Supplementary Information for

Exception Search in Databases for Polymers with Practically Contradictory Properties of Heat Resistance and Transparency

Yasuyuki Nakamura,[a] Alice Gros,[b] Wenhao Zhang,[c] Keitaro Sodeyama,[b] and Masanobu Naito[a,c]

[a] *Data-driven Polymer Design Group, Research Center for Macromolecules & Biomaterials, National Institute for Materials Science (NIMS), Tsukuba, Ibaraki, Japan.*

[b] *Data-driven Materials Design Group, Center for Basic Research on Materials, and Energy Materials Design Group, Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Ibaraki, Japan.*

[c] *Department of Advanced Material Science, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan.*

Table of contents

General Information

Materials. Unless otherwise noted, chemicals obtained from commercial suppliers were used as received. Dimethyl 2,2'-azobis(isobutyrate) (MAIB) and 2,2'-azobis(isobutyronitrile) (AIBN) were recrystallized from methanol.

Instrumental analyses. ^{1}H (400 MHz) NMR spectra were measured with JEOL ECS-400 for a $CDCl_3$ solution of sample and are reported in ppm (δ) from internal Me_4Si. Gel permeation chromatography (GPC) was performed with Viscotek TDA 302 or JASCO EXTREMA system equipped with polystyrene mixed gel columns at 40 °C using RI and viscosity detectors, and THF as an eluent. The GPC was calibrated with PMMA standards. IR spectra were recorded on JASCO FT/IR-6100 for KBr pellet of samples. Differential scanning calorimetry (DSC) was measured on Shimadzu DSC-60Plus under nitrogen atmosphere. The sample was heated at 250 °C for 10 min to erase thermal history, then cooled down, and the temperature cycle was applied at a rate of 10 °C/min. Thermogravimetric (TG) analysis was performed on Hitachi TG/DTA 7300 under nitrogen/oxygen = 8/2 stream at a heat rate of 10 °C/min. In the temperature program for TG, the temperature was kept at 150 °C for 5 h before further temperature rising to remove the adsorbed water in the sample. UV-vis transmittance measurement was performed on JASCO V-600. Films of polymers for a transmittance measurement were prepared by casting the polymer aqueous solution on quartz glass plate and subsequent slow solvent evaporation (thickness of 9-19 μm). Total transmittance and haze measurements were performed on NDH 7000 SP II Haze meter. The transmittance of diffused light is measured by directly measuring transmittance of parallel light, and then subtracting the parallel light from the total light to obtain diffused light. The percentage of light that deviates from the incident beam by greater than 2.5 degrees on average is defined as haze. The samples are prepared by the same method for transmittance measurement.

Synthetic procedures.

Di-tert-butyl fumarate: To a solution of *tert*-butyl alcohol (16.8 g, 21.6 mL, 0.22 mol) in THF cooled at -78°C with stirring under nitrogen atmosphere, *n*-BuLi (1.6 M in hexane, 141 mL) was added dropwisely. After mixing the solution for 2 hours at the same temperature, fumaryl dichloride (16.8 g, 12 mL, 0.11 mol) was added slowly. Then the cold bath was removed, and the mixture was stirred for overnight. The resulting solution was washed with a saturated NH_4Cl aqueous solution and AcOEt, and the organic phase was further washed with a saturated $NaHCO_3$ aqueous solution and saturated NaCl aqueous solution. The organic phase was dried with $MgSO_4$, and the solvent was evaporated. The residue was purified by column chromatography on silica gel using hexane/EtOAc = 95/5 mixture as an eluent to give the target compound as white solid (21.2 g, 82%). ^{1}H NMR ($CDCl_3$) δ: 6.67 (s, 2H), 1.50 (s, 18H).

Polymerization of di-*tert*-butyl fumarate: A mixture of di-*tert*-butylk fumarate (5.0 g, 21.9 mmol) and dimethyl azobis(isobutyrate) (100 mg, 0.43 mmol) in a flask was stirred at 75 °C under nitrogen for 16 h. The resulting mixture was diluted with THF and poured into a vigorously stirred methanol. The precipitate was collected by centrifuge and decantation, and dried in vacuo to give poly(di-*tert*-butyl fumarate) in 82% yield. The molecular weight was analyzed with GPC (M_n = 11200, $Đ$ = 2.69). ^{1}H NMR ($CDCl_3$) δ: 3.15 (br, 2H), 1.49 (br, 18H). The tacticity of PFA-*t*Bu agreed well with those of previous reports analysed by ^{13}C NMR spectroscopy (m ~ 0.85).[1]

Table S1. Free-radical polymerization of di-*tert*-butylfumarate.

