Supporting Information for

"CopDDB: a descriptor database for copolymers and its applications to the machine learning"

*Takayoshi Yoshimura,[a] Hiromoto Kato,[a] Shunto Oikawa,[b] Taichi Inagaki,[a] Shigehito Asano,[c] Tetsunori Sugawara,[c] Tomoyuki Miyao,[b] Takamitsu Matsubara,[b] Hiroharu Ajiro,[b] Mikiya Fujii,[b] Yu-ya Ohnishi,[d] and Miho Hatanaka[a,e]**

[a] Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8521, Japan.

[b] Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan.

[c] Fine Chemical Process Dept., JSR Corporation, 100 Kawajiri-cho, Yokkaichi, Mie, 510-8552, Japan.

[d] Materials Informatics Initiative, RD technology and digital transformation center, JSR Corporation, 3-103-9 Tonomachi, Kawasaki-ku, Kawasaki, Kanagawa, 210-0821, Japan.

[e] Institute for Molecular Science, 38 NishigoNaka, Myodaiji, Okazaki, Aichi 444-8585 Japan.

*E-mail: hatanaka@chem.keio.ac.jp

Contents

1. List of Monomers in CopDDB.

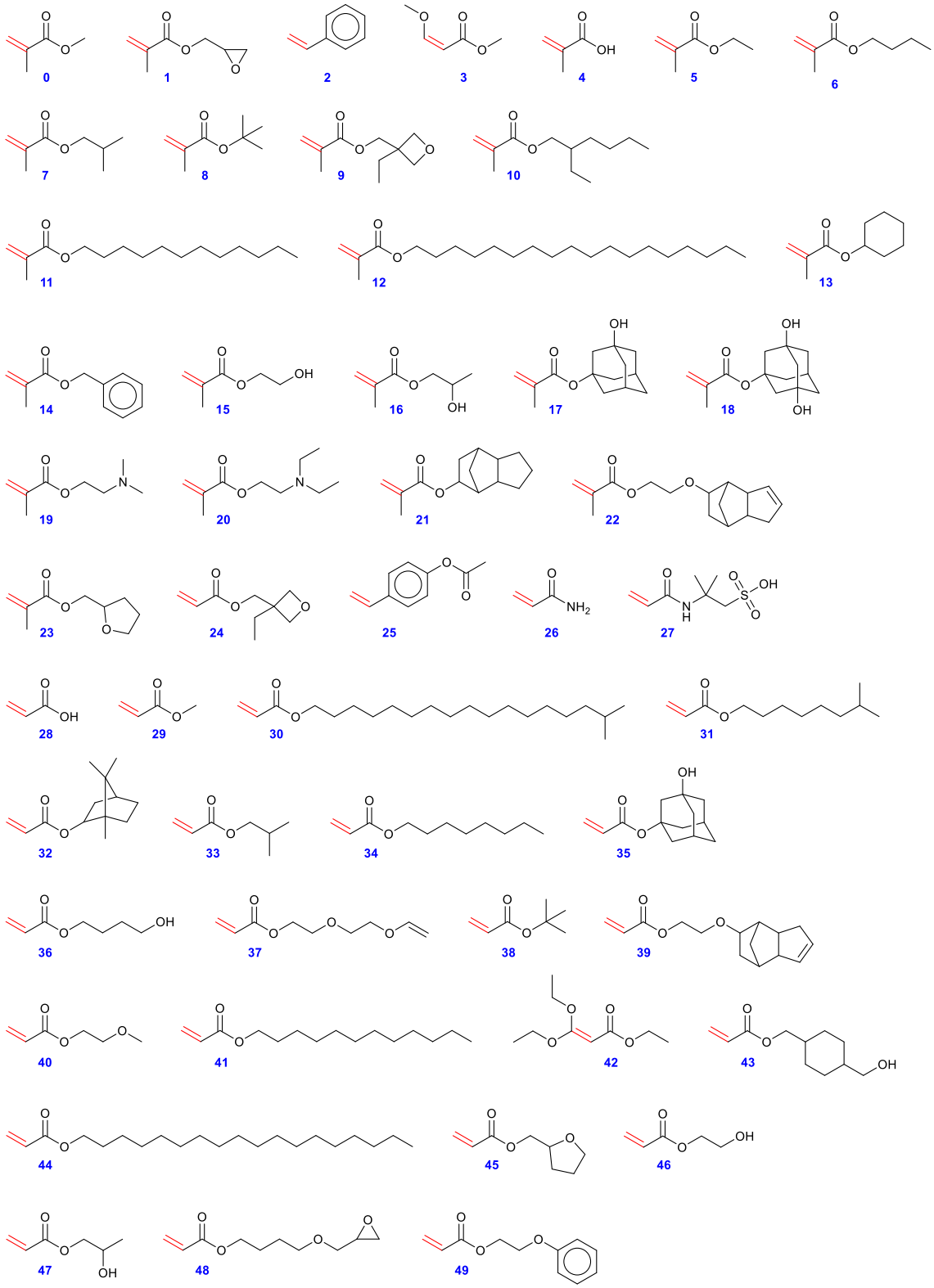

Figure S1. List of 50 monomers involved in CopDDB. The reactive carbons are shown in red. Numbers in blue correspond to the monomer ID in Table S1.

Table S1. The CAS registry numbers (RNs) and names of 50 monomers involved in the CopDDB.

Monomer ID [a)]	CAS RN	Name
0	80-62-6	Methyl methacrylate (MMA)
1	106-91-2	Glycidyl methacrylate (GMA)
2	100-42-5	Styrene (St)
3	5739-81-1	Methyl (Z)-3-methoxyacrylate
4	79-41-4	Methacrylic acid
5	97-63-2	Ethyl methacrylate
6	97-88-1	Butyl methacrylate
7	97-86-9	Isobutyl methacrylate
8	585-07-9	tert-Butyl methacrylate
9	37674-57-0	(3-Ethyloxetan-3-yl)methyl methacrylate
10	688-84-6	2-Ethylhexyl methacrylate
11	142-90-5	Dodecyl methacrylate
12	32360-05-7	Stearyl methacrylate
13	101-43-9	Cyclohexyl methacrylate (CHMA)
14	2495-37-6	Benzyl methacrylate
15	868-77-9	2-Hydroxyethyl methacrylate (HEMA)
16	923-26-2	2-Hydroxypropyl methacrylate
17	115372-36-6	3-Hydroxy-1-methacryloyloxyadamantane
18	115522-15-1	3,5-Dihydroxy-1-adamantyl methacrylate
19	2867-47-2	2-(Dimethylamino)ethyl Methacrylate
20	105-16-8	2-(Diethylamino)ethyl Methacrylate
21	34759-34-7	Dicyclopentanyl Methacrylate
22	68586-19-6	Dicyclopentenyloxyethyl Methacrylate
23	2455-24-5	Tetrahydrofurfuryl Methacrylate (THFMA)
24	41988-14-1	(3-ethyloxetan-3-yl)methyl acrylate
25	2628-16-2	4-Acetoxystyrene (PACS)
26	79-06-1	Acrylamide
27	15214-89-8	(1 1-Dimethyl-2-sulfoethyl)acrylamide
28	79-10-7	Acrylic Acid
29	96-33-3	Methyl Acrylate
30	93841-48-6	Isooctadecyl Acrylate
31	51952-49-9	Isononyl Acrylate
32	5888-33-5	Isobornyl Acrylate

