



Bioplastic design using multitask deep neural networks

Christopher Kuenneth ^{1,2}, Jessica Lalonde^{3,4}, Babetta L. Marrone³, Carl N. Iverson⁵, Rampi Ramprasad² & Ghanshyam Pilania ¹✉

Non-degradable plastic waste jeopardizes our environment, yet our modern lifestyle and current technologies are impossible to sustain without plastics. Bio-synthesized and biodegradable alternatives such as polyhydroxyalkanoates (PHAs) have the potential to replace large portions of the world's plastic supply with cradle-to-cradle materials, but their chemical complexity and diversity limit traditional resource-intensive experimentation. Here, we develop multitask deep neural network property predictors using available experimental data for a diverse set of nearly 23,000 homo- and copolymer chemistries. Using the predictors, we identify 14 PHA-based bioplastics from a search space of almost 1.4 million candidates which could serve as potential replacements for seven petroleum-based commodity plastics that account for 75% of the world's yearly plastic production. We also discuss possible synthesis routes for the identified promising materials.

¹ Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ² School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. ³ Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ⁴ Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA. ⁵ Chemistry Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ✉email: gpilania@lanl.gov

Plastics are an integral part of our everyday life and modern technology. Their simple, yet diverse, chemistries and tunable properties make plastics versatile and desirable; plastics display high or low flexibility, strength, thermal, or electronic conductivity along with low cost, low weight, and abundance¹. The global plastic production of 2019 amounts to an unimaginable 368 million tonnes and is expected to further increase in the coming years. About 40% (145 million tonnes) of the yearly plastic production accounts for packaging products such as bags, food containers, cutlery, or bottles, which have very short service lifetimes and often end up in landfills, seawater, or other natural environments^{2–4}. It is therefore not surprising that packaging plastics are one of the largest polluters of our world's ecosystems, severely threatening the existence of animals and humans through waste and microplastic particles on land and in oceans that last for decades or longer⁵. Finding eco-friendly plastics (bioplastics) with properties akin to conventional plastics but with sustainable recycling options is therefore of utmost importance for a circular economy.¹ The bio-derived and biodegradable family of polyhydroxyalkanoates (PHAs) is a promising cradle-to-cradle material that can be synthesized by several microorganisms directly using sunlight and CO₂ from the environment or industrial point sources^{3,6,7}. Existing commercial entities have already studied the biosynthesis, processing, and industrialization of PHA-based polymers and blends⁸. Diverse chemistries harbored in PHAs span a large property space with ample opportunities to design mechanical and thermal properties

such as the Young's modulus (E), tensile strength (σ), elongation (ϵ), glass transition temperature (T_g), melting temperature (T_m), and degradation temperature (T_d)^{3,9–14}.

PHAs provide copious opportunities for chemical modification and property modulation^{3,15}. Key parameters of these modifications are the numbers of carbons in the main-chain and side-chain (n and m in Fig. 1a), and the terminating functional groups of the side-chain (R in Fig. 1a)^{15,16}. For instance, the most widely known PHA, poly-3-hydroxybutyrate (P3HB, with $n = 1$, $m = 1$, and no R), is brittle and inflexible¹⁷. As the number of carbon atoms in the backbone increases, the resulting polymers tend to display higher elongation at break (ϵ_b) combined with improved mechanical strength and enhanced tendency for degradability³. Also, PHAs with side-chain-terminating phenyl groups exhibit higher T_g s because of increased rigidity due to enhanced inter-chain interactions resulting from the polar side chain functional groups^{15,18}. Besides systematic structural and chemical alterations, copolymers provide an additional knob to grow the accessible property space by not only combining multiple PHA-based motifs but also PHAs with conventional polymers¹⁹. In the past, PHA-only copolymers have been found to improve mechanical properties while keeping high T_m and low T_g values, which is ideal for applications that require large temperature operation windows^{3,14}. By forming copolymers of PHAs with conventional polymers, one may harness synergistic effects, potentially leading to recyclable polymers with enhanced mechanical strength and improved gas permeability. Polymers with low gas permeabilities are of

