MRS Communications (2019), *9*, 537–544

doi:10.1557/mrc.2019.54

MRS Communications

Artificial Intelligence Prospective

Challenges and opportunities of polymer design with machine learning and high throughput experimentation

Jatin N. Kumar, Institute of Materials Research & Engineering, 2 Fusionopolis Way, #08-03, 138634, Singapore
Qianxiao Li, and **Ye Jun,** Institute of High-Performance Computing, 1 Fusionopolis Way, #16-16, 138632, Singapore
Address all correspondence to Jatin N. Kumar at **jatinkumar@mac.com** and **kumarjn@imre.a-star.edu.sg**

(Received 18 January 2019; accepted 17 April 2019)

Abstract

In this perspective, the authors challenge the status quo of polymer innovation. The authors first explore how research in polymer design is conducted today, which is both time consuming and unable to capture the multi-scale complexities of polymers. The authors discuss strategies that could be employed in bringing together machine learning, data curation, high-throughput experimentation, and simulations, to build a system that can accurately predict polymer properties from their descriptors and enable inverse design that is capable of designing polymers based on desired properties.

Introduction

Polymers are ubiquitous in their use spanning a vast range of materials due to their highly tunable physical and chemical properties. They are critical enablers of many modern and emerging technologies today, ranging from structural material used to build modern aircraft, to flexible electronics, and even emulsifiers in personal care products—arguably making them one of the most important classes of material that exist. The properties that govern their specific use are influenced by the architecture or structure,[1] made up of the following four parameters: topology, composition, functionality, and size (Fig. 1). Fine control of polymer architecture and a deep understanding of its relationship with physiochemical properties is what allows for the innovation of materials with novel properties with important end-applications.[2–7]

Polymer properties are influenced by either polymer dynamics or chemo-functionality or both. Polymer dynamics are governed by the entanglement and flexibility of the polymer chain, which in turn influence physical properties such as rheology, glass transition temperature, and mechanical properties. Chemo-functionality affords the polymer chains' chemical response, biologic activity, and electro-chemical activity.

A combination of the two results in interesting morphological and phase characteristics, as well as stimuli-responsive characteristics, which have been thoroughly exploited in drug delivery.[8]

Apart from understanding structure–property relationships, the mechanism, and kinetics of polymerization, the process of polymer synthesis is also a critical area of study of importance when preparing tailored material to fine-tune polymer properties.

For instance, copolymerization compositional relationships between your reaction feed and final product is governed by the Mayo–Lewis equation, which relates the two monomers with their reactivity ratio, a function of their propagation rate constants. The ability to use the equation as an accurate calculator depends on several factors, most important of which is the reactivity ratio. Other factors include the type of polymerization and the conversion. Reactivity ratios are not readily available for all monomer combinations, which have to typically be calculated through extensive and systematically controlled experiments.

State-of-the-art in polymer design

The unique behavior of polymers comes about due to the complexity of the material across multiple length scales. When focusing on a molecular level, these are interactions of a single repeating unit on a polymer to: (1) the unit next to it; (2) the penultimate unit next to it; (3) other units on the polymer chain in the event of the chain coiling on itself, inclusive of the end-group; (4) the solvent or other molecules that are mixed with the polymer; and (5) repeat units of other polymers. The extent of these interactions is often multiplied with increasing polymer size, concentration of the polymer and components, other dissolved compounds in solution and the temperature.

In certain cases, forward empirical, physical, and kinetic models, or mathematical equations governing a particular relationship, provide an excellent predictive tool with high accuracy. Examples of these include the Flory–Fox, Mark–Houwink,

