Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Molecular Partition Coefficient from Machine Learning with Polarization and Entropy Embedded Atom-Centered Symmetry Functions

Qiang Zhu, Qingqing Jia, Ziteng Liu, Yang Ge, Xu Gu, Ziyi Cui, Mengting Fan, and Jing Ma*

Key Laboratory of Mesoscopic Chemistry of Ministry of Education Institute of Theoretical and Computational Chemistry School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, P. R. China

E-mail: majing@nju.edu.cn

Contents

S1 Details of 100 selected descriptors S3

S2 Homemade dataset S6

S3 Summary of molecular entropy S8

S4 Atom-centered Symmetry Functions (ACSFs) S9

S5 Grid search of optimal parameters S10

S6 Distribution of chemical elements and partition coefficient among 6 datasets S12

S7 Computational Details S13

S7.1 Molecular Dynamics Simulations S13

S7.2 Quantum Mechanisms S14

S8 Evaluation metrics S15

S9 Principle Component Analysis S16

S10 Generation of Standard Descriptors S17

S11 Feature selection S18

S12 An illustration of 5-5 neural network S20

S13 Distribution of contribution from 4 distinct elements with different environment over datasets Star & Non-Star S22

References S27

S1 Details of 100 selected descriptors

Table S1: 100 descriptors extracted from RDKit

No.	Name	Description
0	MaxEStateIndex	N/A
1	MinEStateIndex	N/A
2	MaxAbsEStateIndex	N/A
3	MinAbsEStateIndex	N/A
4	qed	Calculate the weighted sum of ADS mapped properties
5	MolWt	The average molecular weight of the molecule
6	HeavyAtomMolWt	The average molecular weight of the molecule ignoring hydrogens
7	ExactMolWt	The exact molecular weight of the molecule
8	NumValenceElectrons	The number of valence electrons the molecule has
9	MaxPartialCharge	N/A
10	MinPartialCharge	N/A
11	MaxAbsPartialCharge	N/A
12	MinAbsPartialCharge	N/A
13	FpDensityMorgan1	N/A
14	FpDensityMorgan2	N/A
15	FpDensityMorgan3	N/A
16	BalabanJ	Calculate Balaban's J value for a molecule
17	BertzCT	A topological index meant to quantify "complexity" of molecules.
18	Chi0	From equations (1),(9) and (10) of Rev. Comp. Chem. vol 2, 367-422, (1991)
19	Chi0n	N/A
20	Chi0v	N/A
21	Chi1	From equations (1),(11) and (12) of Rev. Comp. Chem. vol 2, 367-422, (1991)
22	Chi1n	N/A
23	Chi1v	N/A
24	Chi2n	N/A
25	Chi2v	N/A
26	Chi3n	N/A
27	Chi3v	N/A
28	Chi4n	N/A
29	Chi4v	N/A
30	HallKierAlpha	N/A
31	Ipc	This returns the information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule.
32	Kappa1	N/A
33	Kappa2	N/A
34	Kappa3	N/A
35	LabuteASA	N/A
36	PEOE_VSA1	MOE Charge VSA Descriptor 1 (-inf < x < -0.30)
37	PEOE_VSA10	MOE Charge VSA Descriptor 10 (0.10 <= x < 0.15)
38	PEOE_VSA11	MOE Charge VSA Descriptor 11 (0.15 <= x < 0.20)
39	PEOE_VSA12	MOE Charge VSA Descriptor 12 (0.20 <= x < 0.25)
40	PEOE_VSA13	MOE Charge VSA Descriptor 13 (0.25 <= x < 0.30)
41	PEOE_VSA14	MOE Charge VSA Descriptor 14 (0.30 <= x < inf)
42	PEOE_VSA2	MOE Charge VSA Descriptor 2 (-0.30 <= x < -0.25)
43	PEOE_VSA3	MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20)
44	PEOE_VSA4	MOE Charge VSA Descriptor 4 (-0.20 <= x < -0.15)
45	PEOE_VSA5	MOE Charge VSA Descriptor 5 (-0.15 <= x < -0.10)
46	PEOE_VSA6	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)
47	PEOE_VSA7	MOE Charge VSA Descriptor 7 (-0.05 <= x < 0.00)
48	PEOE_VSA8	MOE Charge VSA Descriptor 8 (0.00 <= x < 0.05)

No.	Name	Description
49	PEOE_VSA9	MOE Charge VSA Descriptor 9 (0.05 <= x < 0.10)
50	SMR_VSA1	MOE MR VSA Descriptor 1 (-inf < x < 1.29)
51	SMR_VSA10	MOE MR VSA Descriptor 10 (4.00 <= x < inf)
52	SMR_VSA2	MOE MR VSA Descriptor 2 (1.29 <= x < 1.82)
53	SMR_VSA3	MOE MR VSA Descriptor 3 (1.82 <= x < 2.24)
54	SMR_VSA4	MOE MR VSA Descriptor 4 (2.24 <= x < 2.45)
55	SMR_VSA5	MOE MR VSA Descriptor 5 (2.45 <= x < 2.75)
56	SMR_VSA6	MOE MR VSA Descriptor 6 (2.75 <= x < 3.05)
57	SMR_VSA7	MOE MR VSA Descriptor 7 (3.05 <= x < 3.63)
58	SMR_VSA8	MOE MR VSA Descriptor 8 (3.63 <= x < 3.80)
59	SMR_VSA9	MOE MR VSA Descriptor 9 (3.80 <= x < 4.00)
60	TPSA	N/A
61	EState_VSA1	EState VSA Descriptor 1 (-inf < x < -0.39)
62	EState_VSA10	EState VSA Descriptor 10 (9.17 <= x < 15.00)
63	EState_VSA11	EState VSA Descriptor 11 (15.00 <= x < inf)
64	EState_VSA2	EState VSA Descriptor 2 (-0.39 <= x < 0.29)
65	EState_VSA3	EState VSA Descriptor 3 (0.29 <= x < 0.72)
66	EState_VSA4	EState VSA Descriptor 4 (0.72 <= x < 1.17)
67	EState_VSA5	EState VSA Descriptor 5 (1.17 <= x < 1.54)
68	EState_VSA6	EState VSA Descriptor 6 (1.54 <= x < 1.81)
69	EState_VSA7	EState VSA Descriptor 7 (1.81 <= x < 2.05)
70	EState_VSA8	EState VSA Descriptor 8 (2.05 <= x < 4.69)
71	EState_VSA9	EState VSA Descriptor 9 (4.69 <= x < 9.17)
72	VSA_EState1	VSA EState Descriptor 1 (-inf < x < 4.78)
73	VSA_EState10	VSA EState Descriptor 10 (11.00 <= x < inf)
74	VSA_EState2	VSA EState Descriptor 2 (4.78 <= x < 5.00)
75	VSA_EState3	VSA EState Descriptor 3 (5.00 <= x < 5.41)
76	VSA_EState4	VSA EState Descriptor 4 (5.41 <= x < 5.74)
77	VSA_EState5	VSA EState Descriptor 5 (5.74 <= x < 6.00)
78	VSA_EState6	VSA EState Descriptor 6 (6.00 <= x < 6.07)
79	VSA_EState7	VSA EState Descriptor 7 (6.07 <= x < 6.45)
80	VSA_EState8	VSA EState Descriptor 8 (6.45 <= x < 7.00)
81	VSA_EState9	VSA EState Descriptor 9 (7.00 <= x < 11.00)
82	FractionCSP3	CalcFractionCSP3((Mol)mol) -> float : returns the fraction of C atoms that are SP3 hybridized
83	HeavyAtomCount	Number of heavy atoms a molecule.
84	NHOHCount	Number of NHs or OHs
85	NOCount	Number of Nitrogens and Oxygens
86	NumAliphaticCarbocycles	CalcNumAliphaticCarbocycles((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) carbocycles for a molecule
87	NumAliphaticHeterocycles	CalcNumAliphaticHeterocycles((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) heterocycles for a molecule
88	NumAliphaticRings	CalcNumAliphaticRings((Mol)mol) -> int : returns the number of aliphatic (containing at least one non-aromatic bond) rings for a molecule
89	NumAromaticCarbocycles	CalcNumAromaticCarbocycles((Mol)mol) -> int : returns the number of aromatic carbocycles for a molecule
90	NumAromaticHeterocycles	CalcNumAromaticHeterocycles((Mol)mol) -> int : returns the number of aromatic heterocycles for a molecule
91	NumAromaticRings	CalcNumAromaticRings((Mol)mol) -> int : returns the number of aromatic rings for a molecule
92	NumHAcceptors	Number of Hydrogen Bond Acceptors
93	NumHDonors	Number of Hydrogen Bond Donors
94	NumHeteroatoms	Number of Heteroatoms
95	NumRotatableBonds	Number of Rotatable Bonds
96	NumSaturatedCarbocycles	CalcNumSaturatedCarbocycles((Mol)mol) -> int : returns the number of saturated carbocycles for a molecule

