

# A Recommender System for Inverse Design of Polycarbonates and Polyesters

Nathaniel H. Park,<sup>\*</sup> Dmitry Yu. Zubarev,<sup>\*\*</sup> James L. Hedrick, Vivien Kiyek,<sup>†</sup> Christiaan Corbet,<sup>††</sup> and Simon Lottier<sup>†††</sup>

IBM Research–Almaden, 650 Harry Rd. San Jose, CA 95120

<sup>\*</sup>email: [npark@us.ibm.com](mailto:npark@us.ibm.com)

<sup>\*\*</sup>email: [dmitry.zubarev@ibm.com](mailto:dmitry.zubarev@ibm.com)

<sup>†</sup>Present address: Peter Grünberg Institute (PGI-9), Forschungszentrum Jülich 52425 Jülich, Germany

<sup>††</sup>Present address: Institute for Complex Molecular Systems and Laboratory of Macromolecular and Organic Chemistry, Department of Chemical Engineering and Chemistry, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

<sup>†††</sup>Present address: Arts et Métiers ParisTech – PIMM 151 bd de l'Hôpital, 75013 Paris, France

## ABSTRACT

The convergence of artificial intelligence and machine learning with material science holds significant promise to rapidly accelerate development timelines of new high-performance polymeric materials. Within this context, we report an inverse design strategy for polycarbonate and polyester discovery based on a recommendation system that proposes polymerization experiments that are likely to produce materials with targeted properties. Following recommendations of the system driven by the historical ring-opening polymerization results, we carried out experiments targeting specific ranges of monomer conversion and dispersity of the polymers obtained from cyclic lactones and carbonates. The results of the experiments were in close agreement with the recommendation targets with few false negatives or positives obtained for each class.

## Introduction

Accelerated discovery of new materials has potential to broad economic impacts arising from both decreased cost of development timelines and improved material performance.<sup>1–5</sup> To meet this challenge, traditional, labor-intensive research workflows have been redesigned to utilize a combination of automated or autonomous experimentation, accelerated characterization, high-performance computing, and artificial

intelligence to improve experimental quality and rate.<sup>6–12</sup> In this context, new automated synthesis platforms for polymeric materials are critical and offer significant benefits by enabling programmatic control over polymer properties such as monomer conversion (MC), the degree of polymerization (DP), dispersity (Đ), and resulting architecture of the polymeric materials.<sup>13–17</sup> Additionally, these systems can be merged with in-line characterization to enable real-time feedback and optimization of polymer characteristics.<sup>14,15</sup> Coupling of these systems with advances in predictive models for polymer properties will undoubtedly afford immense progress towards rapid development and commercialization of new materials with improved performance characteristics.<sup>10,11,18</sup>

However, the merging of laboratory automation and artificial intelligence does not automatically accrue or guarantee benefits from the reduction of experimental overhead. Frequently overlooked is the significant amount of background experimentation that must be completed before an automated synthetic platform can be used. For the ring-opening polymerization (ROP) of lactones and cyclic carbonates, it is imperative that the polymerization conditions of a particular monomer with a chosen ROP catalyst are investigated prior to use in an automated system—especially systems using continuous-flow. Failure to do so can lead to improper choice of conditions and residence times for the automated polymerization, leading to reactor fouling, undesired broadening of dispersity, or potentially no reaction at all.<sup>19</sup> And while failed results are necessary for the development of improved predictive models, the costs of forgoing background experimentation can be significant, particularly when custom monomers or catalysts are used.

The controlled ROP of lactones and cyclic carbonates is a result of the interplay between monomer reactivity<sup>20–23</sup> and catalyst activity.<sup>13,24–27</sup> While the thermodynamics and kinetics of ROP have been well-studied,<sup>20–23</sup> these insights do not provide a robust theoretical framework to enable *a priori* catalyst selection. Particularly since even relatively small changes in monomer structure can have an outsized influence on reactivity.<sup>20,28</sup> Instead, experimentalists must rely on these principles in addition to the available literature and their own experience to guide catalyst choice and experimental design. Design of experiments—a statistical method to help determine the influence of different experimental factors—can

be a powerful tool to help arrive at an optimized set of reaction conditions. Although this approach still necessitates a significant number of experiments to be run.<sup>29,30</sup> Consequently, in order to truly accelerate materials development in combination with using automated platforms, the amount of background experimentation must be minimized. Ultimately, developing predictive models based on historical data which can be translated into experimentally actionable hypotheses could significantly reduce the time needed to determine the requisite reaction parameters for use of new monomers with automated synthetic platforms. Thus, given the significant breadth of catalysts available for ROP,<sup>31–33</sup> a predictive framework capable of matching catalysts to lactone and carbonate monomers for controlled polymerization would both dramatically reduce experiment overhead and accelerate material development timeframes using automated synthetic platforms.

