Supporting Information for functional monomer design for synthetically accessible polymers

Seonghwan Kim,[†] Charles M. Schroeder,[†,‡,¶,§] and Nicholas E. Jackson[*,‡,¶]

[†]*Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

[‡]*Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

[¶]*Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

[§]*Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

E-mail: jacksonn@illinois.edu

1. Generation of Atomic Coordinates for OMG CRUs

The atomic coordinates of the OMG constitutional repeating units (CRUs)[1] were generated for monomer-level property calculations with their 3D molecular geometries. The OMG CRUs[2] were represented by SMILES strings[3] with asterisk symbols denoting repeating units (e.g., *CC* for polyethylene). The OMG CRUs were terminated with methyl groups to replace asterisks for atomistic calculations. While the atomistic calculations of several monomer units (e.g., trimer) can be considered for enhanced polymer property estimation,[4,5] we focused on single OMG monomer unit calculations to mitigate a computational cost. First, a set of up to 30 OMG CRU diverse conformers was generated using a genetic algorithm implemented in OpenBabel[6] maximizing a diversity score estimated through root-mean-square deviation (RMSD) of atomic positions between conformers. The geometries of these OMG CRU conformers were then optimized using Universal Force Field (UFF).[7] Out of the 30 conformers, up to 15 conformers with low UFF energy were selected and further optimized with a semi-empirical quantum chemical method, GFN2-xTB (XTB2).[8] In geometry optimization with XTB2, the implicit solvation (domain decomposition conductor-like screening model[9]) was employed to solvate OMG CRU conformers with toluene of a dielectric constant ($\epsilon = 2.4$) similar to conventional polymers at room temperature.[10] After the geometry optimizations of 15 conformers with XTB2, up to 5 distinct conformers with low XTB2 energy were chosen for monomer-level property calculations. Especially, the RMSD and energy criteria from Grimme[11] was adopted to select up to 5 distinct, low XTB2 energy conformers with adjacency matrices consistent with original OMG CRUs. These 5 distinct, low XTB2 energy conformers were anticipated to have low DFT single-point energies, thereby contributing a large Boltzmann weight for subsequent analysis of OMG CRU properties at room temperature. With DFT single-point energies, we obtained Boltzmann averaged values for monomer-level properties (mean values for Flory-Huggins χ parameters).

2. Mathematical Definitions for Geometry Descriptors

The following geometry descriptors were computed using RDKit[12] applied to the optimized XTB2 geometries of OMG CRUs.

(1) Ashpericity ($\Omega_\mathbf{A}$)

Ω_A is defined[13] from a gyration tensor S_{mn}.

$$S_{mn} = \frac{1}{M}\sum_{i=1}^{A}(r_{i,m} - r_{CM,m})(r_{i,n} - r_{CM,n})\, m_i$$
$$\Omega_A = \frac{1}{2}\,\frac{(t_3 - t_1)^2 + (t_1 - t_2)^2 + (t_2 - t_3)^2}{(t_1 + t_2 + t_3)^2}$$

where M is the total mass of a molecule including hydrogen, A is the total number of atoms in a molecule, $r_{i,m}$ is the m component of an atom i in a molecule (e.g., $m = 1$ corresponds to the x component), r_{CM} is the center of mass position of a molecule, m_i is the mass of the atom i, and t_i is the i-th diagonal component of the diagonalized gyration tensor S_{mn}.

(2) Eccenctricity (ϵ)

ϵ is defined[14] from principal moments of inertia tensor.

$$I = \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix}$$

$$I_{xx} = \sum_{i=1}^{A} (y_i^2 + z_i^2)\, m_i$$

$$I_{yy} = \sum_{i=1}^{A} (x_i^2 + z_i^2)\, m_i$$

$$I_{zz} = \sum_{i=1}^{A} (x_i^2 + y_i^2)\, m_i$$

$$I_{xy} = I_{yx} = -\sum_{i=1}^{A} x_i y_i\, m_i$$

$$I_{yz} = I_{zy} = -\sum_{i=1}^{A} y_i z_i\, m_i$$

$$I_{zx} = I_{xz} = -\sum_{i=1}^{A} z_i x_i\, m_i$$

$$I_A \leq I_B \leq I_C$$

where an inertia tensor is calculated with the reference point of the center of mass of a molecule. I_A, I_B, and I_C are principal moments of inertia obtained by diagonalizing the inertia tensor. ϵ is defined as

$$\epsilon = \frac{\sqrt{I_C^2 - I_A^2}}{I_C}$$

$$0 \leq \epsilon \leq 1$$

(3) Inertial shape factor (S_I)

S_I is defined[15] from principal moments of inertia.

$$S_\mathrm{I} = \frac{I_B}{I_A \, I_C}$$

(4) Radius of gyration (R_g)

R_g is defined[14] from a gyration tensor S_{mn}.

$$R_\mathrm{g}^2 = \mathrm{Tr}(S_{mn})$$

(5) Spherocity (Ω_S)

Ω_S is defined[15] from the eigenvalues of the covariance matrix of the atomic coordinates (a gyration tensor without mass weights).

$$\begin{aligned} \lambda_{mn} &= \frac{1}{N}\sum_{i=1}^{A}(r_{i,m} - r_{center,m})(r_{i,n} - r_{center,n}) \\ \Omega_\mathrm{S} &= \frac{3\,\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} \\ \lambda_1 &\le \lambda_2 \le \lambda_3 \end{aligned}$$

where N is the total number of atoms in a molecule including hydrogen, r_{center} is the center position of a molecule, and λ_i is the i-th diagonal component of the diagonalized covariance matrix λ_{mn}.

3. Experimental Correlation of Mean Squared End-to-End Distance per Mass of Polymers in the Melt and Glass Transition Temperature with Φ index.

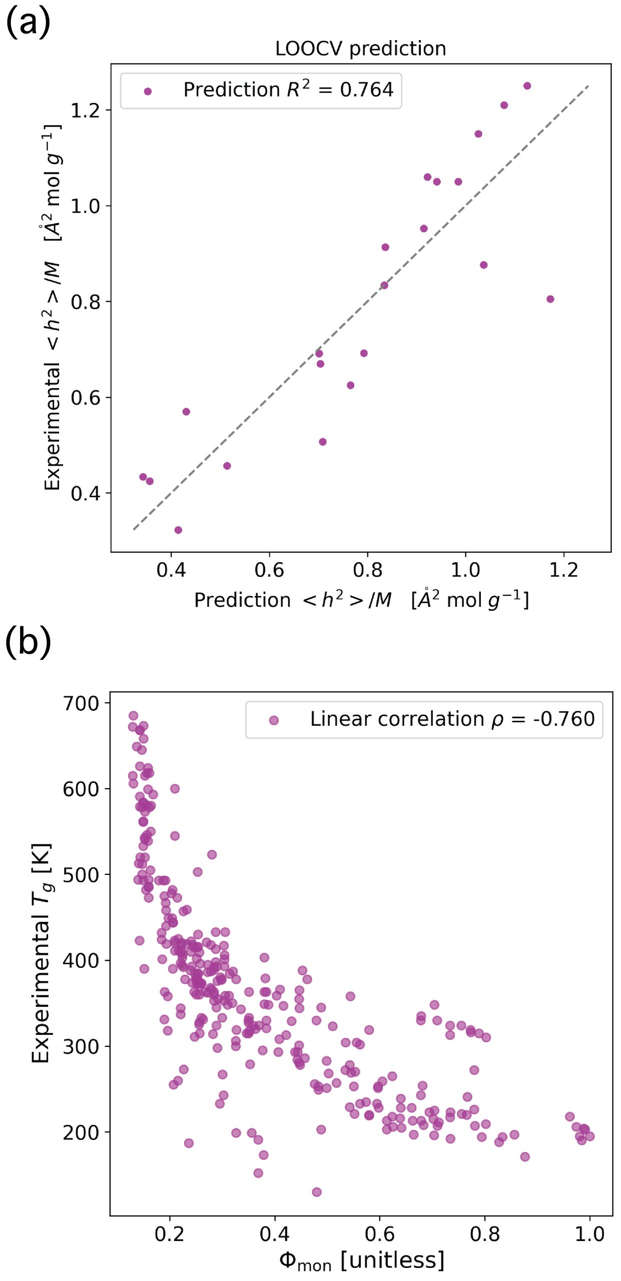


Figure S1: (a) Leave-one-out cross-validation (LOOCV) linear regression for experimental $\langle h^2\rangle_0 / M$ of polymers in the melt. The $\langle h^2\rangle_0 / M$ was predicted with linear regression from Φ_{mon} and Φ_{bb}. (b) The correlation between experimental glass transition temperatures and Φ index

The Phi (Φ) index showed a positive linear correlation with a computational molecular conformational entropy predicted with a group additive fashion.[16] The Φ index estimates molecular flexibility by counting the number of length-1 and length-2 paths in a 2D molecular graph without hydrogen. Additionally, the Φ index was normalized by the number of atoms (not including hydrogen) in OMG CRUs because the Φ index tends to increase with the number of atoms.

Figure S1a shows the leave-one-out cross-validation (LOOCV) linear regression for the experimental $\langle h^2 \rangle_0 / M$ of polymers in the melt. We used experimental $\langle h^2 \rangle_0 / M$ values measured at 413K for polymers in the melt from the literature.[17] We manually processed SMILES strings[3] for polymer constitutional repeating units[1] (CRUs) in the literature[17] and obtained 21 $\langle h^2 \rangle_0 / M$ values for polymers that had no ambiguity in calculating Φ_{mon} and Φ_{bb} from polymer CRUs. The Φ_{mon} values were prepared using RDKit[12] by calculating the molecular flexibility Phi (Φ) index followed by a normalization with the number of heavy atoms in a molecule. Additionally, we applied our own program to remove the side chains deeper than length-1 in a molecule and estimated normalized molecular flexibility for the backbone Φ_{bb}. The processed experimental $\langle h^2 \rangle_0 / M$ values and Python implementation of Φ estimation is available at `https://github.com/TheJacksonLab/OMG_PhysicalProperties`.

Figure S1b shows the correlation between experimental glass transition temperatures and Φ_{mon}. The experimental glass transition temperatures were obtained from in the Bicerano Handbook.[18,19] The strong negative linear correlation ($\rho \approx -0.76$) indicates that Φ_{mon} can capture the chain stiffness.[20]

4. DFT Calculations

The electronic properties were calculated with DFT single-point calculations implemented in Orca[21] (functional: revPBE-D3/ basis set: def2-SVP) performed on optimized geometries for OMG CRUs. In DFT calculations, the revPBE-D3 functional was adopted due to its high accuracy among generalized gradient approximation (GGA) functionals.[22,23] We also used the implicit solvation (CPCM[24]) for toluene as a solvent having a dielectric constant ($\epsilon = 2.4$) similar to that of conventional polymers at room temperature.[10] The same functional, basis set, and solvation as in the electronic property calculations are used in TD-DFT calculations.

5. Prediction of Experimental Flory-Huggins χ Parameters of Polymer Solutions

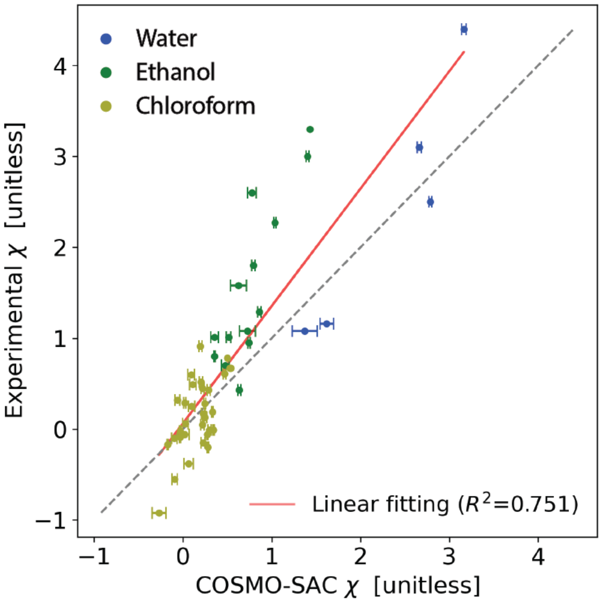


Figure S2: Predictions for experimental Flory-Huggins χ parameters for polymer solutions (volume fraction $\phi_{\text{polymer}} \geq 0.2$) with color representing three different solvents of water ($\epsilon = 80.4$), ethanol ($\epsilon = 24.3$), and chloroform ($\epsilon = 4.9$).

Flory-Huggins χ interaction parameters for OMG CRUs were calculated with three different solvents of varying dielectric constants: water (ϵ = 80.4), ethanol (ϵ = 24.3), and chloroform (ϵ = 4.9). Flory-Huggins χ interaction parameters describe thermodynamics of binary mixture.[25–28] Flory-Huggins χ interaction parameters of a polymer solution can be predicted using analytical functional forms of $\chi(\phi, \mathrm{T})$ with the dependency of a polymer volume fraction (ϕ) and temperature (T).[29–32] However, these analytical functions require experimental data to fit adjustable parameters for a specific polymer solution. Alternatively, Flory-Huggins χ parameters can be calculated from activity coefficients of solute and solvent molecules obtained from COSMO-SAC calculations[33] applied to σ-profiles[34] describing the distribution of the surface screening charges of a molecule. Activity coefficients of solute and solvent molecules obtained from COSMO-SAC[33] can be used to calculate a chemical potential difference during mixing, thereby estimating Flory-Huggins χ parameters.[35] Similarly, Yoshida et al.[5] demonstrated that experimental χ values can be predicted ($R^2 \approx 0.62$) with COSMO-RS calculations.[36] The COSMO-SAC calculations rather than fitting of an analytical function to experimental data were adopted to estimate Flory-Huggins χ interaction parameters of OMG polymer solutions given the extensive chemical space of OMG CRUs.[2] We modified the open benchmark implementation of COSMO-SAC[37] to be compatible with DFT output files from Orca.[21] We converted the Orca output files (.cpcm) to σ-profiles describing the surface-charge distribution of a molecule with a conductor-like solvent ($\epsilon \sim \infty$), and the σ-profile was used for COSMO-SAC calculations. This modified implementation was tested with experimental χ values of polymers solutions[38] (volume fraction $\phi_{\text{polymer}} \geq 0.2$ to avoid a critical regime ($\phi_c \approx 0$) of polymer solutions with water, ethanol, and chloroform as a solvent. The χ values from the COSMO-SAC calculations showed a strong linear correlation with experimental χ values ($R^2 \approx 0.75$) as detailed in **Figure S2**. We calculated Flory-Huggins χ interaction parameters for polymer solutions with a fixed polymer volume fraction $\phi = 0.2$ with the number of backbone atoms in a polymer $\approx 1,000$. The COSMO-SAC implementation is available at `https://github.com/TheJacksonLab/OMG_PhysicalProperties`

Figure S2 shows the predictions for experimental Flory-Huggins χ parameters for polymer solutions using COSMO-SAC calculations.[33] We extracted 52 experimental χ values for polymer solutions with water, ethanol, and chloroform from the literature[38] with a polymer volume fraction $\phi_{\text{polymer}} \geq 0.2$ to avoid a critical regime $\phi_c \approx 0$ where the Flory-Huggins equation is not valid. To predict experimental χ values, we first obtained σ-profiles[34] of methyl-terminated monomers and solvent molecules under a conductor-like ideal solvent ($\epsilon \sim \infty$). We then multiplied the σ-profile of a methyl-terminated monomer by a constant to approximate the σ-profile of a polymer with the number of backbone atoms $\approx 1,000$. From the σ-profiles of a polymer and a solvent molecule, the COSMO-SAC calculations estimated activity coefficients of a polymer and a solvent molecule. Flory-Huggins χ parameters were obtained by comparing mixing free energies from activity coefficients and Flory-Huggins equations.[35] The implementation for the COSMO-SAC calculation is available at `https://github.com/TheJacksonLab/OMG_PhysicalProperties`.