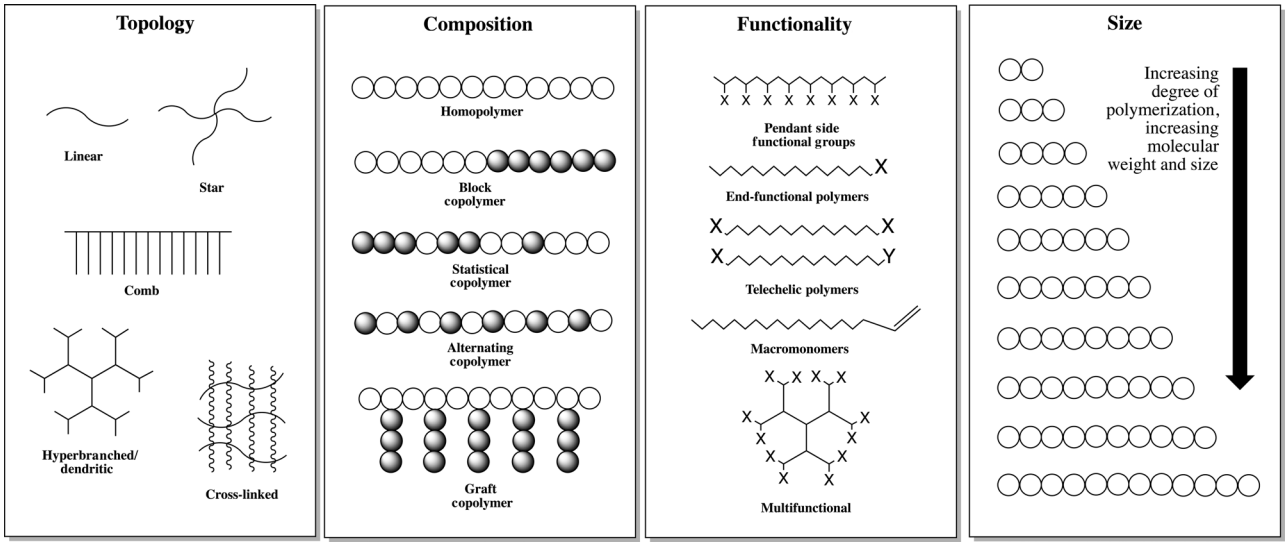


Figure 1. An illustration of the four fundamental parameters of polymer architecture, namely, topology, composition, functionality, and size that influence the properties of a polymer.

Carothers, and Mayo–Lewis equations.

$$T_g = T_{g,\infty} - \frac{K}{M_n}$$

The above Flory–Fox equation describes an empirical forward model, which is commonly used to relate the molecular weight dependence of the glass transition temperature (T_g) of polymers. The T_g of a polymer is a function of the glass transition temperature of the polymer at infinite molecular weight ($T_{g,\infty}$), the molecular weight of the polymer and a constant (K).

$$[\eta] = KM^a$$

The above Mark–Houwink equation relates the intrinsic viscosity $[\eta]$ to molecular weight (M) with the coefficients K and a unique for each polymer and architecture. One of the primary uses of this equation is in the determination of molecular weight distribution *via* size exclusion chromatography using a refractive index detector. Typically, this model generalizes well when calibrating instruments with a dissimilar polymer to the analyte.

$$\bar{X}_n = \frac{1}{1-p} \quad \bar{X}_n = \frac{1+r}{1+r-2rp}$$

The above equations are part of the Carothers equations that relate the degree of polymerization from fractional monomer conversion for step-growth polymerization. The two equations are meant for linear copolymerization where the two monomer constituents are either equal or unequal in ratio respectively, where the average degree of polymerization, $\bar{X}_n$ is a function of the conversion, p, and the stoichiometric ratio of reactants, r. Derivations of the two equations have also been made to capture the complexity of branched polymers with multifunctional monomers.

$$F_1 = 1 - F_2 = \frac{r_1 f_1^2 + f_1 f_2}{r_1 f_1^2 + 2 f_1 f_2 + r_2 f_2^2}$$

$$\frac{d[M_1]}{d[M_2]} = \frac{[M_1](r_1[M_1] + [M_2])}{[M_2]([M_1] + r_2[M_1])}; \quad r_1 = \frac{k_{11}}{k_{12}} \quad r_2 = \frac{k_{22}}{k_{21}}$$

The Mayo–Lewis equation much like the Carothers equations is part of a larger group of equations that govern or are governed by the kinetics polymerization. This equation calculates the distribution of monomers in a two monomer batch copolymerization. The final composition, F_1 is a function of the reactivity ratios (which are in turn a function of the reactivity rate coefficients) and the feed composition of the monomers.

However, beyond this, simple forward models for structure–property relationships are rare. Thus, either molecular dynamics simulations or experimentation have been traditionally employed to map the relationship in this multi-variable parameter space. Bicerano demonstrates how with an increasing amount of available data, efforts have also been made towards the quantitative prediction of polymer properties based on the topological information through their connectivity indices.[9]

Stuart et al. highlighted a vast number of experiments where phase behavior of polymers in solution was modulated by external stimuli, which shed light on structure–property relationships, reported in the form of plots and tables.[10] While these carefully conducted experiments have been the starting point for numerous synthetic chemists, the method itself is archaic, having been the only way of dissemination for multiple decades.[11–13] However, this method does not scale, nor does it

https://doi.org/10.1557/mrc.2019.54

succinctly capture the multi-variable parameter space of phase properties, as is the case with many other physical properties of polymers. In fact Halperin et al. discuss that after several decades, this area of research is now in its maturity yet we have still not been able to derive a clear, quantitative understanding of not only phase behavior, but also many other structure-property relationships.[13]

An expected evolution from the norm was the attempt by Hoogenboom et al., who demonstrated that a quadratic equation was adequate in predicting the temperature responsive phase behavior of a particular copolymer based on its molecular weight and composition.[14] While definitely an advance, this technique is limited by its simplicity, restricted to a variable space of three. The approach also makes the assumption that this relationship is quadratic, which may contribute to a poor fit of the data.

