# CHEMICAL-REACTION-AWARE MOLECULE REPRESENTATION LEARNING

**Hongwei Wang**[1], **Weijiang Li**[1], **Xiaomeng Jin**[1], **Kyunghyun Cho**[2,3], **Heng Ji**[1], **Jiawei Han**[1], **Martin D. Burke**[1]
[1]University of Illinois Urbana-Champaign, [2]New York University, [3]Genentech
{hongweiw, wl13, xjin17, hengji, hanj, mdburke}@illinois.edu, kyunghyun.cho@nyu.edu

# 1. Introduction and current work

- SMILES (Simplified Molecular-Input Line-Entry System) based LLMs.
- Delicately hand-crafted GNNs.

- Paper proposes using chemical reactions to learn *effective and generalizable* embeddings.
  Eg. $CH_3COOH + C_2H_5OH \rightarrow CH_3COOC_2H_5 + H_2O$

  $\Rightarrow h_{CH3COOH} + h_{C2\ H5\ OH} = h_{CH3\ COOC2\ H5} + h_{H2\ O}$

# 2. Brief results

- *Well-organized embeddings.*
- Can learn *reaction templates* (like esterification) using sumas READOUT.
- Beats many baselines. (Hit@1, AUC, RMSE across datasets and tasks)

# 3. ***Method***

- Initial Encoding - *"Each atom $a_i$ has an initial feature vector $x_i$ encoding its properties. In this work, we use four types of atom properties:* ***element type, charge, whether the atom is an aromatic ring, and the count of attached hydrogen atom(s).*** *Each type of atom properties is represented as a one-hot vector, and we add an additional "**unknown**" entry for each one-hot vector to handle unknown values during inference. The four one-hot vectors are concatenated as the initial atom feature."*

- Bond-types are ignored! Graph becomes homogeneous.

- Preserving equivalence – (results in *equivalence classes*)

**Proposition 1** *Let $M$ be the set of molecules, $R \subseteq M$ and $P \subseteq M$ be the reactant set and product set of a chemical reaction, respectively. If $R \to P \Leftrightarrow \sum_{r \in R} h_r = \sum_{p \in P} h_p$ for all chemical reactions, then "$\to$" is an equivalence relation on $2^M$ that satisfies the following three properties: (1) Reflexivity: $A \to A$, for all $A \in 2^M$; (2) Symmetry: $A \to B \Leftrightarrow B \to A$, for all $A, B \in 2^M$; (3) Transitivity: If $A \to B$ and $B \to C$, then $A \to C$, for all $A, B, C \in 2^M$.*

# 3. Method

- The last condition is *very strong* leads to robust embeddings.
- *Eg.* in organic synthesis, a target compound ***t*** may be made from three different sets of starting materials ***A***, ***B***, and ***C***. Then the sets ***A***, ***B***, ***C*** as well as {***t***} belong to *one equivalence class*, and we have
  $\sum_{a\in A} h_a = \sum_{b\in B} h_b = \sum_{c\in C} h_c = h_t$

- It also ***improves generalizability.***
  Reaction Center – *'The reaction center of R → P is defined as a subgraph of R consisting of atoms whose bonds have changed after reaction'.*

**Proposition 2** *Let $R \to P$ be a chemical reaction where $R$ is the reactant set and $P$ is the product set, and $C$ be its reaction center. Suppose that we use the GNN (whose number of layers is $K$) shown in Eqs. (1) and (2) as the molecule encoder, and set the READOUT function in Eq. (2) as summation. Then for an arbitrary atom $a$ in one of the reactant whose final representation is $h_a^K$, the residual term $\sum_{r\in R} h_r - \sum_{p\in P} h_p$ is a function of $h_a^K$ if and only if the distance between atom $a$ and reaction center $C$ is less than $K$.*

$$R_1\text{-}CH_2COOH + R_2\text{-}CH_2CH_2OH \to R_1\text{-}CH_2COOCH_2CH_2\text{-}R_2 + H_2O$$