6. Details on Active Learning

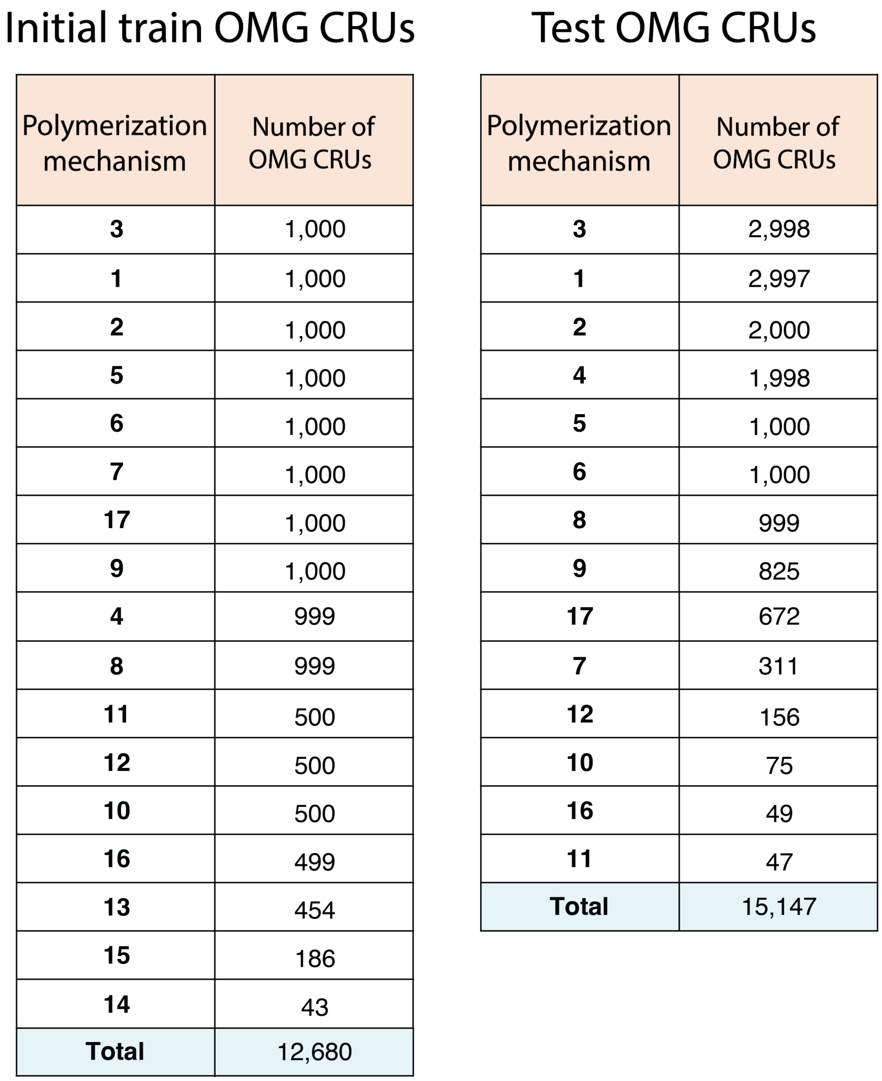
We developed ML models to estimate monomer-level properties for 12M OMG CRUs via uncertainty-guided active learning. The computational calculations of monomer-level properties for 12M OMG CRUs posed an intractable computational cost. Approximately 4.8 CPU hours were needed to perform DFT single-point calculations for one OMG CRU with an average of 23 heavy atoms (standard deviation of 9) consisting of up to 5 distinct conformers. We constructed ML models predicting monomer-level properties for the OMG CRUs to avoid the intractable computational cost of calculating monomer-level properties for 12M OMG CRUs. Especially, we adopted uncertainty-guided active learning to train ML prediction models with a reduced number of quantum chemistry calculations by sampling OMG CRUs with high prediction uncertainties to improve ML prediction models efficiently.[39,40] We combined evidential learning[41] with a directed message-passing 2D graph neural network (D-MPNN)[42] to estimate a ML model predictive uncertainty (Var(μ)) by assuming a normal inverse-Gamma prior distribution for an unknown mean (μ) and variance (σ^2) of a target prediction. The D-MPNN evidential learning was demonstrated as an active learning strategy in building a molecular property prediction model for 12 simultaneous physical properties of QM9 molecules[43] with a reduced computational cost compared to an ensemble method.[44]

In the active learning with evidential regression, we trained four different D-MPNN evidential networks, each for 3D geometry descriptors (5 properties), electronic properties (7 properties), optical properties (4 properties), and Flory-Huggins χ interaction parameters (3 properties). We did not train a D-MPNN evidential network for chemistry descriptors and molecular flexibility because these properties could be easily obtained from SMILES strings for OMG CRUs without 3D geometry. When training D-MPNN evidential networks, 200 RDKit global molecular features were concatenated to OMG monomer embedding vectors from message-passing of D-MPNN evidential networks to overcome a local nature of a message-passing network.[42] It is important to note that D-MPNN evidential learning only needs the 2D molecular graph of a methyl-terminated OMG CRU without 3D geometry for property prediction. Therefore, the D-MPNN evidential network can predict monomer-level properties and corresponding prediction uncertainties for 12M OMG CRUs without 3D atomic coordinates of 12M OMG CRUs.

D-MPNN evidential networks underwent iterative training with OMG CRUs having high ML prediction uncertainties during the active learning campaign. Approximately 12k OMG CRUs were randomly sampled as an initial dataset based on polymerization mechanisms from the 12M OMG CRUs as detailed in (**Figure S3**). Quantum chemistry calculations were applied to the sampled OMG CRUs to obtain 19 different monomer-level properties to train four D-MPNN evidential networks. The trained D-MPNN evidential networks estimated prediction uncertainties for 19 monomer-level properties for the unseen OMG CRUs. To sample OMG CRUs for the next round of active learning, we searched for non-dominated OMG CRUs located on the Pareto front of the 19-dimensional prediction uncertainty space using a non-dominated sorting algorithm.[45] The Pareto front represents the set of non-dominated OMG CRUs where an increase in ML prediction uncertainty for given monomer-level property is only possible by reducing some of the other ML prediction uncertainties. We applied the Pareto front search algorithm[45] for a subset of OMG CRUs with high mean prediction uncertainties to reduce a computational cost for the Pareto front search as detailed in **Figure S4**. The active learning campaign continued with the sampled OMG CRUs from the Pareto front of the uncertainty space until the ML models stopped showing a significant improvement in prediction performance. Especially, we sampled 10k OMG CRUs for the first round of active learning and 5k OMG CRUs afterward. We decreased the sampling size from 10k to 5k from Round 2 to reduce the computational cost for monomer-level property calculations because the active learning tended to sample OMG CRUs with an increased number of heavy atoms possessing high prediction uncertainties (**Figure S5**).

We also tried different sampling strategies on QM9 molecules[43] to choose the best sampling strategy for OMG CRUs. The Pareto uncertainty sampling we adopted showed the best performance for QM9 molecules compared to random and mean uncertainty sampling as detailed in **Figure S6**. After the active learning campaign, the trained D-MPNN evidential networks were used to predict 19 monomer-level properties for the 12M OMG CRUs.

6-1) Compositions of initial train and test OMG CRUs



Initial train OMG CRUs

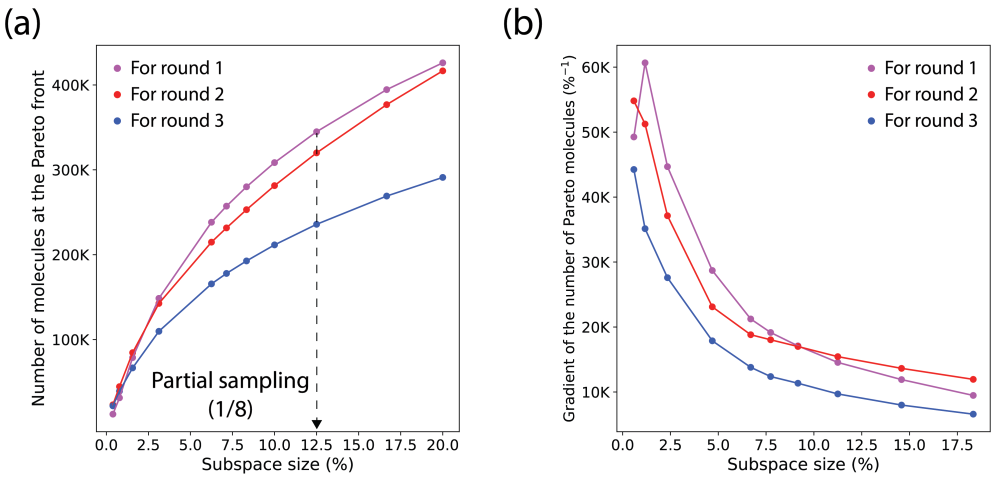
Polymerization mechanism	Number of OMG CRUs
3	1,000
1	1,000
2	1,000
5	1,000
6	1,000
7	1,000
17	1,000
9	1,000
4	999
8	999
11	500
12	500
10	500
16	499
13	454
15	186
14	43
Total	12,680

Test OMG CRUs

Polymerization mechanism	Number of OMG CRUs
3	2,998
1	2,997
2	2,000
4	1,998
5	1,000
6	1,000
8	999
9	825
17	672
7	311
12	156
10	75
16	49
11	47
Total	15,147

Figure S3: Compositions of initial train and test OMG CRUs. The polymerization mechanism indices correspond to Figure 2 in the previous work.[2]

6-2) Pareto front search with high mean prediction uncertainties



(c)

Round 1 (10,000 polymers)

Polymerization mechanism	Number of OMG polymers
1	5,075
3	4,619
2	133
5	71
6	49
4	42
8	5
17	4
9	1
7	1

Round 2 (5,000 polymers)

Polymerization mechanism	Number of OMG polymers
1	2,543
3	2,309
2	79
5	34
6	17
4	16
8	1
17	1

Round 3 (5,000 polymers)

Polymerization mechanism	Number of OMG polymers
1	3,155
3	1,655
2	95
5	53
6	20
4	18
17	4

Figure S4: Sampling OMG CRUs located on the Pareto front of 19 dimensional prediction uncertainties during the active learning campaign. (a) Number of the OMG CRUs located on the Pareto front searched for a subspace of the OMG CRUs selected based on their mean prediction uncertainties for 19 monomer-level properties. (b) Gradient of the number of OMG CRUs (%$^{-1}$) on the Pareto front as the subspace size becomes larger. (c) Compositions of the randomly sampled OMG CRUs located on the Pareto front searched for 12.5% of the available space for active learning.

We sampled OMG CRUs based on prediction uncertainties from trained ML models during the active learning campaign. Especially, we searched for non-dominated OMG CRUs located on the Pareto front of the 19 dimensional prediction uncertainty space using a non-dominated sorting algorithm.[45] The Pareto front search algorithm[45] for approximately 12M OMG CRUs with 19 monomer-level properties, however, required a prohibitive computational cost of $\mathcal{O}(N\log^{K-1} N)$ at worst where where N is the number of molecules available (i.e., 12M), and K is the number of objectives (i.e., 19). To reduce a computational cost for the Pareto front search, we sorted the OMG CRUs based on their mean prediction uncertainties for 19 monomer-level properties from trained ML models and applied the Pareto front search algorithm[45] to selected OMG CRUs with high mean prediction uncertainties.

Figure S4a shows the number of the OMG CRUs located on the Pareto front of the selected OMG CRUs. For example, the subspace size of 12.5% means that the Pareto front was searched for the top 12.5% of the approximately 12M OMG CRUs with high mean prediction uncertainties for the next round of active learning. **Figure S4a** also shows that the number of OMG CRUs on the Pareto front increased when the subspace size became larger. The number of OMG CRUs on the Pareto front in **Figure S4a** is expected to approach that on that of the unfiltered OMG CRUs as the subspace size becomes 100%. We arbitrarily decided to apply the Pareto front search algorithm for the subspace size of 12.5%, which took approximately 17 hours with a single CPU, to detour the computational cost needed for the Pareto front search for the whole space.

Figure S4b displays the gradient of the number of OMG CRUs on the Pareto front with respect to the subspace size. In **Figure S4b**, the gradient was calculated from **Figure S4a** and mapped on the middle point of the two adjacent subspace sizes. **Figure S4b** indicates the increase amount in the number of the OMG CRUs at the Pareto front per subspace gradually decayed as the subspace size increased. We estimated the percentage of the Pareto OMG CRUs with the subspace size of 12.5% to that of the whole space (100%) assuming the gradient value in **Figure S4b** linearly decays with the subspace size afterwards. The subspace size of 12.5% was estimated to include 49.80% (for Round 1), 41.62% (for Round 2), and 49.66% (for Round 3) of the Pareto OMG CRUs for the whole space (100%) for each round. This large percentage of the Pareto OMG CRUs included in the 12.5% subspace implies that the subspace selection based on mean prediction uncertainties might be a computationally efficient strategy to search the Pareto front for a huge chemical space. It is important to note that the number of OMG CRUs on the Pareto front of the 12.5% subspace is still larger than the sampling size (e.g., 10,000 or 5,000). Therefore, we randomly sampled OMG CRUs (10,000 or 5,000) from the OMG CRUs on the Pareto front of the 12.5% subspace for the next round of active learning. **Figure S4c** shows the compositions of the randomly sampled OMG CRUs. The implementation of the Pareto front search with high mean prediction uncertainties is available at `https://github.com/TheJacksonLab/OMG_PhysicalProperties`.

