The Block Copolymer Phase Behavior Database

Nathan J. Rebello,[1] Akash Arora,[1] Hidenobu Mochigase,[1] Tzyy-Shyang Lin,[1] Jiale Shi,[1] Debra J. Audus,[2] Eric S. Muckley,[3] Ardiana Osmani,[1] and Bradley D. Olsen[1]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[2]Materials Science and Engineering Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899, United States

[3]Citrine Informatics, Redwood City, CA, United States

Corresponding Author: Bradley D. Olsen, E-mail: bdolsen@mit.edu

Abstract

The Block Copolymer Database (BCDB) is a platform that allows users to search, submit, visualize, benchmark, and download experimental phase measurements and their associated characterization information for di- and multi-block copolymers. To our knowledge, there is no widely accepted data model for publishing experimental and simulation data on block copolymer self-assembly. This proposed data schema with traceable information can accommodate any number of blocks and at the time of publication contains over 5,400 block copolymer total melt phase measurements mined from literature and manually curated and simulation data points of the phase diagram generated from self-consistent field theory that can rapidly be augmented. This database can be accessed on the Community Resource for Innovation in Polymer Technology (CRIPT) web application and the Materials Data Facility. The chemical structure of the polymer is encoded in BigSMILES, an extension of the Simplified Molecular-Input Line-Entry System (SMILES) into the macromolecular domain, and the user can search repeat units and functional groups using SMARTS search syntax (SMILES Arbitrary Target Specification). The user can also query characterization and phase information using the Structured Query Language (SQL) and download custom sets of block copolymer data to train machine learning models. Finally, a

protocol is presented in which GPT-4, an AI-powered large language model, can be used to rapidly screen and identify block copolymer papers from literature using only the abstract text and determine whether they have BCDB data, allowing the database to grow as the number of published papers on the world wide web increases. The F1-score for this model is 0.74. This platform is an important step in making polymer data more accessible to the broader community.

1. Introduction

The rich and elaborate phase behavior of block copolymers, composed of at least two chemically distinct polymeric blocks, has attracted extensive interest in applications including drug delivery,[1, 2] surfactants,[3] photovoltaics and fuel cells,[4] asphalt modifiers,[5, 6] adhesives,[7] thermoplastic elastomers (TPEs),[8] and lithography.[9] It is well-known that block copolymers can microphase separate into periodically arranged spheres, cylinders, lamellar, and gyroid structures in order to minimize unfavorable energy contacts between thermodynamically incompatible blocks.[10-16] It is critical that models accurately predict the phase formed so that scientists and engineers can design and manipulate phases with a minimal experimental burden.

To address this need, physics-based mean-field modeling and self-consistent field theory (SCFT) apply coarse-grained approaches to predicting the equilibrium and non-equilibrium thermodynamics that drive phase behavior. In these theories, four parameters govern phase behavior: the incompatibility between blocks, volume fraction, polymer architecture, and size. For models of simple AB diblock copolymer melts, these parameters are captured by the Flory-Huggins χ quantifying monomer incompatibility, volume fraction of a single block, f_A, number of statistical segments, N, and Kuhn segment lengths of both blocks, b_A and b_B. Despite a number of seminal advances[10-16] and impressive success at making qualitative predictions and providing physical insight into the design of systems, coarse-grained theories remain unable to quantitatively

predict the variation between different block copolymer systems as chemistry is changed. Asymmetry in experimental phase diagrams often cannot be predicted accurately, and χ is a difficult phenomenological parameter to accurately determine, showing a complex dependence on temperature, composition and chemistry.[17-20] One approach to overcome these shortcomings would be to combine data-driven models with theory-driven models in order to improve predictive capability.

Data-driven modeling in materials science and materials informatics shows great promise due to the exponential growth in published data from experiments and simulations and the proven ability for machine learning models to advance theory and simulation, property prediction in soft matter, and materials design.[21-31] There are several notable polymer databases that have substantial volume. These include the Materials Data Facility,[32, 33] PoLyInfo,[34] Polymer Property Predictor and Database,[35] NanoMine,[36] Chemical Retrieval on the Web (CROW),[37] and Polymer Genome.[38] Data has also been published in several textbooks and handbooks.[39-43] Furthermore, a comprehensive user-friendly polymer representation system to log polymer characterization data has recently been developed.[44]

However, consolidating accurate, precise, and properly contextualized experimental information from literature into a polymer database for data-driven modeling remains challenging. While decades of block copolymer research have led to an impressive assemblage of experimental knowledge, its diffuse state in literature currently makes the consideration of data-driven modeling approaches difficult. Automated extraction and natural language processing from polymer literature are difficult because data tends to be in many formats that are not easily machine-accessible. Moreover, manual extraction is laborious, time-intensive, and prone to mistakes. Even with robust methods for extracting data, measurement context and raw data are not often published,

imposing barriers to model-building and benchmarking against literature data.[45-47] In response to the challenges of consolidating data, high throughput computational data generation has become instrumental to model building,[24, 48] and ingenious approaches to generative models have emerged.[49] Though models trained on these types of data can provide valuable insights, large collections of experimental measurements are invaluable to accelerating innovation in materials informatics.

To address these challenges, experimental data has been mined and consolidated for block copolymer melts from approximately 130 peer-reviewed journal articles into a single platform called the Block Copolymer Database (BCDB). While the database focuses on diblocks to build a critical mass of data, expandability to multiblocks (triblocks, tetrablocks, hexablocks, undecablocks) is demonstrated. The mining process is described in Section 2.1, and a comprehensive data schema with traceable information is presented to organize experimental and simulation information. Challenges are addressed related to quality control, lack of measurement context, inference of missing information, and crowdsourcing that will be valuable towards future database construction in polymer informatics. This platform is designed so that scientists and engineers can submit phase behavior data with quality control when they publish, view updated copies of the database, quickly visualize and download, benchmark their own experiments, and compare and validate physics and chemistry-based models.

First, the construction of the database is described, including how chemistry, characterization results, and phase information is logged. Then, the contents of the database at the time of publication are described. A method is introduced to accurately search stochastic polymer graphs complemented with SQL so that the user can search and visualize the data by chemistry, characterization, and structure measurements. Finally, a protocol is introduced that uses GPT-4 to

rapidly screen papers using the abstract text and determine if the paper has data that can be added to BCDB, allowing the database to grow at scale as the number of published papers increases. Though this database stores information for block copolymer melts at the time of publication, the database schema was designed to be flexible so that it can potentially expand to store information for more complex systems, as mentioned in the Discussion.

2. Methods

2.1. Data Curation

The BCDB was constructed by manual extraction of block copolymer phase behavior data from the literature. At first, the search was restricted to pure diblock copolymers (no additives or blends with solvent, homopolymers, or nanoparticles), with functionality later added to accommodate multiblock copolymers. Data was extracted from comprehensive and well-cited works from research groups studying the experimental self-assembly of block copolymers, predominantly in the polymer physics and materials science communities, using scholarly literature search engines like Google Scholar and Web of Science (rather than solely focusing on citations and research group, the authors relied on domain knowledge to determine whether the data was credible). However, these works tended to focus on a relatively narrow space of polymer chemistry. Therefore, a concerted effort was subsequently made to expand the chemical diversity of the database by combing the literature for less-studied systems (Figures 1-3 in Section 2 reveal the breakdown in phase measurements, characterization, and chemistry). Most publications in this database contain a relatively low density of phase behavior data but focused on chemistries that are much less studied. From these articles, data entries are extracted from tables, text, and figures.

Each entry in the database must contain the block copolymer chemistry, overall number-average molar mass (M_n), the volume fraction of the blocks, temperature, and one or two measured

phases for points within a single-phase region or on a phase boundary, respectively. Molar masses were accepted from measurements using membrane osmometry, light scattering, size exclusion chromatography (SEC) or gel permeation chromatography (GPC), and nuclear magnetic resonance (NMR) spectroscopy. Volume fraction is usually calculated and published in the articles using the measured individual block molar masses and specific volumes of homopolymers. Microdomain structures are most commonly deduced and analyzed through small-angle X-ray or neutron scattering (SAXS or SANS), transmission electron microscopy (TEM), and rheology. It was assumed that the appropriate measurement techniques were employed for determining polymer characterization and phase behavior, and the methods are logged in the database for the user; hence, characterization data were not modified during extraction.

Data collected from X-ray and neutron scattering, depolarized light scattering,[50, 51] and rheology are often acquired for the same polymer as a function of temperature, imparting structure to the overall data set where there are clusters of data points with the same M_n and f_A at varying temperature, enabling rapid exploration of wider ranges of parameter space. For example, in rheological or depolarized light scattering temperature scans, changes in morphology are generally accompanied by changes in mechanical or optical properties during heating or cooling, which can signify an order-to-order or order-to-disorder (ODT) transition. These scans and scattering experiments can sometimes reveal the coexistence of ordered and disordered structures (depending on the temperature resolution and molar mass distribution). Importantly, only data from the heating curves is extracted because for phase transitions, in particular the ODT, heating or melting a phase is expected to yield an ODT value closer to the true value, in contrast to the cooling cycle which often exhibits significant supercooling due to high nucleation barrier to form ordered phases.

Extracting data from figures is a crucial step towards building the database. Though many articles report temperature scans of scattering measurements, much of the data is extracted from rheology curve figures due to the large number of rheological measurements reported for each chemistry. Data is extracted from scan images using WebPlotDigitizer.[52] A scatter plot image is uploaded, axes are calibrated, points are classified by phase, and the data is downloaded and added to the database. Transition regions are all data points in the temperature range in which the mechanical or optical properties change, though there must be evidence from other techniques that this in fact a transition region for the data to be accepted into the database. If data points in a figure are too close to discriminate by temperature because temperature increments are below 1 °C, then temperatures are logged in increments of 1 °C from the minimum to the maximum temperature in the plot. Though initially data extraction from images was done manually, it is anticipated that automated data mining systems from scatter plots[53] and other data charts will be crucial, and training these algorithms to associate sets of data in temperature scans with phases and characterization information in the text will have invaluable impact. However, this will require domain knowledge of the system studied.

