Supporting Information

Discovery of Innovative Polymers for Next-Generation Gas-Separation Membranes using Interpretable Machine Learning

Jason Yang [a,†], Lei Tao [b,†], Jinlong He [b,†], Jeffrey R McCutcheon [c,d], and Ying Li [b,d*]

[a] Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States
[b] Department of Mechanical Engineering, University of Connecticut, Storrs, Connecticut 06269, United States
[c] Department of Chemical & Biomolecular Engineering, Center for Environmental Sciences and Engineering, University of Connecticut, Storrs, Connecticut 06269, United States
[d] Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, Connecticut 06269, United States

[†]Equal contribution

*Corresponding Author: Ying Li; Email: ying.3.li@uconn.edu; Tel: (860) 483-7110. Fax: (860) 486-5088.

MaxEStateIndex	PEOE_VSA9	NumAliphaticHeterocycles	fr_benzodiazepine
MinEStateIndex	SMR_VSA1	NumAliphaticRings	fr_bicyclic
MaxAbsEStateIndex	SMR_VSA10	NumAromaticCarbocycles	fr_diazo
MinAbsEStateIndex	SMR_VSA2	NumAromaticHeterocycles	fr_dihydropyridine
qed	SMR_VSA3	NumAromaticRings	fr_epoxide
MolWt	SMR_VSA4	NumHAcceptors	fr_ester
HeavyAtomMolWt	SMR_VSA5	NumHDonors	fr_ether
ExactMolWt	SMR_VSA6	NumHeteroatoms	fr_furan
NumValenceElectrons	SMR_VSA7	NumRotatableBonds	fr_guanido
NumRadicalElectrons	SMR_VSA8	NumSaturatedCarbocycles	fr_halogen
MaxPartialCharge	SMR_VSA9	NumSaturatedHeterocycles	fr_hdrzine
MinPartialCharge	SlogP_VSA1	NumSaturatedRings	fr_hdrzone
MaxAbsPartialCharge	SlogP_VSA10	RingCount	fr_imidazole
MinAbsPartialCharge	SlogP_VSA11	MolLogP	fr_imide
FpDensityMorgan1	SlogP_VSA12	MolMR	fr_isocyan
FpDensityMorgan2	SlogP_VSA2	fr_Al_COO	fr_isothiocyan
FpDensityMorgan3	SlogP_VSA3	fr_Al_OH	fr_ketone
BalabanJ	SlogP_VSA4	fr_Al_OH_noTert	fr_ketone_Topliss
BertzCT	SlogP_VSA5	fr_ArN	fr_lactam
Chi0	SlogP_VSA6	fr_Ar_COO	fr_lactone
Chi0n	SlogP_VSA7	fr_Ar_N	fr_methoxy
Chi0v	SlogP_VSA8	fr_Ar_NH	fr_morpholine
Chi1	SlogP_VSA9	fr_Ar_OH	fr_nitrile
Chi1n	TPSA	fr_COO	fr_nitro
Chi1v	EState_VSA1	fr_COO2	fr_nitro_arom
Chi2n	EState_VSA10	fr_C_O	fr_nitro_arom_nonortho
Chi2v	EState_VSA11	fr_C_O_noCOO	fr_nitroso
Chi3n	EState_VSA2	fr_C_S	fr_oxazole
Chi3v	EState_VSA3	fr_HOCCN	fr_oxime
Chi4n	EState_VSA4	fr_Imine	fr_para_hydroxylation
Chi4v	EState_VSA5	fr_NH0	fr_phenol
HallKierAlpha	EState_VSA6	fr_NH1	fr_phenol_noOrthoHbond
Ipc	EState_VSA7	fr_NH2	fr_phos_acid
Kappa1	EState_VSA8	fr_N_O	fr_phos_ester
Kappa2	EState_VSA9	fr_Ndealkylation1	fr_piperdine
Kappa3	VSA_EState1	fr_Ndealkylation2	fr_piperzine
LabuteASA	VSA_EState10	fr_Nhpyrrole	fr_priamide
PEOE_VSA1	VSA_EState2	fr_SH	fr_prisulfonamd
PEOE_VSA10	VSA_EState3	fr_aldehyde	fr_pyridine
PEOE_VSA11	VSA_EState4	fr_alkyl_carbamate	fr_quatN
PEOE_VSA12	VSA_EState5	fr_alkyl_halide	fr_sulfide

PEOE_VSA13	VSA_EState6	fr_allylic_oxid	fr_sulfonamd
PEOE_VSA14	VSA_EState7	fr_amide	fr_sulfone
PEOE_VSA2	VSA_EState8	fr_amidine	fr_term_acetylene
PEOE_VSA3	VSA_EState9	fr_aniline	fr_tetrazole
PEOE_VSA4	FractionCSP3	fr_aryl_methyl	fr_thiazole
PEOE_VSA5	HeavyAtomCount	fr_azide	fr_thiocyan
PEOE_VSA6	NHOHCount	fr_azo	fr_thiophene
PEOE_VSA7	NOCount	fr_barbitur	fr_unbrch_alkane
PEOE_VSA8	NumAliphaticCarbocycles	fr_benzene	fr_urea

Table S1. List of function names for molecular descriptor calculation in RDKit.

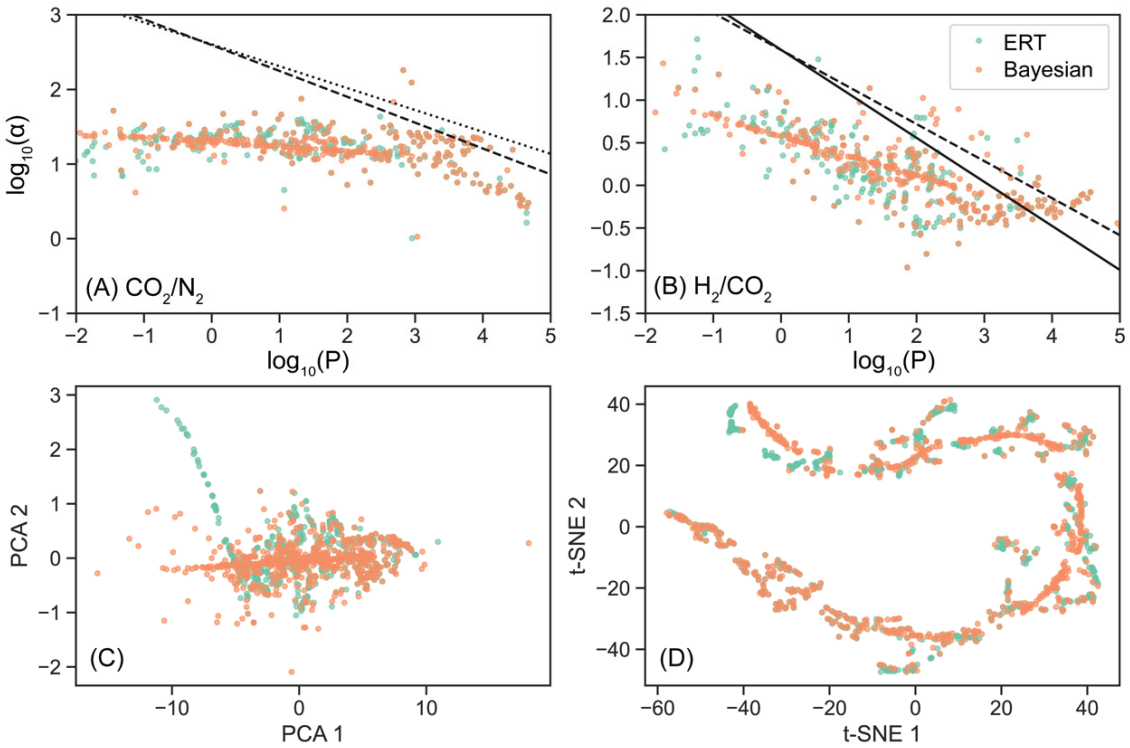


Fig. S1. Further visualization of imputed permeabilities in the training set, Dataset A. (a) CO_2/N_2 and (b) H_2/CO_2 Robeson plots after imputation using extremely randomized trees (ERT) and Bayesian linear regression, with permeabilities averaged across entries that correspond to the same polymer. Solid, dashed, and dotted lines refer to the 1991, 2008, and 2015/2019 upper bounds, respectively. The gas permeabilities of all six gases (He, H_2, O_2, N_2, CO_2, and CH_4) as visualized in lower-dimensional space using (e) principal components analysis (PCA) and (f) t-distributed stochastic neighbor embedding (t-SNE) show that linear and non-linear imputation yield similar outputs.

For the generation of Dataset C, polyimides can be synthesized through the reaction between a dianhydride and a diamine with the elimination of water molecules[1], or through the reaction between a dianhydride and a diisocyanate with the elimination of carbon dioxide molecules[2], shown in **Fig. S2(a)**. All of the polyimides in Dataset C are generated via this mechanism from binary pairs of known chemicals in PubChem[3].

Double-stranded ladder polymers can be obtained through the binary reaction of a tetra-halogenated aromatic monomer and a tetra-hydroxy aromatic monomer[4], as shown in **Fig. S2(b)**. We obtain ~500 ladder polymers from different binary combinations of selected reacting monomers, which defines a chemical space of ladder polymers with sparse data points. To better explore this chemical space, we train a RNN model to learn its structure patterns and populate another 500 ladder polymers to supplement the chemical space. Together, the ladder polymers from monomer combinations and RNN generation comprise Dataset D.