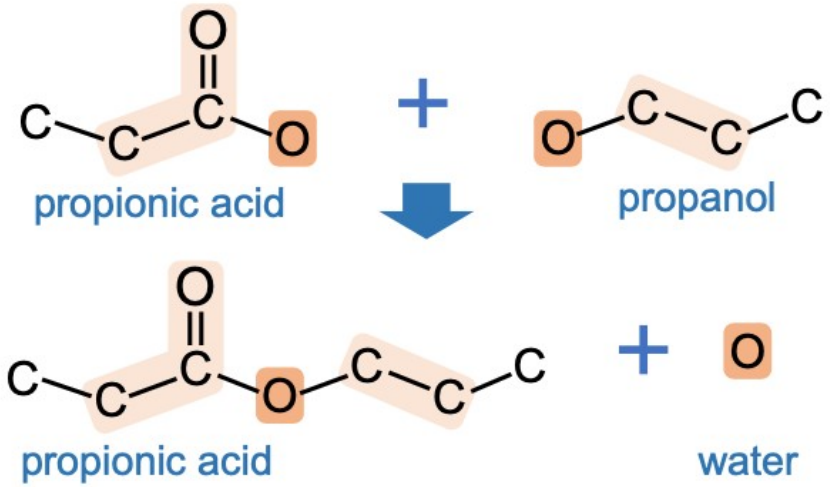
# 3. ***Method***

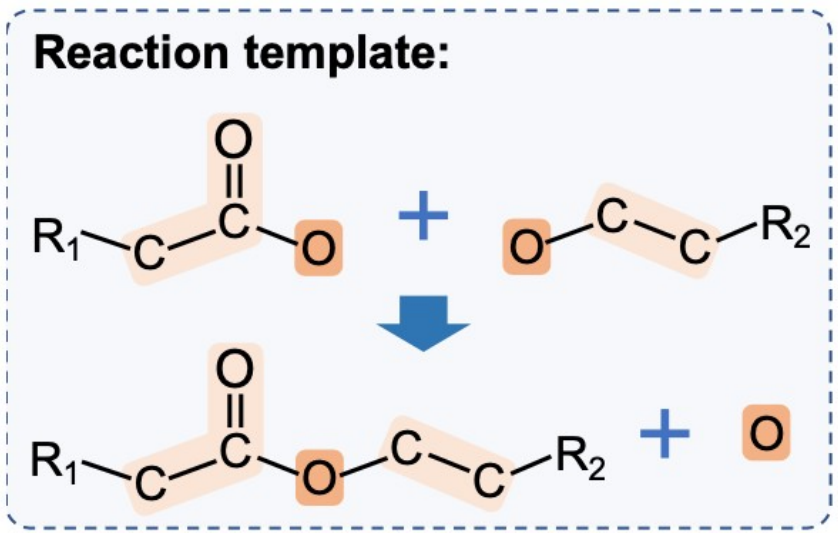
- Important implications –

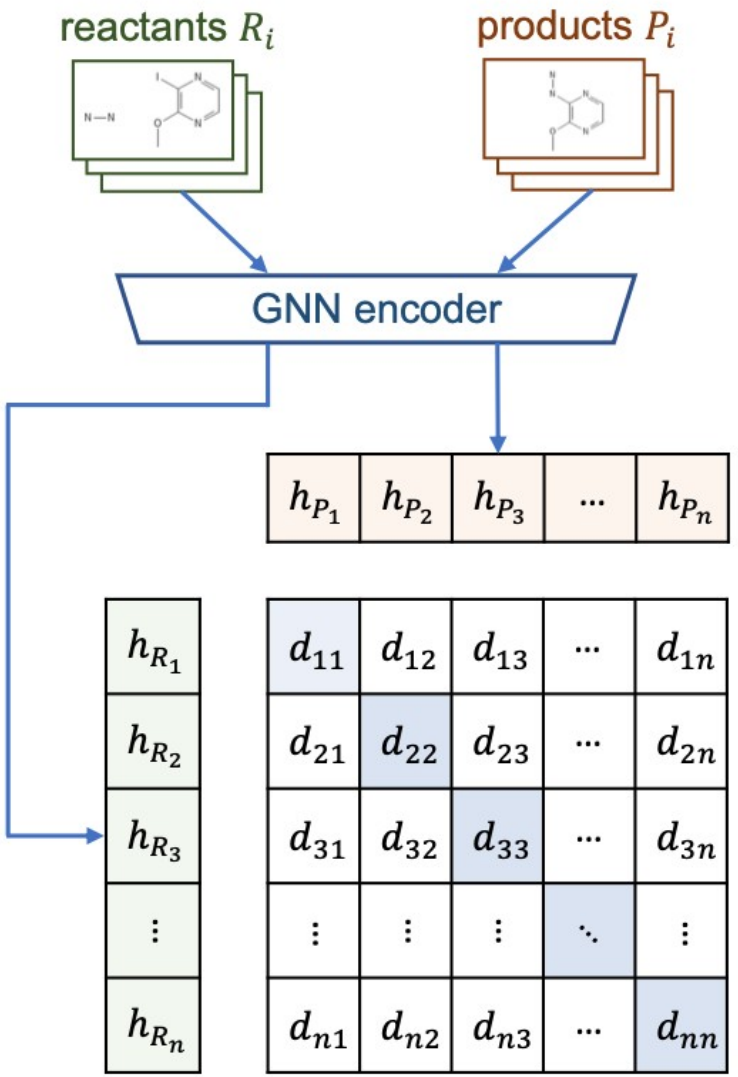
  1. No complicated networks required unlike previous work + *few-shot* adaptation.
  2. Can learn strong properties even though reactions can be unbalanced, eg. $H_2O$ is omitted from all Esterification reactions.
  3. Because of Proposition-2, number of GNN layers is a ***very important factor.***