From sources other than rheology, the relationship between the data and the temperature can be more complex because measurements are often not performed *in situ*. When the measurement temperature at which the morphology observation was conducted is reported, this temperature is logged. The annealing temperature, the temperature usually just below the ODT and above the glass transition temperature, can also be logged. As block copolymer morphologies are known to reconstruct in directed self-assembly applications in thin films, only bulk morphologies are reported.

For many studies, one or more of the parameters required for the database are not directly reported but may be inferred from reported data. In particular, many M_n and f_A values are inferred because many articles do not explicitly report this information. If the dispersity (Đ) and mass average molar mass are reported (M_w), then number average molar mass (M_n) can be computed as:

$$M_n = \frac{M_w}{Đ} \tag{1}$$

If the mass fraction (w_i) and the homopolymer density (ρ_i) for block *i* are reported, then the volume fraction for any arbitrary block *i* in an *n*-block chemistry is:

$$f_i = \frac{\frac{w_i}{\rho_i}}{\sum_n \left(\frac{w_n}{\rho_n}\right)} \tag{2}$$

In Equation 2, the densities that are not reported in the article are extracted from the Polymer Database. (it is assumed that density is weakly dependent on temperature, and thus densities extracted from the Polymer Database are not always at the same temperature at which the melt phases are examined).[37] In the future, densities could also be estimated using group contribution theory for polymers not available in the database.[43] Furthermore, if the overall number of segments N are reported along with a reference volume v_{ref}, with the default set to 100 Å^3 per segment (noted in the *notes* column), then for any *n*-block melt:

$$M_n = N v_{ref} N_{Avg} \sum_n (\rho_n f_n) \tag{3}$$

where N_{Avg} is Avogadro's number.

BCDB can also supports computational, simulation, and theoretical data, with extra columns for parameters that describe blocks like the statistical segment length and for parameters that describe block interactions like the Flory-Huggins interaction parameter. Self-consistent field

theory can be used to predict the phase behavior of a block copolymer melt as a function of the volume fraction and segregation strength χN, where χ is the Flory-Huggins interaction parameter.[16] Over 100 data points of the transition boundaries of the phase diagram have been generated and input into BCDB, but more data points can rapidly be generated and added.

2.2. Data Schema

Table 1. BCDB Data Schema

Category	Column	Meaning	Data Type or Units
Quality Control	ORCID	digital identifier for article authors	string
	ORCID_sub	digital identifier for submitter	string
	DOI	digital object identifier for article	string
System	phase	dictionary of phases	string
	phase_method	method for phase measurement	string
	T	temperature for visualization	Celsius
	T_meas	measurement temperature	Celsius
	T_anneal	annealing temperature	Celsius
	T_alt	alternate or inferred temperature	Celsius
	T_describe	description of T_alt	string
	notes	notes from article or logger	string
Overall Polymer [a]	BigSMILES	structure notation with end groups	string
	Mn	number-average molar mass	g/mol
	Mw	mass average molar mass	g/mol
	D	dispersity	double
	N	number of segments	double
	chiN	segregation strength	double
Individual Block	name	dictionary of abbreviations	string
	Mn	number-average molar mass	g/mol
	Mw	mass average molar mass	g/mol
	D	dispersity	double
	N	number of segments	double
	b	statistical segment length	double
	f	volume fraction of individual block	double
	f_tot	total volume fraction in multiblock	double
	w	mass fraction	double
	rho	density	g/mL
Uncertainty [b]	std	standard deviation	double
	se	standard error	double
	unc_other	alternate uncertainty metric	double
	unc_description	description of alternate uncertainty metric	string

[a] For entries *Mn* through *N* in the "Overall Polymer" category and *Mn* through *rho* in the "Individual Block" category, there is a method entry. [b] For every quantitative column in the database, there are four uncertainty columns.

Table 1 displays a list of all columns in BCDB categorized into quality control, system, overall polymer, and individual blocks. The submitter's Open Researcher and Contributor ID (ORCID) and the manuscript digital object identifier (DOI) are required for quality control, clearly identifying the provenance of data and maintaining peer review as a key quality control metric. The ORCID column can have multiple IDs for each co-author and the principal investigator in the article. The ORCID_sub column can have multiple IDs for each submitter to the database, and the submitters can be different from the authors of the article. Entries that characterize the system are the phase(s) measured, measurement technique (phase_method) (at the time of publication, the strings in this column are "SAXS", "SANS", "TEM", "rheology", "DPLS", "AFM/SFM"), temperature, and any notes written by the submitter. The notes column can contain short quotations from the article specifying how the phase was measured, assumptions made regarding the phase behavior, evidence from SAXS, SANS, TEM, or another measurement technique that a set of data points in temperature scans can be mapped to a measured phase, annealing conditions, and insights into why expected and unexpected phase behavior are observed. Since annealing and measurement temperatures are both critical to the measured morphology, these entries (columns T_meas and T_anneal) are distinguished. If neither is reported, then an alternate temperature can be logged (T_alt) with a description (T_describe). For example, room temperature, assumed to be 25°C, can be logged in the T_alt column. The entry T is a required entry for visualization purposes; it stores a single temperature (either measured or annealing, if neither is stated, then T_alt).

Entries that characterize the overall polymer are the BigSMILES string, molar mass (M_n or M_w), dispersity (Đ), and the total number of segments (N). The *BigSMILES* column contains the polymer line notation string[54] that extends SMILES[55] to encode polymers as ensembles of molecular graphs. The BigSMILES string can include end groups (initiators and terminating

groups) if they are reported, allowing the user to explore these effects on phase behavior. Following entries for the overall polymer are sets of entries for each block in the order in which they are written in the BigSMILES. BCDB accommodates any number of blocks: a diblock would have two sets of parameters, and a triblock would have three. The user can enter volume fractions for each individual block in the di- or multiblock in the *f* entry. Alternatively, the user can enter the total volume fraction of a block if it appears multiple times in the polymer in the *f_tot* entry, like poly(styrene) in poly(styrene)-*b*-poly(isoprene)-*b*-poly(styrene)-*b*-poly(ethylene oxide). For each column in Table 1 with a numerical data type, the user has the option to specify any of the four entries under "Uncertainty": standard deviation, standard error, different metric, and string description of that metric. For example, M_n has four extra entries for the user to specify uncertainty. The columns for *names* and *phases* have dictionaries specified by BCDB for easy, standardized search (see Supporting Information for these dictionaries which show some but not all of the phase and block information). As the database grows and new chemistries and phases are added, the curator will add to the dictionary accordingly.

2.3. Quality Control

Enforcing quality control is important to addressing the challenge of collecting high-quality, high-volume data, especially when data are not reported with uncertainty and can come from a variety of experimental techniques. The data must be from a peer-reviewed manuscript with a DOI, and the submitter of the data must submit a verifiable ORCID for digital identification. Furthermore, a series of checks are applied by the database curator before each data point is added to the database. For many systems, the phase can be checked with SCFT prediction (already in BCDB) to determine whether it is reasonable given the molecular design of the polymer. Data for the ODT or OOT that is extracted from tables should agree with data extracted from figures,

validating that the axes for extracting data from figures using WebPlotDigitizer are calibrated correctly. Experimental characterization methods are generally, but not always, prioritized over calculations. For instance, molar mass reported by SEC and NMR is prioritized over molar mass predicted from the reaction stoichiometry; this does not imply that one is more accurate than the other, so the DOI is provided to enable the user to access more complete information on methods. When submitting new information into the database, the user should (and the database curator will) check to ensure that these quality criteria are met. To be accepted, all data in the BCDB must be peer-reviewed; however, users of the data are free to build their own tools to detect and interpret outliers in the data.

3. Results and Discussion

3.1. BCDB Content

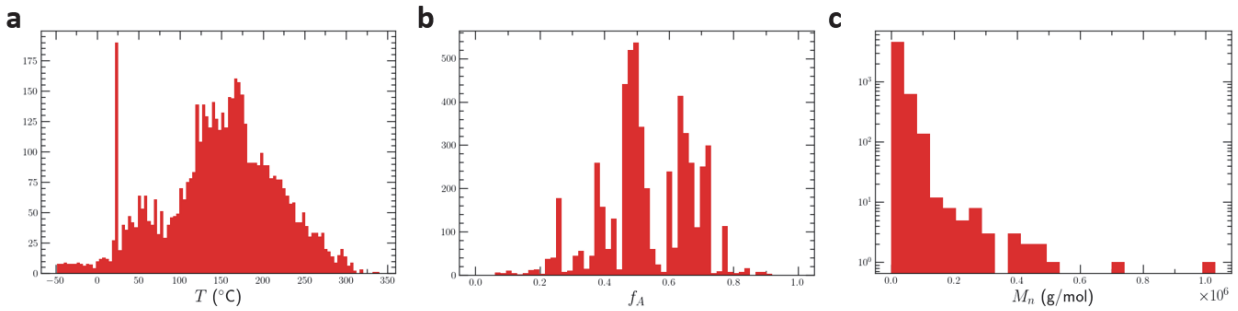


Figure 1. Histograms of the database for (a) $T(°C)$ with 100 bins, (b) f_A with 50 bins, and (c) M_n with 25 bins (y-axis is log-scale). The temperature distribution has a peak at $120 \leq T(°C) \leq 180$, above the glass transition temperature of many blocks, and a sharp peak at room temperature. The block fraction has peaks at 0.5 and 0.65 for lamellar and cylinder phases. Most polymers have $M_n < 200{,}000$ g/mol.