33	106-63-8	Isobutyl Acrylate
34	2499-59-4	*n*-Octyl Acrylate
35	216581-76-9	3-Hydroxy-1-adamantyl Acrylate
36	2478-10-6	4-Hydroxybutyl Acrylate
37	86273-46-3	Vinyl ethoxyethyl acrylate
38	1663-39-4	*tert*-Butyl Acrylate
39	65983-31-5	Dicyclopentyloxyethyl acrylate
40	3121-61-7	2-Methoxyethyl Acrylate
41	2156-97-0	Dodecyl Acrylate
42	32002-24-7	Ethyl 3,3-Diethoxyacrylate
43	23117-36-4	1,4-Cyclohexanedimethanol monoacrylate
44	4813-57-4	Stearyl Acrylate
45	2399-48-6	Tetrahydrofurfuryl Acrylate
46	818-61-1	2-Hydroxyethyl Acrylate
47	999-61-1	2-Hydroxypropyl Acrylate
48	119692-59-0	4-(2,3-epoxypropoxy)butylacrylate
49	999-61-1	2-Hydroxypropyl Acrylate

a) Monomer ID corresponds to the number shown in Figure S1.

2. Activation barriers for C-C bond formation to different chains.

Table S2. Comparison of activation barriers of head-to-tail attack of monomers

Entry	Propagating end of radical	Monomer attacking to the radical	Activation barrier $\Delta G^{\ddagger}$ [kJ/mol] [a)]
1a	Me-MMA・	MMA	53.2
1b	Me-MMA-MMA・	MMA	52.5
1c	Me-St-MMA・	MMA	49.2
2a	Me-St・	MMA	60.5
2b	Me-MMA-St・	MMA	55.6
2c	Me-St-St・	MMA	54.2
3a	Me-MMA・	St	52.1
3b	Me-MMA-MMA・	St	52.5
3c	Me-St-MMA・	St	53.6
4a	Me-St・	St	63.9
4b	Me-MMA-St・	St	61.0
4c	Me-St-St・	St	57.6

a) All the calculations were at the B3LYP-D3/def2-SVP level of theory.

3. Details of descriptors in CopDDB

Table S3. Descriptors for radical (M_1*) and monomer (M_2) pairs available in CopDDB

Descriptor name	Notation in Fig. 1	Description [a)]
DE_tail	ΔE_{tail}	Reaction energy for the addition of a model initiator radical (Me*) to M_1 at the tail position
DE_head	ΔE_{head}	Reaction energy for the addition of a model initiator radical (Me*) to M_1 at the head position, which affords the radical M_1*
DE_precursor	$\Delta E_{precursor}$	Relative energy of the precursor from the dissociation limit (M_1* and M_2)
DE_TS	ΔE_{TS}	Relative energy of the TS of C-C bond formation from the dissociation limit (M_1* and M_2)
DE_product	$\Delta E_{product}$	Relative energy of the product from the dissociation limit (M_1* and M_2)
DE_barrier	$\Delta E_{barrier}$	Activation barrier for the C-C bond formation (*i.e.*, the energy difference between the precursor and the TS)
DE_reaction	$\Delta E_{reaction}$	Reaction energy for the C-C bond formation (*i.e.*, the energy difference between the precursor and the product)
E_Rad_SOMO		SOMO energy of M_1*
E_Rad_LUMO		LUMO energy of M_1*
E_Mon_HOMO		HOMO energy of M_2
E_Mon_LUMO		LUMO energy of M_2
DE_SHgap		Energy difference between SOMO of M_1* and HOMO of M_2
DE_SLgap		Energy difference between SOMO of M_1* and LUMO of M_2
VBur_R228_Rad		$\%V_{Bur}$ within 2.28 Å of the reactive carbon atom of M_1*
VBur_R350_Rad		$\%V_{Bur}$ within 3.50 Å of the reactive carbon atom of M_1*
VBur_R228_Mon		$\%V_{Bur}$ within 2.28 Å of the reactive carbon atom of M_2
VBur_R350_Mon		$\%V_{Bur}$ within 3.50 Å of the reactive carbon atom of M_2
Volume_Rad		Volume of M_1*
Volume_Mon		Volume of M_2
CCdist_TS	C2-C3	Reactive C-C bond distance at the TS structure
Dihedral_TS	<C1-C2-C3-C4	Dihedral angle around the reactive C-C at the TS structure
Sum_MW		Sum of molecular weights of M_1* and M_2
logP_Rad		Partition coefficient log*P* of M_1
logP_Mon		Partition coefficient log*P* of M_2

a) Units of energy, distance, angle, and volume are in Hartree, Å, degree, and cm^3/mol, respectively.