- Training –

$$L_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_i \left\| \sum_{r\in R_i} h_r - \sum_{p\in P_i} h_p \right\|_2 + \frac{1}{|\mathcal{B}|(|\mathcal{B}|-1)}\sum_{i\neq j} \max\left(\gamma - \left\| \sum_{r\in R_i} h_r - \sum_{p\in P_j} h_p \right\|_2, 0\right),$$


propionic acid
propanol
propionic acid
water


Reaction template:


reactants $R_i$
products $P_i$
GNN encoder
$h_{P_1}$ $h_{P_2}$ $h_{P_3}$ ... $h_{P_n}$
$h_{R_1}$ $h_{R_2}$ $h_{R_3}$ ⋮ $h_{R_n}$
$d_{11}$ $d_{12}$ $d_{13}$ ... $d_{1n}$
$d_{21}$ $d_{22}$ $d_{23}$ ... $d_{2n}$
$d_{31}$ $d_{32}$ $d_{33}$ ... $d_{3n}$
⋮ ⋮ ⋮ ⋱ ⋮
$d_{n1}$ $d_{n2}$ $d_{n3}$ ... $d_{nn}$

# 4. Experiments

- *Reaction Product Prediction* - USPTO dataset

| Metrics | MRR | MR | Hit@1 | Hit@3 | Hit@5 | Hit@10 |
|---|---|---|---|---|---|---|
| Mol2vec | 0.681 | 483.7 | 0.614 | 0.725 | 0.759 | 0.798 |
| Mol2vec-FT1 | 0.688 ± 0.000 | 417.6 ± 0.1 | 0.620 ± 0.000 | 0.734 ± 0.000 | 0.767 ± 0.000 | 0.806 ± 0.000 |
| MolBERT | 0.708 | 460.7 | 0.623 | 0.768 | 0.811 | 0.858 |
| MolBERT-FT1 | 0.731 ± 0.000 | 457.9 ± 0.0 | 0.649 ± 0.000 | 0.790 ± 0.000 | 0.831 ± 0.000 | 0.873 ± 0.000 |
| MolBERT-FT2 | 0.776 ± 0.000 | 459.6 ± 0.2 | 0.708 ± 0.000 | 0.827 ± 0.000 | 0.859 ± 0.000 | 0.891 ± 0.000 |
| MolR-GCN | 0.905 ± 0.001 | 34.5 ± 2.4 | 0.867 ± 0.001 | 0.938 ± 0.001 | 0.950 ± 0.001 | 0.961 ± 0.002 |
| MolR-GAT | 0.903 ± 0.002 | 35.3 ± 2.8 | 0.864 ± 0.002 | 0.935 ± 0.003 | 0.948 ± 0.003 | 0.961 ± 0.003 |
| MolR-SAGE | 0.903 ± 0.004 | 53.0 ± 4.6 | 0.865 ± 0.005 | 0.935 ± 0.004 | 0.948 ± 0.004 | 0.961 ± 0.002 |
| MolR-TAG | **0.918** ± 0.000 | **27.4** ± 0.4 | **0.882** ± 0.000 | **0.949** ± 0.001 | **0.960** ± 0.001 | **0.970** ± 0.000 |
| MolR-TAG (1% training data) | 0.904 ± 0.002 | 33.0 ± 3.7 | 0.865 ± 0.003 | 0.937 ± 0.003 | 0.951 ± 0.002 | 0.963 ± 0.002 |