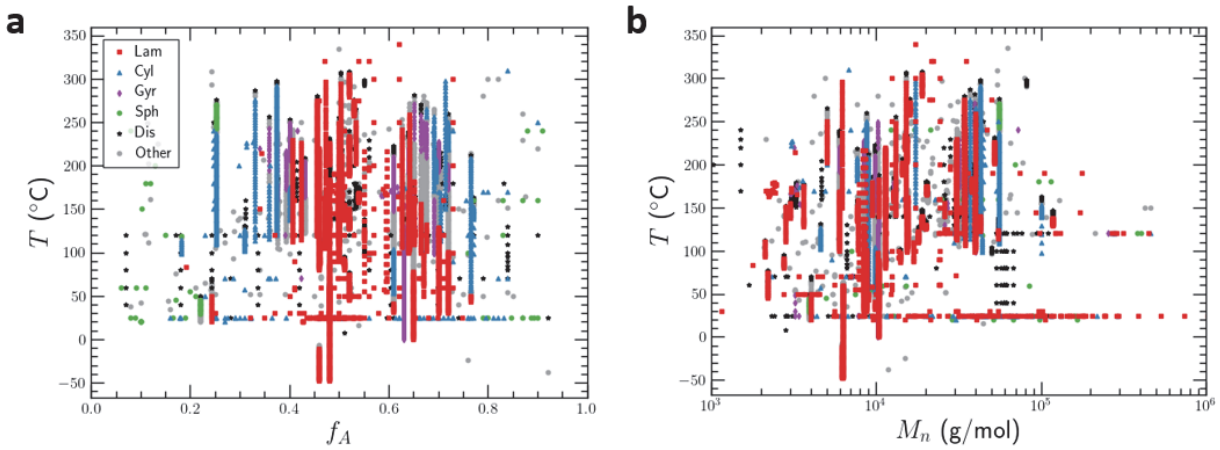


Figure 2. Plots of the entire database for pure phases. "Other" includes coexistence and complex phases. The platform can plot all relationships of M_n, T, and f_A. Vertical points at constant M_n and f_A are rheology scans. The horizontal line at constant T is room temperature.

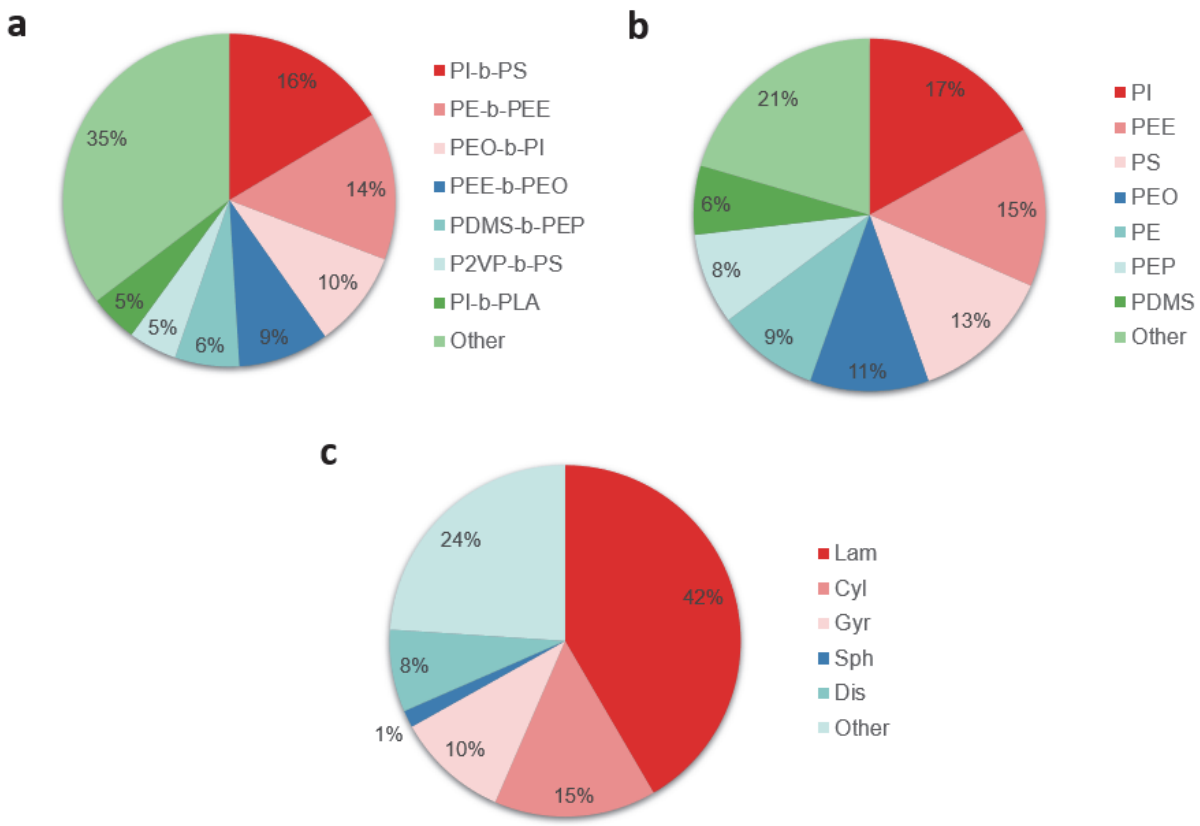


Figure 3. BCDB breakdown by (a) block copolymer (% of total entries); (b) block (% of total blocks); (c) phase (% of total entries). PI is poly(isoprene). PS is poly(styrene). PEO is poly(ethylene oxide). PEE is poly(ethyl ethylene) from hydrogenated polybutadiene. PDMS is poly(dimethyl siloxane). PEP is poly(ethylene-alt-propylene). P2VP is poly(2-vinyl pyridine). PLA is poly(lactic acid). Phase labels are same as in **Figure 2**a.

At the time of publication, BCDB contains a broad range of linear di- and multi-block copolymers, including triblock,[56, 57] tetrablock,[58] hexablock,[59] and undecablock.[59] This data can be downloaded from Zenodo (http://doi.org/10.5281/zenodo.4780309). In the diblock category, there are over 5,400 total entries with more than 60 unique blocks and 60 unique diblocks from approximately 130 peer-reviewed journal articles. Approximately 80% of the data points are pure

phase. To visualize the distributions in M_n, T (column *T* in Table 1), and f_A (for the first block in the *BigSMILES* string), Figure 1 displays histograms, and Figure 2 shows two bivariate plots illustrating the range of parameter space and phases covered by the data. Figures 3a and 3b display breakdowns of diblock and individual block chemistries respectively. The imbalance of chemistries in the literature is clear from these figures: a few well-studied polymers like poly(styrene)-*b*-poly(isoprene), poly(ethylene)-*b*-poly(ethyl ethylene), and poly(ethylene oxide)-*b*-poly(isoprene) dominate the data set. Rheology has a major impact on this skew because each temperature ramp adds many data points for that chemistry. Figure 3c displays the percentages of pure lamellar, cylinder, gyroid, sphere, and disordered. The "other" category includes transition phases and complex phases like hexagonally perforated and modulated layers[60-62] and *Fddd*.[63-66]

BCDB is designed to allow user submission of new data to expand the database. Similar to submitting a manuscript to a peer-reviewed journal, the user is asked to submit all information to the Community Resource for Innovation in Polymer Technology (instructions in the download folder), a centralized database for polymers with a data model[67] and graph-based polymer search algorithms.[68] The database curator will review and verify all information to maintain quality control and will also update the database using large language models (vide infra). Only data from datasets that have completed peer review will be accepted.

3.2. Use Cases

3.2.1. User Interface and Data Availability

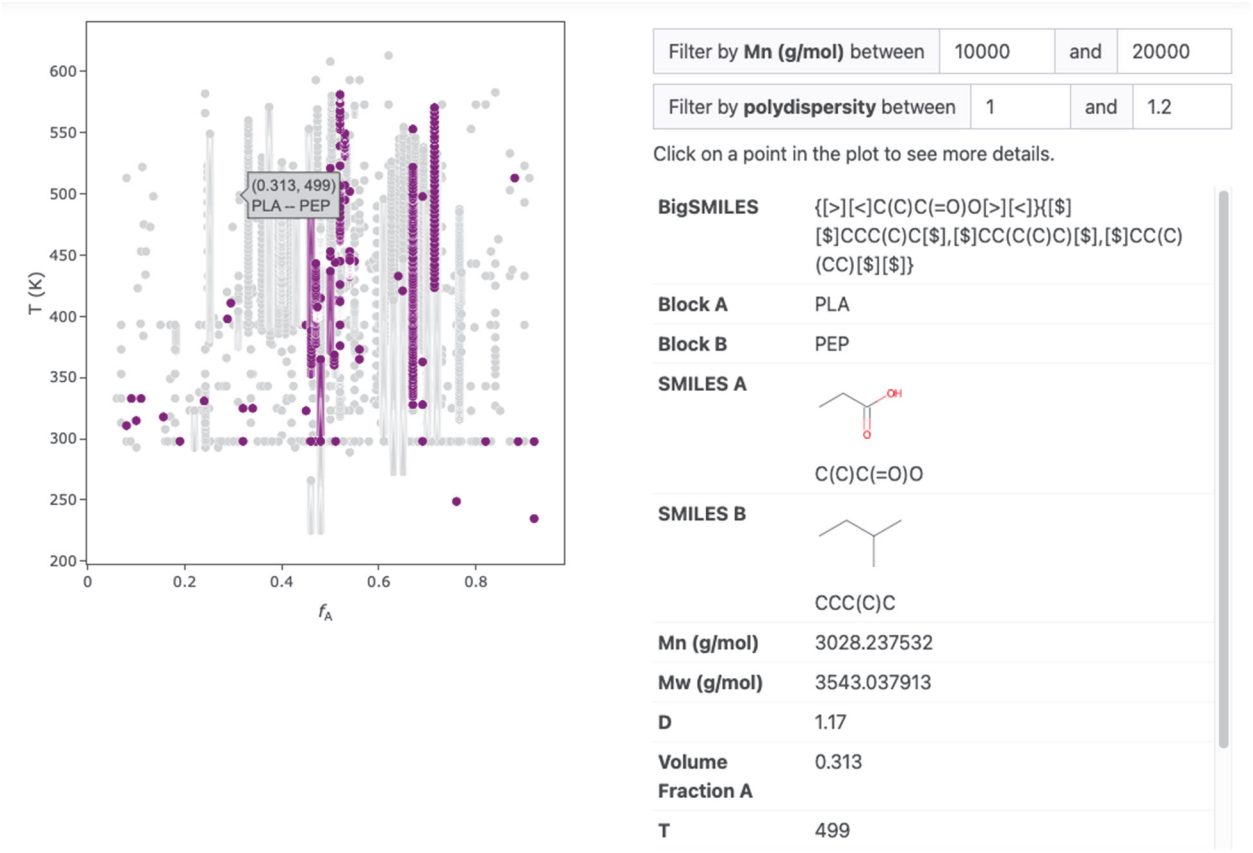
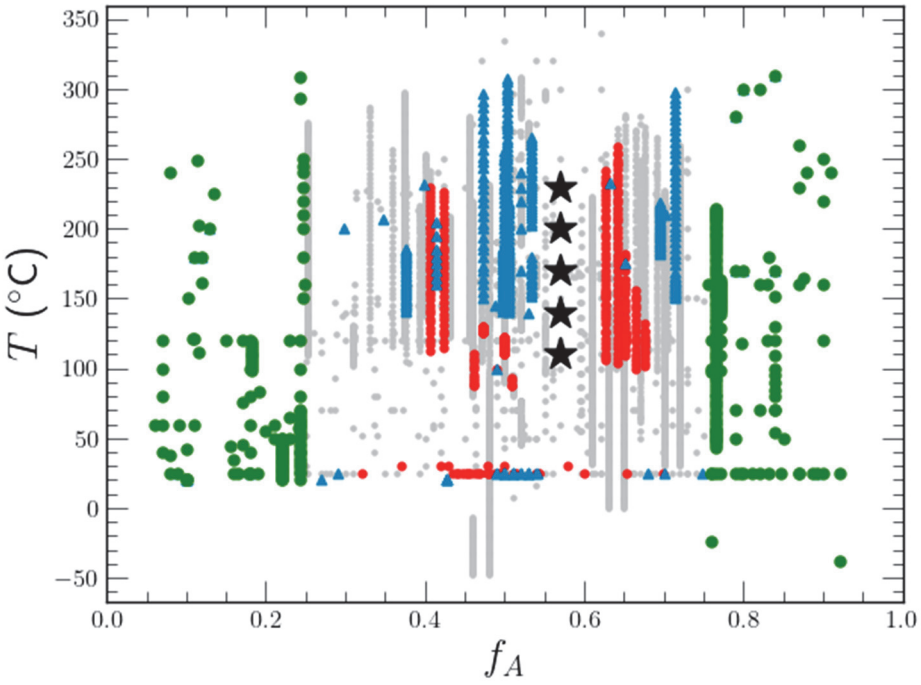


