

Supplementary Information

Using GPT-4 in Parameter Selection of Polymer Informatics: Improving Predictive Accuracy Amidst Data Scarcity and 'Ugly Duckling' Dilemma

Kan Hatakeyama-Sato,* Seigo Watanabe, Naoki Yamanec, Yasuhiko Igarashi, and Kenichi Oyaizu*

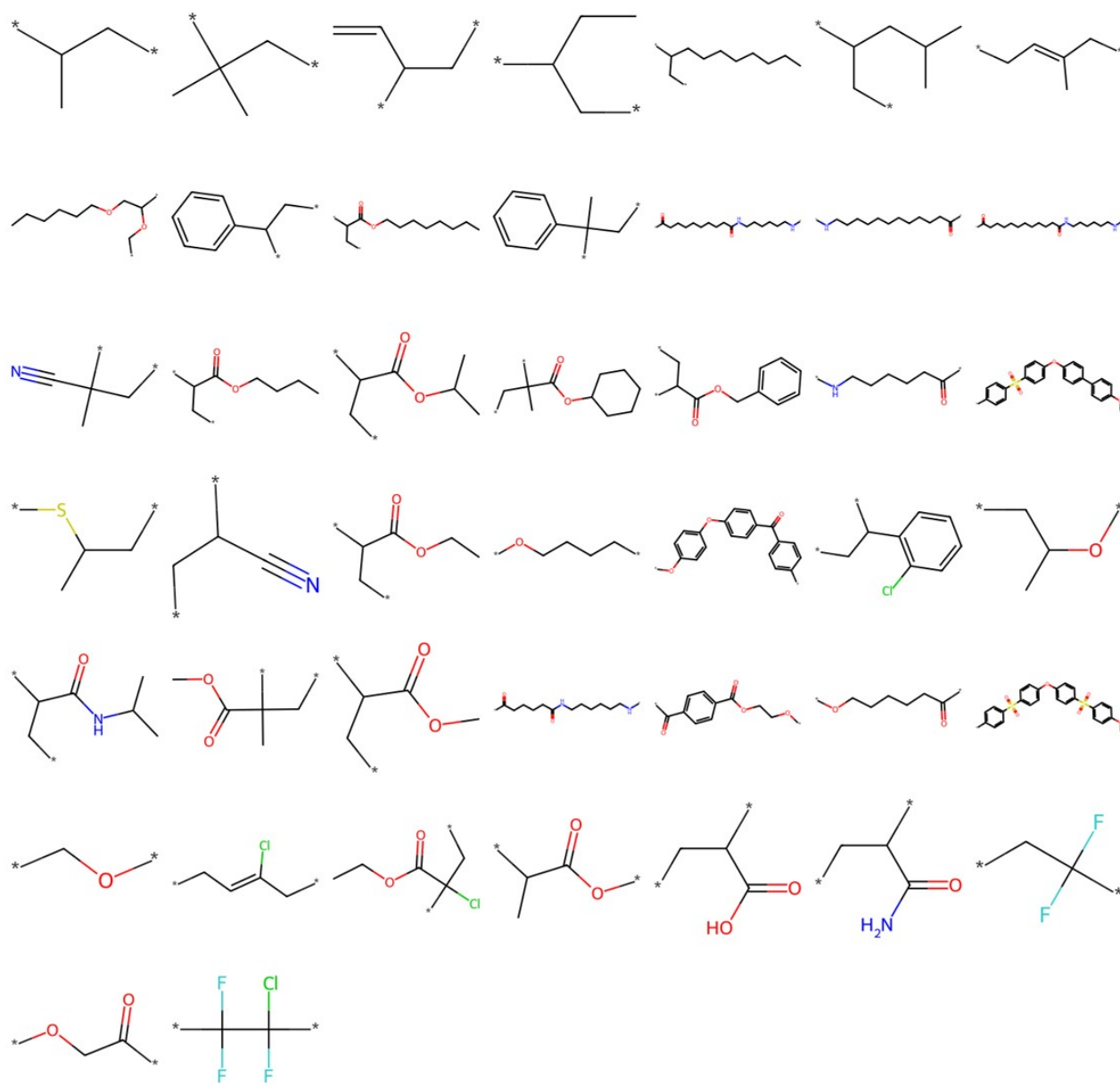


Fig. S1 Chemical structures recorded in the polymer database.

Table S 1 List of molecular descriptors employed in this study.

PM7_energy	rdkit_MaxEStateIndex	rdkit_Kappa1	rdkit_EState_VSA8	rdkit_fr_N_O	rdkit_fr_phos_ester
PM7_dipoleX	rdkit_MinEStateIndex	rdkit_Kappa2	rdkit_EState_VSA9	rdkit_fr_Ndealkylation1	rdkit_fr_piperidine
PM7_dipoleY	rdkit_MaxAbsEStateIndex	rdkit_Kappa3	rdkit_VSA_EState1	rdkit_fr_Ndealkylation2	rdkit_fr_piperzine
PM7_dipoleZ	rdkit_MinAbsEStateIndex	rdkit_LabuteASA	rdkit_VSA_EState10	rdkit_fr_Nhpyrrole	rdkit_fr_priamide
PM7_dipoleTot	rdkit_qed	rdkit_PEOE_VSA1	rdkit_VSA_EState2	rdkit_fr_SH	rdkit_fr_prisulfonamd
PM7_HOMO	rdkit_MolWt	rdkit_PEOE_VSA10	rdkit_VSA_EState3	rdkit_fr_aldehyde	rdkit_fr_pyridine