6-3) Distributions of the number of heavy atoms for the sampled OMG polymers during active learning

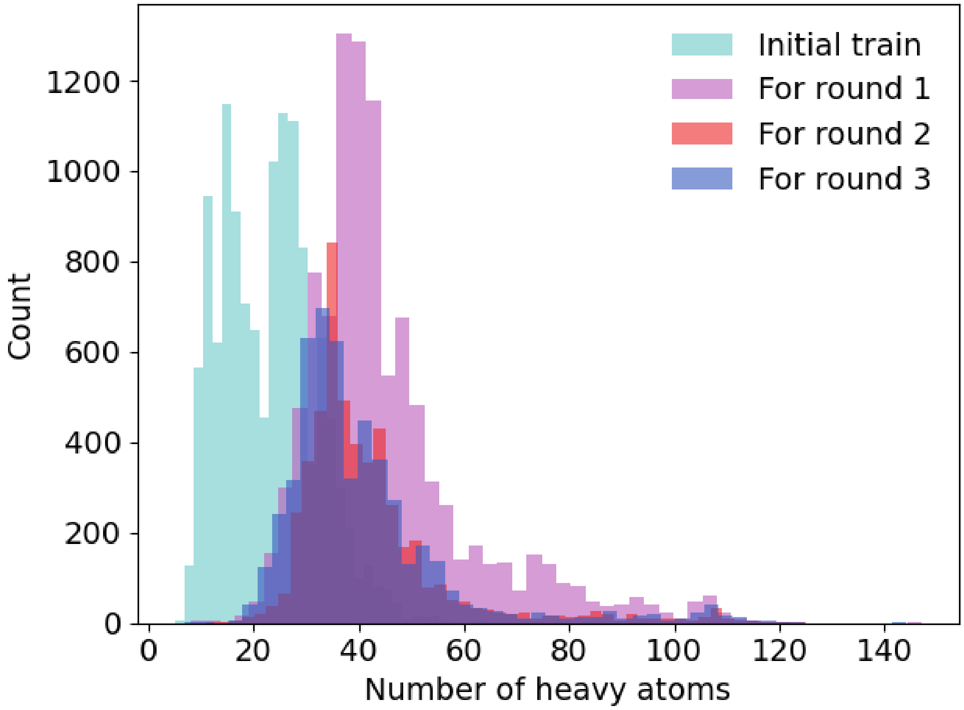


Figure S5: Distributions of the number of heavy atoms in the sampled OMG CRUs during the active learning campaign.

Figure S5 shows the histogram of the number of heavy atoms in OMG CRUs sampled during active learning. The initial train OMG CRUs (12,683 CRUs) were randomly sampled based on polymerization mechanisms. The OMG CRUs for Round 1 (10,000 CRUs), Round 2 (5,000 CRUs), and Round 3 (5,000 CRUs) were randomly sampled from the Pareto front of the 12.5% subspace of the OMG CRUs for each round. **Figure S5** shows that the number of heavy atoms in the sampled OMG CRUs for Round 1, Round 2, and Round 3 is larger than that of the initial train. This implies that OMG CRUs with the more number of heavy atoms have higher prediction uncertainties for molecular-size relevant properties such as radius of gyration and polarizability. We decreased the sampling size from 10,000 to 5,000 for Round 2 and Round 3 to reduce a computational cost for the following quantum chemistry calculations.

6-4) Active learning sampling strategy tested on QM9

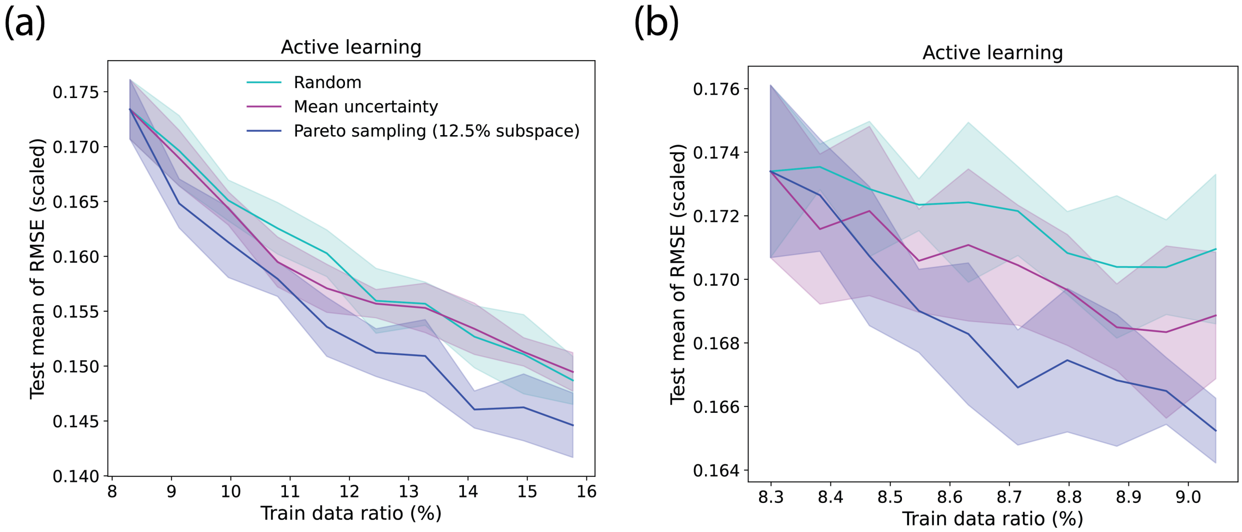


Figure S6: Active learning strategies tested on QM9 with (a) 1,000 molecules or (b) 100 molecules added for each round of active learning. The average values of root mean square errors for each molecular property in QM9 are plotted with respect to the ratio of the train data. The root mean square errors for each property were scaled with their mean and standard deviation. The error bars are from five different random seeds for train-test split. The solid line represents the average value from five random seeds, and the shade denotes one standard deviation away from the average value.

Figure S6 shows the active learning results of QM9 with different sampling strategies including random, mean uncertainty, and Pareto front for a 12.5% subspace. The prediction uncertainties from a single D-MPNN network[42] for 12 molecular properties in QM9 were utilized to sample molecules during active learning. As a baseline, the random strategy sampled the molecules randomly for the next round of active learning ignoring prediction uncertainties from a D-MPNN network. The mean uncertainty strategy sorted the molecules based on their scaled prediction mean uncertainties and chose either the top 1,000 (**Figure S6a**) or 100 molecules (**Figure S6b**) with high mean prediction uncertainties for the next round of active learning. The Pareto sampling for the 12.5% subspace strategy picked the top 12.5% of the molecules with high mean prediction uncertainties and sampled molecules on the Pareto front for the subspace. The subsequent Pareto front was also considered if the number of molecules on the first Pareto front was not larger than the sampling size (i.e., 1,000 or 100). **Figure S6a** and **Figure S6b** imply that the Pareto sampling for the subspace might be a effective active learning strategy to increase a model prediction accuracy for a multi-task learning.

7. Active Learning Performance

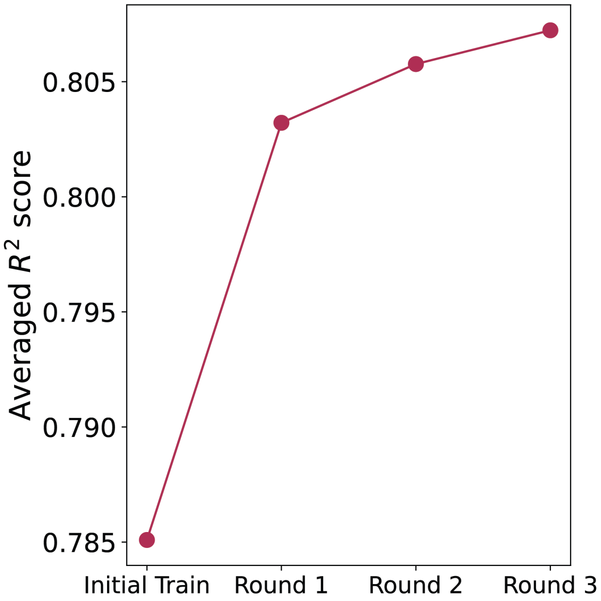


Figure S7: The averaged R^2 score over monomer-level properties are plotted during the active learning campaign. The active learning campaign was stopped at Round 3 when the averaged R^2 score was not expected to increase significantly.

Figure S7 shows averaged R^2 scores over 19 monomer-level properties for the test OMG CRUs during the active learning campaign. For example, Round 1 in **Figure S7** denotes that the D-MPNN evidential networks were trained on the initial train OMG CRUs plus 10k OMG CRUs sampled from the Pareto front search for Round 1. The D-MPNN evidential networks exhibited increasing averaged R^2 scores and achieved the averaged $R^2 \approx 0.807$ at Round 3

8. Active Learning Stopping Criterion with Uncertain OMG Polymers

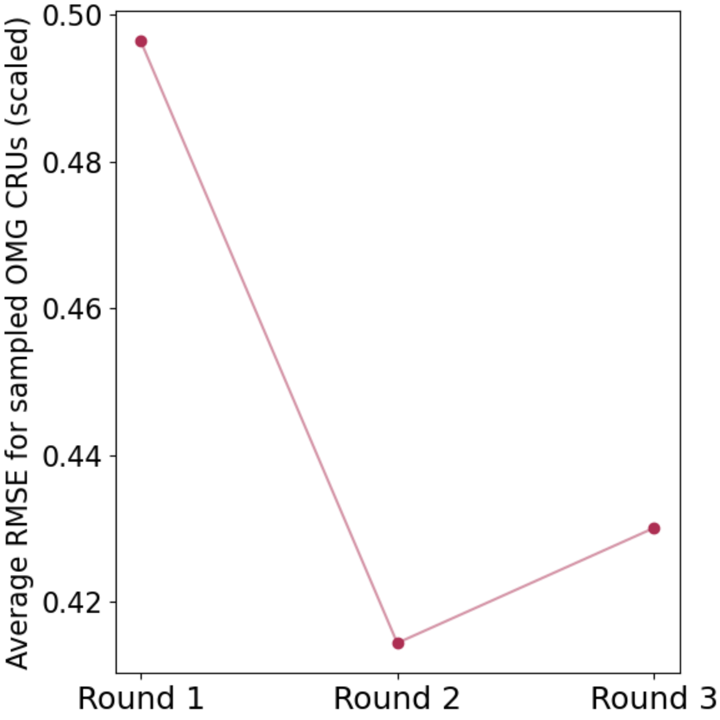
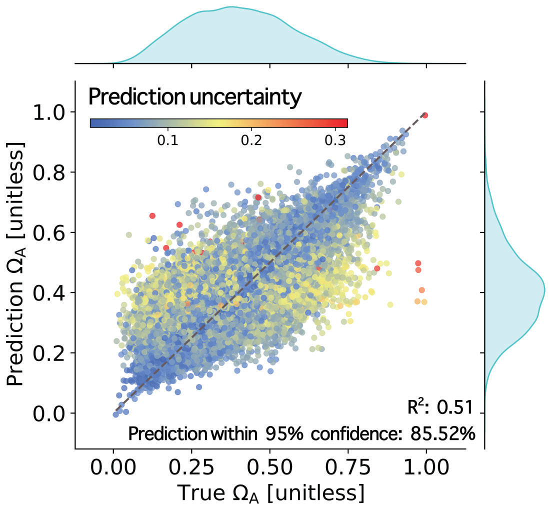


Figure S8: Averaged root mean square errors for 19 monomer-level properties of the sampled OMG CRUs during the active learning campaign. The root mean square errors for each monomer-level property were normalized with their mean and standard deviation.

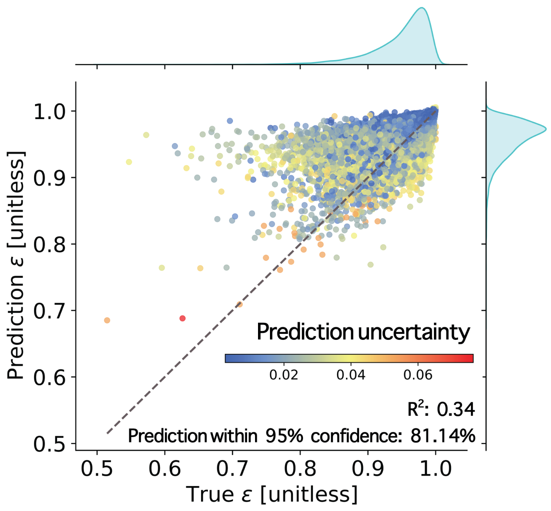
Figure S8 shows the average of root mean square errors for 19 monomer-level properties of the sampled OMG CRUs during the active learning campaign. The training data can be used to decide when to stop active learning.[46] The sampled OMG CRUs for each round represent molecules where the trained ML models are anticipated to exhibit high prediction errors. Therefore, the prediction errors for the sampled OMG CRUs can indicate the prediction accuracy of the trained ML models.[46] **Figure S8** displays that the average error at Round 1 is the largest, which is consistent with that the trained models showed a huge increase in the test R^2 score at Round 1 in **Figure S7**. The relatively small average error at Round 2 and Round 3 in **Figure S8** is also consistent with that the improvement in test prediction accuracy of trained ML models of Round 2 or Round 3 may not be significant as that of Round 1 as described in **Figure S7**. Overall, **Figure S8** implies that the improvement in the prediction accuracy of trained ML models may not be significant after Round 3 assuming that the prediction error for the sampled OMG CRUs does not increase significantly afterwards.