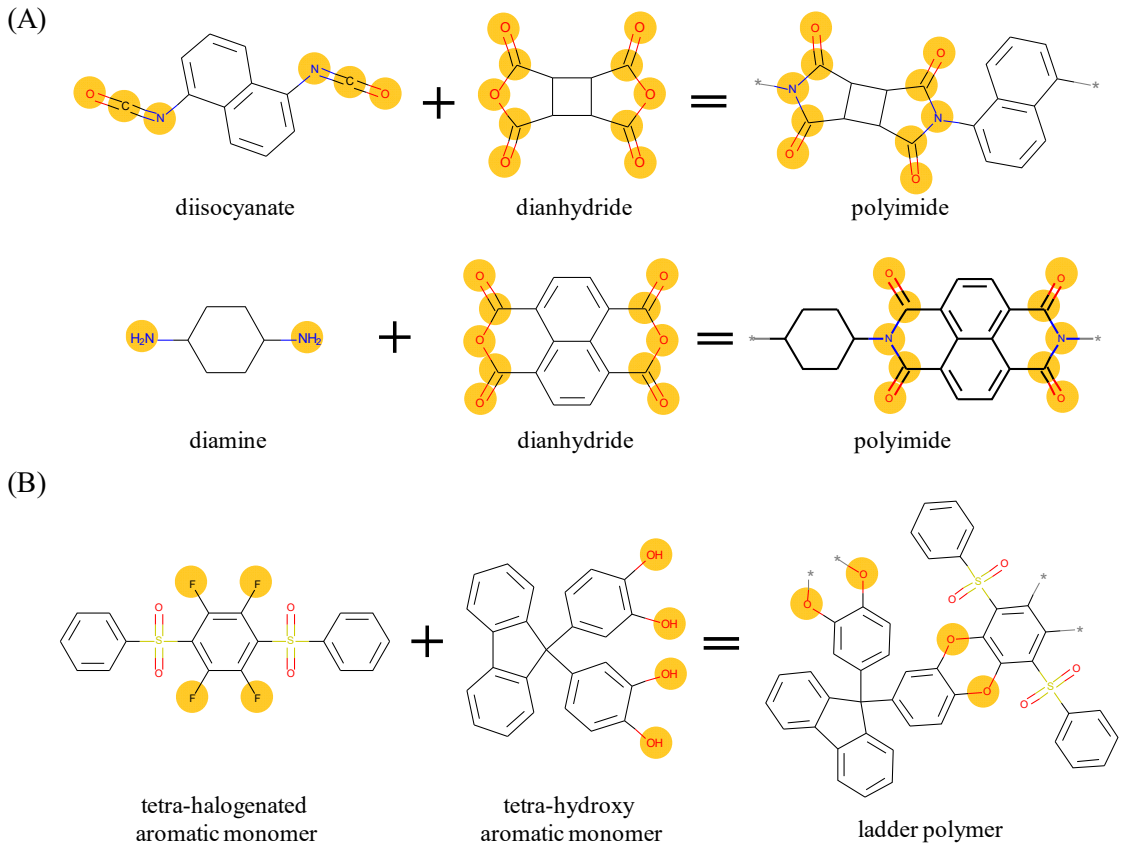


Fig. S2. In-silico reaction routes for the generation of polyimides (Dataset C) and ladder polymers (Dataset D). (a) The polycondensation of diisocyanate/diamine and dianhydride to form a polyimide. Each generated polyimide features two functional imide groups. (b) The formula for ladder polymer formation from a tetra-halogenated aromatic monomer and a tetra-hydroxy aromatic monomer. Each ladder polymer repeating unit features four connection points along its double stranded backbone. Reacting groups are highlighted in yellow.

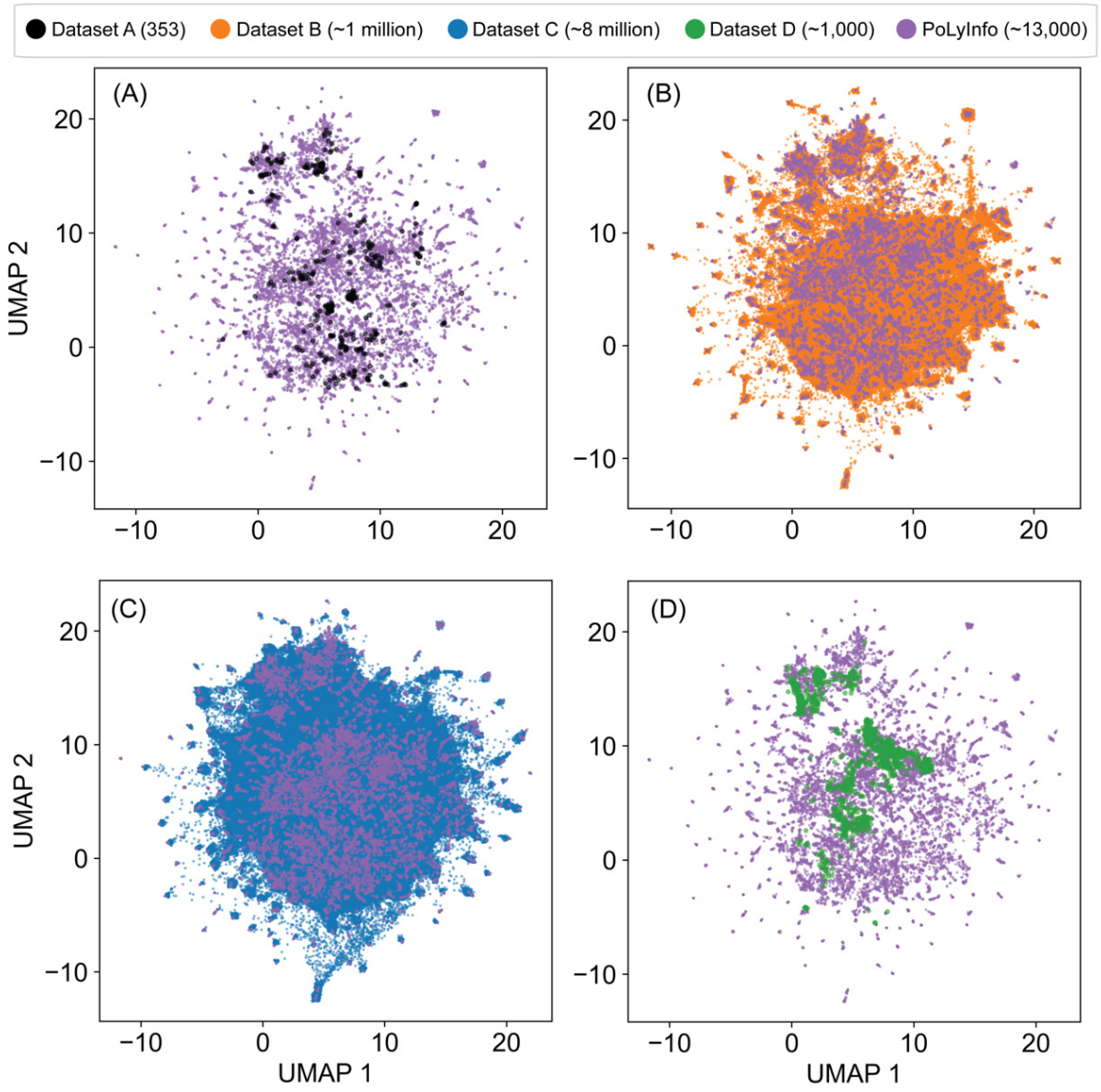


Fig. S3. Visualization of the MFF chemical space for the datasets explored in this work, using uniform manifold approximation and projection (UMAP), which captures local and global structure in the data. (a-d) correspond to Datasets A-D respectively, each compared to the PoLyInfo database, one of the largest polymer databases. Dataset A (training set) generally spans the feature space occupied by Datasets B, C, and D.

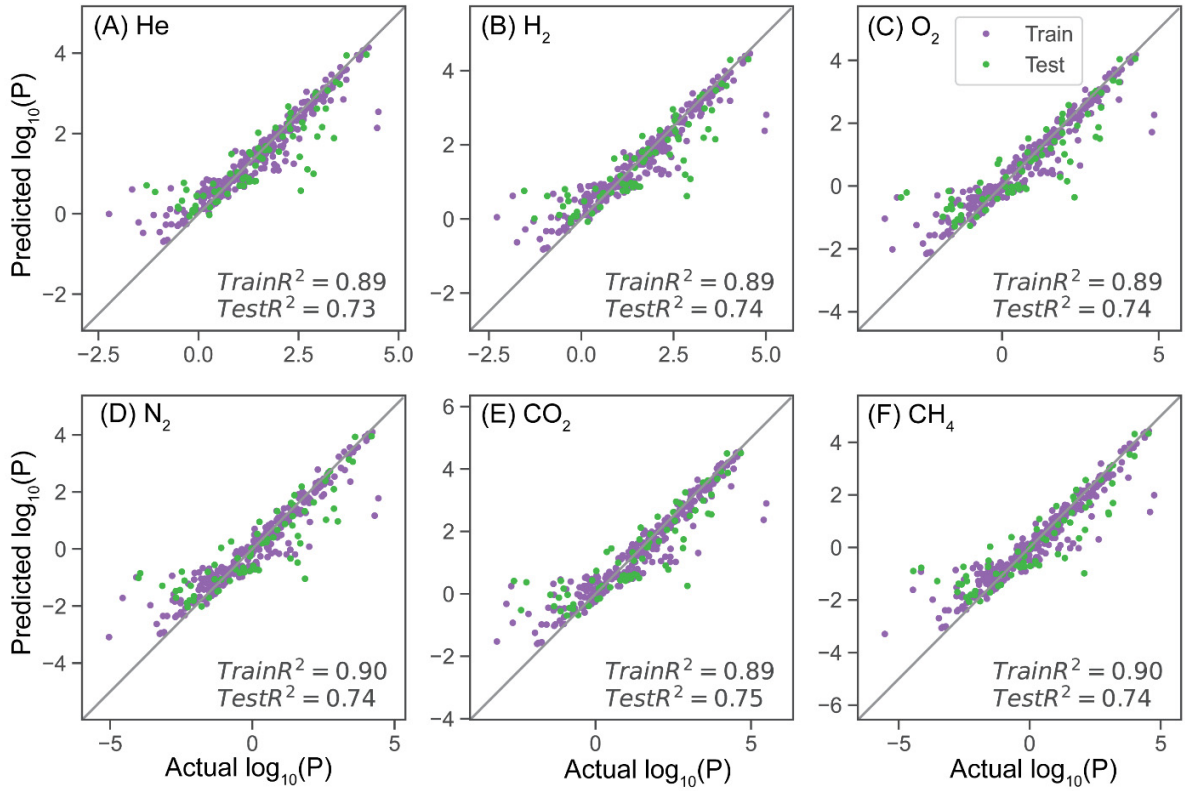


Fig. S4. Summary of the performance of the random forest model trained on MFFs and gas permeabilities imputed using BLR. The predicted and actual permeabilities in Barrers for six gases are plotted for the train and test datasets.

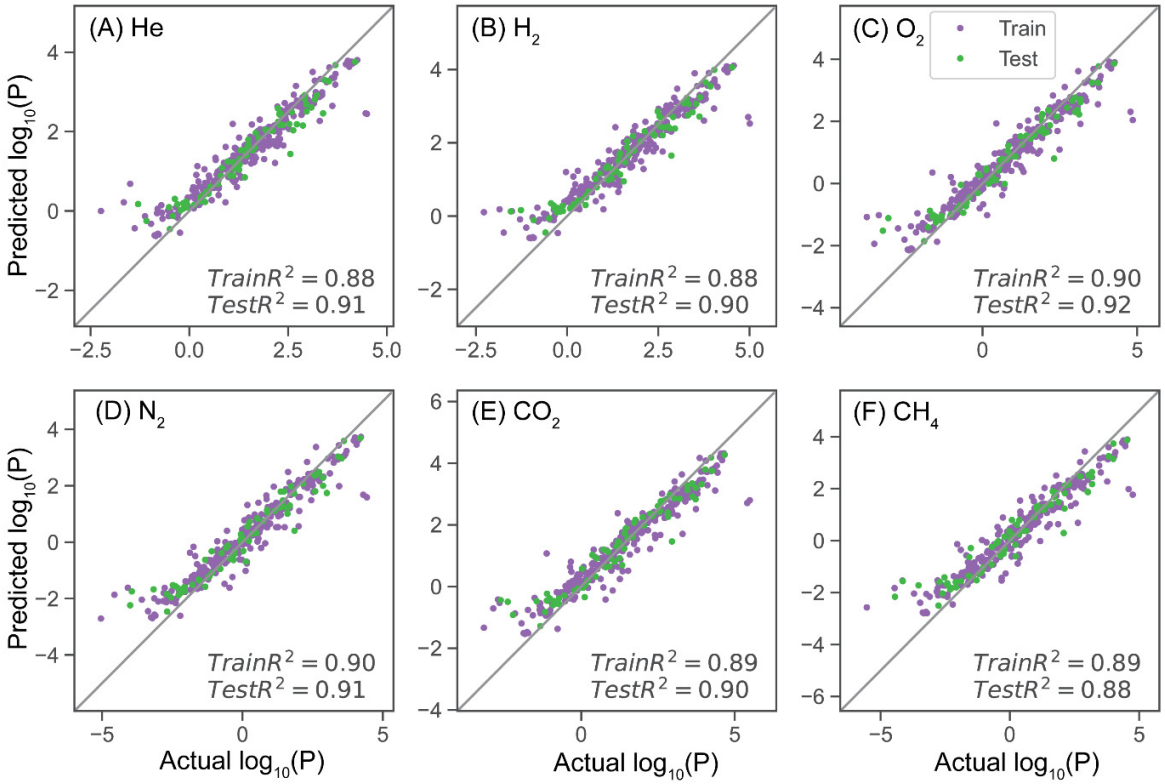


Fig. S5. Summary of the performance of the ensemble of DNNs trained on MFFs and gas permeabilities imputed using BLR. The predicted and actual permeabilities in Barrers for six gases are plotted for the train and test datasets.