PM7_LUMO	rdkit_HeavyAtomMolWt	rdkit_PEOE_VSA11	rdkit_VSA_EState4	rdkit_fr_alkyl_carbamate	rdkit_fr_quatN
PM7_alpha656nm	rdkit_ExactMolWt	rdkit_PEOE_VSA12	rdkit_VSA_EState5	rdkit_fr_alkyl_halide	rdkit_fr_sulfide
JR_BoilingPoint	rdkit_NumValenceElectrons	rdkit_PEOE_VSA13	rdkit_VSA_EState6	rdkit_fr_allylic_oxid	rdkit_fr_sulfonamd
JR_MeltingPoint	rdkit_NumRadicalElectrons	rdkit_PEOE_VSA14	rdkit_VSA_EState7	rdkit_fr_amide	rdkit_fr_sulfone
JR_CriticalTemp	rdkit_MaxPartialCharge	rdkit_PEOE_VSA2	rdkit_VSA_EState8	rdkit_fr_amidine	rdkit_fr_term_acetylene
JR_CriticalPress	rdkit_MinPartialCharge	rdkit_PEOE_VSA3	rdkit_VSA_EState9	rdkit_fr_aniline	rdkit_fr_tetrazole
JR_CriticalVolume	rdkit_MaxAbsPartialCharge	rdkit_PEOE_VSA4	rdkit_FractionCSP3	rdkit_fr_aryl_methyl	rdkit_fr_thiazole
JR_EnthalpyForm	rdkit_MinAbsPartialCharge	rdkit_PEOE_VSA5	rdkit_HeavyAtomCount	rdkit_fr_azide	rdkit_fr_thiocyan
JR_GibbsEnergy	rdkit_FpDensityMorgan1	rdkit_PEOE_VSA6	rdkit_NHOHCount	rdkit_fr_azo	rdkit_fr_thiophene
JR_HeatCapacity	rdkit_FpDensityMorgan2	rdkit_PEOE_VSA7	rdkit_NOCCount	rdkit_fr_barbitur	rdkit_fr_unbrch_alkane
JR_EnthalpyVap	rdkit_FpDensityMorgan3	rdkit_PEOE_VSA8	rdkit_NumAliphaticCarbocycles	rdkit_fr_benzene	rdkit_fr_urea
JR_EnthalpyFus	rdkit_BCUT2D_MWHI	rdkit_PEOE_VSA9	rdkit_NumAliphaticHeterocycles	rdkit_fr_benzodiazepine	
JR_LiquidVisco	rdkit_BCUT2D_MWLOW	rdkit_SMR_VSA1	rdkit_NumAliphaticRings	rdkit_fr_bicyclic	
JR_CrystalSolub_1	rdkit_BCUT2D_CHGHI	rdkit_SMR_VSA10	rdkit_NumAromaticCarbocycles	rdkit_fr_diazo	
JR_CrystalSolub_2	rdkit_BCUT2D_CHGLO	rdkit_SMR_VSA2	rdkit_NumAromaticHeterocycles	rdkit_fr_dihydropyridine	
JR_AmorphSolub_1	rdkit_BCUT2D_LOGPHI	rdkit_SMR_VSA3	rdkit_NumAromaticRings	rdkit_fr_epoxide	
JR_AmorphSolub_2	rdkit_BCUT2D_LOGPLOW	rdkit_SMR_VSA4	rdkit_NumHAcceptors	rdkit_fr_ester	
DFTMD_vol	rdkit_BCUT2D_MRHI	rdkit_SMR_VSA5	rdkit_NumHDonors	rdkit_fr_ether	
rdkit_vol	rdkit_BCUT2D_MRLOW	rdkit_SMR_VSA6	rdkit_NumHeteroatoms	rdkit_fr_furan	
	rdkit_BalabanJ	rdkit_SMR_VSA7	rdkit_NumRotatableBonds	rdkit_fr_guanido	
	rdkit_BertzCT	rdkit_SMR_VSA8	rdkit_NumSaturatedCarbocycles	rdkit_fr_halogen	
	rdkit_Chi0	rdkit_SMR_VSA9	rdkit_NumSaturatedHeterocycles	rdkit_fr_hdrzine	
	rdkit_Chi0n	rdkit_SlogP_VSA1	rdkit_NumSaturatedRings	rdkit_fr_hdrzone	
	rdkit_Chi0v	rdkit_SlogP_VSA10	rdkit_RingCount	rdkit_fr_imidazole	
	rdkit_Chi1	rdkit_SlogP_VSA11	rdkit_MolLogP	rdkit_fr_imide	
	rdkit_Chi1n	rdkit_SlogP_VSA12	rdkit_MolMR	rdkit_fr_isocyan	
	rdkit_Chi1v	rdkit_SlogP_VSA2	rdkit_fr_Al_COO	rdkit_fr_isothiocyan	
	rdkit_Chi2n	rdkit_SlogP_VSA3	rdkit_fr_Al_OH	rdkit_fr_ketone	
	rdkit_Chi2v	rdkit_SlogP_VSA4	rdkit_fr_Al_OH_noTert	rdkit_fr_ketone_Topless	
	rdkit_Chi3n	rdkit_SlogP_VSA5	rdkit_fr_ArN	rdkit_fr_lactam	
	rdkit_Chi3v	rdkit_SlogP_VSA6	rdkit_fr_Ar_COO	rdkit_fr_lactone	
	rdkit_Chi4n	rdkit_SlogP_VSA7	rdkit_fr_Ar_N	rdkit_fr_methoxy	
	rdkit_Chi4v	rdkit_SlogP_VSA8	rdkit_fr_Ar_NH	rdkit_fr_morpholine	
	rdkit_HallKierAlpha	rdkit_SlogP_VSA9	rdkit_fr_Ar_OH	rdkit_fr_nitrile	
	rdkit_Ipc	rdkit_TPSA	rdkit_fr_COO	rdkit_fr_nitro	
		rdkit_EState_VSA1	rdkit_fr_COO2	rdkit_fr_nitro_arom	
		rdkit_EState_VSA10	rdkit_fr_C_O	rdkit_fr_nitro_arom_nonortho	

		rdkit_EState_VSA11	rdkit_fr_C_O_noCOO	rdkit_fr_nitroso
		rdkit_EState_VSA2	rdkit_fr_C_S	rdkit_fr_oxazole
		rdkit_EState_VSA3	rdkit_fr_HOCCN	rdkit_fr_oxime
		rdkit_EState_VSA4	rdkit_fr_Imine	rdkit_fr_para_hydroxylation
		rdkit_EState_VSA5	rdkit_fr_NH0	rdkit_fr_phenol
		rdkit_EState_VSA6	rdkit_fr_NH1	rdkit_fr_phenol_noOrthoHbond
		rdkit_EState_VSA7	rdkit_fr_NH2	rdkit_fr_phos_acid

Table S 2 Molecular descriptor sets for Random-10 and 20.