9. Predictions for 19 Monomer-Level Properties

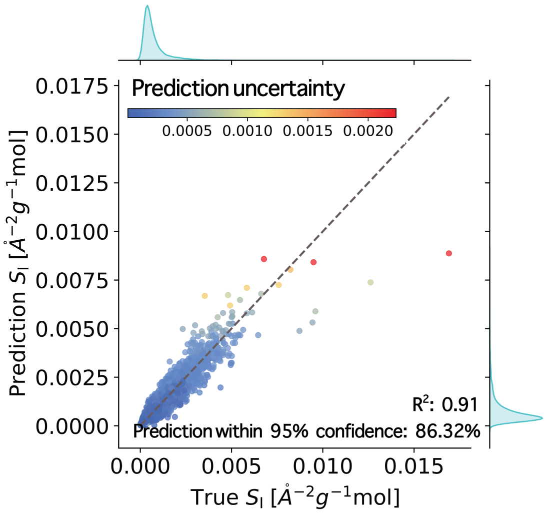
(1) Asphericity



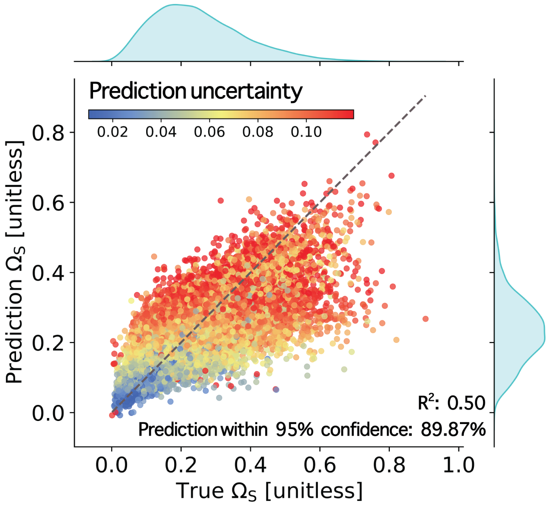
(2) Eccentricity



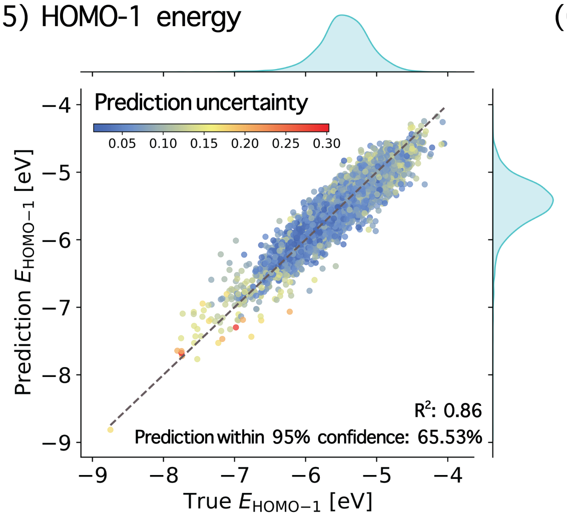
(3) Inertial shape factor



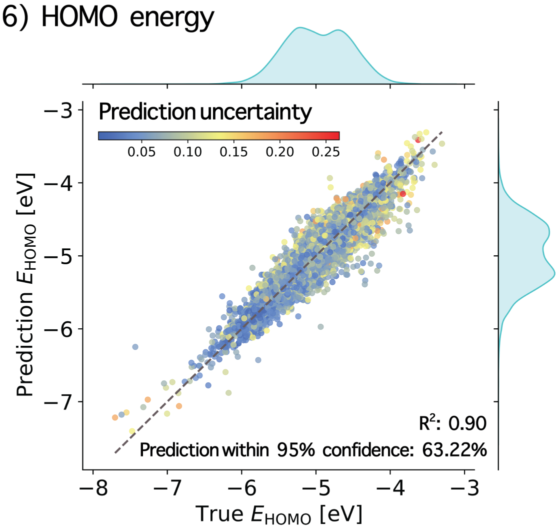
(4) Spherocity

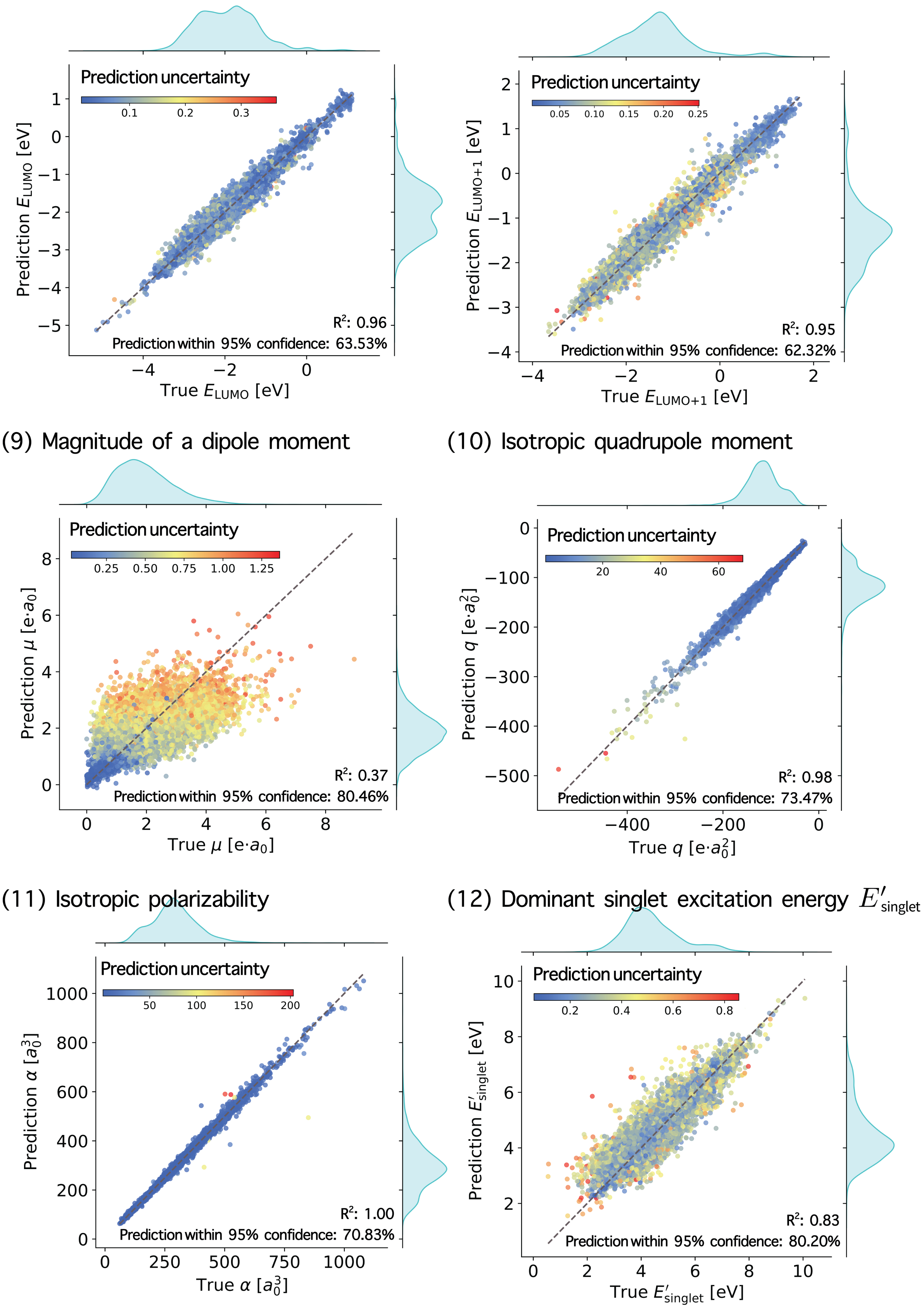


(5) HOMO-1 energy



(6) HOMO energy





(7) LUMO energy

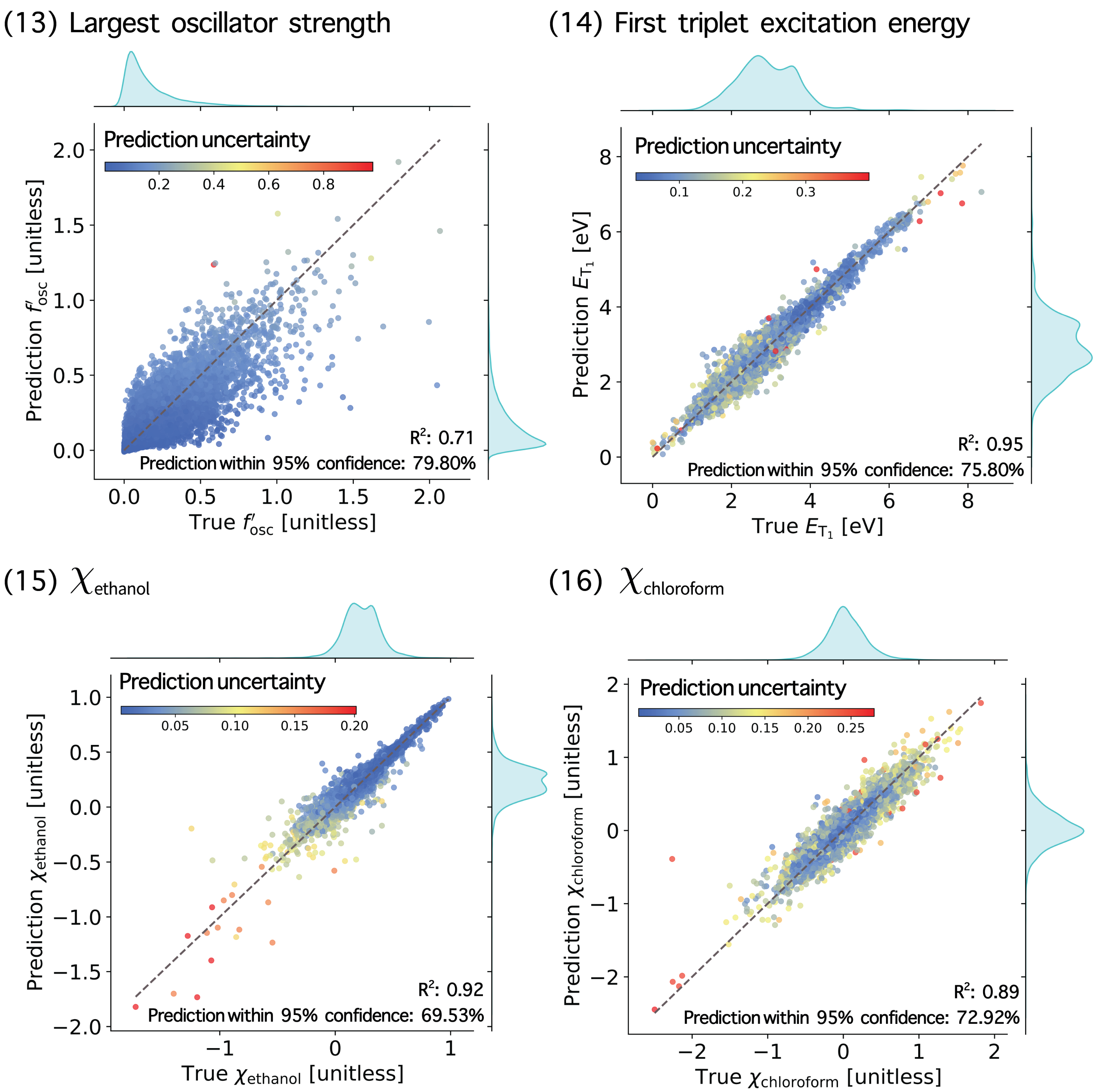
(8) LUMO+1 energy

(9) Magnitude of a dipole moment

(10) Isotropic quadrupole moment

(11) Isotropic polarizability

(12) Dominant singlet excitation energy E'_{singlet}



(13) Largest oscillator strength

(14) First triplet excitation energy

(15) χ_{ethanol}

(16) $\chi_{\text{chloroform}}$

Figure S9: Monomer-level property prediction for the test OMG polymers after the active learning campaign. The colorbar indicates prediction uncertainties. The prediction R^2 score and the percentage of test OMG CRUs within a predictive Gaussian distribution $N(\hat{y}_{i,\text{prediction}}, \sigma^2_{i,\text{calibrated uncertainty}})$ are computed.

10. Uncertainty Calibration

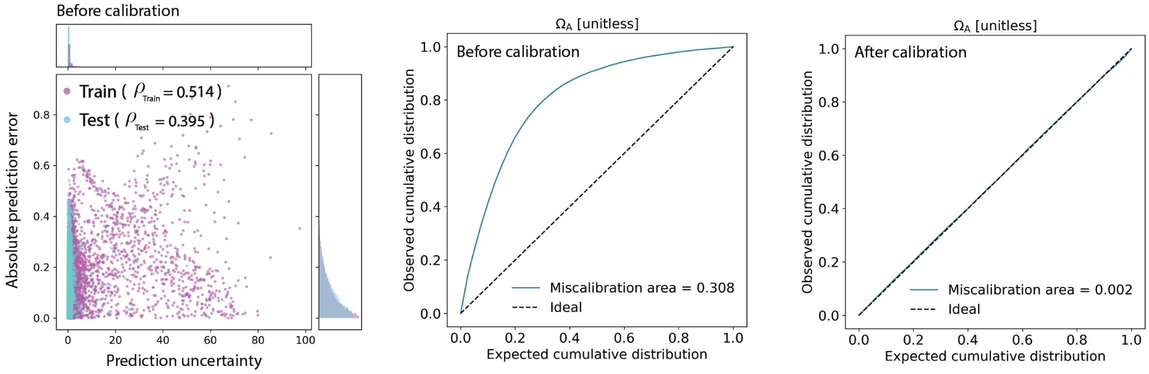
After the active learning campaign, we scaled prediction uncertainties from trained D-MPNN evidential networks[42] for monomer-level properties using a quantile-based calibration.[47] Deep evidential regression[41] has a regularization term to assign a high prediction uncertainty to a prediction with a large error. We used a regularization coefficient (λ) of 0.2 for D-MPNN evidential networks of all 19 OMG monomer properties during active learning because the regularization coefficient of 0.2 was suggested for molecular property predictions.[44] We did not perform additional regularization coefficient optimization because further regularization coefficient regularization might be biased toward a small portion of the OMG chemical space (the initial training set occupies a only small portion ≈0.1% of the total OMG CRUs).

The uncertainties from D-MPNN evidential networks generally increase with absolute prediction errors with decent rank correlations as shown in **Figure S10**. However, quantile-based calibration plots[47] in **Figure S10** show that the trained D-MPNN evidential networks after the active learning campaign estimated larger uncertainty (underconfident) than calibrated uncertainty estimates on the training set (≈32k monomers) with the assumption that target OMG monomer-level properties from quantum chemistry calculations (y_i) follow a Gaussian distribution of $N(\hat{y}_i, \sigma^2)$ where $\hat{y}_i$ is a prediction value, and σ is a prediction uncertainty from deep evidential regression.[44] The predicted uncertainties were also larger than absolute prediction errors in order of magnitudes as displayed in **Figure S10**.

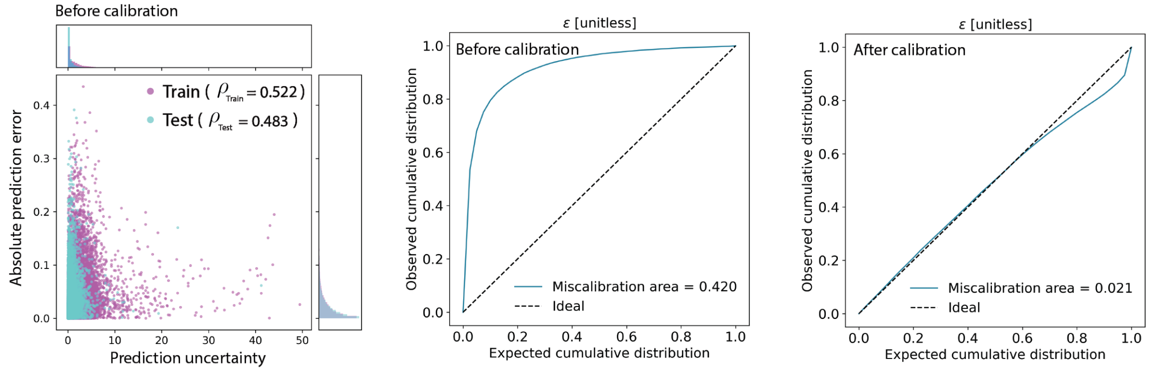
We calibrated uncertainty estimates from D-MPNN evidential networks to achieve better interpretability. Especially, we scaled prediction uncertainties with non-linear scaling[48] with a non-decreasing isotonic regression to match the scale between prediction uncertainty and absolute prediction error. It is important to note that a non-decreasing isotonic regression (f) preserves the rank order of the uncertainty estimates; that is $f(\sigma_A) \geq f(\sigma_B)$ if $\sigma_A > \sigma_B$ where σ_A and σ_B are prediction uncertainties for a OMG CRU A and a OMG CRU B, respectively. Each of 19 OMG monomer-level properties was calibrated with a non-decreasing isotonic regression followed by linear scaling to minimize the miscalibration area[49] on the training dataset (≈32k

OMG monomers) from the active learning campaign. We also tried a linear scaling[50,51] of prediction uncertainties without an isotonic regression, but a non-linear scaling with an isotonic regression outperformed the linear scaling in terms of several criteria[49] including miscalibration area, sharpness, and negative log-likelihood. The detailed uncertainty calibration procedure is available at `https://github.com/TheJacksonLab/OMG_PhysicalProperties`.