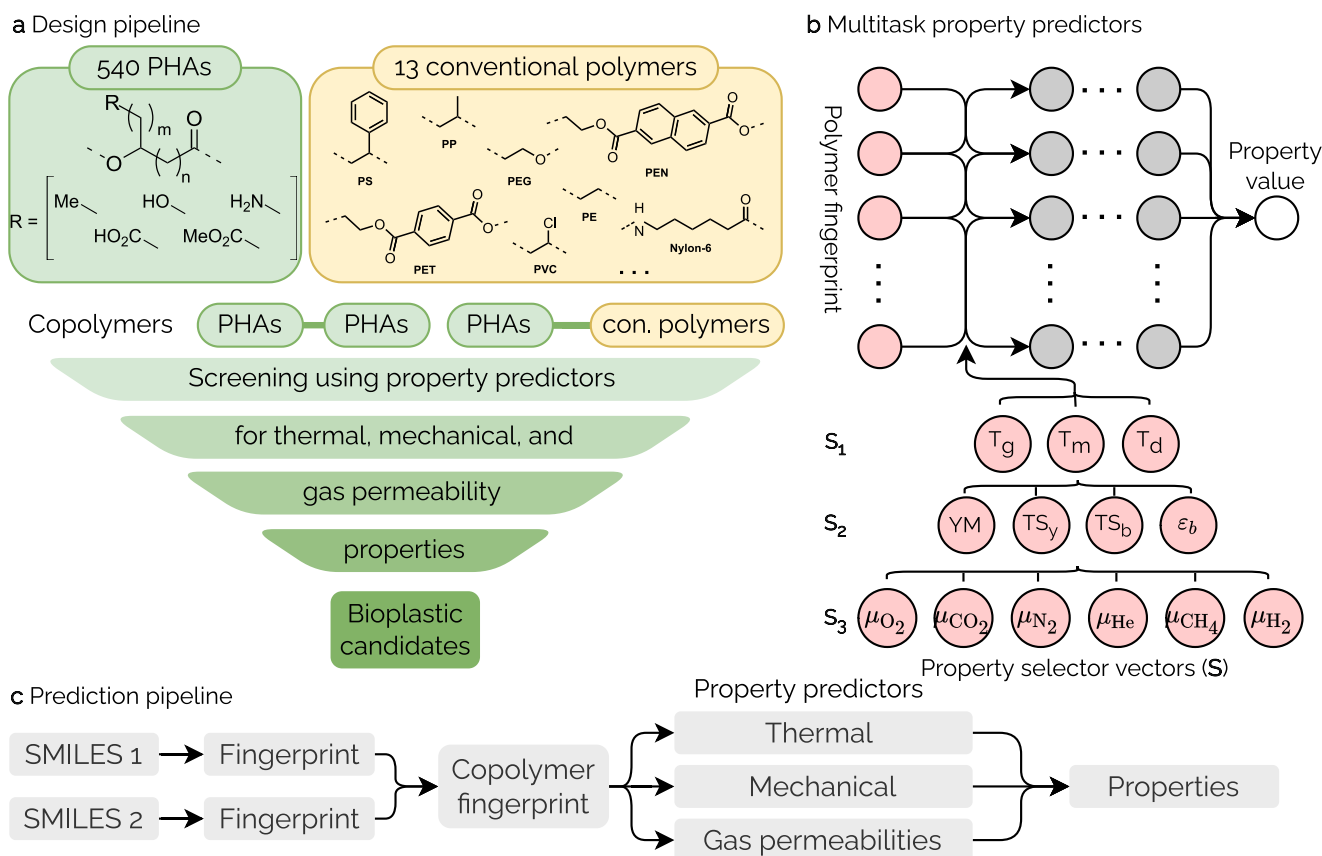


Fig. 1 Bioplastic design using multitask deep learning predictors. **a** Design pipeline. A large search space is created by combining 540 polyhydroxyalkanoates (PHAs) and 13 conventional polymers to copolymers. Property predictors and property requirements of commonly used polymers allow us to identify bioplastic candidates within the search space. **b** Architecture of the multitask neural network predictors. Three separate predictors are trained; one for each of the selector vectors (S_1 , S_2 , S_3). S is a binary vector (1 for the selected component and 0 everywhere else) that selects the output property. **c** Prediction pipeline. The two SMILES⁵⁸ strings belong to comonomers in a copolymer.

particular interest to the food packaging industry where the containment of gases such as CO₂, or the prevention of gases such as O₂ from permeating into the food products, is an important factor in deciding application-specific suitability of a polymer.

Just contemplating copolymer compositions on a rather coarse composition grid ($c = 0, 0.1, \dots, 1$), the total number of distinct PHA-only copolymer possibilities is far beyond a million, effectively rendering trial and error-based high-throughput experiments an impractical route of searching for application-specific candidate materials²⁰. Also, the sheer size of the search space disqualifies time-consuming computational methods such as density functional theory (DFT) or even classical molecular dynamics (MD) simulations. The burgeoning field of polymer informatics^{21–26} offers an exciting alternative route to address such search problems by using modern data-driven machine learning approaches^{18,27–29}.

The present study, with the details of the workflow and machine learning framework outlined in Fig. 1, has several vital elements. First, we develop efficient multitask deep neural network-based multiproperty predictors for copolymers that forecast three different thermal (T_g , T_m , and T_d), four different mechanical (E , σ_y , σ_b , and ϵ_b), and six gas permeability ($\mu_{gl} \in \{O_2, CO_2, N_2, H_2, He, CH_4\}$) properties using nearly 23,000 experimental data points pertaining to a diverse range of homo- and copolymer chemistries. Here, T_g , T_m , T_d , E , σ_y , σ_b , and ϵ_b are the glass transition temperature, melting temperature, degradation temperature, Young's modulus, tensile strength at yield, tensile strength at break, and elongation at break, respectively. μ_{O_2} , μ_{CO_2} , μ_{N_2} , μ_{H_2} , μ_{He} , and μ_{CH_4} stand for the gas permeabilities of O₂, CO₂, N₂, H₂, He, and CH₄. The thermal, mechanical, and gas permeability properties are selected as they play a critical role in the design and selection of plastics for packaging and other large-scale industrial applications. Second, we create a bioplastic candidate space of nearly 1.4 million bioplastics, which is spanned by 540 PHAs and 13 conventional polymer chemistries. Third, we follow a two-step protocol to find several PHA-only and PHA-conventional polymer bio-replacements in the candidate space for seven petroleum-based and commonly used plastics. Possible synthesis routes of the bio-replacements are discussed. This work represents the state-of-the-art in polymer informatics, and contributes to and accelerates the identification of sustainable functional polymer candidate materials.

Results and discussion

Data Set. Our data set for training of the multitask property predictors includes a total of 22,731 homopolymer ($\approx 60\%$) and copolymer ($\approx 30\%$) data points of the thermal, mechanical, and the small molecule gas permeability properties as reported in Table 1. Each of the 7512 copolymer data points involves two distinct comonomers at various compositions while spanning over 1440 distinct copolymer chemistries. Homo- and copolymer data points of T_g , T_m , and T_d , and homopolymer data points of μ_{gs} , E , and σ_b were already utilized in previous studies^{30–35}. The copolymer data points belonging to μ_{gs} , E , σ_y , σ_b , and ϵ_b , and homopolymer data points of σ_y and ϵ_b were collected from the PoLyInfo³⁵ repository for this study. If multiple data points were available for a single polymer in PoLyInfo, we used the average values after a manual curation. For consistency and uniformity, only T_g and T_m data points measured via differential scanning calorimetry (DSC), T_d data points measured via thermogravimetric analysis (TGA), and mechanical data points recorded around room temperature (300 K) were included in the data set. Moreover, for configurational consistency, all copolymer data points in this study are from random copolymers. As part of an additional curation step and our due diligence strategy, we employed a clustering algorithm (DBSCAN using standard parameters as implemented in Scikit-learn³⁶) to identify outliers and select suspicious data points for manual inspection. The degree of polymerization and molecular weight were not taken into account because they were not uniformly available for all data points. Mandated by the multitask method, all property values were scaled to the range of [0, 1] (min-max scaling) for training and transformed back to the actual ranges before computing the respective error metrics. Additionally, ϵ_b and the gas permeabilities were transformed to the log base 10 scale ($x \mapsto \log_{10}(x + 1)$) before training because of their power-law-shaped data distributions (see Supplementary Figs. 4–6).

Property predictors. Multitask deep neural networks with meta learners have shown best-in-class performance in past polymer informatics studies^{33,34} due to their ability to utilize inherent correlations in data that helps to overcome data sparsity. Here, we create three multiproperty predictors (one for each category in Table 1) to predict, in total, 13 polymer properties using the data

Table 1 Overview of our data set used for training the multitask predictors.

| | Symbol | Unit | Homopolymer | Copolymer | Range | Total |
|------------------------------------|--------------|--------|-------------|-----------|-----------------------------|--------|
| Thermal properties | | | | | | |
| Glass transition temp. | T_g | K | 5183 | 3312 | [80,873] | 8495 |
| Melting temp. | T_m | K | 2132 | 1523 | [215,860] | 3655 |
| Degradation temp. | T_d | K | 3584 | 1064 | [291,1173] | 4648 |
| Mechanical properties | | | | | | |
| Young's modulus | E | MPa | 592 | 322 | [0.2,4000] | 914 |
| Tensile strength at yield | σ_y | MPa | 216 | 78 | [0.01,132] | 294 |
| Tensile strength at break | σ_b | MPa | 663 | 318 | [0.04,200] | 981 |
| Elongation at break | ϵ_b | | 868 | 260 | [0.3, 995] | 1128 |
| Gas permeability properties | | | | | | |
| O ₂ | μ_{O_2} | barrer | 420 | 210 | [5·10 ⁻⁶ ,1000] | 630 |
| CO ₂ | μ_{CO_2} | barrer | 313 | 119 | [10 ⁻⁶ , 4756] | 432 |
| N ₂ | μ_{N_2} | barrer | 417 | 99 | [3·10 ⁻⁵ , 480] | 516 |
| H ₂ | μ_{H_2} | barrer | 266 | 46 | [2·10 ⁻² , 5000] | 312 |
| He | μ_{He} | barrer | 261 | 58 | [5·10 ⁻² ,1950] | 319 |
| CH ₄ | μ_{CH_4} | barrer | 360 | 47 | [4·10 ⁻⁴ , 1690] | 407 |
| Total | | | 15,275 | 7456 | | 22,731 |

The property portfolio for the three subgroups, associated property ranges, units of measurements, as well as the number of homo- and copolymers included in the dataset are outlined.