No.	Name	Description
97	NumSaturatedHeterocycles	CalcNumSaturatedHeterocycles((Mol)mol) -> int : returns the number of saturated heterocycles for a molecule
98	NumSaturatedRings	CalcNumSaturatedRings((Mol)mol) -> int : returns the number of saturated rings for a molecule
99	RingCount	N/A

S2 Homemade dataset

Table S2: Collection of molecules of datasets n-carboxylic acids and Solv-54.

Name	SMILES	$\log P_{exp}$	$\langle q\!-\!ACSFs \rangle_{conf}$
	n-carboxylic acids		
acetic acid	CC(=O)O	-0.17	-0.27
propionic acid	CCC(=O)O	0.33	0.26
butyric acid	CCCC(=O)O	0.79	0.73
valeric acid	CCCCC(=O)O	1.39	1.25
caproic acid	CCCCCC(=O)O	1.92	1.79
enanthic acid	CCCCCCC(=O)O	2.42	2.38
caprylic acid	CCCCCCCC(=O)O	3.05	2.92
pelargonic acid	CCCCCCCCC(=O)O	3.42	3.36
capric acid	CCCCCCCCCC(=O)O	4.09	3.88
undecanoic acid	CCCCCCCCCCC(=O)O	4.42	4.40
lauric acid	CCCCCCCCCCCC(=O)O	4.60	4.76
tridecanoic acid	CCCCCCCCCCCCC(=O)O	5.49	5.57
myristic acid	CCCCCCCCCCCCCC(=O)O	6.11	6.12
	Solv-54		
butanal	CCCC=O	0.88	0.80
5-Nonanone	CCCCC(=O)CCCC	2.88	3.01
hexadecanoic acid	CCCCCCCCCCCCCCCC(=O)O	7.17	6.96
octadecanoic acid	CCCCCCCCCCCCCCCCCC(=O)O	8.35	8.12
methane	C	1.09	0.97
ethane	CC	1.81	1.52
propane	CCC	2.36	2.07
butane	CCCC	2.89	2.61
pentane	CCCCC	3.26	3.14
2-methyl-butane	CCC(C)C	2.72	3.16
hexane	CCCCCC	3.90	3.67
cyclohexane	C1CCCCC1	3.44	3.31
heptane	CCCCCCC	4.66	4.19
octane	CCCCCCCC	4.78	4.73
oct-1-ene	CCCCCCC=C	4.57	4.34
nonane	CCCCCCCCC	5.45	5.27
decane	CCCCCCCCCC	5.01	5.79
dodecane	CCCCCCCCCCCC	6.10	6.86
tetradecane	CCCCCCCCCCCCCC	7.20	7.90
hexadecane	CCCCCCCCCCCCCCCC	8.2	8.95
octadecane	CCCCCCCCCCCCCCCCCC	9.32	10.02
3,6,9-trioxa-undecan-1,11-diol	OCCOCCOCCOCCO	-2.02	-1.35
ethane-1,2-diol	OCCO	-1.36	-1.40
butane-1,4-diol	OCCCCO	-0.83	-0.45
methanol	CO	-0.69	-0.64
cyclopentanol	OC1CCCC1	0.71	0.82
butan-1-ol	CCCCO	0.88	0.88
3-methyl-butan-1-ol	CC(C)CCO	1.16	1.30
cyclohexanol	OC1CCCCC1	1.23	1.20
octan-1-ol	CCCCCCCCO	3.00	2.89
N,N-dimethyl-formamide	CN(C)C=O	-1.01	-0.82
N-methyl-formamide	CNC=O	-0.97	-1.06
N,N-dimethyl-acetamide	CN(C)C(C)=O	-0.77	-0.89
nitromethane	C[N+]([O-])=O	-0.35	-0.03
diethylamine	CCNCC	0.66	0.65

Name	SMILES	log P_{exp}	$\langle q{-}ACSFs\rangle_{conf}$
4-methyl-pyridine	Cc1ccncc1	1.22	1.34
triethylamine	CCN(CC)CC	1.65	1.46
1,4-Dioxane	C1COCCO1	-0.27	-0.38
fromic acid ethyl ester	CCOC=O	0.23	0.20
1,2,3-triacetoxy-propane	CC(=O)OCC(COC(C)=O)OC(C)=O	0.25	0.29
heptan-2-one	CCCCCC(C)=O	1.98	1.96
phenol	Oc1ccccc1	1.46	1.39
benzonitrile	N#Cc1ccccc1	1.56	1.58
benzene	c1ccccc1	2.13	2.19
toluene	CC1=CC=CC=C1	2.73	2.68
o-xylene	CC1=CC=CC=C1C	3.12	3.17
p-xylene	CC1=CC=C(C=C1)C	3.15	3.19
acetonitrile	CC#N	-0.33	-0.39
m-xylene	CC1=CC(=CC=C1)C	3.20	3.17
naphthalene	C1=CC=C2C=CC=CC2=C1	3.30	3.37
1-methyl-naphthalene	Cc1cccc2ccccc12	3.87	3.86
pyrene	C1=CC2=C3C(=C1)C=CC4=CC=CC(=C43)C=C2	4.88	5.06
p-terphenyl	C1=CC=C(C=C1)C2=CC=C(C=C2)C3=CC=CC=C3	6.03	5.74
benzo(e)pyrene	C1=CC=C2C(=C1)C3=CC=CC4=C3C5=C(C=CC=C25)C=C4	6.44	6.22

S3 Summary of molecular entropy

Table S3: Properties of molecules for building the correlation between partition coefficient ($\log P$) and entropy (S) measured experimentally or calculated by quantum mechanism at b3lyp/6-31g(d) level.