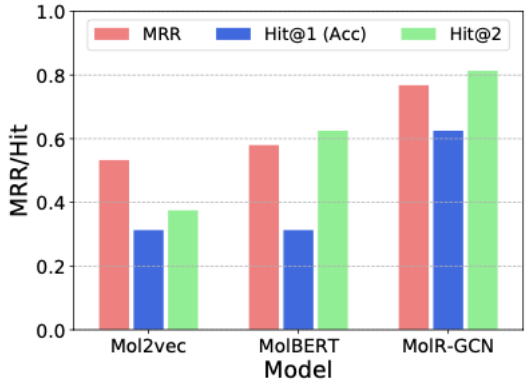
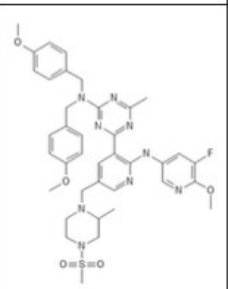
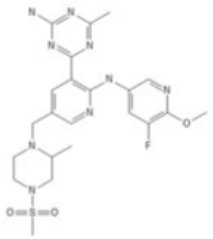
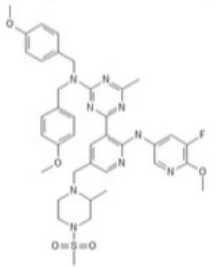
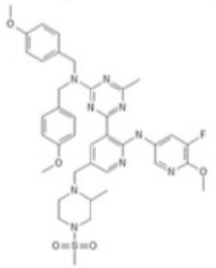
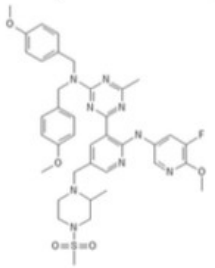
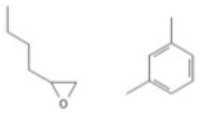
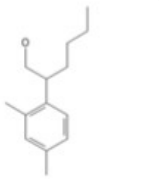
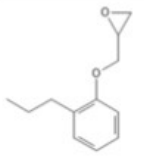
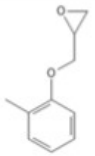
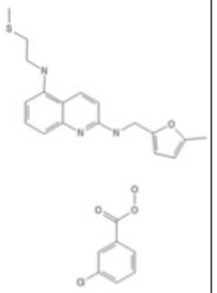
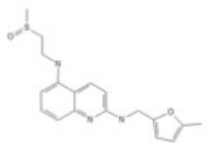
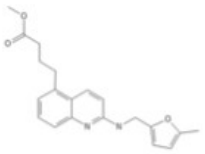
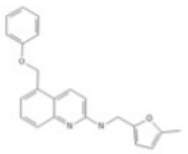
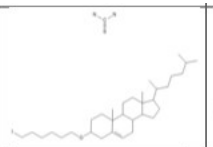
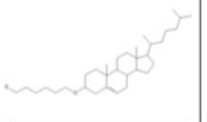
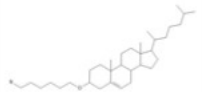
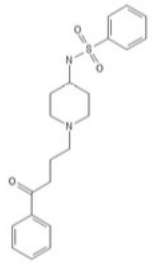
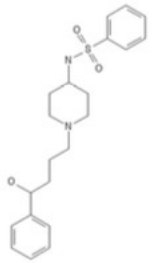
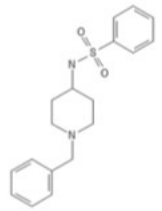
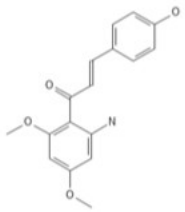
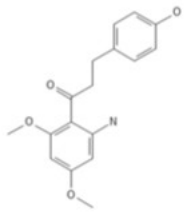
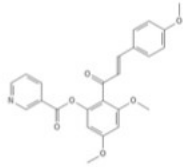
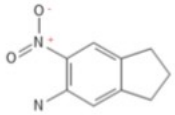
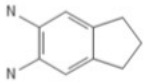
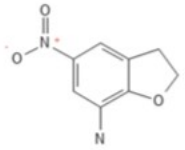
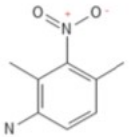



Figure 2: Result of answering real multi-choice questions on product prediction.