	He	H_2	O_2	N_2	CO_2	CH_4
Mean single-model test R^2	0.69	0.68	0.71	0.68	0.70	0.70
Std of R^2	0.043	0.070	0.049	0.098	0.043	0.054
Ensemble test R^2	0.91	0.90	0.92	0.91	0.90	0.88
Mean variance of normalized test predictions	0.071	0.086	0.072	0.101	0.073	0.074

Table S2. Summary of uncertainty quantification of the ensemble of 16 DNNs trained on MFFs and permeabilities imputed using BLR. The last row refers to variance on predicted permeability values that have been normalized to a standard deviation of 1 and a mean of 0.

ID	Descriptor Name	Description
101	VSA_EState8	Hybrid MOE-type descriptor using EState indices and VSA contributions for $6.45 < x < 7.00$
107	NumAliphaticCarbocycles	Number of carbocycles that contain at least one non-aromatic bond
109	NumAliphaticRings	Number of rings that contain at least one non-aromatic bond
15	FpDensityMorgan2	Approximated density of the molecule
75	SlogP_VSA4	Sum of the approximate accessible VDW surface areas of atoms in the molecule with contributions between 0 and 0.1 to the molecule octanol/water partition coefficient calculation method proposed by Crippen
79	SlogP_VSA8	Sum of the approximate accessible VDW surface areas of atoms in the molecule with contributions between 0.25 and 0.3 to the molecule octanol/water partition coefficient calculation method proposed by Crippen
103	FractionCSP3	Fraction of C atoms that are sp3 hybridized
91	EState_VSA8	Hybrid MOE-type descriptor using EState indices and VSA contributions for $2.05 < x < 4.69$
16	FpDensityMorgan3	Approximated density of the molecule
69	SlogP_VSA1	Sum of the approximate accessible VDW surface areas of atoms in the molecule with contributions less than -0.40 to the molecule octanol/water partition coefficient calculation method proposed by Crippen
92	EState_VSA9	Hybrid MOE-type descriptor using EState indices and VSA contributions for $4.69 < x < 9.17$
65	SMR_VSA6	MOE-type descriptor using surface area contributions where $2.75 < x < 3.05$

Table S3. Definitions of the top molecular descriptors as identified using SHAP on the DNN model trained on descriptors with permeabilities imputed using BLR. All descriptors are calculated based on the graph structure and atom and bond types within the molecule. EState refers to the electrotopological state indices calculated as proposed by Hall and Kier[5]. Large positive values indicate atoms of high electronegativity and/or terminal atoms, while small or negative EState values indicate atoms with only σ electrons, interior atoms, or atoms near electronegative atoms. VSA refers to the accessible van der Waals surface area.

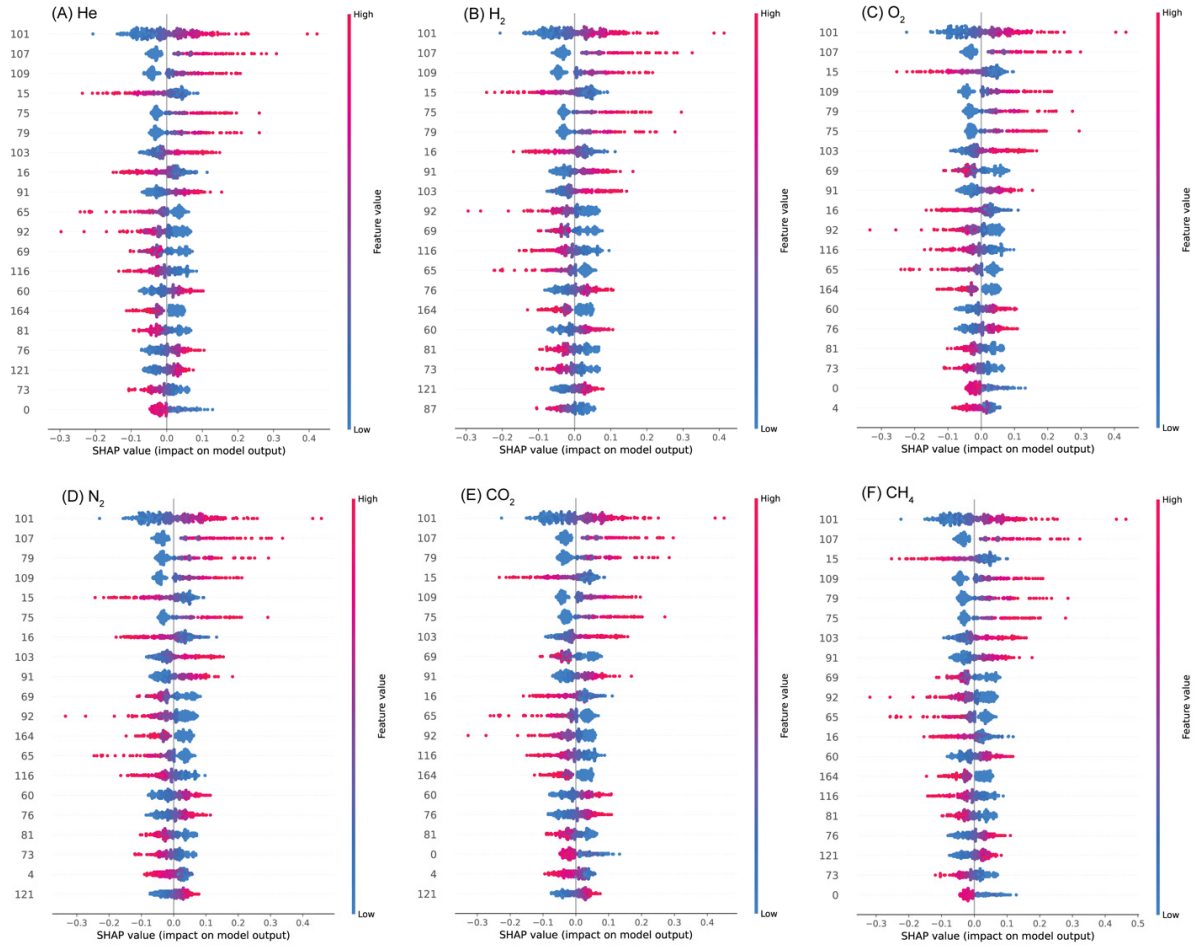


Fig. S6. SHAP summary plots showing the impacts of the twenty most important chemical descriptors for predicting (a) He, (b) H_2, (c) O_2, (d) N_2, (e) CO_2, and (f) CH_4 gas permeabilities in the DNN ensemble model trained on descriptors with permeabilities imputed via BLR.

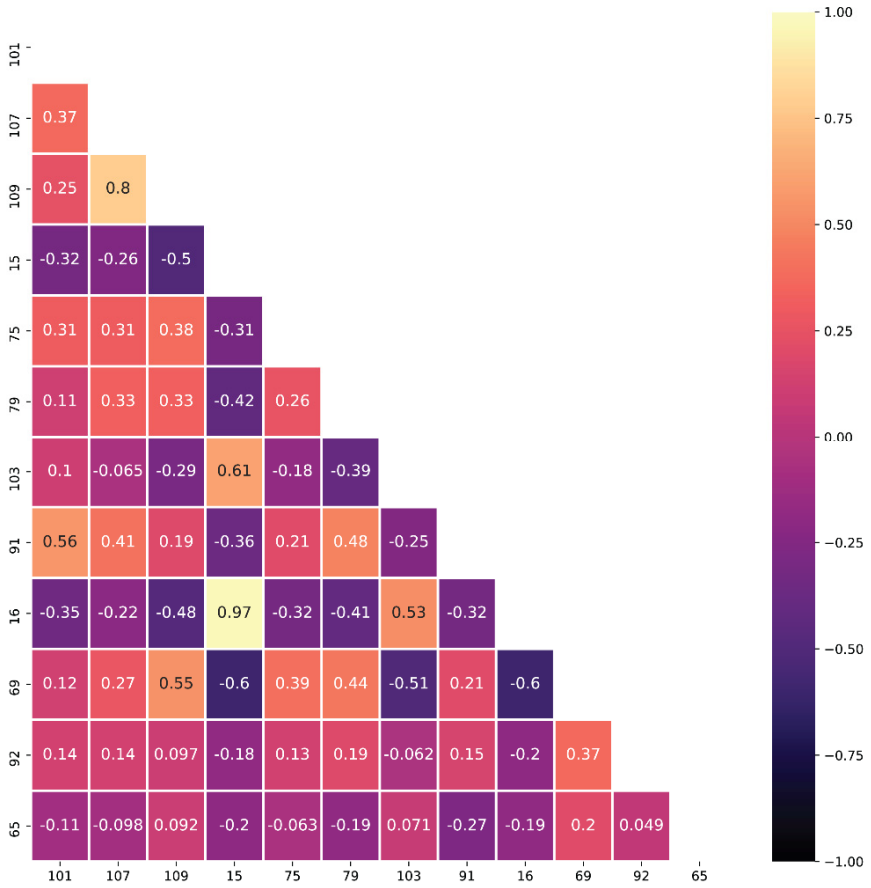


Fig. S7. Correlations of top molecular descriptors. The descriptors can be broken down into two main opposing correlation groups: [15, 16] are anti-correlated with [107, 109], whereas the other descriptors are relatively independent.