Figure 4. The BCDB has been integrated with the Community Resource for Innovation in Polymer Technology (CRIPT) to enable polymer data visualization and a user-friendly interactive dashboard using Plotly. In this example, the user can search for block copolymers with an overall molar mass between 10,000 and 20,000 g/mol, and a dispersity between 1.0 and 1.2. For each data point, the user can access the ORCID and DOI.

There are two ways users can search chemistry and property information. The first is through the web. An interactive dashboard for polymer data visualization was constructed (Figure

4), which has been integrated as an application with the Community Resource for Innovation in Polymer Technology (CRIPT) with search capabilities for the molar mass and dispersity, an open API, and metadata including when and how the data was collected, descriptions of the column names, and data provenance. The database can be downloaded in JSON format.



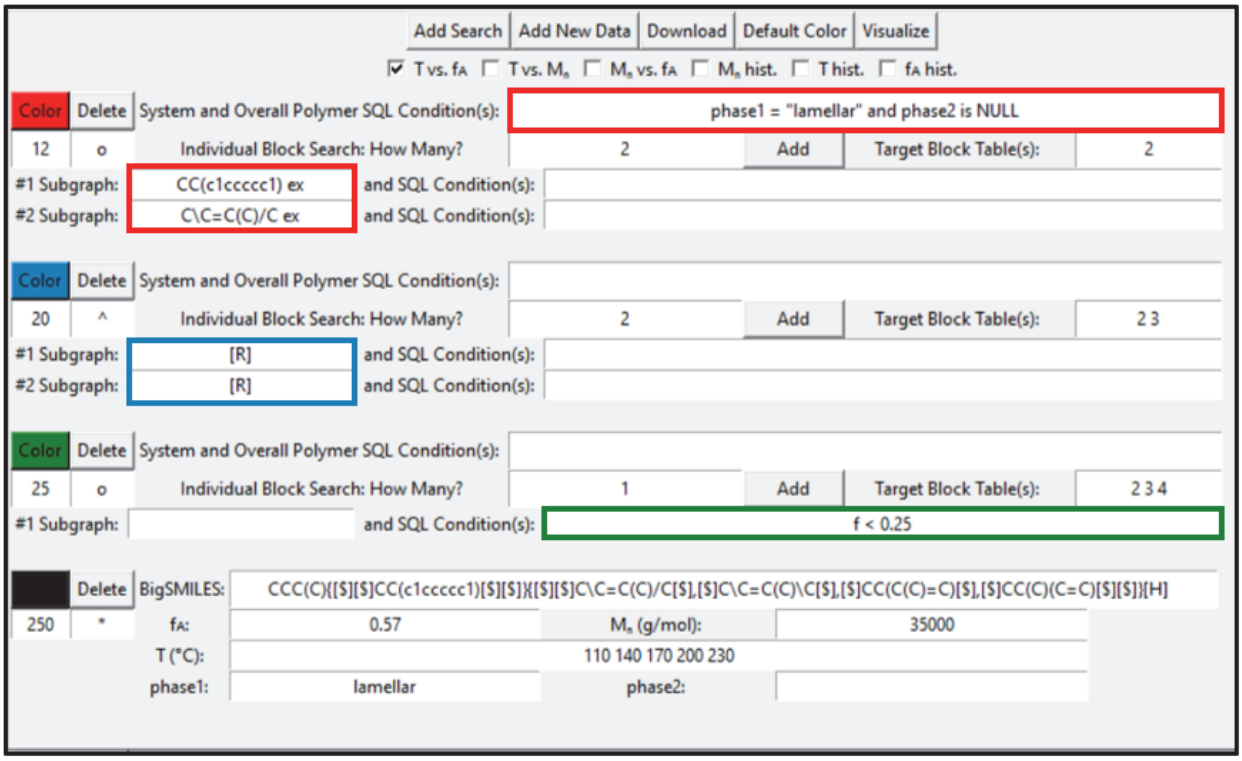


Figure 5. The user can query chemistry and properties, specifically properties of the system and overall polymer using SQL (red), properties of the individual blocks using SQL (green), and

functional group and repeat unit subgraphs using the SMARTS query language (red and blue). Shown is a T versus f_A visualization for different queries the user has entered; red square "s": pure lamellar poly(styrene)-b-poly(isoprene), with repeat units written in SMARTS as "CC(c1ccccc1)" and "C\C=C(C)/C", searched in the diblock table ("2") with the keyword "ex" for exact repeat unit; blue triangle "^": at least two blocks must have rings (the SMARTS query is "[R]") searched in the diblock and triblock tables ("2 3"); green circle "o": at least one block has a volume fraction less than 0.25 searched in the diblock, triblock, and tetrablock tables ("2 3 4"); black star "*": new lamellar data for sec-butyllithium initiated poly(styrene)-b-poly(isoprene) ($f_{PI} = 0.57, M_n = 35$ kDa); grey circle "o": data not hit.

Alternatively, this tool can be downloaded to the user's desktop from Zenodo (http://doi.org/10.5281/zenodo.4780309) and run locally. To increase accessibility and dissemination, BCDB can also be downloaded from the Materials Data Facility (https://acdc.alcf.anl.gov/mdf/detail/bcdb_v1.3/), which uses a non-profit service called Globus to securely store, share, and discover data.[32, 33] The user interface on the user's desktop is shown in Figure 5, built using Python. On the user interface, the user can search all di- and multi-block information, download all of the search hits, and visualize the diblock hits. There is no restriction on the number of searches that the user can enter. BCDB supports two types of searches: Structured Query Language (SQL) (3.2.2), which searches the system, overall polymer, and block columns in Table 1, and chemistry search (3.2.3), which targets the *BigSMILES* column.

For each search, the query and target must be defined. To enter a search, the user clicks the *Add Search* button. The user can write a SQL for the system and overall polymer categories in the *System and Overall Polymer SQL condition(s)* entry, shown in the first query in Figure 5 that

targets pure lamellar systems. The user can also write a block search, shown in the second and third queries in Figure 5, by first entering the number of block searches and clicking *Add*. Each block search has a *Subgraph* and *SQL* entry, allowing block stochastic graphs to be targeted with their characterization information. If the user enters more block searches than there are blocks in the target, then a match is not possible, and the platform will not return any hits. For example, three block searches will not hit to any diblock, whereas three block searches must hit to each block in the triblock. After entering the search, the user chooses the target chemistries in the *Target Block(s)* entry. For example, if the user enters "2 3" in Figure 5, then this will target searches to the diblock and triblock copolymers.

After the query and target are defined, the user can choose to visualize and download the data. To visualize the data, the user chooses the color, shape (entry under the *Delete* button), and size (entry under the *Color* button), and checks any number of bivariate plots to visualize the data. Then, the user clicks *Visualize,* and the platform uses MatPlotLib[69] to generate scatter plots with each single point representing each sample hit. To download the data, the user clicks the *Download* button, and the platform generates CSV files of all hits, with all columns printed, of the target chemistries specified by the user. For hits the user finds interesting, the user can use the *DOI* column to read more information on the phase behavior. Supporting Information shows examples of important uses cases that would be of benefit to the polymer community: search, visualization of the search hits, and download.

3.2.2. Structured Query Language (SQL)

SQL enables searching characterization and strings for the system, overall polymer, and individual block columns in Table 1. The template for extracting data using SQL is:[70]

SELECT *column(s)* FROM *table(s)* WHERE *condition(s)*

User-friendly tutorials can be found in reference 68. The *column(s)* are listed in Table 1, and the *table(s)* refers to each block table that the user can search on the user interface in the *Target Block(s)* entry. On the interface, the user only has to specify the *condition(s)*. These conditions are queries for exact values (string or quantitative) or value ranges (quantitative). For example, to search the pure lamellar systems (Figure 5), the user can enter in the *System and Overall Polymer SQL Condition(s)* entry: "phase1 = "lamellar" and phase2 is NULL". This means that *phase2* is not populated (only the *phase1* column is populated for pure phases in the database). The user is provided a dictionary of *names* and *phases* that is automatically displayed on the platform with the click of a button. Moreover, to conduct a block search for a volume fraction less than 0.25, the user can enter in the block SQL search: "f $<$ 0.25", invoking the volume fraction column f, shown in Figure 5. To execute the SQL queries, the platform uses the sqlite3 in Python to run the queries against the database.

