A 3D-Shape Similarity-based Contrastive Approach to Molecular Representation Learning

Austin Atsango,[1] **Nathaniel Diamant,**[2] **Ziqing Lu,**[2]
Tommaso Biancalani,[2] **Gabriele Scalia,**[2†] **Kangway V. Chuang**[2†]
[1]Department of Chemistry, Stanford University
[2]Department of Artificial Intelligence and Machine Learning,
Genentech Research and Early Development
[†]{scalia.gabriele,chuang.kangway}@gene.com

Abstract

Molecular shape and geometry dictate key biophysical recognition processes, yet many graph neural networks disregard 3D information for molecular property prediction. Here, we propose a new contrastive-learning procedure for graph neural networks, **Mol**ecular **C**ontrastive **Lea**rning from **S**hape **S**imilarity (**MolCLaSS**), that implicitly learns a three-dimensional representation. Rather than directly encoding or targeting three-dimensional poses, **MolCLaSS** matches a similarity objective based on Gaussian overlays to learn a meaningful representation of molecular shape. We demonstrate how this framework naturally captures key aspects of three-dimensionality that two-dimensional representations cannot and provides an inductive framework for scaffold hopping.

1 Introduction and Background

Molecular shape is critical for biophysical processes, yet encoding relevant three-dimensional features remains challenging for many molecular property prediction tasks, especially when an understanding of three-dimensional shape is limited or unknown [1]. Numerous methods have been developed to effectively encode individual conformers that are both appropriate and highly-effective for conformer-level prediction tasks such as predicting quantum chemical properties of single conformational poses [2; 3; 4; 5; 6]. However, these approaches are poorly suited for representing complete molecules since relying on a single low-energy conformer to represent a diverse conformational ensemble is inherently limiting. Dietterich et al. [7] first recognized and addressed this challenge in the development of the multiple-instance learning framework [8]. Recent studies have explored deep multiple-instance learning approaches for learning on conformational ensembles [9; 10], yet are computationally-demanding due to the need to encode each conformer independently. Furthermore, methods that encode three-dimensional information often do not provide a strong performance benefit over 2D baselines [10].

Herein, we propose to inject graph neural networks with *implicit* 3D shape information through a supervised contrastive approach. Learning implicit 3D representations have been recently explored by Stärk et al. [11] and Liu et al. [12], but these approaches do not consider three-dimensional relationships between molecules and hence do not learn a direct measure of molecular shape similarity. In contrast, we propose to learn key features of molecular shape through direct comparison with the use of 3D molecular similarity kernels based on Gaussian overlays [13]. Specifically, methods such as the Rapid Overlay of Chemical Structures (ROCS) [14] provide a fast and scalable method for matching molecular shapes based both on molecular volumes and electrostatic matching on predefined pharmacophore features.

NeurIPS 2022 AI for Science Workshop.

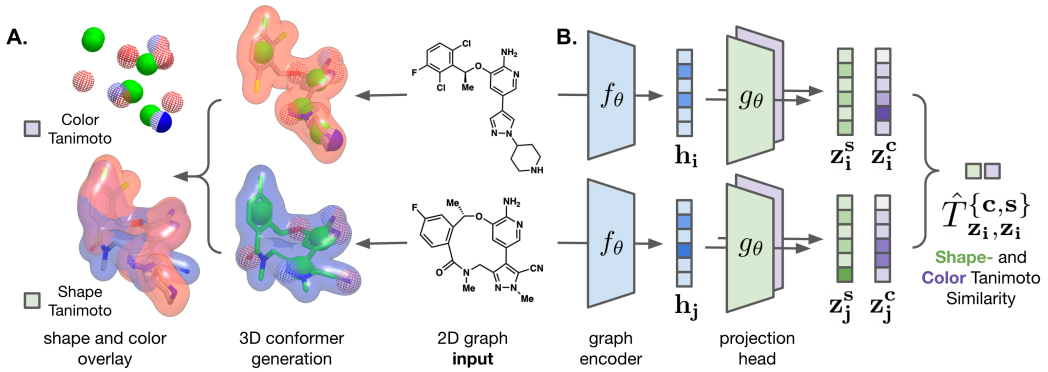

Figure 1: Our proposed model couples a pairwise, Gaussian shape- and pharmacophore similarity objective to a graph contrastive learning procedure to learn molecular embeddings with implicit 3D-information. **A.** For each input molecule, we sample a conformational ensemble and perform OpenEyeROCS to find a maximum overlay via ShapeTanimoto and ColorTanimoto similarities. **B.** The resulting pairwise, 3D shape similarities become the objective of our graph contrastive learning procedure on 2D graphs *only*, which are approximated using a Tanimoto kernel function.