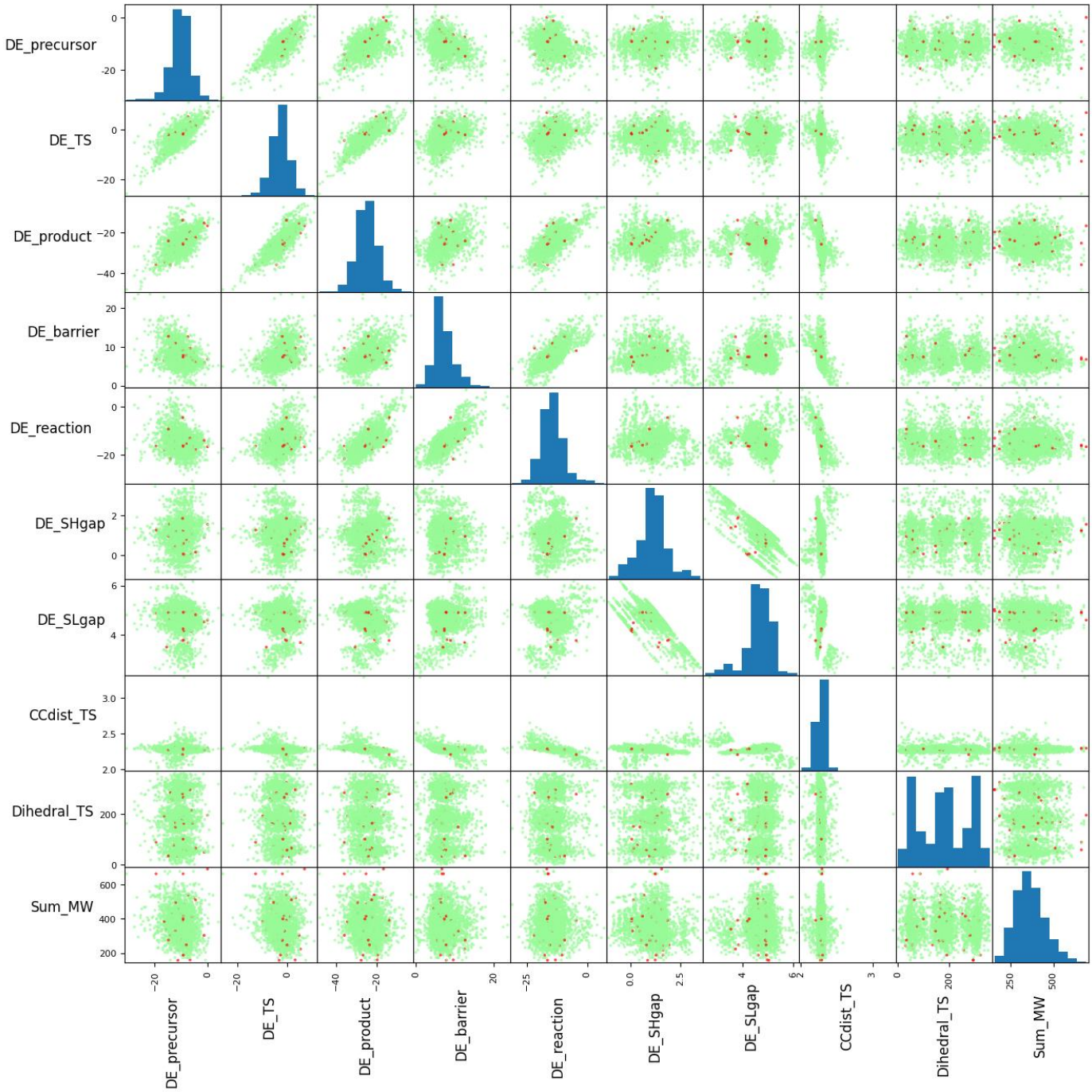

Figure S2. Histograms and scatter plots of the descriptors for 2500 radical-monomer pairs in CopDDB. The names of the descriptors are shown in Table S3. Data for homogeneous radical-monomer pairs (M_1*, M_1) and heterogeneous pairs (M_1*, M_2) pairs are shown in red and light green, respectively, in the scatter plots. Relative energies (DE_X; X = precursor, TS, product, barrier, and reaction) and orbital energy gaps (DE_SHgap and DE_SLgap) are in kcal/mol and eV, respectively.

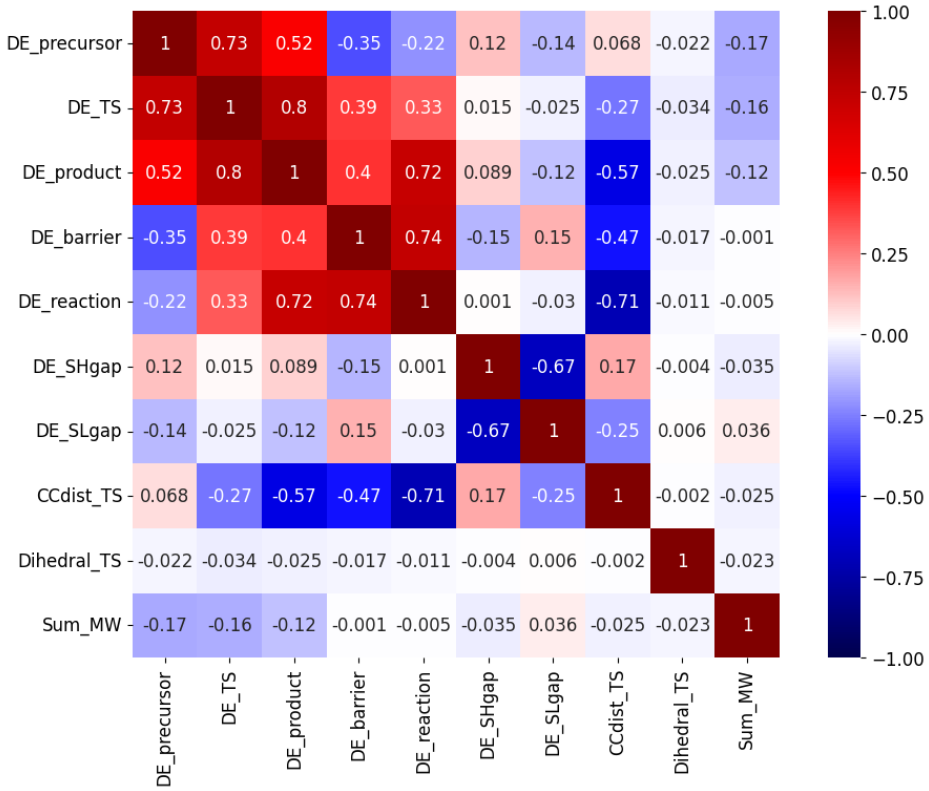

Figure S3. Heatmap of correlation matrix for the descriptors of 2500 radical-monomer pairs in CopDDB. The names of descriptors are shown in Table S3.