Run	Initiator[a]	(mol%)	Time (h)	Yield (%)	M_n (10^4)	*Đ*
1	MAIB	2	16	82	1.12	2.69
2	MAIB	1	16	60	4.18	3.70
3	MAIB	0.2	48	66	5.49	3.25
4	MAIB	0.04	72	55	13.8	1.71
5	AIBN	2	16	78	1.62	1.70
6[b]	MAIB	2	16	38	0.30	1.75

[a] MAIB: dimethyl 2,2'-azobis(isobutyrate), AIBN: 2,2'-azobis(isobutyronitrile). [b] Toluene was used as a solvent (100 vol% to the monomer).

The molecular weight of PFA-*t*Bu increased in line with a decrease in the amount of azo initiator. This is likely related to the slow termination reaction of the monomer, di-*tert*-butylfumarate. The use of the solvent decreased the rate of polymerisation and resulted in a low yield.

Pyrolysis of poly(di-tert-butyl fumarate): Poly(di-*tert*-butyl fumarate) powder (2.0 g) in an opened glass tube was placed in a vacuum open, and heated at 160 °C for 10 h. The resulting slight yellowish powder was dissolved in deionized water and poured into a vigorously stirred 1,4-dioxane. The precipitate was collected by centrifuge and decantation, and dried to give poly(fumaric acid) in 85% yield. An unidentified impurity(es) was observed in ^{1}H NMR analysis, however, the amount was expected to be not significant, and the product was used for further experiments. ^{1}H NMR (D_2O) δ: 3.22 (br, 2H).

Poly(dimetal fumarate), typical procedure for a sodium salt: A solution of poly(fumaric acid) in deionized water was added a aqueous NaOH, and stirred at rt for 24 hours. The resulting solution was added into methanol. The precipitate was collected by centrifuge, washed with water, and dried in vacuo to give the target polymer. Disodium product: ^{1}H NMR (D_2O) δ: 2.91 (br, 2H). Dilithium product: ^{1}H NMR (D_2O) δ: 2.87 (br, 2H). Dipotassium product: ^{1}H NMR (D_2O) δ: 3.01 (br, 2H).

Computational method

MD simulations. Full atomistic molecular dynamics simulations were performed with the large atomic molecular massively parallel simulator (LAMMPS) with GPU package.[2,3] Winmostar V11 software was used for the pre and post treatments of simulations. The atomic interactions are described by the DREIDING force field[4] in this study with atomic charges estimated from DFT calculation in B3LYP/6-31G(d,p) level of theory. The charge was calculated by Gaussian 16 rev C.01 suit.[5] Each polymer has 50 atoms in the main chain, which correspond to 25 repeating units [$(CR^1R^2)2$] of C1-polymers or ($CH_2CR^1R^2$) of C2-polymers. The simulation cell contains 15 or 20 chains for poly(fumaric acid) derivatives and vinyl polymers or heteroatom-substituted methylene polymers, respectively, which were randomly arranged to give an amorphous polymer with a density of 0.5 g/cm^3. The simulation cell was relaxed at 873 K for 100 ps in the NVT ensemble, subsequently quenched from 873 K to 773 K at a rate of 1 K/ps at 1 atm in the NPT ensemble, and then equilibrated at 773 K at 1 atm for 0.5 ns. A time step of 1 fs was used. Nosé-Hoover thermostat and Parrinello-Rahamn borostat were employed for temperature and pressure control.[6] Long-range electrostatic interactions were calculated by using particle-particle particle-mesh (PPPM) method with an accuracy of 10^{-5}.[7] For the calculation of glass transition temperature (T_g), the simulation system was quenched in NPT ensemble from 773 K at the rate of 1 K/ps followed by the equilibration in NPT ensemble for 200 ps under 1 atm at every 50 K step until 73 K. The T_g was obtained from the specific volume as a function of temperature with a bilinear fitting. The condition of the simulation for T_g calculation is discussed below in the separated section.

TD-DFT calculation. All computational calculations were performed using Gaussian 16 rev C.01 suite.[5] The geometry optimizations and frequency calculations were carried out at the B3LYP/6-31G(d,p) level of theory. The excited state energies were calculated by using the time-dependent DFT (TD-DFT) calculations at the B3LYP/aug-cc-pVDZ level of theory. As a model of PFA-Na, Me-$(CHCO_2Na)_6$-Me was applied for the calculation of absorption spectrum.