Name	SMILES	$\log P_{exp}$	S_{exp} [a]	S_{trans}^{QM} [a]	S_{rot}^{QM} [a]	S_{vib}^{QM} [a]	S_{total}^{QM} [a]
Methane	C	1.09	187.46	143.41	42.44	0.32	186.17
Cyclohexane	C1CCCCC1	3.44	298.19	164.09	110.41	37.62	312.12
1,4-Dioxane	C1COCCO1	-0.27	299.91	164.67	103.12	30.45	298.23
Butanal	CCCC=O	0.88	344.80	162.16	107.50	58.75	328.41
Pentane	CCCCC	3.39	347.82	162.17	103.44	66.23	331.85
o-xylene	CC1=CC=CC=C1C	3.12	353.60	166.99	117.19	68.11	352.29
m-xylene	CC1=CC(=CC=C1)C	3.2	358.20	166.99	117.86	100.34	385.19
Hexane	CCCCCC	3.9	388.82	164.39	114.24	90.81	369.43
Octane	CCCCCCCC	5.18	467.06	167.90	122.47	142.72	433.09
Nonane	CCCCCCCCC	5.45	506.50	169.35	120.13	169.64	459.12
decane	CCCCCCCCCC	5.01	545.80	170.64	123.11	197.42	491.17
dodecane	CCCCCCCCCCCC	6.1	622.50	172.89	128.35	252.61	553.85
Tetradecane	CCCCCCCCCCCCCC	7.2	700.40	174.79	132.80	311.75	619.34
Hexadecane	CCCCCCCCCCCCCCCC	8.2	778.31	176.44	136.65	368.64	681.74
5-Nonanone	CCCCC(=O)CCCC	2.88	-	170.64	127.84	199.79	498.28
hexadecanoic acid	CCCCCCCCCCCCCCCC(=O)O	7.17	-	177.99	146.48	405.60	730.08
Octadecanoic acid	CCCCCCCCCCCCCCCCCC(=O)O	8.35	-	179.29	149.56	463.79	792.63
Ethane	CC	1.81	-	151.25	68.21	8.09	227.54
Propane	CCC	2.36	-	156.03	89.23	22.87	268.12
Butane	CCCC	2.89	-	159.47	96.95	43.44	299.86
Heptane	CCCCCCC	4.66	-	166.27	112.96	116.26	395.49
Octadecane	CCCCCCCCCCCCCCCCCC	9.32	-	177.90	140.05	427.34	745.29
Benzene	C1=CC=CC=C1	2.13	-	163.16	107.41	18.35	288.93
Toluene	CC1=CC=CC=C1	2.73	-	165.22	112.96	54.06	332.25
Naphthalene	C1=CC=C2C=CC=CC2=C1	3.3	-	169.34	121.15	52.60	343.09
p-terphenyl	C1=CC=C(C=C1)C2=CC=C(C=C2)C3=CC=CC=C3	6.03	-	176.65	139.15	177.44	493.24
Benzo(e)pyrene	C1=CC=C2C(=C1)C3=CC=CC4=C3C5=C(C=CC=C25)C=C4	6.44	-	177.79	139.32	145.66	462.78
Pyrene	C1=CC2=C3C(=C1)C=CC4=CC=CC(=C43)C=C2	4.88	-	175.03	127.13	99.15	401.31
p-xylene	CC1=CC=C(C=C1)C	3.15	-	166.99	111.42	95.29	373.70

[a] in the unit of $J \cdot mol^{-1} K^{-1}$

S4 Atom-centered Symmetry Functions (ACSFs)

Atom-centered symmetry functions describe the local chemical environment of atom i with two sets of parameters, namely, radial and angular distributions. The radial distribution is expressed as below:

$$G_i^{rad} = \sum_{j \neq i}^{all} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{S1}$$

where R_{ij} is the distance between atom i and j, parameters η and R_s determines the width and peak position of the Gaussian function. f_c is a cutoff function that utilized here to only take atoms that within the local environment into consideration. It is a function related to the distance R_{ij} and take the form

$$f_c(R_{ij}) = \begin{cases} 0.5 \times [\cos(\frac{\pi R_{ij}}{R_c}) + 1] & for\ R_{ij}\ \leq\ R_c; \\ 0 & for\ R_{ij}\ >\ R_c \end{cases} \tag{S2}$$

where R_c is the distance that specifying how large the region size that should be considered. In this work, the cutoff distance was all set to be 6.0 Å.

The angular symmetry functions are described in eq. S3:

$$G_i^{ang} = 2^{1-\zeta} \sum_{j,k \neq i}^{all} (1 + \lambda \cos(\theta_{ijk}))^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{S3}$$

Here, θ_{ijk} is the angle formed between the atoms i, j and k. R is the distance between any two atoms specified in the subscripts. Parameter η again determines the width of the Gaussian function, the peak position of Gaussian function is set to be 0. λ is a parameter which could only take the value of $+1$ and -1 so as to shift the maxima of the cosine function to 0° and 180°, respectively. ζ is a parameter controls the angular resolution. High ζ values result in a narrow distribution of angular symmetry function values.

S5 Grid search of optimal parameters

The Adam optimizer[1] was utilized here for the gradient descent updates. Four different learning rates ($1e-2$, $1e-3$, $1e-4$ and $1e-5$) together with four different architectures ($25-25-25$, $50-50-50$, $75-75-75$ and $100-100-100$) were tested. Detailed results could be found in Table S4-S7. In four different learning rates, a factor of 0.999 was applied and the learning rate decayed every 10000 steps. A gradient norm clipping strategy was employed so as to avoid exploding gradient problems.[2] For the reduction of overfitting and the generalization of models, early stopping strategy was applied with maximum number of training steps set to be 1000. The total trainings steps were all set to be 200000. The high-dimensional neural networks mentioned above was built with the Tensorflow.[3]

Table S4: Grid search for different architectures and learning rates, where the polarization effects and entropy are encoded into the conventional atom-centered symmetry functions ($\langle q - ACSFs \rangle_{conf}$)

Datasets	lr	1e-2			1e-3			1e-4			1e-5		
	arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Martel	100-100-100	1.51	3.36	1.83	0.87	1.35	1.16	0.98	1.80	1.34	0.81	1.16	1.07
Star&Nonstar		0.86	1.20	1.09	0.59	0.70	0.84	0.60	0.78	0.88	0.63	0.70	0.84
Huuskonen		0.56	0.54	0.74	0.20	0.13	0.35	0.23	0.14	0.37	0.35	0.26	0.51
Martel	75-75-75	0.79	1.00	1.00	0.92	1.56	1.25	0.87	1.43	1.20	0.82	1.19	1.09
Star&Nonstar		1.24	2.83	1.68	0.56	0.57	0.76	0.59	0.69	0.83	0.60	0.65	0.81
Huuskonen		0.55	0.61	0.78	0.21	0.12	0.35	0.27	0.17	0.41	0.35	0.25	0.50
Martel	50-50-50	0.89	1.30	1.14	0.91	1.54	1.24	0.90	1.50	1.22	0.84	1.25	1.12
Star&Nonstar		0.58	0.59	0.77	0.52	0.51	0.72	0.60	0.69	0.83	0.60	0.64	0.80
Huuskonen		0.33	0.23	0.48	0.21	0.12	0.35	0.29	0.19	0.44	0.35	0.25	0.50
Martel	25-25-25	1.07	1.87	1.37	**0.91**	**1.50**	**1.23**	0.87	1.34	1.16	0.86	1.31	1.15
Star&Nonstar		0.59	0.57	0.75	**0.48**	**0.44**	**0.66**	0.57	0.60	0.78	0.59	0.63	0.80
Huuskonen		0.31	0.19	0.43	**0.22**	**0.12**	**0.35**	0.30	0.20	0.45	0.36	0.26	0.51

Table S5: Grid search for different architectures and learning rates, where only polarization effects and the stablest conformation are encoded into the conventional atom-centered symmetry functions ($q - ACSFs^{max}$)

Datasets	lr	1e-2			1e-3			1e-4			1e-5		
	arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Martel	100-100-100	0.73	0.88	0.94	0.87	1.29	1.14	0.94	1.66	1.29	0.89	1.41	1.19
Star&Nonstar		1.05	1.90	1.38	0.64	0.80	0.89	0.58	0.72	0.85	0.59	0.63	0.79
Huuskonen		0.54	0.56	0.75	0.21	0.14	0.37	0.23	0.14	0.37	0.35	0.25	0.50
Martel	75-75-75	1.67	3.69	1.92	**0.90**	**1.53**	**1.23**	0.92	1.58	1.26	0.86	1.33	1.15
Star&Nonstar		2.44	9.22	3.04	**0.54**	**0.54**	**0.74**	0.60	0.73	0.85	0.60	0.64	0.80
Huuskonen		1.08	1.75	1.32	**0.22**	**0.13**	**0.37**	0.27	0.17	0.41	0.35	0.25	0.50
Martel	50-50-50	0.90	1.26	1.12	0.94	1.61	1.27	0.90	1.45	1.21	0.87	1.36	1.17
Star&Nonstar		0.65	0.76	0.87	0.53	0.52	0.72	0.58	0.64	0.80	0.59	0.64	0.80
Huuskonen		0.37	0.27	0.52	0.21	0.12	0.35	0.29	0.19	0.43	0.35	0.25	0.50
Martel	25-25-25	0.89	1.34	1.16	0.95	1.65	1.28	0.86	1.32	1.15	0.87	1.33	1.15
Star&Nonstar		0.55	0.56	0.75	0.52	0.51	0.71	0.58	0.64	0.80	0.60	0.64	0.80
Huuskonen		0.27	0.18	0.42	0.23	0.13	0.37	0.31	0.21	0.45	0.36	0.26	0.51

