Macromolecules

pubs.acs.org/Macromolecules Article

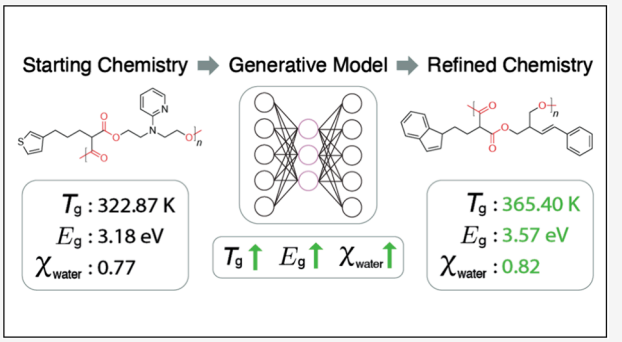
# Generative Multi-Property Refinement of Polymer Chemistries

Seonghwan Kim, Charles M. Schroeder, and Nicholas E. Jackson*

Cite This: Macromolecules 2026, 59, 6759−6771 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** We present a generative, multiobjective optimization method for polymer chemistry. By leveraging monomer-level properties that are correlated with polymer properties, we design step-growth polymers with targeted glass transition temperatures ($T_g$), band gaps ($E_g$), and Flory−Huggins interaction parameters with water ($\chi_{water}$) across a broad chemical space. Generative design is accomplished using a variational autoencoder integrated with linear property prediction heads. Linear organization of the latent space enables the identification of a single latent vector to steer the simultaneous optimization of multiple polymer property objectives. Subsequent Bayesian optimization within the latent space allows further enhancement of $T_g$, $E_g$, and $\chi_{water}$ relative to a reference polymer chemistry. We then apply the generative model to design per- and polyfluoroalkyl substances (PFAS)-based polymers with reduced fluorine content but comparable physical properties. Overall, this work establishes a generative, multiobjective approach for navigating early stage polymer design, prior to experimental validation or computationally expensive simulations.





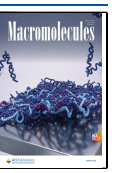

## INTRODUCTION

Practical polymer design necessitates the simultaneous optimization of numerous physical and chemical properties, spanning mechanical, thermodynamic, and electronic attributes, in addition to sustainability and economic considerations.[1−8] Many of these target properties, such as long-term mechanical stability or electronic transport, require laborious experiments or costly simulations, limiting the pace of chemical space exploration.[9−11] Although high-throughput experimentation[12−14] and computational modeling[15−18] have accelerated property prediction, exploration of the vast polymer chemical space remains infeasible due to monomer identity, copolymer sequence, and architectural variations. Moreover, simultaneous optimization of multiple functionalities is challenging because many key attributes[19,20] are intrinsically coupled and often exhibit trade-offs. These challenges highlight the need for computational models capable of guiding early stage polymer discovery across multiple property objectives.

In recent years, generative machine learning (ML) has emerged as a powerful approach for materials discovery by enabling inverse design and efficient navigation of large molecular structure−property landscapes. A variety of generative models—including variational autoencoders (VAEs),[21−23] generative adversarial networks (GANs),[24−26] autoregressive models,[27,28] large language models (LLMs),[29,30] and diffusion models[31,32]—have been developed for molecular design. Although generative ML was first widely developed for relatively data-rich small organic molecules, recent work has extended this approach to functional polymer design. For example, polymer dielectrics with high dielectric breakdown strength have been discovered by simultaneously optimizing the glass transition temperature ($T_g$), band gap ($E_g$), and electron injection barrier.[33] Diverse polymer architectures such as linear, cyclic, branched, comb, star, and dendrimer structures have been designed to achieve a targeted mean-square radius of gyration.[34] Conjugated copolymers have been designed as organic photocatalysts by tuning monomer chemistries and stoichiometries to control their electronic properties.[35] LLM-based generative models have been demonstrated for functional polymer design including the optimization of $E_g$,[36] $T_g$,[37] and ionic conductivity.[38] Despite these advances, existing generative polymer models remain constrained by the limited availability of polymer structure−property data, which hinders the development of reliable, scalable tools for designing polymers subject to multiple physical and chemical property objectives.[39]

Over the past decade, the polymer community has made substantial efforts to construct comprehensive databases of polymer structures and properties to support data-driven design of functional polymers.[40−65] Several resources such as PolyInfo,[40] PolyID,[53] and Polymer Scholar[59] have curated

Received: February 20, 2026
Revised: April 10, 2026
Accepted: May 15, 2026
Published: June 1, 2026



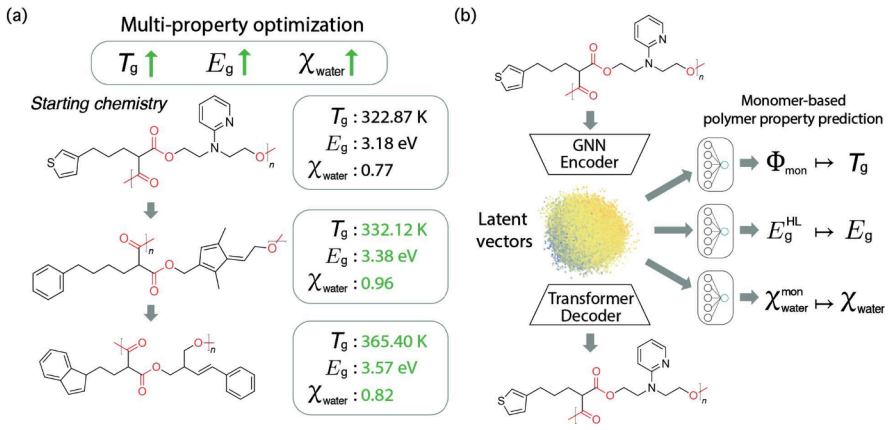



**Figure 1.** Multiproperty design of linear homopolymers using variational autoencoder (VAE)-based generative machine learning (ML). (a) Gradual molecular structure refinement for a simultaneous increase in glass transition temperature ($T_g$), band gap ($E_g$), and Flory–Huggins $\chi$ interaction parameter with water as a solvent ($\chi_{water}$). Green text indicates enhanced property values. (b) VAE architecture comprising a graph neural network (GNN) encoder and a Transformer decoder with property predictions enabled from latent vectors. The VAE was trained on computed monomer-level properties—structural flexibility ($\Phi_{mon}$), HOMO and LUMO energies ($E_{HOMO}$, $E_{LUMO}$), and Flory–Huggins $\chi$ interaction parameter from monomer-based calculations ($\chi_{water}^{mon}$)—to estimate polymer properties either from experiments or computation. The monomer HOMO–LUMO gap ($E_g^{HL}$) was computed as the difference between $E_{HOMO}$ and $E_{LUMO}$.

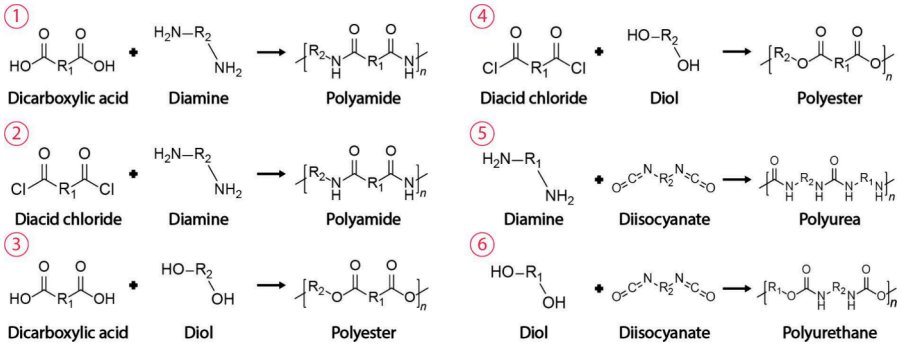



**Figure 2.** Six polymerization mechanisms spanning the chemical space of approximately 600000 step-growth polymers within the Open Macromolecular Genome (OMG) for generative multiproperty polymer design. These polymerization mechanisms enable the generation of hypothetically synthesizable linear homopolymers by identifying certain functional groups (e.g., dicarboxylic acid) in commercially available reactants and applying rule-based polymerization reactions between compatible reactants.

experimental polymerization data from peer-reviewed literature. Complementing these efforts, databases including the Open Macromolecular Genome (OMG),[54] PolyUniverse,[56] and SMiPoly[55] have enumerated large sets of hypothetically synthesizable polymers to expand the accessible chemical design space. Recently, large-scale computational initiatives have emerged to address the scarcity of polymer data. For example, PolyOmics[65] reports 62 physical properties for more than 100000 polymers, and OPoly26[66] provides diverse polymer structures to support the development of ML interatomic potentials. Despite these active efforts, the inherent complexity of polymers (e.g., monomer chemistry, sequence, architecture, and morphology) continues to impede the construction of databases that map physical properties across chemical space.