Machine-learning prediction

For the analysis and machine-learning study of T_g and T_{deg} extracted in PoLyInfo[8], the median values were used. A prediction model was build using SMILES-X[9,10], an autonomous pipeline for molecular compounds characterization based on a neural architecture with a data-specific Bayesian hyper-parameters optimization using simplified molecular input line entry system (SMILES) as the input. And the model was build using T_g and T_{deg} datasets, with a split of each by 90%, 5%, 5%, for training, validation and test. The median value was used as T_g and T_{deg} data. No augmentation of data was performed as it seems that there were enough datapoints. Bayesian optimization is performed at each fold to aim for the most performant architecture size. Training/validation/test steps were performed for 10 folds of 400 epochs at most, without augmentation of data. Parameters and inputs used to run SMILES-X are indicated in Figure S6, with the inputs name as required to run the main function of SMILES-X. The predictions scores as indicated by the SMILES-X are summarized in Table S2. Using the resulting model with a fair root-mean-square-error (RMSE) value of 36.3°C +/- 3.1, prediction was then performed using the SMILES of the C1 polymers of interest.

Table S2. The predictions scores by SMILES-X for polymers listed in Chart 1.

Polymer	Predicted T_g (°C)	Error (°C)	Polymer	Predicted T_g (°C)	Error (°C)
M1	140.25	15.49	**H1**	97	27.29
M2	214.98	47.72	**H2**	58.43	26.08
M3	66.84	8.75	**H3**	51.44	9.83
M4	25.84	9.77	**H4**	181.16	47.44
M5	2.18	8.27	**H5**	75.04	31.29
M6	60.72	9.24	**H6**	76.2	8.76
M7	-13.12	8.64	**H7**	75.32	22.12
M8	89.1	16.84	**V1**	107.13	20.01
M9	-23.58	10.2	**V2**	151.1	28
M10	61.41	14.59	**A1**	-21.19	23.88
M11	114.8	19.72	**A2**	31.41	23.45
M12	55.64	15.15	**A3**	10.78	23.04
M13	-39.76	14.65	**A4**	-6.99	17.09
M14	183.15	18.66	**A5**	-12.58	14.83
M15	144.64	12.59	**A6**	-17.82	13.39
M16	91.49	11.85	**A7**	34.61	35.07
M17	113.17	22.18	**A8**	55.29	29.11
M18	150.52	24.24			

Additional figures and table

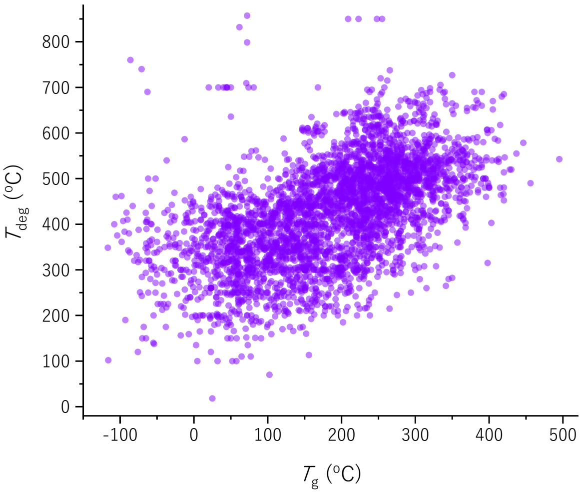

Figure S1. Correlation of T_g and T_{deg} from the entries in PoLyInfo database.

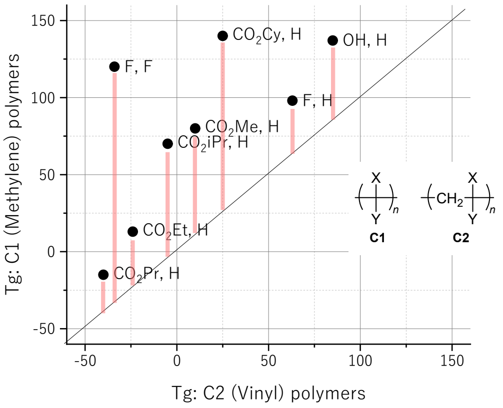

Figure S2. Tendency of higher glass transition temperature for all backbone-carbon substituted polymers (C1 polymers) than the corresponding vinyl polymers (C2 polymers).

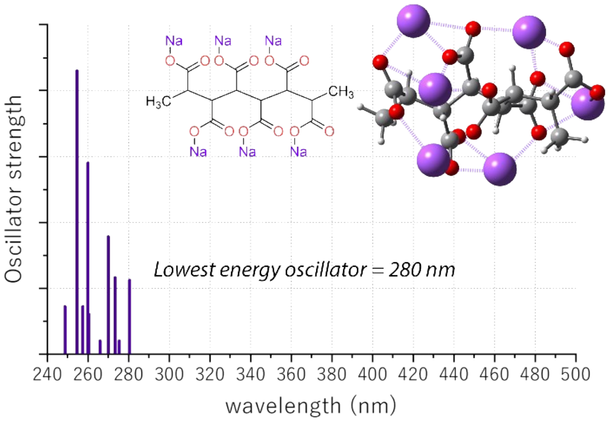

Figure S3. Geometry optimized structure and calculated absorption spectrum of the model compound of **M2** (PFA-Na).