Table S6: Grid search for different architectures and learning rates, where only entropy effects are encoded into the conventional atom-centered symmetry functions ($\langle ACSFs\rangle_{conf}$)

Datasets	lr	1e-2			1e-3			1e-4			1e-5		
	arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Martel Star&Nonstar Huuskonen	100-100-100	1.85 0.96 0.91	4.25 1.39 1.28	2.06 1.18 1.13	1.04 0.84 0.58	1.84 1.29 0.61	1.36 1.14 0.78	1.03 0.87 0.48	1.8 1.36 0.45	1.34 1.16 0.67	1.06 0.79 0.52	1.89 1.17 0.49	1.38 1.08 0.70
Martel Star&Nonstar Huuskonen	75-75-75	0.83 1.36 0.67	1.12 3.46 0.78	1.06 1.86 0.88	0.95 0.92 0.58	1.58 1.54 0.62	1.26 1.24 0.79	1.05 0.77 0.49	1.89 1.06 0.44	1.37 1.03 0.66	1.03 0.84 0.54	1.84 1.34 0.53	1.36 1.16 0.73
Martel Star&Nonstar Huuskonen	50-50-50	1.90 2.81 1.08	4.80 13.13 1.87	2.19 3.62 1.37	0.99 0.82 0.57	1.67 1.19 0.57	1.29 1.09 0.75	1.08 0.77 0.51	1.96 1.12 0.47	1.40 1.06 0.69	**0.97** **0.83** **0.54**	**1.66** **1.27** **0.53**	**1.29** **1.13** **0.73**
Martel Star&Nonstar Huuskonen	25-25-25	0.82 1.28 0.66	1.10 3.11 0.76	1.05 1.76 0.87	1.01 0.83 0.56	1.77 1.27 0.56	1.33 1.13 0.75	1.04 0.83 0.54	1.84 1.30 0.54	1.36 1.14 0.73	0.98 0.83 0.57	1.64 1.30 0.60	1.28 1.14 0.77

Table S7: Grid search for different architectures and learning rates, where only the stablest conformation was encoded into the conventional atom-centered symmetry functions ($ACSFs^{max}$)

Datasets	lr	1e-2			1e-3			1e-4			1e-5		
	arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Martel Star&Nonstar Huuskonen	100-100-100	1.10 0.98 0.71	1.94 1.71 0.82	1.39 1.31 0.91	0.96 0.96 0.58	1.63 1.62 0.62	1.28 1.27 0.79	1.15 0.79 0.48	2.19 1.16 0.45	1.48 1.08 0.67	1.10 0.79 0.52	2.00 1.12 0.48	1.42 1.06 0.70
Martel Star&Nonstar Huuskonen	75-75-75	1.35 2.21 0.87	2.63 8.38 1.30	1.62 2.89 1.14	1.00 0.83 0.58	1.70 1.26 0.60	1.30 1.12 0.78	1.05 0.83 0.50	1.88 1.26 0.49	1.37 1.12 0.70	0.98 0.84 0.54	1.65 1.34 0.54	1.29 1.16 0.74
Martel Star&Nonstar Huuskonen	50-50-50	1.41 2.28 0.89	2.86 8.79 1.35	1.69 2.97 1.16	1.09 0.77 0.57	2.03 1.10 0.57	1.42 1.05 0.75	1.06 0.79 0.51	1.88 1.17 0.48	1.37 1.08 0.69	0.98 0.83 0.55	1.69 1.30 0.57	1.30 1.14 0.75
Martel Star&Nonstar Huuskonen	25-25-25	1.02 0.94 0.71	1.63 1.51 0.82	1.28 1.23 0.91	1.01 0.79 0.56	1.80 1.15 0.55	1.34 1.07 0.74	**0.96** **0.82** **0.54**	**1.60** **1.27** **0.53**	**1.27** **1.13** **0.73**	0.97 0.83 0.57	1.57 1.26 0.60	1.25 1.12 0.77

Table S8: Grid search for different architectures and learning rates with all four public datasets taking into consideration, where the polarization effects and the ensemble effects are encoded into the conventional atom-centered symmetry functions ($\langle q - ACSFs\rangle_{conf}$)

lr	1e-2			1e-3			1e-4			1e-5		
arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
100-100-100	2.15	7.42	2.72	0.35	0.27	0.52	**0.35**	**0.26**	**0.51**	0.42	0.33	0.57
75-75-75	1.76	4.79	2.19	0.37	0.30	0.54	0.36	0.28	0.53	0.42	0.33	0.57
50-50-50	0.89	1.39	1.18	0.37	0.31	0.56	0.37	0.28	0.53	0.42	0.33	0.58
25-25-25	0.49	0.44	0.67	0.37	0.30	0.55	0.38	0.29	0.54	0.43	0.34	0.58

Table S9: Grid search for different architectures and learning rates with all four public datasets taking into consideration, where the ensemble effects are encoded into the conventional atom-centered symmetry functions ($\langle ACSFs\rangle_{conf}$)

lr	1e-2			1e-3			1e-4			1e-5		
arch.	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
100-100-100	1.94	5.53	2.35	0.69	0.83	0.91	0.62	0.70	0.83	0.62	0.68	0.83
75-75-75	2.54	9.95	3.15	0.68	0.82	0.91	0.62	0.68	0.83	0.62	0.68	0.83
50-50-50	1.93	5.79	2.41	0.66	0.74	0.86	**0.61**	**0.68**	**0.82**	0.62	0.69	0.83
25-25-25	1.23	2.72	1.65	0.63	0.70	0.84	0.62	0.68	0.82	0.64	0.72	0.85

S6 Distribution of chemical elements and partition coefficient among 6 datasets

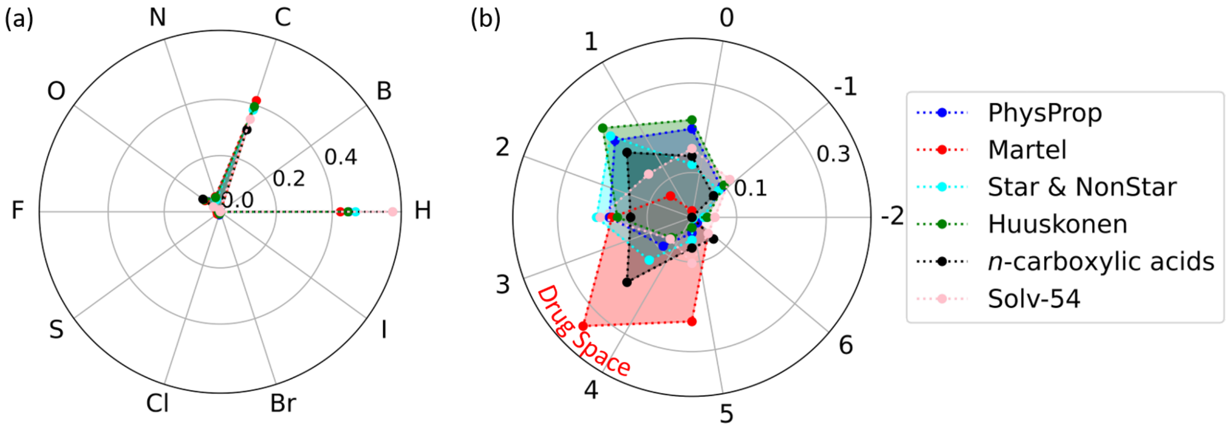

Figure S1: Distribution of (a) chemical elements and (b) partition coefficient ($\log P$) among 6 public databases.

