# A Probability-Based Contrastive Learning Framework for 3D Molecular Representation Learning

Jiayu Qin,  Jian Chen, Rohan Sharma, JIngchen Sun, Changyou Chen

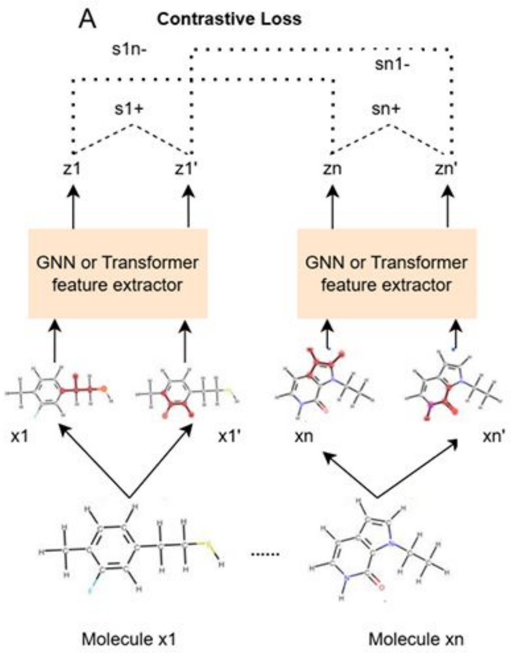
University at Buffalo The State University of New York

# Abstract

- The role of contrastive learning (CL) in molecular representation learning
  - Contrastive Learning (CL) enables unsupervised learning from large-scale, unlabeled molecular datasets.
- The problem of false positive and false negative pairs in molecular datasets
  - Existing methods often introduce false positive and false negative pairs due to conventional augmentations, limiting their effectiveness.
- Our proposed framework and its achievements
  - We propose a probability-based contrastive learning framework, optimized through a stochastic expectation-maximization process, achieving state-of-the-art results in multiple benchmarks.

# Contrastive molecular learning

- Molecular contrastive learning Molecules are represented as 2D or 3D molecule graphs.
- Two stochastic augmentation strategies are applied to each graph, resulting in two aug mentations.
- A feature extractor is used to extract features and contrastive loss is used to maximize the similarity of positive pairs and minimize the similarity of negative pairs



# Motivation

- Contrastive Learning is essential for unsupervised learning from large-scale unlabeled molecular datasets.
- Existing methods often generate false positive and false negative pairs due to conventional graph augmentations, such as node masking and subgraph removal. These issues can reduce the effectiveness of CL on molecular datasets.
- Our approach introduces a probability-based method that assigns dynamic weights to pairs to reduce this issue.

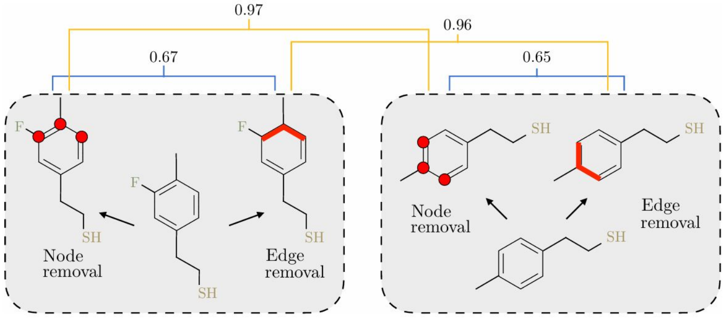


Figure 1: **Existing problem in molecular contrastive learning.** Adopt node removal and edge removal for molecular contrastive learning can lead to false positive and false negative problems. Blue lines indicate positive pairs and yellowing lines indicate negative pairs. The numbers on each line indicate the chemical similarity between the augmented pair of molecules. In this case, positive pairs indeed have lower similarity than negative pairs.