Table 2 The RMSEs and R^2 s averages of the five cross-validation models and meta learner from predictions on the respective validation data sets. The cross-validation deviations show one standard deviation (1σ).

| Symbol ^a | Unit | Cross-validation | | Meta learner | |
|--|--------|------------------|----------------|--------------|----------------|
| | | RMSE | R ² | RMSE | R ² |
| Thermal properties | | | | | |
| T _g | K | 29.78 ± 1.26 | 0.92 ± 0.01 | 13.04 | 0.98 |
| T _m | K | 40.17 ± 0.83 | 0.84 ± 0.01 | 16.67 | 0.97 |
| T _d | K | 62.16 ± 2.52 | 0.72 ± 0.02 | 23.84 | 0.96 |
| Mechanical properties | | | | | |
| E | MPa | 475.34 ± 31.84 | 0.78 ± 0.03 | 237.2 | 0.94 |
| σ _y | MPa | 15.43 ± 3.81 | 0.79 ± 0.12 | 7.1 | 0.96 |
| σ _b | MPa | 18.82 ± 1.00 | 0.77 ± 0.02 | 9.81 | 0.94 |
| ε _b ^b | | 0.43 ± 0.04 | 0.59 ± 0.10 | 0.2 | 0.91 |
| Gas permeability properties | | | | | |
| μ _{O₂} ^b | barrer | 0.13 ± 0.02 | 0.97 ± 0.02 | 0.07 | 0.99 |
| μ _{CO₂} ^b | barrer | 0.20 ± 0.04 | 0.96 ± 0.02 | 0.11 | 0.99 |
| μ _{N₂} ^b | barrer | 0.12 ± 0.04 | 0.96 ± 0.03 | 0.05 | 0.99 |
| μ _{H₂} ^b | barrer | 0.14 ± 0.02 | 0.97 ± 0.01 | 0.06 | 1.0 |
| μ _{He} ^b | barrer | 0.14 ± 0.01 | 0.96 ± 0.01 | 0.06 | 0.99 |
| μ _{CH₄} ^b | barrer | 0.16 ± 0.03 | 0.96 ± 0.01 | 0.06 | 0.99 |

^a See Table 1 for symbol definition.^b Trained on log base 10 scale ($x \rightarrow \log_{10}(x + 1)$). RMSE and R^2 values are reported on this scale.

set and categories profiled in Table 1 and fingerprints outlined in the Methods section. Figure 1b schematically shows the architecture of the multitask predictors, while implementation details are given in the Methods section and in Supplementary Fig. 1.

The developed meta learners and cross-validation predictors display outstanding overall coefficient of determination (R^2) values of 0.97 and 0.93, respectively. The root-mean-square error (RMSE) and R^2 values of all properties are reported in Table 2. The meta learner (cross-validation) predictors of the thermal properties perform very well with R^2 s as high as 0.98 (0.92), 0.97 (0.84), and 0.96 (0.72) for T_g , T_m , and T_d , respectively. This is expected because of the large number of data points and high data fidelity of the thermal property data points. The reported validation metrics here are slightly better than those reported in Ref. 35, which uses a very similar thermal property data set. This is because of the extra data curation and cleaning steps adopted in this work, as discussed above in the Data Set section. Wu et al.³⁷ and Tao et al.³⁸ report similar R^2 values of 0.91 and 0.93, respectively, for a similar-sized T_g data set. Chen et al.³⁹ and Tao et al.⁴⁰ obtain slightly worse R^2 values for learning T_g . We note that different training and evaluation schemes lead to different metrics and only the exact same data sets (and splits) should be compared for benchmarking purposes. The mechanical and gas permeability meta learners show very high R^2 s of 0.94, 0.96, 0.94, and 0.91 for E , σ_y , σ_b , and ϵ_b , respectively, and 0.99, 0.99, 0.99, 1.00, 0.99, and 0.99 for the six gases $g \in \{O_2, CO_2, N_2, H_2, He, CH_4\}$, respectively. The overall performance of the three developed meta-learners with averaged R^2 s of 0.97, 0.94, and 0.99 is exceptional and may be credited to the large data set of almost 23,000 data points, additional data curation measures, well-conditioned and smooth fingerprints, and fully-hyperparameter-optimized multitask deep neural networks. We use the meta learners for property predictions. The cross-validation metrics indicate the generalization error for learning the data set (see also Supplementary Discussion on “Generalization and data set errors”). The individual parity plots of the meta learners for each property can be found in the Supplementary Figs. 7–9. Supplementary Fig. 11 shows a good agreement of predictions and data points of four copolymers, which are included in the test data set only.

Bioplastic search space. In the next step, we consider a bioplastic space that can be searched using the property predictors developed in the last section. As shown in Fig. 1a, 540 PHAs and 13 conventional polymers define and bound this space. The 540 PHAs are devised through variations of the number of carbon atoms in the main-chain and side-chain from 1 to 6 (n and m in Fig. 1a), and by terminating the side-chains with 17 different functional groups (see Supplementary Fig. 2). The biocopolymers of this space are generated by the outer product of PHAs and conventional polymers at eleven different compositions ($c = 0, 0.1, \dots, 1$). The total number of bioplastics in the search space amounts to 1 373 503 and is composed of 553 homopolymers, 146 070 copolymers of PHA-only, and 7 033 copolymers of PHAs and conventional polymers. The 13 conventional polymers were selected according to the list of most commonly used plastics and are documented in Supplementary Fig. 3.

Figure 2a, b display the 2D uniform manifold approximations and projections (UMAPs)⁴¹ of two different fingerprint subspaces. These subspaces contain all fingerprint components of the selected polymers. The fingerprint subspace of Fig. 2a contains five PHAs (green dots, $c = 1$), the remaining 548 polymers (blue dots, $c = 0$), and copolymers ($c = 0.1, 0.2, \dots, 0.9$) that connect the five PHAs and the remaining 548 polymers. Interestingly, UMAP has identified similar polymers and agglomerated them into the shape of stars. The corner vertices of these stars are the fingerprints of the copolymers closest to the five PHAs ($c = 0.9$), while all other dots in the stars indicate fingerprints of polymers of similar chemistry. For example, the dots of the topmost star in Fig. 2a show fingerprints of PHA copolymers containing a nitro phenyl functional group, while the rightmost star includes fingerprints of PHA copolymers containing benzonitrile. The different clusters in Fig. 2b have 13 corner vertices (some of them are hidden) that indicate the fingerprints of the copolymers closest to the 13 conventional polymers ($c = 0.9$, see Fig. 1a), which are included in the fingerprint subspace of Fig. 2b, instead of the five PHAs as in Fig. 2a. Equivalent to the stars, the different clusters contain fingerprints of polymers of similar chemistry. The agglomeration to stars or clusters illustrates that the used fingerprints (i) unambiguously

distinguish polymers; (ii) position related polymers of similar chemistries in the vicinity (i.e., pack all copolymers with a specific side-chain functional group in the same part of the space); (iii) and thus create a physically meaningful learning space with chemical similarity integrated that is well-suited for machine learning.