One strategy for overcoming data scarcity is to approximate polymer functional properties using computationally accessible monomer-level descriptors. Because monomer chemistry fundamentally affects polymer conformational, electronic, and solubility behavior, monomer properties may serve as efficient proxies for polymer properties when direct polymer data are unavailable. This approach enables rapid screening across vast chemical spaces by filtering candidate polymers based on inexpensive monomer calculations before pursuing more costly polymer experiments or computational simulations. In prior work,[67] we estimated diverse monomer-level properties for approximately 12 million synthetically accessible linear homopolymers[54] using quantum chemistry calculations and uncertainty-based active learning. By comparing these monomer-level descriptors against experimentally measured and computationally predicted polymer properties, we identified strong correlations between several monomer-level features and bulk polymer properties. Although monomer-level properties cannot fully capture polymer functionality, which ultimately depends on molecular weight distributions and morphology, they nonetheless provide an efficient and scalable framework for screening vast chemical spaces in early stage, multiobjective polymer design.

In this work, we develop a generative framework for multiobjective polymer design using monomer-level properties to efficiently navigate a chemical space of approximately 600000 hypothetically synthesizable step-growth polymers including polyesters, polyamides, polyureas, and polyurethanes. We demonstrate this generative approach through multiobjective polymer design of high $T_g$, high $E_g$, and enhanced hydrophobicity (characterized by the Flory–Huggins param-


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759–6771

eter $\chi_{water}$), starting from a chosen reference polymer chemistry (Figure 1a). These properties are selected because they respectively capture thermal stability ($T_g$), electronic stability ($E_g$), and polymer−solvent phase behavior ($\chi_{water}$), which together govern materials performance and processability in functional polymers. We first introduce our VAE-based generative model with a graph neural network (GNN) encoder and a Transformer-based decoder, coupled with linear prediction heads for monomer-level properties. We then describe our latent space search strategy that relies on a combination of Bayesian optimization (BO) and linear organization of the latent space. We then assess the performance of the model and the transferability by applying it to the multiobjective design of per- and polyfluoroalkyl substances (PFAS)-based polymers aimed at reducing fluorine content. Overall, this work demonstrates that generative modeling combined with monomer-level properties provides an efficient and scalable route to explore large chemical spaces, thereby enabling early stage multiobjective polymer design.

## METHODS

### Scoping Chemical Space and Functionality

For generative polymer design, we utilized the chemical space of approximately 600000 hypothetical step-growth polymers constructed via six distinct polymerization mechanisms that are widely used in academia and industry (Figure 2).[68] These step-growth polymers were selected from approximately 12 million hypothetically synthesizable linear homopolymers in the OMG.[54] The OMG database includes hypothetical step-growth polymers generated from commercially available reactants that conform to one of six polymerization mechanisms: the reactions between (1) dicarboxylic acid and diamine (polyamide), (2) diacid chloride and diamine (polyamide), (3) dicarboxylic acid and diol (polyester), (4) diacid chloride and diol (polyester), (5) diamine and diisocyanate (polyurea), and (6) diol and diisocyanate (polyurethane). Although the polymers in the OMG database were derived from polymerization reactions involving commercially available reactants, these chemistries are not guaranteed to be synthesizable due to the practical complexities of polymer synthesis. For example, the intended polymerization mechanism might not proceed as expected when monomer reactants contain more than one type of polymerizable functional group—such as both amine and hydroxyl groups capable of nucleophilic attack.[54] To improve synthetic feasibility, we filtered the OMG step-growth polymers to retain only those formed from monomer reactants containing a single type of polymerization functional group, thereby enforcing a single intended polymerization mechanism. In addition, we restricted the monomer reactant structures to be symmetric around their polymerization functional groups to eliminate ambiguity in reaction ordering (e.g., A−A rather than A−A′, where A and A′ represent polymerizable groups). Following this filtering process, approximately 750000 OMG step-growth polymers were selected to form a full data set representing a hypothetical chemical space of synthetically accessible step-growth polymers. These polymers were partitioned into training (600000), validation (75000), and test (75000) data sets for generative modeling. Details of the chemical composition of the training, validation, and test data sets are provided in the Supporting Information (Figure S1).

We subsequently approximated the properties of the entire data set of 750000 OMG step-growth polymers using previously computed monomer-level properties,[67] enabling efficient exploration of the polymer property space during generative modeling. Notably, several of these monomer-level properties were shown to exhibit strong linear correlations with experimentally measured or computationally predicted polymer properties.[67] For example, monomer structural flexibility Φ index ($\Phi_{mon}$) exhibits a strong negative linear correlation ($\rho \approx -0.76$) with experimental $T_g$, where a higher $\Phi_{mon}$ indicates increased structural flexibility (Figure S2). Similarly, the energy gap between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of a monomer ($E_g^{HL}$) also possesses a strong positive linear correlation ($\rho \approx 0.89$) with the polymer bulk $E_g$ computed using density functional theory (DFT; Figure S3). In addition, Flory−Huggins $\chi$ interaction parameters with water as a solvent estimated from monomer-level calculations ($\chi_{water}^{mon}$) show a strong positive linear correlation with experimental $\chi_{water}$ ($\rho \approx 0.94$; Figure S4). These correlations between monomer-level and polymer properties were used to approximate $T_g$, $E_g$, and $\chi_{water}$ of the entire data set of 750000 OMG step-growth polymers. This strategy enables early stage functional polymer design by efficiently exploring a large polymer chemical and property space based on computationally inexpensive monomer-level properties, without the need for experimental or computational determination of polymer properties.

### Generative Model

A generative ML model based on a VAE was employed to design linear homopolymers simultaneously optimized for $T_g$, $E_g$, and $\chi_{water}$ (Figure 1b). The VAE[21,22,34,35,54] constructs a low-dimensional, continuous latent space by compressing molecular structures into latent vectors through an encoder network and reconstructing the original molecular structures from these latent vectors using a decoder network. By jointly learning to predict molecular properties from the latent vectors, the VAE organizes the latent space according to structure−property relationships. Once trained, the VAE enables generative molecular design by identifying latent vectors associated with desirable predicted properties and decoding those latent vectors into corresponding molecular structures.

Following prior work,[35] our VAE architecture integrates a GNN encoder with a Transformer decoder. The GNN encoder, which was implemented using Chemprop,[69] encodes methyl-terminated polymer constitutional repeating units (CRUs)[70] into continuous latent vectors by propagating atomic and bonding information through message passing. The Transformer decoder built with OpenNMT[71] generates sequential Group SELFIES representations[72] via cross-attention with the latent vectors. We used Group SELFIES as the decoding representation to treat a polymerizable linkage (e.g., ester or urethane group) as a single token, thereby preserving essential polymerization motifs required for step-growth chemistries during CRU generation. For property prediction, we implemented a separate linear regression head for each monomer-level property including $\Phi_{mon}$, $E_{HOMO}$, $E_{LUMO}$, and $\chi_{water}^{mon}$. The HOMO−LUMO gap, $E_g^{HL}$, was computed as the difference between $E_{HOMO}$ and $E_{LUMO}$. This linear property prediction enforces a linear organization of the latent space with respect to monomer-level properties. Although not explicitly optimized in this work, additional monomer-level properties such as the lowest singlet excited-state energy ($E_{S1}$) were incorporated as additional tasks during model training to support comprehensive property coverage (Table S1). The resulting linearly organized latent space[73] enhanced the interpretability of learned monomer structure−property relationships (vide infra). Details on the use of Group SELFIES are provided in the Supporting Information.

We trained the VAE on approximately 600000 OMG step-growth polymer CRUs from the training data set with their monomer-level properties. Similar to prior work,[54] the VAE was trained to (i) maximize reconstruction accuracy of polymer CRUs via encoding and decoding, (ii) regularize the latent space to a prior Gaussian distribution, and (iii) minimize prediction error for monomer-level properties. Model hyperparameters were optimized using Optuna[74] parallelized across multiple GPUs using Dask.[75] The explored hyperparameter search space is summarized in the Supporting Information (Figure S5). It is important to note that the trained VAE should learn meaningful latent representations that can be decoded into chemically valid CRUs during the generation phase. To achieve high generative performance, after hyperparameter optimization we selected the final model by jointly considering reconstruction accuracy on the validation set and its ability to generate chemically valid CRUs when sampling from the prior Gaussian distribution. This model selection strategy balances reconstruction fidelity with


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

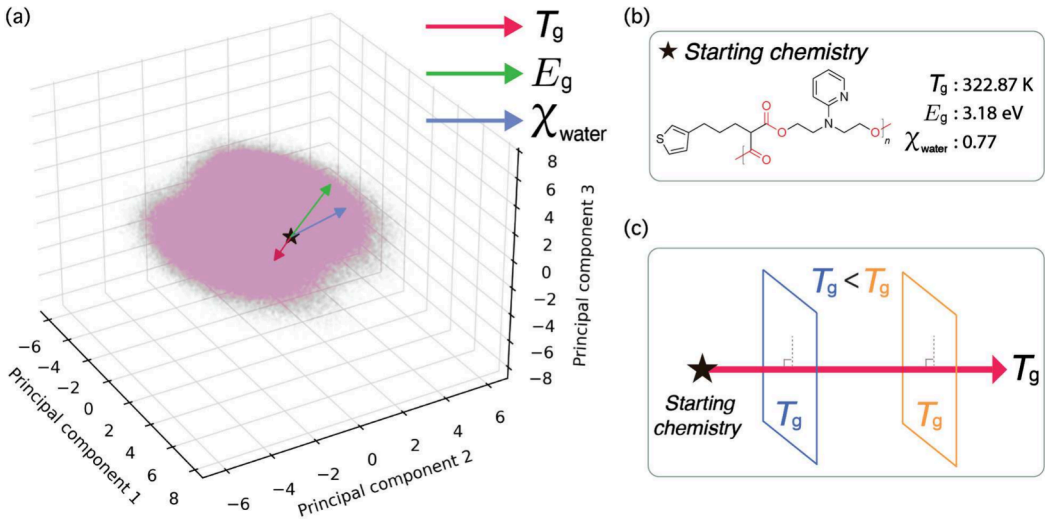



**Figure 3.** Linearly organized latent space of the VAE trained on 600000 step-growth polymers and their monomer-level properties. (a) 3D projection of the latent space via principal component analysis (PCA) with arrows indicating increasing directions for $T_g$ (red), $E_g$ (green), and $\chi_{water}$ (blue). (b) Example starting polymer chemistry with estimated $T_g$, $E_g$, and $\chi_{water}$ from monomer-level properties predicted by the trained VAE. (c) Schematic illustrating the linear organization of the latent space with respect to monomer-level properties (e.g., $\Phi_{mon}$), yielding a monotonic trend in polymer properties (e.g., $T_g$).