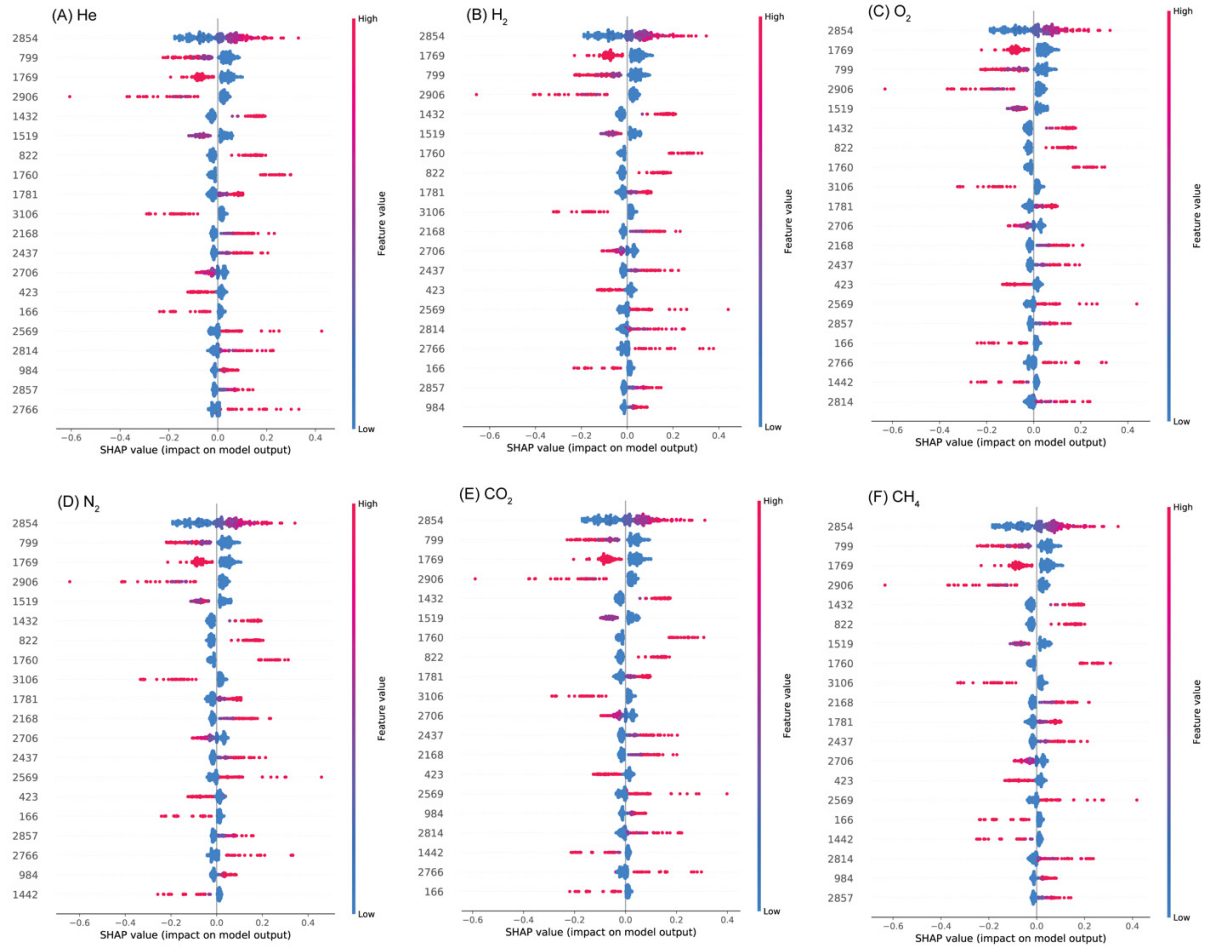


Fig. S8. SHAP summary plots showing the impacts of the twenty most important chemical substructures for predicting (a) He, (b) H_2, (c) O_2, (d) N_2, (e) CO_2, and (f) CH_4 gas permeabilities in the DNN ensemble model trained on MFFs with permeabilities imputed via BLR.

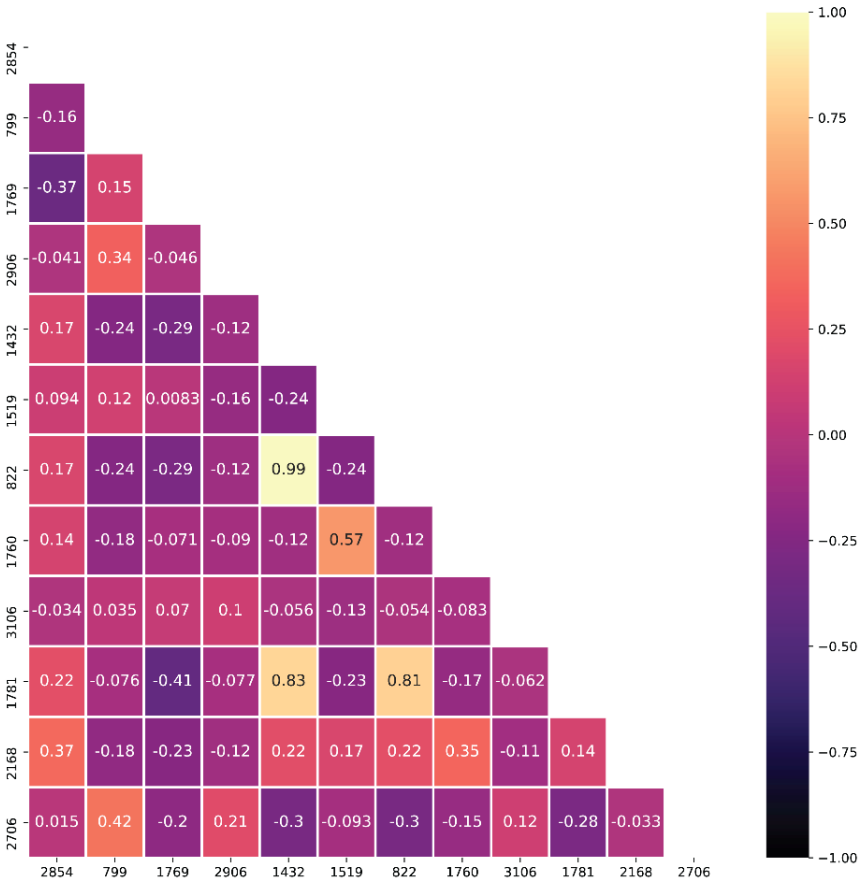


Fig. S9. Correlations of top molecular fingerprints. The main correlation group is [822, 1432, 1781].

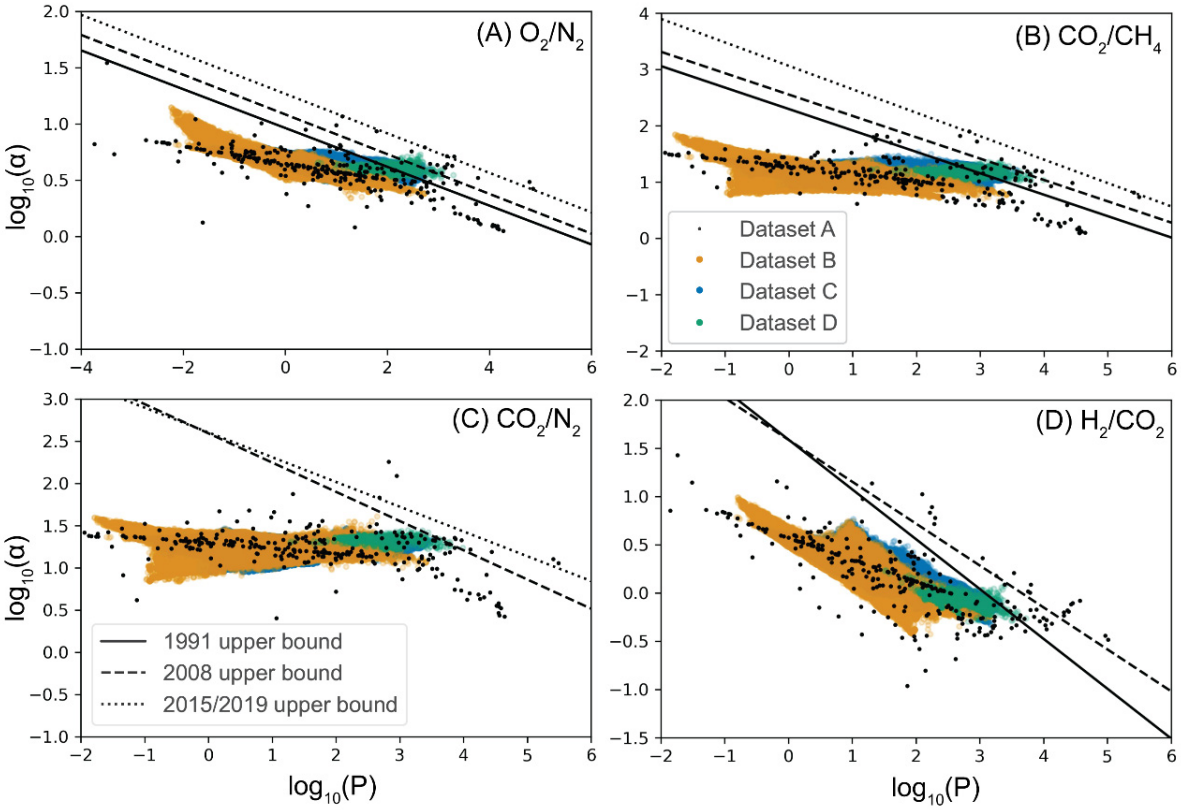


Fig. S10. Visualization of predicted permeabilities for polymers in Datasets B, C, and D, based on the RF trained on MFFs with BLR-imputed permeabilities. The training dataset (Dataset A) is overlaid on the predicted permeabilities. The data is visualized for (a) O_2/N_2, (b) CO_2/CH_4, (c) CO_2/N_2, and (d) H_2/CO_2 separations. Units of permeability are Barrers.

Polymer Candidate	SMILES String
P-DNN-C1	*C1CCC(C(C)(C)c2cccc(C(C)(C)C3CCC(N4C(=O)c5ccc6c7c(C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)cc8c9c(cc(C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)c(c%10ccc(c5c6%10)C4=O)c97)C(=O)N(*)C8=O)CC3)c2)CC1
P-DNN-C2	*N1C(=O)c2cc(C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)c3c4ccc5c6c(ccc(c7c(C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)cc(c2c37)C1=O)c64)C(=O)N(C(C)(C)C(C)(C)C(C)(C)C1C(C)(C)C(C)(C)C(C)(C)C(*)(C)C)C(C)(C)C1(C)C)C5=O
P-DNN-C3	*C1CCCCCN1C(=O)N1CCN(CC)C(=O)NC[Si]2(C(C)(C)C)COC[Si](C(C)(C)C)(CCCNC(=O)N(N(C)C(=O)N3CCCCCC3N3C(=O)C4CCC(C5CCC6C(=O)N(*)C(=O)C6C5)CC4C3=O)CC1)OC(C)C(C)O2
P-DNN-C4	*N(C(C)=O)C(=O)CS(=O)(=O)CCS(=O)(=O)CC(=O)N(C(C)=O)C(C)C1(*)CCCCCCCCCCCC12CCCCCCCCCCC21CCCCCCCCCCC12CCCCCCCCCC2
P-DNN-D1	*Oc1cc2c(cc1O*)C1(c3cc4c(cc3-c3cc5c(cc31)C(C)(C)CCC5(C)C)Oc1c(c(C(=O)c3ccc(-c5ccccc5)cc3)c(*)c(*)c1C(=O)c1ccc(-c3ccccc3)cc1)O4)c1cc3c(cc1-2)C(C)(C)CCC3(C)C
P-DNN-D2	*Oc1cc2c(cc1O*)C1(c3cc4c(cc3-c3cc5c(cc31)C(C)(C)CCC5(C)C)Oc1c(c(C(=O)c3cc(F)c(F)c(F)c3)c(*)c(*)c1C(=O)c1cc(F)cc(F)c1)O4)c1cc3c(cc1-2)C(C)(C)CCC3(C)C
P-RF-C	*c1c(C)cc2ccccc2c1-c1c(N2C(=O)c3cc4c(cc3C2=O)C2(C)CCC4(C)c3cc4c(cc32)C(=O)N(*)C4=O)c(C)cc2ccccc12
P-RF-D	*Oc1cc2c(cc1O*)C1(c3ccccc3-2)c2ccccc2-c2cc3c(cc21)Oc1cc2c(cc1O3)C(=O)N(c1c(C)cc(C)c(N3C(=O)c4cc(*)c(*)cc4C3=O)c1C)C2=O

Table S4. Selected top polymer candidates and their associated SMILES strings.