Random-10	Random-20
rdkit_fr_urea	rdkit_fr_phos_ester
rdkit_EState_VSA6	rdkit_fr_thiazole
rdkit_fr_oxazole	rdkit_SMR_VSA9
JR_LiquidVisco	rdkit_NumHAcceptors
rdkit_VSA_EState5	rdkit_EState_VSA1
rdkit_MinEStateIndex	rdkit_fr_NH2
rdkit_PEOE_VSA8	JR_CrystalSolub_1
rdkit_MolLogP	PM7MD_vol
rdkit_NumHeteroatoms	rdkit_PEOE_VSA10
rdkit_VSA_EState10	rdkit_RingCount
	rdkit_BertzCT
	rdkit_SMR_VSA4
	rdkit_fr_nitroso
	rdkit_FpDensityMorgan2
	rdkit_fr_allylic_oxid
	rdkit_fr_prisulfonamd
	rdkit_PEOE_VSA6
	rdkit_fr_Ar_COO
	rdkit_fr_imide
	rdkit_fr_sulfide

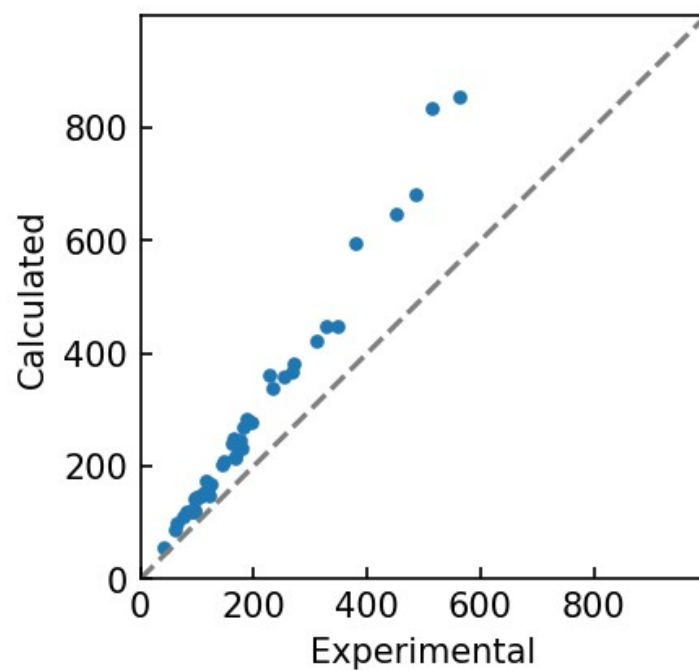


Fig. S2 Relationship between calculated and experimental molecular volume (angstrom unit). Calculations were conducted by Matlantis. Experimental values were derived from actual density and molecular weight.

Chat log between the authors and GPT-4.

Note: In the manuscript, we use the notation PM7_*** to denote parameters calculated using the semi-empirical method PM7. However, in our interactions with GPT-4, we conveniently treated these as if they were calculated using the more widely known method, Density Functional Theory (DFT).

(1) Parameter selection

Question by Author

Tasks

1: consider theoretical equation of refractive index of molecules

2: select essential descriptors from below by yourself

Especially, I want to predict refractive index of aliphatic and aromatic polymers.

#Followings are list of descriptors

DFT_energy,DFT_dipoleX,DFT_dipoleY,DFT_dipoleZ,DFT_dipoleTot,DFT_HOMO,DFT_LUMO,DFT_alp
ha656nm,rdkit_vol,rdkit_MaxEStateIndex,rdkit_MinEStateIndex,rdkit_MaxAbsEStateIndex,rdkit_MinAbsES
tateIndex,rdkit_qed,rdkit_MolWt,rdkit_HeavyAtomMolWt,rdkit_ExactMolWt,rdkit_NumValenceElectrons,r
dkit_NumRadicalElectrons,rdkit_MaxPartialCharge,rdkit_MinPartialCharge,rdkit_MaxAbsPartialCharge,rdki
t_MinAbsPartialCharge,rdkit_FpDensityMorgan1,rdkit_FpDensityMorgan2,rdkit_FpDensityMorgan3,rdkit_B
CUT2D_MWHI,rdkit_BCUT2D_MWLOW,rdkit_BCUT2D_CHGHI,rdkit_BCUT2D_CHGLO,rdkit_BCUT2
D_LOGPHI,rdkit_BCUT2D_LOGPLOW,rdkit_BCUT2D_MRHI,rdkit_BCUT2D_MRLOW,rdkit_BalabanJ,r
dkit_BertzCT,rdkit_Chi0,rdkit_Chi0n,rdkit_Chi0v,rdkit_Chi1,rdkit_Chi1n,rdkit_Chi1v,rdkit_Chi2n,rdkit_Chi
2v,rdkit_Chi3n,rdkit_Chi3v,rdkit_Chi4n,rdkit_Chi4v,rdkit_HallKierAlpha,rdkit_Ipc,rdkit_Kappa1,rdkit_Kap
pa2,rdkit_Kappa3,rdkit_LabuteASA,rdkit_PEOE_VSA1,rdkit_PEOE_VSA10,rdkit_PEOE_VSA11,rdkit_PE
OE_VSA12,rdkit_PEOE_VSA13,rdkit_PEOE_VSA14,rdkit_PEOE_VSA2,rdkit_PEOE_VSA3,rdkit_PEOE_
VSA4,rdkit_PEOE_VSA5,rdkit_PEOE_VSA6,rdkit_PEOE_VSA7,rdkit_PEOE_VSA8,rdkit_PEOE_VSA9,rd
kit_SMR_VSA1,rdkit_SMR_VSA10,rdkit_SMR_VSA2,rdkit_SMR_VSA3,rdkit_SMR_VSA4,rdkit_SMR_V
SA5,rdkit_SMR_VSA6,rdkit_SMR_VSA7,rdkit_SMR_VSA8,rdkit_SMR_VSA9,rdkit_SlogP_VSA1,rdkit_Sl
ogP_VSA10,rdkit_SlogP_VSA11,rdkit_SlogP_VSA12,rdkit_SlogP_VSA2,rdkit_SlogP_VSA3,rdkit_SlogP_
VSA4,rdkit_SlogP_VSA5,rdkit_SlogP_VSA6,rdkit_SlogP_VSA7,rdkit_SlogP_VSA8,rdkit_SlogP_VSA9,rdk
it_TPSA,rdkit_EState_VSA1,rdkit_EState_VSA10,rdkit_EState_VSA11,rdkit_EState_VSA2,rdkit_EState_V
SA3,rdkit_EState_VSA4,rdkit_EState_VSA5,rdkit_EState_VSA6,rdkit_EState_VSA7,rdkit_EState_VSA8,rd
kit_EState_VSA9,rdkit_VSA_EState1,rdkit_VSA_EState10,rdkit_VSA_EState2,rdkit_VSA_EState3,rdkit_V
SA_EState4,rdkit_VSA_EState5,rdkit_VSA_EState6,rdkit_VSA_EState7,rdkit_VSA_EState8,rdkit_VSA_ES
tate9,rdkit_FractionCSP3,rdkit_HeavyAtomCount,rdkit_NHOHCount,rdkit_NOCCount,rdkit_NumAliphaticCa