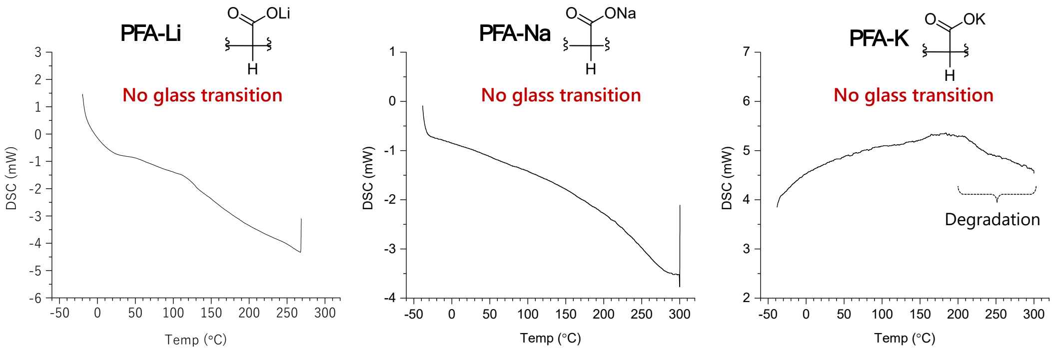

Figure S4. DSC analysis of PFA-Metals.

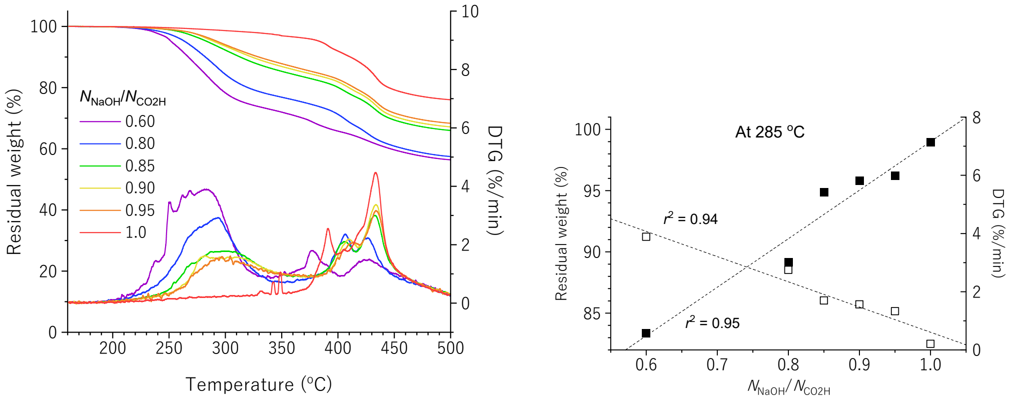

Figure S5. (left) Correlation between the equivalence of NaOH and T_{deg} value of PFA-Na. (right) Effect of equivalence at 285 °C.

```
bayopt_bounds = [{'name': 'lstmunits', 'type': 'discrete', 'domain’:
      [8, 16, 32, 64, 128, 256, 512, 1024]},
      {'name': 'denseunits', 'type': 'discrete', 'domain’:
      [8, 16, 32, 64, 128, 256, 512, 1024]},
      {'name': 'embedding', 'type': 'discrete', 'domain’:
      [8, 16, 32, 64, 128, 256, 512, 1024]}],
data_units = '',
bayopt_bounds=bounds,
k_fold_number = 10,
augmentation =False,
n_seeds = 1,
bayopt_n_rounds = 25,
bayopt_on = True,
lstmunits_ref = 16,
denseunits_ref = 128,
embedding_ref = 256,
seed_ref = None,
n_gpus = 1,
gpus_list = None,
gpus_debug = False,
patience = 50,
n_epochs = 400,
batchsize_pergpu = None,
lr_schedule = None,
lr_min = 1e-5,
```

Figure S6. Parameters and inputs used to run SMILES-X.

Table S3. Results of haze measurements of film samples on quartz glass plates.