Polymer	# Gas mols:	Gas Number/ Diffusivity [10^{-8} cm^2/s]					Solubility [cm^3 (STP) /cm^3 bar]
		10	20	30	40	50	
P-DNN-C1	CH_4	44.6067	60.5667	59.4783	58.085	99.6317	21.473
	CO_2	79.935	190.8	181.9	190.9933	261.8417	69.456
	N_2	53.1483	79.9667	82.9083	82.6283	124.0167	13.161
	O_2	283.4	439.5333	454.55	445.4833	658.0833	10.67
	H_2	5085.1768	8706.2073	8612.0009	8318.7770	9630.0227	0.827
P-DNN-C2	CH_4	1066.467	2940.167	3008.667	3069.833	4436.917	33.832
	CO_2	2457.167	9288.333	8900	8613.333	11448.17	75.4
	N_2	1186.283	2080	2329.5	2203.667	2603.667	13.843
	O_2	3436.667	5883.833	5820.167	5622	6508.167	9.081
	H_2	18510.4800	25402.7283	25005.3350	25673.7933	31672.1650	1.359
P-DNN-C3	CH_4	607.9195	934.1333	974.1057	900.75	1510.1495	57.225
	CO_2	16001.2953	27005.769	28008.8731	27506.6922	45003.2976	1.419
	N_2	365.1723	632.5788	613.1065	653.1013	904.6467	10.017
	O_2	4008.6947	6204.4736	6608.0365	6406.684	9503.2363	6.088
	H_2	450.3511	709.6833	632.9442	669.0403	1258.4709	33.926
P-DNN-C4	CH_4	2710.163	4709.488	4501.601	4654.203	8001.208	34.534
	CO_2	30007.729	47002.976	45508.835	47506.752	70008.972	1.983
	N_2	347.1227	852.2008	822.1247	862.3635	1509.3613	8.135
	O_2	7008.975	10503.245	10005.407	9504.424	18007.930	5.424
	H_2	457.240	1359.953	1282.746	1486.643	1808.290	14.943
P-DNN-D1	CH_4	290.7	250.95	252.375	270.325	399.325	5.700
	CO_2	125.1567	163.38	154.735	156.5817	163.5967	75.101
	N_2	202.4	421.8833	414.8667	432.25	791.7167	3.68
	O_2	619.4833	1899	1863.5	1949	2496.5	1.518
	H_2	9999.3383	15892.3700	15706.9133	16479.6533	19980.1867	0.284
P-DNN-D2	CH_4	303.4167	480.8667	428.8333	537	549.9	5.922
	CO_2	652.225	1122.87	1081.99	1012.725	3161.725	17.876
	N_2	482.5	1052.2	1023.517	1046.45	1834.833	3.581
	O_2	1226.667	2994.667	2666.667	2752.167	3828.833	2.175
	H_2	14993.9500	21997.9900	20999.5850	23004.1967	27019.6850	0.389
P-RF-C	CH_4	12.9317	12.33	11.9883	13.2017	35.76	16.050
	CO_2	30.8567	30.74	32.5	34.9483	101.59	70.120
	N_2	5.9833	12.7667	10.8333	12.0333	19.4883	10.941
	O_2	36.65	64.86	73.9517	70.12	78.1167	6.575
	H_2	1001.2217	1902.2850	2058.2783	2002.8350	2890.8383	0.716
P-RF-D	CH_4	1.8483	2.3267	3.02	2.745	6.98	52.742
	CO_2	8.9933	9.8783	8.4167	8.4933	29.4	218.717
	N_2	1.9	3.0042	3.9547	3.3403	6.6538	29.255
	O_2	15.8738	33.63	33.0467	37.5233	64.1417	13.437
	H_2	644.5450	1201.9217	1185.6833	1244.1850	1654.6150	1.203

Table S5. Intermediate values in the calculation of permeability as diffusivity times solubility for selected top polymer candidates, via MD simulations. The diffusivity was calculated as the average of the shaded values, which refer to simulations with 20, 30, and 40 gas molecules.

Polymer	Gas	MD-Simulated Permeability (P_{MD})	Standard Deviation of P_{MD}	ML-Predicted Permeability (P_{ML})	$\log(P_{ML})$	Standard Deviation of $\log(P_{ML})$ from Bootstrapping	%Difference $(P_{ML}-P_{MD})/P_{MD}$
P-DNN-C1	CH_4	1657.512	84.81211	1653.12891	3.218306	2.019868	-0.26%
	CO_2	16965.81	602.6442	17662.1029	4.247042	1.803747	4.10%
	N_2	1400.094	46.43858	1367.46	3.135914	2.210223	-2.33%
	O_2	6193.831	158.9193	6251.10322	3.795956	1.777441	0.92%
	H_2	4727.856	123.3328	4685.6193	3.670767	1.803201	-0.89%
P-DNN-C2	CH_4	52486.33	1830.314	59462.9225	4.774246	2.756414	13.29%
	CO_2	155897.4	5214.004	339634.203	5.531011	2.413820	117.86%
	N_2	38486.6	3429.304	49003.579	4.69022	2.302411	27.33%
	O_2	100832.1	8783.601	125618.927	5.09905	2.180247	24.58%
	H_2	44794.08	1108.444	45040.1678	4.6536	1.950639	0.55%
P-DNN-C3	CH_4	69655.9	2230.851	66988.4609	4.826	3.704102	-3.83%
	CO_2	50742.37	755.4335	51789.4144	4.714241	2.526167	2.06%
	N_2	8242.063	212.6468	7448.64026	3.872077	3.014252	-9.63%
	O_2	50702.8	1303.929	53790.1803	4.730703	3.087153	6.09%
	H_2	29574.06	1382.519	32884.0273	4.516985	2.570558	11.19%
P-DNN-C4	CH_4	207490.4	3946.849	215341.632	5.333128	4.503192	3.78%
	CO_2	120318	2187.07	126352.547	5.101584	3.771231	5.02%
	N_2	8942.252	180.6803	9605.36782	3.982514	4.410758	7.42%
	O_2	70542.73	2875.251	80745.9968	4.907121	3.758369	14.46%
	H_2	26738.73	1632.816	22549.9197	4.353145	3.595450	-15.67%
P-DNN-D1	CH_4	1911.04	70.41435	1709.35828	3.23283	0.799550	-10.55%
	CO_2	15448.42	371.7476	14724.1268	4.168029	0.800241	-4.69%
	N_2	2023.817	60.26185	2117.81715	3.325888	0.920442	4.64%
	O_2	3757.197	83.67529	3675.05404	3.565263	0.824867	-2.19%
	H_2	5915.074	158.5987	5925.44539	3.772721	0.567156	0.18%
P-DNN-D2	CH_4	3712.361	354.0196	4066.00857	3.609168	0.862484	9.53%
	CO_2	24924.16	1507.761	26067.0875	4.416092	0.792417	4.59%
	N_2	4844.493	100.834	4733.66914	3.67519	0.924626	-2.29%
	O_2	7929.709	402.0509	7832.41445	3.893895	0.754249	-1.23%
	H_2	11111.7	461.5145	11813.6745	4.072385	0.628416	6.32%
P-RF-C	CH_4	260.9449	14.06993	164.128095	2.215182	n/a	-37.10%
	CO_2	2983.488	168.798	2713.03121	3.433454	n/a	-9.07%
	N_2	168.9471	11.53365	126.766191	2.103003	n/a	-24.97%
	O_2	595.2628	34.40819	514.335227	2.711246	n/a	-13.60%
	H_2	1849.185	70.50671	1645.7581	3.21636	n/a	-11.00%
P-RF-D	CH_4	184.9357	21.46799	118.081848	2.072183	n/a	-36.15%
	CO_2	2538.926	218.0743	2123.91166	3.327136	n/a	-16.35%
	N_2	130.5626	14.98025	92.1239731	1.964372	n/a	-29.44%
	O_2	606.7425	35.17128	413.90048	2.616895	n/a	-31.78%
	H_2	1893.513	73.42637	1791.57503	3.253235	n/a	-5.38%

Table S6. Comparison of gas permeabilities calculated from MD simulations with permeabilities predicted from ML models, with error and uncertainty quantification, of selected top polymer candidates. Units of permeability are Barrers.