3.2.3. Molecular Search

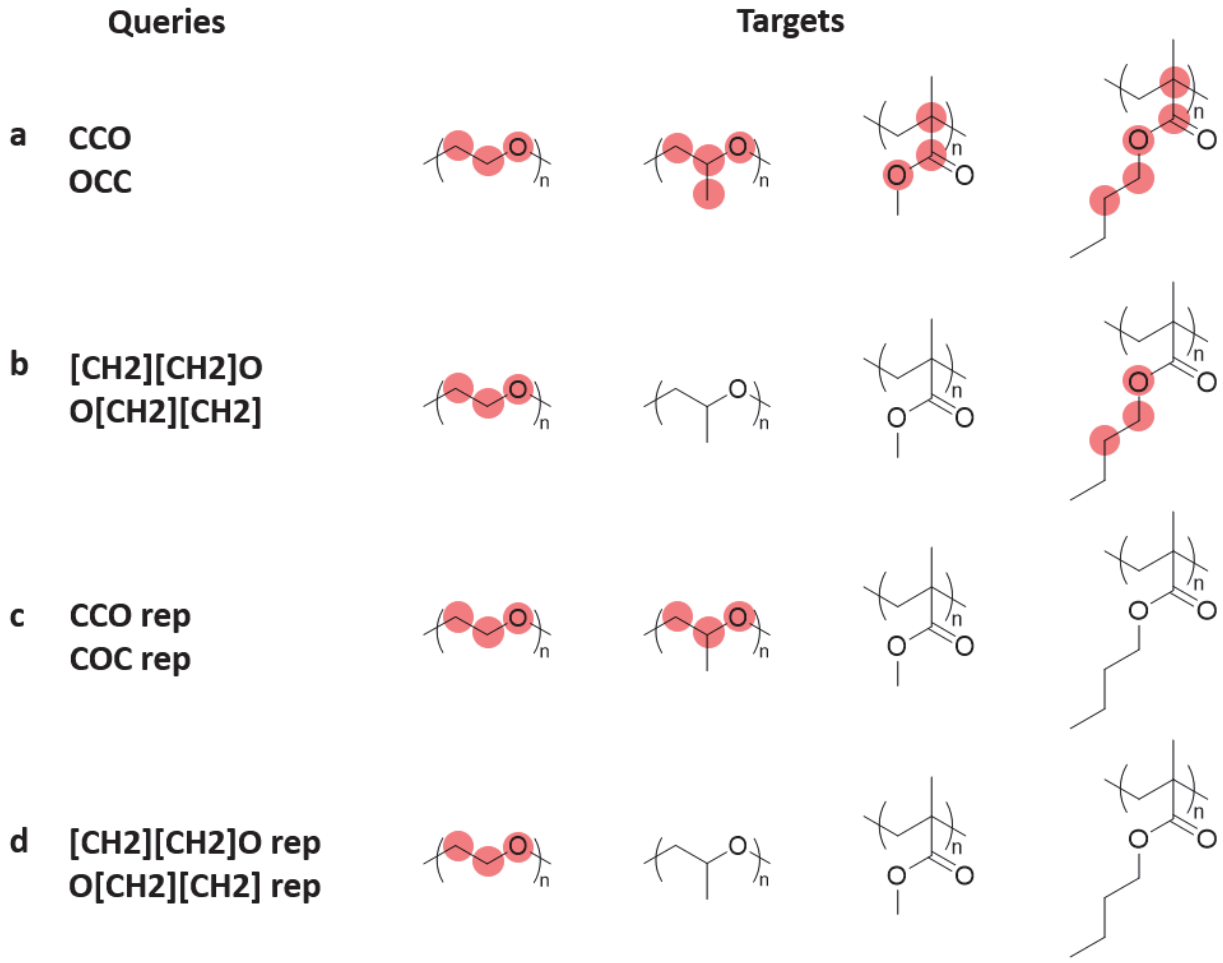


Figure 6. The user can search for functional groups (a-b) in the backbone and sidechains and repeat unit "rep" backbones (c-d) using the SMARTS query language with increasing level of restriction on the target, adding the string "rep" for repeat unit backbone searches. BCDB considers frame shifts in the repeat unit backbone search, and the user can add hydrogens to the query. These searches target poly(ethylene oxide), or "{[][<]CCO[>][]}", poly(propylene oxide), or "{[][<]CC(C)O[>][]}", poly(methyl methacrylate), or "{[][$]CC(C)(C(=O)OC)[$][]}", and poly(butyl methacrylate), or "{[][$]CC(C)(C(=O)OCCCC)[$][]}", blocks.

To identify different polymer blocks, BCDB enables searching polymer chemical structures via line notation. This can be written in the Subgraph entry of the individual block search in Figure 5. The user can search functional groups, repeat unit backbones, and exact repeat units as shown in Figures 6 and 7 by entering the query as SMARTS strings,[71-73] a popular line notation in digital chemistry to encode molecular patterns or subgraphs, and user-friendly tutorials can be found in references 71 and 72. These SMARTS strings are fed to RDKit's[74] subgraph search subroutines. The targets are encoded as BigSMILES strings in the BigSMILES column in BCDB. The repeat units and end groups in these strings are parsed using the open-source BigSMILES parser, discussed in the Supporting Information. Specifically, in each target block, the platform will concatenate allowed combinations of three repeat units to form a set of SMILES targets to feed to RDKit and concatenate end groups if they are present. Chain-growth blocks have more combinations of three repeat units due to head-to-head polymerization than step-growth blocks. In summary, users can use SMARTS to search for:

1) Functional groups in each stochastic object and in connecting end groups
2) Repeat unit backbones by specifying the "rep" keyword
3) Exact repeat units by adding hydrogens or by specifying the "ex" keyword

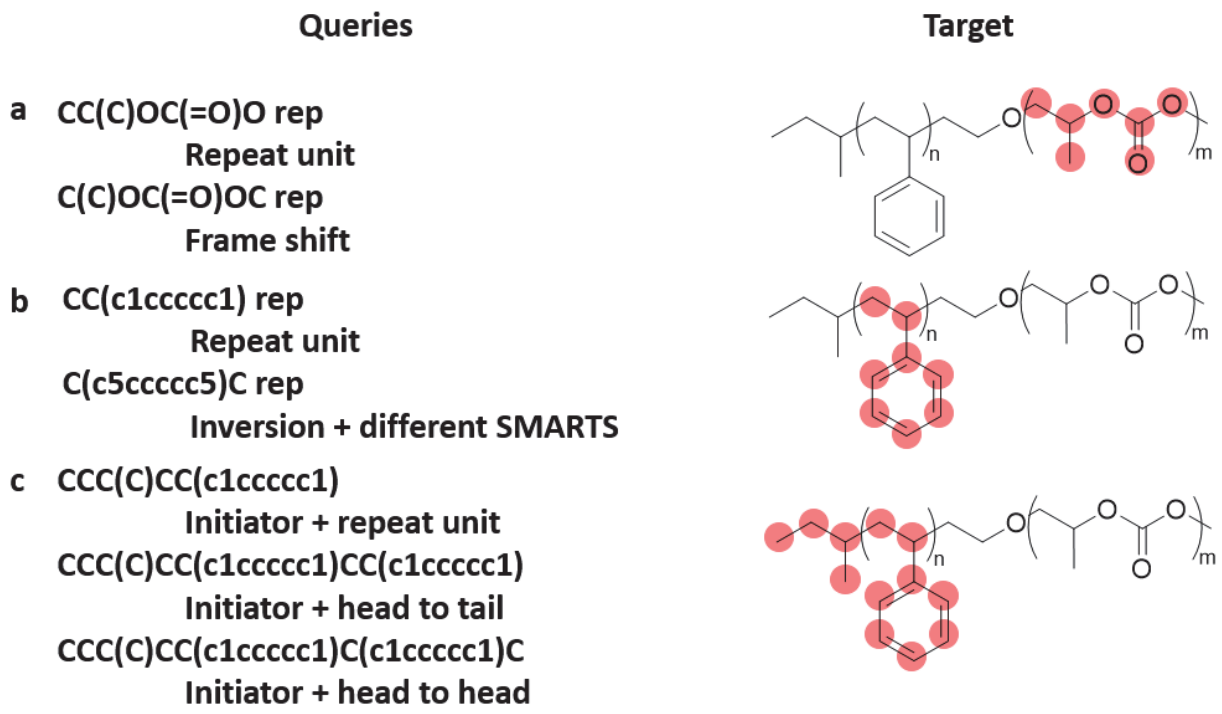


Figure 7. The user can search for (a-b) repeat units with different transformations and (c) a specific initiator and repeat unit using the SMARTS query language, which all match to sec-butyl lithium initiated poly(isoprene)-*b*-poly(propylene carbonate), encoded in BigSMILES as "CCC(C){[$][$]CC(c1ccccc1)[$][$]}CCO{[>][<]CC(C)COC(=O)O[>][<]}[H]" in the database. When executing a substructure search, the platform concatenates deterministic end groups to the stochastic set of SMILES realizations generated by each object so that initiators and realizations of the repeat units can be matched during search.