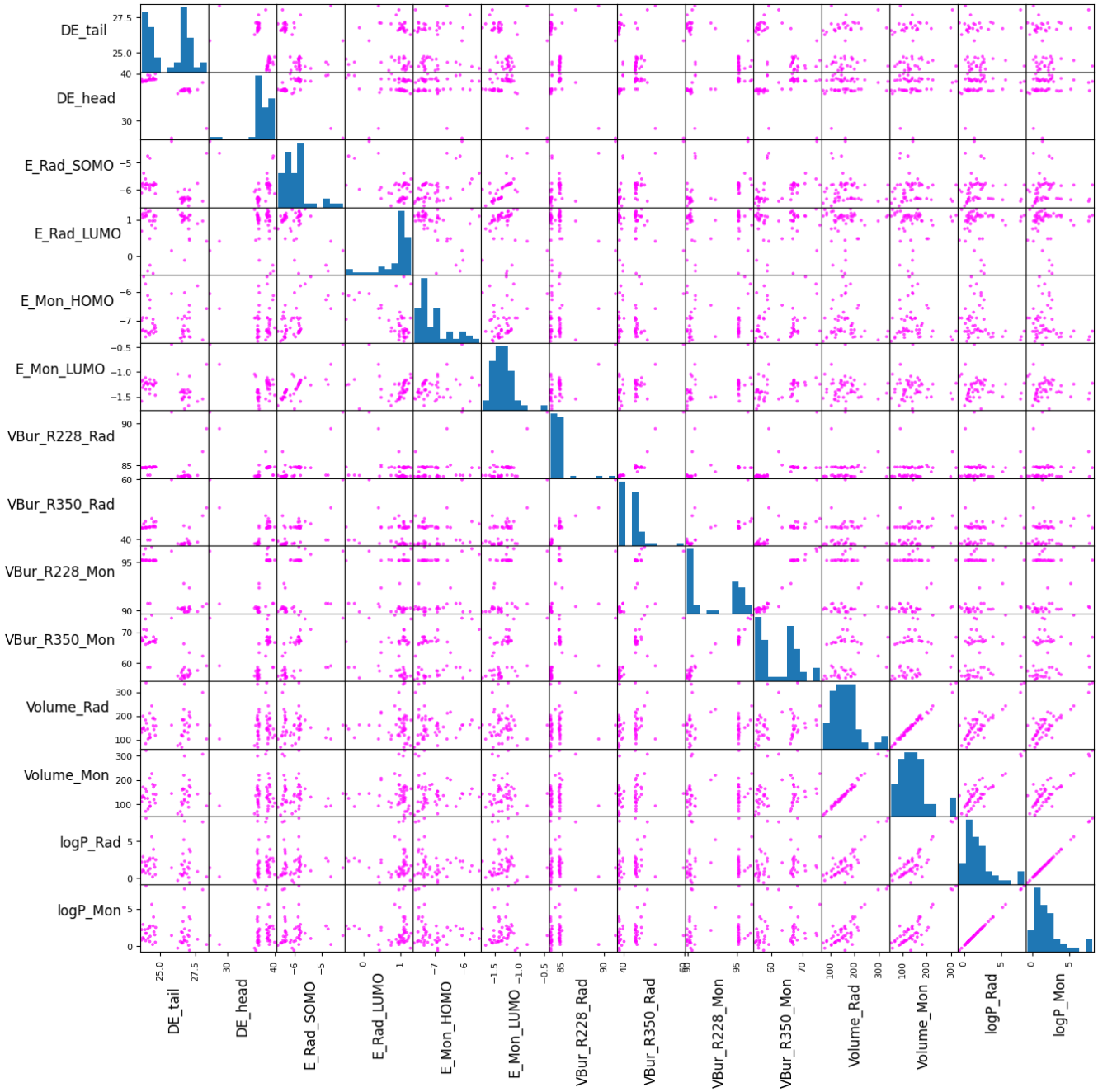

Figure S4. Histograms and scatter plots of the descriptors for 50 monomers M_1 and their corresponding radicals M_1* in CopDDB. The names of the descriptors are shown in Table S3. Relative energies (DE_tail and DE_head) and orbital energies (E_X_Y; X = Rad, Mon; Y = HOMO, SOMO, LUMO) are in kcal/mol and eV, respectively. Note that logP_Rad and logP_Mon are the same values.

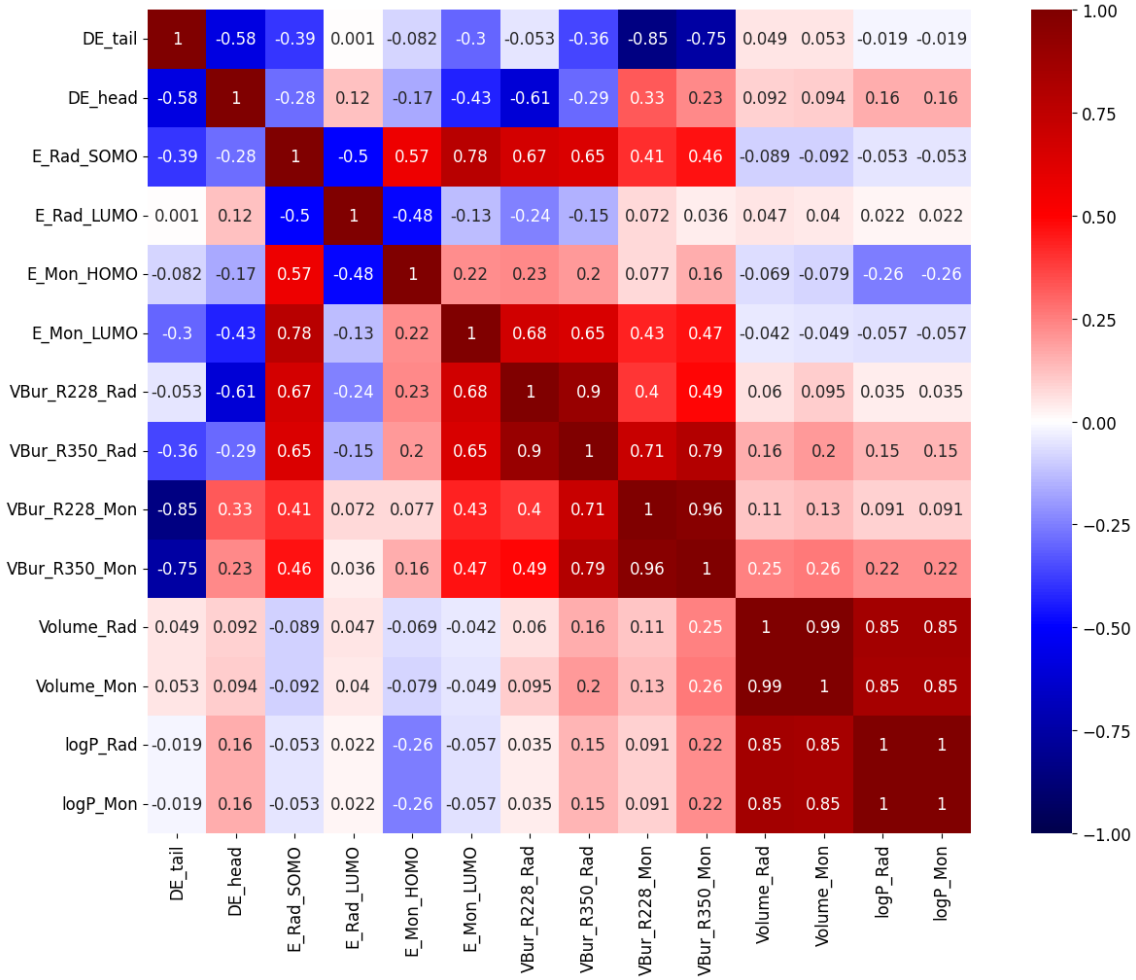

Figure S5. Heatmap of correlation matrix for the descriptors for 50 monomers M_1 and their corresponding radicals M_1* in CopDDB. The names of descriptors are shown in Table S3.

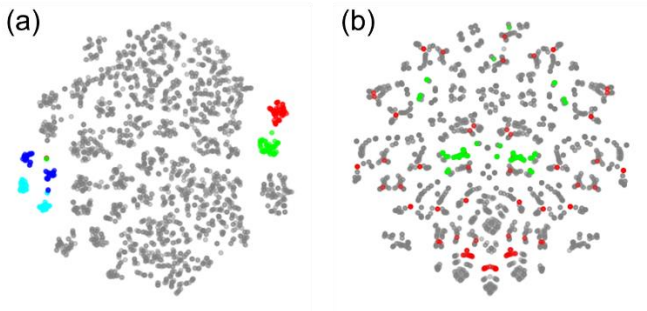

Figure S6. Distribution of data points for alkoxyaclylates (Monomer ID = 3 and 42; named M3 and M42, respectively) with 2500 radical-monomer pair data points within the chemical space spanned by the CopDDB descriptors (a) and the RDKit descriptors (b). The data points were projected into two dimensions as in Figure 3. In Panel (a), the data points for (M3*, M_1), (M42*, M_1), (M_1*, M3), (M_1*, M42), and the others are shown in red, green, blue, light blue, and gray, respectively. In Panel (b), the data points for (M3, M_1) and (M_1, M3) are in red, (M42, M_1) and (M_1, M42) are in blue, and the others are in gray. Here, M_1 and M_1* represent 50 monomers in the CopDDB and their corresponding radicals, respectively.

4. Details for the first case study.

Table S3. Detailed r_1 data from Polymer Handbook.