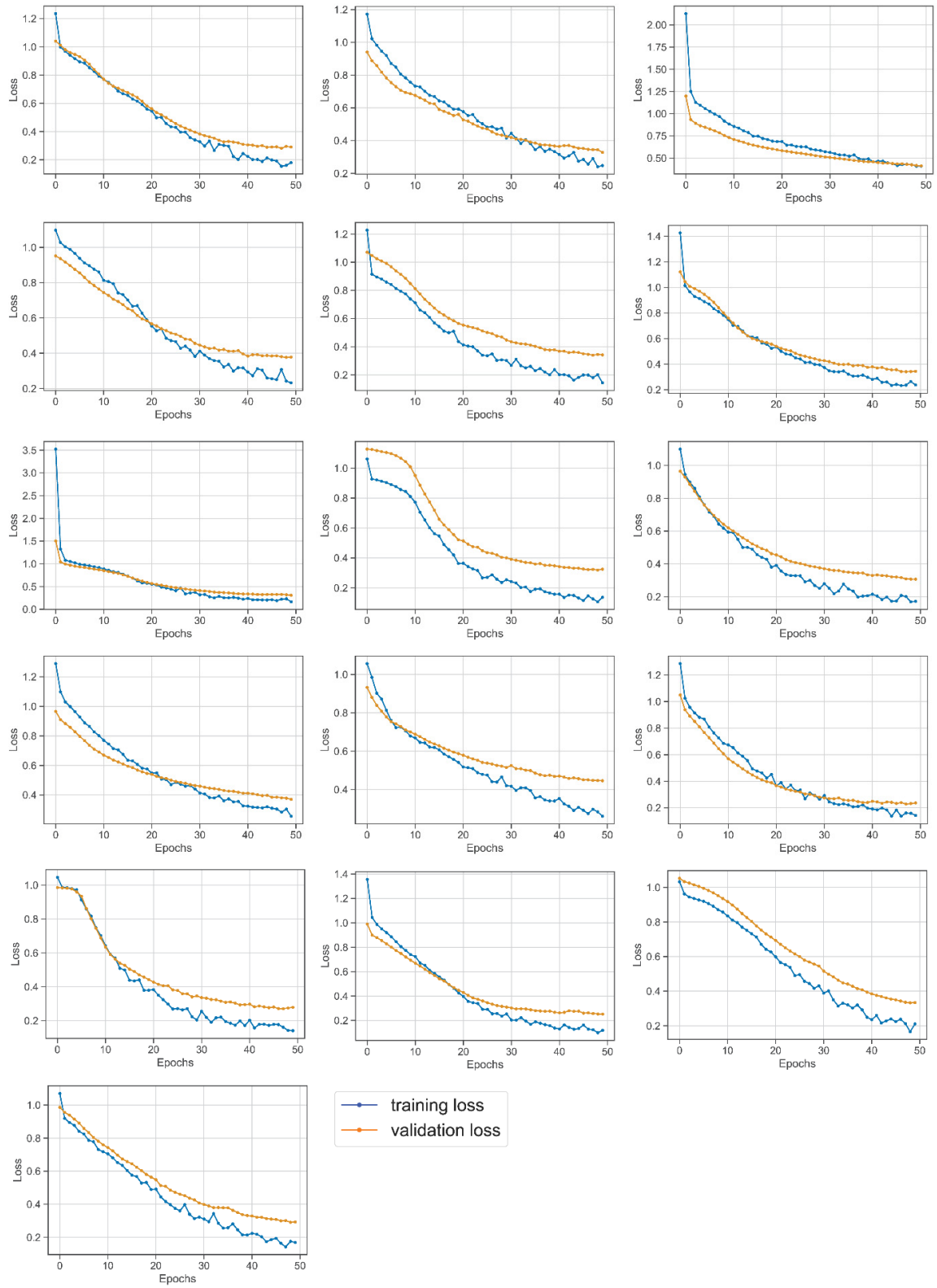


Fig. S11. Training model verification for the ensemble of 16 DNNs trained on MFFs and gas permeabilities imputed using BLR.

In-silico crosslinking of ladder polymers and polyimides

The multi-step cross-linking reaction for ladder polymers is exemplified by the case of PIM-1 (**Fig. S12(a)**). After 45 of each of the two reactants are packed into a 3D-periodic amorphous cell, there are 180 potential reaction sites in the total system (each monomer has four potential reaction sites). As we increase the cutoff distance from 4.5 to 9.5 Å step by step, the cross-linking degree increases from 33.3% to 92.8% (**Fig. S12(b)).** The final cross-linked system after relaxation contains 2501 atoms with a density of 1.07 g/cm3, and the box length is 31.87 Å (**Fig. S12(c)**). During the cross-linking process, extra hydrogen atoms are removed, and partial charges are updated to satisfy charge-neutrality. Then the generated cross-linked polymer structure is used for subsequent calculations of solubility and diffusivity. The ladder polymers in Dataset D of our work are subject to a similar cross-linking procedure, before further MD simulations.

The multi-step crosslinking strategy used for PIM-1 and other ladder polymers is also applied to novel polyimides, namely P-DNN-C1 through P-DNN-C4, whose case studies are shown in **Fig. S13-S16**, respectively. The difference for polyimides, compared to ladder polymers, is that after 45 of each component are packed into the 3D-periodic amorphous cell, there are 90 potential reaction sites in the total system (each monomer has two potential reaction sites). Differently, the reactive atoms for polyimides are carbon and nitrogen.

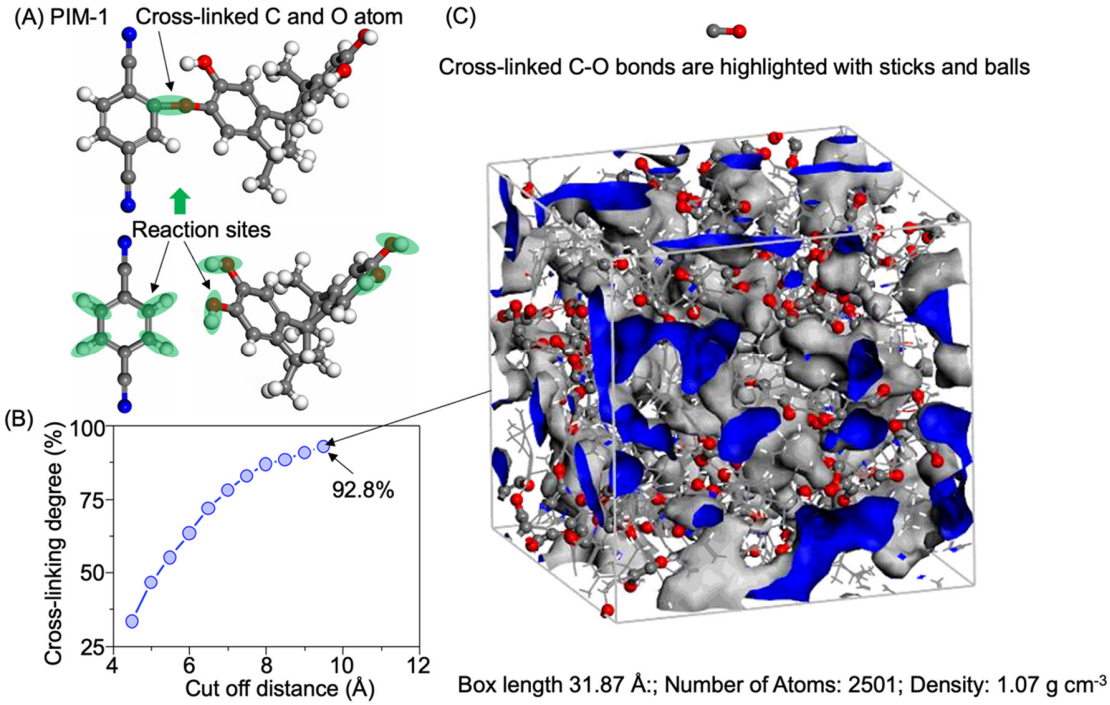


Fig. S12. Cross-linking steps for the generation of an atomistic model of PIM-1. (a) The reaction mechanism between 5,5',6,6'-tetrahydroxy-3,3,3',3'-tetramethyl-1,1'-spirobisindane (PubChem CID 66162) and Tetrafluoroterephthalonitrile (PubChem CID 104426). Carbon-oxygen bonds (C-O) are formed through the crosslinking of highlighted carbon and oxygen accompanied by the elimination of hydrogen molecules. (b) The cross-linking degree increases as the cutoff distance increases from 4.5 to 9.5 Å. The final cross-linking degree reaches 92.8%. (c) Snapshot of the final cross-linked network of PIM-1. Cross-linked C and O atoms are highlighted with sticks and balls, while other atoms are drawn as lines.

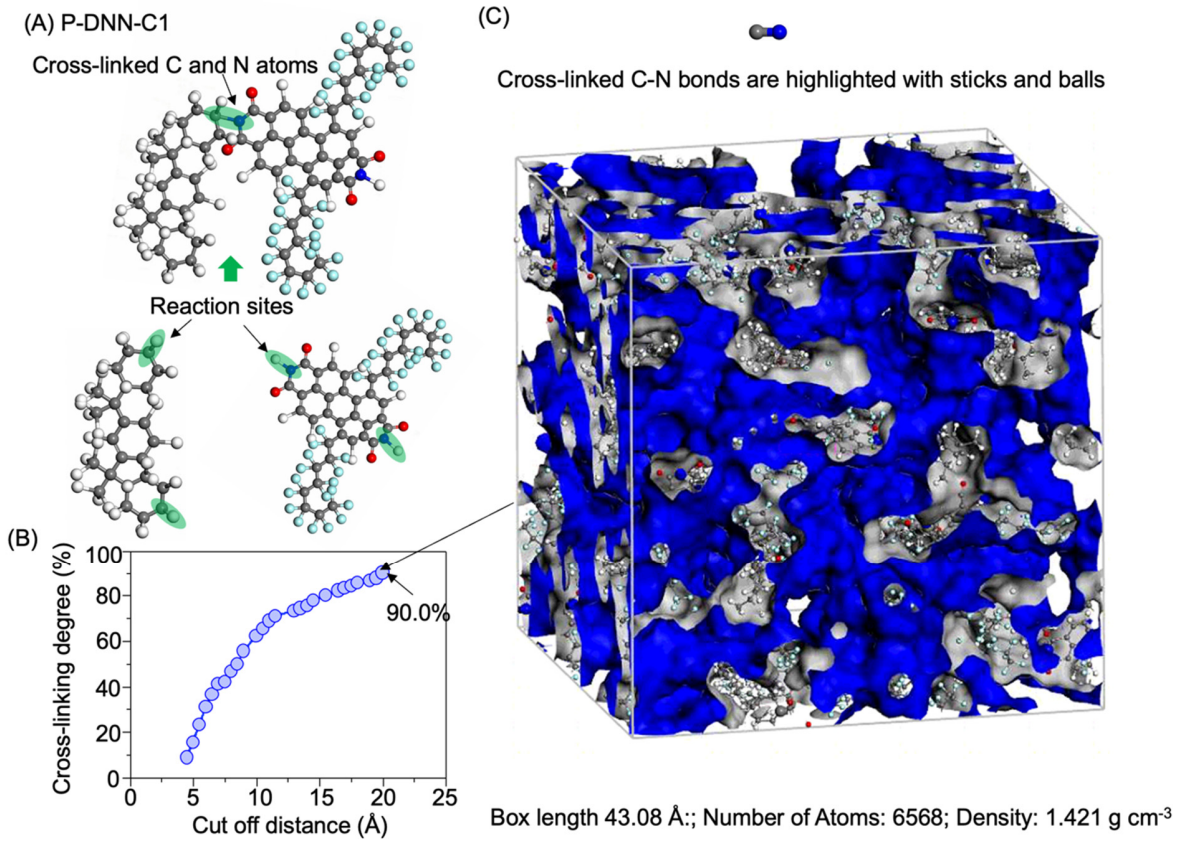


Fig. S13. Cross-linking steps for the generation of an atomistic model of P-DNN-C1. (a) The reaction mechanism between the diamine component '4-[2-[3-[2-(4-aminocyclohexyl)propan-2-yl]phenyl]propan-2-yl]cyclohexan-1-amine' (PubChem CID 54351456) and the dianhydride component '1,6-Bis(Heptadecafluorooctyl)-3,4,9,10-Perylenetetracarboxylic Acid 3,4:9,10-Dianhydride' (PubChem CID 46217647). The imide groups are formed through the crosslinking of highlighted carbons and nitrogens, accompanied by eliminating hydrogen molecules. (b) The cross-linking degree increases as the cutoff distance is increased from 4.5 to 20 Å. The final cross-linking degree reaches 90.0%. (c) Snapshot of the final cross-linked network of P-DNN-C1. Cross-linked C and N atoms are highlighted with sticks and balls, while other atoms are drawn as lines.