The standard computational selection strategy for evaluating monomer–catalyst pairs would rely on the direct evaluation of the thermodynamic and kinetic reaction parameters via the state-of-the-art of quantum chemistry. It is straightforward to obtain enthalpic profiles along reaction paths, albeit achieving chemical accuracy would be challenging. Transition from enthalpic to free energy profiles will face a heavy overhead of conformational sampling of the potential surface of the reaction. Correct accounting for the effect of the medium, including the solvent and the growing chain of the polymer, is a steep cost factor as well. The computational task will, effectively, blow up once it is recognized that matching a monomer and a catalyst for polymer design in fact involves selection of a monomer, an initiator, a catalyst, a co-catalyst, a solvent, their respective feed ratios, and reaction time.

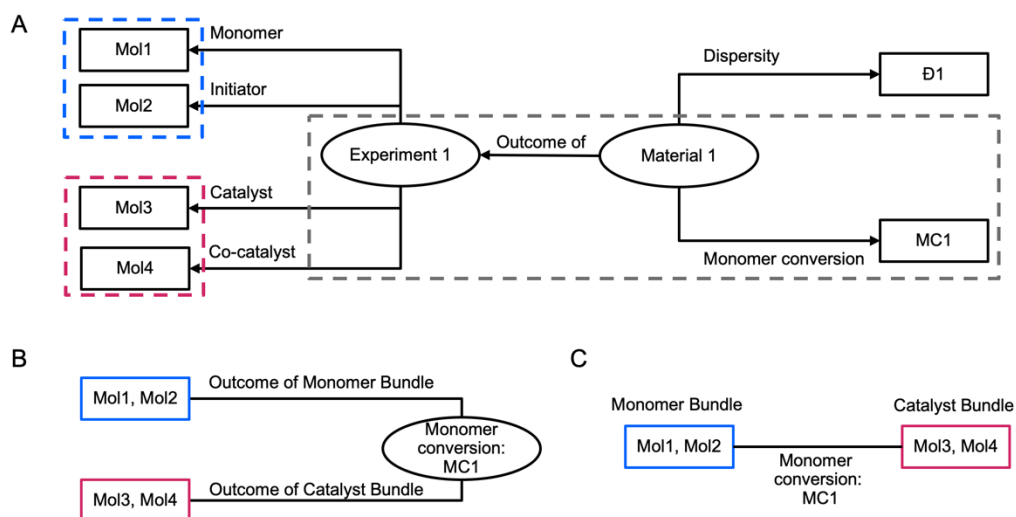
Rapid hypothesis generation can always benefit from first principle methods. However, a complementary strategy using data centric approaches rooted in statistical learning and demonstrating a different balance of advantages and disadvantage has been gaining relevance lately.<sup>34–43</sup> In the chemical domain, the bulk of the machine learning (ML) studies, including deep learning (DL) as a subset of ML, relies on the construction of structure-property and structure-activity relations informed by the structural features of molecular compounds. Polymer materials research encounters easily identifiable unique

challenges that set it apart from other chemical domains, such as drug design. Polymers are stochastic macromolecules built from well-defined, monomeric building blocks. Establishing the exact make-up of the ensemble of polymer chain—particularly those with cross-links or network interpenetration—is often impractical both from computational and experimental perspectives. Therefore, it makes sense to find alternatives to the molecular structure of the polymer chains as the source of features in statistical learning tasks involving polymer materials. Assuming that materials are produced in a deterministic manner, there is a level of consistency of the synthetic and processing conditions that guarantees a degree of reproducibility of the measurable properties of the material. For the purposes of data-centric modeling, the degrees of freedom of the experimental procedures—including synthesis and processing—should be minimally sufficient. In molecular design and discovery, models that help to identify molecules with target properties are not directly actionable—they transform the problem of finding a candidate structure into the subsequent problem of finding a way to produce this candidate structure experimentally. Transitioning from the structure-property to experiment-property models in the polymer materials domain helps to avoid this conundrum because the latter has the capability to predict experiments in terms of the control settings for the factually deployed equipment. Evaluation of the experimental conditions by means of ML/DL has been gaining momentum. The spectrum of this effort is quite broad, from the experiments on AI predicting organic synthesis,<sup>12</sup> to retrosynthetic approaches for organic molecules<sup>44–46</sup> to generative modeling of synthetic conditions for inorganics and mixed materials trained from the literature data.<sup>47,48</sup> Experiment-property models require adequate information about capabilities of the available experimental platform. Diversity of the data in published sources does not necessarily inform experimental choices for the specific performance targets of the material pursued in a given project: historical data might be obtained with different, potentially obsolete, methodologies and equipment, they also might miss critical execution details. Training data sets for the development of experiment-property models can be produced via data mining, but the most powerful aspect of experiment-property modeling is its innate capability to proceed via direct access of the data-centric model to the experimental equipment deployed in the project for the model training and target data acquisition stages.

Having identified experiment-property modeling as the overarching theme of our efforts, we should distinguish the cases of the forward and inverse material design. The forward design of polymer material encompasses models that enable one-to-one mapping of polymer structure to the property. The structure can be captured explicitly, such as list of atomic positions in 3 dimensions, or implicitly, such as parameters of the experiments producing the polymer. The inverse material design includes the models that map polymer property on the structure in one-to-many manner. In other words, given the value or the range of property values, inverse design model is expected to produce several distinguishable hypotheses about material structure.<sup>18,49–51</sup> The inverse design strategies offer higher computational efficiency of the design and discovery in terms of the computational cost and time to solution. In practice, the forward and inverse strategies should be seen as complementary in order to ensure flexibility and robustness of the computational design and discovery workflow.