Entry	Monomer ID [a]		r_1 data in Polymer Handbook [b]	Average of r_1 [c]
	M_1*	M_2		
1	0	1	0.688, 0.71, 0.726, 0.8	0.731000
2	0	2	0.22, 0.314, 0.32, 0.409, 0.41, 0.41, 0.418, 0.42, 0.42, 0.422, 0.44, 0.45, 0.45, 0.45, 0.454, 0.46, 0.46, 0.46, 0.46, 0.46, 0.464, 0.47, 0.47, 0.478, 0.48, 0.48, 0.49, 0.49, 0.49, 0.5, 0.5, 0.5, 0.504, 0.54, 0.58, 0.59, 0.6, 0.611, 0.63, 0.64	0.470350
3	0	4	0.1, 0.1, 0.209, 0.27, 0.29, 0.31, 0.32, 0.36, 0.48, 0.55, 0.56, 0.63, 0.77, 0.78, 0.78, 0.87, 1.18, 1.25, 1.28, [1.81]	0.583632
4	0	5	0.81, 1.08, 1.16	1.016667
5	0	6	0.52, 1.27	0.895000
6	0	7	0.62, 0.89, 0.92	0.810000
7	0	8	0.96	0.960000
8	0	14	0.78, 0.808, 0.93	0.839333
9	0	15	[0.192], 0.75, 0.824	0.787000
10	0	16	0.402	0.402000
11	0	19	0.699	0.699000
12	0	20	0.843	0.843000
13	0	26	2.34, 2.53, 3.0, 3.0	2.717500
14	0	28	0.418	0.418000
15	0	29	2.15	2.150000
16	0	33	1.04, 4.1	2.570000
17	0	44	2.477	2.477000
18	1	0	0.501, 0.52, 0.934, 1.05	0.751250
19	1	2	0.16, 0.46, 0.514, 0.539, 0.54, 0.55, 0.73, 0.74	0.529125
20	1	4	1.2	1.200000
21	1	6	0.94	0.940000
22	1	33	1.24	1.240000
23	1	38	2.096	2.096000
24	2	0	0.275, 0.371, 0.38, 0.396, 0.41, 0.42, 0.432, 0.44, 0.44, 0.44, 0.45, 0.47, 0.472, 0.48, 0.48, 0.485, 0.49, 0.49, 0.497, 0.5, 0.5, 0.5, 0.52, 0.52, 0.52, 0.52, 0.52, 0.52, 0.52, 0.53, 0.54, 0.54, 0.55, 0.55, 0.56, 0.564, 0.57, 0.58, 0.585, 0.62	0.491175
25	2	1	0.11, 0.278, 0.31, 0.435, 0.44, 0.45, 0.47, 0.54	0.379125

26	2	4	0.041, 0.067, 0.124, 0.15, 0.17, 0.2, 0.2, 0.21, 0.22, 0.221, 0.38, 0.55, 0.627	0.243077
27	2	5	0.55, 0.55, 0.65, 0.67	0.605000
28	2	6	[0.001], 0.52, 0.54, 0.56, 0.63, 0.74	0.598000
29	2	7	0.47, 0.5, 0.509, 0.55, 0.56	0.517800
30	2	8	0.545	0.545000
31	2	11	0.528	0.528000
32	2	13	0.586	0.586000
33	2	14	0.435,0.45,0.45,0.463,0.48,0.8	0.513000
34	2	15	0.332,0.44,0.5,0.5,0.53,0.57,0.59	0.494571
35	2	25	0.835, 0.887	0.861000
36	2	26	0.65,1.05,1.13,1.17,1.21,1.49	1.116667
37	2	28	0.15,0.15,0.22,0.25,0.25,0.25,0.253,1.1	0.327875
38	2	29	0.192,0.4,0.65,0.68,0.722,0.75,0.75,0.82,0.871,0.9	0.673500
39	2	34	0.39	0.390000
40	2	44	0.44,0.44,0.777,0.79	0.611750
41	2	45	0.475	0.475000
42	4	0	0.25, 0.33, 0.33, 0.46, 0.48, 0.63, 0.63, 0.63, 0.68, 0.78, 0.99, 1.06, 1.06, 1.18, 1.26, 1.38, 1.55, 1.63, 1.84, 2.16	0.965500
43	4	1	0.98	0.980000
44	4	2	0.12, 0.28, 0.39, 0.44, 0.49, 0.55, 0.56, 0.6, 0.602, 0.631, 0.64, 0.66, 0.85	0.524077
45	4	5	0.57	0.570000
46	4	6	0.73, 0.8	0.765000
47	4	7	2.01	2.010000
48	4	16	0.99	0.990000
49	4	20	0.63	0.630000
50	4	26	[0.15], 1.63, 4.4	3.015000
51	5	0	0.86, 1.0, 1.08	0.980000
52	5	2	0.26, 0.29, 0.33, 0.36	0.310000
53	5	4	0.71	0.710000
54	5	16	0.245, 0.267	0.256000
55	6	0	1.2, 2.11	1.655000
56	6	1	0.85	0.850000
57	6	2	0.31, 0.47, 0.59, 0.64, 0.64, [2.52]	0.861667

58	6	4	1.15,1.26	1.205000
59	6	16	0.158, 0.158	0.158000
60	6	28	3.53	3.530000
61	7	0	0.488, 1.2, 1.88	1.189333
62	7	2	0.271, 0.4, 0.42, 0.58, 0.74	0.482200
63	7	4	0.47	0.470000
64	7	28	0.68	0.680000
65	8	0	1.35	1.350000
66	8	2	0.61	0.610000
67	10	16	0.083, 0.083	0.083000
68	11	2	0.3	0.300000
69	13	2	0.57	0.570000
70	14	0	1.05, 1.112, 1.38	1.180667
71	14	2	0.3, 0.42, 0.467, 0.5, 0.51, 0.658	0.475833
72	15	0	0.63, 0.81, 1.5	0.980000
73	15	2	0.53, 0.54, 0.59, 0.65, 0.856, 1.65, 1.65	0.923714
74	15	16	0.55	0.550000
75	15	26	0.98, 1.89	1.435000
76	15	27	0.86	0.860000
77	15	29	8.67	8.670000
78	16	0	1.055	1.055000
79	16	4	0.31	0.310000
80	16	5	1.844, 1.878	1.861000
81	16	6	2.35, 2.35	2.350000
82	16	10	4.56, 4.56	4.560000
83	16	15	1.82	1.820000
84	16	27	0.89	0.890000
85	16	29	7.335	7.335000
86	19	0	1.2	1.200000
87	20	0	1.44	1.440000
88	20	4	2.34	2.340000
89	25	2	1.218, 1.305	1.261500
90	26	0	0.53, 0.82, 0.9, [2.29]	0.750000
91	26	2	0.2, 0.33, 0.58, 0.59, 1.32, [8.97]	0.604000
92	26	4	0.56, 0.57, 0.58	0.570000

93	26	15	0.05, 0.14	0.095000
94	26	28	0.445, 0.58, 0.598, 1.06, 1.08, 1.346	0.851500
95	27	15	0.9	0.900000
96	27	16	1.03	1.030000
97	28	0	1.73	1.730000
98	28	2	0.05, 0.08, 0.136, 0.15, 0.25, 0.25, 0.35, 0.45	0.214500
99	28	6	0.24	0.240000
100	28	7	1.03	1.030000
101	28	26	0.288, 0.29, 0.341, 1.38, 1.644, [3.8]	0.788600
102	29	0	0.4	0.400000
103	29	2	0.07, 0.14, 0.148, 0.16, 0.168, 0.18, 0.18, 0.24, 0.3, 0.8	0.238600
104	29	15	0.001	0.001000
105	29	16	0.013	0.013000
106	29	46	[0.23], 0.94, 1.0	0.970000
107	33	0	0.135, 0.29	0.212500
108	33	1	0.282	0.282000
109	34	2	0.01	0.010000
110	38	1	0.463	0.463000
111	44	0	0.464	0.464000
112	44	2	0.18, 0.18, 0.31, 0.37	0.260000
113	45	2	0.489	0.489000
114	46	29	0.9, 0.94, 1.0	0.946667

a) Monomer ID corresponds to the number shown in Table S1 and Figure S1.

b) The numbers in square brackets are omitted as the outliers.

c) Each average is calculated without the outliers.