For database PhysProp, the $\log P$ values mainly concentrate around 2 (blue shadows). For database Star & Non-Star and Huuskonen, the distribution did not change a lot. (cyan and green shadows) As the Martel is more compatible with pharmaceutical company, the $\log P$ distribution is quite different from the above mentioned three databases and values are almost small than 5 and concentrate around 4, (red shadows) which follows the Lipinski's rule of five.[4]

S7 Computational Details

S7.1 Molecular Dynamics Simulations

For the small organic molecules, RDKit[5] and OpenBabel[6] were utilize to convert molecules from SMILES[7,8] format to PDB format. Subsequentally, GAFF2[9] force fields were applied. The process of generating GAFF2 parameters leads to a final of 4802, 178, 156, and 1122 molecules for PhysProp, Martel, Star & NonStar, and Huuskonen, respectively. With all bonds that involve hydrogen atom are constrained, the integration time step was set to be 2 fs. LangevinMiddleIntegrator[10] method was applied, with temperature set to be 300 K and the friction coefficient to be 1 ps^{-1}. As all simulations were performed in the vacuum and the electrostatic and van der Waals (vdW) interactions were calculated over the whole chemical spaces without cutoff. To sample the conformational space, trajectories which last 1 ns were generated for each small organic molecules following a local energy minimization with tolerance set to be 10.0 kJ/mol. All simulations were performed with package suite OpenMM.[11]

Consdering the large amount of conformations generated from molecular dynamics simulations, K-Means clustering method[12] was utilized here to divide structures into 3 groups. To pick the centroid as the representitive structure for each group, we computed all of the pairwise RMSDs between conformations among a certain group, and transformed these distances into similarity scores according to eq. S4

$$s_{ij} = e^{-d_{ij}/d_{scale}} \tag{S4}$$

Where $d_{ij} = \sqrt{\frac{\sum_{n=1}^{N_{atoms}}(r_i^n - r_j^n)^2}{N_{atoms}}}$ is the pairwise distance between i_{th} and j_{th} conformations, and d_{scale} is a parameter which is the standard deviation of the pairwise distance so as to make the computation scale invariant.

The centroid structure is picked with the highest similarities and the mathematical expression is as follows:

$$argmax_i \sum_j s_{ij} \tag{S5}$$

The probability of selected i_{th} conformation is defined as below

$$p_i = \frac{e^{-\Delta E_i / k_B T}}{\sum_{j=1}^{M} e^{-\Delta E_j / k_B T}} \tag{S6}$$

Where k_B is the Boltzmann constant, T is the temperature where the simulation performed, M is the number of clusters we specified, and ΔE_i is the difference between potential energy of the selected conformation i and the lowest energies of M clusters. Here, K-Means clustering procedure was implemented with scikit-learn,[13] and trajectories were processed with mdtraj.[14]

S7.2 Quantum Mechanisms

The density functional theory (DFT) calculations were utilized here for the estimation of entropy with the help of Gaussian 16 package suite.[15] The Becke three-parameter exchange and Lee-Yang-Parr correlation (B3LYP),[16,17] a hybrid density function, was used here for the geometry optimizations and subsequent frequency calculation. The 6-31G(d) basis set was applied for all the organic molecules.

S8 Evaluation metrics

Performances of models under different descriptors are represented in terms of mean square error (MSE) and mean absolute error (MAE). These two criteria are defined as below:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2 \tag{S7}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_{test}^{(i)} - \hat{y}_{test}^{(i)}| \tag{S8}$$

Where m denotes the number of the data used for test and $y_{test}^{(i)}$ and $\hat{y}_{test}^{(i)}$ denote the actual and predicted ones for the i_{th} molecule, respectively.

S9 Principle Component Analysis

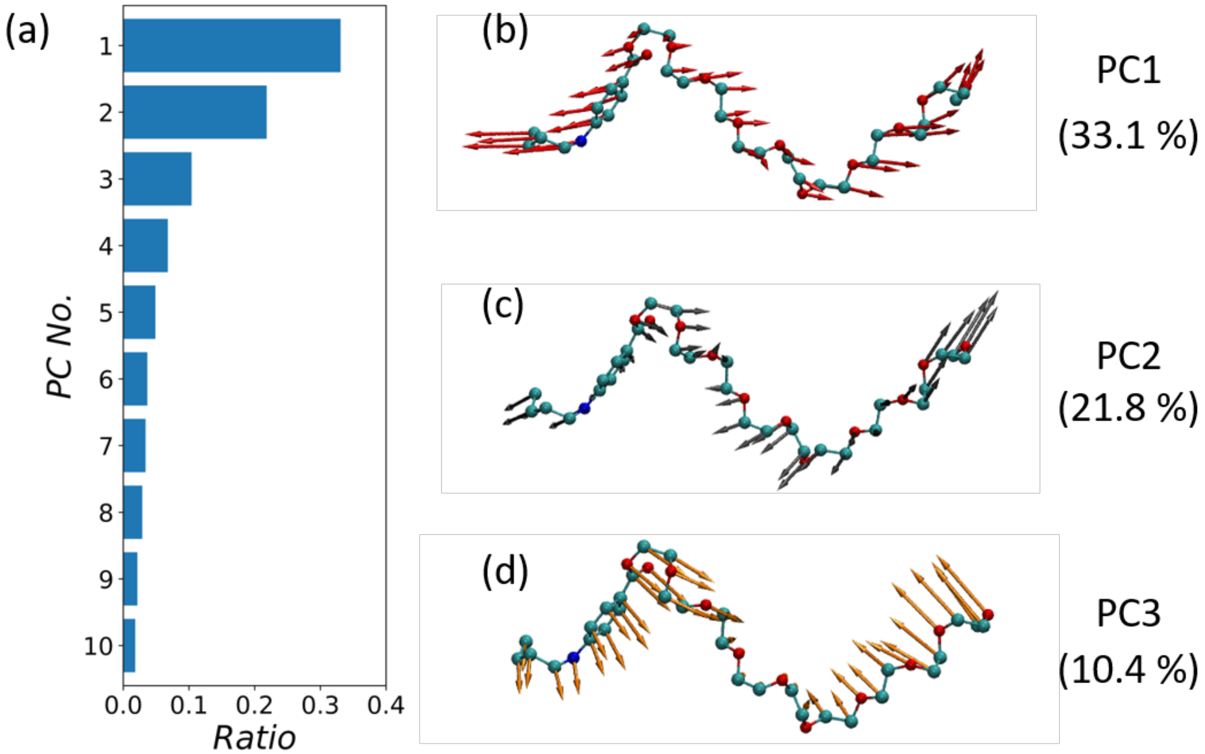

Figure S2: Principle component analysis (PCA). (a) Top 10 ranked principle component ratios; an example of (b) the first (c) second and (d) third mode derived from principle component analysis.

S10 Generation of Standard Descriptors

To disclose which contributes most to the prediction of partition coefficient ($\log P$), same as Gabriele et. al,[18] we also collected 100 “standard” descriptors directly from the RDKit package.[5] We collected descriptors in several aspects, such as the topological or topochemical descriptors (e.g. electrotopological state (EState),[19] BertzCT,[20] Balaban’s J value (BalabanJ),[21] Chi indexes and Kappa shape indexes[22]), surface area based descriptors, such as, Labute’s approximate surface area (LabuteASA) and hybrid descriptors which take the polarizability ($SMR{-}VSA_{k=1,2,\cdots,10}$), direct electrostatic interactions ($PEOE{-}VSA_{k=1,2,\cdots,14}$) and electrotopological state ($EState{-}VSA_{k=1,2,\cdots,11}$) into considerations,[23] topological polar surface area (TPSA),[24] and some simple and transparent descriptors, to name a few here, molecular weight, number of valance electrons, number of heavy atoms, number of NHs or OHs ($N_{NHs/OHs}$), number of Nitrogens and Oxygens, and so on. The full list could be found in Table S1.

S11 Feature selection

To better understand the 100 descriptors generated above, we utilize the univariate feature selection and intrinsic algorithm (Random Forests, RFs) to estimate the importance of each descriptor.

Univariate feature selection is a method to exclude noisy features. It examines each feature individually to estimate the strength between the feature and corresponding response variables. In this section, we utilize the Pearson's Correlation Coefficient as the statistical measures. The Pearson's Correlation Coefficient is defined as below:

$$\rho(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}} \tag{S9}$$

where X denotes the input space and Y denotes the variables we want to predict.

RF is an intrinsic algorithm which are capable of providing a measure of importance for each descriptor via mean decrease in impurity (MDI). It uses an impurity function $i(\tau)$ as a measurement of the probability of incorrectly classfying a randomly chosen element in the dataset. It is defined as below:

$$i(\tau) = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{S10}$$

where C is the number of classed in the dataset, and $p(i)$ is the probability of picking an element of class i.