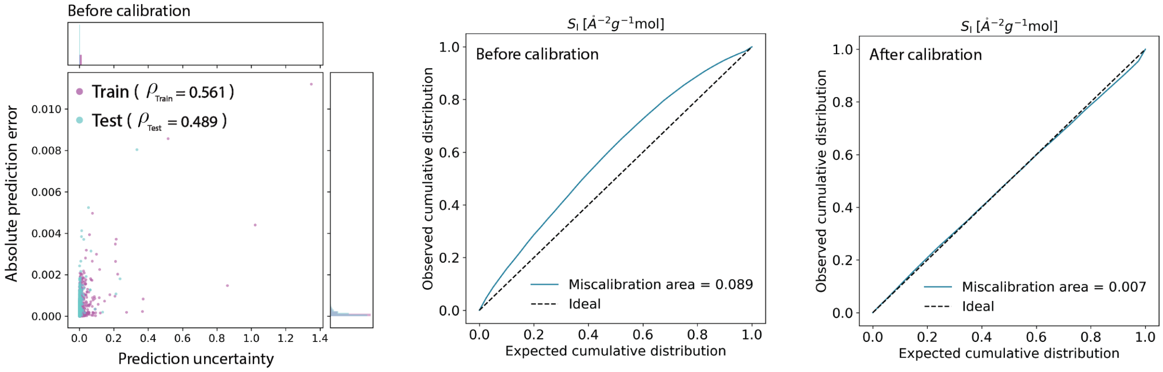
(1) Asphericity



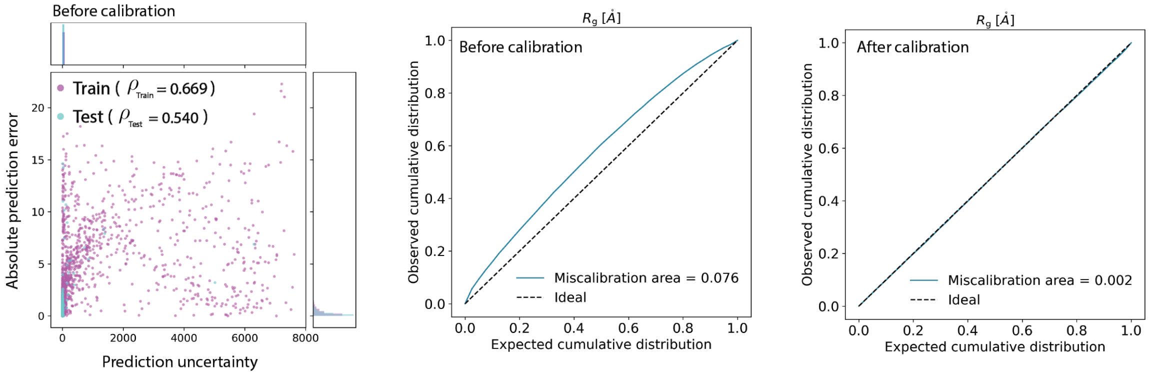
(2) Eccentricity



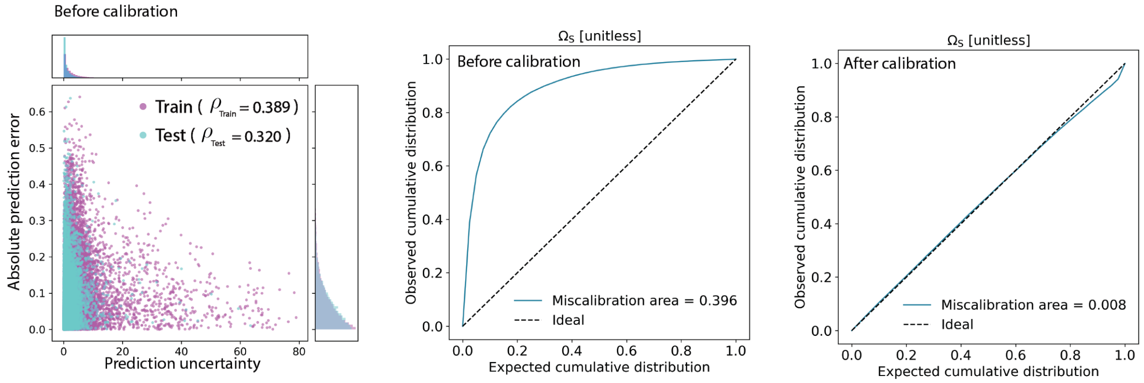
(3) Inertial shape factor



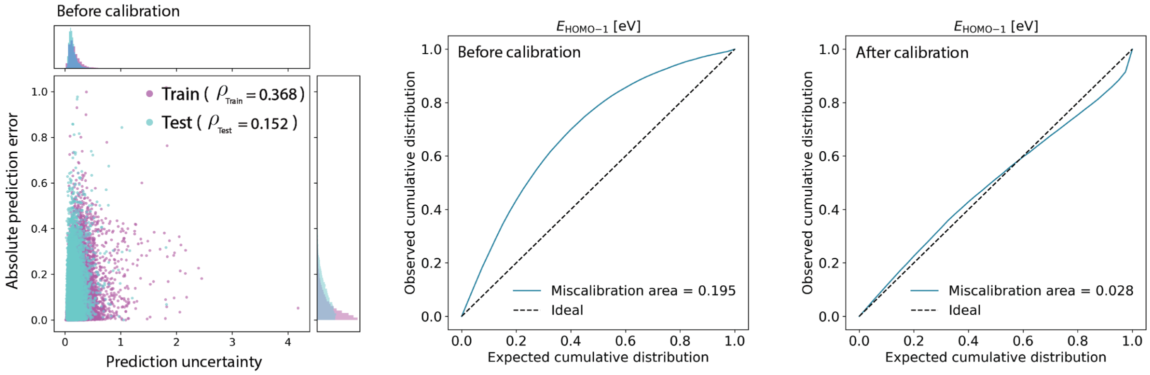
(4) Radius of gyration



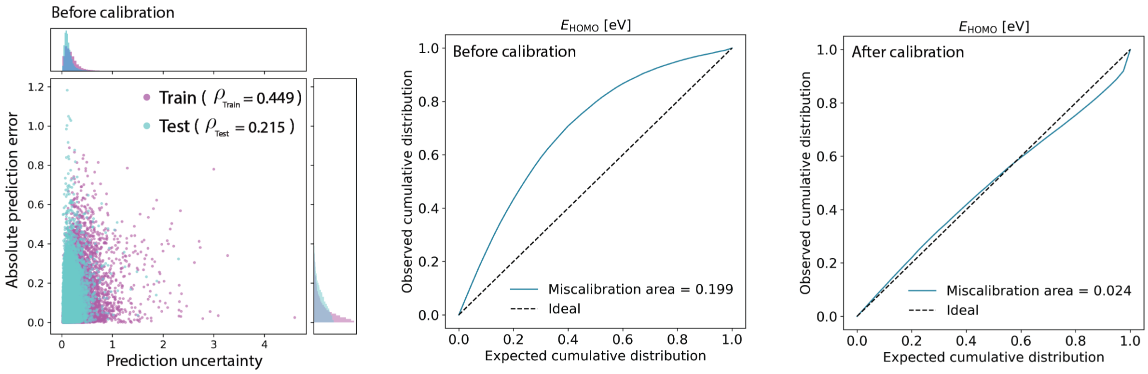
(5) Spherocity



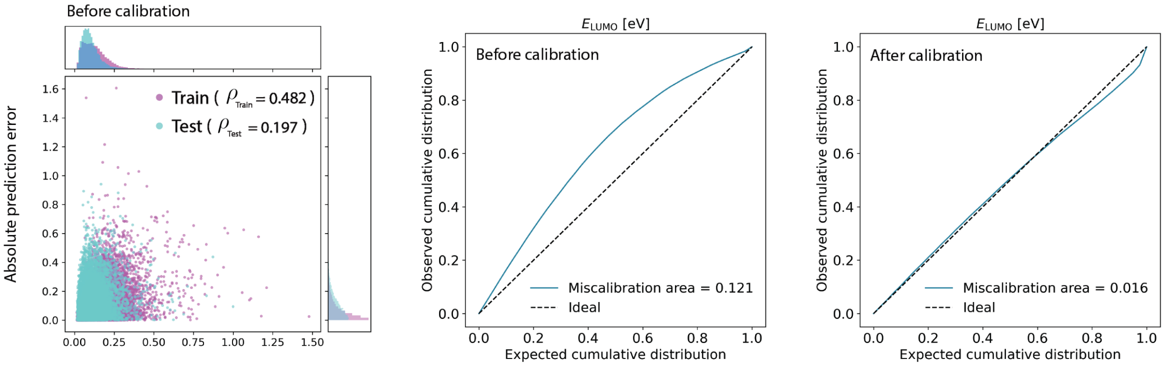
(6) HOMO-1 energy



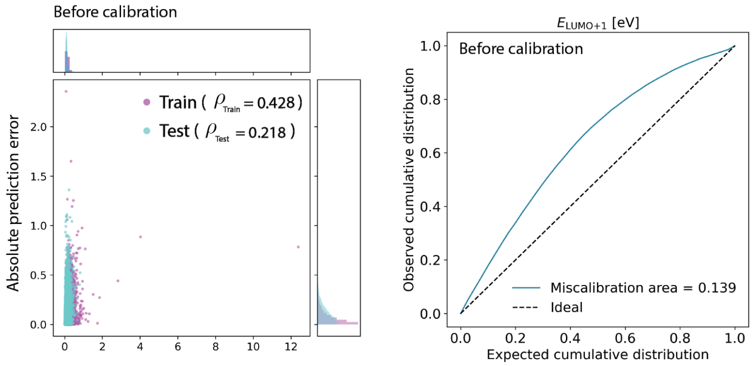
(7) HOMO energy



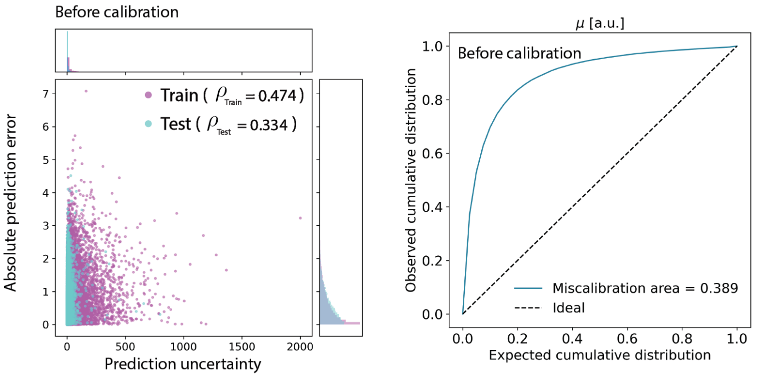
(8) LUMO energy



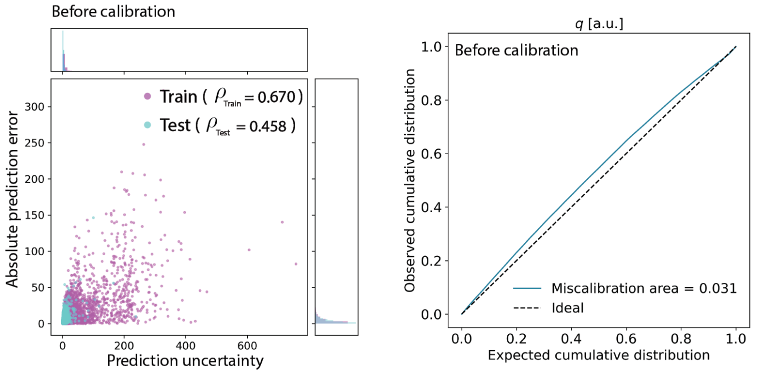
(9) LUMO+1 energy



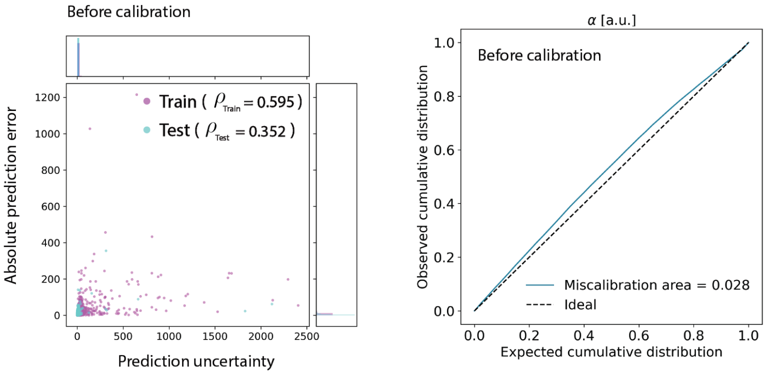
(10) Magnitude of a dipole moment vector