Table S4. List of 26 DFT-based descriptors used for predicting the reactivity ratio r_1 [a)]

Descriptor for X	Name of Descriptor [b)]
M_1	E_Rad_LUMO, E_Mon_HOMO, E_Mon_LUMO VBur_R228_Mon, Volume_Mon DE_head, DE_tail
M_2	E_Rad_SOMO, E_Rad_LUMO, E_Mon_HOMO, E_Mon_LUMO VBur_R228_Mon, Volume_Mon DE_head, DE_tail
(M_1*,M_1)	DE_Precursor, DE_barrier, DE_SHgap, DE_SLgap, CCdist_TS, Dihedral_TS
(M_1*,M_2)	DE_Precursor, DE_barrier, DE_SHgap, CCdist_TS, Dihedral_TS

a) Our procedure for preprocessing the descriptors was as follows: First, the DFT-based descriptors for M_1, M_2, (M_1*, M_1) and (M_1*, M_2), totaling 38, were prepared. Next, we calculated the correlation coefficients among the 38 descriptors. For descriptor pairs with correlation coefficients exceeding 0.9, one descriptor from each pair was removed. Consequently, the number of descriptors was reduced from 38 to 29. For descriptor pairs with correlation coefficients exceeding 0.8, we visualized the corresponding scatter plots and removed one descriptor from each pair if the correlation was deemed high. As a result, 26 descriptors remained.

b) The meaning of each descriptor is shown in Table S3.

Table S5. List of 24 RDKit descriptors used for predicting the reactivity ratio r_1 [a)]

Descriptor for X	Name of Descriptor
M_1	MaxEStateIndex, TPSA, MolWt, FpDensityMorgan2, BCUT2D_CHGHI, BCUT2D_LOGPHI, BCUT2D_LOGPLOW BCUT2D_MRHI, AvgIpc, BalabanJ, BertzCT, HallKierAlpha
M_2	MaxEStateIndex, TPSA, MolWt, FpDensityMorgan2, BCUT2D_CHGHI, BCUT2D_LOGPHI, BCUT2D_MRHI, AvgIpc, BalabanJ, BertzCT, HallKierAlpha, MolLogP

a) First, the RDKit descriptors for M_1 and M_2 without those represent substructures, totaling 64, were prepared. For descriptor pairs with correlation coefficients exceeding 0.9, one descriptor from each pair was removed, which reduced the descriptors from 64 to 30. For descriptor pairs with correlation coefficients exceeding 0.8, we visualized the corresponding scatter plots and removed one descriptor from each pair if the correlation was deemed high. As a result, 24 descriptors remained.

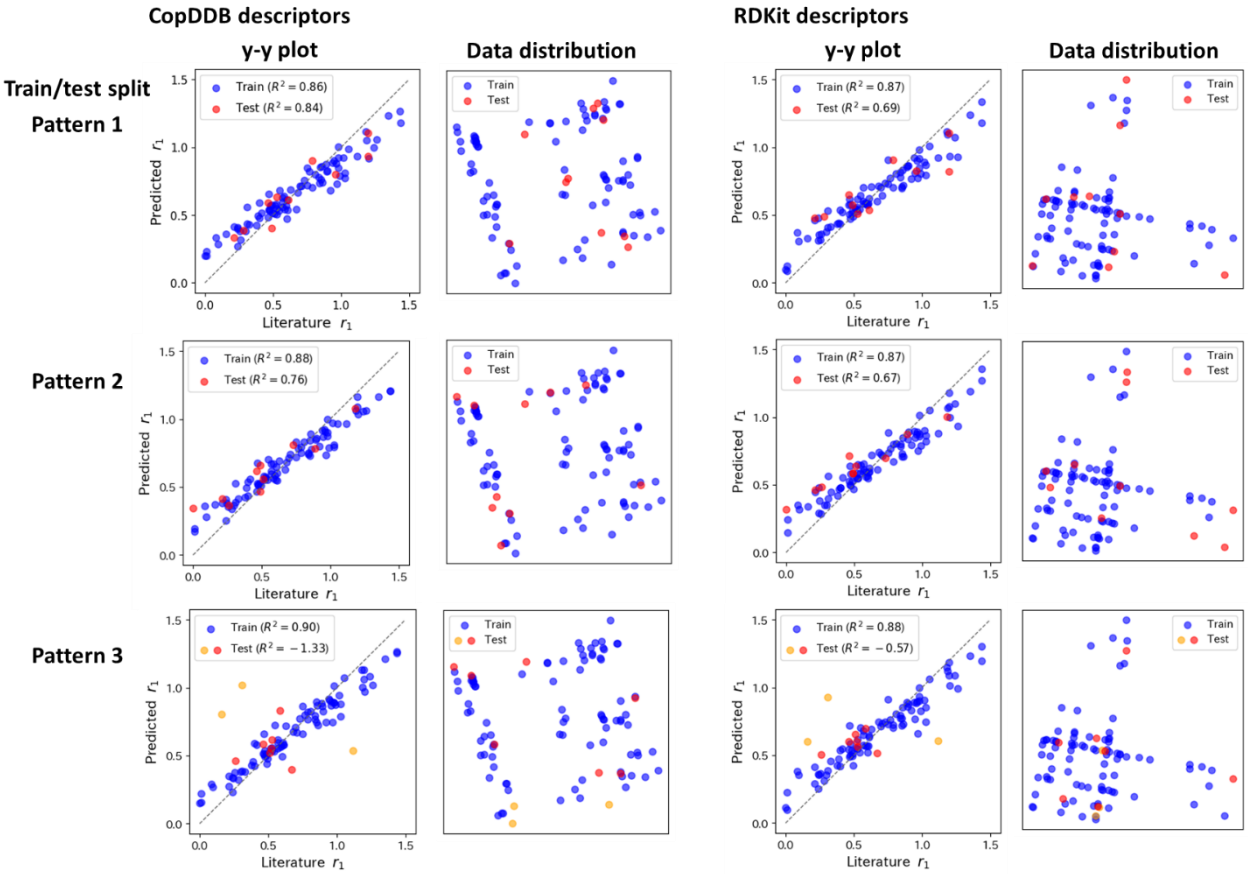

Figure S7. Y-y plots of r_1 for the Random Forest models using three different train/test data splits, along with their two-dimensional data distributions reduced by Principal Component Analysis (PCA). Blue and red dots represent the training and test data, respectively, while test data with poor prediction accuracy are highlighted in orange. The cumulative proportions of the first and second PCA components were 89.2 % and 84.3 % for the DFT-based and RDKit descriptors, respectively. Comparing the data distribution in the DFT-based and RDKit descriptor spaces, it can be observed that whether the test data are located in the interpolation or extrapolation region depends on the descriptor space. As shown in Patterns 1 and 2, prediction accuracy was high for test data located near the interpolation region in the DFT-based descriptor space. Conversely, as shown in the orange dots in Pattern 3, the prediction accuracy was low for test data outside the training data in the DFT-based descriptor space, regardless of whether the DFT-based or RDKit descriptors were used.