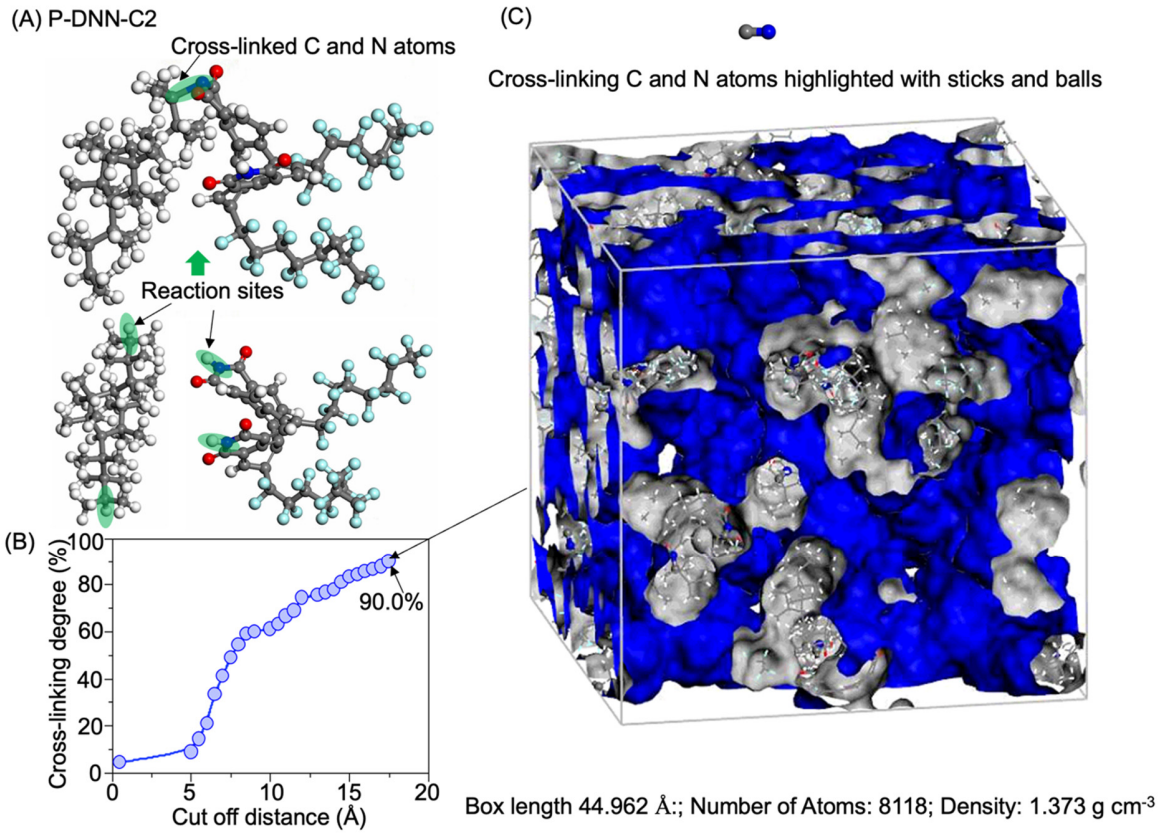


Fig. S14. Cross-linking steps for the generation of an atomistic model of P-DNN-C2. (a) The reaction mechanism between the diamine component '4-[4-(4-amino-2,3,3,4-tetramethylpentan-2-yl)-2,2,3,3,5,5,6,6-octamethylcyclohexyl]-2,3,3,4-tetramethylpentan-2-amine' (PubChem CID 118069048) and the dianhydride component '1,6-Bis(Heptadecafluorooctyl)-3,4,9,10-Perylenetetracarboxylic Acid 3,4:9,10-Dianhydride' (PubChem CID 46217647). The imide groups are formed through the crosslinking of highlighted carbons and nitrogens accompanied by eliminating hydrogen molecules. (b) The cross-linking degree increases as the cutoff distance is increased from 0.5 to 17.5 Å. The final cross-linking degree reaches 90.0%. (c) Snapshot of the final cross-linked network of P-DNN-C2. Cross-linked C and N atoms are highlighted with sticks and balls, while other atoms are drawn as lines.

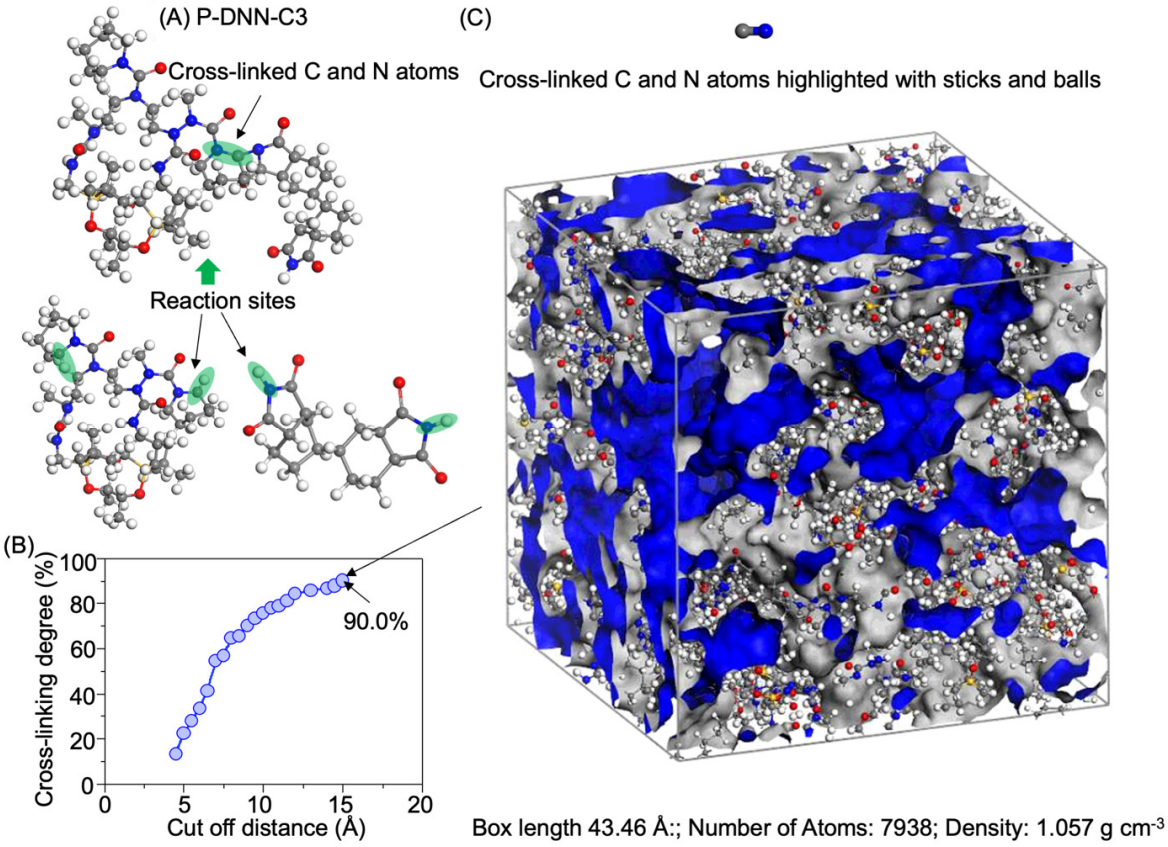


Fig. S15. Cross-linking steps for the generation of an atomistic model of P-DNN-C3. (a) The reaction mechanism between the diisocyanate component 'N-[1,17-ditert-butyl-5-ethyl-8-(2-isocyanatoazepane-1-carbonyl)-19,20-dimethyl-4,12-dioxo-18,21,23-trioxa-3,5,8,11,13-pentaza-1,17-disilabicyclo[15.4.3]tetracosan-11-yl]-2-isocyanato-N-methylazepane-1-carboxamide' (PubChem CID 60139559) and the dianhydride component 'Dodecahydro-[5,5'-biisobenzofuran]-1,1',3,3'-tetraone' (PubChem CID 11077626). The imide groups are formed through the crosslinking of highlighted carbons and nitrogens, accompanied by eliminating hydrogen molecules. (b) The cross-linking degree increases with the increase of the cutoff distance from 4.5 to 15 Å. The final cross-linking degree reaches 90.0%. (c) Snapshot of the final cross-linked network of P-DNN-C3. Cross-linked C and N atoms are highlighted with sticks and balls, while other atoms are drawn as lines.

