Understanding the language of molecules: Predicting pure component parameters for the PC-SAFT equation of state from SMILES

Benedikt Winter
ETH Zurich, EPSE
Email: bewinter@ethz.ch

Philipp Rehner
ETH Zurich, EPSE
Email: prehner@ethz.ch

Timm Esper
University Stuttgart, ITT
Email: esper@itt.uni-stuttgart.de

Johannes Schilling
ETH Zurich, EPSE
Email: jschilling@ethz.ch

André Bardow*
ETH Zurich, EPSE
Email: abardow@ethz.ch

*Corresponding author. Email: abardow@ethz.ch

1 Vocab and token training

Table 1: List of tokens and their frequencies.

Token	Frequency	**Token**	Frequency	**Token**	Frequency
pad	-	**sos**	-	**mos**	-
eos	-	**c**	4169	**C**	13262
(	11020	**)**	11020	**1**	6145
O	8256	**2**	1545	**=**	6534
N	2384	**n**	650	**3**	240
[	1824	**]**	1824	**@**	345
H	609	**F**	1440	**-**	738
4	49	**S**	1413	**l**	1798
/	528	**s**	115	**o**	93
+	588	**5**	14	**#**	750
.	258	**B**	714	**r**	601
****	248	**P**	261	**6**	2
I	191	**a**	1	**i**	716
7	0	**e**	8	**K**	0
8	0	**L**	1	**A**	1
Z	3	**g**	0	**M**	0
T	11	**t**	0	**9**	0
p	0	**%**	0	**0**	0
V	0	**b**	12	**u**	11
R	5	**X**	0	**H2O**	0
d	3	**G**	1	**Y**	2
D	0	**y**	0	**W**	0
E	2	**f**	3	**h**	0
m	0	**U**	0		

2 Derivations of the implicit derivatives of liquid densities and vapor pressures

The equation for the calculation of a liquid density ρ_L for given temperature T and pressure p and model parameters $\boldsymbol{\phi}$ from an equation of state is

$$f(\rho_\mathrm{L}) = p(T, \rho_\mathrm{L}, \boldsymbol{\phi}) - p = 0 \tag{1}$$

It is then straightforward to use a Newton step to calculate the derivatives for an already converged value of the liquid density ρ_L^*

$$\rho_\mathrm{L}(T, p, \boldsymbol{\phi}) = \rho_\mathrm{L}^* - \frac{p(T, \rho_\mathrm{L}^*, \boldsymbol{\phi}) - p}{p_\rho(T, \rho_\mathrm{L}^*, \boldsymbol{\phi})} \tag{2}$$

For the calculation of the vapor pressure p^sat for given temperature T and model parameters $\boldsymbol{\phi}$ the equilibrium condition

$$f(p^\mathrm{sat}) = \mu_\mathrm{V}(T, p^\mathrm{sat}, \boldsymbol{\phi}) - \mu_\mathrm{L}(T, p^\mathrm{sat}, \boldsymbol{\phi}) = 0 \tag{3}$$

needs to be solved. Here, μ_V and μ_L are the chemical potentials of the liquid and the vapor phase. Applying the newton step from a known vapor pressure $p^{\mathrm{sat},*}$ yields

$$p^\mathrm{sat}(T, \boldsymbol{\phi}) = p^{\mathrm{sat},*} - \frac{\mu_\mathrm{V}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi}) - \mu_\mathrm{L}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi})}{v_\mathrm{V}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi}) - v_\mathrm{L}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi})} \tag{4}$$

where the thermodynamic identity $v = \left(\frac{\partial \mu}{\partial p}\right)_T$ was used. The expression can be further simplified by using the molar Helmholtz energy $a = -pv + \mu$, as

$$p^\mathrm{sat}(T, \boldsymbol{\phi}) = -\frac{a_\mathrm{V}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi}) - a_\mathrm{L}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi})}{v_\mathrm{V}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi}) - v_\mathrm{L}(T, p^{\mathrm{sat},*}, \boldsymbol{\phi})} \tag{5}$$

If the vapor pressure is calculated from an equation of state, the molar Helmholtz energy and volume are expressed in terms of the density, which then gives the expression shown in the main manuscript

$$p^\mathrm{sat}(T, \boldsymbol{\phi}) = -\frac{a(T, \rho_\mathrm{V}^*, \boldsymbol{\phi}) - a(T, \rho_\mathrm{L}^*, \boldsymbol{\phi})}{\frac{1}{\rho_\mathrm{V}^*} - \frac{1}{\rho_\mathrm{L}^*}} \tag{6}$$

3 Pseudocode for the PC-SAFT Head

Input Variables

- `x_in`: Shape $[B, E]$, where B is the batch size and E is the embedding dimension. Contains the final token embedding after multi-head attention.
- `T`: A tensor (or array) of shape $[B]$ (or a scalar) specifying the temperature(s) in Kelvin.
- `p`: A tensor (or array) of shape $[B]$ (or a scalar) specifying the pressure(s) in Pascal. Used only if `property` $= 2$ (i.e., for liquid density calculations).
- `property`: An integer flag indicating which thermodynamic property to compute:
  - `1` for vapor pressure
  - `2` for liquid density

Listing 1: Pseudocode for the PC-SAFT Head.

```
function PCSAFT_Head(x_in, T, p, property):

    # 1. Project input to auxiliary parameters
    x_in = layer_norm(x_in)
    a_aux = linear_layer(x_in)

    # 2. Scale and clamp PC-SAFT parameters
    phi_mean = [mean_m, mean_sigma, mean_epsilon, mean_eab,
       mean_kab, mean_u]
    x_physical = (1 + a_aux[:, 0:6] / 10) * phi_mean
    x_physical = clamp(x_physical, lower_bounds, upper_bounds)

    # 3. Compute likelihoods for association & polarity
    lambda_assoc = sigmoid(a_aux[:, 6])
    lambda_polar = sigmoid(a_aux[:, 7])

    # 4. Compute final PC-SAFT parameters
    m       = x_physical[:, 0]
    sigma   = x_physical[:, 1]
    epsilon = x_physical[:, 2]
    eab     = x_physical[:, 3] * lambda_assoc
    kab     = x_physical[:, 4] * lambda_assoc
    mu      = x_physical[:, 5] * (1 - lambda_assoc) *
       lambda_polar

    # 5. Select target property
    if property == 1:
        # Vapor pressure
        psat = pcsaft_vapor_pressure_solver(T, m, sigma, epsilon,
            mu, eab, kab)
        output_property = log(psat)
        status_flag = check_validity(psat)
```

```
    else if property == 2:
        # Liquid density
        rhoL = pcsaft_liquid_density_solver(T, p, m, sigma,
           epsilon, mu, eab, kab)
        output_property = log(rhoL)
        status_flag = check_validity(rhoL)

    # 6. Return property plus likelihoods
    x_out = [output_property, lambda_assoc, lambda_polar]
    return x_out, status_flag
```

Listing 2: PC-SAFT Solver Pseudocode

```
function helmholtz_energy(rho, m, sigma, epsilon, mu, kappa_ab,
    epsilon_ab, T):
    # ... Implementation of PC-SAFT Helmholtz energy ...
    return A(rho)

function pcsaft_liquid_density_solver(T, p, m, sigma, epsilon, mu
    , kappa_ab, epsilon_ab):
    for i in range(batch_size):
        density_guess = FeOs_liquid_density(
            T[i], p[i], m[i], sigma[i], epsilon[i], mu[i],
                kappa_ab[i], epsilon_ab[i]
        )
        if density_guess is valid:
            density[i] = density_guess
            status[i] = 1
        else:
            density[i] = 0
            status[i] = 0

    for i where status[i] == 1:
        _, p_calc, dpd_rho = derivatives(
            density[i], m[i], sigma[i], epsilon[i], mu[i],
                kappa_ab[i], epsilon_ab[i], T[i]
        )
        density[i] = density[i] - (p_calc - convert_pressure(p[i
            ], T[i])) / dpd_rho
        density[i] = convert_density_units(density[i])

    return density, status

function pcsaft_vapor_pressure_solver(T, m, sigma, epsilon, mu,
    kappa_ab, epsilon_ab):
    for i in range(batch_size):
        rhoL, rhoV = FeOs_phase_equilibrium(
            T[i], m[i], sigma[i], epsilon[i], mu[i], kappa_ab[i],
                epsilon_ab[i]
        )
        if valid(rhoV):
            rho_L[i] = rhoL
            rho_V[i] = rhoV
            status[i] = 1
        else:
            rho_L[i] = 0
            rho_V[i] = 0
            status[i] = 0

    for i where status[i] == 1:
        aL = helmholtz_energy(
            rho_L[i], m[i], sigma[i], epsilon[i], mu[i], kappa_ab
                [i], epsilon_ab[i], T[i]
```

```
        ) / rho_L[i]
        aV = helmholtz_energy(
            rho_V[i], m[i], sigma[i], epsilon[i], mu[i], kappa_ab
               [i], epsilon_ab[i], T[i]
        ) / rho_V[i]

        psat[i] = - ((aV - aL) + log(rho_V[i] / rho_L[i])) / ((1
           / rho_V[i]) - (1 / rho_L[i]))
        psat[i] = convert_pressure_units(psat[i], T[i])

    return psat, status
```

4 Example Input Data

Explanation of Columns

- `c0`: Composition (pure, here always 1).
- `c1`: Temperature (K).
- `c2`: Pressure (Pa).
- `c3`: Property (1 = pressure, 2 = density).
- `y0`: Predicted value (ln(p)/ln(roh)).
- `y1`: Associative (1 = associative, 0 = non-associative).
- `y2`: Polar (1 = polar, 0 = non-polar, -inf for missing data).
- `w0`: Weight in loss function (here always 1).