generative quality. The trained VAE achieved a test reconstruction accuracy of 69.13% for polymer CRUs and $R^2$ value above 0.9 for all monomer-level property predictions (Figure S6). For example, the trained model predicted $\Phi_{mon}$ with $R^2 \approx 0.98$, $E_{HOMO}$ with $R^2 \approx 0.96$, $E_{LUMO}$ with $R^2 \approx 0.97$, and $\chi^{mon}_{water}$ with $R^2 \approx 0.97$ on the test data set. To assess generative performance, we decoded 30000 randomly sampled latent vectors from the prior Gaussian distribution. Among the generated polymer CRUs, 90.29% were chemically valid (e.g., satisfying chemical valence rules), 90.26% were chemically valid and unique (not duplicated among generated CRUs), and 90.21% were chemically valid, unique, and novel compared to the trained OMG polymer CRUs. Note that not all generated polymer CRUs were chemically valid when the generated Group SELFIES representations failed to include exactly two repeating points required to define a linear homopolymer CRU (e.g., *CC*, where * denotes a repeating point), despite Group SELFIES enforcing 100% chemical validity for small organic molecules. Among the trained monomer-level properties, we focused on $\Phi_{mon}$, $E_{HOMO}$, $E_{LUMO}$, and $\chi^{mon}_{water}$ as these properties can be converted to polymer properties (i.e., $T_g$, $E_g$, $\chi_{water}$) through their strong correlations with polymer properties (Figures S2–S4) for multiobjective polymer design. The trained model and implementation are available on GitHub (https://github.com/TheJacksonLab/OMG-MultiProperty-GenML).

## RESULTS

Generative molecular design using the VAE architecture structures the latent space according to the targeted molecular properties to obtain molecular structure–property relationships. The organized latent space is then explored to identify latent vectors corresponding to desired functionalities for molecular generation. We first examine the distribution of the trained OMG CRUs within the linearly organized latent space and subsequently perform latent space searches to design polymers optimized for multiple properties ($T_g$, $E_g$, and $\chi_{water}$) simultaneously with BO. We further assess model transferability for multiproperty polymer refinement by applying the trained VAE to optimize PFAS-based polymers.

### Latent Space Organization

Because linear property prediction was incorporated during training, the latent space exhibits a linear organization with respect to monomer-level properties. Each linear regressor head corresponds to a particular monomer-level physical property and possesses linear coefficients that define a hyperplane in the latent space. In this way, the latent space of the trained OMG CRUs can be smoothly navigated to systematically modulate polymer physical properties by leveraging monotonic monomer-polymer property correlations. For example, an increasing direction for $\Phi_{mon}$ corresponds to a decreasing direction for experimental $T_g$ (Figure S2). Similarly, an increasing direction for $E_g^{HL}$ corresponds to an increasing direction for bulk polymer $E_g$ (Figure S3), and an increasing direction for $\chi^{mon}_{water}$ corresponds to an increasing direction for experimental $\chi_{water}$ (Figure S4). Figure 3a illustrates the latent space simultaneously and monotonically organized for $T_g$ (red), $E_g$ (green), and $\chi_{water}$ (blue). For visualization, the high-dimensional mean latent representations of the trained OMG CRUs were projected onto three principal axes corresponding to the largest eigenvalues of the mean latent representations obtained from principal component analysis (PCA). The linear latent space organization for monomer-level properties implies that, for a chosen starting polymer chemistry (Figure 3b), the trained OMG CRUs are monotonically arranged in polymer properties (e.g., $T_g$) along the corresponding direction (e.g., red arrow), as shown in Figure 3c. Note that the identified increasing directions for $T_g$, $E_g$, and $\chi_{water}$ can be applied to any starting polymer chemistry embedded in the latent space.

To probe the organization of the latent space, we explored trajectories corresponding to increasing $T_g$, $E_g$, and $\chi_{water}$. Figure 4a illustrates a trajectory for increasing $T_g$ (red arrow), originating from the initial chemistry in Figure 3b. For visualization, the high-dimensional mean latent representations of the trained OMG CRUs were projected onto principal axes that exhibit both high variance and strong linear alignment with the latent direction vector corresponding to increasing $T_g$. The resulting colormap suggests that trained OMG CRUs are structurally organized to yield higher $T_g$ values along this direction. The slight visual misalignment of the displayed CRUs relative to the projected vector arises from the reduction of the high-dimensional latent space, which is shaped by


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759–6771

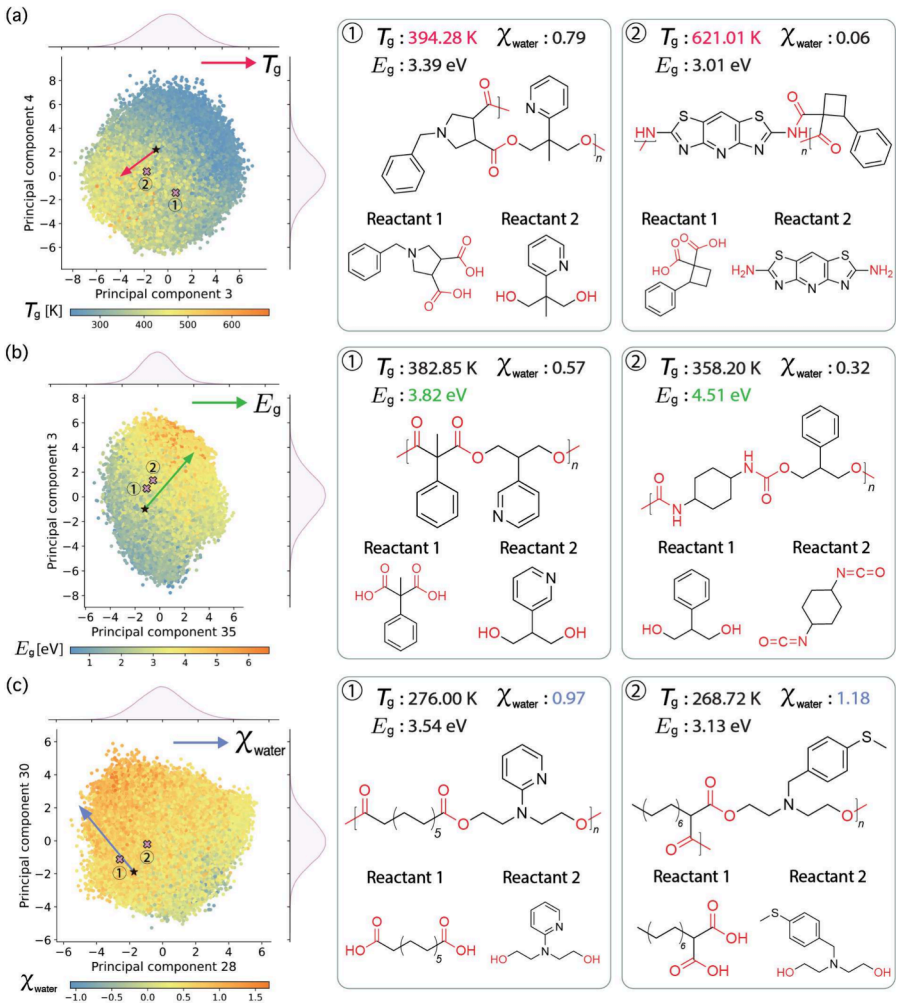



**Figure 4.** Latent space organization for (a) $T_g$, (b) $E_g$, and (c) $\chi_{water}$. The high-dimensional latent space was projected onto two principal axes that exhibit strong linear alignment with the increasing direction for each respective polymer property. The colormap denotes the predicted polymer property values. Two example trained OMG CRUs are presented along with their predicted properties, identified by searching in each property-increasing direction. Red highlights in OMG CRUs represent polymerization linkages, while red highlights in reactants indicate polymerization functional groups.

multiple VAE loss objectives, onto a 2D plane. Physically, the selected CRUs achieve higher $T_g$ values by incorporating rigid ring structures. As this sample trajectory focuses solely on $T_g$, monotonic trends in $E_g$ and $\chi_{water}$ are not expected.