Our framework, **Mol**ecular **C**ontrastive **Lea**rning on **S**hape **S**imilarity (MolCLaSS), naturally aligns a pairwise shape similarity objective with supervised contrastive learning framework to learn meaningful representations of molecular shape (Figure 1). Our work contributes directly in two ways: 1) MolCLaSS provides fast, scalable, and inductive approximation of 3D-shape similarity scores between molecules directly from their 2D graphs, and does not require conformer generation and shape alignment for inference, and 2) we demonstrate how MolCLaSS learns meaningful molecular embeddings that naturally capture 3D-shape and features that 2D, topological methods cannot. The resulting pre-trained encoder can be used for downstream molecular shape tasks.

2 Related Work

Graph Neural Networks for Small Molecules Graph neural networks have been widely developed for predicting small molecule properties and activities [15; 16; 17] on both 2D and 3D tasks. The properties of individual conformers have been effectively modeled by leveraging 3D spatial features for a range of quantum chemical and property prediction tasks [3; 4; 5; 6]. Adams et al. [18] recently described a hybrid approach for conformationally-invariant 3D approaches. Furthermore, "4D" multiple-instance learning methods have recently been developed that operate over sets of conformers that are each modeled as a graph [9; 10]. Our work builds on this prior work to learn improved and compact representations.

Pre-Training Graph Neural Networks Graph pretraining methods are an active area of research. Hu et al. [19] first reported a pretraining strategy based on self-supervised node pre-training, followed by supervised pre-training with masking. You et al. [20] recently demonstrated the effectiveness of a contrastive learning approach on graphs, which was further extended by Wang et al. [21, 22]. Recently, Stärk et al. [11] and Liu et al. [12] developed self-supervised approaches to maximize three-dimensional information for 2D-graph neural network pretraining with promising results. These approaches aim to maximize mutual information with the goal of matching a 2D representation to a three-dimensional pose, but they do not explicitly consider three-dimensional similarity between molecules. Our work here presents a supervised contrastive approach for pretraining neural networks that complements the self-supervised approaches above.

Kernel-Based Approximations of Molecular Shape Our work is closely related to prior work on learning low dimensional embeddings from molecular similarity kernels. Raghavendra and Maggiora [23] introduced a method to learn molecular basis vectors by directly decomposing Tanimoto similarities. The SCISSORS method by Haque & Pande [24; 25; 26] generalizes a kernel PCA approach to molecular similarity measures, including ROCS, that provides a fast approximation for molecular similarity. These prior approaches are naturally transductive; to obtain new scores, molecules must be scored against the resulting basis set and predicted based on a least-squares

estimate. Our approach couples the intuition of Haque and Pande [24] with modern graph neural networks to directly learn a meaningful molecular space that is naturally inductive, i.e. can generalize to new and unseen molecules without the need for additional conformer generation and scoring.

3 Problem Formulation and Methods

We invoke a variation of the classic similar property principle and assume that two molecules are similar if they can adopt similar molecular shapes [27; 28]. Prior approaches for encoding three-dimensional information typically operate on a molecular input x_i and corresponding spatial features s_i that includes explicit atomic coordinate, distance, or angle information. These approaches rely on designing an expressive transformation $f_\theta(x_i, s_i)$ that can accurately learn to map similar molecules to similar parts of chemical space, and assume that a single or several low-energy conformers encode relevant geometry. Rather than operate over explicit spatial representations, we instead adopt an *implicit* strategy that leverages a predefined similarity kernel over pairs of molecular inputs $k(x_i, x_j)$. Critically, by leveraging a well-defined 3D similarity function using inner products, we avoid the need to explicitly encode spatial features s_i, and can focus on learning an invariant function $f_\theta(x_i)$.