SMILES	**c0**	**c1 (K)**	**c2 (Pa)**	**c3**	**y0**	**y1**	**y2**	**w0**
CCCCC(CC)CCC(C)O	1	473.15	0	1	10.8072	1	-inf	1
CCCCC(CC)CCC(C)O	1	498.15	0	1	11.5154	1	-inf	1
CC(=O)C(C)O	1	421.53	0	1	11.5261	1	1	1
CC(=O)C(C)O	1	414.15	0	1	11.5261	1	1	1
N#CCl	1	261.15	0	1	10.3438	0	-inf	1
N#CCl	1	263.15	0	1	10.4725	0	-inf	1
CC(=N)O	1	404.85	101325	2	2.7838	1	-inf	1
CC#N	1	298.15	1778000	2	2.9437	0	-inf	1
c1ccccc1	1	298.15	101325	2	2.4147	1	0	1

Table 2: Example Input Data.

5 Data Cleaning

In the initial dataset processing, data points with a vapor pressure of 1.00(1) bar at 298.15(100) K were excluded. These entries were deemed to have been erroneously entered. Further refinement was made by removing data points incompatible with the PC-SAFT fitting.

For this refinement, eight models were trained on the dataset over 15 epochs utilizing a SmoothL1 loss function. This particular loss function was chosen to mitigate the undue influence of outliers compared to using an MSE loss. Subsequent to this training, any data points that exhibited a training loss greater than 0.5 in the remaining dataset were removed.

As a result of this data cleaning process, 21 456 out of 233 988 data points were removed. This left the remaining dataset with 160 186 vapor pressure data points with to 12 019 molecules and 52 343 liquid density data points corresponding to 2067 molecules. Manual inspection of the discarded data points largely confirmed the removal of data that appeared unreasonable or anomalous. For examples, see Figure S1. In the course of the data cleaning, several types of errors were identified. Primarily were data series that appeared erroneous, either due to their stark deviation from expected values or patterns inconsistent with related data (Figure S1a-S1b). Additionally, we identified instances of data scattering, which could potentially hinder model training (Figure S1c-S1d).

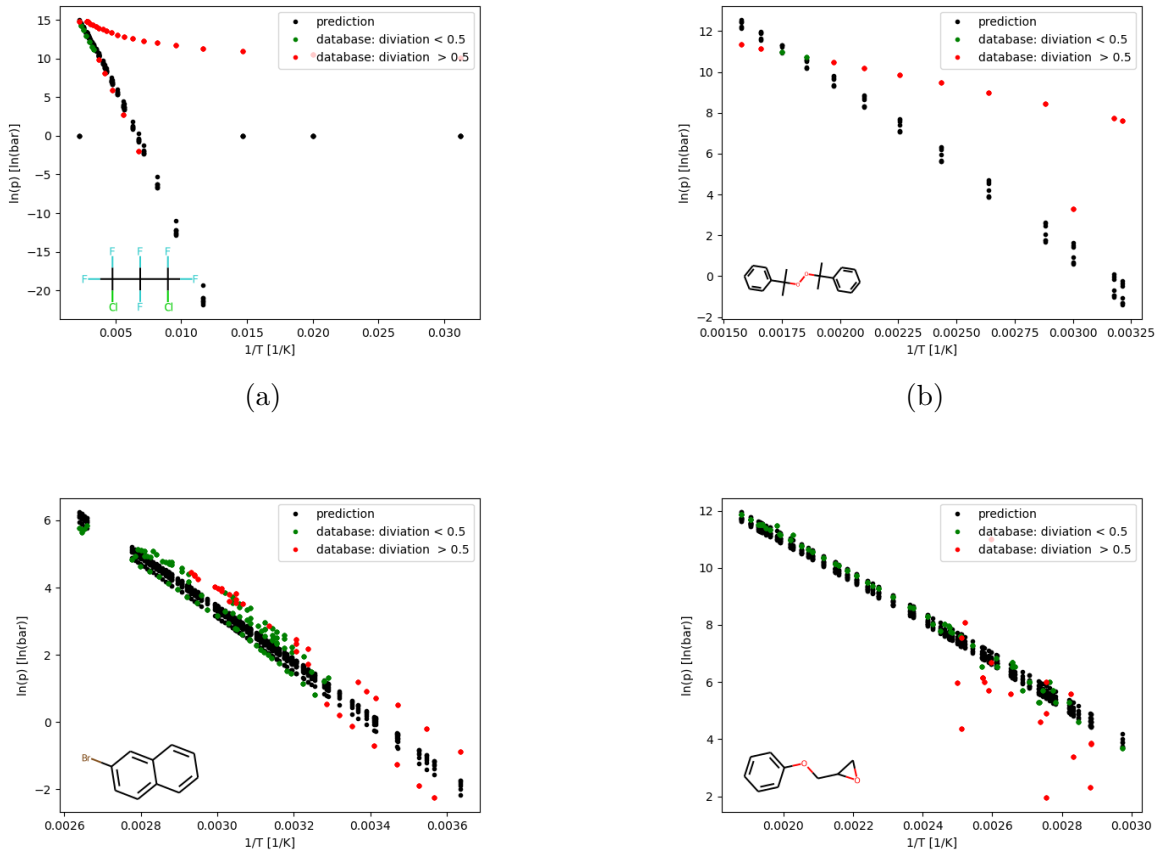
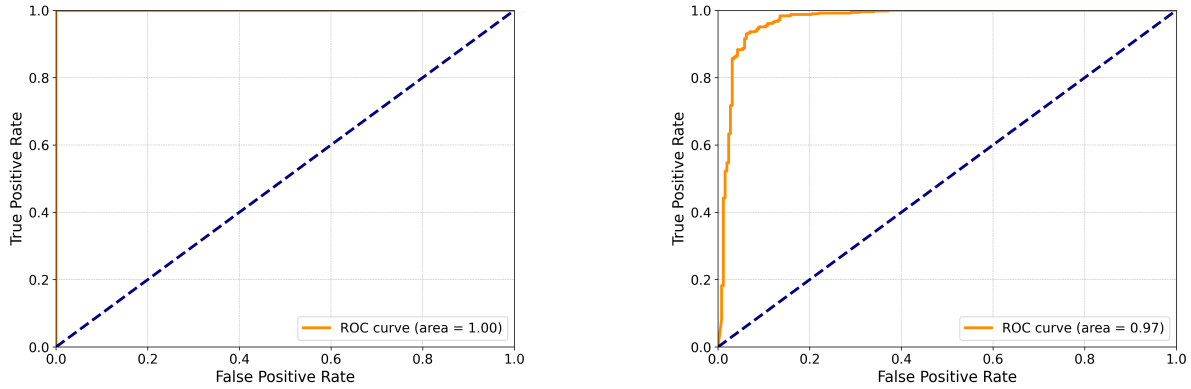


Figure S1: Examples for topical errors in the remaining data identified by our cleaning step. If the prediction for vapor pressure is 0, the predicted state is supercritical.

6 Receiver operating characteristic curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of binary classification models. It plots the true positive rate against the false positive rate at various threshold settings, illustrating the trade-off between sensitivity and specificity. An area under the ROC curve (AUC) value of 1 represents a perfect model, while an AUC value of 0.5 suggests a model no better than random guessing.