In polymerization kinetics, Stockmayer and Flory attempted to solve the Carothers equation for predicting the extent of gelation in a polymerization statistically.[15] A number of assumptions are made to simplify the variable space, including ones that may not necessarily be true such as the absence of intra-polymeric reactions and that the reactivity of all functional groups are the same. The calculations were originally performed by hand and the calculations are iterative in nature. A method such as this could be prone to error, not only due to the assumptions, but also the fact that detailed information and meticulous calculation would be required for accuracy.

These examples are non-exhaustive, and there are numerous others such as the relationship of rheological properties to architecture, and other polymer interactions. A summary of all the forward model and relationships discussed in this section is provided in Table I.

A new approach to polymer design

A key challenge thus is a succinct way to parameterize structure–property relationships. Both the quadratic (or polynomial) and manual statistical methods mentioned above attempt to deal with the multi-parameter space, but often find themselves falling short as the complexities of such a space are often non-trivial and thus require numerous assumptions. This almost always leads to important parameters being overlooked.

A computational data-driven approach to predictive modeling would be able to relate the data in a more rapid and complete manner, and could be a strategy to overcome the aforementioned challenges.[16] There are two broad classes of methods, namely statistical models (such as multivariate analysis[17] and Bayesian inference[18]) and machine learning models (such as support vector machines,[19] decision tree learning,[20] and deep neural networks.[21]) Statistical models usually perform well on relatively small datasets, but require nontrivial domain information, such as statistical priors and the form of the relationship between inputs to outputs such as a mathematical or forward model. These may not be immediately obvious for some applications, but they severely impact model accuracy and may limit applicability. On the other hand, machine learning models tend to make minimal assumptions and directly learn representations and relationships from data. For sufficiently large datasets, it is often more robust when the underlying mechanisms that links the input and the outputs are unclear, or when the dataset has unknown systematic and random noise corruptions.[22] Of course, one disadvantage of machine learning-based methods is that, unlike statistical models, besides the final prediction very little can be inferred about the underlying physical process that produced the data. Nevertheless, in many situations when the key goal is producing accurate predictions, machine learning methods are well-suited.

A further advantage of machine learning-based predictive modeling is fast inference speed. Once a model is trained, it can make predictions quickly on previously unseen data. This is especially important in inverse design, where a (global) optimization algorithm is required to minimize some objective that depends on our data-driven model. This involves repeated calls to the fitted predictive model, which must be fast to ensure an extensive search over the desired parameter space. In this sense, simulation-based predictive models will be prohibitively slow and computationally expensive for such an optimization and design process.

Machine learning serves as an enabling tool that allows us to work in a multi-variable parameter space when mapping polymer architecture to a physical property. In the case of cloud points, the traditional technique would be to generate multiple detailed phase plots for systematically varied polymer composition,[23] an empirical approach similar to examples above. Such curation is often slow, laborious, and could be prone to error with inconsistencies in data quality. In contrast, machine learning is able to map datasets that may not be systematically varied, or consistent in quality, into one unifying algorithm that is agnostic of any assumptions and takes all variables into account.

State-of-the-art for polymer and small molecule design by machine learning

The Aspuru-Guzik team has looked toward incorporating experimental intuition into algorithms capable of predicting a number of key parameters in the synthesis and development of small molecules—the outcome of organic reactions,[24] design of organic light emitting diodes,[25] and energy harvesting materials.[26,27] The approach of this team is focused on an inter-disciplinary cohesion of their core-competence in small molecule synthesis and quantum mechanics, along with high-throughput density functional theory (DFT) calculations of molecular descriptors,[28] enabled by the computing power of machine learning methods.[29] Their powerful data-driven method has shown tremendous promise to not only understand the inherent complex physiochemical properties of an organic molecule, but also how it could be synthesized. While the strategies exhibited in the work might be transferrable to other domains, it seems challenging when facing the complexity of polymers, particularly in structure–property relationships.

https://doi.org/10.1557/mrc.2019.54

Table I. A non-exhaustive summary of polymer relationships and models commonly studied.