# 4. Experiments

- *Case study on first 20 reactions*

| No. | Reactant(s) | Ground-truth product | Predicted product by MolR-GCN | Predicted product by Mol2vec | Predicted product by MolBERT |
|---|---|---|---|---|---|
| 5 |  |  |  (No. 32353) |  (No. 32353) |  (No. 32353) |
| 6 |  |  | Same as ground-truth |  (No. 39181) |  (No. 24126) |
| 8 |  |  | Same as ground-truth |  (No. 11233) |  (No. 17526) |

| | | | | | |
|---|---|---|---|---|---|
| 10 |  |  |  (No. 18889) |  (No. 18889) |  (No. 18889) |
| 13 |  |  | Same as ground-truth | Same as ground-truth |  (No. 37024) |
| 16 |  |  | Same as ground-truth |  (No. 21947) | Same as ground-truth |
| 17 |  |  | Same as ground-truth |  (No. 2029) |  (No. 22247) |

# 4. Experiments

- *Molecule Property Prediction* - 5 datasets *BBBP*, *HIV*, *BACE*, *Tox21*, and *ClinTox*

| Datasets | BBBP | HIV | BACE | Tox21 | ClinTox |
|---|---|---|---|---|---|
| SMILES-Transformers | 0.704 | 0.729 | 0.701 | 0.802 | **0.954** |
| ECFP4 | 0.729 | 0.792 | 0.867 | 0.822 | 0.799 |
| GraphConv | 0.690 | 0.763 | 0.783 | 0.829 | 0.807 |
| Weave | 0.671 | 0.703 | 0.806 | 0.820 | 0.832 |
| ChemBERTa | 0.643 | 0.622 | - | 0.728 | 0.733 |
| D-MPNN | 0.708 | 0.752 | - | 0.688 | 0.906 |
| CDDD | 0.761 ± 0.00 | 0.753 ± 0.00 | 0.833 ± 0.00 | - | - |
| MolBERT | 0.762 ± 0.00 | 0.783 ± 0.00 | 0.866 ± 0.00 | - | - |
| GraphCL | 0.695 ± 0.005 | 0.776 ± 0.009 | 0.782 ± 0.012 | 0.754 ± 0.009 | 0.701 ± 0.019 |
| GraphLoG | 0.725 ± 0.008 | 0.778 ± 0.008 | 0.835 ± 0.012 | 0.757 ± 0.005 | 0.767 ± 0.033 |
| Mol2vec | 0.872 ± 0.021 | 0.769 ± 0.021 | 0.862 ± 0.027 | 0.803 ± 0.041 | 0.841 ± 0.062 |
| MolR-GCN | 0.890 ± 0.032 | **0.802** ± 0.024 | **0.882** ± 0.019 | 0.818 ± 0.023 | 0.916 ± 0.039 |
| MolR-GAT | 0.887 ± 0.026 | 0.794 ± 0.022 | 0.863 ± 0.026 | **0.839** ± 0.039 | 0.908 ± 0.039 |
| MolR-SAGE | 0.879 ± 0.032 | 0.793 ± 0.026 | 0.859 ± 0.029 | 0.811 ± 0.039 | 0.890 ± 0.058 |
| MolR-TAG | **0.895** ± 0.031 | 0.801 ± 0.023 | 0.875 ± 0.023 | 0.820 ± 0.028 | 0.913 ± 0.043 |

# 4. Experiments

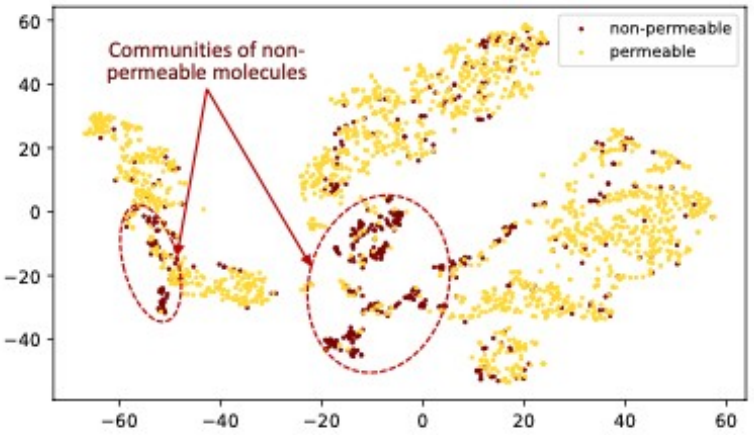
- *Chemical Reaction Classification* – USPTO-1k-TPL dataset

| Methods | Accuracy |
|---|---|
| RXNFP | 0.989 |
| AP3-256-5NN | 0.295 |
| AP3-256-MLP | 0.809 |
| DRFP-5NN | 0.917 |
| DRFP-MLP | 0.977 |
| MolR-GCN | $0.931 \pm 0.022$ |
| MolR-GAT | $0.930 \pm 0.017$ |
| MolR-SAGE | $0.936 \pm 0.025$ |
| MolR-TAG | $0.962 \pm 0.028$ |