5. Details of dimensional compression of descriptors in the second case study.

The scheme of the variational autoencoder (VAE) to extract the three latent variables from the 22 radical-monomer descriptors is shown in Figure S8. The loss function was defined as follows,

$$loss = \sum_{i=1}^{I}\sum_{k=1}^{K}(\hat{y}_{ik} - y_{ik})^2 + -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2\right). \qquad \text{(eq. S1)}$$

Here, the first term is the sum of squared errors of y_{ik} and $\hat{y}_{ik}$, which are the input and output of k-th descriptor in i-th batch. The numbers of descriptors K and batches N are 22 and 20. (Note that among the 24 CopDDB descriptors, two parameters, $\Delta E_{\text{reaction}}$ and $\Delta E_{\text{product}}$, were not included.) The second derm is the KL divergence. μ_j, σ_j, and J are the mean and standard deviation used for j-th latent variable and the number of latent variables, respectively.

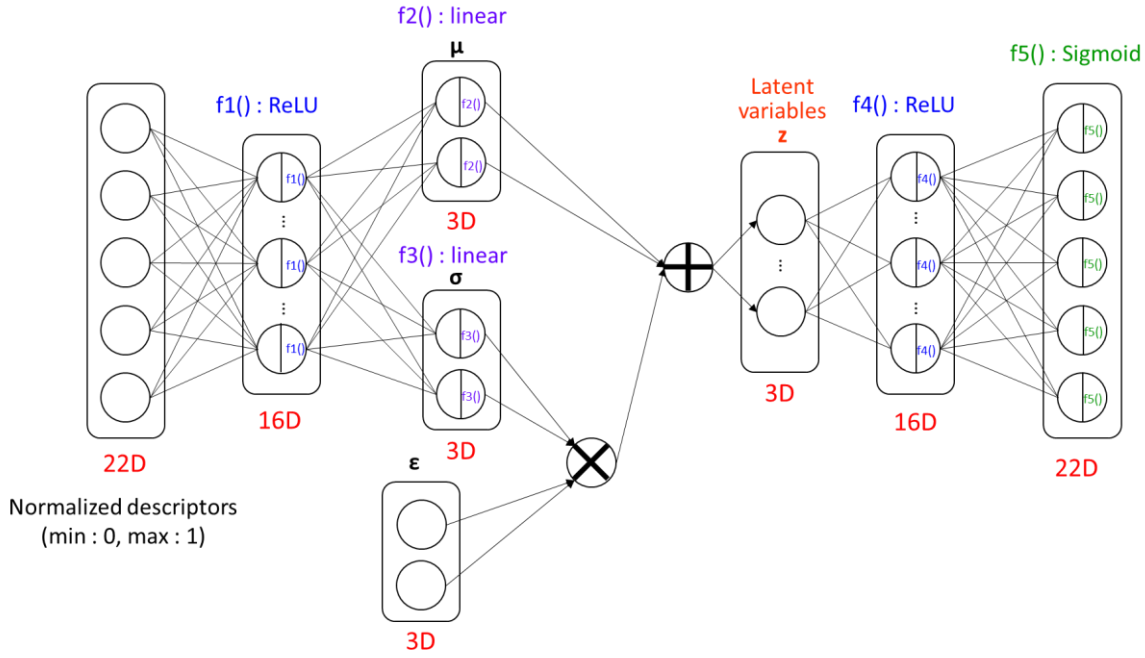

Figure S8. The scheme of the VAE used for the dimensional compression of 22 descriptors. 22D, 16D, and 3D in red represent 22, 16 and 3 dimensions. The three latent variables $\boldsymbol{z}$ are calculated as follows: $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\varepsilon}$ are the mean, standard deviation, and noise, respectively. fi() represents the i-th activation function.

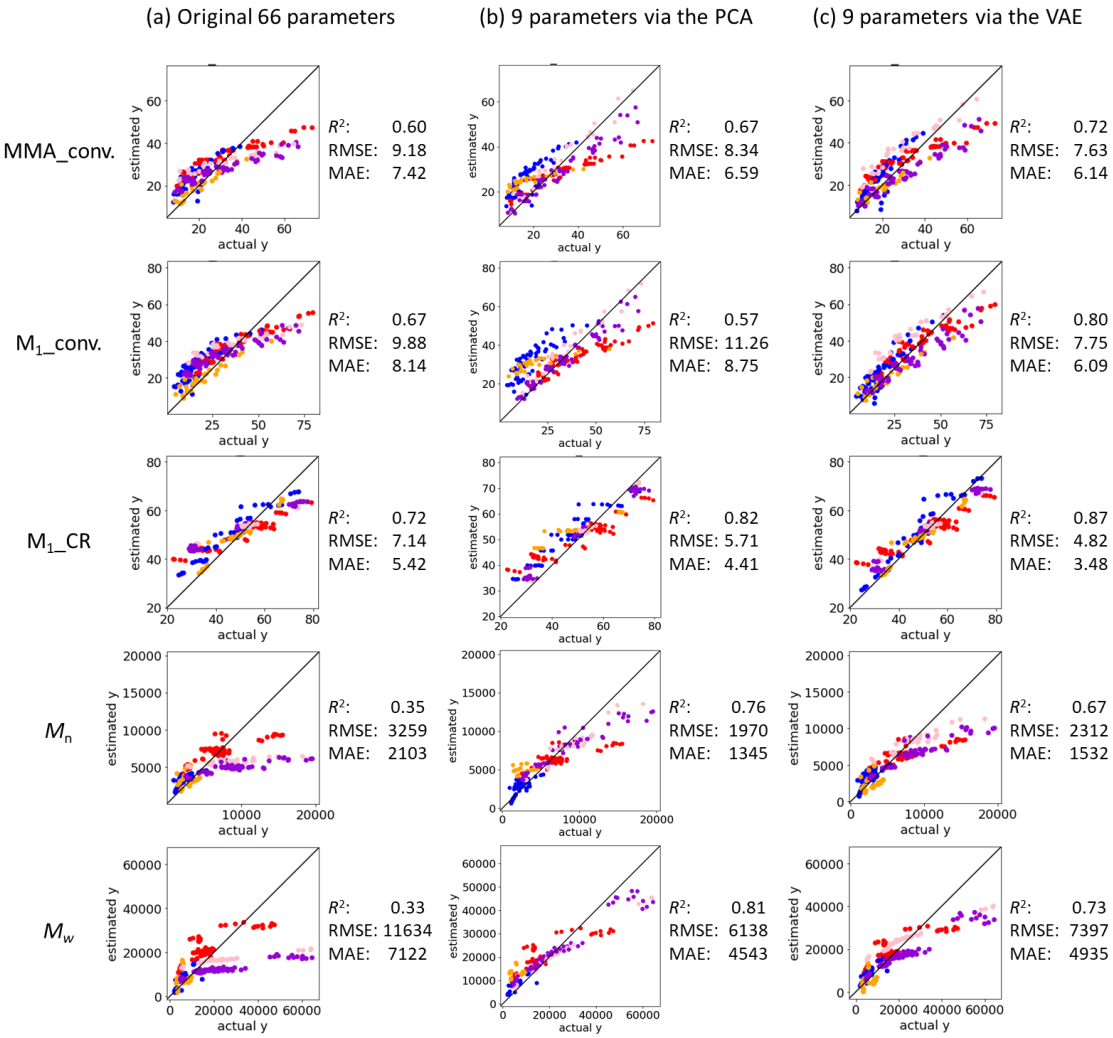

Figure S9. y-y plots of the GPR models for the five properties, including MMA_conv., M_1_conv., M_1_CR, M_n, and M_w constructed using different descriptor sets: 66 descriptors in the CopDDB (a), nine parameters compressed by the PCA (b) and the VAE (c). The data points of copolymers consisting of MMA with M_1, such as GMA, St, PACS, THFMA, CHMA, are shown in red, blue, orange, violet, and pink, respectively. The scores of leave-one-monomer (M_1)-out cross validation (LOOCV), including the coefficient of determination R^2, the root mean square error (RMSE), and the mean absolute error (MAE) are also listed.