To measure how well a potential splits at node τ will seperate the data, a value named Gini Gain $\Delta i(\tau)$ is defined. When a node is splited, it sends sample point to two sub-nodes, named left and right, and corresponding impurity denotes as $i(\tau_l)$ and $i(\tau_r)$. The Gini Gain is defined as below:

$$\Delta i(\tau) = i(\tau) - p_l i(\tau_l) - p_r i(\tau_r) \tag{S11}$$

The importance of each descriptor θ is the summation of Gini Gain of each descriptor ($\Delta i(\tau)$) over all nodes τ and trees T. In the expression of the mathematical form, it is written as below:

$$I(\theta) = \sum_{T} \sum_{\tau} \Delta i_{\theta}(\tau, T) \tag{S12}$$

We have used the model selection module and RandomForestRegressor embedded in the scikit-learn[13] package to implement these approaches.

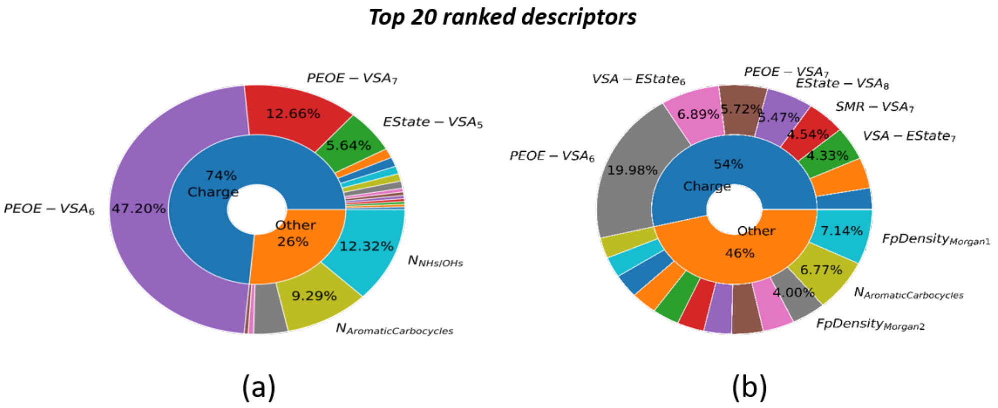

Figure S3: Top 20 ranked descriptors selected based on (a) mean decreased in impurity (MDI) and (b) univariate statistical test, which is implemented with the scikit-learn package.[13]

S12 An illustration of 5-5 neural network

For instance, the neural network shown in Figure S4 has 2 hidden layers and each layer has 5 nodes, short notation $5 - 5$ was utilized here to represent the neural network.

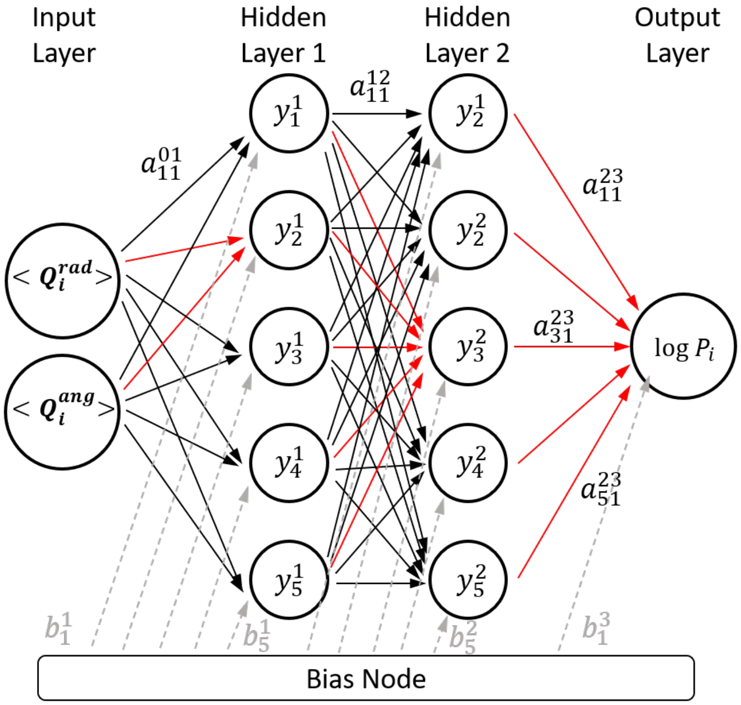

Figure S4: Schematic representation of a feed-forward neural network with 2 hidden layers and each layer has 5 nodes. The function form of the neural network which relates the input layer and output layer is given in eq. S14

As shown in Figure S4, the node i in layer k is connected to the node j in the adjecent layers l where $l = k + 1$ with a weight parameter a_{ij}^{kl} which is represented by the black arrow. The superscript starts from 0 and data flowed from input layer to output layer in one direction. In each node of hidden layer, a bias weight b_i^j was added as an adjustable offset for the activation funcont f_i^j, where i and j denote node and layer, respectively. As the red arrow line shown in Figure S4, the value y_i^j of node i in any hidden layer j was derived from the values of the N_{j-1} nodes in layer $j - 1$ together with the activation function f_i^j and bias weight b_i^j:

$$y_i^j = f_i^j(b_i^j + \sum_{k=1}^{N_{j-1}} a_{k,j}^{j-1,j} \cdot y_k^{j-1}) \tag{S13}$$

The mathematical form between the inpurt layer and output layer is given by the following equation:

$$\log P_i = f_1^3(b_i^3 + \sum_{k=1}^{5} a_{k1}^{23} \cdot f_k^2(b_k^2 + \sum_{j=1}^{5} a_{jk}^{12} \cdot f_j^1(b_j^1 + \sum_{i=1}^{N_{\mathbf{Q_i^{rad}},\mathbf{Q_i^{ang}}}} a_{ij}^{01} \cdot Q_i))) \tag{S14}$$

S13 Distribution of contribution from 4 distinct elements with different environment over datasets Star & Non-Star

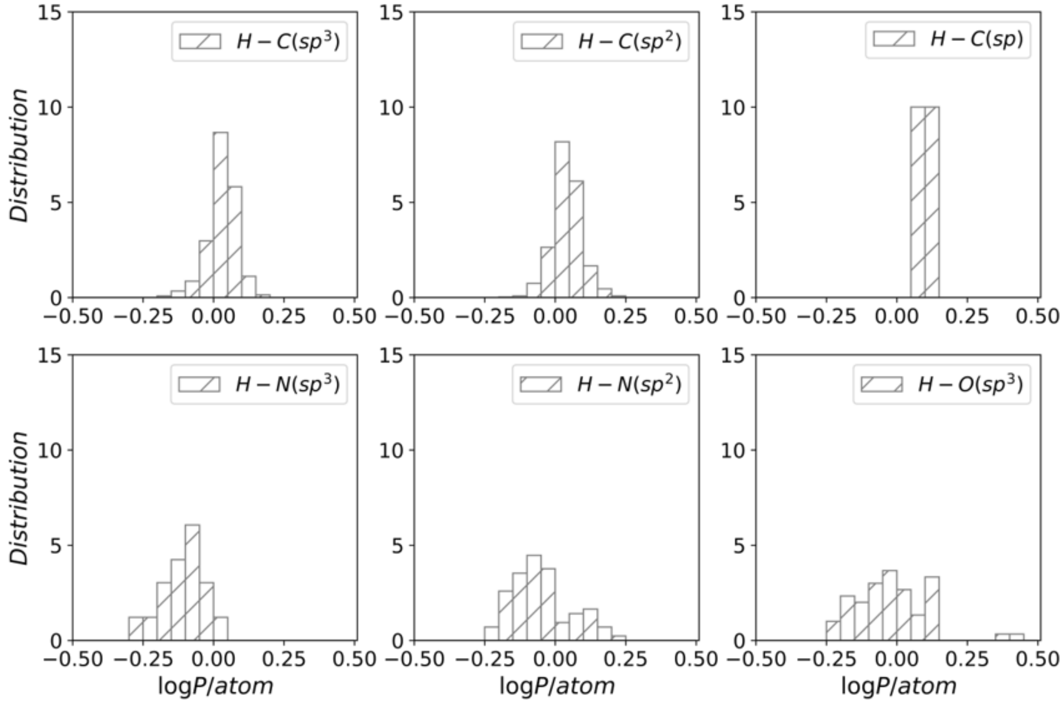

Figure S5: Distribution of contribution from atom H over datasets Star & Non-Star. These contributions were classified according to different surrounding environment.