rbocycles,rdkit_NumAliphaticHeterocycles,rdkit_NumAliphaticRings,rdkit_NumAromaticCarbocycles,rdkit_NumAromaticHeterocycles,rdkit_NumAromaticRings,rdkit_NumHAcceptors,rdkit_NumHDonors,rdkit_NumHeteroatoms,rdkit_NumRotatableBonds,rdkit_NumSaturatedCarbocycles,rdkit_NumSaturatedHeterocycles,rdkit_NumSaturatedRings,rdkit_RingCount,rdkit_MolLogP,rdkit_MolMR,rdkit_fr_Al_COO,rdkit_fr_Al_OH,rdkit_fr_Al_OH_noTert,rdkit_fr_ArN,rdkit_fr_Ar_COO,rdkit_fr_Ar_N,rdkit_fr_Ar_NH,rdkit_fr_Ar_OH,rdkit_fr_COO,rdkit_fr_COO2,rdkit_fr_C_O,rdkit_fr_C_O_noCOO,rdkit_fr_C_S,rdkit_fr_HOCCN,rdkit_fr_Imine,rdkit_fr_NH0,rdkit_fr_NH1,rdkit_fr_NH2,rdkit_fr_N_O,rdkit_fr_Ndealkylation1,rdkit_fr_Ndealkylation2,rdkit_fr_Nhpyrrole,rdkit_fr_SH,rdkit_fr_aldehyde,rdkit_fr_alkyl_carbamate,rdkit_fr_alkyl_halide,rdkit_fr_allylic_oxid,rdkit_fr_amide,rdkit_fr_amidine,rdkit_fr_aniline,rdkit_fr_aryl_methyl,rdkit_fr_azide,rdkit_fr_azo,rdkit_fr_barbitur,rdkit_fr_benzene,rdkit_fr_benzodiazepine,rdkit_fr_bicyclic,rdkit_fr_diazo,rdkit_fr_dihydropyridine,rdkit_fr_epoxide,rdkit_fr_ester,rdkit_fr_ether,rdkit_fr_furan,rdkit_fr_guanido,rdkit_fr_halogen,rdkit_fr_hdrzine,rdkit_fr_hdrzone,rdkit_fr_imidazole,rdkit_fr_imide,rdkit_fr_isocyan,rdkit_fr_isothiocyan,rdkit_fr_ketone,rdkit_fr_ketone_Topliss,rdkit_fr_lactam,rdkit_fr_lactone,rdkit_fr_methoxy,rdkit_fr_morpholine,rdkit_fr_nitrile,rdkit_fr_nitro,rdkit_fr_nitro_arom,rdkit_fr_nitro_arom_nonortho,rdkit_fr_nitroso,rdkit_fr_oxazole,rdkit_fr_oxime,rdkit_fr_para_hydroxylation,rdkit_fr_phenol,rdkit_fr_phenol_noOrthoHbond,rdkit_fr_phos_acid,rdkit_fr_phos_ester,rdkit_fr_piperdine,rdkit_fr_piperzine,rdkit_fr_priamide,rdkit_fr_prisulfonamd,rdkit_fr_pyridine,rdkit_fr_quatN,rdkit_fr_sulfide,rdkit_fr_sulfonamd,rdkit_fr_sulfone,rdkit_fr_term_acetylene,rdkit_fr_tetrazole,rdkit_fr_thiazole,rdkit_fr_thiocyan,rdkit_fr_thiophene,rdkit_fr_unbrch_alkane,rdkit_fr_urea,JR_BoilingPoint,JR_MeltingPoint,JR_CriticalTemp,JR_CriticalPress,JR_CriticalVolume,JR_EnthalpyForm,JR_GibbsEnergy,JR_HeatCapacity,JR_EnthalpyVap,JR_EnthalpyFus,JR_LiquidVisco,JR_CrystalSolub_1,JR_CrystalSolub_2,JR_AmorphSolub_1,JR_AmorphSolub_2,DFTMD_vol

[explanation] DFT_*: parameters calculated by DFT

rdkit_: parameters calculated by RDKit

JR_: parameters estimated by group contribution method

DFTMD_vol: molecular volume estimated by DFT-MD simulation

3: output the descriptors as list data.