Relationship	Name (if any)	Forward model?	Type
Intrinsic viscosity versus M_n	Mark–Houwink	Yes	Physical
Degree of polymerization from fractional monomer conversion for step-growth polymerization	Carothers	Yes	Kinetic
Distribution of monomers in a copolymer	Mayo–Lewis	Yes	Kinetic
T_g versus M_n	Flory–Fox	Yes	Physical/empirical
Phase behavior		No	Physical/empirical
Rheological properties		No	Physical/empirical
M_n and composition of poly (2-ethyl-2-oxazoline-*co*-2-*n*-propyl-2-oxazoline) to LCST		?	Empirical
Extent of gelation (statistical Carothers equation)		Yes	Kinetic/statistic

The first attempt toward a machine learning approach for the prediction of polymer properties was by the Ramprasad group. The polymer genome initiative utilized machine learning for the accelerated prediction of the dielectric properties of bulk polymers.[30] The approach leverages on a "fingerprinting" methodology,[31] whereby DFT calculations for the individual repeat units of the polymer are performed to understand its physical properties, and the polymer properties are derived from a statistical learning algorithm. Their latest strategy leverages at a combination atomic level, quantitative structure–property relationship (QSPR), and morphological descriptors, which is accurate in the prediction of bulk polymer properties calculated by DFT.[32] This approach was also accurate in its prediction of experimentally obtained properties such as glass transition temperature (T_g), solubility, and density, critical descriptors were simplified, which could render these techniques inaccurate when more robustly tested. An example is that for T_g, molecular mass was assumed to be infinite, rather than incorporating the additional descriptors of the Flory–Fox relationship between T_g and M_n. In a similar approach, Zeng et al. demonstrated that crystallography files may be a sufficient descriptor that provides highly accurate prediction of bandgap and dielectric constant of polymers *via* crystal graph convolutional neural networks.[33]

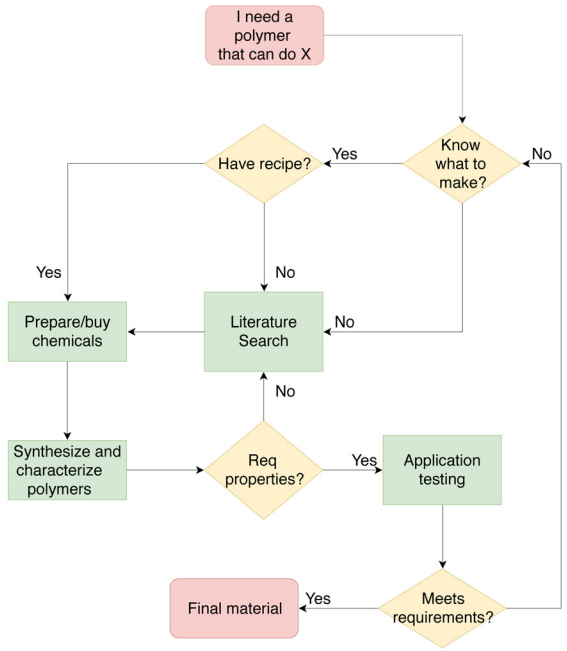
Most of these datasets were generated by high-throughput DFT calculations, with only minimal experimental validation. Wei et al. demonstrated the efficacy of neural networks from simulated data for the inherently more complex scenario of polymer phase behavior in solution.[34] Highly accurate demonstrations of machine learning on simulated data may be fallacious since both the latter and the former are computed relationships to their descriptors, leading to cyclic argument, but their efforts clearly show that machine learning could a powerful tool in polymer property prediction. However, none of these teams demonstrate inverse design where input parameters are predicted based on a desired output. The next milestone thus would be a forward predictive machine learning algorithm trained on experimental data, and the experimental validation of the inverse design of this model.

In work recently conducted by our team,[35] we demonstrate the use of a combination of machine learning methods to map a design space of composition from four possible monomers and the molecular weight of the polymer to the lower critical solubility temperature (LCST) of poly(2-oxazoline). Predictions from gradient boosting with decision trees had a mean square error of 4°C over a temperature range of 24–90°C (6% error). The fast inference speed of the machine learning algorithm allowed for the inverse design *via* particle swarm optimization. When cross validated with an ensemble of neural networks, the inverse design was capable of designing 17 unique polymers, including terpolymers, with a mean squared error of 6.1°C. These polymers were designed for four target LCSTs, with constraints such as maximum molecular weight and preferred composition imposed. This is likely to be the first of many research efforts that demonstrate the efficacy machine learning enabled polymer design. The conceptual advance of all the teams mentioned is the progression from the traditionally accepted forms of data representation toward one that is far more robust and intuitive.