Figure 3 displays property relations for a selected set of properties of the bioplastic search space in order to qualitatively assess our predictions and verify physical trends expected based on chemical intuition. The trend in Fig. 3a is that polymers of high T_g values also have high T_m and room temperature E values. This confirms our chemical intuition that T_g is approximately linearly correlated to T_m , and high T_g and/or T_m polymers have stiffer morphologies thus possessing high E values. Also, we

observe that the correlation of T_g and T_m is not sharp but broad, which arises from the different side-chain functional groups in the search space. In contrast, Fig. 3b suggests little to no correlation of T_m and σ_b , except that the range spanned by σ_b at a given T_m broadens as T_m increases. However, σ_b is intuitively correlated to E , i.e., stiffer materials (high E) break at higher stresses (high σ_b). Figure 3c shows a roughly linear correlation of μ_{CO_2} and μ_{O_2} that again agrees with chemical intuition and lends credibility to the developed predictors.

Bioplastic Replacements. Up to this point, we have discussed the training and validation of three multitask deep neural networks (each targeting separately the thermal, mechanical, and gas permeability properties) to forecast 13 polymer properties, the consideration of a search space of over 1.3 million bioplastic candidates, and predictions for each of the candidates in the search space. Next, we search the candidate set for suitable replacements for seven petroleum-based and commonly used plastics listed in Table 3. The search is performed following a two-step protocol. In the first step, we employ a nearest neighbors search to find the five closest replacements (within the target property space) for each of the seven plastics and in each copolymer subgroup of PHA-only and PHAs with conventional polymers. We employ the nearest neighbors algorithm using standard parameters as implemented in Scikit-learn³⁶. Table 3 reports averaged property values of polymers at standard condition (films or pieces from neat resin). The values of PE match the properties of medium-/ high-density PE the best. Process and manufacturing conditions can impact these property values. We note that the nearest neighbors search may be performed for any polymer with property values that deviate from the average values considered here. The nearest neighbors search algorithm and nearly 1.4 million bioplastic candidates, including the predicted properties, are shared on GitHub (see Code Availability section).

In the second step, we use our domain expertise to pick the most promising bio-replacement from the five candidates based on its potential to be synthesized. The most promising bio-replacements for each commodity plastic and for each of the two copolymer subgroups are reported in Fig. 4. The full list of bio-replacements (70) is provided as Supplementary Data 1.

Figure 5 a shows the property distributions of the bioplastic candidate set along with the properties of the seven commodity plastics (see Table 3) indicated as triangles. As expected, the T_m peak is shifted to higher temperatures (by around 80 K) with respect to the T_g peak. Among the mechanical properties, the densities of E and σ_b demonstrate a peak at around 540 MPa and 20 MPa, while ϵ_b shows a broad distribution, which covers most of the data range. Moreover, the majority of μ_{O_2} and μ_{CO_2} values are below 4 barrer and 20 barrer, which match the expected value range of this polymer class³⁵. Overall, all commodity plastics

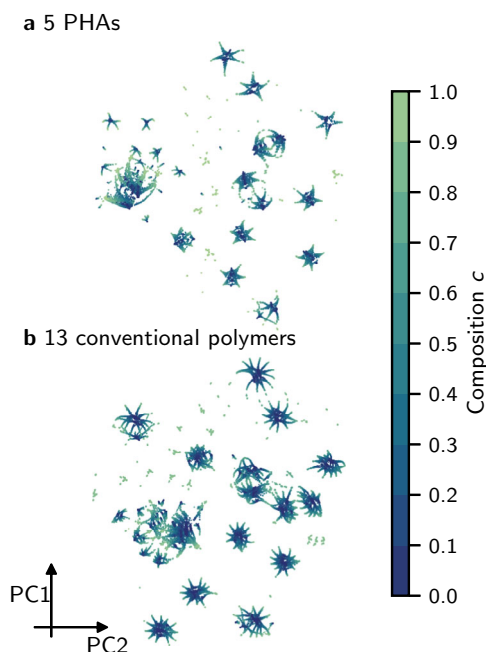


Fig. 2 Two-dimensional UMAP⁴¹ plot of two fingerprint subspaces that contain only fingerprints of selected polymers. a Five PHAs. The dark green dots ($c=1$) show the fingerprints of the SMILES strings [*] OCCC(=O) [*], [*] OC(O) CC(=O) [*], [*] OC(C(=O)O) CC(=O) [*], [*] OC(C(=O)OC) CC(=O) [*], and [*] OC(N) CC(=O) [*]. **b** 13 conventional polymers. The dark green dots show the fingerprints of the 13 conventional polymers. The dark blue dots ($c=0$) in panels **a** and **b** indicate the fingerprints of the remaining 548 and 540 polymers in the bioplastic search space (a total of 553 polymers), respectively. The dots with intermediate colors (green to blue) indicate the fingerprints of connecting copolymers. PC1 and PC2 represent the first and second principal components in the UMAP projection, respectively.

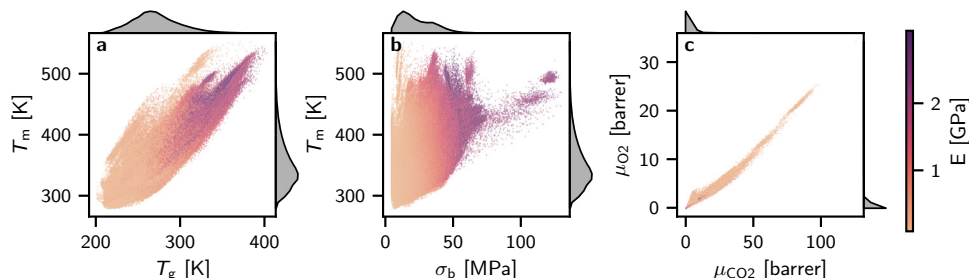
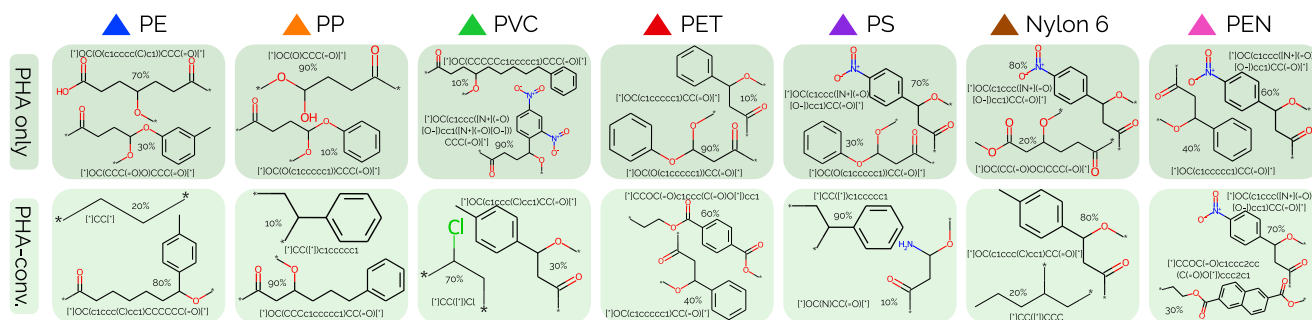


Fig. 3 Property relations of almost 1.4 million bioplastic candidates. T_m , T_g , σ_b , μ_{O_2} , μ_{CO_2} , and E in the panels **a–c** stand for melting temperature, glass transition temperature, tensile strength at break, O_2 gas permeability, CO_2 gas permeability, and Young's modulus, respectively. The data point densities are indicated in the plot margins.