We report an inverse design strategy for hypothesis generation about experiments that are likely to produce the desired outcome. Specifically, given the target range of the measurable property of a polymer material, we generate recommendations for the values of the experimental degrees of freedom, both categorical and quantitative, that are immediately actionable. The recommendation engine performs link completion on the network of the historical experiments. Preparation of the network amenable to link completion can be approached in different ways. We describe an operational approach motivated by potential utilization of the knowledge graphs (KG) in polymer materials domain.<sup>52</sup> The first step is to convert a relational database of historical experiments into resource description framework (RDF) graph (Figure 1A). We refer to this construction as “experiment knowledge graph” (eKG). Prediction of new experiments on eKG is an instance of the network completion problem<sup>53</sup> where one infers a new “experiment” node connected to a new “material” node. On the eKG, the “experiment” node on the eKG is linked to the nodes representing particular values of the control parameters and the “material” node is linked to the nodes representing particular values of measurable properties. The second step is to collapse nodes on the historical eKG forming a network with multiple partitions (Figures 1A and 1B). In the case of ROP,

we consider three partitions. The nodes in the first partition are bundles of control parameter values related to the catalyst utilization, including the catalyst and the co-catalyst (red outlines, Figures 1A and 1B). The nodes in the second partition are bundles of control parameter values related to the monomer utilization, including the monomer, the initiator, the monomer concentration ( $[M]_0$ ), and initial monomer–initiator molar ratio ( $[M]_0:[I]_0$ ) (blue outlines, Figures 1A and 1B). The nodes in the third partition are values of polymer properties (grey outlines, Figures 1A and 1B). Finally, the multi-partite graph is simplified by projecting out the partition of property values (Figure 1C). In the resulting bi-partite network two nodes are connected via a link if the experiment with the respective parameters was carried out; the outcome of the experiment, i.e., measured properties, are assigned to the link as attributes (Figure 1C). Both the original eKG and the generated bi-partite network are suitable inputs for the network completion task. Structuring the data as a bi-partite network (Figure 1C) enables network completion via both simple decomposition methods, such as non-negative matrix factorization (NMF)<sup>54</sup> and advanced representation learning approaches, such as *node2vec*<sup>55,56</sup>. The presented study reports the results obtained with continuous feature representation of the nodes in the bi-partite network (Figure 1C) obtained using *node2vec* framework.



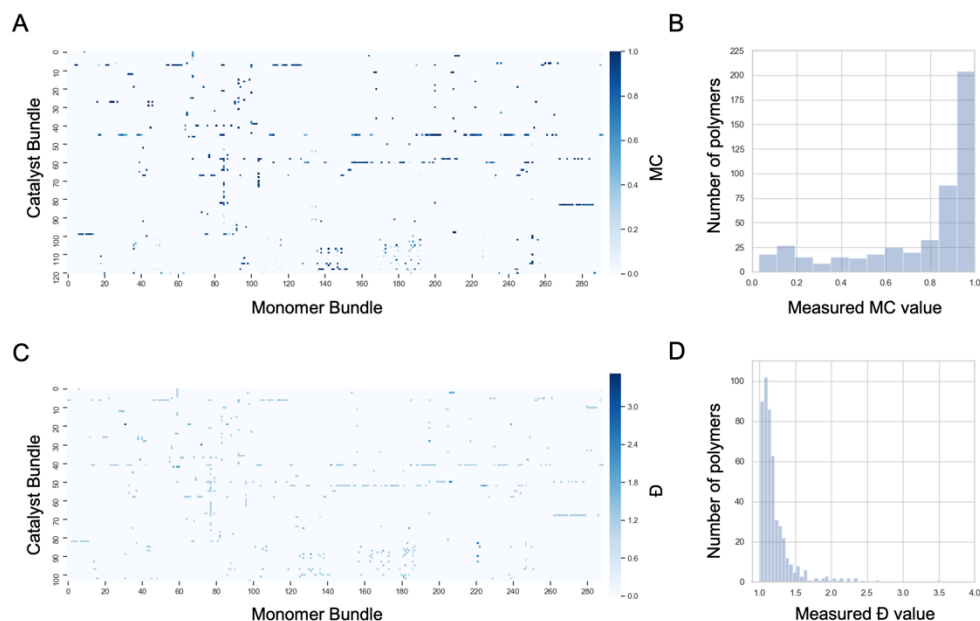
**Figure 1.** Transformation of the “experiment knowledge graph” (eKG) into a network amenable to link prediction via non-negative matrix factorization. Panel A: an RDF graph constructed from the relational database of the experimental parameters and outcomes (not shown). Continuous experimental parameters and measured properties are converted into categorical values via binning. Dashed outlines show the patterns of node collapse. Panel B: simplified RDF graph where nodes of the experimental parameters (objects) are collapsed forming bundles and the categorical value of the measured property is treated as a subject. The same

procedure applies to MC (shown) and Đ (not shown). Panel C: RDF triples are transformed into a bi-partite graph, where nodes representing bundled experimental parameters (RDF objects) are directly connected to each other via a link; the value of the measured property (RDF subject) is assigned to the link as an attribute.