The challenges in machine learning augmented polymer design

Nevertheless, this approach is not without its challenges, the greatest of which is that the application of data driven techniques to polymer science required large datasets. In the previously mentioned examples, the smallest dataset had 171 unique polymers while the rest were at least a factor of ten greater. While the training datasets could be simulated in a larger and more efficient manner by the methods suggested by the Ramprasad group,[32] manual experimentation is typically slower. The flowchart and table in Fig. 2 demonstrate a typical experimentation process from ideation to final product. The

https://doi.org/10.1557/mrc.2019.54



Activity	Total hours	Man hours
Literature search	5	5
Prepare chem icals	168	2
Synthesize and characterize polymers	50	10
Subtotal 1	223	17
Typical # of iterations	5	5
Subtotal 2	1115	85
Application testing	720	480
Subtotal 3	1835	565
Typical # of iterations	5	5
Grand total	9175 (382 days)	2825 (353 eight hour work days

Figure 2. (Left) Flowchart of a typical development process of a polymer from ideation to a final material or product; (right) typical timeline of the completion of a new polymer development process.

length of time it takes to generate one experimental data point can typically be 223 h or about 9 days (subtotal 1), while the time taken to generate one new material could take as much as 353 work days, or about 70 five-day work weeks continuously. The reason for this slow speed is attributed to the limits of the human ability. While the quantity of unique polymers might be low, as is the case in the research of most other materials, the data obtained are often rich in detail. These details comprise of descriptors and attributes—the latter a function of the former. In order to improve the efficacy of machine learning techniques on scarce datasets, two strategies—data curation and data augmentation, should be employed.

Data curation is the necessary to mitigate over-complexity over the search space and optimize the training of the forward predictive model. A prime example of the curation of descriptors was demonstrated by Zeng et al.[33] where they used crystallography data to train a graph convolutional neural network as a predictor for dielectric constant and bandgap of polymers. This was a step forward from the work by the Ramprasad group, which used three sets of descriptors to predict the same properties. Another example is where the shape of the molecular weight distribution was a determining factor in choosing the peak molecular weight rather than the number average molecular weight for thermally responsive polymers.[35] Feature importance rankings, which are typical in machine learning algorithms might also assist in determining the best descriptors for the dataset, in a bid to further refine and tune the algorithm.[36]

While experimental data are the most important, supplementing these data with those from modeling and simulations could also be valuable in providing a well-rounded dataset. However, modeling the key physical and chemical processes in polymers that directly relate to its physical property by modern molecular simulation techniques such as classical molecular dynamics simulation[37] or dissipative particle dynamics simulations[38] come with their own challenges. The length scales for such phase behavior are normally within the range of hundreds of nanometers with timescales within seconds to minutes.[13] Molecular simulations are usually capable of dealing with a simulation system composed of millions of atoms to maximum of billion atoms, equivalent to a small cubic box of water with its box length smaller than 1 μm.[39] In order to obtain usable data resembling the dynamics of the system within 1 ms of real time for conventional simulation systems as large as that with 1 million atoms equivalent to a cubic box of water molecules with box length around 20 nm, it would require expensive high-performance computers to perform simulations with hundreds and thousands of CPUs continuously for weeks or even months.[39] Therefore, while data from simulations cannot be generated rapidly for specific polymer properties such as T_g and T_c, these data, along with experimental data, could piece together a far more valuable data space than just either alone. On the other hand, promoting accuracy of molecular simulations to reach that is comparable with DFT calculations *via* machine learning-based potential and/or accelerating molecular simulations through machine learning

https://doi.org/10.1557/mrc.2019.54

approaches are on-going efforts in the molecular simulation community.

The solution to accelerate the experimentation process lies in high-throughput experimentation. In a recent review article by Oliver et al.,[40] a number of high-throughput experimentation strategies to generate large polymer libraries *via* controlled radical polymerization were summarized. The cited examples focus on the enabling capabilities that high-throughput experimentation have on experimenting over a large parameter space when understanding combinatorial polymer synthesis and its kinetics. The ability to generate vast libraries of polymers with precisely controlled variation of precursors and conditions shows how much promise there is from such a tool. There is only one cited example where high-throughput experimentation in both synthesis and application testing of polymer systems, where the efficacy in nucleic acid complexation and delivery were ascertained for core–shell polymer nanoparticles.