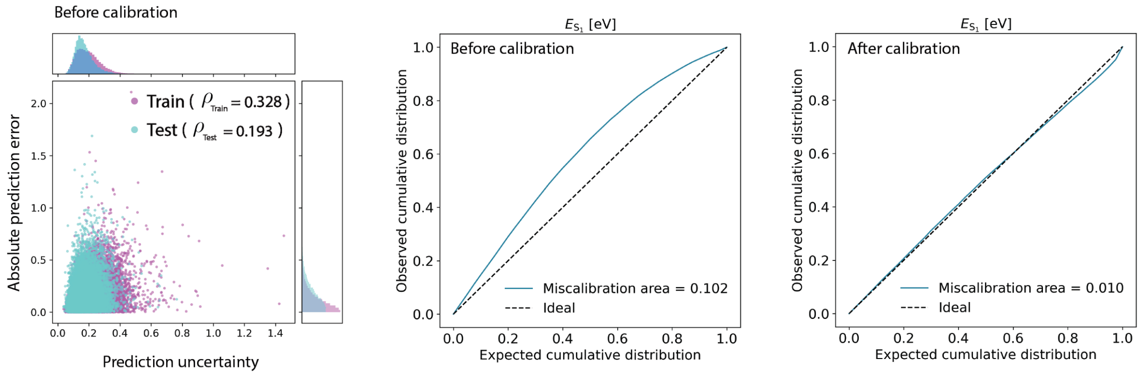
(11) Isotropic quadrupole moment



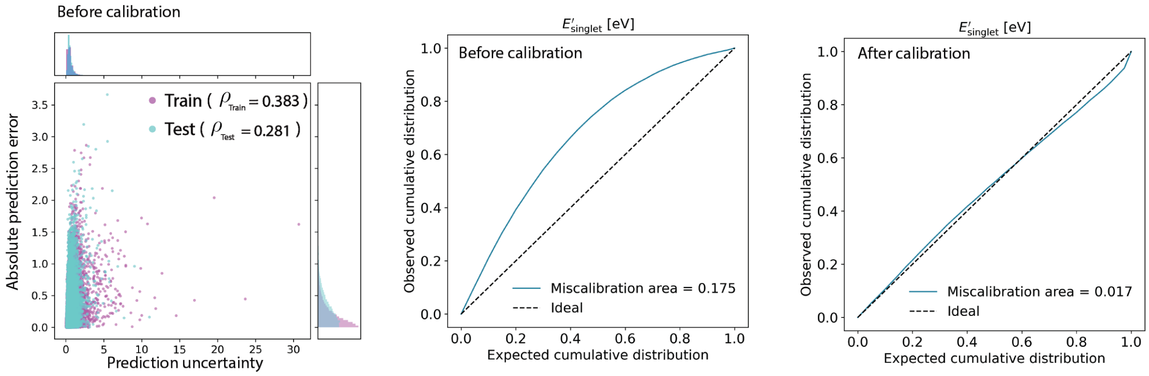
(12) Isotropic polarizability



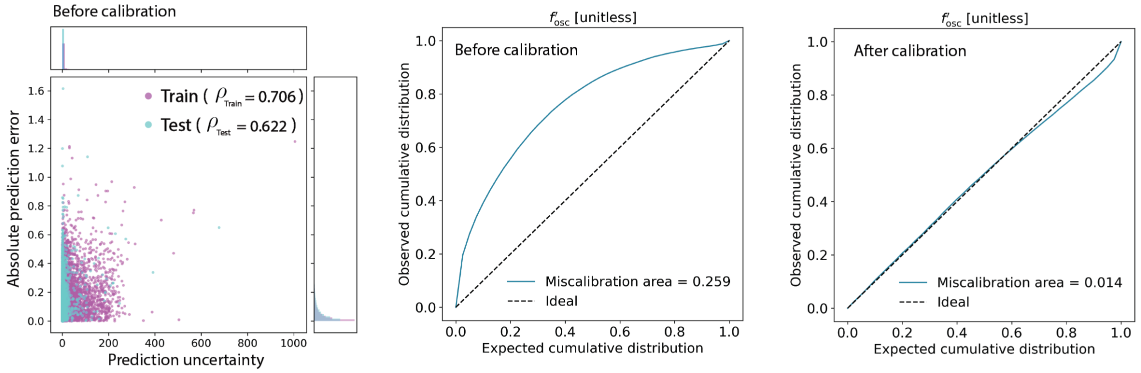
(13) First singlet excitation energy



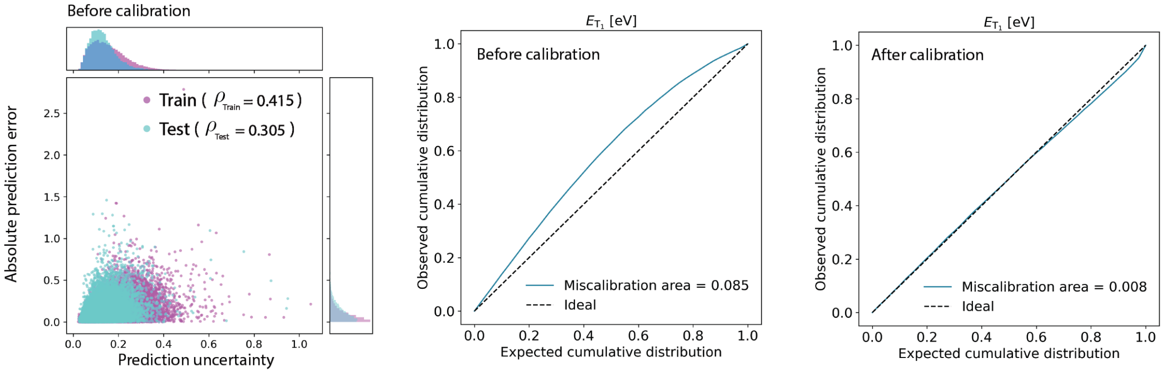
(14) Singlet excitation energy with the largest oscillator strength



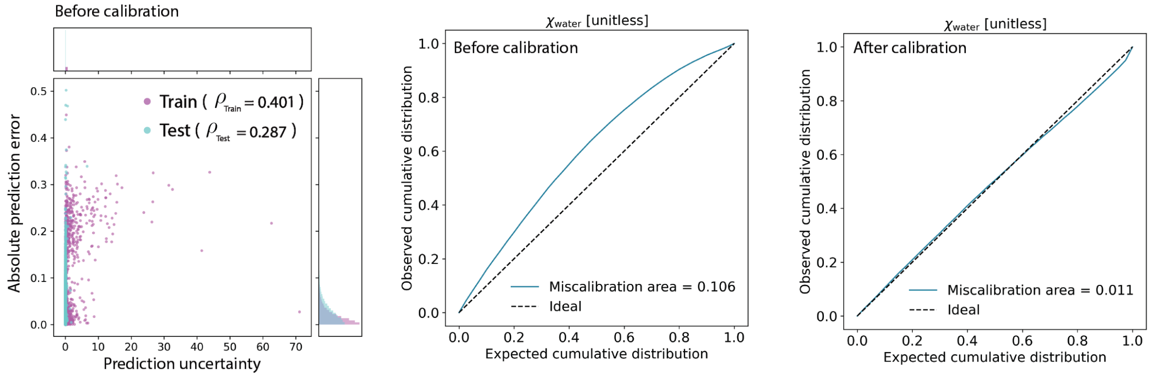
(15) Largest oscillator strength for singlet excitations



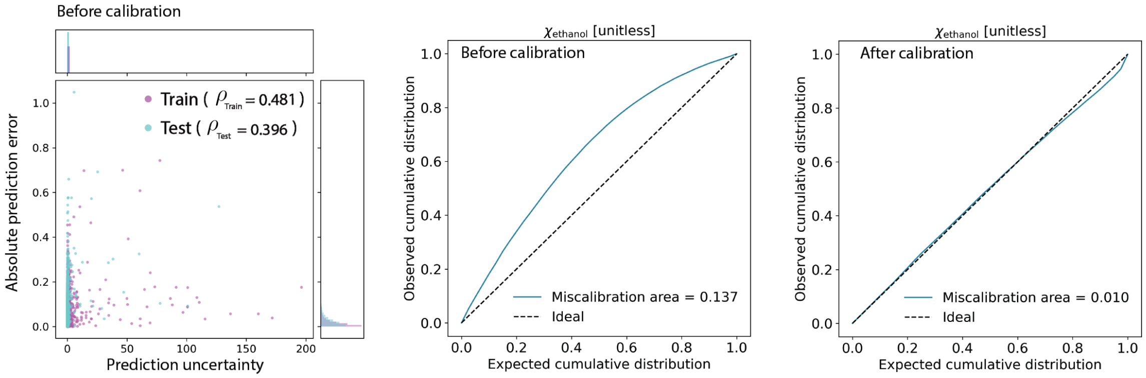
(16) First triplet excitation energy



(17) Flory-Huggins χ parameters of a polymer solution with a water solvent



(18) Flory-Huggins χ parameters of a polymer solution with a ethanol solvent



(19) Flory-Huggins χ parameters of a polymer solution with a chloroform solvent

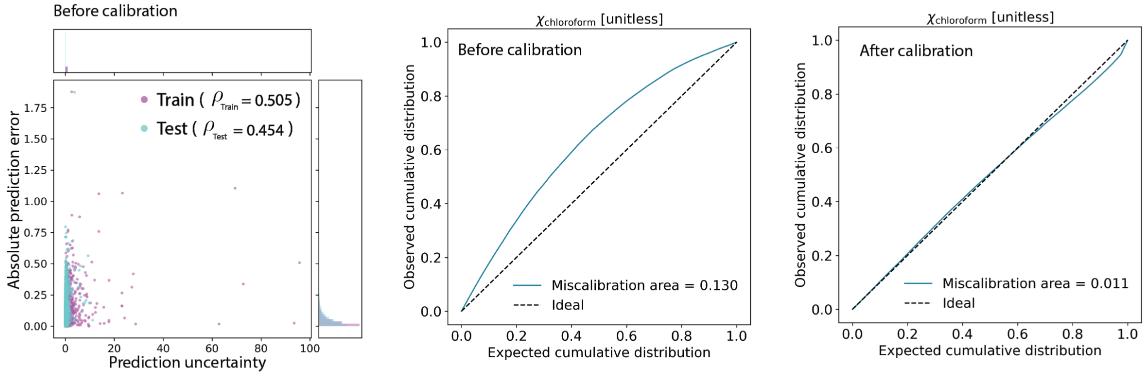


Figure S10: Prediction uncertainties from D-MPNN networks after the active learning campaign. Absolute prediction errors and prediction uncertainties were plotted for 19 monomer-level properties with their rank correlations. The quantile-based calibration curves were displayed before and after the uncertainty calibration.

11. PCA Analysis

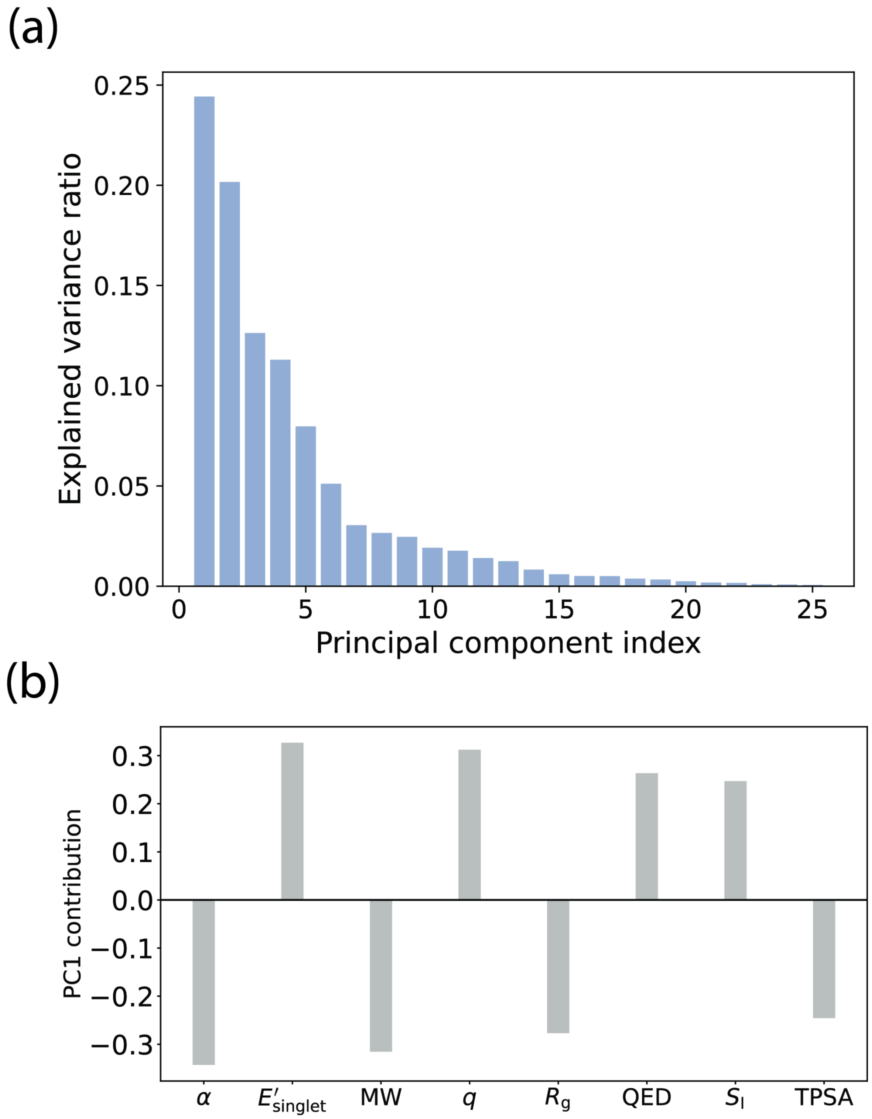


Figure S11: (a) Explained variance for the PCA analysis and (b) Top eight monomer-level properties contributing to the PC1 vector.

12. Molecular Size Effect

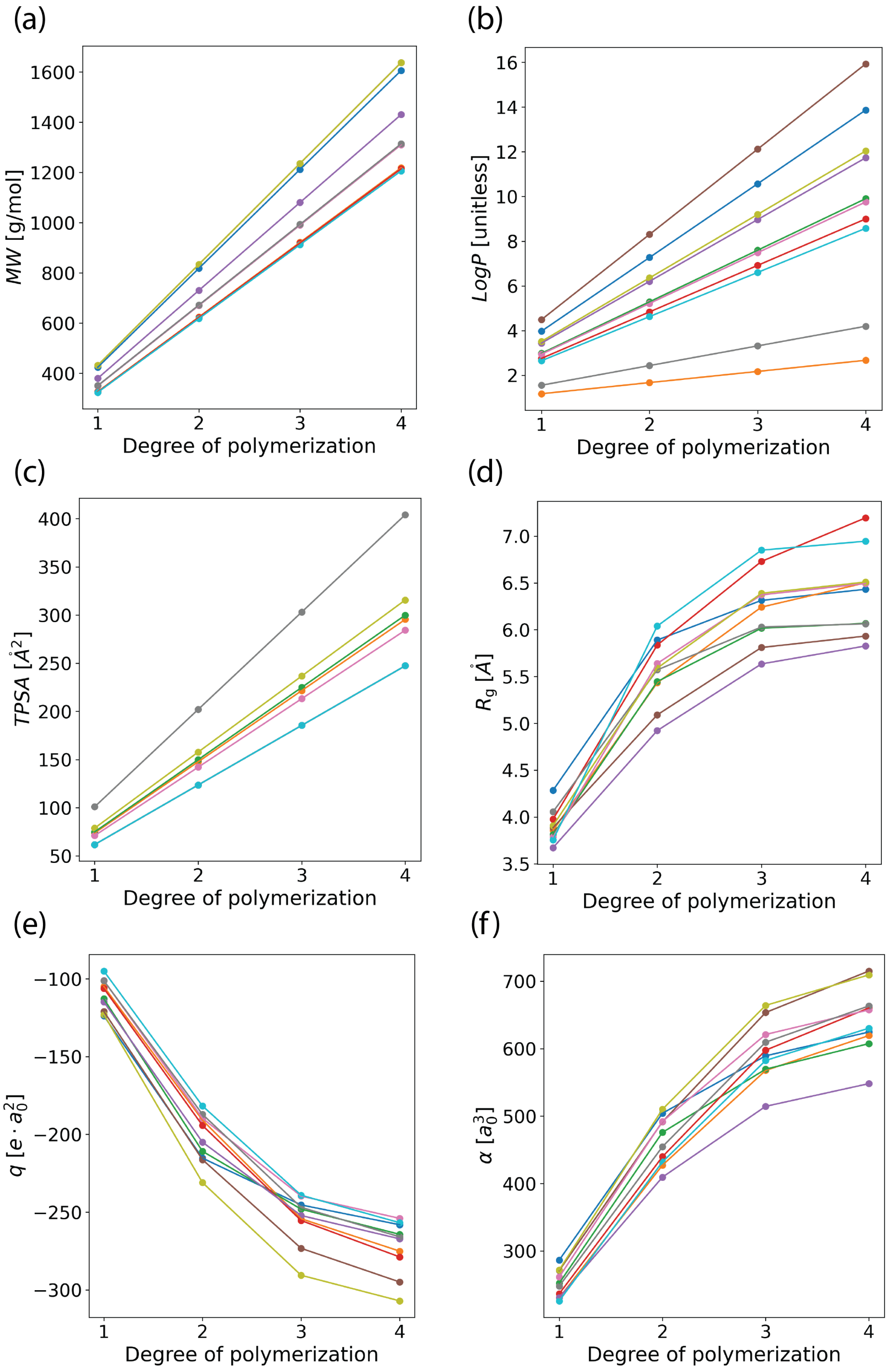


Figure S12: Example extrapolations of 10 OMG CRUs to higher degrees of polymerization for (a) molecular weight (MW), (b) octanol-water partition coefficient (LogP), (c) polar surface area (TPSA), (d) radius of gyration (R_g), (e) isotropic quadrupole moment (q), and (f) isotropic polarizability (α).