Table 3 Measured properties of petroleum-based commodity plastics that in total account for 75.1% of Europe's yearly plastic production in 2019 (see usage column)².

| Polymer | Abb. | Applications | SMILES ^a | T _g [K] | T _m [K] | σ _b [MPa] | ε _b | E [MPa] | μ _{O₂} [barrer] | μ _{CO₂} [barrer] | Usage [%] |
|--------------------------------|--------|---|-------------------------------------|--------------------|--------------------|----------------------|----------------|---------|-------------------------------------|--------------------------------------|-----------|
| Poly(ethylene) | PE | Cloth packaging, shopping bags, waste bags | [*]CC[*] | 220 | 403 | 22.0 | 338.0 | 670 | 2.00 | 27.10 | 29.2 |
| Poly(propylene) | PP | Living hinges, pipes, caps, cutlery | [*]CC([*])C | 287 | 437 | 30.0 | 150.0 | 1600 | 0.76 | 4.40 | 19.3 |
| Poly(vinyl chloride) | PVC | Window frames, cables, pipes, films | [*]CC([*])C1 | 353 | 485 | 36.0 | 29.7 | 1680 | 0.06 | 0.23 | 9.9 |
| Poly(ethylene terephthalate) | PET | Bottles, automotive industry | [*]CCOC(=O)c1ccc(cc1)C(=O)O[*] | 350 | 526 | 119.0 | 65.0 | 2970 | 0.05 | 0.33 | 7.9 |
| Poly(styrene) | PS | Packaging fillers, cutlery, foam cups, take-out boxes | [*]CC([*])c1ccccc1 | 371 | 528 | 34.0 | 2.0 | 2450 | 2.60 | 12.60 | 6.8 |
| Poly(hexano-6-lactam) | Nylon6 | Yarns, fibers | [*]CCCCC(=O)N[*] | 324 | 493 | 60.0 | 61.0 | 1600 | 4.00 | 0.09 | 2 |
| Poly(ethylene 2,6-naphthalate) | PEN | Bottles, scintillators, medical product containers | [*]CCOC(=O)c1ccc2cc3ccccc3cc2c1O[*] | 357 | 541 | 77.0 | 42.0 | 2310 | 0.02 | 0.24 | |

^aThe two stars ([*]) indicate the endpoints of the polymer repeat unit.Property values are averaged over the entries in the PolyInfo repository³⁵ at standard conditions (film or piece from neat resin).**Fig. 4 PHA-only and PHA-conventional bio-replacements for seven commodity plastics.** Full polymer names are listed in Table 3.

(triangles) lie within the property ranges spanned by the bioplastic search space. However, because the triangles often lie in the tails of the property distributions, it is challenging (but possible) to find suitable replacements. Similar to Fig. 5a–h compare the experimental properties of the commodity plastics with their top bio-replacements identified in Fig. 4 in a radar chart. Qualitative graphical pairwise similarities between the property profiles in each radar chart indicate that the identified bio-replacements do indeed exhibit similar sets of properties with respect to the experimental properties. A comparison of the predicted and experimental properties of the seven commodity plastics is shown in Supplementary Fig. 10.

Synthesis Opportunities. It is interesting to note that all PHA-only and PHA-conventional bio-replacements in Fig. 4 contain aromatic groups in the side-chain. The biosynthesis of PHAs containing an aromatic monomer was first reported in 1990 by Fritzsche et al.⁴² for Poly(3-hydroxy-5-phenylvalerate) and since

then a wide range of aromatic side-chain functional groups have been introduced into PHAs through biosynthesis routes. For instance, see a recent comprehensive review by Ishii-Hyakutake et al.⁴³ for a more detailed discussion. Looking at the complexity of the chemistries that are already accessible via biosynthesis, prospects of producing the identified PHA-only bio-replacements are rather optimistic. For example, Aróstegui et al.⁴⁴ reported that the *Pseudomonas oleovorans* bacterium synthesizes PHAs with nitrophenyl side-chain functional groups, which occur in the PHA-only bio-replacements of PVC, PS, and PEN in Fig. 4. Moreover, engineering the bacterial PHA synthesis remains a highly active area of research with novel genome editing techniques, such as clustered regularly interspaced short palindromic repeats interference (CRISPRi),⁴⁵ that carry enormous potential for future breakthroughs in terms of both the accessible chemical diversity in PHA biosynthesis of homo- and copolymers as well as the yield optimization of the produced polymer chemistries.

Chemical synthesis routes for PHAs have been widely reported in the literature as well^{46–49}. The potential for the chemical synthesis

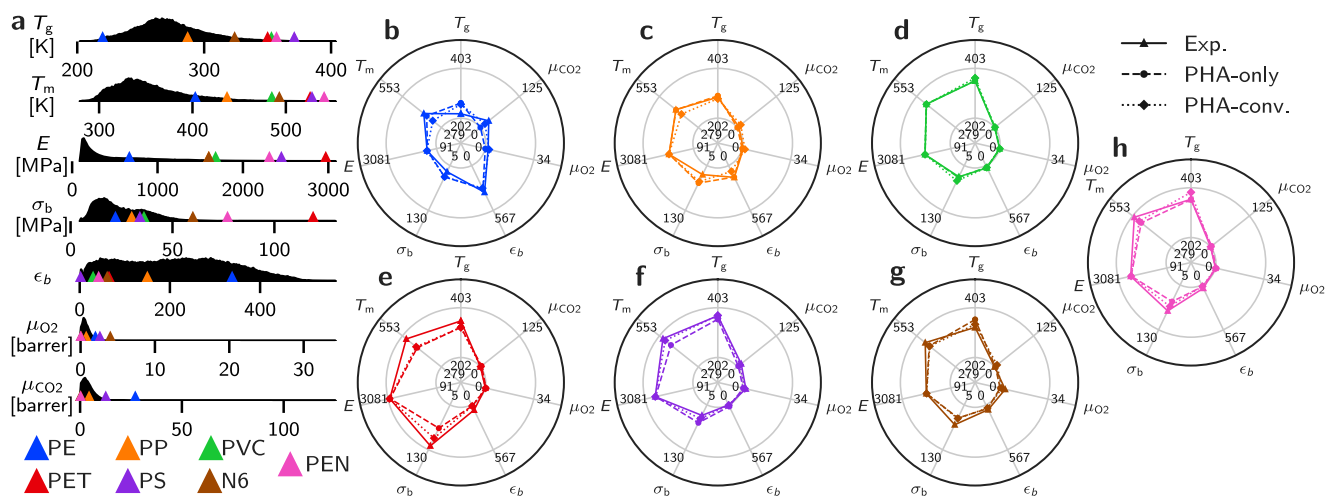


Fig. 5 Experimental and predicted properties. **a** Property density profiles computed over the entire prediction set of bioplastic candidates. Missing x-axes beyond a certain cutoff indicate zero predicted property densities over those property ranges. The triangles show experimental properties of the seven commodity plastics. Full polymer names are listed in Table 3. **b–h** Property radar charts for each commodity plastic. Triangles with solid lines show the experimental properties. Circles with dashed lines and diamonds with dotted lines indicate predicted properties of the bio-replacements in Fig. 4 for the copolymer subgroups of PHA-only and PHA-conventional polymers, respectively.