# Probability contrastive framework

- Our framework uses a Bayesian inference model to dynamically adjust weights for molecular pairs.
  - Original contrastive loss

$$\mathcal{L} = \frac{1}{N}\sum_{k=1}^{N}[\ell(2k-1,2k)+\ell(2k,2k-1)],\ \text{with}\ \ell(i,j) = -\log\frac{s_{i+}}{s_{i+}+\sum_{k=1}^{2N}\mathbb{I}_{[k\neq i,j]}s_{i,k-}}$$

  - Ours weighted loss

$$\mathcal{L}_w = \frac{1}{N}\sum_{k=1}^{N}[\bar{\ell}(2k-1,2k)+\bar{\ell}(2k,2k-1)],\ \ \bar{\ell}(i,j) = -\log\frac{w_i^+ s_{i+}}{w_i^+ s_{i+}+\sum_{k=1}^{2N}\mathbb{I}_{[k\neq i,j]}w_{ik}^- s_{ik-}}$$

- We incorporate Gamma and Bernoulli distributions to represent pair weights, reducing mislabeling effects.
  -

Option 1 - Gamma priors for continuous weighting:

$$w_i^+ \sim \mathrm{Gamma}(a_+, b_+), w_{ik}^- \sim \mathrm{Gamma}(a_-, b_-).$$

Option 2 - Bernoulli priors for selective weighting:

$$w_i^+ \sim \mathrm{Gamma}(a_+, b_+), \quad w_{ik}^- \sim \mathrm{Bernoulli}(a_-^-).$$

- With this formulation, we can define the following distribution:
  -

$$p\left(\{w_i^+\},\{w_{ik}^-\},\boldsymbol{\theta};\mathcal{D}\right) \propto \prod_{\mathbf{x}_i\in\mathcal{D}}\frac{w_i^+ s_{i+}}{w_i^+ s_{ij+}+\sum_{k=1}^{K}w_{ik}^- s_{ik-}}p(\{w_i^+\})p(\{w_{ik}^-\})p(\boldsymbol{\theta}).$$

# Method continued

- $p\left(\{w_i^+\},\{w_{ik}^-\},\boldsymbol{\theta};\mathcal{D}\right) \propto \prod_{\mathbf{x}_i\in\mathcal{D}} \frac{w_i^+ s_{i+}}{w_i^+ s_{ij+} + \sum_{k=1}^{K} w_{ik}^- s_{ik-}} p(\{w_i^+\})p(\{w_{ik}^-\})p(\boldsymbol{\theta}).$
- With this distribution, posterior inference of the weights is challenging, due to the lack of convenience posterior distributions
- We can introduce an augmented variable u to associate with data points x, then we can define an augmented distribution:
    - $p(\boldsymbol{\theta},\mathbf{u},\mathbf{w}\mid\mathcal{D}) \propto \prod_{i:\mathbf{x}_i\in\mathcal{D}} w_i^+ s_i + e^{-\mathbf{u}_i w_i^+ s_i +} \prod_k e^{-u_i w_{ik}^- s_{ik}-} p\left(\{w_i^+\}\right) p\left(\{w_{ik}^-\}\right) p(\boldsymbol{\theta}),$
- Then we can do inference based on this distribution

# Efficient Inference and Learning with Stocastic EM

We alternatively infer the local random variables w and optimize the global model parameter θ

The basic idea is to alternatively

1) optimizing model parameter θ with fixed (u,w) and

2) sampling (u,w) with f ixed θ.

We follow standard procedures in stochastic EM to divide the learning into three steps: Simulation, Stochastic Expectation, and Maximization.

Simulation: based the posterior distribution and the current batch of data, we infer the u and w:

$$u_i \mid \{w_i^+, w_{ik}^-, \boldsymbol{\theta}\} \sim \text{Gamma}\left(a_u, b_u + w_i^+ s_{i+} + \sum w_{ik}^- s_{ik-}\right), \forall i, \text{ and}$$

$$w_i^+ \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Gamma}\left(1 + a_+, u_i s_{i+} + b_+\right), \text{and}$$

$$\text{Option 1: } w_{ik}^- \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Gamma}\left(a_-, u_i s_{ik^-} + b_-\right), \forall i, k$$

$$\text{Option 2: } w_{ik}^- \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Bernoulli}\left(\frac{a_- e^{-u_i s_{ik-}}}{1 - a_- + a_- e^{-u_i s_{ik-}}}\right)$$