The user can query block copolymers with increasing restriction on the target matched. The user can search functional groups anywhere in the polymer, in the backbones, sidechains, and end groups, matching blocks with rings like poly(vinyl cyclohexane), poly(styrene), and poly(2-vinylpyridine), blocks with unsaturated functionality like poly(butadiene) and poly(isoprene), blocks with aromatic groups, and any functional group within a single repeat unit or formed from

the stochastic concatenation of repeat units (Figures 6 and 7). The user can add hydrogens, making the query more specific. Moreover, the user can search for repeat unit backbones in SMARTS using the word "rep". For repeat unit searches, the platform will target every repeat unit in every stochastic object. A block can have multiple isomers like poly(isoprene) with four isomers (repeat units in BigSMILES would be [$]C\C=C(C)/C[$], [$]C\C=C(C)\C[$], [$]CC(C(C)=C)[$], [$]CC(C)(C=C)[$], with [$] operators permitting head-to-head or head-to-tail concatenation of the repeat units), or multiple monomer chemistries like poly(lactic-*co*-glycolic acid) with two monomers ([<]C(=O)C(C)O[>] and [<]C(=O)CO[>] have conjugate descriptors permitting only head to tail concatenation of the repeat units). For repeat units with compatible descriptors, the query is allowed to be a frame shifted version of the target. For example, the user can query either "CCO rep" or frame shifted "COC rep" to hit to a target containing poly(ethylene oxide) substructure, which can be poly(ethylene oxide) or [<]CCO[>] in Figure 6 or poly(propylene oxide) or [<]CC(C)O[>]. The user can also add hydrogens to the repeat units, making the backbone query more specific. To execute this search, the platform will construct a circular permutation of the target with no end groups so that the starting atom in the SMARTS query does not matter but rather the exact arrangement of atoms. For a match, the first and last atoms of the target repeat unit backbone must hit. In essence, BCDB enables rich searches of polymers according to their stochastic graph representations.

3.2.4. Download for Machine Learning

Beyond expanded search functionality, it is envisioned that the data infrastructure provided by BCDB can lead to increased transparency and accessibility for block copolymer data, supporting FAIR data principles that scholarly data should be Findable, Accessible, Interoperable and Reusable.[75] With the search and download features, users can build rich and powerful data-

driven models on subset(s) of the data for materials design and discovery. For example, in a separate work by the coauthors,[76] a subset of the BCDB data was used to build a random forest classification model for block copolymer phase prediction. Because of the growing number of publications in machine learning in materials science, we anticipate that the data search and download features on this platform will be of great benefit to the polymer community.

3.2.5. New Data Benchmarking

In addition to allowing users to readily search and extract block copolymer phase data for data-driven research, BCDB also facilitates benchmarking of new experimental data against literature results. By clicking *Add New Data*, the user can visualize new results plotted against data from BCDB as shown for lamellar poly(styrene)-*b*-poly(isoprene) with the large black stars in Figure 5. For each benchmark, the user enters the BigSMILES, f_A for the first block in the BigSMILES, total M_n, a single temperature or set of temperatures from scans separated by a space or comma, and the phase(s). The user can add any number of benchmarks.

3.3. GPT-4 for Rapid Screening of Published Papers

To grow the database and keep pace with the growing volume of literature on block copolymer phase behavior, natural language processing can be used to quickly identify promising papers with linear block copolymer melt data that have been recently published. Large-language models (LLMs) such as GPT-4 caught the attention of many scientists in chemistry and materials, and LLMs were utilized for various applications, including predicting properties of materials, building novel interfaces for tools, data mining from scientific literatures, and designing new educational applications.[77] [78] A zero-shot prompt was created so that GPT-4,[79] an AI-powered language model, can identify whether an entire paper contains data that fits into BCDB by analyzing only the abstract text, shown in Figure 8. This GPT-4 prompt was created by analyzing

a training dataset of approximately 70 papers directly from BCDB and 30 papers that do not contain data that fit into BCDB. The training dataset of papers that do not contain data was created by entering queries on Google Scholar like “block copolymer melt phase behavior”, and papers that do not contain data deal with nonlinear polymers like miktoarm or graft, electrolytes, solutions, blends, and nanoparticle composites. GPT-4's zero-shot prompt was written by the authors after reading these abstract texts, and the prompt was adjusted to maximize the prediction accuracy 100% on the training dataset. The prompt is shown in the Supporting Information.

There are several reasons why abstracts are chosen instead of the titles or the whole papers. First, the abstracts contain more information than the titles. Second, abstracts have much fewer words or tokens than whole papers; using abstracts can avoid memory limitations and save cost.[79] Third, abstracts can be treated as a handcraft summary of the whole paper. Finally, abstracts are always accessible regardless of whether the papers are open-access. However, this study uses papers that are accessible by the scientific community.

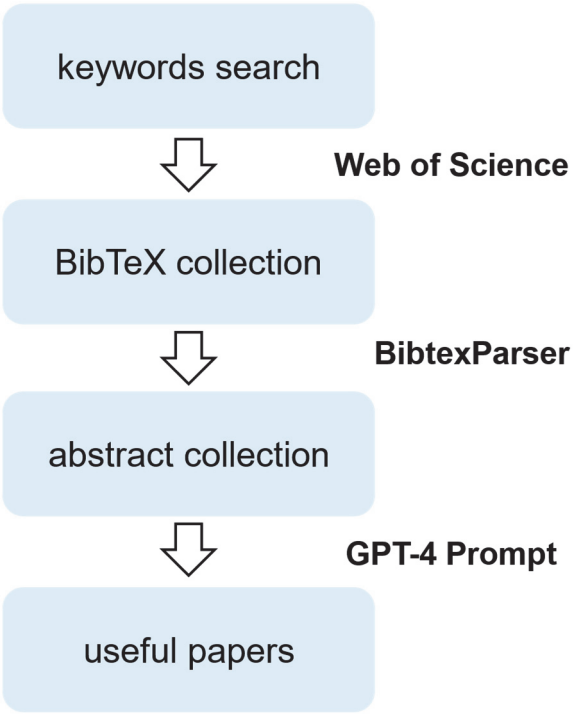


Figure 8. The workflow of utilizing GPT-4 for rapid screening of promising papers.

After the prompt was created using the training dataset, Web of Science was then used to generate a test dataset of block copolymer papers. A keyword search was used on Web of Science with the query, "block copolymer phase behavior," and over 9,000 results were returned. Since OpenAI does not publicize their training dataset for GPT-4 and only stated that the training data for training GPT-4 is before September 2021, to maximally to avoid that the data is touched by ChatGPT and make sure the testing is unbiased, the results from 2022 were chosen, and the number of papers is 305. Then, Web of Science is utilized to output the BibTeX of these 305 papers. BibtexParser, a Python library,[80] is used to extract the abstract text from the BibTeX collection.

GPT-4 is then used with the optimized zero-shot prompt to read the abstracts in this test dataset and manually identify whether the papers contain data that fits into BCDB. Since the large language models' output might have variation, all the abstracts are read by GPT-4 three times

separately. The papers whose three outputs by GPT-4 are all positive or all negative are selected for manual checking of whether they have had BCDB data. For simplicity, only experimental papers are reviewed. In essence, GPT-4 inputs the prompt and an abstract text from the test dataset of papers and outputs a score of either 0 or 1 depending on whether the paper is predicted to have data. This prompt is run repeatedly three different times, and of the 305 papers, 218 were negative all three times. The results of this test are shown in Table 2 and Table 3 (only a fraction of total papers were checked, specifically experimental papers), demonstrating that GPT-4 is decisive for the rapid screening of promising papers. The precision is the ratio of true positives to all positive papers suggested by the model. The recall is the ratio of true positives to all positive papers in the dataset. The F1-score is the harmonic mean of the two. Here, the zero-shot prompt instruction is only preliminary optimized. Therefore, the screening accuracy is anticipated to be further improved if the zero-shot prompt instruction on GPT-4 is further revised or replaced by few-shot prompt instructions with representative shots.

Table 2. GPT-4 results for identifying papers (predicted versus actual).

Category	# Papers	Meaning
True Positive	21	GPT-4 says positive and the paper contains data
False Positive	7	GPT-4 says positive and the paper does not contain data
True Negative	24	GPT-4 says negative and the paper does not data
False Negative	8	GPT-4 says negative and the paper contains data

Table 3. GPT-4 results for identifying papers (metrics).

Metric	Value
Precision	0.75
Recall	0.72
F1 Score	0.74

4. Future Development

As cheminformatics tools continue to grow and develop, there is potential for continued advances in the capabilities of BCDB to enable more efficient research in block copolymers. In the future, the user can execute logical searches for multiple repeat units and end groups, matching statistical copolymer blocks, search functional groups localized to the repeat unit backbones and sidechains, and localize searches to specific stochastic objects, like the middle block of a triblock. This requires more advanced search language tailored for stochastic molecules, but such development would enable users to better extract data sets for machine learning models, enabling structure to property connections and the discovery of new polymers with desired phase behavior.

BCDB is the largest database of its type and at the time of publication contains approximately 130 articles, which is only a fraction of block copolymer papers on Google Scholar or Web of Science. This is in large part because of the initial scope of BCDB restricted to copolymer melts. Furthermore, many papers do not report sufficient data to populate all required fields in BCDB, suggesting the value of the proposed data standard in helping to standardize reporting. It has been previously shown that a simple keyword search for the related quantity χ to be extracted only contained the quantity 38% of the time.[46] Nevertheless, the initial size of BCDB reported in this paper is on par with other polymer data resources. The *Physical Properties of Polymers Handbook*[42] contains a database with approximately 50 peer-reviewed papers for the Flory-Huggins chi (χ) parameter (Table 19.1), a related quantity. Moreover, the PPPDB contains approximately 110 papers that report over 260 χ values.[35]

There are future steps to keeping pace with the exponential growth of chemical literature and supporting search in Big Data space. A natural language processing tool to identify promising papers from literature introduced in the previous section. Ultimately, the GPT-4 prompt can be

modified so that the accuracy is higher, but this work sets the stage for using large language models and artificial intelligence to populate polymer databases, and many other models can be used such as Code Llama.[81] Tools that automatically extract key information from scientific literature using natural language processing continue to develop,[82, 83] and they could be linked to BCDB so that curation becomes faster and more efficient. Moreover, high throughput experimental and simulation methods[24, 84, 85] are becoming popular and instruments from the lab could be connected to the database so that researchers can interact with and analyze their data. BCDB can help provide infrastructure to take full advantage of these advances. As a result, there remains significant potential to easily expand with additional data, and a broader goal is that this database representation expands beyond melts to block copolymers with extra components such as block copolymer and salt mixtures,[86-90] block copolymer and homopolymer blends,[91-93] and amphiphilic block copolymers in solution.[94-96]

5. Conclusions

BCDB is a database of block copolymer phase behavior data that enables users to search, visualize, and download over 5,400 block copolymer melt phase measurements collected from the scientific literature. This database encodes polymers as stochastic molecules in BigSMILES, and the user can apply chemical substructure search and SQL to search chemical structure, characterization, and self-assembled phase helping to make this data more discoverable by the community. This enables facile extraction of data sets for machine learning as well benchmarking of current results against past data. BCDB has been constructed to allow for user-driven addition of data as well as expansion of the type of data that can be received. In these ways, BCDB will facilitate data-driven research in both theory and experimental synthesis that drives forward the field of block copolymers.