Opportunities

All that we have explored point to three key points that should be the tenets of machine learning augmented polymer design: (1) experimental data are the most important data, and could be generated rapidly by high-throughput experimentation; (2) this should be supplemented by simulated data; and (3) to rationalize this data, proper curation of data should be conducted. Therefore, we propose a framework toward machine learning augmented polymer design, comprising discrete process steps illustrated in Fig. 3.

The first and most important step is in data management, where historic and new data should be curated and archived in a manner that adequately captures the detail of the polymer attributes, yet is simple enough and is in a format that lends itself to computation. In this step, datasets can also be augmented by conducting well controlled experiments with diligence paid toward data acquisition.

The second step involves molecular simulations, which is capable of generating data that are not easily accessible by experiments. An example of this is the ability to provide the distribution of monomers in a statistical copolymerization. Yet another would be the physical models such as the QSPR of the repeat unit that could aid in providing a constraint to the machine learning search space.

The third step is machine learning, where the proper selection of the algorithm, along with its tuning, is the primary focus. The improper selection of algorithm or tuning parameters could lead to over-fitting—a common occurrence in relatively small datasets. For a forward predictive model, accuracy of the algorithm is critical. As discussed earlier a short inference time is critical when inverse design is desired. Forward algorithms with slow inference times might be too computationally expensive in such cases.

The final steps involve the experimental validation of the forward model, inverse design, and finally its own experimental validation. When conducted uniformly over the composition, configuration, and attribute space, experimental validation of the forward algorithm has two functions. The first is to verify the accuracy of the machine learning algorithm with a fresh set of "unseen" data. The second is to further train and improve the algorithm. Typically, a well-trained algorithm from the previous step should afford similar accuracy to the original

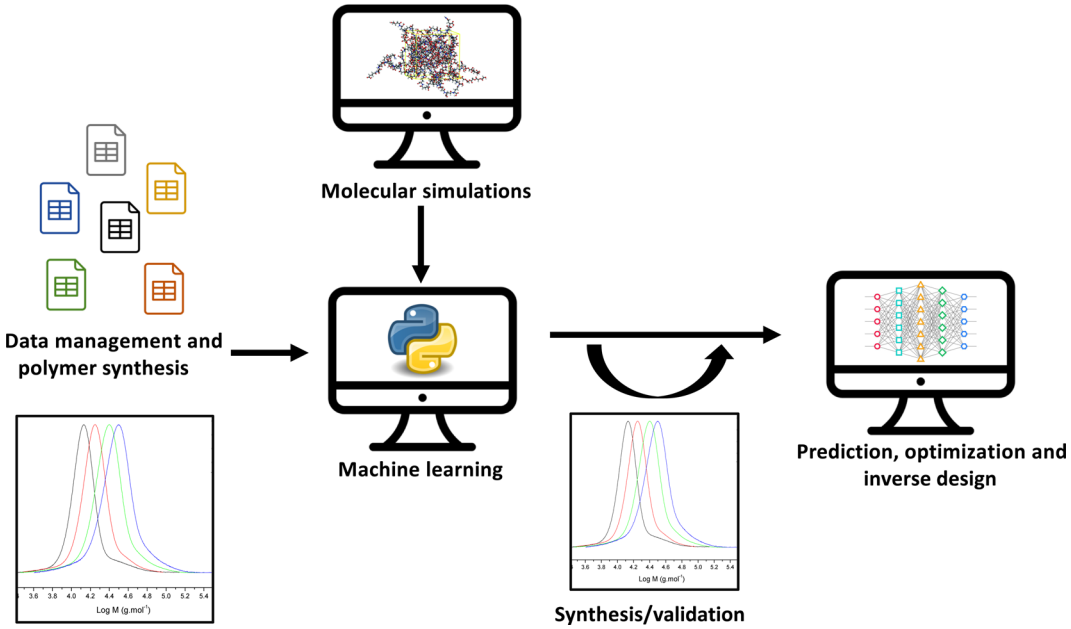


Figure 3. The proposed framework for machine learning augmented polymer design which incorporates the following steps: data management and polymer synthesis; machine learning; molecular simulations; synthesis/validation; and prediction, optimization and inverse design.

https://doi.org/10.1557/mrc.2019.54

training exercise and this new data serves to lower the total error. This loop train-validation loop should be repeated to increase the predictive accuracies of the forward models to desired levels.