## Results

Historical data prepared by the subject matter expert (SME) included 810 experiments involving total of 83 catalysts, 24 co-catalysts, 80 monomers, and 61 initiators. Additionally, experiment specification requires selection of values with at least two continuous degrees of freedom: the  $[M]_0$  and the  $[M]_0:[I]_0$ . In order to make the recommendation of the values of continuous variables tractable, we discretized historical ranges of the respective degrees of freedom, producing 21 sub-ranges, or bins, for each. Approaching combinatorial design of the future experiments by strictly recycling the historical sets of options for the aforementioned degrees of freedom, one would have to navigate a set of 4B possible experiments.

We constrained the space of recommendations by considering only those historical experiments that resulted in the successful polymerization and evaluation of the monomer conversion (741 experiment). Considering monomer conversion (MC) as the target property, the successful historical experiments involved 77 catalysts, 24 co-catalysts, 68 monomers, and 50 initiators. Only 12 sub-ranges of  $[M]_0$  and 11 ranges of  $[M]_0:[I]_0$  were utilized in these experiments. Specifically, 121 unique catalyst/co-catalyst combinations and 292 unique monomer/monomer concentration/initiator/monomer-initiator molar ratio combinations were encountered, shaping up a combinatorial space of 35K possible experiments (Figure 2A). Successful historical experiments that measured dispersity (Đ) comprised 69 catalysts, 20 co-catalysts, 67 monomers, and 53 initiators. There were 104 unique catalyst/co-catalyst combinations and 289 unique monomer/ $[M]_0$ /initiator/ $[M]_0:[I]_0$  combinations with the potential for ~30K experiments (Figure 2C).



**Figure 2.** Historical data selected to train the recommender; the target properties are MC and  $\bar{D}$ . **Panel A:** heatmap of MC values. Vertical axis represents the bundled degree of freedom includes catalyst/co-catalyst combinations designated as “Catalyst Bundle”; horizontal axis represents the bundled degree of freedom that includes monomer/monomer concentration/initiator/monomer-initiator molar ratio combinations designated as “Monomer Bundle”. **Panel B:** histogram of MC values observed in the historical data. **Panel C:** heatmap of  $\bar{D}$  data. **Panel D:** histogram of  $\bar{D}$  values observed in the historical data.

Recommendations regarding experiments that are expected to produce polymers with target properties were generated in the following manner. The target range of MC was set at 90–100%; the target range of  $\bar{D}$  was set at 1.05–1.15. These ranges were dictated by the practical constraints of the ongoing polymer synthesis efforts and consideration of balancing positive and negative classes in the recommendation task. Separate bi-partite networks were constructed for each property (cf. Figure 1C). “Monomer” and “catalyst” nodes in these networks were connected via links of weight 1.0 if the resulting experiments produced polymers with properties in the target ranges. Nodes comprising “unknown” experiments were connected via links of weight 0.5 per imputation strategies outlined in the studies of recommenders (see “Recommender Methodology” section in SI for the discussion). Finally, the nodes were disconnected if the resulting experiments produced polymers with properties outside target ranges. The recommendation workflow comprised the node embedding stage utilizing *node2vec* algorithm, the stage computing edge embedding as Hadamard product of the node embeddings, and the edge evaluation stage



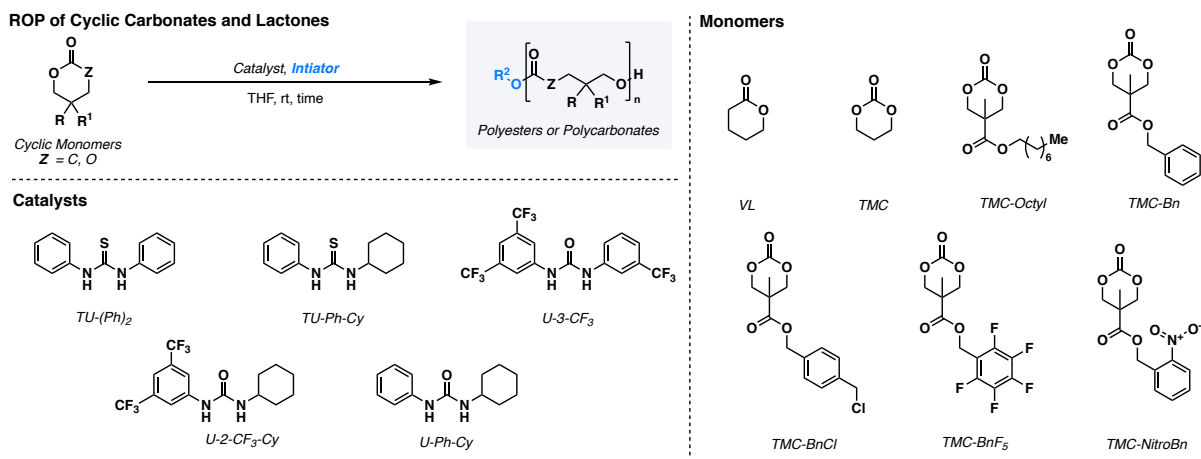
using a random forest classifier (RFC) producing probabilities of the “unknown” edges to have weight 1.0 (see Supporting Information). A cross-validation study was performed to select parameter values for the stages of the recommendation workflow and to establish expectations about its performance (see Supporting Information).