Answer by GPT-4

Task 1:

The refractive index of a material is the ratio of the speed of light in a vacuum to the speed of light in that material. In general, the refractive index of molecules can be theoretically described by the Lorentz-Lorenz equation (also known as the Clausius-Mossotti relation), which is:

$$n^2 - 1 / n^2 + 2 = 4\pi/3 * N * \alpha$$

where:

- n is the refractive index
- N is the number density of the molecules (number of molecules per unit volume)
- α is the polarizability of the molecules

The polarizability describes the ability of a molecule to form instantaneous dipoles, which is directly linked to the electronic structure of the molecules. Therefore, descriptors that reflect the electronic structure, size and shape of the molecules should be relevant for predicting the refractive index.

Task 2 & 3:

Based on the general principles above, the most relevant descriptors from the provided list for predicting the refractive index of aliphatic and aromatic polymers would be:

1. DFT_energy - Reflects overall energy of the system that could correlate with polarizability.
2. DFT_HOMO - Energy of the Highest Occupied Molecular Orbital, related to electron distribution.
3. DFT_LUMO - Energy of the Lowest Unoccupied Molecular Orbital, also related to electron distribution.
4. DFT_dipoleTot - Total dipole moment, directly related to polarizability.
5. rdkit_MolWt - Molecular weight, which could relate to size of the molecule.
6. rdkit_NumValenceElectrons - The number of valence electrons, related to the bonding and potential polarizability.
7. rdkit_NumAromaticRings - Reflects presence of aromatic systems which could alter the refractive index.
8. rdkit_NumAliphaticRings - Reflects presence of aliphatic rings which could alter the refractive index.
9. DFTMD_vol - Molecular volume estimated by DFT-MD simulation, could correlate with size of the molecule.
10. rdkit_HeavyAtomMolWt - Weight of the molecule without hydrogen atoms, could be a measure of molecular size.
11. rdkit_MaxAbsPartialCharge - Could be linked to polarizability.
12. rdkit_MolLogP - Describes the lipophilicity of the molecule which could correlate with polarizability.

13. JR_BoilingPoint - Boiling point can reflect intermolecular forces and thus polarizability.
14. JR_EnthalpyForm - Enthalpy of formation, reflects the stability of the molecule, may correlate with polarizability.

Output list:

```
["DFT_energy", "DFT_HOMO", "DFT_LUMO", "DFT_dipoleTot", "rdkit_MolWt",  
"rdkit_NumValenceElectrons", "rdkit_NumAromaticRings", "rdkit_NumAliphaticRings", "DFTMD_vol",  
"rdkit_HeavyAtomMolWt", "rdkit_MaxAbsPartialCharge", "rdkit_MolLogP", "JR_BoilingPoint",  
"JR_EnthalpyForm"]
```

Please note, however, that the selection of descriptors would ideally be done based on a more detailed understanding of the specific polymers under consideration and the performance of these descriptors in initial models. Additionally, a machine learning model could be employed to handle feature selection and automatically pick out the most informative descriptors.

Discussion about the selected parameters

1. Energy

GPT-4 selected the "energy" parameter, based on the rationale that it reflects the overall energy of the system, which could correlate with polarizability. While molecular energy does not explicitly appear in the Lorentz-Lorenz equation that defines the refractive index, it is one of the most fundamental molecular properties and can have relationships with various physical parameters.^{49, 57} Therefore, it is often chosen in data science research and is a reasonable choice for a predictive variable.

2,3. HOMO, LUMO

GPT-4 selected the "HOMO, LUMO energy" parameters, reasoning they are related to electron distribution. They are not direct variables of polarizability or molecular volume. However, like energy, they are very general variables that characterize molecules and are standard parameters in machine learning, making them reasonable choices.^{49, 57} In the world of refraction, a phenomenon known as anomalous refraction,⁵⁸ which originates from molecular absorption, is known. Absorption wavelengths are greatly influenced by HOMO, LUMO, thus, adding these two parameters could be useful in predicting refractive index. However, it's not entirely clear whether GPT-4 made this parameter selection considering the elements of anomalous refraction.

4. DipoleTot

"DipoleTot" is a parameter that represents the total dipole moment of the molecule. It strongly correlates with polarizability, thus, as GPT-4 argued, it is a valid predictor.

5. MolWt

"MolWt" is a parameter representing molecular weight. Although it doesn't directly correlate with polarizability or volume, it was chosen for its potential indirect influence. In simulations, errors may occur in parameter calculations due to molecular size,⁵⁹ and incorporating molecular weight as a predictor can potentially rectify these influences.

6. rdkit_NumValenceElectrons

This parameter, representing the number of valence electrons, was chosen because it relates to bonding and potential polarizability.²⁸ It could be useful as an auxiliary term when estimating polarizability.

7. rdkit_NumAromaticRings

This parameter was chosen based on its reflection of the presence of aromatic systems which could alter the refractive index. In molecular simulations, the contribution of aromaticity is sometimes underestimated or overestimated. For example, in DFT calculations, the Coulomb-attenuating method is well known.⁵⁹ It could be useful as a term to correct such influences.

8. rdkit_NumAliphaticRings

This parameter was adopted based on the reflection of the presence of aliphatic rings which could alter the refractive index. From our experience, the presence of aliphatic rings seldom has a critical impact on the refractive index or molecular simulation.¹⁵ We are uncertain why this parameter was prioritized, but we believe it won't detrimentally affect regression predictions.

9. DFTMD_vol

This parameter is the molecular volume estimated by DFT-MD simulation. As molecular volume is explicitly included in the Lorentz-Lorenz equation, it is clearly useful as a predictor.

10. rdkit_HeavyAtomMolWt

This parameter is the weight of the molecule without hydrogen atoms. GPT-4 chose it because it could be a measure of molecular size.²⁸ In molecular simulations, heteroatoms often have over- or under-represented influences on property predictions, hence it can be useful as a predictor.

11. rdkit_MaxAbsPartialCharge

This parameter is related to the charge state of the molecule, and GPT-4 selected it with the reasoning that it could be linked to polarizability.²⁸ It is not as straightforward as other parameters and is more of a lower-level feature. Therefore, it's not a feature we prefer, but it could be useful as a corrective term for polarizability.