Figure 4b illustrates the latent space organization along the increasing $E_g$ direction (green). The two trained OMG CRUs identified by searching along this direction exhibit larger $E_g$ values than the starting chemistry by replacing heteroatoms (nitrogen and sulfur) with carbon and shortening alkyl chains, and ultimately substituting pyridine with benzene[76] and introducing a nonconjugated cyclohexane. Figure 4c shows the organization corresponding to the increasing $\chi_{water}$ direction in latent space. More hydrophobic OMG CRUs were identified by searching along this direction by introducing longer alkyl chains to the CRUs and eventually replacing water-soluble pyridine with benzene. Overall, the latent space was simultaneously organized with respect to distinct polymer properties, yielding an interpretable latent space structure enforced by the linear prediction of monomer-level properties.

## Multiproperty Polymer Refinement

After structuring the latent space to enforce monotonic organization of polymer properties, we identified vectors in the latent space that simultaneously optimize multiple properties to enable facile polymer optimization. As an example, we demonstrate the simultaneous maximization of $T_g$, $E_g$, and $\chi_{water}$ for the reference chemistry shown in Figure 3b. Note that the generative design model can be applied to any desired increase or decrease in property values. Although the latent space exhibits linear monomer property landscapes that translate into monotonic polymer-property trends, inherent stochasticity in the latent space renders simple traversal along a linear monomer-property direction unreliable for designing new functional polymers during generation. To be specific, the encoder maps a molecule not to a single latent point but to a posterior Gaussian distribution, introducing prediction uncertainty in linear models. Consequently, a polymer generated by decoding a latent vector chosen along a linear monomer-property direction may possess substantially different properties than predicted by linear models if the selected vector corresponds to a low-probability molecular representation. It is worth noting that the linear traversal shown in Figure 4 was performed only on the trained OMG CRUs using their mean latent representations. Furthermore, the sparse nature of the high-dimensional latent space makes


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

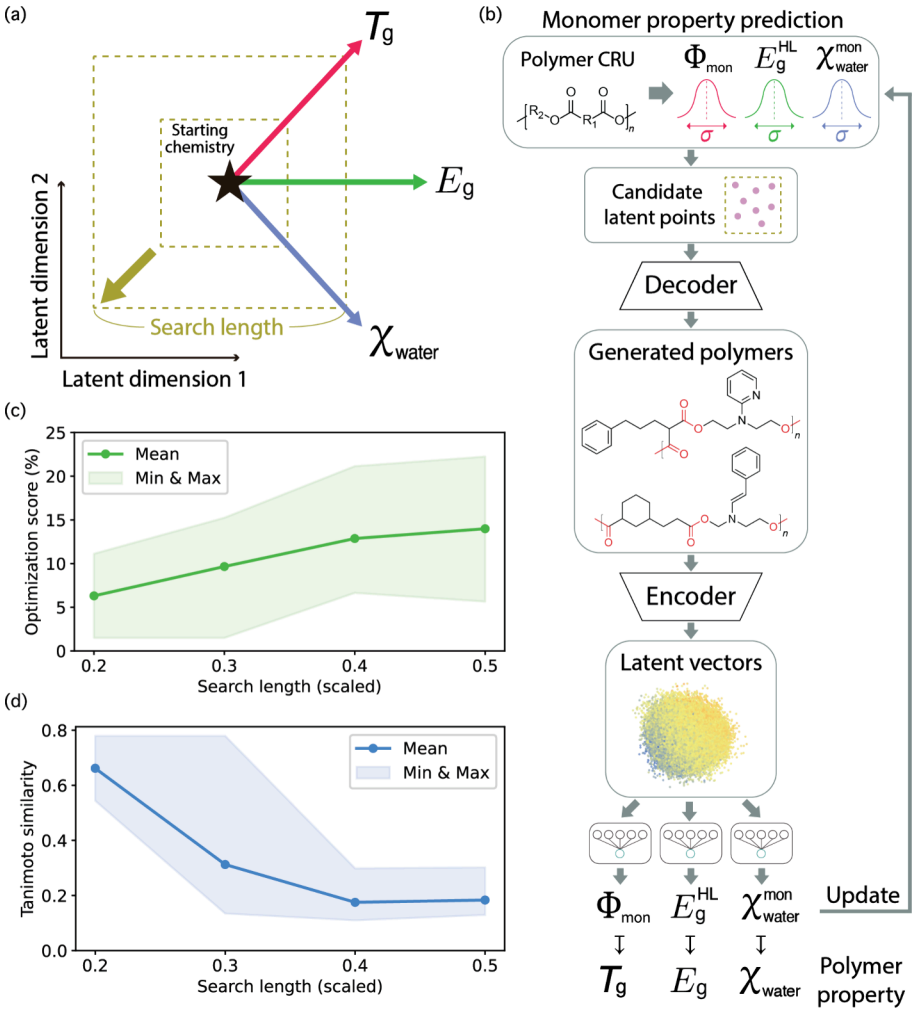



**Figure 5.** Latent space search via Bayesian optimization (BO) for multiproperty polymer refinement. (a) Monotonically organized latent space with respect to polymer properties such as $T_g$, $E_g$, and $\chi_{water}$. The latent space was explored by systematically varying the size of the latent search region to control the extent of property optimization with larger regions expected to yield greater improvements. (b) BO to guide property-driven navigation of the latent space for multiproperty polymer refinement. Gaussian process regressor (GPR) guides selection of latent candidates, for example, to simultaneously maximize $T_g$, $E_g$, and $\chi_{water}$, which are decoded into polymer candidates. (c) Optimization score for the simultaneous increase of $T_g$, $E_g$, and $\chi_{water}$ as the search region increased. (d) Tanimoto similarity between generated CRUs and the reference chemistry.

linear interpolation prone to falling into data-scarce regions[77] that can lead to poor-quality generation. To enable effective latent space search, we adopted BO (Figure 5b) to guide property-driven navigation in the stochastic, high-dimensional latent space. BO explicitly incorporates predictive uncertainty into the iterative optimization process, enabling robust property optimization.

We systematically varied the size of the latent search region as shown in Figure 5a to control the extent of BO-driven property optimization, with larger regions expected to yield greater improvements. A variety of optimization methods has been applied to targeted molecular design, including gradient-based optimization,[22,54] genetic algorithms,[20,35,78] particle-swarm optimization,[79,80] and BO.[23,35,81] Among these, BO has been frequently used to optimize latent vectors of VAEs,[82] providing sample-efficient global optimization with explicit uncertainty quantification.[83] To initiate BO, we first sampled 2000 OMG CRUs from the training set of 600000 OMG CRUs using *k*-means clustering on latent representations, selecting CRUs near cluster centers. These samples were used to train separate Gaussian process regressors (GPRs) to predict monomer properties—$\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$—along with associated prediction uncertainties from latent representations. It is worth noting that *k*-means clustering based on Euclidean distance, while highly scalable for large data sets,[84] is not ideal to ensure diversity in the high dimensional latent space (104 dimensions in this case) where Euclidean distance can fail to capture molecular similarity.[21,85,86] Nevertheless, PCA results indicate that the sampled 2000 OMG CRUs adequately represent the principal component distribution of the trained 600000 OMG CRUs (Figure S7). The trained GPR models then suggested candidate latent points by optimizing a multiobjective acquisition function to simultaneously increase $T_g$, $E_g$, and $\chi_{water}$. Candidate polymer CRUs were obtained by decoding the proposed latent vectors.

To verify the properties of decoded polymers, we re-encoded[87] the decoded CRUs to mean latent vectors and predicted their monomer properties (i.e., $\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$) using the trained linear regression heads. As shown in Figure S8, re-encoded properties provide more accurate estimates than properties predicted directly from candidate latent points using linear predictors. The improved accuracy of re-encoded