Full specifications of the recommended experiments were reviewed by the SMEs to perform the basic sanity checks and provide the initial level of the assessment of the prediction quality. We attempted to predict reaction time for each experiment using several standard regressors and treating the combination of binned continuous parameters and one-hot encoded categorical parameters as the features. In order to facilitate SMEs review and prioritization of the proposed experiments, we introduced the following novelty classes for the proposed experiments. The class of recommendations with the highest novelty comprises pairs of monomers and catalysts that have not been investigated before. Medium novelty class comprises combinations of monomer, initiator, catalyst, and co-catalyst that have not been attempted. This class comprises performance-tuning hypotheses where new initiators and co-catalyst accompany known monomer-catalyst pairs. Finally, the class with the lowest novelty of recommendations comprises combinations of monomer, initiator, catalyst, and co-catalyst that exist in historical data but new quantitative parameters, such as concentrations and ratios. Considerations of the novelty of the proposed experiments were further constrained by the availability and/or accessibility of the required monomers and catalysts. In our case, there were seven monomers and five catalysts available for the immediate recommender-driven action (Figure 3).

The recommender produced approximately 16K hypotheses above 0.51 threshold (see “Recommender Methodology” section in SI for the discussion of the threshold selection) comprising monomer, initiator,  $[M]_0$  (discretized),  $[M]_0:[I]_0$  (discretized), catalyst, and co-catalyst that are expected to produce polymers with the MC within the target range; 199 hypotheses were based on the available/preferred molecular components only and were considered actionable. Further, 52 actionable hypotheses belonged to the highest novelty class and were based on unique combinations of the molecular

components. 48 actionable hypotheses in the medium novelty class were based on unique combinations of the molecular components. In the case of  $\mathcal{D}$  as the target characteristic, approximately 16K experiments were recommended with the threshold set at 0.6 (see “Recommender Methodology” section in SI for the discussion of the threshold selection), including 238 based on the available/preferred components and were considered actionable; 61 actionable hypotheses had the highest novelty and were based on unique combinations of the molecular components. 36 actionable hypotheses had medium novelty and were based on unique combinations of the molecular components. Overall, 14 recommendations were selected by the SME for the experimental follow-up (Table 1) driven primarily by the availability of the monomers and catalysts and novelty class of the experiment. These experiments included ten unique combinations of the components; five of which were ranked as having high novelty and five combinations are medium novelty. The summary of the experimental outcomes is provided in Table 1.

**Table 1: Experimental Evaluation of Recommendations**



Entry	Novelty	Class <sup>a</sup>	Initiator	Catalyst	Co-catalyst	Monomer	Time (s)	Conv. (%) <sup>b</sup>	M <sub>n</sub> (GPC) <sup>c</sup>	$\mathcal{D}^c$	MC <sup>d</sup>	$\mathcal{D}^d$
1	High	C(+),D(+)	BnOH	U-3-CF <sub>3</sub>	KH	TMC-BnF <sub>5</sub>	180	96	11388	1.39	TP	FP
2	High	C(+),D(+)	BnOH	U-3-CF <sub>3</sub>	KH	TMC-BnF <sub>5</sub>	5	91	14288	1.07	TP	TP
3	Medium	C(+),D(+)	Pyrene-Bu-OH	U-3-CF <sub>3</sub>	KH	VL	186	97	14629	1.07	TP	TP
4	High	C(+),D(+)	BnOH	TU-(Ph) <sub>2</sub>	DBU	TMC	300	0	—	—	FP	FP