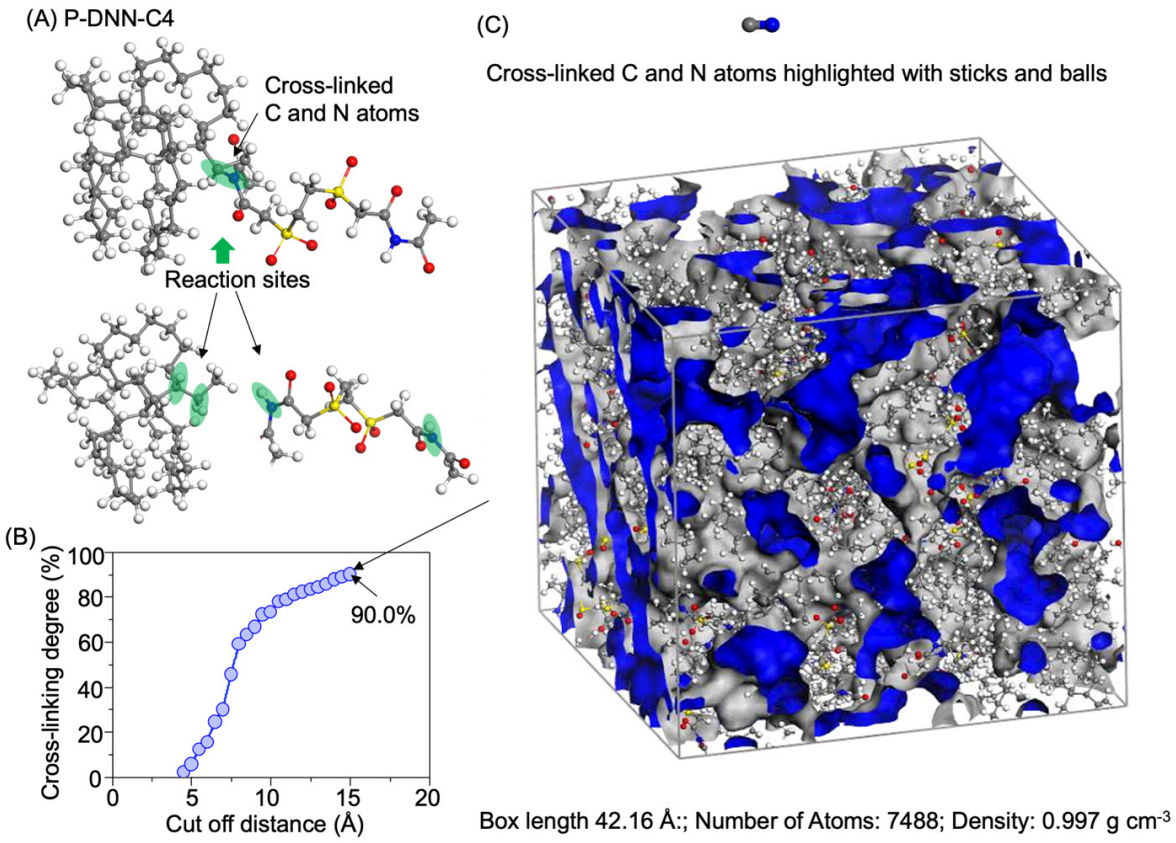
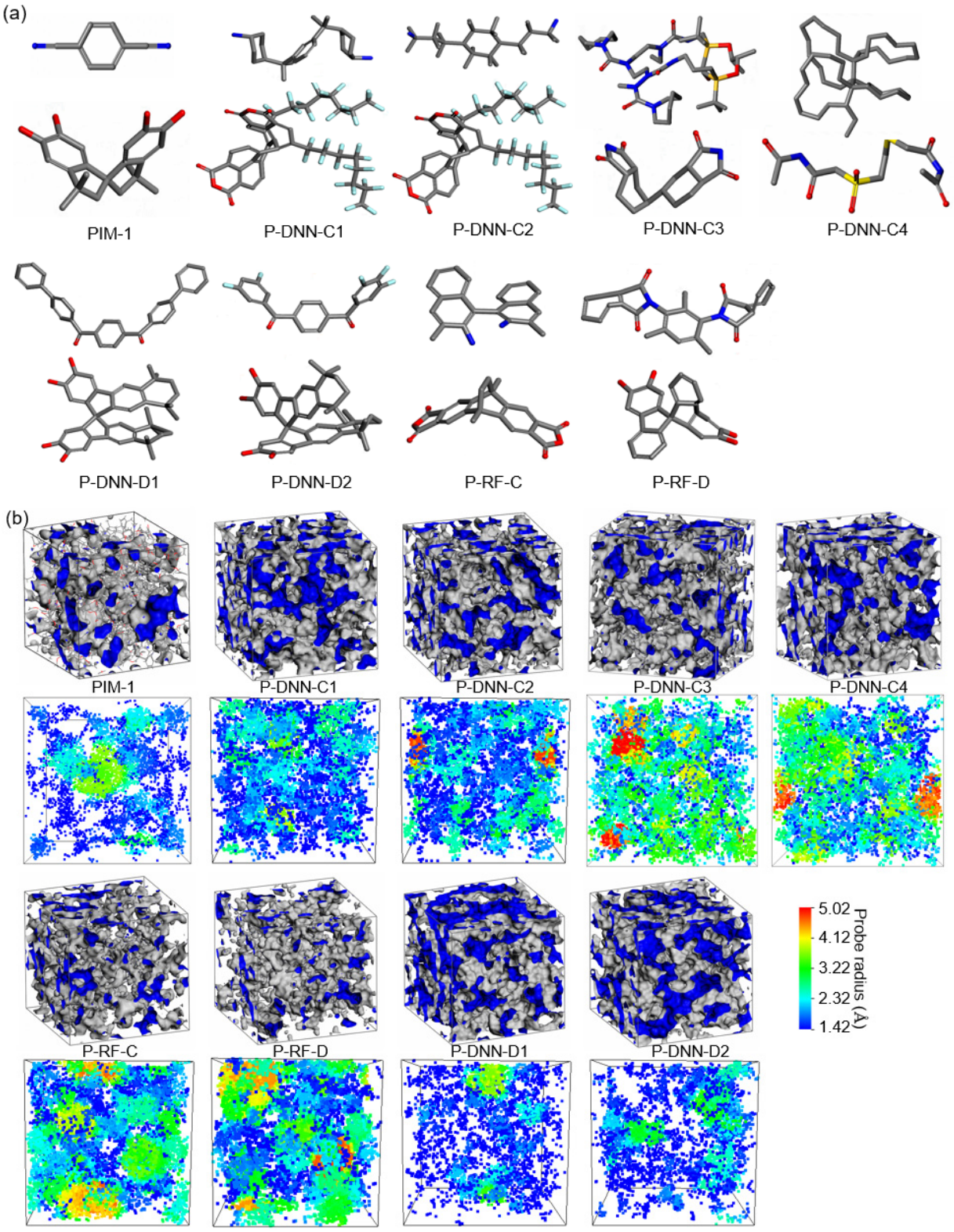


Fig. S16. Cross-linking steps for the generation of an atomistic model of P-DNN-C4. (a) The reaction mechanism between the diisocyanate component '14-Isocyanato-14-(1-isocyanatoethyl)trispiro [10.0.0.1113.1012.1011] tetratetracontane' (PubChem CID 141382908) and the dianhydride component '2,2'-[Ethylenebis(sulfonyl)]bis(acetic acid)diacetic acid dianhydride' (PubChem CID 101718050). The imide groups are formed through the crosslinking of highlighted carbons and nitrogens, accompanied by eliminating hydrogen molecules. (b) The cross-linking degree increases with the increase of the cutoff distance from 4.5 to 15 Å. The final cross-linking degree reaches 90.0%. (c) Snapshot of the final cross-linked network of P-DNN-C4. Cross-linked C and N atoms are highlighted with sticks and balls, while other atoms are drawn as lines.


(a)
PIM-1
P-DNN-C1
P-DNN-C2
P-DNN-C3
P-DNN-C4
P-DNN-D1
P-DNN-D2
P-RF-C
P-RF-D
(b)
PIM-1
P-DNN-C1
P-DNN-C2
P-DNN-C3
P-DNN-C4
P-RF-C
P-RF-D
P-DNN-D1
P-DNN-D2
5.02
4.12
3.22
2.32
1.42
Probe radius (Å)

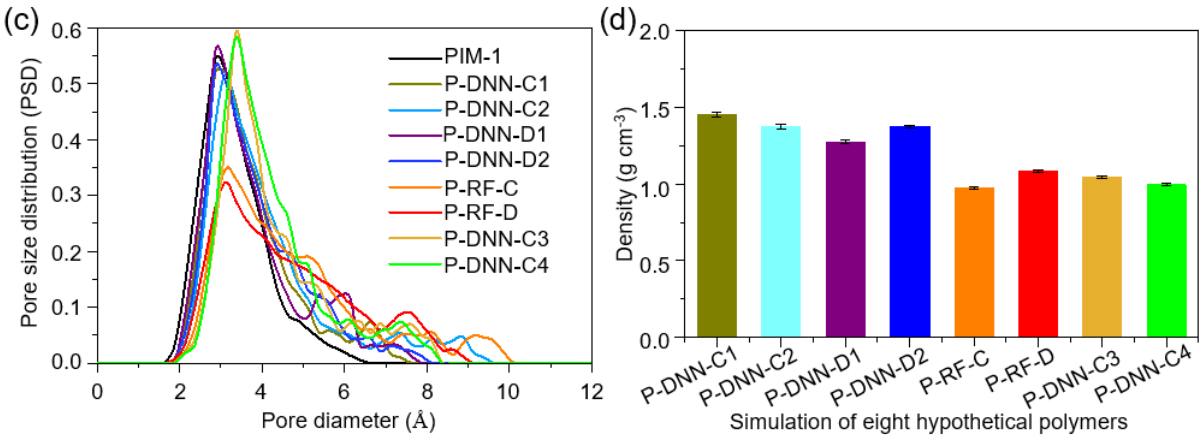


Fig. S17. Microstructure analysis of the eight selected candidate polymer models, with the benchmark of PIM-1 shown for comparison. (a) Three-dimensional molecular structures of monomers used to form crosslinked polymers. (b) Top row shows the free-volume elements, including interconnected and disconnected voids, with respect to a probe radius of 0.80 Å. Bottom row shows voids colored with respect to the largest radius probe that can be inserted. Here, only those voids having a minimum radius of 1.42 Å (the kinetic radius of H_2) are shown. (c) Pore size distribution of the polymers, calculated by averaging across five models for each polymer. (d) The calculated density values of the polymers. P-DNN-C1, P-DNN-C2, P-DNN-D1, and P-DNN-D2 have relatively higher densities, while P-RF-C and P-RF-D, P-DNN-C3 and P-DNN-C4 have relatively lower densities compared with PIM-1 (Fig. S12). For void analysis, we apply the Poreblazer code (https://github.com/SarkisovGroup/PoreBlazer) to directly quantify pore size, and its distribution and interconnectivity, which is based on the Hoshen-Kopelman cluster labeling algorithm. The radius of the probe is chosen to be 1.42 Å, representing the mean Van der Waals radius of hydrogen gas molecules.

	Gas	[cm^3 (STP)/cm^3 (polymer bar)]			[10^{-8} cm^2/s]			Barrer	
		S_{sim}		S_{exp}[6]	D_{sim}		D_{exp}[6]	Permeability	
		Our work	Ref.[7]		Our work	Ref.[7]		Our work	P_{exp}[6]
PIM-1	CH_4	15.1±1.91	14.2±3.1	13.70	8.05±0.23	112±27	6.80	155.147±10.71	121.108
	CO_2	49.93±3.41	50.7±9	66.90	28.45±4.19	151±47	26.00	2150.966±200.718	2261.22
	H_2	0.42±0.06	0.46±0.03	0.58	1826.33±43.33	6630±215	1700.00	983.104±121.417	1281.8
	N_2	7.31±0.89	-	-	31.60±3.51	-	-	289.318±26.73	-
	O_2	3.93±0.68	4.11±0.58	3.50	82.31±5.82	452±81	81.00	400.507±56.293	368.55

Table S7. Benchmark MD simulations for the PIM-1 polymer membrane with simulated solubility coefficients [cm^3 (STP)/cm3 (polymer) bar], diffusion coefficients [10^{-8} cm^2/s] at 300K, and permeabilities in Barrer. The values from our study show good agreement with simulation and experimental values in the literature.

References

1. Liaw, D.-J. *et al.* Advanced polyimide materials: Syntheses, physical properties and applications. *Prog. Polym. Sci.* **37**, 907–974 (2012).
2. Paraskevopoulou, P., Chriti, D., Raptopoulos, G. & Anyfantis, G. C. Synthetic Polymer Aerogels in Particulate Form. *Materials* **12**, 1543 (2019).
3. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
4. Du, N., Guiver, M. D. & Robertson, G. P. Ladder polymers with intrinsic microporosity and process for production thereof. (2016).
5. Hall, L. H., Mohney, Brian. & Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **31**, 76–82 (1991).
6. Budd, P. M. *et al.* Gas separation membranes from polymers of intrinsic microporosity. *J. Membr. Sci.* **251**, 263–269 (2005).
7. Fang, W., Zhang, L. & Jiang, J. Polymers of intrinsic microporosity for gas permeation: a molecular simulation study. *Mol. Simul.* **36**, 992–1003 (2010).