6. Data and Code Availability

The user can download the software and data from Zenodo (http://doi.org/10.5281/zenodo.4780309). A video tutorial is available upon download. The visualization tool is released under the MIT License (https://opensource.org/licenses/MIT). The dataset is released under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0).

Supporting Information Statement

List of chemistries and phases; example chemistry and property searches; discussion of the BigSMILES parser; GPT-4 prompt

Acknowledgements

This work is partially funded by the Community Resource for Innovation in Polymer Technology, a project supported by the National Science Foundation (NSF) Convergence Accelerator program (NSF Convergence Accelerator Research-2040636).

7. References

(1) Pitto-Barry, A.; Barry, N. P. Pluronic® block-copolymers in medicine: from chemical and biological versatility to rationalisation and clinical advances. *Polymer Chemistry* **2014**, *5* (10), 3291-3297.
(2) Kataoka, K.; Harada, A.; Nagasaki, Y. Block copolymer micelles for drug delivery: design, characterization and biological significance. *Advanced drug delivery reviews* **2012**, *64*, 37-48.
(3) Jain, S.; Bates, F. S. On the origins of morphological complexity in block copolymer surfactants. *Science* **2003**, *300* (5618), 460-464.
(4) Orilall, M. C.; Wiesner, U. Block copolymer based composition and morphology control in nanostructured hybrid materials for energy conversion and storage: solar cells, batteries, and fuel cells. *Chemical Society Reviews* **2011**, *40* (2), 520-535.
(5) Benvenuta-Tapia, J. J.; Vivaldo-Lima, E.; Martínez-Estrada, A.; Hernández-Valdez, M.; Paredes-Castañeda, S.; Ramos-Valdes, C. Enhanced asphalt performance upon addition of RAFT-synthesized reactive multi-block copolymers. *Materials Chemistry and Physics* **2019**, *227*, 269-278.
(6) tur Rasool, R.; Song, P.; Wang, S. Thermal analysis on the interactions among asphalt modified with SBS and different degraded tire rubber. *Construction and Building Materials* **2018**, *182*, 134-143.
(7) Sajjad, H.; Tolman, W. B.; Reineke, T. M. Block Copolymer Pressure-Sensitive Adhesives Derived from Fatty Acids and Triacetic Acid Lactone. *ACS Applied Polymer Materials* **2020**, *2* (7), 2719-2728.

(8) Shi, W.; Lynd, N. A.; Montarnal, D.; Luo, Y.; Fredrickson, G. H.; Kramer, E. J.; Ntaras, C.; Avgeropoulos, A.; Hexemer, A. Toward strong thermoplastic elastomers with asymmetric miktoarm block copolymer architectures. *Macromolecules* **2014**, *47* (6), 2037-2043.
(9) Bates, C. M.; Maher, M. J.; Janes, D. W.; Ellison, C. J.; Willson, C. G. Block copolymer lithography. *Macromolecules* **2014**, *47* (1), 2-12.
(10) Helfand, E. Block copolymer theory. III. Statistical mechanics of the microdomain structure. *Macromolecules* **1975**, *8* (4), 552-556.
(11) Helfand, E. Theory of inhomogeneous polymers: Fundamentals of the Gaussian random-walk model. *The Journal of chemical physics* **1975**, *62* (3), 999-1005.
(12) Leibler, L. Theory of microphase separation in block copolymers. *Macromolecules* **1980**, *13* (6), 1602-1617.
(13) Semenov, A. Contribution to the theory of microphase layering in block-copolymer melts. *Zh. Eksp. Teor. Fiz* **1985**, *88* (4), 1242-1256.
(14) Matsen, M. W.; Schick, M. Stable and unstable phases of a diblock copolymer melt. *Physical Review Letters* **1994**, *72* (16), 2660.
(15) Drolet, F.; Fredrickson, G. H. Combinatorial screening of complex block copolymer assembly with self-consistent field theory. *Physical Review Letters* **1999**, *83* (21), 4317.
(16) Arora, A.; Qin, J.; Morse, D. C.; Delaney, K. T.; Fredrickson, G. H.; Bates, F. S.; Dorfman, K. D. Broadly accessible self-consistent field theory for block polymer materials discovery. *Macromolecules* **2016**, *49* (13), 4675-4690.
(17) Maurer, W. W.; Bates, F. S.; Lodge, T. P.; Almdal, K.; Mortensen, K.; Fredrickson, G. H. Can a single function for χ account for block copolymer and homopolymer blend phase behavior? *The Journal of Chemical Physics* **1998**, *108* (7), 2989-3000. DOI: 10.1063/1.475704.
(18) Matsen, M. W. The standard Gaussian model for block copolymer melts. *Journal of Physics: Condensed Matter* **2001**, *14* (2), R21.
(19) Arora, A.; Pillai, N.; Bates, F. S.; Dorfman, K. D. Predicting the phase behavior of ABAC tetrablock terpolymers: Sensitivity to Flory–Huggins interaction parameters. *Polymer* **2018**, *154*, 305-314.
(20) Miquelard-Garnier, G.; Roland, S. Beware of the Flory parameter to characterize polymer-polymer interactions: A critical reexamination of the experimental literature. *European Polymer Journal* **2016**, *84*, 111-124.
(21) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017**, *3* (1), 1-13.
(22) Rickman, J. M.; Lookman, T.; Kalinin, S. V. Materials informatics: From the atomic-level to the continuum. *Acta Materialia* **2019**, *168*, 473-510.
(23) Peerless, J. S.; Milliken, N. J.; Oweida, T. J.; Manning, M. D.; Yingling, Y. G. Soft matter informatics: current progress and challenges. *Advanced Theory and Simulations* **2019**, *2* (1), 1800129.
(24) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C* **2018**, *122* (31), 17575-17585.
(25) Shi, J.; Walsh, D.; Zou, W.; Rebello, N.; Deagen, M.; Fransen, K.; Gao, X.; Olsen, B.; Audus, D. Calculating Pairwise Similarity of Polymer Ensembles via Earth Mover's Distance. *ACS Polym. Au* **2024**. DOI: 10.1021/acspolymersau.3c00029.

(26) Shi, J.; Albreiki, F.; Colón, Y. J.; Srivastava, S.; Whitmer, J. K. Transfer Learning Facilitates the Prediction of Polymer-Surface Adhesion Strength. *J. Chem. Theory Comput.* **2023**, *19* (14), 4631-4640. DOI: 10.1021/acs.jctc.2c01314.
(27) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; de Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. Digital Discovery **2023**, 2 (5), 1233-1250.
(28) Shi, J.; Rebello, N. J.; Walsh, D.; Zou, W.; Deagen, M. E.; Leão, B. S.; Audus, D. J.; Olsen, B. D. Quantifying Pairwise Similarity for Complex Polymers. *Macromolecules* **2023**, *56* (18), 7344-7357. DOI: 10.1021/acs.macromol.3c00761.
(29) Shi, J.; Quevillon, M. J.; Amorim Valença, P. H.; Whitmer, J. K. Predicting Adhesive Free Energies of Polymer–Surface Interactions with Machine Learning. *ACS Applied Materials & Interfaces* **2022**, *14* (32), 37161-37169. DOI: 10.1021/acsami.2c08891.
(30) Tang, D.; Boker, S. M.; Tong, X. Are the Signs of Factor Loadings Arbitrary in Confirmatory Factor Analysis? Problems and Solutions. *arXiv preprint arXiv:2401.12937* **2024**. DOI: 10.48550/arXiv.2401.12937 (acccessed 2024-1-30).
(31) Tang, D.; Tong, X. A Comparison of Full Information Maximum Likelihood and Machine Learning Missing Data Analytical Methods in Growth Curve Modeling. *arXiv preprint arXiv:2312.17363* **2023**. (acccessed 2024-1-20).
(32) Blaiszik, B.; Chard, K.; Pruyne, J.; Ananthakrishnan, R.; Tuecke, S.; Foster, I. The Materials Data Facility: Data services to advance materials science research. *Jom* **2016**, *68* (8), 2045-2052.
(33) Blaiszik, B.; Ward, L.; Schwarting, M.; Gaff, J.; Chard, R.; Pike, D.; Chard, K.; Foster, I. A data ecosystem to support machine learning in materials science. *arXiv preprint arXiv:1904.10423* **2019**.
(34) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011/9//, 2011; IEEE: pp 22-29. DOI: 10.1109/EIDWT.2011.13.
(35) *Polymer Property Predictor and Database*. https://pppdb.uchicago.edu/ (accessed Oct. 2020).
(36) *NanoMine*. https://materialsmine.org/nm#/ (accessed Nov. 2020).
(37) *Welcome to Chemical Retrieval on the Web!* https://polymerdatabase.com/about.html (accessed Oct. 2020).
(38) *Polymer Genome*. www.polymergenome.org (accessed Dec. 2020).
(39) Mark, J. E. *Polymer data handbook*; Oxford university press, 2009.
(40) Brandrup, J.; Immergut, E. H. *Polymer Handbook*; John Wiley & Sons, Inc., 1975.
(41) Brandrup, J.; Immergut, E. H.; Grulke, E. A.; Abe, A.; Bloch, D. R. *Polymer Handbook*; Wiley New York, 1999.
(42) Mark, J. E. *Physical properties of polymers handbook*; Springer, 2007.
(43) Bicerano, J. *Prediction of polymer properties*; cRc Press, 2002.
(44) Lin, T.-S.; Rebello, N. J.; Beech, H. K.; Wang, Z.; El-Zaatari, B.; Lundberg, D. J.; Johnson, J. A.; Kalow, J. A.; Craig, S. L.; Olsen, B. D. PolyDAT: A Generic Data Schema for Polymer Characterization. *Journal of chemical information and modeling* **2021**.