5	Medium	C(+),D(-)	KOMe	U-Ph-Cy	—	TMC-BnCl	720	97	5340	1.7	TP	TN
<b>6</b>	<b>Medium</b>	<b>C(+),D(-)</b>	<b>KOMe</b>	<b>U-Ph-Cy</b>	—	<b>TMC-BnCl</b>	<b>5</b>	<b>96</b>	<b>8729</b>	<b>2.24</b>	<b>TP</b>	<b>TN</b>
7	Medium	C(+),D(-)	KOMe	U-2-CF <sub>3</sub> -Cy	—	VL	15007	98	14672	1.77	TP	TN
<b>8</b>	<b>Medium</b>	<b>C(+),D(-)</b>	<b>KOMe</b>	<b>U-2-CF<sub>3</sub>-Cy</b>	—	<b>VL</b>	<b>1800</b>	<b>98</b>	<b>15791</b>	<b>1.62</b>	<b>TP</b>	<b>TN</b>
9	Medium	C(+),D(-)	KOMe	U-2-CF <sub>3</sub> -Cy	—	TMC	27901	99	7835	2.16	TP	TN
<b>10</b>	<b>Medium</b>	<b>C(+),D(-)</b>	<b>KOMe</b>	<b>U-2-CF<sub>3</sub>-Cy</b>	—	<b>TMC</b>	<b>600</b>	<b>99</b>	<b>8696</b>	<b>2.01</b>	<b>TP</b>	<b>TN</b>
11	High	C(+),D(-)	BnOH	TU-Ph-Cy	DBU	TMC-Octyl	225	38	—	—	FP	TN
12	High	C(-),D(+)	BnOH	U-3-CF <sub>3</sub>	DBU	TMC-Bn	60	95	9276	1.15	FN	TP
13	High	C(-),D(-)	BnOH	U-Ph-Cy	DBU	TMC-Octyl	141	20	—	—	TN	TN
14	Medium	C(-),D(-)	KOMe	U-2-CF <sub>3</sub> -Cy	—	TMC-NitroBn	190	87	2311	1.83	FN	TN

<sup>a</sup>Recommender output classes: **C(+)** — monomer conversion >90% after allotted experiment time; **C(-)** — monomer conversion <90% after allotted experiment time; **D(+)** — dispersity of resulting polymer within 1.05–1.15 range; **D(-)** — dispersity of resulting polymer greater than 1.15. Bold entries are ones that have SME adjusted reaction times. <sup>b</sup>Determined by <sup>1</sup>H NMR spectroscopy of crude reaction mixture. <sup>c</sup>Determined by gel-permeation chromatography using THF as the eluent and calibration with polystyrene standards. THF = tetrahydrofuran; KOMe = potassium methoxide; KH = potassium hydride; DBU = 1,8-diazabicyclo[5.4.0]undec-7-ene; BnOH = benzyl alcohol; Pyrene-Bu-OH = 1-pyrene butanol. <sup>d</sup>Evaluation of the predicted class labels: TP = true positive; FP = false positive, TN = true negative, and FN = false negative. See Supporting Information for details on reaction conditions and characterization data.

## Discussion

The implemented recommender system is intrinsically sensitive to the connectivity of the networks subject to *node2vec* embedding. If only successful historical experiments are considered, the bi-partite networks connecting “monomer bundle” nodes and “catalyst bundle” nodes include multiple connected components (Figure S1). This is a form of the “cold start” problem, i.e., lack of the data required to produce a recommendation. If only successful experiments are expressed as links, there are pairs of monomers and catalysts that do not have any paths between them. Our imputation strategy (expressing unknown experiments as links with weight 0.5) enables *node2vec* to carry out random walks between disconnected components and produce a unified embedding. Of course, the obtained embeddings for the nodes that belong to the different connected components lead to recommendations equivalent to random guesses which drags down the performance metrics. The presence of the multiple connected components in the network of successful experiments is easy to interpret in the context of historical patterns of data acquisition, such as low tolerance for failing experiments leading to “survivor’s bias” and “frozen accidents” in the selection of the catalytic platforms. Exposure of this feature of the historical data topology in the recommender

workflow helps to explicitly address “cold start” problem via acquisition of the critically missing experimental data.

The experiments from the recommendation engine fell into two outcome-based classes based on the Đ or MC. Within each class, the experiment was assigned a positive or negative label based on whether its outcome will be within the targeted ranges for each class (>90 MC and Đ between 1.05–1.15). For each experiment selected for validation, the conditions were tested as given (see Table S1 for additional recommended parameters), meaning no additional fine-tuning of reaction parameters based on the knowledge of SME were performed. In select cases where the provided reaction time was determined by the SME be longer than necessary, a second experiment was performed at a shorter reaction time based on the SME’s estimation (entries 2, 5, 7, and 14; Table 1).

The validation experimental results are summarized in Table 1 and are highly encouraging. Out of the 10 recommended experiments and 40 potential outcomes (positive or negative within each class), only five experiments had outcomes containing either a false positive or negative. Notably, in the four cases where a second experiment was run with the SME adjusted reaction time, the one with a false positive outcome (entry 1, Table 1) could be changed into a true positive (entry 2, Table 1) and for the others true positive outcomes were maintained. It is also important to note that the experimental results are consistent with general reactivity trends for the ROP catalysts and monomers utilized, in particular for combinations classified as having high novelty. For example, the high conversion and broad dispersity resulting from using U-Ph-Cy catalyst with TMC-BnCl (entries 5 and 6, Table 1) is clearly a mismatched catalyst–monomer pair at the given reaction time and the obtained results are not unexpected from the perspective of a SME given the reported trends.<sup>24,25</sup> Together, the experimental validation results demonstrate the overall robustness of the recommendations, with recommender accurately providing true positive results across all Đ and MC class combinations.