(a) association likelihood

(b) polarity likelihood

Figure S2: ROC curves for the association and polarity likelihood

In our analysis, the ROC curve for the association likelihood exhibits an impeccable performance, with an AUC of 1 (Figure S2a). This suggests that the model can flawlessly differentiate between the positive and negative classes for this likelihood. Given that association is determined by the presence of hydrogen bonding groups, which can be easily discerned from SMILES codes, this outcome is not surprising. Conversely, the ROC curve for the polarity likelihood, though not perfect, demonstrates still an outstanding performance, with an AUC of 0.97 (Figure S2b). Since polarity is determined based on discrete decisions within a continuous space, a perfect representation should not be expected.

7 Prediction of dipole moment

Figure S3 presents a comparison of dipole moments predicted by COSMO-RS and SPT-PCSAFT. SPT-PCSAFT consistently predicts higher dipole moments than COSMO-RS. During training, it was observed that the average dipole moment depends on its initialization value, with minimal impact on the final model performance. This indicates that SPT-PCSAFT, in general, exhibits low sensitivity to the dipole moment.

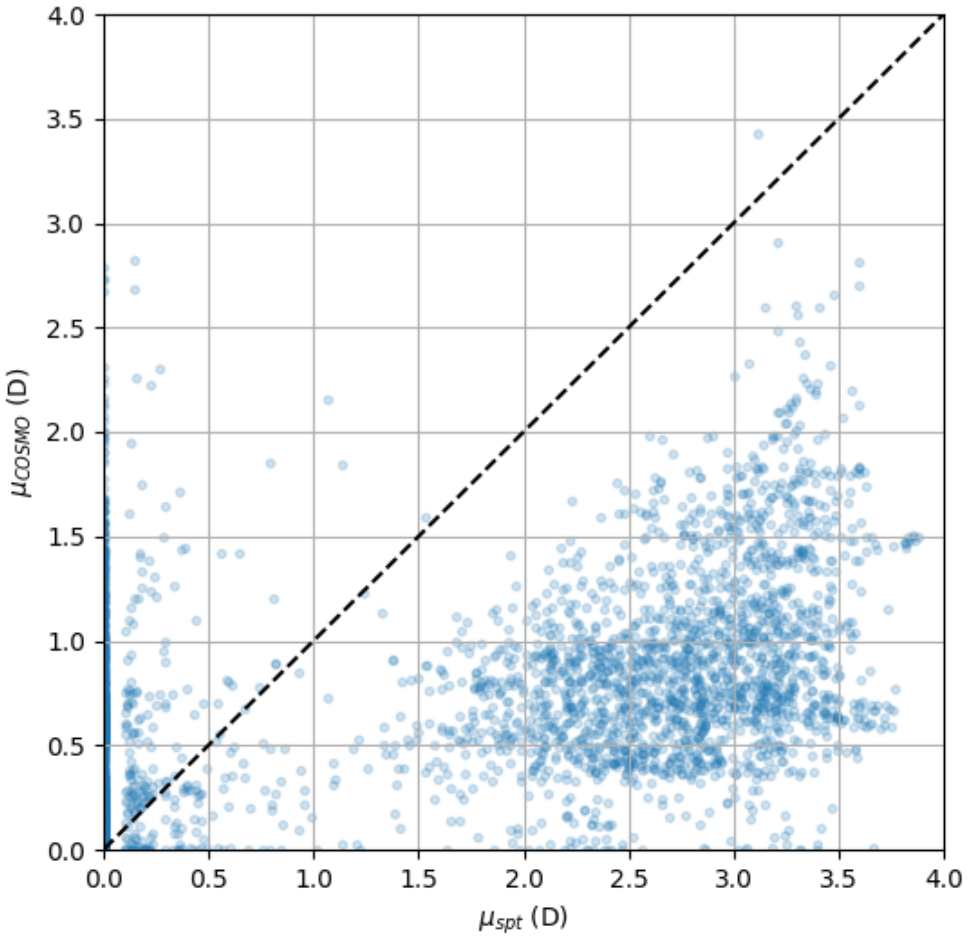


Figure S3: Comparison of dipole moments predicted by COSMO-RS and SPT-PCSAFT.

8 Molecular Families

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
lauricacid	CCCCCCCCCCCC(=O)O	Acids
pentanoicacid	CCCCC(=O)O	Acids
tridecanoicacid	CCCCCCCCCCCCC(=O)O	Acids
heptanoicacid	CCCCCCC(=O)O	Acids
methacrylicacid	C=C(C)C(=O)O	Acids
caprylicacid	CCCCCCCC(=O)O	Acids
nonanoicacid	CCCCCCCCC(=O)O	Acids
capricacid	CCCCCCCCCC(=O)O	Acids
hexanoicacid	CCCCCC(=O)O	Acids
3-methylbutanoicacid	CC(C)CC(=O)O	Acids
arachidicacid	CCCCCCCCCCCCCCCCCCCC(=O)O	Acids
2-ethylhexanoicacid	CCCCC(CC)C(=O)O	Acids
undecanoicacid	CCCCCCCCCCC(=O)O	Acids
myristicacid	CCCCCCCCCCCCCC(=O)O	Acids
pentadecanoicacid	CCCCCCCCCCCCCCC(=O)O	Acids
stearicacid	CCCCCCCCCCCCCCCCCC(=O)O	Acids
2-methylpropanoicacid	CC(C)C(=O)O	Acids
palmiticacid	CCCCCCCCCCCCCCCC(=O)O	Acids
butyricacid	CCCC(=O)O	Acids
pivalicacid	CC(C)(C)C(=O)O	Acids
2-heptanol	CCCCCC(C)O	Alcohols
1,3-butanediol	CC(O)CCO	Alcohols
2-methyl-2,4-pentanediol	CC(O)CC(C)(C)O	Alcohols
2,3-butanediol	CC(O)C(C)O	Alcohols
1-butanol	CCCCO	Alcohols
tert-butanol	CC(C)(C)O	Alcohols
2-octanol	CCCCCCC(C)O	Alcohols
2-methyl-1-propanol	CC(C)CO	Alcohols
2,2,4-trimethyl-1,3-pentanediol	CC(C)C(O)C(C)(C)CO	Alcohols
3-methyl-1-butanol	CC(C)CCO	Alcohols
1-pentanol	CCCCCO	Alcohols
1-hexanol	CCCCCCO	Alcohols
1-heptanol	CCCCCCCO	Alcohols
1-octanol	CCCCCCCCO	Alcohols
1-nonanol	CCCCCCCCCO	Alcohols
1-undecanol	CCCCCCCCCCCO	Alcohols
1-dodecanol	CCCCCCCCCCCCO	Alcohols
n-tridecanol	CCCCCCCCCCCCCO	Alcohols
1-tetradecanol	CCCCCCCCCCCCCCO	Alcohols
1-hexadecanol	CCCCCCCCCCCCCCCCO	Alcohols
1-octadecanol	CCCCCCCCCCCCCCCCCCO	Alcohols
3-methyl-2-butanol	CC(C)C(C)O	Alcohols
4-methyl-2-pentanol	CC(C)CC(C)O	Alcohols
1-decanol	CCCCCCCCCCO	Alcohols
diethyleneglycol	OCCOCCO	Alcohols
triethyleneglycol	OCCOCCOCCO	Alcohols
3-methyl-3-pentanol	CCC(C)(O)CC	Alcohols
2-methyl-1-pentanol	CCCC(C)CO	Alcohols