12. rdkit_MolLogP

This was chosen because it describes the lipophilicity of the molecule, which could correlate with polarizability. MolLogP is a factor frequently used in chemoinformatics⁵ and, being a parameter related to polarity, makes a meaningful predictor.

13. JR_BoilingPoint

This parameter is the boiling point estimated by the group contribution method and was chosen because it can reflect intermolecular forces and thus polarizability.⁵³ Generally, molecules with higher boiling points have stronger intermolecular interactions or higher densities. The former correlates with polarizability, the latter with molecular volume, making this parameter a beneficial choice.

14. JR_EnthalpyForm

This parameter is the enthalpy of formation estimated by the group contribution method.⁵³ Like the boiling point, it reflects the stability of the molecule and may correlate with polarizability.

(2) Causal relationships

Question by Author

Think step by step. Estimate the causal relationship between the following variables and output it in Mermaid syntax.

Refractive_index

DFT_energy

DFT_HOMO

DFT_LUMO

DFT_dipoleTot

rdkit_MolWt

rdkit_NumValenceElectrons

rdkit_NumAromaticRings

rdkit_NumAliphaticRings

DFTMD_vol

rdkit_HeavyAtomMolWt

rdkit_MaxAbsPartialCharge

rdkit_MolLogP

JR_BoilingPoint

JR_EnthalpyForm

Answer by GPT-4

Mermaid is a simple markdown-like syntax that lets you generate flowcharts, sequence diagrams, and other types of diagrams. To write a diagram depicting a causal relationship, we use a flowchart syntax.

However, I should note that in order to map the causal relationships between these variables accurately, I would need to have specific knowledge about the relationships themselves. In the absence of this information, I'll have to make an assumption for the purposes of demonstration.

Let's assume a hypothetical situation where:

'Refractive_index' is influenced by 'DFT_energy', 'DFT_HOMO', 'DFT_LUMO', 'DFT_dipoleTot', and 'DFTMD_vol'.

'DFT_energy', 'DFT_HOMO', 'DFT_LUMO', 'DFT_dipoleTot' and 'DFTMD_vol' are influenced by 'rdkit_MolWt' and 'rdkit_NumValenceElectrons'.

'rdkit_MolWt' and 'rdkit_NumValenceElectrons' are influenced by 'rdkit_NumAromaticRings', 'rdkit_NumAliphaticRings', 'rdkit_HeavyAtomMolWt', 'rdkit_MaxAbsPartialCharge', 'rdkit_MolLogP', 'JR_BoilingPoint', and 'JR_EnthalpyForm'.

The Mermaid syntax would be something like this:

graph LR

JR_BoilingPoint -->|Influence| rdkit_MolWt

JR_BoilingPoint -->|Influence| rdkit_NumValenceElectrons

JR_EnthalpyForm -->|Influence| rdkit_MolWt

JR_EnthalpyForm -->|Influence| rdkit_NumValenceElectrons

rdkit_MolLogP -->|Influence| rdkit_MolWt

rdkit_MolLogP -->|Influence| rdkit_NumValenceElectrons

rdkit_MaxAbsPartialCharge -->|Influence| rdkit_MolWt

rdkit_MaxAbsPartialCharge -->|Influence| rdkit_NumValenceElectrons

rdkit_HeavyAtomMolWt -->|Influence| rdkit_MolWt

rdkit_HeavyAtomMolWt -->|Influence| rdkit_NumValenceElectrons

rdkit_NumAliphaticRings -->|Influence| rdkit_MolWt

rdkit_NumAliphaticRings -->|Influence| rdkit_NumValenceElectrons

rdkit_NumAromaticRings -->|Influence| rdkit_MolWt

rdkit_NumAromaticRings -->|Influence| rdkit_NumValenceElectrons

rdkit_MolWt -->|Influence| DFT_energy

rdkit_MolWt -->|Influence| DFT_HOMO

rdkit_MolWt -->|Influence| DFT_LUMO

rdkit_MolWt -->|Influence| DFT_dipoleTot

rdkit_MolWt -->|Influence| DFTMD_vol

rdkit_NumValenceElectrons -->|Influence| DFT_energy

rdkit_NumValenceElectrons -->|Influence| DFT_HOMO

rdkit_NumValenceElectrons -->|Influence| DFT_LUMO

rdkit_NumValenceElectrons -->|Influence| DFT_dipoleTot

rdkit_NumValenceElectrons -->|Influence| DFTMD_vol

DFT_energy -->|Influence| Refractive_index

DFT_HOMO -->|Influence| Refractive_index

DFT_LUMO -->|Influence| Refractive_index

DFT_dipoleTot -->|Influence| Refractive_index

DFTMD_vol -->|Influence| Refractive_index

Table S 3 Full regression results.

Descriptor	Model	MAE	RMSE	MAE-std	RMSE-std
GPT-4	Ridge	0.0229	0.0310	0.0038	0.0080
GPT-4	Lasso	0.0236	0.0320	0.0035	0.0083
All	Lasso	0.0272	0.0334	0.0038	0.0055
Boruta	GPR	0.0276	0.0339	0.0110	0.0124
Boruta	Ridge	0.0279	0.0345	0.0103	0.0118
Random-20-2	Lasso	0.0280	0.0355	0.0040	0.0075
All	RFR	0.0289	0.0377	0.0067	0.0127
All	GBR	0.0291	0.0379	0.0067	0.0112
All	Ridge	0.0291	0.0373	0.0072	0.0098
Boruta	RFR	0.0295	0.0387	0.0076	0.0132
Boruta	GBR	0.0300	0.0383	0.0059	0.0105
GPT-4	GPR	0.0304	0.0372	0.0124	0.0118
Random-20-2	RFR	0.0310	0.0399	0.0078	0.0144
Boruta	Lasso	0.0314	0.0410	0.0122	0.0191
Random-20-2	GPR	0.0318	0.0427	0.0150	0.0218
GPT-4	RFR	0.0319	0.0406	0.0067	0.0133
GPT-4	GBR	0.0322	0.0413	0.0080	0.0126
Random-20-2	Ridge	0.0331	0.0430	0.0075	0.0128
Random-20-2	GBR	0.0336	0.0431	0.0076	0.0131
Random-10-3	Ridge	0.0340	0.0412	0.0047	0.0037
Random-10-3	GPR	0.0352	0.0430	0.0053	0.0058
Random-20	RFR	0.0357	0.0453	0.0083	0.0130
Random-20-3	GBR	0.0367	0.0446	0.0070	0.0095
Random-10-3	Lasso	0.0371	0.0447	0.0083	0.0063
Random-20-3	RFR	0.0373	0.0466	0.0074	0.0116
Random-10	RFR	0.0380	0.0473	0.0102	0.0162
Random-20	GBR	0.0383	0.0467	0.0087	0.0128
Random-20-1	GBR	0.0385	0.0476	0.0112	0.0159
Random-10-1	Ridge	0.0388	0.0490	0.0106	0.0130
Random-20	GPR	0.0391	0.0514	0.0120	0.0238
Random-20-3	GPR	0.0399	0.0499	0.0147	0.0199
Random-10	GBR	0.0401	0.0510	0.0170	0.0246
Random-10	GPR	0.0404	0.0512	0.0109	0.0173
Random-20-1	Ridge	0.0404	0.0490	0.0088	0.0114