(45) Audus, D. J.; de Pablo, J. J. Polymer informatics: Opportunities and challenges. *ACS macro letters* **2017**, *6* (10), 1078-1082.
(46) Tchoua, R. B.; 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.
(47) Tchoua, R. B.; Chard, K.; Audus, D. J.; Ward, L. T.; Lequieu, J.; De Pablo, J. J.; Foster, I. T. Towards a hybrid human-computer scientific information extraction pipeline. In *2017 IEEE 13th International Conference on e-Science (e-Science)*, 2017; IEEE: pp 109-118.
(48) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S. New frontiers for the materials genome initiative. *npj Computational Materials* **2019**, *5* (1), 41.
(49) Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *Journal of Chemical Information and Modeling* **2020**.
(50) Wang, X.; Li, X.; Loo, W.; Newstein, M. C.; Balsara, N. P.; Garetz, B. A. Depolarized Scattering from Block Copolymer Grains Using Circularly Polarized Light. *Macromolecules* **2017**, *50* (13), 5122-5131.
(51) Chang, M.; Abuzaina, F.; Kim, W.; Gupton, J.; Garetz, B.; Newstein, M.; Balsara, N.; Yang, L.; Gido, S.; Cohen, R. Analysis of grain structure in partially ordered block copolymers by depolarized light scattering and transmission electron microscopy. *Macromolecules* **2002**, *35* (11), 4437-4447.
(52) Rohatgi, A. WebPlotDigitizer user manual version 3.4. *URL http://arohatgi. info/WebPlotDigitizer/app* **2014**, 1-18.
(53) Cliche, M.; Rosenberg, D.; Madeka, D.; Yee, C. Scatteract: Automated extraction of data from scatter plots. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017; Springer: pp 135-150.
(54) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS central science* **2019**, *5* (9), 1523-1531.
(55) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28* (1), 31-36.
(56) Shefelbine, T.; Vigild, M. E.; Matsen, M.; Hajduk, D.; Hillmyer, M.; Cussler, E.; Bates, F. Core− shell gyroid morphology in a poly (isoprene-block-styrene-block-dimethylsiloxane) triblock copolymer. *Journal of the American Chemical Society* **1999**, *121* (37), 8457-8465.
(57) Bailey, T. S.; Hardy, C. M.; Epps, T. H.; Bates, F. S. A noncubic triply periodic network morphology in poly (isoprene-b-styrene-b-ethylene oxide) triblock copolymers. *Macromolecules* **2002**, *35* (18), 7007-7017.
(58) Zhang, J.; Sides, S.; Bates, F. S. Ordering of sphere forming SISO tetrablock terpolymers on a simple hexagonal lattice. *Macromolecules* **2012**, *45* (1), 256-265.
(59) Fleury, G.; Bates, F. S. Structure and properties of hexa-and undecablock terpolymers with hierarchical molecular architectures. *Macromolecules* **2009**, *42* (10), 3598-3610.
(60) Hamley, I. W.; Koppi, K. A.; Rosedale, J. H.; Bates, F. S.; Almdal, K.; Mortensen, K. Hexagonal mesophases between lamellae and cylinders in a diblock copolymer melt. *Macromolecules* **1993**, *26* (22), 5959-5970.
(61) Thomas, E. L.; Anderson, D. M.; Henkee, C. S.; Hoffman, D. Periodic area-minimizing surfaces in block copolymers. *Nature* **1988**, *334* (6183), 598-601.

(62) Hajduk, D. A.; Takenouchi, H.; Hillmyer, M. A.; Bates, F. S.; Vigild, M. E.; Almdal, K. Stability of the perforated layer (PL) phase in diblock copolymer melts. *Macromolecules* **1997**, *30* (13), 3788-3795.
(63) Tyler, C. A.; Morse, D. C. Orthorhombic F d d d Network in Triblock and Diblock Copolymer Melts. *Physical review letters* **2005**, *94* (20), 208302.
(64) Takenaka, M.; Wakada, T.; Akasaka, S.; Nishitsuji, S.; Saijo, K.; Shimizu, H.; Hasegawa, H. Orthorhombic Fddd network in diblock copolymer melts. *arXiv preprint cond-mat/0605268* **2006**.
(65) Kim, M. I.; Wakada, T.; Akasaka, S.; Nishitsuji, S.; Saijo, K.; Hasegawa, H.; Ito, K.; Takenaka, M. Stability of the Fddd phase in diblock copolymer melts. *Macromolecules* **2008**, *41* (20), 7667-7670.
(66) Kim, M. I.; Wakada, T.; Akasaka, S.; Nishitsuji, S.; Saijo, K.; Hasegawa, H.; Ito, K.; Takenaka, M. Determination of the Fddd phase boundary in polystyrene-block-polyisoprene diblock copolymer melts. *Macromolecules* **2009**, *42* (14), 5266-5271.
(67) Walsh, D. J.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M. E.; Mysona, J.; Lin, T.-S.; de Pablo, J. J.; Jensen, K. F.; Audus, D. J. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. ACS Publications: 2023.
(68) Rebello, N. J.; Lin, T.-S.; Nazeer, H.; Olsen, B. D. BigSMARTS: A Topologically Aware Query Language and Substructure Search Algorithm for Polymer Chemical Structures. *Journal of Chemical Information and Modeling* **2023**, *63* (21), 6555-6568.
(69) Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **2007**, *9* (3), 90-95.
(70) *SQL Tutorial*. Refsnes Data, https://www.w3schools.com/sql/default.asp (accessed Oct. 2020).
(71) *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Dec. 2020).
(72) *SMARTS Tutorial*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/ (accessed Dec. 2020).
(73) *SMARTS Examples*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html (accessed Dec. 2021).
(74) Landrum, G. Rdkit documentation. *Release* **2013**, *1*, 1-79.
(75) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3* (1), 1-9.
(76) Arora, A.; Lin, T.-S.; Rebello, N. J.; Av-Ron, S. H.; Mochigase, H.; Olsen, B. D. Random forest predictor for diblock copolymer phase behavior. *ACS Macro Letters* **2021**, *10* (11), 1339-1345.
(77) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digital Discovery* **2023**.
(78) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis. *arXiv preprint arXiv:2306.11296* **2023**.
(79) OpenAI. *GPT-4 Technical Report* 2023. https://arxiv.org/abs/2303.08774 (accessed 2023 August 21).

(80) Weiss, M. *Welcome to BibtexParser's documentation!* 2023. https://bibtexparser.readthedocs.io/en/main/ (accessed 2023 August 21).
(81) Roziere, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X. E.; Adi, Y.; Liu, J.; Remez, T.; Rapin, J. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* **2023**.
(82) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Automated Chemical Reaction Extraction from Scientific Literature. *Journal of Chemical Information and Modeling* **2021**.
(83) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications* **2020**, *11* (1), 1-11.
(84) Liu, Y.; Hu, Z.; Suo, Z.; Hu, L.; Feng, L.; Gong, X.; Liu, Y.; Zhang, J. High-throughput experiments facilitate materials innovation: A review. *Science China Technological Sciences* **2019**, *62* (4), 521-545.
(85) 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. *Journal of the American Chemical Society* **2019**, *141* (22), 8921-8927.
(86) Wanakule, N. S.; Panday, A.; Mullin, S. A.; Gann, E.; Hexemer, A.; Balsara, N. P. Ionic conductivity of block copolymer electrolytes in the vicinity of order− disorder and order− order transitions. *Macromolecules* **2009**, *42* (15), 5642-5651.
(87) Ruzette, A.-V. G.; Soo, P. P.; Sadoway, D. R.; Mayes, A. M. Melt-formable block copolymer electrolytes for lithium rechargeable batteries. *Journal of The Electrochemical Society* **2001**, *148* (6), A537.
(88) Epps, T. H.; Bailey, T. S.; Waletzko, R.; Bates, F. S. Phase behavior and block sequence effects in lithium perchlorate-doped poly (isoprene-b-styrene-b-ethylene oxide) and poly (styrene-b-isoprene-b-ethylene oxide) triblock copolymers. *Macromolecules* **2003**, *36* (8), 2873-2881.
(89) Young, W.-S.; Epps III, T. H. Salt doping in PEO-containing block copolymers: counterion and concentration effects. *Macromolecules* **2009**, *42* (7), 2672-2678.
(90) Loo, W. S.; Galluzzo, M. D.; Li, X.; Maslyn, J. A.; Oh, H. J.; Mongcopa, K. I.; Zhu, C.; Wang, A. A.; Wang, X.; Garetz, B. A. Phase behavior of mixtures of block copolymers and a lithium salt. *The Journal of Physical Chemistry B* **2018**, *122* (33), 8065-8074.
(91) Takagi, H.; Yamamoto, K. Phase Boundary of Frank–Kasper σ Phase in Phase Diagrams of Binary Mixtures of Block Copolymers and Homopolymers. *Macromolecules* **2019**, *52* (5), 2007-2014.
(92) Kinning, D. J.; Thomas, E. L.; Fetters, L. J. Morphological studies of micelle formation in block copolymer/homopolymer blends. *The Journal of chemical physics* **1989**, *90* (10), 5806-5825.
(93) Schwahn, D.; Mortensen, K.; Frielinghaus, H.; Almdal, K.; Kielhorn, L. Thermal composition fluctuations near the isotropic Lifshitz critical point in a ternary mixture of a homopolymer blend and diblock copolymer. *The Journal of Chemical Physics* **2000**, *112* (12), 5454-5472.
(94) Ivanova, R.; Lindman, B.; Alexandridis, P. Effect of glycols on the self-assembly of amphiphilic block copolymers in water. 1. Phase diagrams and structure identification. *Langmuir* **2000**, *16* (8), 3660-3675.
(95) Barnhill, S. A.; Bell, N. C.; Patterson, J. P.; Olds, D. P.; Gianneschi, N. C. Phase diagrams of polynorbornene amphiphilic block copolymers in solution. *Macromolecules* **2015**, *48* (4), 1152-1161.

(96) Wanka, G.; Hoffmann, H.; Ulbricht, W. Phase diagrams and aggregation behavior of poly (oxyethylene)-poly (oxypropylene)-poly (oxyethylene) triblock copolymers in aqueous solutions. *Macromolecules* **1994**, *27* (15), 4145-4159.

TOC Figure