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
2-methyl-1-butanol	CCC(C)CO	Alcohols
2-ethyl-1-hexanol	CCCCC(CC)CO	Alcohols
2-pentanol	CCCC(C)O	Alcohols
3-hexanol	CCCC(O)CC	Alcohols
cyclohexanol	OC1CCCCC1	Alcohols
1,2-dihydroxybenzene¡pyrocatechol¿	Oc1ccccc1O	Alcohols
1,10-decanediol	OCCCCCCCCCCO	Alcohols
2-butanol	CCC(C)O	Alcohols
1,3-dihydroxybenzene¡resorcinol¿	Oc1cccc(O)c1	Alcohols
1,6-hexanediol	OCCCCCCO	Alcohols
1,2-butanediol	CCC(O)CO	Alcohols
1,4-butanediol	OCCCCO	Alcohols
2-methyl-1-hexanol	CCCCC(C)CO	Alcohols
3-pentanol	CCC(O)CC	Alcohols
2-hexanol	CCCCC(C)O	Alcohols
tetraethyleneglycol	OCCOCCOCCOCCO	Alcohols
tert-pentanol	CCC(C)(C)O	Alcohols
3-methylbutyraldehyde	CC(C)CC=O	Aldehydes
2-methylpropanal	CC(C)C=O	Aldehydes
n-decanal	CCCCCCCCCC=O	Aldehydes
valeraldehyde	CCCCC=O	Aldehydes
methacrolein	C=C(C)C=O	Aldehydes
nonanal	CCCCCCCCC=O	Aldehydes
hexanal	CCCCCC=O	Aldehydes
undecanaldehyde	CCCCCCCCCCC=O	Aldehydes
octanal	CCCCCCCC=O	Aldehydes
heptanal	CCCCCCC=O	Aldehydes
2-methylbutyraldehyde	CCC(C)C=O	Aldehydes
2-phenylpropionaldehyde	CC(C=O)c1ccccc1	Aldehydes
butyraldehyde	CCCC=O	Aldehydes
2,3-dimethylbutane	CC(C)C(C)C	Alkanes
dodecane	CCCCCCCCCCCC	Alkanes
2-methylpropane	CC(C)C	Alkanes
n-butane	CCCC	Alkanes
n-undecane	CCCCCCCCCCC	Alkanes
2,3-dimethylpentane	CCC(C)C(C)C	Alkanes
2,2-dimethylpentane	CCCC(C)(C)C	Alkanes
2,2,3-trimethylbutane	CC(C)C(C)(C)C	Alkanes
2,4-dimethylhexane	CCC(C)CC(C)C	Alkanes
3-methylpentane	CCC(C)CC	Alkanes
2,3-dimethylhexane	CCCC(C)C(C)C	Alkanes
2-methylpentane	CCCC(C)C	Alkanes
hexane	CCCCCC	Alkanes
2-methylheptane	CCCCCC(C)C	Alkanes
3-methylheptane	CCCCC(C)CC	Alkanes
2-methylhexane	CCCCC(C)C	Alkanes
pentane	CCCCC	Alkanes
heptane	CCCCCCC	Alkanes
2-methyloctane	CCCCCCC(C)C	Alkanes
octane	CCCCCCCC	Alkanes
nonane	CCCCCCCCC	Alkanes

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
2-methylbutane	CCC(C)C	Alkanes
4-methylheptane	CCCC(C)CCC	Alkanes
decane	CCCCCCCCCC	Alkanes
3,3-dimethylpentane	CCC(C)(C)CC	Alkanes
2,7-dimethyloctane	CC(C)CCCCC(C)C	Alkanes
squalane	CC(C)CCCC(C)CCCC(C)CCCCC(C)CCCC(C)CCCC(C)C	Alkanes
3-methylhexane	CCCC(C)CC	Alkanes
2,5-dimethylhexane	CC(C)CCC(C)C	Alkanes
2,2,5-trimethylhexane	CC(C)CCC(C)(C)C	Alkanes
2,4-dimethylpentane	CC(C)CC(C)C	Alkanes
2,2,4-trimethylpentane	CC(C)CC(C)(C)C	Alkanes
2,2-dimethylbutane	CCC(C)(C)C	Alkanes
2,2,4,4,6,8,8-heptamethylnonane	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C	Alkanes
octadecane	CCCCCCCCCCCCCCCCCC	Alkanes
tetradecane	CCCCCCCCCCCCCC	Alkanes
tridecane	CCCCCCCCCCCCC	Alkanes
octacosane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	Alkanes
heptacosane	CCCCCCCCCCCCCCCCCCCCCCCCCCCC	Alkanes
pentacosane	CCCCCCCCCCCCCCCCCCCCCCCCCC	Alkanes
tetracosane	CCCCCCCCCCCCCCCCCCCCCCCCC	Alkanes
eicosane	CCCCCCCCCCCCCCCCCCCC	Alkanes
nonadecane	CCCCCCCCCCCCCCCCCCC	Alkanes
3-ethylpentane	CCC(CC)CC	Alkanes
heptadecane	CCCCCCCCCCCCCCCCC	Alkanes
2,2-dimethylpropane	CC(C)(C)C	Alkanes
hexadecane	CCCCCCCCCCCCCCCC	Alkanes
pentadecane	CCCCCCCCCCCCCCC	Alkanes
3,3-diethylpentane	CCC(CC)(CC)CC	Alkanes
4-methyl-1-pentene	C=CCC(C)C	Alkenes
1-butene	C=CCC	Alkenes
1,3-butadiene	C=CC=C	Alkenes
3-methyl-1-hexene	C=CC(C)CCC	Alkenes
3-methyl-1-butene	C=CC(C)C	Alkenes
2,3-dimethyl-2-butene	CC(C)=C(C)C	Alkenes
dimethylallene	C=C=C(C)C	Alkenes
2,4,4-trimethyl-1-pentene	C=C(C)CC(C)(C)C	Alkenes
1,4-pentadiene	C=CCC=C	Alkenes
2-methyl-1-butene	C=C(C)CC	Alkenes
2,3-dimethyl-1-butene	C=C(C)C(C)C	Alkenes
2-methyl-2-butene	CC=C(C)C	Alkenes
isobutylene	C=C(C)C	Alkenes
1,4-cyclohexadiene	C1=CCC=CC1	Alkenes
1,3-cyclopentadiene	C1=CCC=C1	Alkenes
2-methyl-2-pentene	CCC=C(C)C	Alkenes
2-methyl-1-pentene	C=C(C)CCC	Alkenes
1-pentene	C=CCCC	Alkenes
isoprene	C=CC(=C)C	Alkenes

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
1-hexene	C=CCCCC	Alkenes
1,5-hexadiene	C=CCCC=C	Alkenes
1-hexadecene	C=CCCCCCCCCCCCCCC	Alkenes
2,4,4-trimethyl-2-pentene	CC(C)=CC(C)(C)C	Alkenes
1-tetradecene	C=CCCCCCCCCCCCC	Alkenes
1-dodecene	C=CCCCCCCCCCC	Alkenes
1-undecene	C=CCCCCCCCCC	Alkenes
1-octadecene	C=CCCCCCCCCCCCCCCCC	Alkenes
1-decene	C=CCCCCCCCC	Alkenes
1-nonene	C=CCCCCCCC	Alkenes
1-octene	C=CCCCCC	Alkenes
1-heptene	C=CCCCC	Alkenes
2-butyne	CC#CC	Alkynes
1-butyne	C#CCC	Alkynes
1-pentyne	C#CCCC	Alkynes
1-hexyne	C#CCCCC	Alkynes
3-hexyne	CCC#CCC	Alkynes
1-heptyne	C#CCCCCC	Alkynes
1-octyne	C#CCCCCC	Alkynes
1-nonyne	C#CCCCCCC	Alkynes
1-buten-3-yne	C#CC=C	Alkynes
2-hexyne	CC#CCCC	Alkynes
diethylenetriamine	NCCNCCN	Amines
n-(2-aminoethyl)ethanolamine	NCCNCCO	Amines
pyrrole	c1cc[nH]c1	Amines
triethanolamine(tea)	OCCN(CCO)CCO	Amines
2,2'-diethanolamine(dea)	OCCNCCO	Amines
1-octanamine	CCCCCCCCN	Amines
n-(2-aminoethyl)piperazine	NCCN1CCNCC1	Amines
indole	c1ccc2[nH]ccc2c1	Amines
diglycolamine(dga)	NCCOCCO	Amines
1,6-hexanediamine	NCCCCCCN	Amines
n-methylcyclohexylamine	CNC1CCCCC1	Amines
cyclopentylamine	NC1CCCC1	Amines
tributylamine	CCCCN(CCCC)CCCC	Amines
dibutylamine	CCCCNCCCC	Amines
dipropylamine	CCCNCCC	Amines
methyldiethylamine	CCN(C)CC	Amines
triethylamine	CCN(CC)CC	Amines
n,n-diethylethanolamine(demea)	CCN(CC)CCO	Amines
n,n-diethylamine	CCNCC	Amines
3-n,n-dimethylaminopropylamine	CN(C)CCCN	Amines
n,n,n',n'-tetramethylethylene diamine	CN(C)CCN(C)C	Amines
n,n-dimethylethanolamine(dmmea)	CN(C)CCO	Amines
butylamine	CCCCN	Amines
1-aminopentane	CCCCCN	Amines
hexylamine	CCCCCCN	Amines
heptylamine	CCCCCCCN	Amines