Figure S12 presents example extrapolations of 10 OMG CRUs to higher degrees of polymerization. For MW, LogP, and TPSA, we calculated properties using RDKit. For R_g, q, and α, the trained ML models from the active learning campaign were utilized to predict properties across varying degrees of polymerization. We extrapolated OMG CRUs with fewer than 25 heavy atoms to ensure that extended OMG CRUs remained within the training regime in terms of the number of heavy atoms.

We also provide normalized properties to approximately compensate for the molecular size effect. For the cheminformatics-derived properties (i.e., MW, LogP, and TPSA), we report the increase per degree of polymerization normalized by the number of heavy atoms in CRUs as these properties exhibit a linear increase with the degree of polymerization. For the DFT-derived properties (i.e., R_g, q, and α), we normalize the value for a monomer (n=1) by the number of heavy atoms in methyl-terminated monomers to approximately compensate for the molecular size effect.

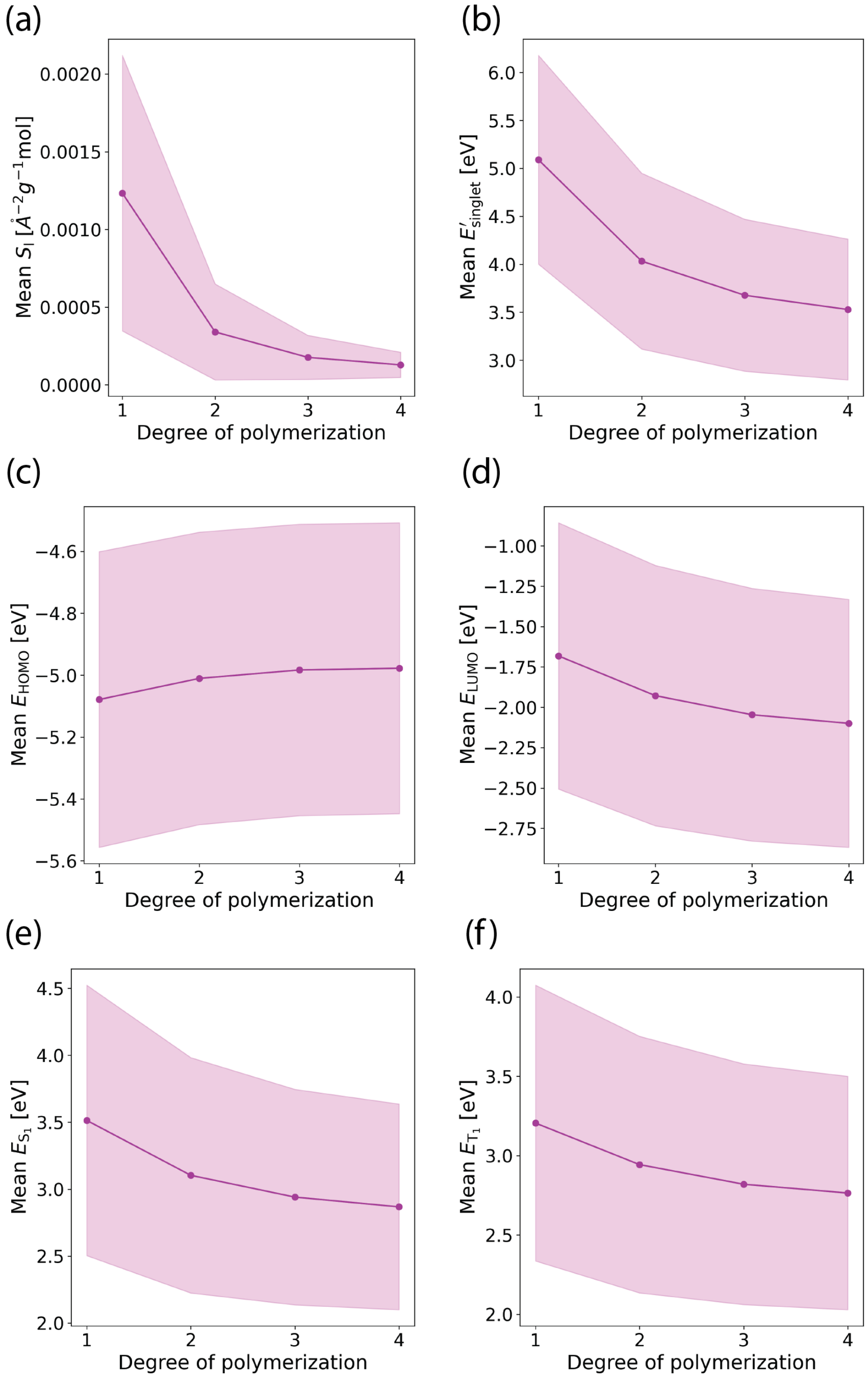
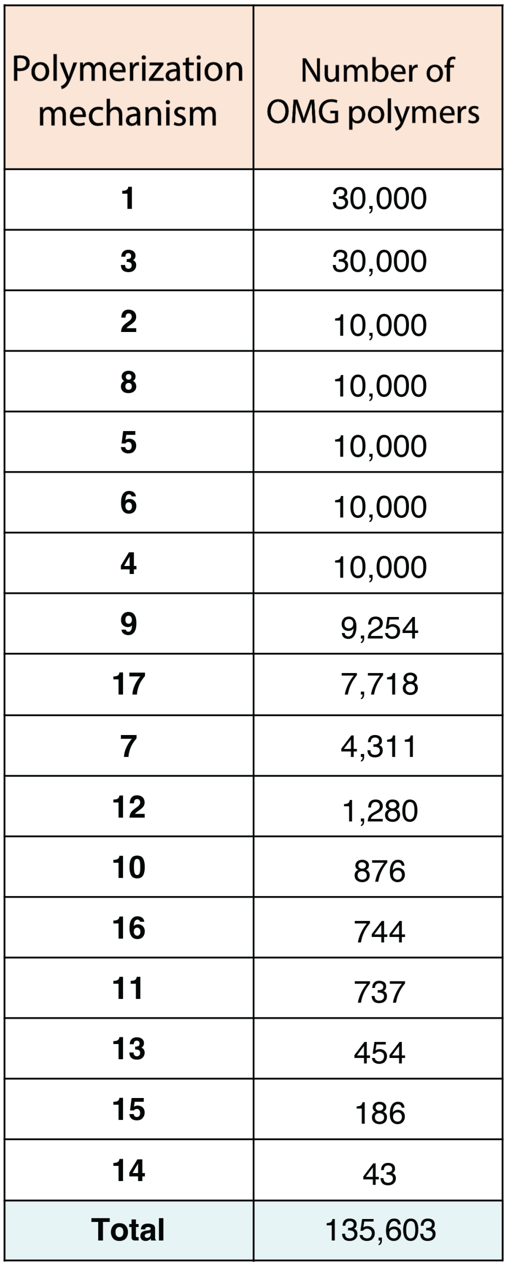


Figure S13: Extrapolations of 68,375 OMG CRUs to higher degrees of polymerization for (a) inertial shape factor (S_{I}), (b) singlet excitation energy with the largest oscillator strength ($E^{'}_{\mathrm{singlet}}$), (c) HOMO energy (E_{HOMO}), (d) LUMO energy (E_{LUMO}), (e) energy of the lowest singlet excited state (E_{S_1}), and (f) energy of the lowest triplet excited state (E_{T_1}).

Figure S13 presents extrapolations of 68,375 OMG CRUs to higher degrees of polymerization. The trained ML models from the active learning campaign were utilized to predict properties across varying degrees of polymerization. We extrapolated OMG CRUs with fewer than 25 heavy atoms to ensure that extended OMG CRUs remained within the training regime in terms of the number of heavy atoms. We did not normalize electronic and optical properties that are not significantly sensitive to varying degrees of polymerization. In addition, inertial shape factors were also not normalized as they exhibit a sharp decrease with increasing degrees of polymerization.

13. Compositions for the Randomly Sampled 135k OMG Polymers Based on Polymerization Mechanisms



Polymerization mechanism	Number of OMG polymers
1	30,000
3	30,000
2	10,000
8	10,000
5	10,000
6	10,000
4	10,000
9	9,254
17	7,718
7	4,311
12	1,280
10	876
16	744
11	737
13	454
15	186
14	43
Total	135,603

Figure S14: Compositions for the randomly sampled 135k OMG polymers based on polymerization mechanisms. The polymerization mechanism indices correspond to Figure 2 in the previous work.[2]

14. Correlations with Polymer Properties

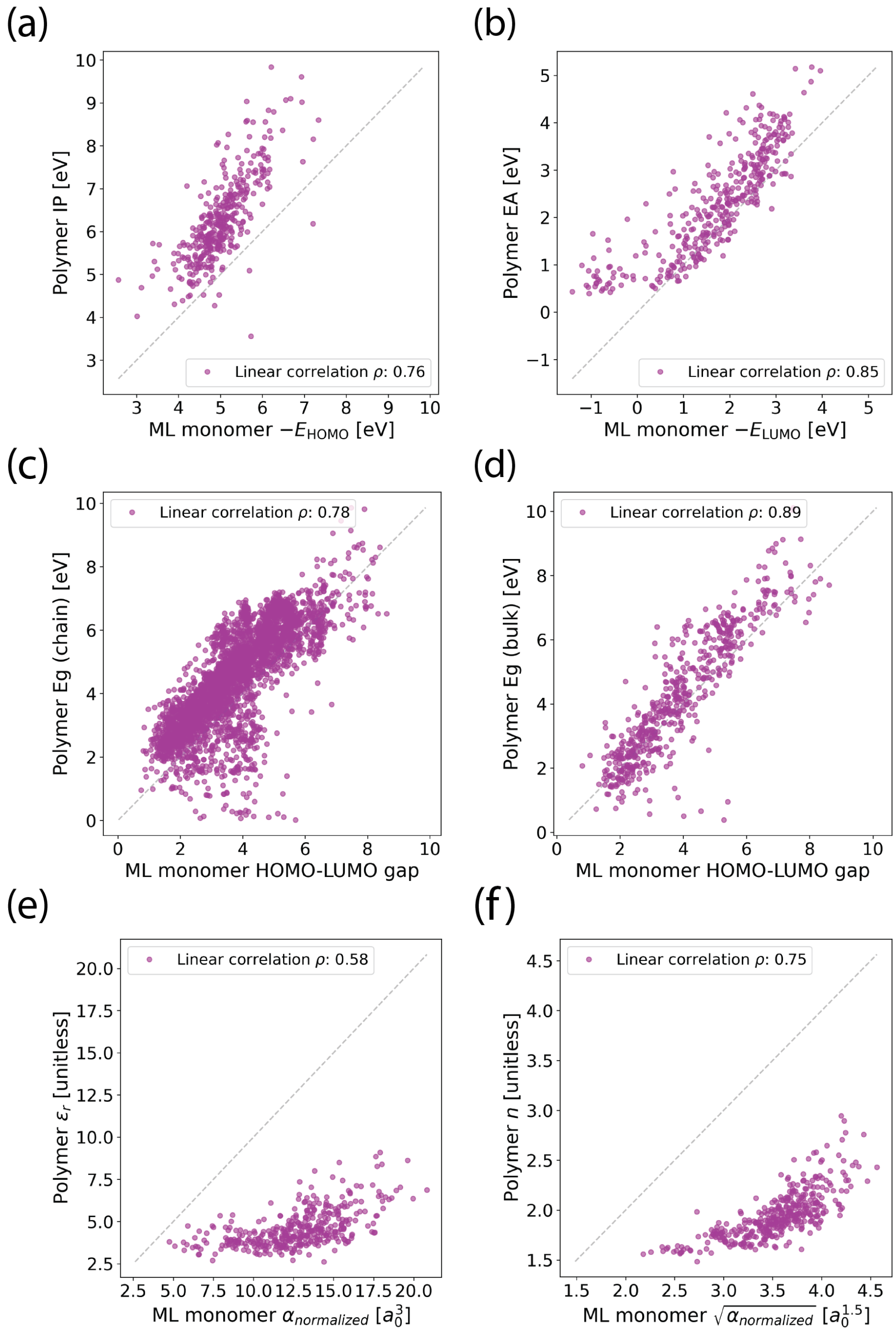


Figure S15: Correlations between ML-based monomer properties and polymer properties of (a) ionization potential (IP), (b) electron affinity (EA), (c) band gap of a polymer chain, (d) band gap of a polymer bulk, (e) dielectric constant (ϵ_r), and (f) refractive index (n). The polymer properties are from DFT calculations.[52]

Figure S15 demonstrates the correlations between ML-based monomer property predictions and polymer properties from DFT calculations.[52] The trained ML models after the active learning campaign were utilized to predict monomer properties. For dielectric constant (ϵ_r) and refractive index (n), the normalized polarizability by the number of heavy atoms in methyl-terminated monomers was used. The following equations from electromagnetism were used to estimate polymer properties from monomer properties.[53]

$$\begin{aligned}\vec{P} &= \epsilon_0 \kappa \vec{E} \\ \epsilon_r &= (1+\kappa) \\ &\approx \alpha_{normalized}\end{aligned}$$

$$\begin{aligned}n &= \frac{\sqrt{\mu\epsilon}}{\sqrt{\mu_0\epsilon_0}} \\ &\approx \sqrt{\epsilon_r} \\ &\approx \sqrt{\alpha_{normalized}}\end{aligned}$$