## Conclusion

In this work we demonstrated that a recommender system constructed from sparse historical experimental data can accurately provide recommendations to match catalysts with monomers for ROP. The process starts with curating a suitable dataset as well as selecting the target property class and proceeds to infer the parameters of the experiments that are expected to produce polymers with the target property. The pattern of our recommendation workflow matches the pattern of the inverse design strategies actively pursued in the context of “structure–property” relationship (where “property–structure” is the inverse of “structure–property”). By exploring “experimental parameters–property” paradigm of discovery, we ensure that the computational predictions are “embarrassingly actionable”, as in immediately transferrable into the experimental phase on the factually available equipment under realistic experimental conditions. This approach is more streamlined compared to typical protocols that rely on a SME to *de novo* construct a polymerization experiment using experience and knowledge by matching a ROP catalyst, monomer, and reaction conditions, particularly in cases of high novelty. Overall, this work highlights the potential for AI systems informed by historical data to reduce experimental workloads by enabling focus towards desired outcome classes. Elaboration of these systems and their combination with automated platforms will significantly enhance and accelerate materials development endeavors.

## **Acknowledgements**

We thank Dr. Tim Erdmann (IBM) and Dr. Pedro Arrechea (IBM) for helpful discussions and assistance in editing this manuscript.

## References

1. Correa-Baena, J.-P. *et al.* Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule* **2**, 1410–1420 (2018).
2. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.* anie.201909987 (2020) doi:10.1002/anie.201909987.
3. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part II: Outlook. *Angew. Chem. Int. Ed.* anie.201909989 (2020) doi:10.1002/anie.201909989.
4. MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *ArXiv190605398 Cond-Mat Physicsphysics* (2020).
5. Häse, F., Roch, L. M. & Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* (2019) doi:10.1016/j.trechm.2019.02.007.
6. Caramelli, D. *et al.* Networking chemical robots for reaction multitasking. *Nat. Commun.* **9**, 3406 (2018).
7. Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 15733 (2017).
8. Bai, Y. *et al.* Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. *J. Am. Chem. Soc.* **141**, 9063–9071 (2019).
9. Kearsey, R. J., Alston, B. M., Briggs, M. E., Greenaway, R. L. & Cooper, A. I. Accelerated robotic discovery of type II porous liquids. *Chem. Sci.* **10**, 9454–9465 (2019).
10. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **6**, 20952 (2016).

11. Chen, G. *et al.* Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges. *Polymers* **12**, 163 (2020).
12. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
13. Lin, B., Hedrick, J. L., Park, N. H. & Waymouth, R. M. Programmable High-Throughput Platform for the Rapid and Scalable Synthesis of Polyester and Polycarbonate Libraries. *J. Am. Chem. Soc.* **141**, 8921–8927 (2019).
14. Rubens, M., Vrijsen, J. H., Laun, J. & Junkers, T. Precise Polymer Synthesis by Autonomous Self-Optimizing Flow Reactors. *Angew. Chem. Int. Ed.* **58**, 3183–3187 (2019).
15. Rubens, M. & Junkers, T. Comprehensive control over molecular weight distributions through automated polymerizations. *Polym. Chem.* **10**, 6315–6323 (2019).
16. Zhou, Y., Gu, Y., Jiang, K. & Chen, M. Droplet-Flow Photopolymerization Aided by Computer: Overcoming the Challenges of Viscosity and Facilitating the Generation of Copolymer Libraries. *Macromolecules* **52**, 5611–5617 (2019).
17. Walsh, D. J. & Guironnet, D. Macromolecules with programmable shape, size, and chemistry. *Proc. Natl. Acad. Sci.* **116**, 1538–1542 (2019).
18. Kumar, J. N. *et al.* Machine learning enables polymer cloud-point engineering via inverse design. *Npj Comput. Mater.* **5**, 1–6 (2019).
19. Reis, M. H., Varner, T. P. & Leibfarth, F. A. The Influence of Residence Time Distribution on Continuous-Flow Polymerization. *Macromolecules* **52**, 3551–3557 (2019).

20. Olsén, P., Odelius, K. & Albertsson, A.-C. Thermodynamic Presynthetic Considerations for Ring-Opening Polymerization. *Biomacromolecules* **17**, 699–709 (2016).
21. Duda, A. & Kowalski, A. Thermodynamics and Kinetics of Ring-Opening Polymerization. in *Handbook of Ring-Opening Polymerization* 1–51 (John Wiley & Sons, Ltd, 2009).  
doi:10.1002/9783527628407.ch1.
22. Duda, A., Kowalski, A., Libiszowski, J. & Penczek, S. Thermodynamic and Kinetic Polymerizability of Cyclic Esters. *Macromol. Symp.* **224**, 71–84 (2005).
23. Olsén, P., Undin, J., Odelius, K., Keul, H. & Albertsson, A.-C. Switching from Controlled Ring-Opening Polymerization (cROP) to Controlled Ring-Closing Depolymerization (cRCDP) by Adjusting the Reaction Parameters That Determine the Ceiling Temperature. *Biomacromolecules* **17**, 3995–4002 (2016).
24. Lin, B. & Waymouth, R. M. Urea Anions: Simple, Fast, and Selective Catalysts for Ring-Opening Polymerizations. *J. Am. Chem. Soc.* **139**, 1645–1652 (2017).
25. Lin, B. & Waymouth, R. M. Organic Ring-Opening Polymerization Catalysts: Reactivity Control by Balancing Acidity. *Macromolecules* **51**, 2932–2938 (2018).
26. Pothupitiya, J. U., Hewawasam, R. S. & Kieseewetter, M. K. Urea and Thiourea H-Bond Donating Catalysts for Ring-Opening Polymerization: Mechanistic Insights via (Non)linear Free Energy Relationships. *Macromolecules* **51**, 3203–3211 (2018).
27. Spink, S. S., Kazakov, O. I., Kieseewetter, E. T. & Kieseewetter, M. K. Rate Accelerated Organocatalytic Ring-Opening Polymerization of L-Lactide via the Application of a Bis(thiourea) H-bond Donating Cocatalyst. *Macromolecules* **48**, 6127–6131 (2015).