# Stochastic Expectation and Maximization

We use the sampled auxiliary random variables to update the model parameter θ by maximizing a stochastic objective Q(θ), defined as:

$$Q_{t+1}(\boldsymbol{\theta}) = Q_t(\boldsymbol{\theta}) + \lambda_t \left(\log p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \mid \mathcal{D}) - Q_t(\boldsymbol{\theta})\right)$$

Here, t is iteration step, and $\{\lambda_t\}$ is a sequence of decreasing weights

by decomposing the recursion, we have:

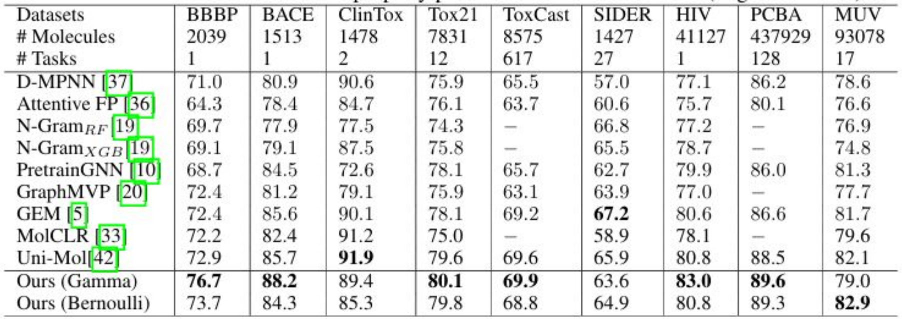
$$Q_{t+1}(\boldsymbol{\theta}) = \sum_{\tau=0}^{t} \tilde{\lambda}_\tau \log p\left(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \mid \mathcal{D}_\tau\right), \text{ where } \tilde{\lambda}_\tau \triangleq \lambda_\tau \prod_{t'=\tau+1}^{t} (1 - \lambda_{t'})$$

At each time t, we can initialize the parameter θ from the last step, and update it by stochastic gradient ascent on the log-likelihood, log p(θ,uт,wт | Dт) calculated from the current batch of data.

To reduce variance, we propose to optimize a marginal version by integrating out uт from p(θ,uт,wт | Dт), which essentially reduces to our original weighted contrastive loss.

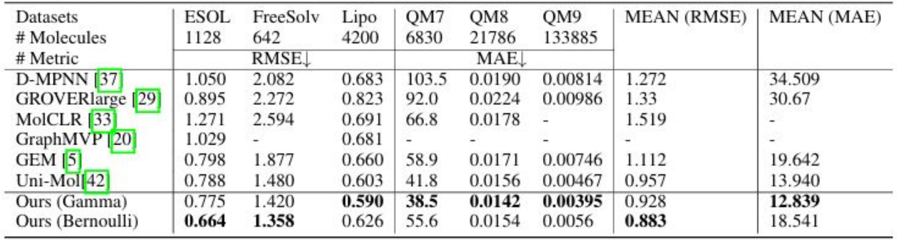
# Experimant results

Table 1: ROC_AUC on molecular property prediction classification tasks (Higher is better)