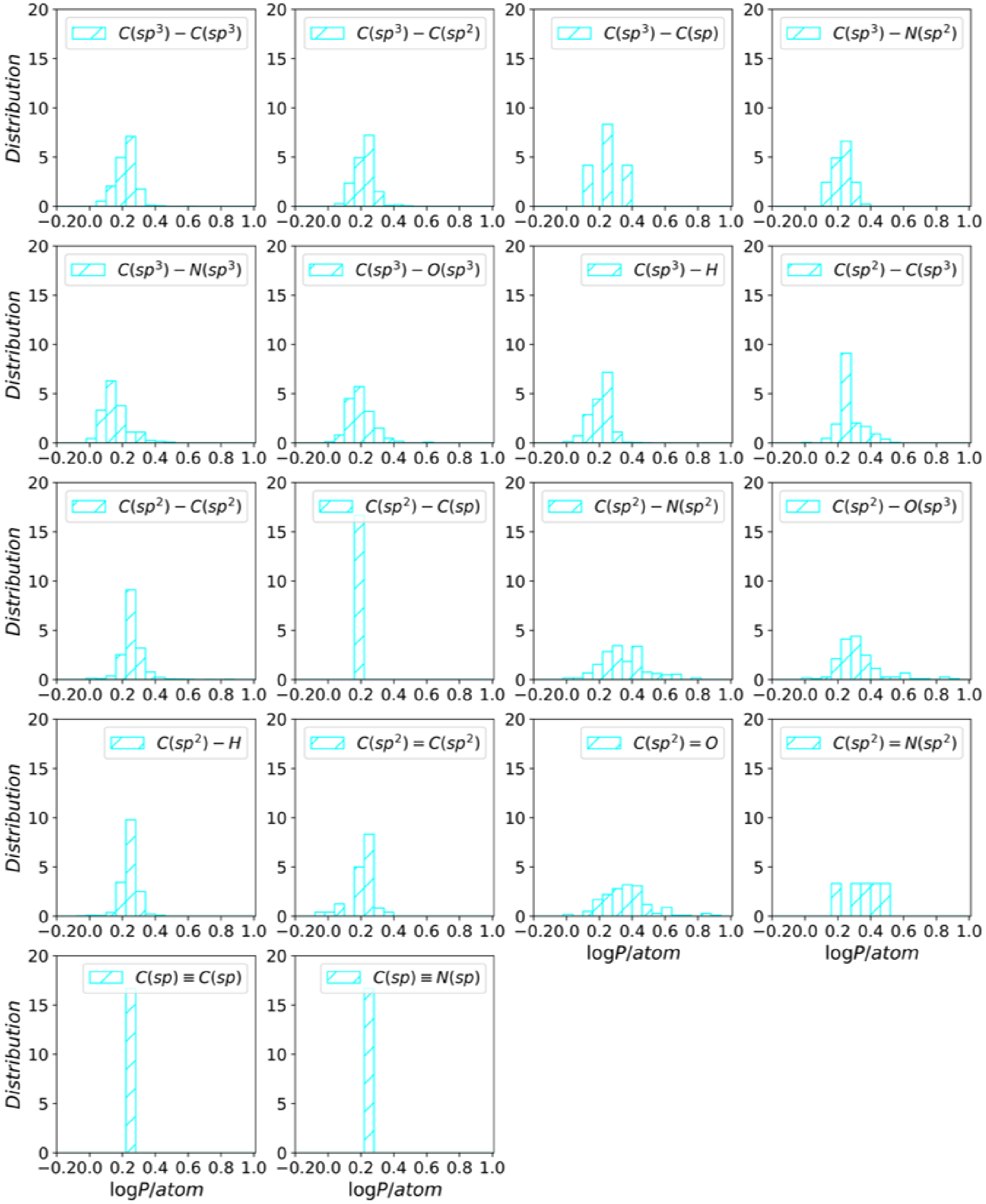

Figure S6: Distribution of contribution from atom C over datasets Star & Non-Star. These contributions were classified according to different surrounding environment.

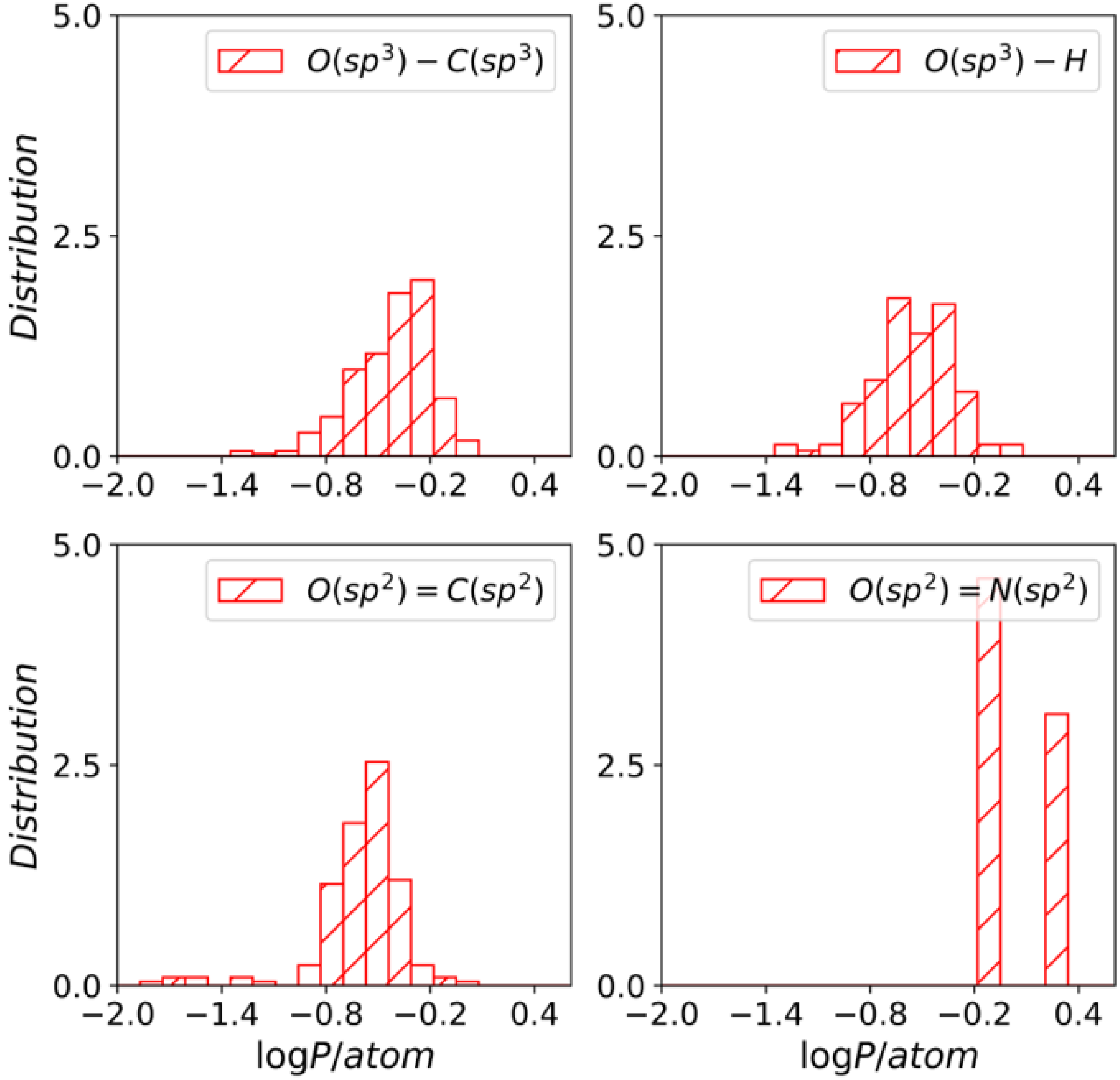

Figure S7: Distribution of contribution from atom O over datasets Star & Non-Star. These contributions were classified according to different surrounding environment.