28. Weilandt, K. D., Keul, H. & Höcker, H. Synthesis and ring-opening polymerization of 2-acetoxymethyl-2-alkyltrimethylene carbonates and of 2-methoxycarbonyl-2-methyltrimethylene carbonate; a comparison with the polymerization of 2,2-dimethyltrimethylene carbonate. *Macromol. Chem. Phys.* **197**, 3851–3868 (1996).
29. Bowden, G. D., Pichler, B. J. & Maurer, A. A Design of Experiments (DoE) Approach Accelerates the Optimization of Copper-Mediated <sup>18</sup>F-Fluorination Reactions of Arylstannanes. *Sci. Rep.* **9**, 11370 (2019).
30. Shear, T. A., Lin, F., Zakharov, L. N. & Johnson, D. W. “Design of Experiments” as a Method to Optimize Dynamic Disulfide Assemblies: Cages and Functionalizable Macrocycles. *Angew. Chem. Int. Ed.* **59**, 1496–1500 (2020).
31. Zhang, X., Fevre, M., Jones, G. O. & Waymouth, R. M. Catalysis as an Enabling Science for Sustainable Polymers. *Chem. Rev.* **118**, 839–885 (2018).
32. Kamber, N. E. *et al.* Organocatalytic Ring-Opening Polymerization. *Chem. Rev.* **107**, 5813–5840 (2007).
33. Kieseewetter, M. K., Shin, E. J., Hedrick, J. L. & Waymouth, R. M. Organocatalysis: Opportunities and Challenges for Polymer Synthesis. *Macromolecules* **43**, 2093–2107 (2010).
34. Kuhn, C. & Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **100**, 10595–10599 (1996).
35. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).

36. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
37. Brocchini, S., James, K., Tangpasuthadol, V. & Kohn, J. A Combinatorial Approach for Polymer Design. *J. Am. Chem. Soc.* **119**, 4553–4554 (1997).
38. Kumar, J. N., Li, Q. & Jun, Y. Challenges and opportunities of polymer design with machine learning and high throughput experimentation. *MRS Commun.* **9**, 537–544 (2019).
39. Paradiso, S. P., Delaney, K. T. & Fredrickson, G. H. Swarm Intelligence Platform for Multiblock Polymer Inverse Formulation Design. *ACS Macro Lett.* **5**, 972–976 (2016).
40. de Almeida, A. F., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
41. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
42. Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
43. Hautier, G., Jain, A. & Ong, S. P. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.* **47**, 7317–7340 (2012).
44. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).
45. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).

46. Nair, V. H., Schwaller, P. & Laino, T. Data-driven Chemical Reaction Prediction and Retrosynthesis. *Chim. Int. J. Chem.* **73**, 997–1000 (2019).
47. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
48. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *Npj Comput. Mater.* **3**, 1–9 (2017).
49. Takeda, S. *et al.* AI-driven Inverse Design System for Organic Molecules. *ArXiv200109038 Cs* (2020).
50. Takeda, S. *et al.* Molecular Inverse-Design Platform for Material Industries. *ArXiv200411521 Phys.* (2020).
51. Freeze, J. G., Kelly, H. R. & Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **119**, 6595–6612 (2019).
52. Zubarev, D. Y. & Pitera, J. W. Cognitive Materials Discovery and Onset of the 5th Discovery Paradigm. in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions* vol. 1326 103–120 (American Chemical Society, 2019).
53. Kim, M. & Leskovec, J. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. in *Proceedings of the 2011 SIAM International Conference on Data Mining* 47–58 (Society for Industrial and Applied Mathematics, 2011).  
doi:10.1137/1.9781611972818.5.
54. Chen, B., Li, F., Chen, S., Hu, R. & Chen, L. Link prediction based on non-negative matrix factorization. *PLOS ONE* **12**, e0182968 (2017).

55. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (Association for Computing Machinery, 2016).  
doi:10.1145/2939672.2939754.
56. Palumbo, E. *et al.* Knowledge Graph Embeddings with node2vec for Item Recommendation. in *The Semantic Web: ESWC 2018 Satellite Events* (eds. Gangemi, A. *et al.*) 117–120 (Springer International Publishing, 2018). doi:10.1007/978-3-319-98192-5\_22.