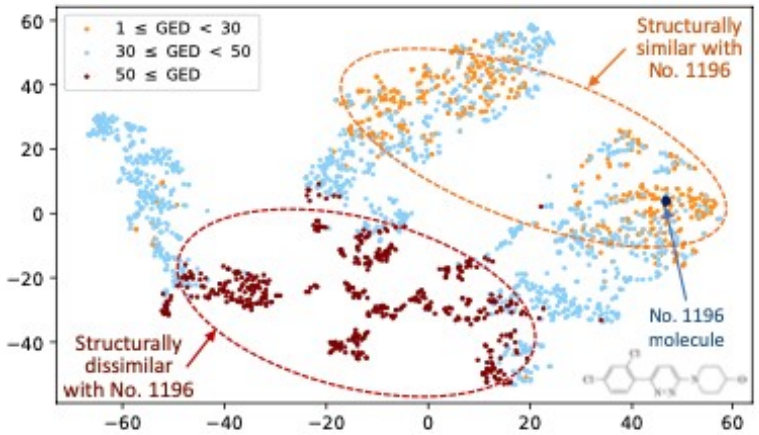
- *Graph Edit Distance Prediction* – subset of the QM9 dataset.

| Feature mode | Concat | Subtract |
|---|---|---|
| Mol2vec | $1.140 \pm 0.041$ | $0.995 \pm 0.034$ |
| MolBERT | $1.127 \pm 0.042$ | $0.937 \pm 0.029$ |
| MolR-GCN | $0.976 \pm 0.026$ | $0.922 \pm 0.019$ |
| MolR-GAT | $1.007 \pm 0.021$ | $0.943 \pm 0.016$ |
| MolR-SAGE | $\mathbf{0.918} \pm 0.028$ | $\mathbf{0.817} \pm 0.013$ |
| MolR-TAG | $0.960 \pm 0.027$ | $0.911 \pm 0.027$ |

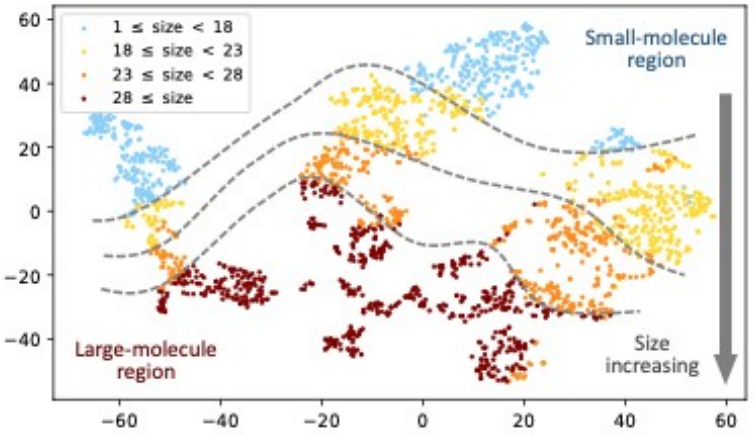
# *5. Visualization of embeddings* (using t-SNE)



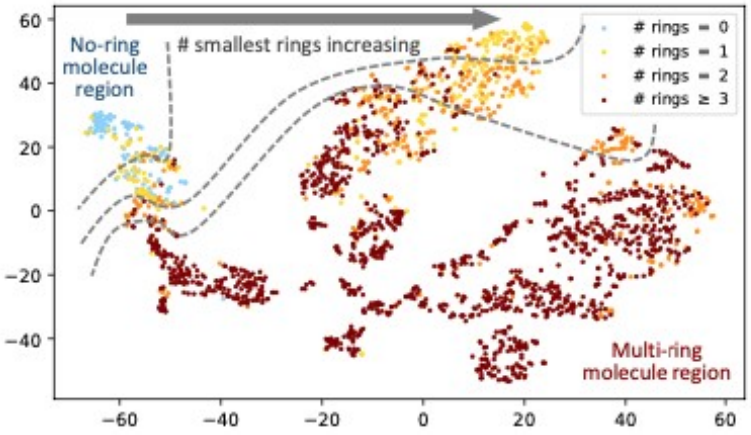

(a) Molecule property




(b) GED w.r.t. No. 1196 molecule




(c) Molecule size




(d) # smallest rings