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

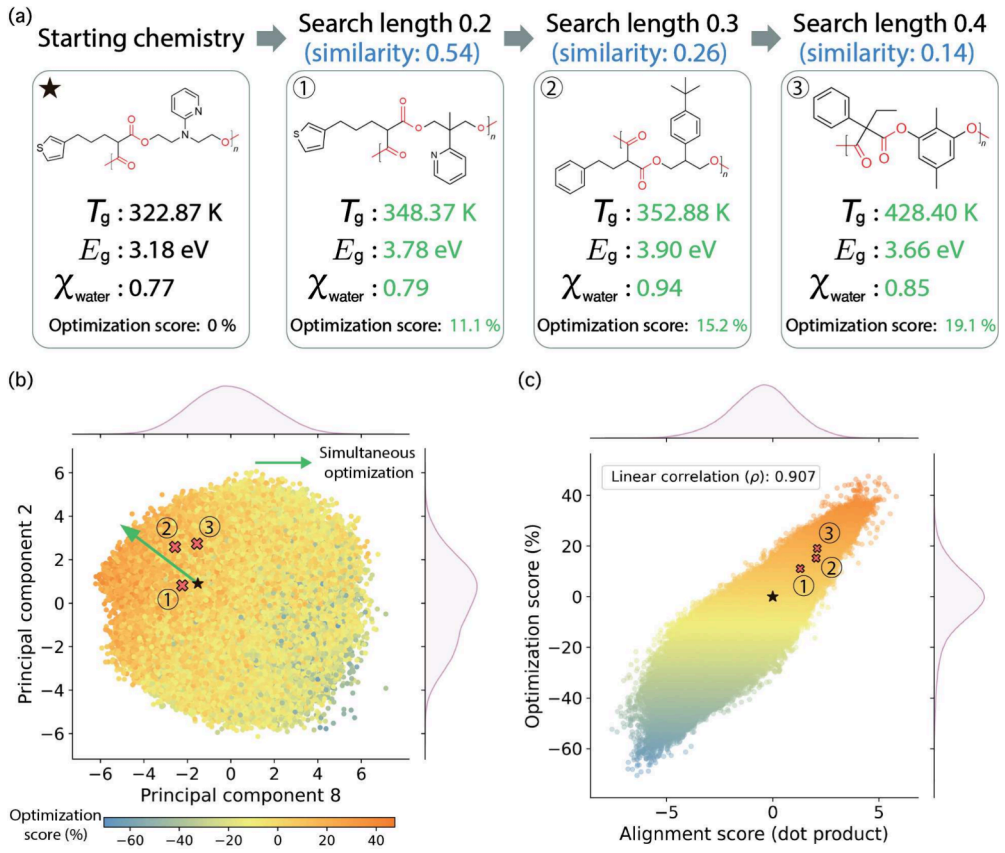



**Figure 6.** Multiproperty refinement and interpretability. (a) Examples of chemically valid, unique, and novel multiproperty optimized polymers discovered through BO. The generated polymers consistently retained an ester group that was treated as a single token in Group SELFIES to preserve the essential polymerization motif during generation. Green text indicates enhanced property values and optimization scores. (b) Increase in optimization score along the simultaneous optimization direction. This trend demonstrates that the linearly organized latent space provides a clear and interpretable direction for multiobjective improvement. (c) Quantitative relationship between the optimization score and the alignment score obtained using the simultaneous optimization direction.

properties likely arises because re-encoding recovers a more comprehensive stochastic representation (i.e., a posterior Gaussian distribution) rather than relying on a single latent point obtained through BO. The re-encoded property estimates were then used to update the GPR models, and the optimization loop was repeated iteratively by maximizing the expected hypervolume improvement for the Pareto frontier using gradient-based optimization over the continuous latent space. After optimization, the monomer properties of generated polymers (i.e., $\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$) were converted to polymer properties (i.e., $T_g$, $E_g$, and $\chi_{water}$). Additional details regarding the BO algorithm are provided in the Supporting Information.

Figure 5c presents the optimization score for the simultaneous increase of $T_g$, $E_g$, and $\chi_{water}$ as the size of the latent space search region was systematically increased. The optimization scores are shown for chemically valid, unique, and novel CRUs whose re-encoded properties satisfied the multiproperty optimization criteria (i.e., simultaneously increased $T_g$, $E_g$, and $\chi_{water}$ relative to the reference chemistry). The optimization score was defined as the average relative improvement (%) across $\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$. The search length denotes the scaled latent space distance between 0 and 1 for each latent space dimension, computed relative to the distribution of 2000 latent points sampled via *k*-means to initiate BO. As illustrated in Figure 5c, larger search lengths enable more effective simultaneous optimization of monomer-level properties, resulting in greater increases in $T_g$, $E_g$, and $\chi_{water}$. This trend is consistent with the Similar Property Principle[88] that structurally similar molecules tend to have similar properties. A narrow latent-space region is likely to contain CRUs that are structurally close to the reference chemistry, limiting potential improvements in achievable properties. In contrast, expanding the search region may access increasingly dissimilar CRUs, allowing for larger property enhancements. This interpretation is further supported by Figure 5d showing that the Tanimoto similarity between generated CRUs and the reference chemistry decreases with increasing search size. Importantly, the observed increase in optimization score with larger search regions is not unique to a linear property prediction model. Replacing the linear regression heads with nonlinear regression heads also exhibits similar behavior (Figure S9). However, the linear model offers improved interpretability of the multi-objective optimization landscape, as discussed in Figure 6.

Figure 6a shows examples of chemically valid, unique, and novel multiproperty optimized polymers discovered through BO. The generated polymers consistently retained an ester group that was treated as a single token in Group SELFIES to preserve the essential polymerization motif during generation. As mentioned previously, these examples illustrate the Similar Property Principle,[88] wherein achieving greater property improvements (i.e., higher optimization score) typically requires greater structural deviation from the reference


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

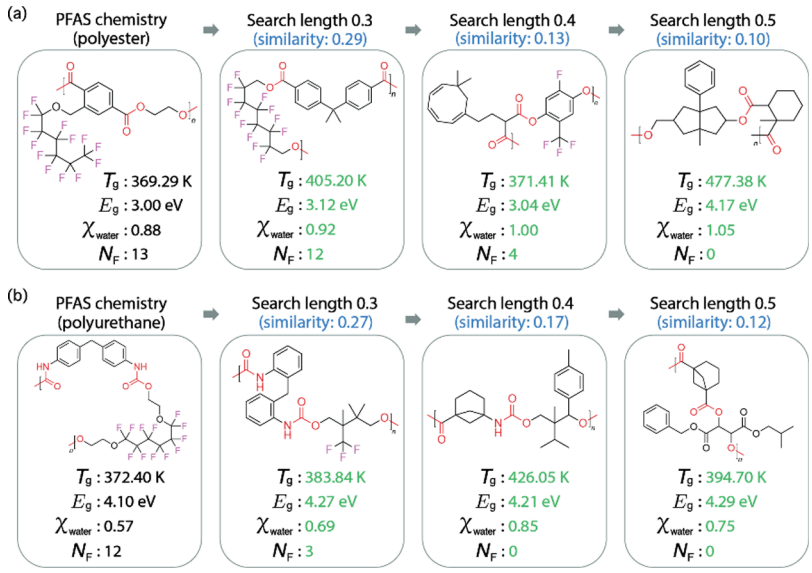



**Figure 7.** Multiproperty refinement of hypothetical, industrially relevant PFAS-based (a) polyester and (b) polyurethane to simultaneously increase $T_g$ (thermal stability), $E_g$ (electronic stability), and $\chi_{water}$ (enhanced hydrophobicity) while reducing $N_F$ (fluorine content). Green text indicates improved property values.

chemistry (i.e., lower Tanimoto similarity). In particular, the optimization score shows a negative linear correlation ($\rho \approx -0.51$) with Tanimoto similarity as shown in the Supporting Information (Figure S10). A complete list of generated polymers is provided in the Supporting Information (Figure S11). To better interpret multiproperty optimization within the linearly organized latent space, similar to the approach shown in Figure 4, we projected the high-dimensional mean latent representations of the trained OMG CRUs onto two principal axes exhibiting both high variance and strong linear alignment with the simultaneous optimization direction vector (Figure 6b). The multiproperty optimized polymers shown in Figure 6a were also projected onto the same principal axes for comparison. The simultaneous optimization direction vector was defined as the linear combination of the directions corresponding to increasing $T_g$ (i.e., decreasing $\Phi_{mon}$), increasing $E_g$ (i.e., increasing $E_g^{HL}$), and increasing $\chi_{water}$ (i.e., increasing $\chi_{water}^{mon}$).

The PCA projection in Figure 6b illustrates that the optimization score increases along this simultaneous optimization direction. This trend demonstrates that the linearly organized latent space provides an interpretable direction for multiobjective improvement. For comparison, when visualized using principal components with the largest variance, the latent space did not exhibit a clear trend of increasing optimization scores (Figure S12). Note that the identified polymer 2 appears to be more aligned with the simultaneous optimization direction than the identified polymer 3 (i.e., exhibiting a smaller principal component 8), despite having a lower optimization score. This apparent discrepancy may arise from the information loss during dimensionality reduction. To further investigate this effect, we performed a quantitative analysis of the relationship between the optimization score and the alignment score, as shown in Figure 6c.

Figure 6c further quantifies the relationship between the optimization score and the simultaneous optimization direction. The alignment score is defined as the dot product between the simultaneous optimization direction and the change in latent vectors relative to the reference chemistry. To provide context, the dot product between a single monomer property direction (e.g., $\Phi_{mon}$) and a latent vector difference relative to the reference chemistry corresponds to the change in the corresponding monomer property—scaled to $N(0, 1)$ using the mean and the standard deviation of the training data distribution—relative to the reference chemistry. Thus, the alignment score aggregates the total normalized change in $\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$ with respect to the reference chemistry. Figure 6c shows a strong linear correlation ($\rho \approx 0.907$) between the optimization score and the alignment score. This strong correlation indicates that a linear monomer-level property prediction model enables the identification of a simultaneous optimization direction as a linear vector in the latent space. The identified linear vector provides a clear and interpretable axis along which the multiobjective optimization score increases within the high-dimensional latent space. Notably, when the optimization score is defined using a standard-deviation–based metric, its correlation with the alignment score becomes even stronger (Figure S13). This reflects that the optimization score (%) was defined in terms of relative percent differences rather than difference normalized by standard deviations.

## Latent Transfer for PFAS Optimization

The trained VAE can be leveraged to design polymers with target properties that were not part of the original training set of monomer-level properties. To demonstrate this transferability for multiobjective polymer design, we applied the trained VAE to PFAS-based polymers. PFAS have been widely used in consumer products and industrial applications with their exceptional chemical and thermal stability.[89] However, PFAS have raised increasing concerns regarding their potential adverse environmental and health impacts throughout their life cycle.[90,91] In this context, we explored whether the trained VAE could optimize hypothetical PFAS-based polymers to achieve enhanced stability while reducing fluorine content. Specifically, we aimed to increase $T_g$ (thermal stability), $E_g$ (electronic stability), and $\chi_{water}$ (enhanced hydrophobicity) while simultaneously decreasing the number of fluorine atoms


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

in generated CRUs ($N_F$). To incorporate fluorine content as an additional design objective, the BO workflow (Figure 5b) was extended to include $N_F$. GPR models were trained on the latent representations of CRUs to predict $N_F$ as well as $\Phi_{mon}$, $E_g^{HL}$, and $\chi_{water}^{mon}$. The GPR achieved high accuracy in predicting $N_F$ ($R^2 \approx 0.943$) for the test OMG CRUs (Figure S14), demonstrating that the learned latent embeddings transfer effectively to a previously unseen monomer-level property.