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
methyldiethanolamine(mdea)	CN(CCO)CCO	Amines
n-methylpiperidine	CN1CCCCC1	Amines
trioctylamine	CCCCCCCCN(CCCCCCCC)CCCCCCCC	Amines
cyclohexylamine	NC1CCCCC1	Amines
tripropylamine	CCCN(CCC)CCC	Amines
pyrazine	c1cnccn1	Amines
tert-butylamine	CC(C)(C)N	Amines
piperidine	C1CCNCC1	Amines
diisopropanolamine(dipa)	CC(O)CNCC(C)O	Amines
diisopropylamine	CC(C)NC(C)C	Amines
dicyclohexylamine	C1CCC(NC2CCCCC2)CC1	Amines
hexamethyleneimine	C1CCCNCC1	Amines
n,n-dimethylacetamide	CC(=O)N(C)C	Amines
2-amino-2-methyl-1-propanol (amp)	CC(C)(N)CO	Amines
2-methylpropylamine	CC(C)CN	Amines
pyrrolidine	C1CCNC1	Amines
sec-butylamine	CCC(C)N	Amines
alpha-methylstyrene	C=C(C)c1ccccc1	Aromatic hydrocarbons
p-methylaniline	Cc1ccc(N)cc1	Aromatic hydrocarbons
3,5-dimethylphenol	Cc1cc(C)cc(O)c1	Aromatic hydrocarbons
phenylhydrazine	NNc1ccccc1	Aromatic hydrocarbons
hexylbenzene	CCCCCCc1ccccc1	Aromatic hydrocarbons
aniline	Nc1ccccc1	Aromatic hydrocarbons
2,4-dimethylaniline	Cc1ccc(N)c(C)c1	Aromatic hydrocarbons
p-chlorotoluene	Cc1ccc(Cl)cc1	Aromatic hydrocarbons
heptylbenzene	CCCCCCCc1ccccc1	Aromatic hydrocarbons
p-xylene	Cc1ccc(C)cc1	Aromatic hydrocarbons
aceticacidbenzylester	CC(=O)OCc1ccccc1	Aromatic hydrocarbons
benzoicacid	O=C(O)c1ccccc1	Aromatic hydrocarbons
p-methyl-styrene	C=Cc1ccc(C)cc1	Aromatic hydrocarbons
octylbenzene	CCCCCCCCc1ccccc1	Aromatic hydrocarbons
1,3-diethylbenzene	CCc1cccc(CC)c1	Aromatic hydrocarbons
3-methylstyrene¡3-vinyltoluene¿	C=Cc1cccc(C)c1	Aromatic hydrocarbons
styrene	C=Cc1ccccc1	Aromatic hydrocarbons
benzylbenzoate	O=C(OCc1ccccc1)c1ccccc1	Aromatic hydrocarbons
nonylbenzene	CCCCCCCCCc1ccccc1	Aromatic hydrocarbons
1,2,4-trimethylbenzene	Cc1ccc(C)c(C)c1	Aromatic hydrocarbons
1-phenyldecane	CCCCCCCCCCc1ccccc1	Aromatic hydrocarbons
4-isopropyltoluene	Cc1ccc(C(C)C)cc1	Aromatic hydrocarbons
1-phenylundecane	CCCCCCCCCCCc1ccccc1	Aromatic hydrocarbons
2-ethylphenol	CCc1ccccc1O	Aromatic hydrocarbons
1-phenyldodecane	CCCCCCCCCCCCc1ccccc1	Aromatic hydrocarbons
1-ethyl-2-methylbenzene	CCc1ccccc1C	Aromatic hydrocarbons
2,4,6-trimethylpyridine	Cc1cc(C)nc(C)c1	Aromatic hydrocarbons
1-ethyl-4-methylbenzene	CCc1ccc(C)cc1	Aromatic hydrocarbons
1-ethyl-3-methylbenzene	CCc1cccc(C)c1	Aromatic hydrocarbons
1-ethylnaphthalene	CCc1cccc2ccccc12	Aromatic hydrocarbons
2,5-dimethylphenol	Cc1ccc(C)c(O)c1	Aromatic hydrocarbons
alpha-aminotoluene	NCc1ccccc1	Aromatic hydrocarbons
pentylbenzene	CCCCCc1ccccc1	Aromatic hydrocarbons

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
phthalicaciddiethylester	CCOC(=O)c1ccccc1C(=O)OCC	Aromatic hydrocarbons
ethylbenzoate	CCOC(=O)c1ccccc1	Aromatic hydrocarbons
2,6-dimethylpyridine	Cc1cccc(C)n1	Aromatic hydrocarbons
phenylaceticacidethylester	CCOC(=O)Cc1ccccc1	Aromatic hydrocarbons
3-methylpyridine	Cc1cccnc1	Aromatic hydrocarbons
2-methylpyridine	Cc1ccccn1	Aromatic hydrocarbons
2-methylphenol	Cc1ccccc1O	Aromatic hydrocarbons
1,2,3-trimethylbenzene	Cc1cccc(C)c1C	Aromatic hydrocarbons
o-methylaniline	Cc1ccccc1N	Aromatic hydrocarbons
o-xylene	Cc1ccccc1C	Aromatic hydrocarbons
propylbenzene	CCCc1ccccc1	Aromatic hydrocarbons
m-methylaniline	Cc1cccc(N)c1	Aromatic hydrocarbons
n,n-diethylaniline	CCN(CC)c1ccccc1	Aromatic hydrocarbons
toluene	Cc1ccccc1	Aromatic hydrocarbons
1-methylnaphthalene	Cc1cccc2ccccc12	Aromatic hydrocarbons
n-ethylaniline	CCNc1ccccc1	Aromatic hydrocarbons
o-chlorotoluene	Cc1ccccc1Cl	Aromatic hydrocarbons
sec-butylbenzene	CCC(C)c1ccccc1	Aromatic hydrocarbons
m-xylene	Cc1cccc(C)c1	Aromatic hydrocarbons
butylbenzene	CCCCc1ccccc1	Aromatic hydrocarbons
4-methylphenol	Cc1ccc(O)cc1	Aromatic hydrocarbons
o-dichlorobenzene	Clc1ccccc1Cl	Aromatic hydrocarbons
chlorobenzene	Clc1ccccc1	Aromatic hydrocarbons
3,4-dimethylphenol	Cc1ccc(O)cc1C	Aromatic hydrocarbons
2,7-dimethylnaphthalene	Cc1ccc2ccc(C)cc2c1	Aromatic hydrocarbons
2-methylnaphthalene	Cc1ccc2ccccc2c1	Aromatic hydrocarbons
phthalicaciddibutylester	CCCCOC(=O)c1ccccc1C(=O)OCCCC	Aromatic hydrocarbons
4-methylpyridine	Cc1ccncc1	Aromatic hydrocarbons
quinaldine	Cc1ccc2ccccc2n1	Aromatic hydrocarbons
1-chloronaphthalene	Clc1cccc2ccccc12	Aromatic hydrocarbons
1,2,3-trichlorobenzene	Clc1cccc(Cl)c1Cl	Aromatic hydrocarbons
m-dichlorobenzene	Clc1cccc(Cl)c1	Aromatic hydrocarbons
p-dichlorobenzene	Clc1ccc(Cl)cc1	Aromatic hydrocarbons
1,2,4-trichlorobenzene	Clc1ccc(Cl)c(Cl)c1	Aromatic hydrocarbons
1,3,5-trichlorobenzene	Clc1cc(Cl)cc(Cl)c1	Aromatic hydrocarbons
benzylchloride	ClCc1ccccc1	Aromatic hydrocarbons
3-isopropyltoluene	Cc1cccc(C(C)C)c1	Aromatic hydrocarbons
2,4-dimethylphenol	Cc1ccc(O)c(C)c1	Aromatic hydrocarbons
ethylbenzene	CCc1ccccc1	Aromatic hydrocarbons
3-methylphenol	Cc1cccc(O)c1	Aromatic hydrocarbons
phenol	Oc1ccccc1	Aromatic hydrocarbons
isoquinoline	c1ccc2cnccc2c1	Aromatic hydrocarbons
phenylacetylene	C#Cc1ccccc1	Aromatic hydrocarbons
isobutylbenzene	CC(C)Cc1ccccc1	Aromatic hydrocarbons
p-terphenyl	c1ccc(-c2ccc(-c3ccccc3)cc2)cc1	Aromatic hydrocarbons
1,3,5-triisopropylbenzene	CC(C)c1cc(C(C)C)cc(C(C)C)c1	Aromatic hydrocarbons
naphthalene	c1ccc2ccccc2c1	Aromatic hydrocarbons
1,3-diisopropylbenzene	CC(C)c1cccc(C(C)C)c1	Aromatic hydrocarbons
m-terphenyl	c1ccc(-c2cccc(-c3ccccc3)c2)cc1	Aromatic hydrocarbons
acridine	c1ccc2nc3ccccc3cc2c1	Aromatic hydrocarbons
isopropylbenzene	CC(C)c1ccccc1	Aromatic hydrocarbons
triphenylmethane	c1ccc(C(c2ccccc2)c2ccccc2)cc1	Aromatic hydrocarbons
phenylcyclohexane	c1ccc(C2CCCCC2)cc1	Aromatic hydrocarbons