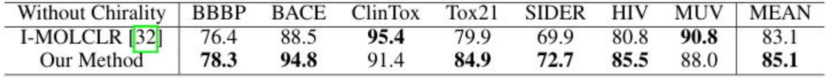
| Datasets<br># Molecules<br># Tasks | BBBP<br>2039<br>1 | BACE<br>1513<br>1 | ClinTox<br>1478<br>2 | Tox21<br>7831<br>12 | ToxCast<br>8575<br>617 | SIDER<br>1427<br>27 | HIV<br>41127<br>1 | PCBA<br>437929<br>128 | MUV<br>93078<br>17 |
|---|---|---|---|---|---|---|---|---|---|
| D-MPNN [37] | 71.0 | 80.9 | 90.6 | 75.9 | 65.5 | 57.0 | 77.1 | 86.2 | 78.6 |
| Attentive FP [36] | 64.3 | 78.4 | 84.7 | 76.1 | 63.7 | 60.6 | 75.7 | 80.1 | 76.6 |
| N-Gram$_{RF}$ [19] | 69.7 | 77.9 | 77.5 | 74.3 | – | 66.8 | 77.2 | – | 76.9 |
| N-Gram$_{XGB}$ [19] | 69.1 | 79.1 | 87.5 | 75.8 | – | 65.5 | 78.7 | – | 74.8 |
| PretrainGNN [10] | 68.7 | 84.5 | 72.6 | 78.1 | 65.7 | 62.7 | 79.9 | 86.0 | 81.3 |
| GraphMVP [20] | 72.4 | 81.2 | 79.1 | 75.9 | 63.1 | 63.9 | 77.0 | – | 77.7 |
| GEM [5] | 72.4 | 85.6 | 90.1 | 78.1 | 69.2 | **67.2** | 80.6 | 86.6 | 81.7 |
| MolCLR [33] | 72.2 | 82.4 | 91.2 | 75.0 | – | 58.9 | 78.1 | – | 79.6 |
| Uni-Mol[42] | 72.9 | 85.7 | **91.9** | 79.6 | 69.6 | 65.9 | 80.8 | 88.5 | 82.1 |
| Ours (Gamma) | **76.7** | **88.2** | 89.4 | **80.1** | **69.9** | 63.6 | **83.0** | **89.6** | 79.0 |
| Ours (Bernoulli) | 73.7 | 84.3 | 85.3 | 79.8 | 68.8 | 64.9 | 80.8 | 89.3 | **82.9** |

Table 2: Performance on molecular property prediction regression tasks (Lower is better)



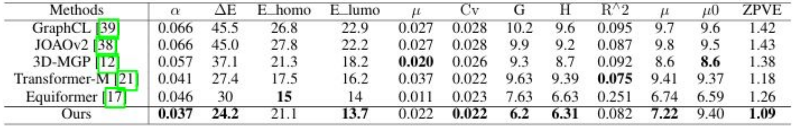
| Datasets | ESOL | FreeSolv | Lipo | QM7 | QM8 | QM9 | MEAN (RMSE) | MEAN (MAE) |
|---|---|---|---|---|---|---|---|---|
| # Molecules | 1128 | 642 | 4200 | 6830 | 21786 | 133885 | | |
| # Metric | | RMSE↓ | | | MAE↓ | | | |
| D-MPNN [37] | 1.050 | 2.082 | 0.683 | 103.5 | 0.0190 | 0.00814 | 1.272 | 34.509 |
| GROVERlarge [29] | 0.895 | 2.272 | 0.823 | 92.0 | 0.0224 | 0.00986 | 1.33 | 30.67 |
| MolCLR [33] | 1.271 | 2.594 | 0.691 | 66.8 | 0.0178 | - | 1.519 | - |
| GraphMVP [20] | 1.029 | - | 0.681 | - | - | - | - | - |
| GEM [5] | 0.798 | 1.877 | 0.660 | 58.9 | 0.0171 | 0.00746 | 1.112 | 19.642 |
| Uni-Mol[42] | 0.788 | 1.480 | 0.603 | 41.8 | 0.0156 | 0.00467 | 0.957 | 13.940 |
| Ours (Gamma) | 0.775 | 1.420 | **0.590** | **38.5** | **0.0142** | **0.00395** | 0.928 | **12.839** |
| Ours (Bernoulli) | **0.664** | **1.358** | 0.626 | 55.6 | 0.0154 | 0.0056 | **0.883** | 18.541 |