Inverse design is the step where the polymer is designed based on the desired attributes. In many cases, the prediction of descriptors based on desired attributes can have numerous solutions. In order to streamline these, constraints need to be imposed on the search space. These constraints can be made up of experimental experience, domain knowledge, or physical models that can be imposed by simulated data. In addition to these, one or an ensemble of other suitable, but slower predictive algorithms could be used to down-select only the predictions with the highest level of agreement across all algorithms. This would be especially useful in cases where there is extrapolation outside of the original design space. After down-selection, the designs with the highest chance of accuracy should be synthesized to verify the accuracy of the algorithm.

Conclusion and vision

High-throughput experimentation for both synthesis and application testing, combined with a well-designed machine learning algorithm of robust architecture, high performance computing and a scalable data management strategy based on the framework presented above, could pave the way toward an autonomous lab,[8] similar to the success story shown previously for carbon nanotube growth by Nikolaev et al.[41] The combination of all the tools which are now of increasingly easy to access, would provide more value than the sum of its parts and a remarkable advance to the archaic Edisonian research methodology carried out today.

Our vision thus is to enable the translation of the present manual research process into a fully automated system requiring little to no human intervention after the experimental space has been designed. The challenge however, will be the algorithmization of experimental intuition, and the insertion of random error to allow for serendipitous discovery.

Acknowledgments

J.N.K. and Q.L. are supported by the AME Programmatic Fund from the Agency for Science, Technology and Research under Grant No. A1898b0043. The concepts put forward in this paper were developed through discussions with Prof. Tonio Buonassisi, Dr. Kedar Hippalgaonkar, and Dr. Anibal L. Gonzalez-Oyarce.

References

1. A. Gregory and M.H. Stenzel: Complex polymer architectures via RAFT polymerization: From fundamental process to extending the scope using click chemistry and nature's building blocks. *Prog. Polym. Sci.* **37**, 38 (2012).
2. S.J. Garcia: Effect of polymer architecture on the intrinsic self-healing character of polymers. *Eur. Polym. J.* **53**, 118 (2014).
3. A.C. Rinkenauer, S. Schubert, A. Traeger and U.S. Schubert: The influence of polymer architecture on in vitro pDNA transfection. *J. Mater. Chem. B* **3**, 7477 (2015).
4. A. Dag, M. Callari, H. Lu and M.H. Stenzel: Modulating the cellular uptake of platinum drugs with glycopolymers. *Polymer Chemistry* **7**, 1031 (2016).
5. D. Paramelle, S. Gorelik, Y. Liu and J. Kumar: Photothermally responsive gold nanoparticle conjugated polymer-grafted porous hollow silica nanocapsules. *Chem. Commun.* **52**, 9897 (2016).
6. J. Kumar, A. Bousquet and M.H. Stenzel: Thiol-alkyne Chemistry for the Preparation of Micelles with Glycopolymer Corona: Dendritic Surfaces versus Linear Glycopolymer in Their Ability to Bind to Lectins. *Macromol. Rapid Commun.* **32**, 1620 (2011).
7. J. Kumar, L. McDowall, G. Chen and M.H. Stenzel: Synthesis of thermoresponsive glycopolymersviacopper catalysed azide-alkyne 'click' chemistry for inhibition of ricin: the effect of spacer between polymer backbone and galactose. *Polymer Chemistry* **2**, 1879 (2011).
8. J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V.R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha and T. Buonassisi: Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule* **2**, 1410 (2018).
9. J. Bicerano: *Prediction of Polymer Properties*, (Taylor & Francis Inc, Bosa Roca, United States, 2002).
10. M.A.C. Stuart, W.T.S. Huck, J. Genzer, M. Muller, C. Ober, M. Stamm, G. B. Sukhorukov, I. Szleifer, V.V. Tsukruk, M. Urban, F. Winnik, S. Zauscher, I. Luzinov and S. Minko: Emerging applications of stimuli-responsive polymer materials. *Nat. Mater.* **9**, 101 (2010).
11. R. Jiang, Q. Jin, B. Li, D. Ding and A.-C. Shi: Phase Diagram of Poly(ethylene oxide) and Poly(propylene oxide) Triblock Copolymers in Aqueous Solutions. *Macromolecules* **39**, 5891 (2006).
12. H.S. Ashbaugh and M.E. Paulaitis: Monomer Hydrophobicity as a Mechanism for the LCST Behavior of Poly(ethylene oxide) in Water. *Ind. Eng. Chem. Res.* **45**, 5531 (2006).
13. A. Halperin, M. Kröger and F.M. Winnik: Poly(N-isopropylacrylamide) Phase Diagrams: Fifty Years of Research. *Angew. Chem. Int. Ed.* **54**, 15342 (2015).
14. R. Hoogenboom, H.M.L. Thijs, M.J.H.C. Jochems, B.M. van Lankvelt, M. W.M. Fijten and U.S. Schubert: Tuning the LCST of poly(2-oxazoline)s by varying composition and molecular weight: alternatives to poly(N-isopropylacrylamide)? *Chem. Commun.* **0**, 5758 (2008).
15. G. Odian: *Principles of Polymerization*, Fourth Edition ed. (John Wiley & Sons, New York, United States, 2004).
16. J.S. Smith, O. Isayev and A.E. Roitberg: ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci* **8**, 3192 (2017).
17. T.W. Anderson: *An Introduction To Multivariate Statistical Analysis*, (Wiley, New York, 1958).
18. G.E.P. Box and G.C. Tiao: *Bayesian Inference in Statistical Analysis*, (John Wiley & Sons, New York, United States, 2011).
19. C. Cortes and V. Vapnik: Support-Vector Networks. *Machin. Learn.* **20**, 273 (1995).
20. L. Rokach and O. Maimon: *Data Mining With Decision Trees: Theory and Applications*, (World Scientific Publishing Co., Inc.2014).
21. Y. LeCun, Y. Bengio and G. Hinton: Deep learning. *Nature* **521**, 436 (2015).
22. J.H. Friedman: Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, 1189 (2001).
23. V. Aseyev, H. Tenhu and F.M. Winnik: Non-ionic Thermoresponsive Polymers in Water, in Self Organized Nanostructures of Amphiphilic Block Copolymers II, edited by A. H. E. Müller and O. Borisov (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), pp. 29.
24. J.N. Wei, D. Duvenaud and A. Aspuru-Guzik: Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci* **2**, 725 (2016).
25. R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T.D. Hirzel, D. Duvenaud, D. Maclaurin, M.A. Blood-Forsythe, H.S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S.I. Hong, M. Baldo, R.P. Adams and A. Aspuru-Guzik: Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120 (2016).
26. F. Häse, C. Kreisbeck and A. Aspuru-Guzik: Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. *Chem. Sci* **8**, 8419 (2017).