Figure 7a,b illustrates examples of chemically valid, unique, and novel multiproperty optimized polymers identified for the simultaneous maximization of $T_g$, $E_g$, and $\chi_{water}$ while minimizing $N_F$. Figure 7a presents multiproperty optimized polymers derived from a hypothetical side-chain fluorinated polyester built on a polyethylene terephthalate (PET) backbone. At the search length 0.3, the generated CRU retained substantial fluorine content, reflecting the tendency of latent regions near the reference chemistry to produce fluorinated structures. At larger search lengths (0.4 and 0.5), however, BO identified CRUs with significantly reduced fluorine content while still achieving the targeted multiproperty optimization. It is important to note that all discovered polymers preserved their ester-based polymerization motif that was explicitly represented as a single token using the Group SELFIES representation.

Figure 7b shows analogous results for a hypothetical PFAS-based polyurethane derived from an industrially relevant monomer,[92,93] 4,4′-methylene diphenyl diisocyanate (MDI), combined with a fluorinated diol. Similar to the polyester case shown in Figure 7a, the number of fluorine atoms in the generated CRUs remained relatively high at a search length of 0.3 but further decreased at longer search lengths, yielding fluorine-free polymers. However, ester groups frequently appeared in place of urethane linkages during generative design, leading the optimization to produce polyesters rather than polyurethanes. This behavior reflects the underlying training-data distribution, which contains a large majority of polyester chemistries. Nevertheless, all generated polymers maintained valid polymerization motifs, either ester or urethane linkages, throughout the optimization process. A complete list of discovered CRUs is provided in the Supporting Information (Figures S15−S16).

## DISCUSSION

In this work, we introduce a framework for multiproperty polymer refinement that leverages monomer-level properties to efficiently explore a large chemical space. Our generative model and BO operate entirely in the space of monomer-level properties, and the monotonic correlations between monomer and polymer properties are then used to translate these monomer properties into polymer properties. However, given the limited availability of polymer property data, the specific functional relationships between monomer and polymer properties may not strictly hold across all step-growth chemistries investigated in this work. Importantly, the present framework remains effective as long as qualitative monotonic trends between monomer and polymer properties hold (e.g., decrease in $\Phi_{mon}$ leading to increase in $T_g$), because the framework is designed to optimize polymer properties relative to a reference chemistry by systematically increasing or decreasing monomer-level properties. As more accurate monomer-polymer property correlations become available, this monomer-based generative approach can yield increasingly reliable multiproperty polymer design by improving the fidelity of translation from monomer to polymer properties. More generally, the framework is extensible to cases where polymer properties arise from nonlinear relationships involving multiple monomer properties. In such cases, BO can be performed to identify monomer property targets that optimize the desired polymer properties. We acknowledge that establishing robust monomer−polymer property correlations is challenging due to additional complexities of polymers involving morphology, molecular weight, and polydispersity. Prior studies have shown that polymer properties can be predicted from the structural and physicochemical characteristics of their monomeric (or oligomeric) precursors, such as glass transition temperatures,[67,94−96] electronic properties of conjugated polymers,[97−101] and Flory−Huggins interaction parameters.[67,102] These findings highlight the potential of the present VAE framework for transfer learning applications, particularly through the use of latent embeddings that encode monomer-level structural and property information. However, it remains difficult to determine *a priori* whether a given polymer property can be reliably approximated by monomer-level descriptors due to the complex interplay of intermolecular interactions, molecular packing, and morphology. Nevertheless, when correlations exist, monomer-level properties can enable broad chemical space exploration for early stage polymer design. For example, although not utilized in the present study, normalized isotropic polarizability ($\alpha_{norm}$)—one of the trained 11 monomer-level properties (Table S1)—has been shown to correlate positively with polymer dielectric constant and refractive index from DFT calculations.[67] We anticipate this monomer-based generative framework to be applied to a broader array of polymer functionalities as additional monomer−polymer property correlations continue to emerge.

The linear monomer-level property prediction model enhances interpretability in multiproperty polymer design by providing clear increasing or decreasing directions for each property within the latent space. This interpretability is effective for both the trained and generated CRUs once encoded into latent representations. However, when generating new multiproperty optimized CRUs, simply moving along a linear property direction and decoding the resulting latent points does not guarantee optimal polymer candidates. This limitation arises from intrinsic stochasticity, where each CRU is represented as a distribution in latent space rather than a single point, introducing property prediction uncertainty (Figure S8). If interpretability is not a primary requirement, nonlinear property prediction models can be adopted as they may achieve higher optimization scores than linear property prediction models at large search lengths (Figure S9). Nonetheless, the linear framework offers an interpretable latent space where differences in monomer-level properties correspond directly to linear projections along the property-guided vectors. This interpretable latent space provides insight into why a given generated polymer exhibits a particular optimization score by constructing a linear vector that directs multiproperty polymer optimization.

The experimental synthetic realizability of the generated CRUs remains an open challenge. Although Group SELFIES enables representation of polymerization motifs as single tokens, this does not guarantee that the generated CRUs are synthetically feasible in a lab synthesis, especially without explicitly considering the reactivity of monomer reactants. Incorporating detailed reactivity information into computa-


https://doi.org/10.1021/acs.macromol.6c00564
Macromolecules 2026, 59, 6759−6771

tional modeling, however, would require computationally expensive transition state calculations. A pragmatic, though imperfect, strategy would be to incorporate handcrafted synthetic accessibility heuristics derived from experimentally preferred reference chemistries. For example, one could restrict the generative process to modifying only side chains while preserving the polymer backbone, thereby maintaining compatibility with user-preferred polymerization chemistry and ensuring that essential reactive precursors remain intact. We anticipate that the growing availability of experimental polymerization data sets will enable more accurate modeling of reactivity constraints, ultimately allowing generative polymer design frameworks to account more reliably for synthetic feasibility.

## CONCLUSION

This work establishes a generative, multiobjective design framework that leverages computationally accessible monomer-level descriptors to navigate the vast, sparse chemical space of functional polymers. By integrating a variational autoencoder with linear property prediction heads, we engineered a linearly organized latent space that offers interpretable trajectories for the simultaneous property optimization. Coupling this structured latent representation with Bayesian optimization enabled the efficient discovery of polymer chemistries with simultaneously enhanced glass transition temperatures, band gaps, and hydrophobicity. The transferability of this approach was further demonstrated through the targeted design of PFAS-based alternatives, successfully reducing the fluorine content while maintaining critical stability metrics. While currently reliant on monomer-polymer correlations, this monomer-centric strategy provides a scalable solution to the persistent challenge of data scarcity in polymer informatics. Ultimately, this framework accelerates early stage materials discovery, offering a robust pathway to identify high-performance candidates prior to computationally expensive simulations or experimental synthesis.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.macromol.6c00564.

Data compositions; Correlations between monomer-level and polymer properties; Monomer-level properties; Group SELFIES; Hyperparameter optimization; Predicted monomer-level properties by the generative model; Principal component analysis of sampled OMG CRUs via *k*-means; Re-encoding strategy; Bayesian optimization; Nonlinear predictor model; Relationship between optimization score and Tanimoto similarity; Multiproperty optimized polymers; Optimization score and PCA; Quantitative relationships between the standard-deviation-based optimization score and the alignment score; Latent transfer for predicting the number of fluorine atoms; PFAS optimization: polyester; PFAS optimization: polyurethane (PDF)

## AUTHOR INFORMATION


### Corresponding Author

**Nicholas E. Jackson** − *Department of Chemistry and Beckman Institute for Advanced Science and Technology, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1470-1903; Email: jacksonn@illinois.edu

### Authors

**Seonghwan Kim** − *Department of Materials Science and Engineering, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States*

**Charles M. Schroeder** − *Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States; Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0001-6023-2274

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.macromol.6c00564


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

N.E.J. thanks the 3M Nontenured Faculty Award for support of this research.

## REFERENCES


(1) Kern, J.; Su, Y.-L.; Gutekunst, W.; Ramprasad, R. An informatics framework for the design of sustainable, chemically recyclable, synthetically accessible, and durable polymers. *npj Comput. Mater.* **2025**, *11*, 182.

(2) Ethier, J.; Antoniuk, E. R.; Brettmann, B. Predicting polymer solubility from phase diagrams to compatibility: a perspective on challenges and opportunities. *Soft Matter* **2024**, *20*, 5652−5669.

(3) Ethier, J. G.; Audus, D. J.; Ryan, D. C.; Vaia, R. A. Integrating theory with machine learning for predicting polymer solution phase behavior. *Giant* **2023**, *15*, No. 100171.

(4) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785−796.

(5) Mortimer, R. J.; Dyer, A. L.; Reynolds, J. R. Electrochromic organic and polymeric materials for display applications. *Displays* **2006**, *27*, 2−18.