Polymer	Total transmittance (%)	Haze (%)
PFA	91.6	0.34
PFA-Li	90.8	0.92
PFA-Na	91.1	34.8
PFA-K	90.9	0.59

Characterization of compounds

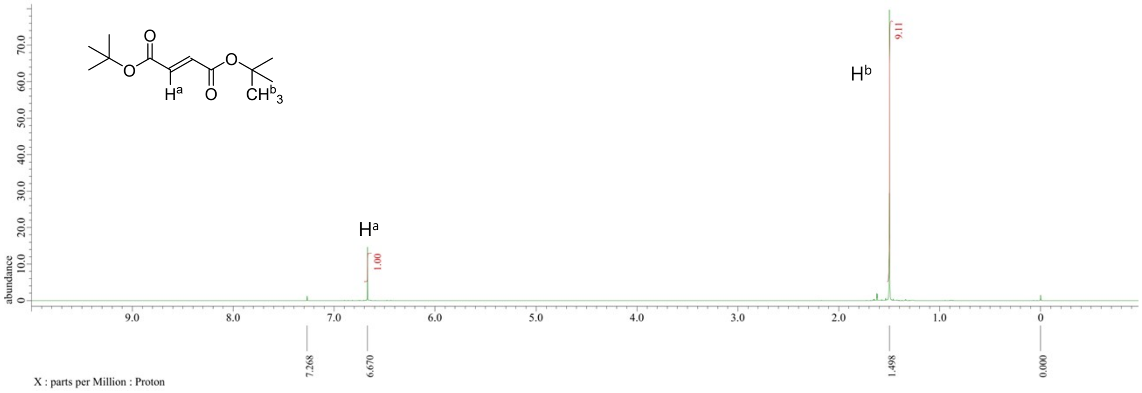

Figure S7. ^{1}H NMR spectrum of di-*tert*-butyl fumarate in $CDCl_3$.

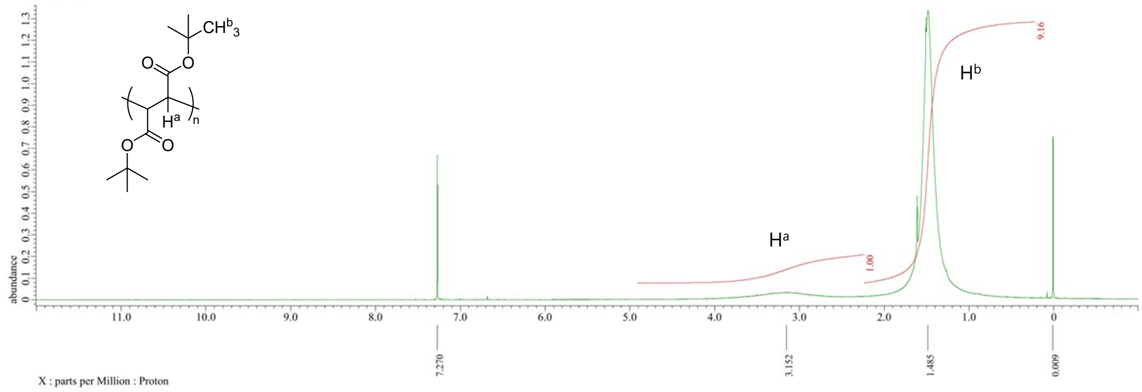

Figure S8. ^{1}H NMR spectrum of poly(di-*tert*-butyl fumarate) in $CDCl_3$.

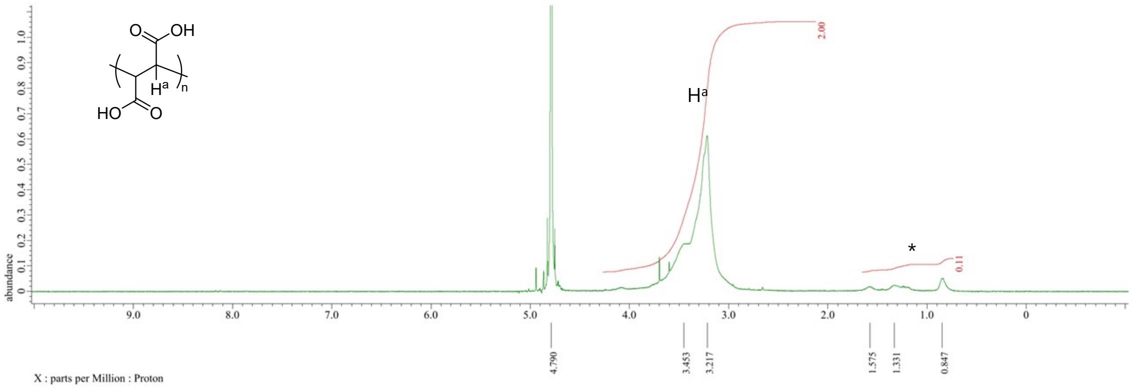

Figure S9. ^{1}H NMR spectrum of poly(fumaric acid) in D_2O. Asterisk indicate the unidentified impurity(es) derived from the pyrolysis of poly(di-*tert*-butyl fumarate).