15. Pair Correlations Including Normalized Properties

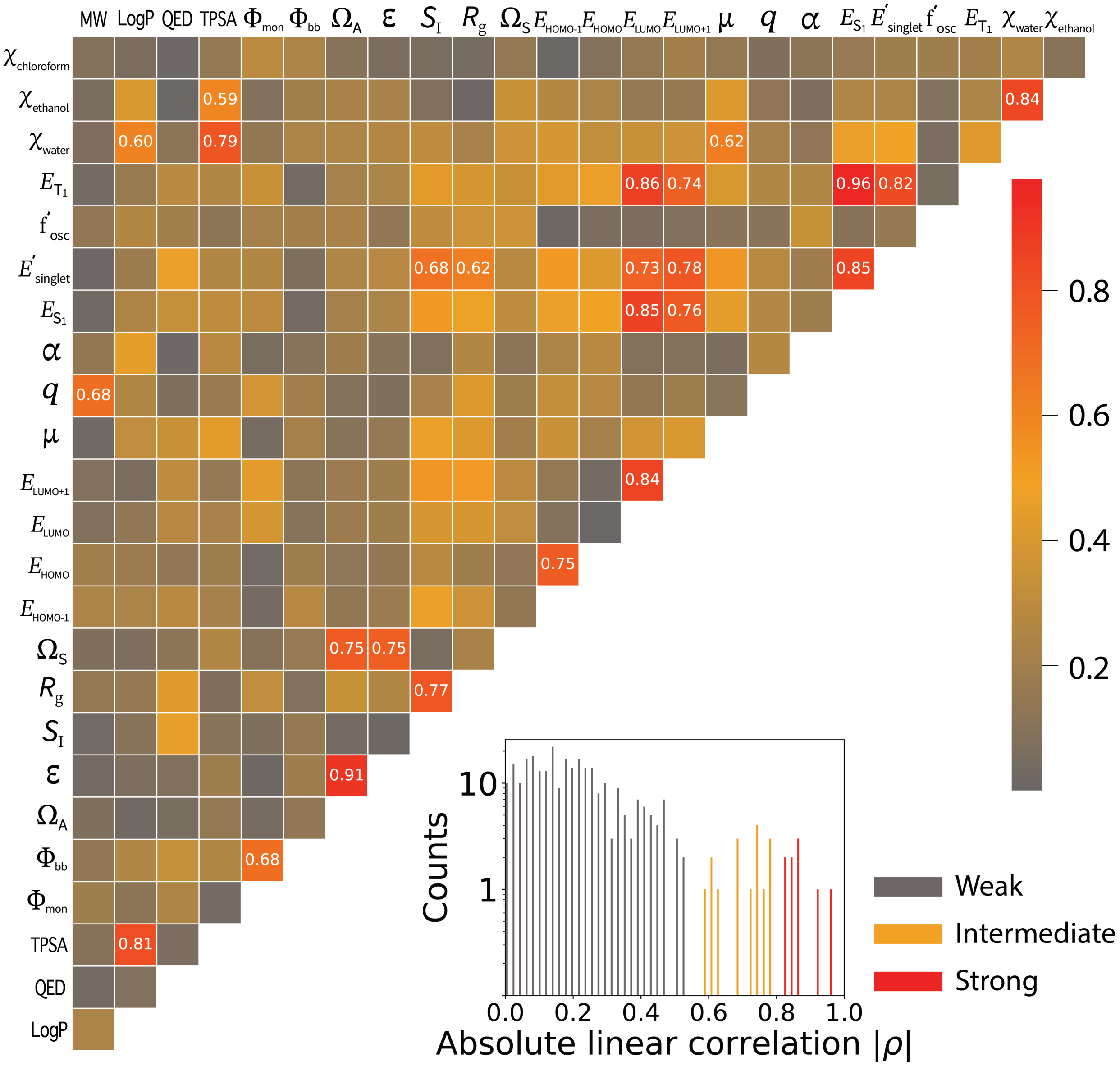


Figure S16: Property pair correlations between 25 monomer-level properties including normalized molecular weight (MW), octanol-water partition coefficient (LogP), polar surface area (TPSA), radius of gyration (R_g), isotropic quadrupole moment (q), and isotropic polarizability (α). These 6 properties were normalized to approximately compensate for the molecular size effect as described in **Figure S12**. The histogram shows the distributions of absolute linear correlation coefficients ($|\rho|$) between monomer-level property pairs. The three regimes are defined based on $|\rho|$: a weak regime ($|\rho| < 0.57$), an intermediate regime ($0.57 \leq |\rho| < 0.80$), and a strong regime ($|\rho| \geq 0.80$) as in the main text. After normalization, the monomer-level properties exhibit weaker correlations without the molecular size effect (intermediate correlations: 16 pairs / strong correlations: 9 pairs).

16. High and Low of Φ_{mon}, E'_{singlet}, and χ_{water}

The four different regimes in **Figure 5** were defined based on Φ_{mon} and E'_{singlet}. Especially, we used the mean (μ_{Φ} and $\mu_{E'}$)and standard deviation (σ_{Φ} and $\sigma_{E'}$) of Φ_{mon} and E'_{singlet} to decide the four regimes.

(1) Low Φ_{mon} and high E'_{singlet} (529 OMG CRUs)

$$\mu_{\Phi} - 1.2\sigma_{\Phi} \le \Phi_{\mathrm{monomer}} < \mu_{\Phi} - 0.8\sigma_{\Phi}$$

$$\mu_{E'} + 0.8\sigma_{E'} \le \quad E'_{\mathrm{singlet}} \quad < \mu_{E'} + 1.2\sigma_{E'}$$

(2) Low Φ_{mon} and low E'_{singlet} (3,015 OMG CRUs)

$$\mu_{\Phi} - 1.2\sigma_{\Phi} \le \Phi_{\mathrm{monomer}} < \mu_{\Phi} - 0.8\sigma_{\Phi}$$

$$\mu_{E'} - 1.2\sigma_{E'} \le \quad E'_{\mathrm{singlet}} \quad < \mu_{E'} - 0.8\sigma_{E'}$$

(3) High Φ_{mon} and high E'_{singlet} (443 OMG CRUs)

$$\mu_{\Phi} + 0.8\sigma_{\Phi} \le \Phi_{\mathrm{monomer}} < \mu_{\Phi} + 1.2\sigma_{\Phi}$$

$$\mu_{E'} + 0.8\sigma_{E'} \le \quad E'_{\mathrm{singlet}} \quad < \mu_{E'} + 1.2\sigma_{E'}$$

(4) High Φ_{mon} and low E'_{singlet} (362 OMG CRUs)

$$\mu_{\Phi} + 0.8\sigma_{\Phi} \le \Phi_{\text{monomer}} < \mu_{\Phi} + 1.2\sigma_{\Phi}$$

$$\mu_{E'} - 1.2\sigma_{E'} \le \; E^{'}_{\text{singlet}} \; < \mu_{E'} - 0.8\sigma_{E'}$$

A low χ_{water} and a high χ_{water} in **Figure 5** were determined based on the mean and standard deviation of χ_{water} for the sampled 135k OMG CRUs in **Figure S14**. A low χ_{water} is around the mean - 1.5 times of the standard deviation, and a high χ_{water} is around the mean + 1.5 times of the standard deviation.

17. The Number of Molecule Sets Sharing Properties

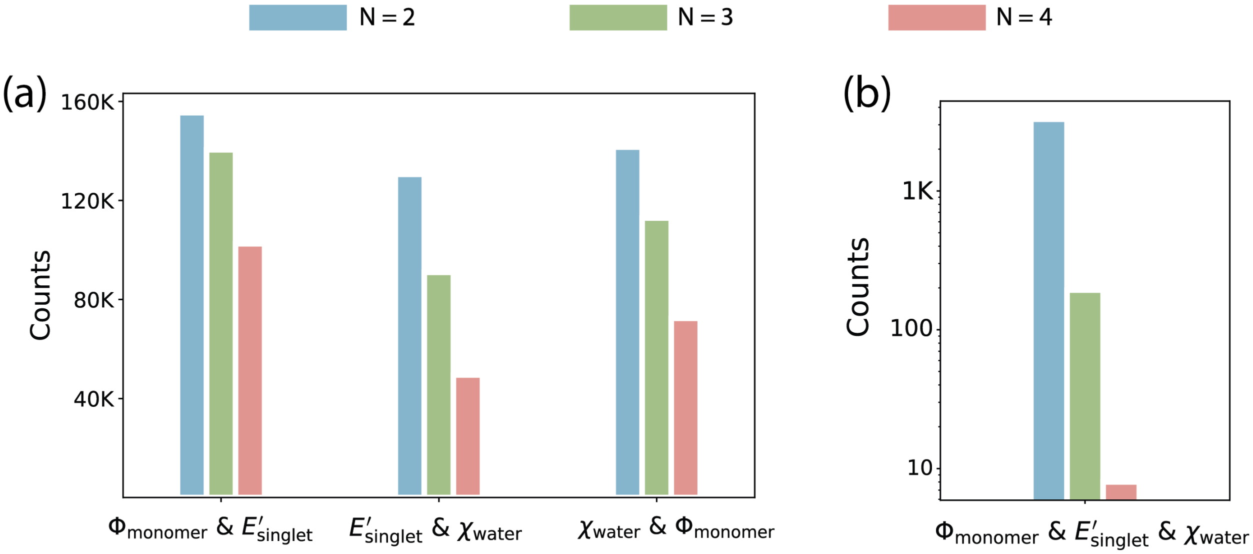


Figure S17: The number of molecule sets sharing (a) two and (b) three properties among Φ_{mon}, $E^{'}_{\text{singlet}}$, and χ_{water}. The different numbers (N) for a molecule set are considered. For example, $N = 3$ means that the number of molecule sets for Molecule 1, Molecule 2, Molecule 3 are counted sharing properties. The distance threshold for sharing properties was set to 0.01 in the standardized space with their mean and standard deviation.

18. Low χ_{water} and $\chi_{\mathrm{chloroform}}$

The OMG CRUs with low χ_{water} (or $\chi_{\mathrm{chloroform}}$) in **Figure 6** have χ_{water} (or $\chi_{\mathrm{chloroform}}$) less than the mean - one standard deviation of χ_{water} (or $\chi_{\mathrm{chloroform}}$).

References

(1) Kahovec, J.; Fox, R. B.; Hatada, K. (IUPAC Recommendations 2002). *Pure and Applied Chemistry* **2002**,

(2) Kim, S.; Schroeder, C. M.; Jackson, N. E. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. *ACS Polym. Au* **2023**, *3*, 318–330.

(3) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(4) Loschen, C.; Klamt, A. Prediction of Solubilities and Partition Coefficients in Polymers Using COSMO-RS. *Ind. Eng. Chem. Res.* **2014**, *53*, 11478–11487.

(5) 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.

(6) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J Cheminform* **2011**, *3*, 33.

(7) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(8) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics

and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(9) Stahn, M.; Ehlert, S.; Grimme, S. Extended Conductor-like Polarizable Continuum Solvation Model (CPCM-X) for Semiempirical Methods. *J. Phys. Chem. A* **2023**, *127*, 7036–7043.

(10) Zha, J.-W.; Zheng, M.-S.; Fan, B.-H.; Dang, Z.-M. Polymer-based dielectrics with high permittivity for electric energy storage: A review. *Nano Energy* **2021**, *89*, 106438.

(11) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(12) RDKit. `https://www.rdkit.org/`.

(13) Baumgärtner, A. Shapes of flexible vesicles at constant volume. *J. Chem. Phys.* **1993**, *98*, 7496–7501.

(14) Arteca, G. A. *Reviews in Computational Chemistry*; John Wiley & Sons, Ltd, 1996; pp 191–253.

(15) Todeschini, R.; Consonni, V. *Handbook of Chemoinformatics*; John Wiley & Sons, Ltd, 2003; pp 1004–1033.

(16) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. Molecular modeling of polymers. IV. Estimation of glass transition temperatures. *J. Polym. Sci. B Polym. Phys.* **1988**, *26*, 2007–2028.

(17) Fetters, L. J.; Lohse, D. J.; Richter, D.; Witten, T. A.; Zirkel, A. Connection between Polymer Molecular Weight, Density, Chain Dimensions, and Melt Viscoelastic Properties. *Macromolecules* **1994**, *27*, 4639–4647.

(18) Afzal, M. A. F.; Browning, A. R.; Goldberg, A.; Halls, M. D.; Gavartin, J. L.; Morisato, T.; Hughes, T. F.; Giesen, D. J.; Goose, J. E. High-throughput molecular dynamics simula-

tions and validation of thermophysical properties of polymers for various applications. *ACS Appl. Polym. Mater.* **2020**, *3*, 620–630.

(19) Bicerano, J. *Prediction of polymer properties*; cRc Press, 2002.

(20) Boyer, R. F. The relation of transition temperatures to chemical structure in high polymers. *Rubber Chemistry and Technology* **1963**, *36*, 1303–1421.

(21) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*, 224108.

(22) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

(23) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(24) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(25) Flory, P. J. Thermodynamics of High Polymer Solutions. *J. Chem. Phys.* **1941**, *9*, 660.

(26) Flory, P. J. Thermodynamics of High Polymer Solutions. *J. Chem. Phys.* **1942**, *10*, 51–61.

(27) Huggins, M. L. Solutions of Long Chain Compounds. *J. Chem. Phys.* **1941**, *9*, 440.

(28) Huggins, M. L. Theory of Solutions of High Polymers1. *J. Am. Chem. Soc.* **1942**, *64*, 1712–1719.

(29) Sanchez, I. C.; Lacombe, R. H. Statistical Thermodynamics of Polymer Solutions. *Macromolecules* **1978**, *11*, 1145–1156.

(30) Bae, Y. C.; Shim, J. J.; Soane, D. S.; Prausnitz, J. M. Representation of vapor–liquid and liquid–liquid equilibria for binary systems containing polymers: Applicability of an extended flory–huggins equation. *J. Appl. Polym. Sci.* **1993**, *47*, 1193–1206.

(31) Qian, C.; Mumby, S. J.; Eichinger, B. E. Phase diagrams of binary polymer solutions and blends. *Macromolecules* **1991**, *24*, 1655–1661.

(32) Knychała, P.; Timachova, K.; Banaszak, M.; Balsara, N. P. 50th Anniversary Perspective: Phase Behavior of Polymer Solutions and Blends. *Macromolecules* **2017**, *50*, 3051–3065.

(33) Lin, S.-T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Ind. Eng. Chem. Res.* **2002**, *41*, 899–913.

(34) Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C.-C.; Zwolak, M.; Seavey, K. C. Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. *Ind. Eng. Chem. Res.* **2006**, *45*, 4389–4415.

(35) Sanchez, I. C. Relationships between polymer interaction parameters. *Polymer* **1989**, *30*, 471–475.

(36) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(37) Bell, I. H.; Mickoleit, E.; Hsieh, C.-M.; Lin, S.-T.; Vrabec, J.; Breitkopf, C.; Jäger, A. A Benchmark Open-Source Implementation of COSMO-SAC. *J. Chem. Theory Comput.* **2020**, *16*, 2635–2646.

(38) Orwoll, R. A.; Arnold, P. A. In *Physical Properties of Polymers Handbook*; Mark, J. E., Ed.; Springer: New York, NY, 2007; pp 233–257.

(39) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(40) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput Mater* **2020**, *6*, 1–11.

(41) Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. Deep Evidential Regression. Advances in Neural Information Processing Systems. 2020; pp 14927–14937.

(42) 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.

(43) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **2014**, *1*, 140022.

(44) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **2021**, *7*, 1356–1367.

(45) Buzdalov, M.; Shalyto, A. A Provably Asymptotically Fast Version of the Generalized Jensen Algorithm for Non-dominated Sorting. Parallel Problem Solving from Nature – PPSN XIII. Cham, 2014; pp 528–537.

(46) Zhu, J.; Wang, H.; Hovy, E.; Ma, M. Confidence-based stopping criteria for active learning for data annotation. *ACM Trans. Speech Lang. Process.* **2010**, *6*, 1–24.

(47) Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. Proceedings of the 35th International Conference on Machine Learning. 2018; pp 2796–2804.

(48) Busk, J.; Bjørn Jørgensen, P.; Bhowmik, A.; Schmidt, M. N.; Winther, O.; Vegge, T. Calibrated

uncertainty for molecular property prediction using ensembles of message passing neural networks. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015012.

(49) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025006.

(50) Levi, D.; Gispan, L.; Giladi, N.; Fetaya, E. Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. *Sensors* **2022**, *22*, 5540.

(51) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906–915.

(52) Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer informatics with multi-task learning. *Patterns* **2021**, *2*, 100238.

(53) Griffiths, D. J. *Introduction to electrodynamics*; Cambridge University Press, 2023.