Random-10-3	RFR	0.0406	0.0513	0.0102	0.0171
Random-20-1	RFR	0.0407	0.0504	0.0109	0.0161
Random-20-1	GPR	0.0413	0.0518	0.0123	0.0181
Random-10-1	GPR	0.0415	0.0518	0.0105	0.0153
Random-10-1	Lasso	0.0422	0.0520	0.0074	0.0105
Random-10-3	GBR	0.0427	0.0531	0.0095	0.0140
Random-10	Lasso	0.0430	0.0548	0.0119	0.0161
Random-10-2	GBR	0.0435	0.0554	0.0114	0.0186
Random-10-2	RFR	0.0435	0.0554	0.0095	0.0191
Random-20-1	Lasso	0.0436	0.0546	0.0117	0.0169
Random-10-2	Lasso	0.0437	0.0550	0.0160	0.0230
Random-10-2	GPR	0.0447	0.0544	0.0130	0.0180
Random-10-2	Ridge	0.0451	0.0551	0.0152	0.0213
Random-20	Ridge	0.0460	0.0633	0.0232	0.0435
Random-10-1	RFR	0.0474	0.0565	0.0111	0.0168
Random-20-3	Ridge	0.0494	0.0700	0.0347	0.0635
Random-10-3	SVR	0.0528	0.0635	0.0045	0.0115
Random-20-1	SVR	0.0534	0.0640	0.0044	0.0116
Random-20-3	SVR	0.0535	0.0640	0.0039	0.0110
Random-20	SVR	0.0536	0.0641	0.0056	0.0125
Random-20-2	SVR	0.0539	0.0643	0.0041	0.0115
Random-20-3	Lasso	0.0540	0.0759	0.0424	0.0732
Boruta	SVR	0.0546	0.0650	0.0038	0.0109
GPT-4	SVR	0.0553	0.0655	0.0029	0.0104
Random-10	SVR	0.0557	0.0662	0.0033	0.0100
Random-10-1	GBR	0.0560	0.0661	0.0119	0.0171
Random-10	Ridge	0.0564	0.0844	0.0434	0.0855
All	SVR	0.0567	0.0669	0.0033	0.0098
Random-20	Lasso	0.0572	0.0820	0.0422	0.0802
Random-10-1	SVR	0.0584	0.0685	0.0025	0.0085
Random-10-2	SVR	0.0585	0.0693	0.0024	0.0077
All	GPR	0.0664	0.1310	0.0831	0.2021

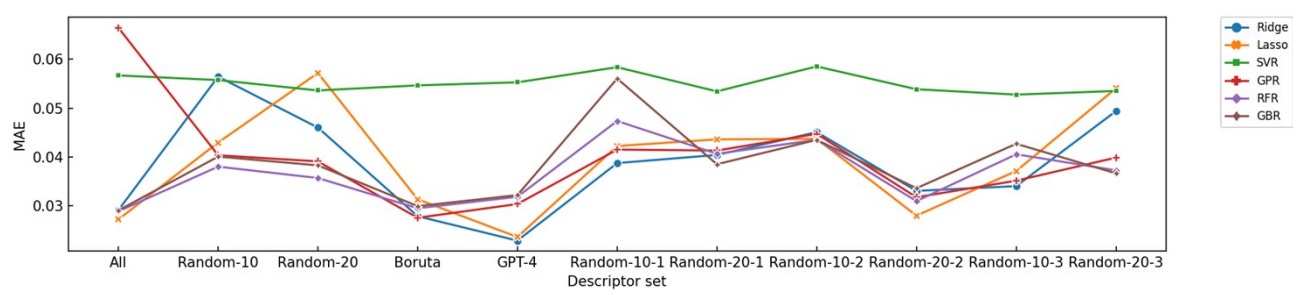


Fig. S3 Comparison of MAE with different descriptor sets and regression models.