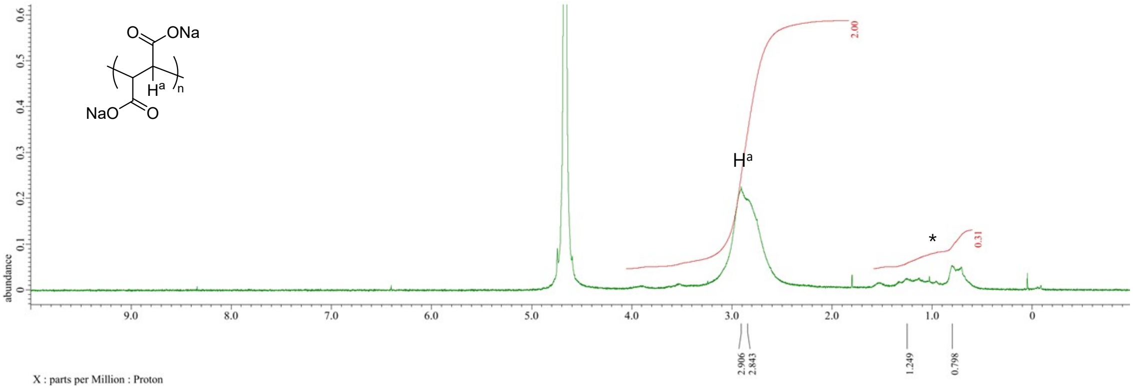

Figure S10. ^{1}H NMR spectrum of poly(disodium fumarate) in D_2O. Asterisk indicate the unidentified impurity(es) originated from the precursor poly(fumaric acid).

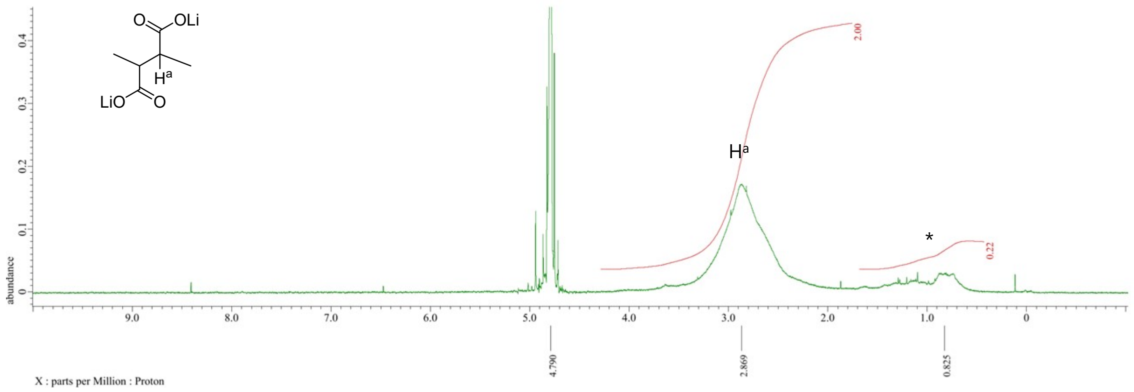

Figure S11. ^{1}H NMR spectrum of poly(dilithium fumarate) in D_2O. Asterisk indicate the unidentified impurity(es) originated from the precursor poly(fumaric acid).

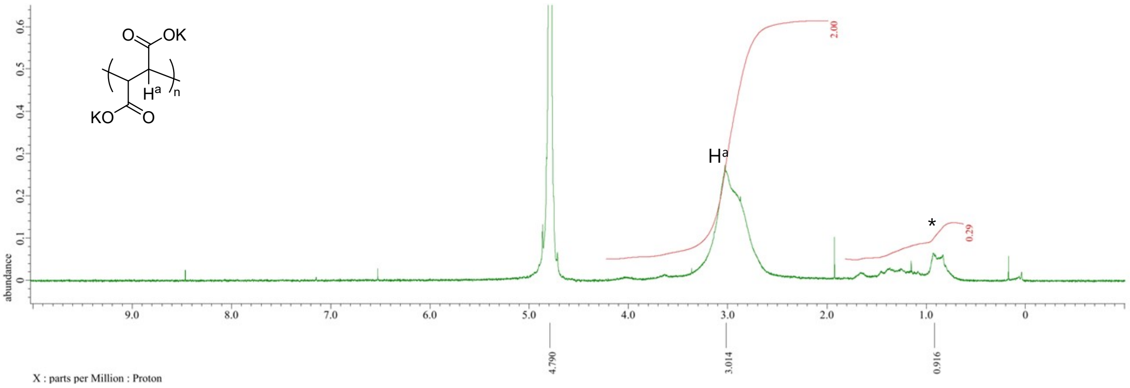

Figure S12. ^{1}H NMR spectrum of poly(dipotassium fumarate) in D_2O. Asterisk indicate the unidentified impurity(es) originated from the precursor poly(fumaric acid).