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
1-phenyl-2-propanol	CC(O)Cc1ccccc1	Aromatic hydrocarbons
diphenylmethane	c1ccc(Cc2ccccc2)cc1	Aromatic hydrocarbons
anthracene	c1ccc2cc3ccccc3cc2c1	Aromatic hydrocarbons
phenanthrene	c1ccc2c(c1)ccc1ccccc12	Aromatic hydrocarbons
1,2,3,4-tetrahydronaphthalene	c1ccc2c(c1)CCCC2	Aromatic hydrocarbons
indane¡hydrindene¿	c1ccc2c(c1)CCC2	Aromatic hydrocarbons
bis(2-ethylhexyl)phthalate	CCCCC(CC)COC(=O)c1ccccc1C(=O)OCC(CC)CCCC	Aromatic hydrocarbons
biphenyl	c1ccc(-c2ccccc2)cc1	Aromatic hydrocarbons
hexamethylbenzene	Cc1c(C)c(C)c(C)c(C)c1C	Aromatic hydrocarbons
n-methylaniline	CNc1ccccc1	Aromatic hydrocarbons
quinoline	c1ccc2ncccc2c1	Aromatic hydrocarbons
1,3,5-trimethylbenzene	Cc1cc(C)cc(C)c1	Aromatic hydrocarbons
indene	C1=Cc2ccccc2C1	Aromatic hydrocarbons
pyridine	c1ccncc1	Aromatic hydrocarbons
tert-butylbenzene	CC(C)(C)c1ccccc1	Aromatic hydrocarbons
2-phenyl-2-propanol	CC(C)(O)c1ccccc1	Aromatic hydrocarbons
p-cumylphenol	CC(C)(c1ccccc1)c1ccc(O)cc1	Aromatic hydrocarbons
phthalicaciddimethylester	COC(=O)c1ccccc1C(=O)OC	Aromatic hydrocarbons
2-phenylethanol	OCCc1ccccc1	Aromatic hydrocarbons
n,n-dimethylaniline	CN(C)c1ccccc1	Aromatic hydrocarbons
1,3-benzenedicarboxylicaciddimethylester	COC(=O)c1cccc(C(=O)OC)c1	Aromatic hydrocarbons
dimethylterephthalate	COC(=O)c1ccc(C(=O)OC)cc1	Aromatic hydrocarbons
1,2,4,5-tetramethylbenzene	Cc1cc(C)c(C)cc1C	Aromatic hydrocarbons
benzene	c1ccccc1	Aromatic hydrocarbons
benzoicacidmethylester	COC(=O)c1ccccc1	Aromatic hydrocarbons
benzylalcohol	OCc1ccccc1	Aromatic hydrocarbons
butylchloride	CCCCCl	Chlorides
tert-butylchloride	CC(C)(C)Cl	Chlorides
1,4-dichlorobutane	ClCCCCCl	Chlorides
sec-butylchloride	CCC(C)Cl	Chlorides
1-chloropentane	CCCCCCl	Chlorides
2,3-dichlorobutane	CC(Cl)C(C)Cl	Chlorides
1-chloro-2-methylpropane	CC(C)CCl	Chlorides
sec-butylcyclohexane	CCC(C)C1CCCCC1	Cyclic compounds
n-propylcyclohexane	CCCC1CCCCC1	Cyclic compounds
furan	c1ccoc1	Cyclic compounds
methylcyclohexane	CC1CCCCC1	Cyclic compounds
.alpha.-pinene	CC1=CCC2CC1C2(C)C	Cyclic compounds
methylcyclopentane	CC1CCCC1	Cyclic compounds
dibenzo[b,d]furan	c1ccc2c(c1)oc1ccccc12	Cyclic compounds
2-methyl-1,3-dioxolane	CC1OCCO1	Cyclic compounds
1,2-epoxybutane	CCC1CO1	Cyclic compounds
.beta.-pinene	C=C1CCC2CC1C2(C)C	Cyclic compounds
ethylcyclohexane	CCC1CCCCC1	Cyclic compounds
.gamma.-terpinene	CC1=CCC(C(C)C)=CC1	Cyclic compounds
bicyclohexyl	C1CCC(C2CCCCC2)CC1	Cyclic compounds
.alpha.-terpinene	CC1=CC=C(C(C)C)CC1	Cyclic compounds
2,2-dimethyl-3-methylenebicyclo[2.2.1]heptane	C=C1C2CCC(C2)C1(C)C	Cyclic compounds
cyclopentane	C1CCCC1	Cyclic compounds

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
2,5-dihydrofuran	C1=CCOC1	Cyclic compounds
2-methylbenzofuran	Cc1cc2ccccc2o1	Cyclic compounds
cyclohexane	C1CCCCC1	Cyclic compounds
cycloheptane	C1CCCCCC1	Cyclic compounds
cyclooctane	C1CCCCCCC1	Cyclic compounds
isopropylcyclohexane	CC(C)C1CCCCC1	Cyclic compounds
cyclohexene	C1=CCCCC1	Cyclic compounds
terpinolene	CC1=CCC(=C(C)C)CC1	Cyclic compounds
cyclopentene	C1=CCCC1	Cyclic compounds
2,5-dimethylfuran	Cc1ccc(C)o1	Cyclic compounds
2,2-dimethyloxirane	CC1(C)CO1	Cyclic compounds
4-ethenylcyclohexene	C=CC1CC=CCC1	Cyclic compounds
1,1-dimethylcyclopentane	CC1(C)CCCC1	Cyclic compounds
tetrahydrofuran	C1CCOC1	Cyclic compounds
tetrahydropyran	C1CCOCC1	Cyclic compounds
n-butylcyclohexane	CCCCC1CCCCC1	Cyclic compounds
1,4-dioxane	C1COCCO1	Cyclic compounds
ethylacetoacetate	CCOC(=O)CC(C)=O	Esters
methyllactate	COC(=O)C(C)O	Esters
methylacetoacetate	COC(=O)CC(C)=O	Esters
isopropenylacetate	C=C(C)OC(C)=O	Esters
salicylicacidmethylester	COC(=O)c1ccccc1O	Esters
ethyleneglycolmonoethyletheracetate	CCOCCOC(C)=O	Esters
diethylsuccinate	CCOC(=O)CCC(=O)OCC	Esters
4-oxo-pentanoicacidethylester	CCOC(=O)CCC(C)=O	Esters
alpha-hydroxy-isobutyricacidmethylester	COC(=O)C(C)(C)O	Esters
methacrylicacid,butylester	C=C(C)C(=O)OCCCC	Esters
diethyleneglycolmonoethyletheracetate	CCOCCOCCOC(C)=O	Esters
2-methylpropanoicacidmethylester	COC(=O)C(C)C	Esters
ethylacetate	CCOC(C)=O	Esters
ethylmethacrylate	C=C(C)C(=O)OCC	Esters
dimethylglutarate	COC(=O)CCCC(=O)OC	Esters
dimethylsuccinate	COC(=O)CCC(=O)OC	Esters
diethylmalonate	CCOC(=O)CC(=O)OCC	Esters
methacrylicacidmethylester	C=C(C)C(=O)OC	Esters
1-methoxyisopropylacetate	COCC(C)OC(C)=O	Esters
ethyllactate	CCOC(=O)C(C)O	Esters
cyclohexylacetate	CC(=O)OC1CCCCC1	Esters
isobutylacetate	CC(=O)OCC(C)C	Esters
1,2,3-propanetriol-triacetate	CC(=O)OCC(COC(C)=O)OC(C)=O	Esters
isoamylacetate	CC(=O)OCCC(C)C	Esters
ethyleneglycoldiacetate	CC(=O)OCCOC(C)=O	Esters
octadecanoicacidmethylester	CCCCCCCCCCCCCCCCCC(=O)OC	Esters
aceticacidphenylester	CC(=O)Oc1ccccc1	Esters
hexadecanoicacidethylester	CCCCCCCCCCCCCCCC(=O)OCC	Esters
hexadecanoicacidmethylester	CCCCCCCCCCCCCCCC(=O)OC	Esters
tetradecanoicacidethylester	CCCCCCCCCCCCCC(=O)OCC	Esters
tetradecanoicacidmethylester	CCCCCCCCCCCCCC(=O)OC	Esters

