1. National Science and Technology Council. Materials Genome Initiative for Global Competitiveness. 2011.
2. Nosengo N. Can artificial intelligence create the next wonder material? *Nature* (London, U K). 2016; 533:22–25. [PubMed: 27147015]
3. Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bull.* 2016; 41:399–409.
4. [accessed Mar. 2017] Materials Genome Initiative. <https://www.mgi.gov/>
5. [accessed Mar. 2017] “Materials Research by Information Integration” Initiative. <http://www.nims.go.jp/MII-I/en/>
6. [accessed Mar. 2017] The NOMAD Laboratory: a European Centre for Excellence. <https://nomad-coe.eu/>
7. Tolle KM, Tansley DSW, Hey AJG. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proc IEEE.* 2011; 99:1334–1337.
8. Kuehmann C, Tufts B, Trester P. Computational Design for Ultra High-Strength Alloy. *Adv Mater Processes.* 2008; 166:37–40.
9. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM.* 2013; 65:1501–1509.
10. van Krevelen, DW., te Nijenhuis, K. Properties of Polymers: their Correlation with Chemical Structure; Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions. 4. Elsevier; Amsterdam: 2009.
11. Bicerano, J. Prediction of Polymer Properties. Marcel Dekker; New York: 2002.
12. [accessed Mar. 2017] Dassault Systèmes BIOVIA, BIOVIA Materials Studio Synthia. <http://accelrys.com/products/datasheets/synthia.pdf>
13. Le T, Epa VC, Burden FR, Winkler DA. Quantitative Structure Property Relationship Modeling of Diverse Materials Properties. *Chem Rev* (Washington, DC, U S). 2012; 112:2889–2919.
14. Sumpter BG, Noid DW. Neural networks and graph theory as computational tools for predicting polymer properties. *Macromol Theory Simul.* 1994; 3:363–378.
15. Mannodi-kanakithodi A, Pilania G, Huan TD, Lookman T, Ramprasad R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci Rep.* 2016; 1–10. [PubMed: 28442746]
16. Jabeen F, Chen M, Rasulev B, Ossowski M, Boudjouk P. Refractive indices of diverse data set of polymers: A computational QSPR based study. *Comput Mater Sci.* 2017; 137:215–224.
17. Cravero F, Martínez MJ, Vazquez GE, Díaz MF, Ponzoni I. Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design. *Journal of Integrative Bioinformatics.* 2016
18. de Pablo JJ, Jones B, Kovacs CL, Ozolins V, Ramirez AP. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr Opin Solid State Mater Sci.* 2014; 18:99–117.
19. Warren JA, Boisvert RF. Workshop Report: Building the Materials Innovation Infrastructure: Data and Standards a Materials Genome Initiative Workshop. 2012 Report number NISTIR-7898.