of the copolymers in this work (see Fig. 4 or Supplementary Data 1) lies in the ability to combine two comonomers of differing composition. Methods used in the synthesis of polystyrene-co-lactone copolymers^{50,51} can lead to the desired formation of the PHA-conventional bio-replacements of PP and PS. A chemical synthesis route for the PE and Nylon 6 PHA-conventional bio-replacements may follow similar steps used for the copolymerization of PHA/PEO (polyethylene oxide) copolymers¹⁹. The PVC, PET and PEN bio-replacements from PHA-conventional polymers may be produced via a reactive twin-screw extrusion process to form block structures containing phthalate-co-lactones.⁵² Alternatively, a copolymer of repeating phthalate/lactone units has recently been produced via a copolymerization reaction of ϵ -lactone with degraded PET^{53,54}. We believe the predictions based on our work have potential to translate into new PHA biopolymers or copolymers and will inspire development of new PHA-only or hybrid conventional-PHA polymer synthesis routes.

The developed property predictors used in this work have several limitations which are largely tied to the availability of training data. The property predictors ignore process and manufacturing conditions as well as certain details of morphology (e.g., % crystallinity). Furthermore, other relevant factors such as the molecular weight distribution, form factors (e.g., linear versus branched), additives, etc. are currently not integrated in the property prediction pipeline. In addition, low level configurational trends across diverse chemistries such as subtle variations stemming from chain morphologies or different relative placements of distinct monomer units across the chains are not accounted for. If such information becomes available in the future, the developed predictors can easily be updated to account for these additional effects. The proposed informatics pipeline should be considered an essential first step to solve the sustainable polymer design problem that calls for future improvements such as the development of improved property predictors, integrated synthesizability criteria, and above all, process optimization protocols to make the designed polymers a practical reality.

Conclusion

We have developed an informatics-based bioplastic design pipeline, which has identified promising PHA-based bioplastic replacements for seven petroleum-based commodity plastics. Our study starts

with the data collection and curation of approximately 23,000 homo- and copolymer data points spanning 13 properties critical for everyday applications and use. Multitask neural networks with a meta learner, pioneered by us for polymer informatics, forecast thermal, mechanical, and gas permeability properties for polymers over a broad chemical space with unprecedented performance. Using the trained models, we predict the 13 key thermal, mechanical, and gas permeability properties of all polymers in a bioplastic search space of almost 1.4 million polymers. The property predictions are validated and subsequently utilized to find bio-replacements for seven commodity plastics that, in total, account for more than 75% of the yearly plastic production. Using a two-step selection protocol of a nearest neighbors search and synthesizability criteria, we propose two bio-replacements for each commodity plastic and discuss chemical synthesis and biosynthesis routes for these promising polymer replacements. Informatics can help to identify suitable synthesis strategies as well⁵⁵.

The implications of this work are far-reaching. We currently produce by far more plastics than we can recycle,⁴ and the demand for plastics is expected to continue to grow at an annual rate of 4%.⁵⁶ As countries begin to implement restrictions on plastic use, there is an urgent need for bioplastic alternatives to conventional plastics. Yet, the options of commercially available biopolymers are currently very limited.⁵⁷ Our approach to design and discover functional biopolymers can be applied to greatly accelerate the replacement of conventional plastic materials with more sustainable alternatives, and with possibly even greater performance advantages. The candidate biopolymers, in particular PHAs, might be synthesized by known chemical or biosynthetic routes, hybrid routes, or routes yet to be developed. Our approach can augment conventional empirically based design approaches by guiding the way to more targeted experiments, fewer experimental trials, or shorter times to market. Our work provides an informatics-based screening tool for researchers and developers aiming to produce bioplastics with improved thermomechanical and transport properties for better performance in specific applications, thus accelerating the transition to a circular economy.

Methods

Fingerprinting. The fingerprinting process converts geometric and chemical information of polymers to machine-readable numerical representations for

training machine learning models. Homopolymer repeat units are represented as simplified molecular-input line-entry system (SMILES)⁵⁸ strings that use two stars to indicate the two endpoints of the repetitive unit of the polymers, but otherwise follow the SMILES syntax. The fingerprint vector (**F**) in this work has 849 components and is calculated based upon the SMILES string following a two-step protocol³⁴. First, we compute hierarchical fingerprints that capture structural and key chemical features of each comonomer at three different length scales^{59,60}. At the atomic scale, our fingerprints track the occurrence of a fixed set of atomic fragments (or motifs)⁶¹. For example, the fragment “C3-S2-C3” is composed of three contiguous atoms, in this order, a three-fold coordinated carbon, a two-fold coordinated sulfur, and a three-fold coordinated carbon. A vector of such triplets represents the fingerprint components at the lowest hierarchy. The next level uses the quantitative structure-property relationship (QSPR) fingerprints⁶² to capture features on larger length-scales. QSPR fingerprints are often used in chemical and biological sciences, and used here as implemented in the chem informatics toolkit RDKit⁶³. Examples of such fingerprints are the van der Waals surface area⁶⁴, the topological polar surface area (TPSA)^{65,66}, the fraction of atoms that are part of rings (i.e., the number of atoms associated with rings divided by the total number of atoms in the formula unit), and the fraction of rotatable bonds. The highest length-scale fingerprint components in our polymer fingerprints deal with “morphological descriptors”. They include features such as the shortest topological distance between rings, the fraction of atoms that are part of side-chains, and the length of the largest side-chain³⁰.

Second, we sum the composition-weighted comonomer fingerprints to compute the total copolymer fingerprint vector $\mathcal{F} = \sum_i^N \mathbf{F}_i c_i$, where N is the number of comonomers in the copolymer, \mathbf{F}_i the i th comonomer fingerprint, and c_i the fraction of the i th comonomer. This copolymer fingerprint satisfies the two main demands of uniqueness and invariance to different (but equivalent) periodic unit specifications and renders the fingerprinting routine invariant to the order in which one may sort the comonomers. Contrary to homopolymer fingerprints, copolymer fingerprints may not be interpretable (e.g., the composition-weighted sum of the fingerprint component “shortest distance between rings” of two homopolymers has no physical meaning). In our work, all copolymer data points are of random copolymers, and alternating copolymers were treated as homopolymers.