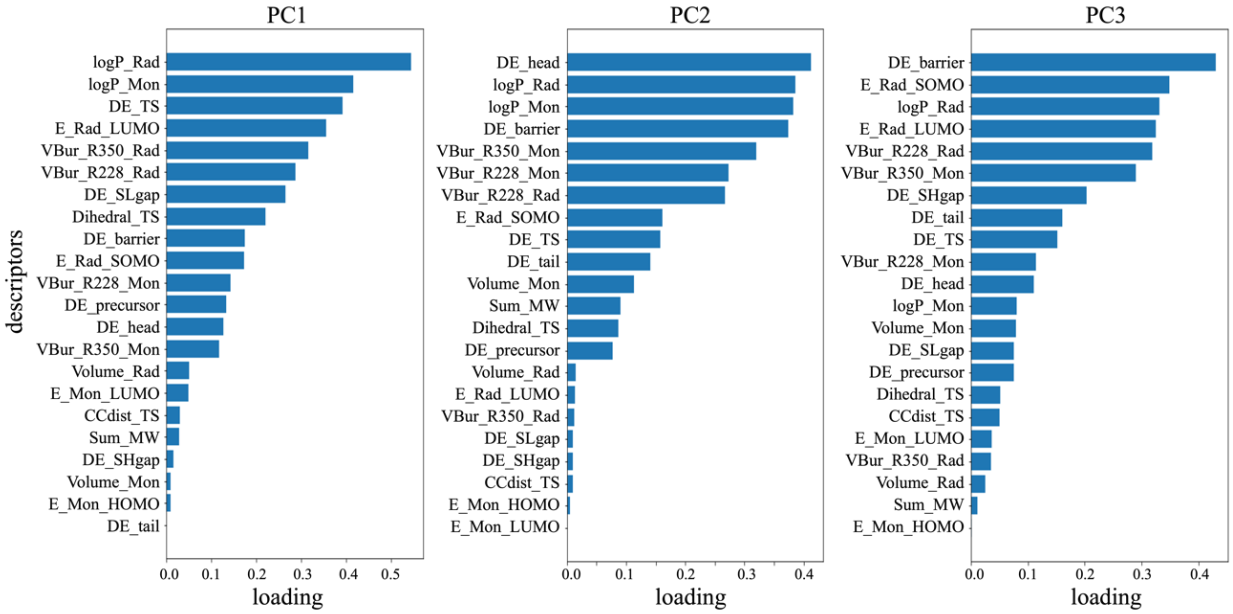

Figure S10. Loadings of the compressed DFT-based descriptors by PCA. The cumulative contribution of the first, second, and third principal components (noted as PC1, PC2, and PC3, respectively) was ca. 47.6%.

6. Details of parameters and results in the third case study.

Table S6. Design space for the optimization of process variables for the copolymerization of MMA and M_1 (=HEMA)

	Initiator concentration (mol%)	M_1 (=HEMA) proportion (mol%) [a)]	Reaction temperature (°C)	Solvent-to-monomer (SM) ratio	Reaction time (min.)
Minimum	1	20	50	1	2
Maximum	10	80	90	10	30

a) M_1 proportion represents molar ratio (in %) of M_1 to M_1 and MMA in the preparation.

Table S7. Four candidates proposed by one-shot Bayesian optimization shown in Figure 7.[a)]

	Initiator concentration (mol%)	M_1 (=HEMA) proportion (mol%) [b)]	Reaction temperature (°C)	Solvent-to-monomer (SM) ratio	Reaction time (min.)
Candidate-1	2.83	46.99	74.25	6.32	23.85
Candidate-2	2.77	47.24	72.42	2.54	9.73
Candidate-3	2.72	45.91	80.02	10.0	30.0
Candidate-4	2.54	47.45	77.38	2.33	30.0

a) The gamma prior in which the concentration and rate were set with 6.0 and 3.0 respectively is used. The detailed settings for the Bayesian optimization are the same as in Ref 51.

b) M_1 proportion represents molar ratio (in %) of M_1 to M_1 and MMA in the preparation.

Table S8. Experimental conditions and results for the validation of the four candidates in Figure 7.

Sample	Experimental conditions (process variable sets) [a]					Experimental results					
	Initiator concentration (mol%)	HEMA proportion (mol%) [b]	Reaction temperature (°C)	SM ratio	Reaction time (min)	HEMA-CR (mol%) [c]	HEMA conversion (%)	MMA conversion (%)	Mn	Mw	Mw/Mn
1-1 [d]	2.83	46.99	74.25	6.32	23.80	49.21	61.17	55.88	2456	3687	1.5012
1-2 [e]	2.83	46.99	74.25	6.32	11.95	49.13	44.43	40.72	2389	3618	1.5144
1-3 [f]	2.83	46.99	74.25	6.32	7.95	48.87	36.63	33.91	2437	3618	1.4846
2-1 [d]	2.77	47.24	72.42	2.54	9.73	49.91	50.66	45.52	4176	8101	1.9399
2-2 [e]	2.77	47.24	72.42	2.54	4.87	51.11	39.55	33.88	3774	8294	2.1977
2-3 [f]	2.77	47.24	72.42	2.54	3.25	51.73	32.68	27.30	2408	3645	1.5137
3-1 [d]	2.72	45.91	80.02	10.00	29.92	47.81	63.23	58.75	6688	14433	2.158
3-2 [e]	2.72	45.91	80.02	10.00	15.03	48.16	48.77	44.69	6522	14631	2.2433
3-3 [f]	2.72	45.91	80.02	10.00	10.01	47.69	41.45	38.71	3815	7543	1.9772
4-1 [d]	2.54	47.45	77.38	2.33	29.92	49.21	81.27	75.79	2783	4960	1.7822
4-2 [e]	2.54	47.45	77.38	2.33	15.03	49.80	67.66	61.61	2881	5023	1.7435
4-3 [f]	2.54	47.45	77.38	2.33	10.01	50.27	57.85	51.69	6324	13614	2.1528

a) Detailed experimental procedure is explained in Ref. 51.
b) HEMA proportion represents molar ratio (in mol%) of HEMA to HEMA and MMA in the preparation.
c) HEMA-CR represents molar ratio (in mol%) of HEMA to HEMA and MMA in the synthesized polymer.
d) Experimental condition of sample *n*-1 (*n* = 1, 2, 3, 4) corresponds to candidate-*n* shown in Table S7.
e) In sample *n*-2 (*n* = 1, 2, 3, 4), only the reaction time was changed to 1/2 of the experimental condition of candidate-*n*.
f) In sample *n*-3 (*n* = 1, 2, 3, 4), only the reaction time was changed to 1/3 of the experimental condition of candidate-*n*.