(6) Hong, M.; Chen, E. Y.-X. Chemically recyclable polymers: a circular economy approach to sustainability. *Green Chem.* **2017**, *19*, 3692−3706.

(7) Fagnani, D. E.; Tami, J. L.; Copley, G.; Clemons, M. N.; Getzler, Y. D. Y. L.; McNeil, A. J. 100th Anniversary of Macromolecular Science Viewpoint: Redefining Sustainable Polymers. *ACS Macro Lett.* **2021**, *10*, 41−53.

(8) Miller, S. A. Sustainable Polymers: Opportunities for the Next Decade. *ACS Macro Lett.* **2013**, *2*, 550−554.

(9) Chamas, A.; Moon, H.; Zheng, J.; Qiu, Y.; Tabassum, T.; Jang, J. H.; Abu-Omar, M.; Scott, S. L.; Suh, S. Degradation rates of plastics in the environment. ACS Sustainable. *Chem. Eng.* **2020**, *8*, 3494−3511.

(10) Strapasson, R.; Amico, S.; Pereira, M.; Sydenstricker, T. Tensile and impact behavior of polypropylene/low density polyethylene blends. *Polym. Test.* **2005**, *24*, 468−473.

(11) Friday, D. M.; Jackson, N. E. Electron and ion transport in semi-dilute conjugated polyelectrolytes: view from a coarse-grained tight binding model. *Mol. Syst. Des. Eng.* **2023**, *8*, 743−755.

(12) Oliver, S.; Zhao, L.; Gormley, A. J.; Chapman, R.; Boyer, C. Living in the Fast Lane−High Throughput Controlled/Living Radical Polymerization. *Macromolecules* **2019**, *52*, 3−23.

(13) Hoogenboom, R.; Meier, M. A. R.; Schubert, U. S. Combinatorial Methods, Automated Synthesis and High-Throughput Screening in Polymer Research: Past and Present. *Macromol. Rapid Commun.* **2003**, *24*, 15−32.

(14) Reis, M. H.; Leibfarth, F. A.; Pitet, L. M. Polymerizations in Continuous Flow: Recent Advances in the Synthesis of Diverse Polymeric Materials. *ACS Macro Lett.* **2020**, *9*, 123−133.
(15) Dhamankar, S.; Webb, M. A. Chemically specific coarse-graining of polymers: Methods and prospects. *J. Polym. Sci.* **2021**, *59*, 2613−2643.
(16) Gartner, T. E. I.; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* **2019**, *52*, 755−786.
(17) Jackson, N. E. Coarse-Graining Organic Semiconductors: The Path to Multiscale Design. *J. Phys. Chem. B* **2021**, *125*, 485−496.
(18) Wang, C.-I.; Jackson, N. E. Bringing Quantum Mechanics to Coarse-Grained Soft Materials Modeling. *Chem. Mater.* **2023**, *35*, 1470−1486.
(19) Kumar, A.; Weig, A. R.; Agarwal, S. Balancing Degradability and Physical Properties of Amorphous Poly (D, L-lactide) by Making Blends. *Macromol. Mater. Eng.* **2022**, *307*, No. 2100602.
(20) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep* **2016**, *6*, No. 20952.
(21) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(22) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M.; Hernández-Lobato, J. M. A model to search for synthesizable molecules. *arXiv:1906.05221 [cs.LG]* **2019**, .na
(23) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, PMLR80, 2018.
(24) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **2017**, *8*, No. 10883.
(25) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharmaceutics* **2017**, *14*, 3098−3104.
(26) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv:1705.10843 [stat.ML]* **2018**, na.
(27) Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. MolecularRNN: Generating realistic molecular graphs with optimized properties. *arXiv:1905.13372 [cs.LG]* **2019**, na.
(28) Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; Tang, J. GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. *arXiv:2001.09382 [cs.LG]* **2020**, na.
(29) Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2022**, *62*, 2064−2076.
(30) Born, J.; Manica, M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mach Intell* **2023**, *5*, 432−444.
(31) Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; Frossard, P. DiGress: Discrete Denoising diffusion for graph generation. *arXiv:2209.14734 [cs.LG]* **2023**, na.
(32) Hua, C.; Luan, S.; Xu, M.; Ying, Z.; Fu, J.; Ermon, S.; Precup, D. MUDiff: Unified Diffusion for Complete Molecule Generation. *arXiv:2304.14621 [cs.LG]* **2024**, na.
(33) Gurnani, R.; Kamal, D.; Tran, H.; Sahu, H.; Scharm, K.; Ashraf, U.; Ramprasad, R. polyG2G: A Novel Machine Learning Algorithm Applied to the Generative Design of Polymer Dielectrics. *Chem. Mater.* **2021**, *33*, 7008−7016.
(34) Jiang, S.; Dieng, A. B.; Webb, M. A. Property-guided generation of complex polymer topologies using variational autoencoders. *npj Comput. Mater.* **2024**, *10*, 139.
(35) Vogel, G.; Weber, J. M. Inverse design of copolymers including stoichiometry and chain architecture. *Chem. Sci.* **2025**, *16*, 1161−1178.
(36) Qiu, H.; Sun, Z.-Y. On-demand reverse design of polymers with PolyTAO. *npj Comput. Mater.* **2024**, *10*, 273.
(37) Savit, A.; Sahu, H.; Shukla, S. S.; Xiong, W.; Ramprasad, R. polyBART: A Chemical Linguist for Polymer Property Prediction and Generative Design. *arXiv:2506.04233 [cond-mat.soft]* **2025**, na.
(38) Yang, Z.; Ye, W.; Lei, X.; Schweigert, D.; Kwon, H.-K.; Khajeh, A. De novo design of polymer electrolytes using GPT-based and diffusion-based generative models. *npj Comput. Mater.* **2024**, *10*, 296.
(39) Audus, D. J.; de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **2017**, *6*, 1078−1082.
(40) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. *2011 International Conference on Emerging Intelligent Data and Web Technologies*; IEEE, 2011; pp 22−29.
(41) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, No. 160012.
(42) 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* **2018**, *122*, 17575−17585.
(43) Ma, R.; Luo, T. PI1M: a benchmark database for polymer informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684−4690.
(44) Xu, J.; Liu, G.; Guo, R.; Jiang, M.; Luo, T. POINT$^2$: A Polymer Informatics Training and Testing Database. *arXiv:2503.23491 [cond-mat.mtrl-sci]* **2025**, na.
(45) Polymer Property Predictor and Database. https://pppdb.uchicago.edu/ (accessed: 2026−04−10).
(46) MALDI Recipes. https://maldi.nist.gov/ (accessed: 2026−01−30).
(47) 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.; Olsen, B. D. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. *ACS Cent. Sci.* **2023**, *9*, 330−338.
(48) 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. *J. Chem. Inf. Model.* **2021**, *61*, 1150−1163.
(49) Hayashi, Y.; Shiomi, J.; Morikawa, J.; Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput. Mater.* **2022**, *8*, 1−15.
(50) Brinson, L. C.; Deagen, M.; Chen, W.; McCusker, J.; McGuinness, D. L.; Schadler, L. S.; Palmeri, M.; Ghumman, U.; Lin, A.; Hu, B. Polymer Nanocomposite Data: Curation, Frameworks, Access, and Potential for Discovery and Design. *ACS Macro Lett.* **2020**, *9*, 1086−1094.
(51) 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.; Jensen, K. F.; Olsen, B. D. BigSMILES: AStructurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523−1531.
(52) Schneider, L.; Walsh, D.; Olsen, B.; de Pablo, J. Generative BigSMILES: an extension for polymer informatics, computer simulations & ML/AI. *Digital Discovery* **2024**, *3*, 51−61.
(53) Wilson, A. N.; St John, P. C.; Marin, D. H.; Hoyt, C. B.; Rognerud, E. G.; Nimlos, M. R.; Cywar, R. M.; Rorrer, N. A.; Shebek, K. M.; Broadbelt, L. J.; et al. PolyID: Artificial intelligence for discovering performance-advantaged and sustainable polymers. *Macromolecules* **2023**, *56*, 8547−8557.
(54) Kim, S.; Schroeder, C. M.; Jackson, N. E. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. *ACS Polym. Au* **2023**, *3*, 318−330.
(55) Ohno, M.; Hayashi, Y.; Zhang, Q.; Kaneko, Y.; Yoshida, R. SMiPoly: Generation of a Synthesizable Polymer Virtual Library


https://doi.org/10.1021/acs.macromol.6c00564
*Macromolecules* 2026, 59, 6759−6771