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
dodecanoicacidethylester	CCCCCCCCCCCC(=O)OCC	Esters
dodecanoicacidmethylester	CCCCCCCCCCCC(=O)OC	Esters
propanoicacidethylester	CCOC(=O)CC	Esters
decanoicacidethylester	CCCCCCCCCC(=O)OCC	Esters
isobutylisobutyrate	CC(C)COC(=O)C(C)C	Esters
isobutylformate	CC(C)COC=O	Esters
octanoicacidethylester	CCCCCCCC(=O)OCC	Esters
octanoicacidmethylester	CCCCCCCC(=O)OC	Esters
2-ethylhexylacetate	CCCCC(CC)COC(C)=O	Esters
tetrahydro-5-methyl-2-furanone	CC1CCC(=O)O1	Esters
methylpropanoate	CCC(=O)OC	Esters
propionicanhydride	CCC(=O)OC(=O)CC	Esters
isoamylbutyrate	CCCC(=O)OCCC(C)C	Esters
ethylbutyrate	CCCC(=O)OCC	Esters
butyricanhydride	CCCC(=O)OC(=O)CCC	Esters
methylbutanoate	CCCC(=O)OC	Esters
sec-butylacetate	CCC(C)OC(C)=O	Esters
decanoicacidmethylester	CCCCCCCCCC(=O)OC	Esters
aceticanhydride	CC(=O)OC(C)=O	Esters
aceticacidisopropylester	CC(=O)OC(C)C	Esters
cyclohexylformate	O=COC1CCCCC1	Esters
ethylisobutyrate	CCOC(=O)C(C)C	Esters
diethyloxalate	CCOC(=O)C(=O)OCC	Esters
2-propenoicacidmethylester	C=CC(=O)OC	Esters
2-propenoicacidethylester	C=CC(=O)OCC	Esters
propylacrylate	C=CC(=O)OCCC	Esters
formicacidpropylester	CCCOC=O	Esters
propylacetate	CCCOC(C)=O	Esters
butanoicacidpropylester	CCCOC(=O)CCC	Esters
propanoicacidpropylester	CCCOC(=O)CC	Esters
propylisobutyrate	CCCOC(=O)C(C)C	Esters
2-propenoicacidbutylester	C=CC(=O)OCCCC	Esters
2-butoxyethylacetate	CCCCOCCOC(C)=O	Esters
tert-butylacetate	CC(=O)OC(C)(C)C	Esters
butylacetate	CCCCOC(C)=O	Esters
formicacidbutylester	CCCCOC=O	Esters
butanoicacidbutylester	CCCCOC(=O)CCC	Esters
maleicanhydride	O=C1C=CC(=O)O1	Esters
octylacetate	CCCCCCCCOC(C)=O	Esters
heptylacetate	CCCCCCCOC(C)=O	Esters
vinylacetate	C=COC(C)=O	Esters
vinylpropionate	C=COC(=O)CC	Esters
decanedioicaciddibutylester	CCCCOC(=O)CCCCCCCCCC(=O)OCCCC	Esters
hexylacetate	CCCCCCOC(C)=O	Esters
amylacetate	CCCCCOC(C)=O	Esters
amylformate	CCCCCOC=O	Esters
aceticacid2-propenylester	C=CCOC(C)=O	Esters
3-buten-2-olacetate	C=CC(C)OC(C)=O	Esters
propanoicacidbutylester	CCCCOC(=O)CC	Esters
amylbutyrate	CCCCCOC(=O)CCC	Esters
1,2-dimethoxyethane	COCCOC	Ethers

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
diphenylether	c1ccc(Oc2ccccc2)cc1	Ethers
methoxybenzene	COc1ccccc1	Ethers
triethyleneglycoldimethylether	COCCOCCOCCOC	Ethers
diethyleneglycoldimethylether	COCCOCCOC	Ethers
di-sec-butylether	CCC(C)OC(C)CC	Ethers
diethyleneglycoldiethylether	CCOCCOCCOCC	Ethers
di-n-propylether	CCCOCCC	Ethers
ethoxybenzene	CCOc1ccccc1	Ethers
methylbutylether	CCCCOC	Ethers
diethoxymethane	CCOCOCC	Ethers
ethylbutylether	CCCCOCC	Ethers
ethyltert-butylether(etbe)	CCOC(C)(C)C	Ethers
dibutylether	CCCCOCCCC	Ethers
1,2-diethoxyethane	CCOCCOCC	Ethers
diethyleneglycoldibutylether	CCCCOCCOCCOCCCC	Ethers
ethylpropylether	CCCOCC	Ethers
diethylether	CCOCC	Ethers
dipentylether	CCCCCOCCCCC	Ethers
methylpentylether	CCCCCOC	Ethers
di-tert-butylether	CC(C)(C)OC(C)(C)C	Ethers
diisobutylether	CC(C)COCC(C)C	Ethers
diisopropylether	CC(C)OC(C)C	Ethers
triethylorthoformate¡triethoxymethane¿	CCOC(OCC)OCC	Ethers
1,1-diethoxyethane	CCOC(C)OCC	Ethers
methyltert-butylether(mtbe)	COC(C)(C)C	Ethers
2,5-dimethyltetrahydrofuran	CC1CCC(C)O1	Ethers
ethyltert-amylether	CCOC(C)(C)CC	Ethers
2,4,6-trimethyl-1,3,5-trioxane	CC1OC(C)OC(C)O1	Ethers
1,1-dimethoxyethane	COC(C)OC	Ethers
methyltert-amylether(tame)	CCC(C)(C)OC	Ethers
butylvinylether	C=COCCCC	Ethers
perfluorononane	FC(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	Fluorides
perfluoro-n-octane	FC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	Fluorides
1,3-difluorobenzene	Fc1cccc(F)c1	Fluorides
perfluorohexane	FC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	Fluorides
perfluoropentane	FC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	Fluorides
perfluorobutane	FC(F)(F)C(F)(F)C(F)(F)C(F)(F)F	Fluorides
perfluorocyclobutane[rc318]	FC1(F)C(F)(F)C(F)(F)C1(F)F	Fluorides
1,4-difluorobenzene	Fc1ccc(F)cc1	Fluorides
fluorobenzene	Fc1ccccc1	Fluorides
1,2-difluorobenzene	Fc1ccccc1F	Fluorides
1,1,1,3,3-pentafluorobutane	CC(F)(F)CC(F)(F)F	Fluorides
perfluoro-n-heptane	FC(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	Fluorides
hexafluorobenzene	Fc1c(F)c(F)c(F)c(F)c1F	Fluorides
iodobenzene	Ic1ccccc1	Halogenated compounds