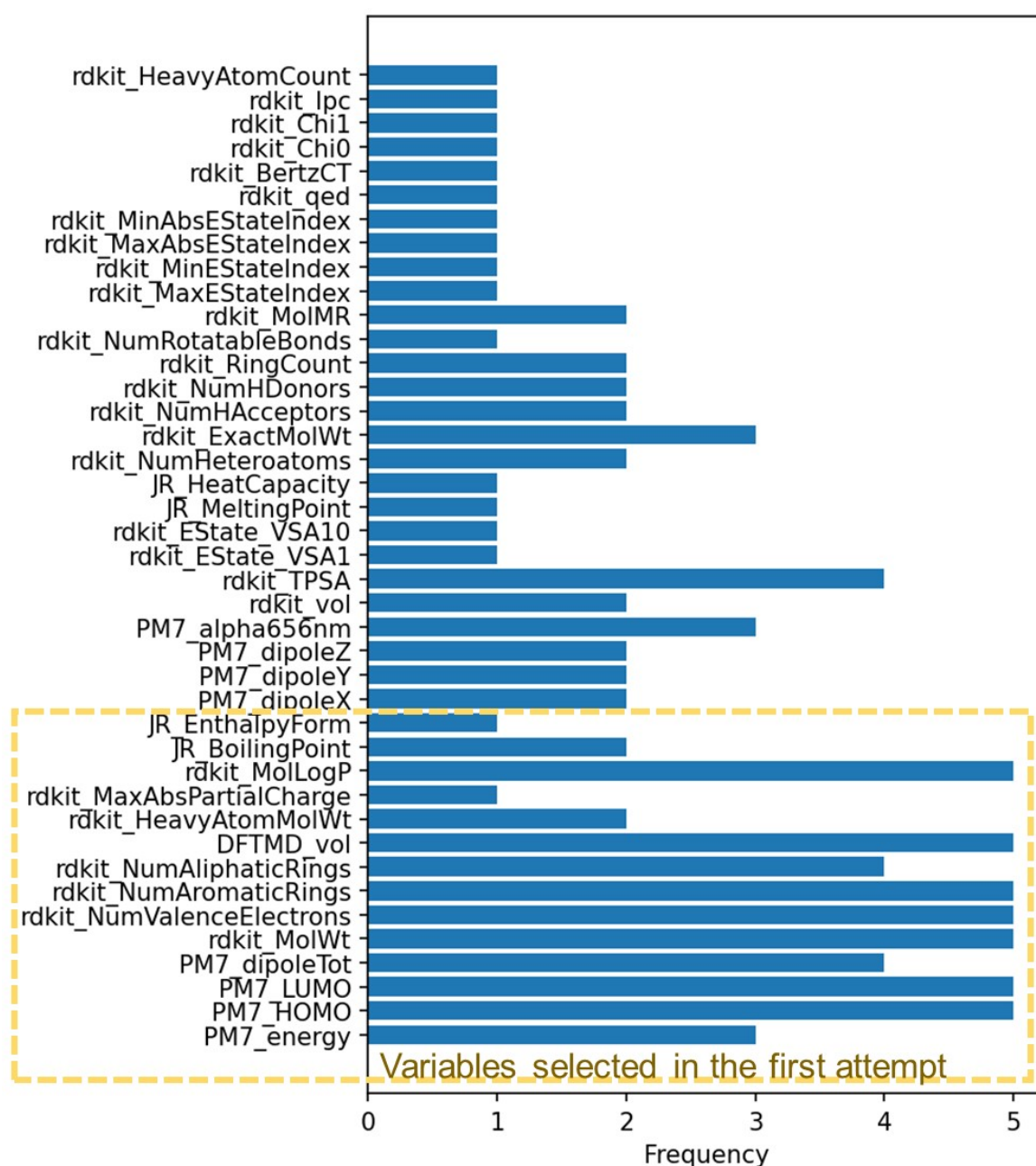
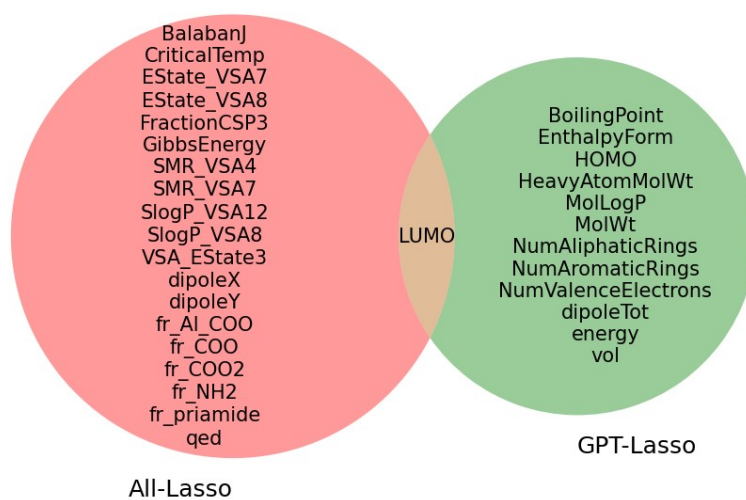
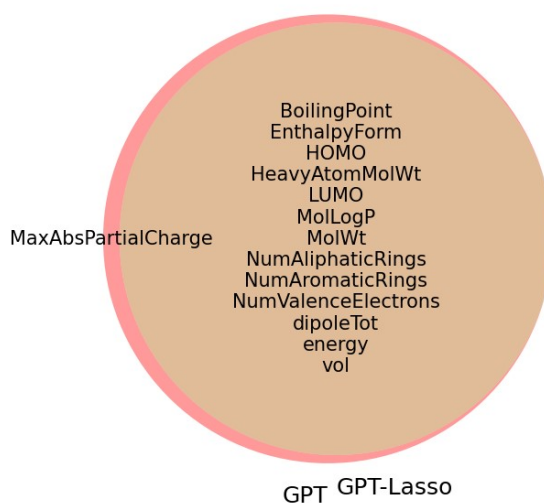


Fig. S4 Results of parameter selection from five independent question-and-answer processes with GPT-4. In this figure, we recorded the frequency of the proposed parameters as a result of independently asking GPT-4 for an explanation selection task five times. The parameters encased within the border up to the 14th from the bottom were extracted in the first trial. We find that generic parameters such as HOMO, LUMO, and energy are frequently chosen, whereas parameters with low interpretability like MaxAbsPartialCharge are less likely to be selected. Whether to use the values that appeared in the initial answer or the results of multiple samplings can be considered as a type of hyperparameter when using machine learning, which is one of the future research topics.



a)



b)

Fig. S5 Venn's diagram for the descriptor sets. a) All-Lasso and GPT-Lasso. b) GPT and GPT-Lasso.

GPT: Descriptors selected by GPT-4.

GPT-Lasso: Descriptors selected by Lasso from the GPT set.

All-Lasso: Descriptors selected by Lasso from the All set.

For GPT-Lasso and All-Lasso, all 44 polymers were used for regression.