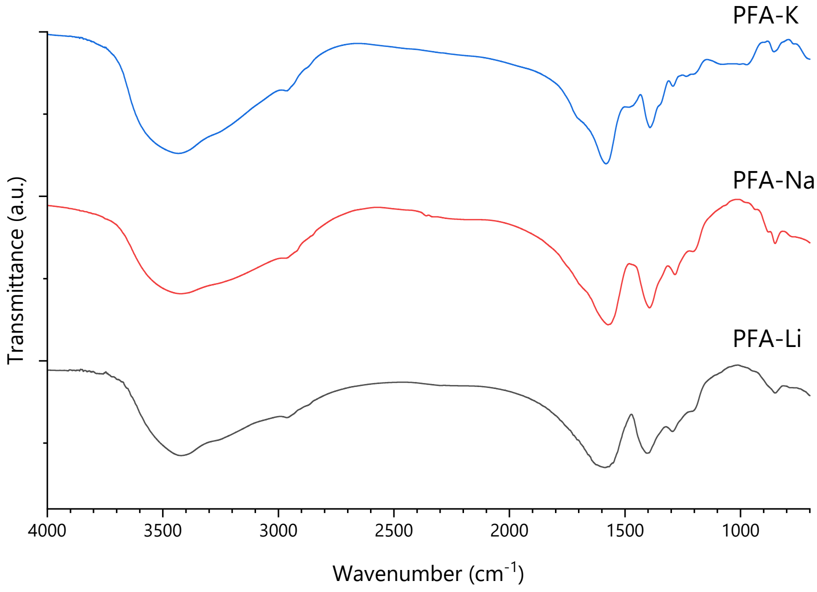

Figure S13. ATR-IR spectra of PFA-metals.

Simulation conditions in the estimation of T_g.

The conditions for molecular dynamics (MD) simulation of the quenching process were investigated to reduce the computational load and to quickly obtain a reasonable glass transition temperature (T_g) for the estimation of the range of T_g of the candidate polymers and filtering them. A common method for the estimation of T_g is the determination of the density of the system at different temperatures in each 20-50 K steps, and mostly the system was quenched from high to low temperature in this process. The T_g is estimated from the intersection of the bilinear fit for the plot of inverse of the density and the temperature. In one of the methods, the simulation for quenching process (temperature changing process) is not performed, and only isothermal equilibration is performed, and in another method, the temperature is descended over time slowly and then the system is equilibrated. Equilibration often takes more than 1 ns to make the density of the system constant, but this also depends on the quenching process. Alternatively, the temperature of the system is lowered slowly, and the T_g is estimated from the density during this process.

In this study, the T_g and density were investigated for a simulation in which the system is quenched at a rate of 1 °C/ps over 50 ps with NPT ensemble and then equilibrated for 200 ps with isothermal NPT ensemble (Run 1), equilibrated for 1000 ps (Run 2), or quenched at a rate of 1 °C/50 ps over 1000 ps (Run 3) (Table S4). In Run 1, the density reached almost constant by equilibration for 200 ps. The results showed that the effects of the simulation conditions on the T_g and the density at 273 K were less than 1.7% (corresponding to 7 K) and 2.7% (corresponding to 0.03 g/cm^3), respectively, which indicated that the effect is not significant and the estimated T_g is reasonable for the filtering of the candidate polymers. On the other hand, calculated density was lower than reported experimental values in these simulation conditions regardless of the simulated time (quenching condition) and polymer structures (chain length or number of chains, see below).

Table S4. Effect of the time program on the estimation of T_g of **M3**.[1]

Run	Simulation time program		T_g (K) (T_g in °C)	Density (g/cm^3)	
	Quenching time (ps)[2]	Equilibration time (ps)[3]		323 K (g/cm^3)	273 K (g/cm^3)
1	50	200	406 (133)	1.13	1.14
2	50	1000	413 (140)	1.13	1.14
3	1000	100	408 (135)	1.15	1.17
Max deviation (K or g/cm^3) [%Deviation][4]			7 [1.7]	0.02 [1.8]	0.03 [2.6]

[1] The number of [$(CHCO_2Me)_2$] repeating units is 25 and the number of chains in a simulation cell is 20. [2] Time for quenching NPT ensemble of 50 K descending. [3] Time for isothermal NPT ensemble at each temperature of 50 K step. [4] Deviation of the values divided by the value in run 1.