20. James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning with Applications in R. Springer-Verlag; New York: 2013. <http://www-bcf.usc.edu/~garth/ISL/> [accessed Aug. 2017]
21. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. Springer-Verlag; New York: 2009. <http://web.stanford.edu/~hastie/ElemStatLearn/> [accessed Aug. 2017]
22. Mueller, T., Kusne, AG., Ramprasad, R. Reviews in Computational Chemistry. John Wiley & Sons, Inc; Hoboken, NJ: 2016. p. 186-273. <http://rampi.ims.uconn.edu/wp-content/uploads/sites/486/2016/12/154.pdf> [accessed Aug. 2017]
23. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., Kim, C. [accessed Aug. 2017] Machine Learning and Materials Informatics: Recent Applications and Prospects. 2017. arXiv: 1707.07294 [cond-mat.mtrl-sci]. arXiv.org e-Print archive. <https://arxiv.org/abs/1707.07294>
24. Based on a search of Web of Science (<https://webofknowledge.com> (accessed Mar. 2017)) specifying date range and the research area of polymer science. 53,884 articles published between 1990 and 1994. 98,968 articles published between 2010 and 2014.
25. Tchoua RB, Chard K, Audus D, Qin J, de Pablo J, Foster I. A hybrid human-computer approach to the extraction of scientific facts from the literature. *Procedia Computer Science*. 2016; 80:386–397. [PubMed: 28649288]
26. Chu-Carroll J, Fan J, Boguraev B, Carmel D, Sheinwald D, Welty C. Finding needles in the haystack: Search and candidate generation. *IBM J Res Dev*. 2012; 56:6:1–6:12.
27. Swain MC, Cole JM. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model*. 2016:1894–1904. [PubMed: 27669338]
28. [accessed Mar. 2017] Polymer Database (PoLyInfo). http://polymer.nims.go.jp/index_en.html
29. [accessed Mar. 2017] CHEMnetBASE - Polymers: a Property Database. <http://poly.chemnetbase.com>
30. [accessed Mar. 2017] Springer Materials. <http://materials.springer.com/>
31. [accessed Mar. 2017] Polymer Property Predictor and Database. <http://pppdb.uchicago.edu>
32. Brandrup, J.Immergut, EH., Grulke, EA., editors. Polymer Handbook. 4. Wiley-Interscience; New York: 1999.
33. Mark, JE., editor. Physical Properties of Polymers Handbook. Springer; New York: 2007.
34. [accessed Mar. 2017] NanoMine. <http://nanomine.northwestern.edu>
35. [accessed Mar. 2017] Heat Capacities of Solid Polymers (the Advanced THERmal Analysis System, ATHAS). 1990. <http://www.osti.gov/servlets/purl/7021212-WabTRM/>
36. [accessed Aug. 2017] Khazana: A Computational Materials Knowledgebase. <http://khazana.uconn.edu/>
37. [accessed Mar. 2017] National Science Foundation Workshop: Frontiers in Polymer Science and Engineering. Aug 17–18. 2016 <https://sites.google.com/a/umn.edu/nsf-polymer-workshop/>
38. [accessed Mar. 2017] Based on CAS entry 9003-53-6. See <http://www.commonchemistry.org/>
39. Lutz JF. Aperiodic Copolymers. *ACS Macro Lett*. 2014; 3:1020–1023. [PubMed: 25419485]
40. Rowan SJ, Barner-Kowollik C, Klumperman B, Gaspard P, Grubbs RB, Hillmyer MA, Hutchings LR, Mahanthappa MK, Moatsou D, O'Reilly RK, Ouchi M, Sawamoto M, Lodge TP. Discussion on “Aperiodic Copolymers”. *ACS Macro Lett*. 2016; 5:1–3.
41. [accessed Mar. 2017] Download InChI version 1 (software version 1.05) for Standard and Non-Standard InChI/InChIKey. Jan 27. 2017 <http://www.inchi-trust.org/downloads/>
42. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. *Nucleic Acids Res*. 2016; 44:D1202–D1213. [PubMed: 26400175]
43. Matsen MW. The standard Gaussian model for block copolymer melts. *J Phys : Condens Matter*. 2002; 14:R21–R47.
44. Bates FS, Schulz MF, Khandpur AK, Forster S, Rosedale JH, Almdal K, Mortensen K. Fluctuations, conformational asymmetry and block copolymer phase behaviour. *Faraday Discuss*. 1994; 98:7–18.

45. Fredrickson GH, Helfand E. Fluctuation effects in the theory of microphase separation in block copolymers. *J Chem Phys.* 1987; 87:697–705.
46. Miquelard-Garnier G, Roland S. Beware of the Flory parameter to characterize polymer-polymer interactions: A critical reexamination of the experimental literature. *Eur Polym J.* 2016; 84:111–124.
47. Bueche, F. *Physical Properties of Polymers.* Interscience Publishers; New York: 1962.
48. Tchoua RB, Qin J, Audus DJ, Chard K, Foster IT, de Pablo J. Blending Education and Polymer Science: Semiautomated Creation of a Thermodynamic Property Database. *J Chem Educ.* 2016; 93:1561–1568. [PubMed: 27795574]
49. [accessed Mar. 2017] Research Collaboratory for Structural Bioinformatics Protein Data Bank. www.rcsb.org
50. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, PEB. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28:235–242. [PubMed: 10592235]
51. [accessed June 2017] Schema Repository and Registry. <https://schemas.nist.gov/>
52. [accessed June 2017] Materials Data Curation System. <https://github.com/usnistgov/MDCS>, <https://mgi.nist.gov/materials-data-curation-system>
53. Dima A, Bhaskarla S, Becker C, Brady M, Campbell C, Dessauw P, Hanisch R, Kattner U, Kroenlein K, Newrock M, Peskin A, Plante R, Li SY, Rigodiat PF, Amaral GS, Trautt Z, Schmitt X, Warren J, Youssef S. Informatics Infrastructure for the Materials Genome Initiative. *JOM.* 2016; 68:2053–2064.
54. [accessed June 2017] Materials Data Facility. <https://materialsdatafacility.org/>
55. Blaiszik B, Chard K, Pruyn J, Ananthakrishnan R, Tuecke S, Foster I. The Materials Data Facility: Data Services to Advance Materials Science Research. *JOM.* 2016; 68:2045–2052.
56. Center for Hierarchical Materials Design. [accessed Aug. 2017] <http://chimad.northwestern.edu>
57. [accessed June 2017] Galaxy Project. <https://galaxyproject.org>
58. SPI: The Plastic Industry Trade Association. [accessed June 2017] SPI, Size and Impact of the Plastics Industry on the US Economy. 2015. <http://www.plasticsindustry.org/sites/plastics.dev/files/U.S.%20Size%20and%20Impact%202015.pdf>

Figure 1.

The framework for materials discovery that is enabled by materials informatics, the fourth paradigm. The other three paradigms, which still play a critical role and are used to inform materials informatics, are experiment, simulation and theory.