Using Rule-Based Polymerization Reactions. *J. Chem. Inf. Model.* **2023**, *63*, 5539−5548.
(56) Yue, T.; He, J.; Li, Y. Polyuniverse: generation of a large-scale polymer library using rule-based polymerization reactions for polymer informatics. *Digital Discovery* **2024**, *3*, 2465−2478.
(57) Wang, J.-F.; Sun, Y.-B.; Chen, Q.-T.; Ji, F.-F.; Song, Y.-Y.; Ruan, M.-Y.; Wang, Y. OpenPoly: a polymer database empowering benchmarking and multi-property predictions. *Chin. J. Polym. Sci.* **2025**, *43*, 1749−1760.
(58) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14*, 4099.
(59) Shetty, P.; Rajan, A. C.; Kuenneth, C.; Gupta, S.; Panchumarti, L. P.; Holm, L.; Zhang, C.; Ramprasad, R. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Comput. Mater.* **2023**, *9*, 1−12.
(60) Xie, T.; Kwon, H.-K.; Schweigert, D.; Gong, S.; France-Lanord, A.; Khajeh, A.; Crabb, E.; Puzon, M.; Fajardo, C.; Powelson, W.; Shao-Horn, Y.; Grossman, J. C. A cloud platform for sharing and automated analysis of raw data from high throughput polymer MD simulations. *APL Mach. Learn.* **2023**, *1*, No. 046108.
(61) Tiwari, S. P.; Shi, W.; Budhathoki, S.; Baker, J.; Sekizkardes, A. K.; Zhu, L.; Kusuma, V. A.; Hopkinson, D. P.; Steckel, J. A. Creation of polymer datasets with targeted backbones for screening of high-performance membranes for gas separation. *J. Chem. Inf. Model.* **2024**, *64*, 638−652.
(62) Ferrari, B. S.; Manica, M.; Giro, R.; Laino, T.; Steiner, M. B. Predicting polymerization reactions via transfer learning using chemical language models. *npj Comput. Mater.* **2024**, *10*, 119.
(63) Reese, R.; Ganose, A.; Romain, C. Thermodynamics of Ring-Opening Polymerisation Informatics Collection (TROPIC): a database to enable polymer chemical recycling. *Faraday Discuss.* **2026**, *262*, 455.
(64) Chen, J.; Jackson, N. E. A Machine Learning Model for Copolymer Radical Reactivity Ratio Predictions with Frontier-Orbital Insights. *Chem. Mater.* **2025**, *37*, 7894−7907.
(65) Yoshida, R. Omics-scale polymer computational database transferable to real-world artificial intelligence applications. *arXiv:2511.11626 [physics.chem-ph]* **2025**, na.
(66) Levine, D. S.; Liesen, N.; Chua, L.; Diffenderfer, J.; Ingolfsson, H.; Kroonblawd, M. P.; Kumar, N.; Maiti, A.; Mohottalalage, S. S.; Shuaibi, M.; Essen, B. V.; Wood, B. M.; Zitnick, C. L.; Blau, S. M.; Antoniuk, E. R. The Open Polymers 2026 (OPoly26) Dataset and Evaluations. *arXiv:2512.23117 [physics.chem-ph]* **2026**, na.
(67) Kim, S.; Schroeder, C. M.; Jackson, N. E. Functional monomer design for synthetically accessible polymers. *Chem. Sci.* **2025**, *16*, 4755−4767.
(68) Hiemenz, P. C.; Lodge, T. P. *Polymer Chemistry*; CRC Press, 2007.
(69) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(70) Kahovec, J.; Fox, R. B.; Hatada, K. IUPAC Recommendations 2002. *Pure Appl. Chem.* **2002**, *74*, 1921.
(71) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv:1701.02810 [cs.CL]* **2017**, na.
(72) Cheng, A. H.; Cai, A.; Miret, S.; Malkomes, G.; Phielipp, M.; Aspuru-Guzik, A. Group SELFIES: a robust fragment-based molecular string representation. *Digital Discovery* **2023**, *2*, 748−758.
(73) Iovanac, N. C.; Savoie, B. M. Simpler is better: how linear prediction tasks improve transfer learning in chemical autoencoders. *J. Phys. Chem. A* **2020**, *124*, 3679−3685.
(74) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv:1907.10902 [cs.LG]* **2019**, na.
(75) Dask. https://www.dask.org/ (accessed: 2026−01−30).
(76) Krotkus, S.; Matulaitis, T.; Diesing, S.; Copley, G.; Archer, E.; Keum, C.; Cordes, D. B.; Slawin, A. M.; Gather, M. C.; Zysman-Colman, E.; et al. Fast delayed emission in new pyridazine-based compounds. *Front. Chem.* **2021**, *8*, No. 572862.
(77) White, T. Sampling Generative Networks. *arXiv:1609.04468 [cs.NE]* **2016**, na.
(78) Kim, C.; Batra, R.; Chen, L.; Tran, H.; Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Comput. Mater. Sci.* **2021**, *186*, No. 110067.
(79) Hartenfeller, M.; Proschak, E.; Schüller, A.; Schneider, G. Concept of Combinatorial De Novo Design of Drug-like Molecules by Particle Swarm Optimization. *Chemical Biology & Drug Design* **2008**, *72*, 16−26.
(80) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016−8024.
(81) Jiang, S.; Webb, M. A. Generative active learning across polymer architectures and solvophobicities for targeted rheological behavior. *npj Comput. Mater.* **2026**, *12*, 28.
(82) Gao, W.; Fu, T.; Sun, J.; Coley, C. W. Sample efficiency matters: a benchmark for practical molecular optimization. *Proceedings of the 36th International Conference on Neural Information Processing Systems*; Red Hook, NY, USA, 2022; Article No. 1551, pp 21342−21357.
(83) Balandat, M.; Karrer, B.; Jiang, D. R.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Advances in Neural Information Processing Systems* **2020**, *33*, na.
(84) Wani, A. A. Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ. Computer Science* **2024**, *10*, No. e2286.
(85) Domingos, P. A few useful things to know about machine learning. *Communications of the ACM* **2012**, *55*, 78−87.
(86) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. *ICDT.* **2001**, *1973*, 420−434.
(87) Batra, R.; Dai, H.; Huan, T. D.; Chen, L.; Kim, C.; Gutekunst, W. R.; Song, L.; Ramprasad, R. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chem. Mater.* **2020**, *32*, 10489−10500.
(88) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley, 1990.
(89) Glüge, J.; Scheringer, M.; Cousins, I. T.; DeWitt, J. C.; Goldenman, G.; Herzke, D.; Lohmann, R.; Ng, C. A.; Trier, X.; Wang, Z. An overview of the uses of per-and polyfluoroalkyl substances (PFAS). *Environ. Sci.: Processes Impacts* **2020**, *22*, 2345−2373.
(90) Fenton, S. E.; Ducatman, A.; Boobis, A.; DeWitt, J. C.; Lau, C.; Ng, C.; Smith, J. S.; Roberts, S. M. Per-and polyfluoroalkyl substance toxicity and human health review: Current state of knowledge and strategies for informing future research. *Environmental toxicology and chemistry* **2020**, *40*, 606−630.
(91) Lohmann, R.; Letcher, R. J. The universe of fluorinated polymers and polymeric substances and potential environmental impacts and concerns. *Current opinion in green and sustainable chemistry* **2023**, *41*, No. 100795.
(92) Delebecq, E.; Pascault, J.-P.; Boutevin, B.; Ganachaud, F. On the versatility of urethane/urea bonds: reversibility, blocked isocyanate, and non-isocyanate polyurethane. *Chem. Rev.* **2013**, *113*, 80−118.
(93) Duan, C.; You, J.; Liu, B.; Ma, J.; Zhou, H.; Zhang, H.; Zhang, J. Ionic liquid-mediated solvothermal synthesis of 4, 4'-methylenediphenyl diisocyanate (MDI): an efficient and environment-friendly process. *New J. Chem.* **2018**, *42*, 12243−12255.
(94) Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*; Elsevier, 2009.
(95) Bicerano, J. *Prediction of polymer properties*; CRC Press, 2002.
(96) Xie, R.; Weisen, A. R.; Lee, Y.; Aplan, M. A.; Fenton, A. M.; Masucci, A. E.; Kempe, F.; Sommer, M.; Pester, C. W.; Colby, R. H.;

Gomez, E. D. Glass transition temperature from the chemical structure of conjugated polymers. *Nat. Commun.* **2020**, *11*, 893.
(97) Larsen, R. E. Simple extrapolation method to predict the electronic structure of conjugated polymers from calculations on oligomers. *J. Phys. Chem. C* **2016**, *120*, 9650−9660.
(98) Hutchison, G.; Zhao, Y.-J.; Delley, B.; Freeman, A.; Ratner, M.; Marks, T. Electronic structure of conducting polymers: Limitations of oligomer extrapolation approximations and effects of heteroatoms. *Phys. Rev. B* **2003**, *68*, No. 035204.
(99) Gierschner, J.; Cornil, J.; Egelhaaf, H.-J. Optical Bandgaps of -Conjugated Organic Materials at the Polymer Limit: Experiment and Theory. *Adv. Mater.* **2007**, *19*, 173−191.
(100) Yang, S.; Olishevski, P.; Kertesz, M. Bandgap calculations for conjugated polymers. *Synth. Met.* **2004**, *141*, 171−177.
(101) Zade, S. S.; Bendikov, M. From Oligomers to Polymer: Convergence in the HOMO-LUMO Gaps of Conjugated Oligomers. *Org. Lett.* **2006**, *8*, 5243−5246.
(102) Aoki, Y.; Wu, S.; Tsurimoto, T.; Hayashi, Y.; Minami, S.; Tadamichi, O.; Shiratori, K.; Yoshida, R. Multitask Machine Learning to Predict Polymer−Solvent Miscibility Using Flory—Huggins Interaction Parameters. *Macromolecules* **2023**, *56*, 5446−5456.





https://doi.org/10.1021/acs.macromol.6c00564
*Macromolecules* 2026, 59, 6759−6771