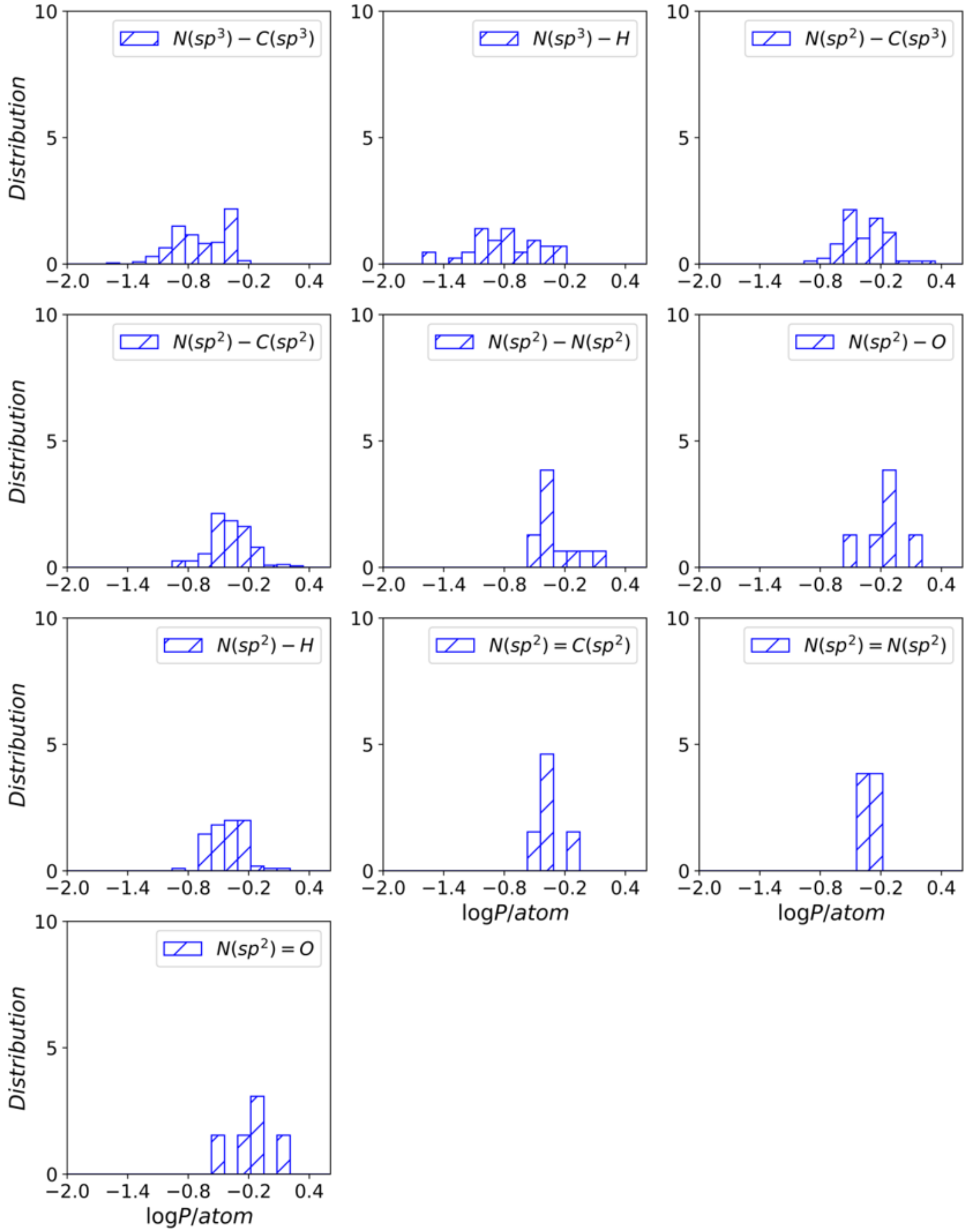

Figure S8: Distribution of contribution from atom N over datasets Star & Non-Star. These contributions were classified according to different surrounding environment.

Table S10: Contribution of 4 elements with distinct environments over datasets Star & Non-Star.

center Element	bonded Element	bond Type	mean	std.
H	C(sp^3)	single	0.030	0.050
	C(sp^2)	single	0.041	0.051
	C(sp)	single	0.114	0.026
	N(sp^3)	single	-0.108	0.073
	N(sp^2)	single	-0.051	0.105
	O(sp^3)	single	-0.016	0.127
C(sp^3)	C(sp^3)	single	0.221	0.059
	C(sp^2)	single	0.222	0.061
	C(sp)	single	0.239	0.090
	N(sp^2)	single	0.225	0.054
	N(sp^3)	single	0.159	0.081
	O(sp^3)	single	0.199	0.078
	H	single	0.209	0.063
C(sp^2)	C(sp^3)	single	0.276	0.076
	C(sp^2)	single	0.258	0.069
	C(sp)	single	0.179	0.000
	N(sp^2)	single	0.349	0.139
	O(sp^3)	single	0.329	0.140
	H	single	0.244	0.049
	C(sp^2)	double	0.215	0.074
	O	double	0.367	0.146
	N(sp^2)	double	0.345	0.101
C(sp)	C(sp)	triple	0.239	0.008
	N(sp)	triple	0.256	0.000
O(sp^3)	C(sp^3)	single	-0.422	0.226
	H	single	-0.554	0.233
O(sp^2)	C(sp^2)	double	-0.592	0.256
	N(sp^2)	double	0.059	0.185
N(sp^3)	C(sp^3)	single	-0.674	0.262
	H	single	-0.781	0.337
N(sp^2)	C(sp^3)	single	-0.372	0.210
	C(sp^2)	single	-0.392	0.202
	N(sp^2)	single	-0.280	0.185
	O	single	-0.170	0.195
	H	single	-0.406	0.170
	C(sp^2)	double	-0.371	0.128
	N(sp^2)	double	-0.347	0.061
	O	double	-0.172	0.213

References

(1) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(2) Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. International conference on machine learning. 2013; pp 1310–1318.

(3) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; `https://www.tensorflow.org/`, Software available from tensorflow.org.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **1997**, *23*, 3–25.

(5) Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

(6) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 1–14.

(7) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(8) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.

(9) Vassetti, D.; Pagliai, M.; Procacci, P. Assessment of GAFF2 and OPLS-AA general force fields in combination with the water models TIP3P, SPCE, and OPC3 for the solvation free energy of druglike organic molecules. *Journal of chemical theory and computation* **2019**, *15*, 1983–1995.

(10) Zhang, Z.; Liu, X.; Yan, K.; Tuckerman, M. E.; Liu, J. Unified efficient thermostat scheme for the canonical ensemble with holonomic or isokinetic constraints via molecular dynamics. *The Journal of Physical Chemistry A* **2019**, *123*, 6056–6079.

(11) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13*, e1005659.

(12) Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* **1982**, *28*, 129–137.

(13) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(14) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, *109*, 1528 – 1532.

(15) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian16 Revision A.01. 2016; Gaussian Inc. Wallingford CT.

(16) Beck, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys* **1993**, *98*, 5648–6.

(17) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **1988**, *37*, 785.

(18) Barnard, T.; Hagan, H.; Tseng, S.; Sosso, G. C. Less may be more: an informed reflection on molecular descriptors for drug design and discovery. *Molecular Systems Design & Engineering* **2020**,

(19) Hall, L. H.; Mohney, B.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *Journal of chemical information and computer sciences* **1991**, *31*, 76–82.

(20) Bertz, S. H. The first general index of molecular complexity. *Journal of the American Chemical Society* **1981**, *103*, 3599–3601.

(21) Balaban, A. T. Highly discriminating distance-based topological index. *Chemical physics letters* **1982**, *89*, 399–404.

(22) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in computational chemistry* **1991**, 367–422.

(23) Labute, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 464–477.

(24) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of medicinal chemistry* **2000**, *43*, 3714–3717.