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
1-iodobutane	CCCCI	Halogenated compounds
bromobenzene	Brc1ccccc1	Halogenated compounds
butylbromide	CCCCBr	Halogenated compounds
p-dibromobenzene	Brc1ccc(Br)cc1	Halogenated compounds
benzophenone	O=C(c1ccccc1)c1ccccc1	Ketones
2-cyclohexen-1-one	O=C1C=CCCC1	Ketones
methylisopropylketone	CC(=O)C(C)C	Ketones
3,3-dimethyl-2-butanone	CC(=O)C(C)(C)C	Ketones
4-methyl-3-penten-2-one	CC(=O)C=C(C)C	Ketones
4-methyl-2-pentanone	CC(=O)CC(C)C	Ketones
acetophenone	CC(=O)c1ccccc1	Ketones
2,4-dimethyl-3-pentanone	CC(C)C(=O)C(C)C	Ketones
cyclohexanone	O=C1CCCCC1	Ketones
2-nonanone	CCCCCCCC(C)=O	Ketones
2-octanone	CCCCCCC(C)=O	Ketones
2-heptanone	CCCCCC(C)=O	Ketones
3,5,5-trimethyl-2-cyclohexen-1-one	CC1=CC(=O)CC(C)(C)C1	Ketones
2-hexanone	CCCCC(C)=O	Ketones
diisobutylketone	CC(C)CC(=O)CC(C)C	Ketones
5-nonanone	CCCCC(=O)CCCC	Ketones
2-butanone	CCC(C)=O	Ketones
4-heptanone	CCCC(=O)CCC	Ketones
3-heptanone	CCCCC(=O)CC	Ketones
2-pentanone	CCCC(C)=O	Ketones
methylvinylketone	C=CC(C)=O	Ketones
3-hexanone	CCCC(=O)CC	Ketones
3-methoxypropionitrile	COCCC#N	Nitriles
allylcyanide	C=CCC#N	Nitriles
benzonitrile	N#Cc1ccccc1	Nitriles
benzylcyanide	N#CCc1ccccc1	Nitriles
glutaronitrile	N#CCCCC#N	Nitriles
hexanenitrile	CCCCCC#N	Nitriles
isobutyronitrile	CC(C)C#N	Nitriles
diphenylcarbonate	O=C(Oc1ccccc1)Oc1ccccc1	Other compounds
nitrobenzene	O=[N+]([O-])c1ccccc1	Other compounds
m-nitrotoluene	Cc1cccc([N+](=O)[O-])c1	Other compounds
di-tert.butylperoxide	CC(C)(C)OOC(C)(C)C	Other compounds
hexamethylphosphoricacidtriamide	CN(C)P(=O)(N(C)C)N(C)C	Other compounds
tert.butylhydroperoxide	CC(C)(C)OO	Other compounds
triethylphosphate	CCOP(=O)(OCC)OCC	Other compounds
2,4-toluenediisocyanate	Cc1ccc(N=C=O)cc1N=C=O	Other compounds
carbonicaciddiethylester	CCOC(=O)OCC	Other compounds
isocyanicacidbutylester	CCCCN=C=O	Other compounds
triethoxyphosphine	CCOP(OCC)OCC	Other compounds
hexamethylenediisocyanate	O=C=NCCCCCCN=C=O	Other compounds
o-nitrotoluene	Cc1ccccc1[N+](=O)[O-]	Other compounds
diethyleneglycolmonomethylether	COCCOCCO	Polyfunctional compounds
2-methoxyphenol	COc1ccccc1O	Polyfunctional compounds
triphenylphosphine	c1ccc(P(c2ccccc2)c2ccccc2)cc1	Polyfunctional compounds
3,4-dichlorophenylisocyanate	O=C=Nc1ccc(Cl)c(Cl)c1	Polyfunctional compounds

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
1-tert-butoxy-2-propanol	CC(O)COC(C)(C)C	Polyfunctional compounds
2-methoxy-1-propanol	COC(C)CO	Polyfunctional compounds
5-hydroxymethylfurfural	O=Cc1ccc(CO)o1	Polyfunctional compounds
salicylaldehyde	O=Cc1ccccc1O	Polyfunctional compounds
furfural	O=Cc1ccco1	Polyfunctional compounds
sulfolane	O=S1(=O)CCCC1	Polyfunctional compounds
3,4-dichloronitrobenzene	O=[N+]([O-])c1ccc(Cl)c(Cl)c1	Polyfunctional compounds
p-chlorophenol	Oc1ccc(Cl)cc1	Polyfunctional compounds
tetrahydrofurfurylalcohol	OCC1CCCO1	Polyfunctional compounds
diacetonealcohol	CC(=O)CC(C)(C)O	Polyfunctional compounds
p-methoxyphenol	COc1ccc(O)cc1	Polyfunctional compounds
levulinicacid	CC(=O)CCC(=O)O	Polyfunctional compounds
methylethylketoxime	CCC(C)=NO	Polyfunctional compounds
8-hydroxyquinoline	Oc1cccc2cccnc12	Polyfunctional compounds
o-chlorophenol	Oc1ccccc1Cl	Polyfunctional compounds
furfurylalcohol	OCc1ccco1	Polyfunctional compounds
cyanoaceticacidethylester	CCOC(=O)CC#N	Polyfunctional compounds
1-methoxy-2-propanol	COCC(C)O	Polyfunctional compounds
p-phenetidine	CCOc1ccc(N)cc1	Polyfunctional compounds
1-propoxy-2-propanol	CCCOCC(C)O	Polyfunctional compounds
ethyleneglycolmonopropyleth er	CCCOCCO	Polyfunctional compounds
bis(2-chloroethyl)ether	ClCCOCCCl	Polyfunctional compounds
1,2-propyleneglycol1-monobu tylether	CCCCOCC(C)O	Polyfunctional compounds
diethyleneglycolmonopropyle ther	CCCOCCOCCO	Polyfunctional compounds
n-methyl-2-pyrrolidone	CN1CCCC1=O	Polyfunctional compounds
1,3-dimethylimidazolidin-2-o ne	CN1CCN(C)C1=O	Polyfunctional compounds
chloroaceticacidethylester	CCOC(=O)CCl	Polyfunctional compounds
morpholine	C1COCCN1	Polyfunctional compounds
dipropyleneglycolbutylether	CCCCOCC(C)OCC(C)O	Polyfunctional compounds
2-butoxyethanol	CCCCOCCO	Polyfunctional compounds
n-ethyl-2-pyrrolidone	CCN1CCCC1=O	Polyfunctional compounds
diethyleneglycolmonobutylet her	CCCCOCCOCCO	Polyfunctional compounds
4-ethylmorpholine	CCN1CCOCC1	Polyfunctional compounds
eugenol	C=CCc1ccc(O)c(OC)c1	Polyfunctional compounds
p-chloroaniline	Nc1ccc(Cl)cc1	Polyfunctional compounds
diethyleneglycoln-hexylether	CCCCCCOCCOCCO	Polyfunctional compounds
benzoylchloride	O=C(Cl)c1ccccc1	Polyfunctional compounds
2-ethoxyethanol	CCOCCO	Polyfunctional compounds
diethyleneglycolethylether	CCOCCOCCO	Polyfunctional compounds
1-ethoxy-2-propanol	CCOCC(C)O	Polyfunctional compounds
trichlorophenylsilane	Cl[Si](Cl)(Cl)c1ccccc1	Silanes/siloxanes
tetramethylsilane	C[Si](C)(C)C	Silanes/siloxanes
hexamethyldisilazane	C[Si](C)(C)N[Si](C)(C)C	Silanes/siloxanes
tetramethoxysilane	CO[Si](OC)(OC)OC	Silanes/siloxanes
tetraethylsilane	CC[Si](CC)(CC)CC	Silanes/siloxanes
hexamethyldisiloxane	C[Si](C)(C)O[Si](C)(C)C	Silanes/siloxanes
dichlorodiethylsilane	CC[Si](Cl)(Cl)CC	Silanes/siloxanes
octamethyltrisiloxane	C[Si](C)(C)O[Si](C)(C)O[Si](C)(C)C	Silanes/siloxanes

Table 3: List of names, their SMILES notations, and families of the validation set that could be grouped into families

Name	SMILES	Family
octamethylcyclotetrasiloxane	C[Si]1(C)O[Si](C)(C)O[Si](C)(C)O[Si](C)(C)O1	Silanes/siloxanes
dichloromethylphenylsilane	C[Si](Cl)(Cl)c1ccccc1	Silanes/siloxanes
dodecamethylpentasiloxane	C[Si](C)(C)O[Si](C)(C)O[Si](C)(C)O[Si](C)(C)O[Si](C)(C)C	Silanes/siloxanes
decamethyltetrasiloxane	C[Si](C)(C)O[Si](C)(C)O[Si](C)(C)O[Si](C)(C)C	Silanes/siloxanes
tetraethoxysilane	CCO[Si](OCC)(OCC)OCC	Silanes/siloxanes
benzo(b)thiophene	c1ccc2sccc2c1	Sulfides/thiophenes
dipropylsulfide	CCCSCCC	Sulfides/thiophenes
2-methylthiophene	Cc1cccs1	Sulfides/thiophenes
thiophene	c1ccsc1	Sulfides/thiophenes
2,5-dimethylthiophene	Cc1ccc(C)s1	Sulfides/thiophenes
diethylsulfide	CCSCC	Sulfides/thiophenes
diethyldisulfide	CCSSCC	Sulfides/thiophenes
2-ethylthiophene	CCc1cccs1	Sulfides/thiophenes
tetrahydrothiophene	C1CCSC1	Sulfides/thiophenes
allylsulfide	C=CCSCC=C	Sulfides/thiophenes
3-methylthiophene	Cc1ccsc1	Sulfides/thiophenes
n-butylmercaptan	CCCCS	Thiols
benzenethiol	Sc1ccccc1	Thiols
2-methyl-2-propanethiol	CC(C)(C)S	Thiols
1-pentanethiol	CCCCCS	Thiols
iso-butylmercaptan	CC(C)CS	Thiols