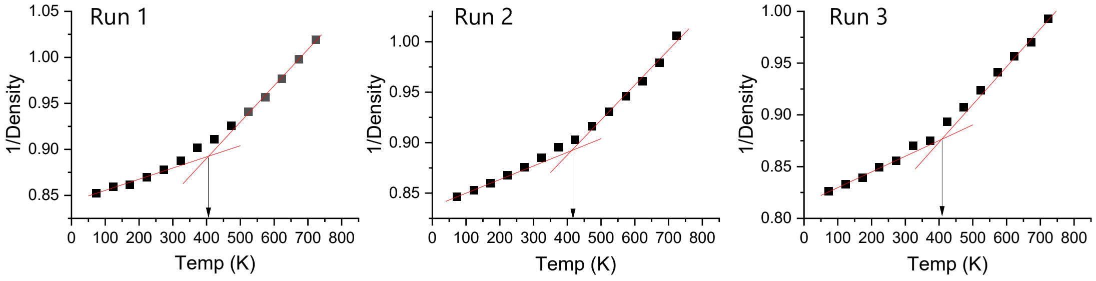

Figure S14. Plots for the determination of T_g.

The effect of the number of repeating monomer units in one polymer chain (degree of polymerization) and the number of chains in a simulation cell for the determination of glass transition temperature (T_g) in MD simulation was examined. When the number of atoms were almost fixed, the number of repeating units showed only a little effect on the determined T_g in the range of the number = 25 to 70, in which the deviation is less than 6.2% (PA-Me) (Table S5). Although the deviation of density depending on the number of repeating unit for **M2** was larger than others, T_g was not affected so much. The number of atoms in the simulation cells were not exactly same due to the difference of atoms in each chain. The trend of the effect was not same depending on the polymers examined. While PA-Me showed a larger T_g difference due to the number of repeating units, the trend of the effect is not simple, and it is difficult to interpret as the effect. Increase of the number of chains showed a little increase of T_g (Table S6). Consequently, these simulations indicates that simulation cells with 15 and more chains with repeating monomer units of 25 give reasonable T_g for this study.

Table S5. Effect of the chain length on the estimation of T_g.

Run	Polymer	Number of repeating units[1]	Number of chains[2]	Number of total atoms in a cell	T_g (K) (T_g in °C)	Density at 323 K (g/cm^3)
1	**M3**	25	20	9040	406 (133)	1.13
2		40	13	9386	404 (131)	1.14
3		70	8	10096	402 (129)	1.14
Max deviation (K or g/cm^3) [%Deviationi][3]					5 [1.2]	0.01 [0.1]
4	**M2** (PFA-Na)	25	20	6060	595 (322)	1.74
5		40	13	6266	593 (321)	1.51
6		70	8	6736	581 (308)	1.57
Max deviation (K or g/cm^3) [%Deviation][3]					11 [1.8]	0.23 [13.5]
7	Poly(methyl acrylate)[4]	25	20	4540	291 (18)	1.13
8		40	13	4706	287 (14)	1.14
9		70	8	5056	305 (32)	1.14
Max deviation (K or g/cm^3) [%Deviation][3]					18 [6.2]	0.01 [0.1]

1 Number of [$(CR_1R_2)_2$] or ($CH_2CR_1R_2$) repeating unit in a polymer chain. [2] Number of polymer chain in a simulation cell. [3] Deviation of the values divided by the value in run 1. [4] Poly(methyl acrylate) is the C2-polymer analogue of **M3** polymer.

Table S6. Effect of the chain length on the estimation of T_g of **M3**.

Run	Number of repeating units[1]	Number of chains[2]	Number of total atoms in a cell	T_g (K) (T_g in °C)	Density at 323 K (g/cm^3)
1	25	15	6780	415 (142)	1.11
2	25	20	9040	406 (133)	1.13
3	25	40	18080	410 (137)	1.12
4	40	13	9386	404 (131)	1.14
5	40	26	18772	413 (140)	1.12
Max deviation (K or g/cm^3) [%Deviation][3]				8 [1.9]	0.03 [2.7]

1 Number of [$(CHCO_2Me)_2$] repeating unit in a polymer chain. [2] Number of polymer chain in a simulation cell. [3] Deviation of the values divided by the value in run 2.

References

1 M. Yoshioka, A. Matsumoto and T. Otsu, *Polym. J.*, 1991, **23**, 1191–1196.

2 A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott and S. J. Plimpton, *Comput. Phys. Commun.*, 2022, **271**, 108171.

3 W. M. Brown, P. Wang, S. J. Plimpton and A. N. Tharrington, *Comput. Phys. Commun.*, 2011, **182**, 898–911.

4 S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.

5 Gaussian 16, Revision C.01.

6 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.

7 R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, Adam Hilger, New York, 1988.

8 S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu and M. Yamazaki, in *2011 International Conference on Emerging Intelligent Data and Web Technologies*, IEEE, Tirana, Albania, 2011, pp. 22–29.

9 G. Lambard and E. Gracheva, *Mach. Learn. Sci. Technol.*, 2020, **1**, 025004.

10 SMILES-X, (https://github.com/GLambard/SMILES-X).