Multitask predictor and meta learner. Multitask deep neural networks simultaneously learn multiple polymer properties to utilize possible inherent correlations in data. Figure 1b schematically portrays the architecture of the three concatenation-conditioned multitask predictors: the copolymer fingerprint and selector vector are fed to a feed-forward deep neural network that outputs a single property. The selector vector is a binary vector and specifies the output property. For instance, the selector vector of the thermal properties predictor (\mathbf{S}_1) has three components and encodes T_g as [100], T_m as [010], and T_d as [001]. All parameters of the neural networks, such as the number of layers, number of nodes, dropout rates, and activation functions, are optimized using the Hyperband method⁶⁷ of the Python package KerasTuner⁶⁸. Final parameters are reported in Supplementary Table 1. All models were implemented using the Python API of TensorFlow⁶⁹.

The training protocol of the predictors follows state-of-the-art techniques involving five-fold cross-validation and a meta learner that forecasts the final property values based upon the ensemble of cross-validation predictors³⁴ (see also Supplementary Fig. 1). The parameters of the cross-validation models are fixed when used in the meta learner. The meta learner has the same network architecture as the multitask predictors but receives the five outputs of the multitask predictors as inputs (rather than the copolymer fingerprint). The cross-validation process ensures that each data point has at least once been in the validation data set and allows us to report the generalization error as averaged RMSEs and R^2 s of the validation data sets. The three meta learners operate as production predictors. After shuffling, the data set was split into two parts. 20% of the data set was set aside for training the meta learners, while the remaining 80% was used for five-fold cross-validation and the validation of the meta learner. All data set splits were stratified by the properties.

Data availability

All the polymer data used in this work to train the various property prediction models can be found in the PoLYInfo database <https://polymer.nims.go.jp/en/> (National Institute for Materials Science (NIMS) holds the copyright of this database system).

Code availability

The Python code for creating the 1 373 503 bioplastic candidates, the bioplastic candidates with property predictions, the code for searching the bio-replacements, and the 70 predicted bio-replacements for seven commodity plastics are available at https://github.com/Ramprasad-Group/bioplastic_design. Production models are deployed at <https://PolymerGenome.org>.

Received: 23 March 2022; Accepted: 10 November 2022;

Published online: 03 December 2022

References

- Satti, S. M. & Shah, A. A. Polyester-based biodegradable plastics: an approach towards sustainable development. *Lett. Appl. Microbiol.* **70**, 413–430 (2020).
- Plasticseurope. <https://www.plasticseurope.org/en/resources/market-data>.
- Naser, A. Z., Deiab, I. & Darras, B. M. Poly(lactic acid) (PLA) and polyhydroxyalkanoates (PHAs), green alternatives to petroleum-based plastics: a review. *RSC Adv.* **11**, 17151–17196 (2021).
- Geyer, R., Jambeck, J. R. & Law, K. L. Production, use, and fate of all plastics ever made. *Sci. Adv.* **3**, 25–29 (2017).
- Lim, X. Microplastics are everywhere—but are they harmful? *Nature* **593**, 22–25 (2021).
- Poltronieri, P. & Kumar, P. *Polyhydroxyalkanoates (PHAs) in Industrial Applications*; Springer International Publishing: pp 2843–2872, https://doi.org/10.1007/978-3-319-68255-6_70 (2019).
- Go!PHA: The Global Organisation for PHA. <https://www.gopha.org>.
- Koller, M. & Mukherjee, A. A new wave of industrialization of PHA biopolyesters. *Bioengineering* **9**, 74 (2022).
- Follain, N. et al. Structure and barrier properties of biodegradable polyhydroxyalkanoate films. *J. Phys. Chem. C* **118**, 6165–6177 (2014).
- Pilania, G., Iverson, C. N., Lookman, T. & Marrone, B. L. Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. *J. Chem. Inf. Model.* **59**, 5013–5025 (2019).
- Bejagam, K. K., Iverson, C. N., Marrone, B. L. & Pilania, G. Molecular dynamics simulations for glass transition temperature predictions of polyhydroxyalkanoate biopolymers. *Phys. Chem. Chem. Phys.* **22**, 17880–17889 (2020).
- Bejagam, K. K., Iverson, C. N., Marrone, B. L. & Pilania, G. Composition and configuration dependence of glass-transition temperature in binary copolymers and blends of polyhydroxyalkanoate biopolymers. *Macromolecules* **54**, 5618–5628 (2021).
- Bejagam, K. K. et al. Predicting the mechanical response of polyhydroxyalkanoate biopolymers using molecular dynamics simulations. *Polymers* **14**, 345 (2022).
- Bejagam, K. K., Lalonde, J., Iverson, C. N., Marrone, B. L. & Pilania, G. Machine learning for melting temperature predictions and design in polyhydroxyalkanoate-based biopolymers. *J. Phys. Chem. B* **126**, 934–945 (2022).
- Sharma, V., Sehgal, R. & Gupta, R. Polyhydroxyalkanoate (PHA): Properties and modifications. *Polymer* **212**, 123161 (2021).
- Pryadko, A., Surmeneva, M. A. & Surmenev, R. A. Review of hybrid materials based on polyhydroxyalkanoates for tissue engineering applications. *Polymers* **13**, 1738 (2021).
- McAdam, B., Brennan Fournet, M., McDonald, P. & Mojicevic, M. Production of Polyhydroxybutyrate (PHB) and factors impacting its chemical and mechanical characteristics. *Polymers* **12**, 2908 (2020).
- Jiang, Z., Hu, J., Marrone, B. L., Pilania, G. & Yu, X. B. A deep neural network for accurate and robust prediction of the glass transition temperature of polyhydroxyalkanoate homo- and copolymers. *Materials* **13**, 5701 (2020).
- Winnacker, M. & Rieger, B. Copolymers of polyhydroxyalkanoates and polyethylene glycols: recent advancements with biological and medical significance. *Polym. Int.* **66**, 497–503 (2017).
- Albright, V. C. & Chai, Y. Knowledge gaps in polymer biodegradation research. *Environ. Sci. Technol.* **55**, 11476–11488 (2021).
- Chen, L. et al. Polymer informatics: Current status and critical next steps. *Mater. Sci. Eng. R: Rep.* **144**, 100595 (2021).
- Batra, R., Song, L. & Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **6**, 655–678 (2021).
- Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
- Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
- Peerless, J. S., Milliken, N. J. B., Oweida, T. J., Manning, M. D. & Yingling, Y. G. Soft matter informatics: current progress and challenges. *Adv. Theory Simul.* **2**, 1800129 (2019).
- Adams, N. & Murray-Rust, P. Engineering polymer informatics: towards the computer-aided design of polymers. *Macromol. Rapid Commun.* **29**, 615–632 (2008).
- Zhong, S. et al. Machine learning: new ideas and tools in environmental science and engineering. *Environ. Sci. Technol.*, acs.est.1c01339, <https://doi.org/10.1021/acs.est.1c01339>, (2021).

28. Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B., & Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **12**, 2312 (2021).
29. Barnett, J. W. et al. Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.*, **6**, <https://doi.org/10.1126/sciadv.aaz4301>, (2020).
30. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).
31. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).
32. Jha, A., Chandrasekaran, A., Kim, C. & Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures. *Model. Simul. Mater. Sci. Eng.* **27**, 24002 (2019).
33. Kuenneth, C. et al. Polymer informatics with multi-task learning. *Patterns* **2**, 100238 (2021).
34. Kuenneth, C., Schertzer, W. & Ramprasad, R. Copolymer informatics with multitask deep neural networks. *Macromolecules* **54**, 5957–5961 (2021).
35. PoLyInfo. <https://polymer.nims.go.jp/en/>.
36. Varoquaux, G. et al. Scikit-learn. *GetMobile: Mob. Comput. Commun.* **19**, 29–33 (2015).
37. Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **5**, 66 (2019).
38. Tao, L., Byrnes, J., Varshney, V. & Li, Y. Machine learning strategies for the structure-property relationship of copolymers. *iScience* **25**, 104585 (2022).
39. Chen, G., Tao, L. & Li, Y. Predicting polymers' glass transition temperature by a chemical language processing model. *Polymers* **13**, 1898 (2021).
40. Tao, L., Varshney, V. & Li, Y. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *J. Chem. Inf. Model.* **61**, 5395–5413 (2021).
41. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction, <https://doi.org/10.48550/arXiv.1802.03426> (2018).
42. Fritzsche, K., Lenz, R. W. & Fuller, R. C. An unusual bacterial polyester with a phenyl pendant group. *Die Makromolekulare Chemie* **191**, 1957–1965 (1990).
43. Ishii-Hyakutake, M., Mizuno, S. & Tsuge, T. Biosynthesis and characteristics of aromatic polyhydroxyalkanoates. *Polymers* **10**, 1267 (2018).
44. Aróstegui, S. M., Aponte, M. A., Díaz, E. & Schröder, E. Bacterial polyesters produced by pseudomonas oleovorans containing nitrophenyl groups. *Macromolecules* **32**, 2889–2895 (1999).
45. Meng, D.-C. & Chen, G.-Q. Synthetic biology of polyhydroxyalkanoates (PHA). *Synthetic Biology-metab. Eng.*, 147–174, https://doi.org/10.1007/10_2017_3 (2017).
46. Westlie, A. H. & Chen, E. Y.-X. Catalyzed chemical synthesis of unnatural aromatic polyhydroxyalkanoate and aromatic-aliphatic PHAs with record-high glass-transition and decomposition temperatures. *Macromolecules* **53**, 9906–9915 (2020).
47. Okada, M. Chemical syntheses of biodegradable polymers. *Prog. Polym. Sci.* **27**, 87–133 (2002).
48. Shi, C. et al. Hybrid monomer design for unifying conflicting polymerizability, recyclability, and performance properties. *Chem* **7**, 670–685 (2021).
49. Tang, X. et al. Biodegradable polyhydroxyalkanoates by stereoselective copolymerization of racemic diolides: stereocontrol and polyolefin-like properties. *Angewandte Chemie Int. Ed.* **59**, 7881–7890 (2020).
50. Herman, J.-J., Jérôme, R., Teyssié, P., Gervais, M. & Gallot, B. Structural Study on styrene/ ϵ -caprolactone block copolymers in absence and in presence of a solvent of the polystyrene block. *Die Makromolekulare Chemie* **182**, 997–1008 (1981).
51. Zalusky, A. S., Olayo-Valles, R., Wolf, J. H. & Hillmyer, M. A. Ordered nanoporous polymers from polystyrene-poly lactide block copolymers. *J. Am. Chem. Soc.* **124**, 12761–12773 (2002).
52. Tang, W., Murthy, N. S., Mares, F., McDonnell, M. E. & Curran, S. A. Poly(ethylene terephthalate)-poly(caprolactone) block copolymer. I. Synthesis, reactive extrusion, and fiber morphology. *J. Appl. Polym. Sci.* **74**, 1858–1867 (1999).
53. Espinoza-García, K. et al. Polymerization of ϵ -caprolactone with degraded PET for its functionalization. *J. Polym. Res.* **26**, 180 (2019).
54. Ben Gara, M., Kammoun, W., Delaite, C., Abid, S. & El Gharbi, R. Synthesis and characterization of aliphatic-aromatic copolyesters from pet waste and ϵ -Caprolactone. *Journal of Macromolecular Science, Part A* **52**, 454–464 (2015).
55. Chen, L., Kern, J., Lightstone, J. P. & Ramprasad, R. Data-assisted polymer retrosynthesis planning. *Appl. Phys. Rev.* **8**, 031405 (2021).
56. Rosenboom, J.-G., Langer, R. & Traverso, G. Bioplastics for a circular economy. *Nat. Rev. Mater.* **7**, 117–137 (2022).
57. Ghosh, K. & Jones, B. H. Roadmap to biodegradable plastics—current state and research needs. *ACS Sustain. Chem. Eng.* **9**, 6170–6187 (2021).
58. Weininger, D. SMILES, a chemical language and information system: 1: Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
59. Mannodi-Kanakthodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Scientific Reports* **6**, 20952 (2016).
60. Doan Tran, H. et al. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **128**, 171104 (2020).
61. Huan, T. D., Mannodi-Kanakthodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B - Condens. Matter Mater. Phys.* **92**, 1–10 (2015).
62. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative structure-property relationship modeling of diverse materials properties. *Chem. Rev.* **112**, 2889–2919 (2012).
63. Landrum, G. RDKit. <http://www.rdkit.org>.
64. Iler, N., Rowitch, D. H., Echelard, Y., McMahon, A. P. & Abate-Shen, C. A single homeodomain binding site restricts spatial expression of Wnt-1 in the developing brain. *Mech. Dev.* **53**, 87–96 (1995).
65. Ertl, P., Rohde, B. & Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **43**, 3714–3717 (2000).
66. Prasanna, S. & Doerksen, R. Topological polar surface area: a useful descriptor in 2D-QSAR. *Curr. Med. Chem.* **16**, 21–41 (2008).
67. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 1–52 (2018).
68. O'Malley, T. Keras Tuner. <https://github.com/keras-team/keras-tuner>, (2019).
69. Martin, A. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, <https://www.tensorflow.org/>, (2015).

Acknowledgements

C.K. thanks the Alexander von Humboldt Foundation for financial support. R.R. acknowledges support from the Office of Naval Research through a Multi-University Research Initiative (MURI) grant (N00014-17-1-2656) and a regular grant (N00014-20-2175). G.P., B.L.M., and C.N.I. acknowledge support from the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) program's project titled Bio-Manufacturing with Intelligent Adaptive Control (BioManIAC) #20190001DR. J.L. gratefully acknowledges support via a LANL Center for Nonlinear Studies (CNLS) Summer 2021 Fellowship Award. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

Author contributions

C.K. designed, trained and evaluated the machine learning models. Numerous discussions with C.N.I. and B.L.M. at an early stage helped in defining the scope of the study. J.L. provided input for the background information and copolymer examples. The work was conceived and guided by R.R. and G.P. All authors discussed the results and commented on the manuscript.

Competing interests

Ghanshyam Pilania is a Guest Editor for Communications Materials and was not involved in the editorial review of, or the decision to publish, this Article. The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43246-022-00319-2>.

Correspondence and requests for materials should be addressed to Ghanshyam Pilania.

Peer review information Communications Materials thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: John Plummer. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022