Table 3: Comparison against i-MolCLR on non-chirality MoleculeNet dataset



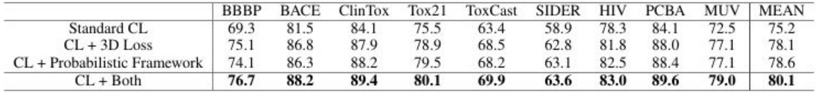
| Without Chirality | BBBP | BACE | ClinTox | Tox21 | SIDER | HIV | MUV | MEAN |
|---|---|---|---|---|---|---|---|---|
| I-MOLCLR [32] | 76.4 | 88.5 | **95.4** | 79.9 | 69.9 | 80.8 | **90.8** | 83.1 |
| Our Method | **78.3** | **94.8** | 91.4 | **84.9** | **72.7** | **85.5** | 88.0 | **85.1** |

Table 4: Experiment results on QM9 dataset



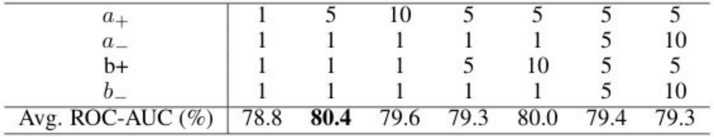
| Methods | $\alpha$ | $\Delta$E | E_homo | E_lumo | $\mu$ | Cv | G | H | R^2 | $\mu$ | $\mu$0 | ZPVE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GraphCL [39] | 0.066 | 45.5 | 26.8 | 22.9 | 0.027 | 0.028 | 10.2 | 9.6 | 0.095 | 9.7 | 9.6 | 1.42 |
| JOAOv2 [38] | 0.066 | 45.0 | 27.8 | 22.2 | 0.027 | 0.028 | 9.9 | 9.2 | 0.087 | 9.8 | 9.5 | 1.43 |
| 3D-MGP [12] | 0.057 | 37.1 | 21.3 | 18.2 | **0.020** | 0.026 | 9.3 | 8.7 | 0.092 | 8.6 | **8.6** | 1.38 |
| Transformer-M [21] | 0.041 | 27.4 | 17.5 | 16.2 | 0.037 | 0.022 | 9.63 | 9.39 | **0.075** | 9.41 | 9.37 | 1.18 |
| Equiformer [17] | 0.046 | 30 | **15** | 14 | 0.011 | 0.023 | 7.63 | 6.63 | 0.251 | 6.74 | 6.59 | 1.26 |
| Ours | **0.037** | **24.2** | 21.1 | **13.7** | 0.022 | **0.022** | **6.2** | **6.31** | 0.082 | **7.22** | 9.40 | **1.09** |

Table 5: Ablation Study on MoleculeNet Classification Datasets



| | BBBP | BACE | ClinTox | Tox21 | ToxCast | SIDER | HIV | PCBA | MUV | MEAN |
|---|---|---|---|---|---|---|---|---|---|---|
| Standard CL | 69.3 | 81.5 | 84.1 | 75.5 | 63.4 | 58.9 | 78.3 | 84.1 | 72.5 | 75.2 |
| CL + 3D Loss | 75.1 | 86.8 | 87.9 | 78.9 | 68.5 | 62.8 | 81.8 | 88.0 | 77.1 | 78.1 |
| CL + Probabilistic Framework | 74.1 | 86.3 | 88.2 | 79.5 | 68.2 | 63.1 | 82.5 | 88.4 | 77.1 | 78.6 |
| CL + Both | **76.7** | **88.2** | **89.4** | **80.1** | **69.9** | **63.6** | **83.0** | **89.6** | **79.0** | **80.1** |

Table 6: Abalation studies on hyperparameters for MoleculeNet classification tasks



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $a_+$ | 1 | 5 | 10 | 5 | 5 | 5 | 5 |
| $a_-$ | 1 | 1 | 1 | 1 | 1 | 5 | 10 |
| b+ | 1 | 1 | 1 | 5 | 10 | 5 | 5 |
| $b_-$ | 1 | 1 | 1 | 1 | 1 | 5 | 10 |
| Avg. ROC-AUC (%) | 78.8 | **80.4** | 79.6 | 79.3 | 80.0 | 79.4 | 79.3 |