27. S.-L. Benjamin, O. Carlos, G. Gabriel L. and A.-G. Alan: Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC), (*ChemRxiv*, 2017), p. 10.26434/chemrxiv.5309668.v3.
28. D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gomez-Bombarelli, T. Hirzel, A. Aspuru-Guzik and R.P. Adams: Convolutional networks on graphs for learning molecular fingerprints, in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (MIT Press, Montreal, Canada, 2015), pp. 2224.
29. R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams and A. Aspuru-Guzik: Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci* **4**, 268 (2018).
30. T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania and R. Ramprasad: A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
31. A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman and R. Ramprasad: Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **6**, 20952 (2016).
32. C. Kim, A. Chandrasekaran, T.D. Huan, D. Das and R. Ramprasad: Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **122**, 17575 (2018).
33. M. Zeng, J.N. Kumar, Z. Zeng, S. Ramasamy, V.R. Chandrasekhar and K. Hippalgaonkar: Graph Convolutional Neural Networks for Polymers Property Prediction. *arXiv*, **1811**.06231 (2018).
34. Q. Wei, R.G. Melko and J.Z.Y. Chen: Identifying polymer states by machine learning. *Physical Review E* **95**, 032504 (2017).
35. J. Kumar, Q. Li, K.Y.T. Tang, T. Buonassisi, A.L. Gonzalez-Oyarce and J. Ye: Machine Learning Enables Polymer Cloud-Point Engineering via Inverse Design, (*ChemRxiv*, 2018), p. 10.26434/chemrxiv.7528343.v1.
36. M.G. Luca, V. Jan, A. Emre, O. Runhai, V.L. Sergey, D. Claudia and S. Matthias: Learning physical descriptors for materials science by compressed sensing. *New Journal of Physics* **19**, 023017 (2017).
37. B. Dünweg and K. Kremer: Molecular dynamics simulation of a polymer chain in solution. *The Journal of Chemical Physics* **99**, 6983 (1993).
38. R.D. Groot and P.B. Warren: Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. *The Journal of Chemical Physics* **107**, 4423 (1997).
39. C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B.L. de Groot and H. Grubmüller: Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of Computational Chemistry* **36**, 1990 (2015).
40. S. Oliver, L. Zhao, A.J. Gormley, R. Chapman and C. Boyer: Living in the Fast Lane—High Throughput Controlled/Living Radical Polymerization. *Macromolecules* **52**, 3 (2018).
41. P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto and B. Maruyama: Autonomy in materials research: a case study in carbon nanotube growth. *Npj Computational Materials* **2**, 16031 (2016).