As illustrated in Figure 1, our approach naturally fits a supervised contrastive learning framework: given a set of molecules, our goal is to learn an expressive representation satisfies a pairwise similarity constraint $k(x_i, x_j)$. Here, we decompose this representation into graph encoder f_θ and projection heads g_θ to flexibly model multiple outputs. In this study, we use Gaussian shape and color overlays as an intuitive measure of shape similarity [13; 1], and follow the insight of Haque and Pande [24, 25] to leverage the Tanimoto kernel [29]:

$$T(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\mathbf{z}_i \cdot \mathbf{z}_i + \mathbf{z}_j \cdot \mathbf{z}_j - \mathbf{z}_i \cdot \mathbf{z}_j} \tag{1}$$

This interpretation defines a molecular embedding space, where three-dimensional shape similarities can be conveniently modeled based on inner products. Rather than approximate embeddings $\mathbf{z_i}$ via linear decomposition methods, our objective is to learn an inductive model that generates $\mathbf{z_i}$ directly from a molecular graph. Here, we use an encoder based on the Graph Isomorphism Network with Edge features [19] followed by two projection heads to model Shape and Color separately:

$$\mathbf{v}_p^{t+1} = q_\theta((1+\epsilon) \cdot \mathbf{v}_p^t + \sum_{q \in \mathcal{N}(p)} \sigma(\mathbf{v}_q^t + \mathbf{e}_{p,q})) \tag{2}$$

$$\text{with} \quad f_\theta(\mathbf{x}_i) = \mathbf{h}_i = \sum_{p \in G} \mathbf{v}_p^T \tag{3}$$

$$\mathbf{z}_i^{\{c,s\}} = g_\theta(\mathbf{h}_i) = U^{\{c,s\}} \sigma(V^{\{c,s\}} \mathbf{h}_i) \tag{4}$$

Above, $\mathbf{v}_p^t$ corresponds to the hidden state of the node p at step t (final step T), with $\mathbf{h}_i$ as the final graph representation of molecule i using sum pooling. The projection heads $g_\theta(\mathbf{h}_i)$ are parameterized by Color- and Shape-dependent MLPs with trainable weight matrices U and V, and σ is a ReLU nonlinearity. Finally, we optimize the network to directly predict ShapeTanimoto and ColorTanimoto scores via minimization of the following loss function:

$$\mathcal{L}_{\{s,c\}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(T_{\{s,c\}}(\mathbf{z}_i, \mathbf{z}_j) - k_{\{s,c\}}(\mathbf{x}_i, \mathbf{x}_j) \right)^2 \quad \text{with } \mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_c \tag{5}$$

Here, we define the loss as the mean squared error over all the pairs in a batch (size N) between the predicted Color and Shape scores based on the Tanimoto kernel (Eq. 1) and the calculated Gaussian overlay $k_{\{s,c\}}(\mathbf{x}_i, \mathbf{x}_j)$ based on conformer generation and alignment. We balance the individual objectives L_s and L_c with the tuneable hyperparameter λ to adjust the influence of pharmacophore features (but typically set to 1).

		ShapeTanimoto			ColorTanimoto		
Model	n	r	R^2	MAE	r	R^2	MAE
2D Tanimoto	–	0.000	–	–	0.000	–	–
ECFP4 + FF-NN	10k	0.733	0.513	0.0473	0.502	0.144	0.0484
ECFP4 + FF-NN	50k	0.793	0.610	0.0419	0.593	0.243	0.0456
ECFP4 + FF-NN	100k	0.822	0.660	0.0388	0.636	0.327	0.0428
MolCLaSS	10k	0.818	0.653	0.0396	0.608	0.295	0.0442
MolCLaSS	50k	0.876	0.757	0.0327	0.718	0.484	0.0374
MolCLaSS	**100k**	**0.893**	**0.794**	**0.0301**	**0.748**	**0.537**	**0.0353**

Table 1: Model performance for prediction of ShapeTanimoto and ColorTanimoto scores from 2D graphs. All results are reported on a random, independent test set of 49,621 molecules corresponding to 1.23 billion pairwise similarity scores. For each training set size n, we report Pearson's correlation coefficient r, the coefficient of determination R^2, and the mean absolute error (MAE) for all 1.23 billion pairwise scores.

4 Experiments and Results

Prediction of Shape- and Color-Tanimoto Scores from 2D Graphs We systematically investigated the ability of contrastive models to directly predict ShapeTanimoto and ColorTanimoto scores of drug-like molecules from the ChEMBL database [30; 31]. For these studies, we generated a complete all-by-all similarity matrix of Shape- and ColorTanimoto scores [13; 14] for 100k molecules from ChEMBL, with maximum overlay scores recorded from pairwise comparisons of up to 10 conformers generated per molecule [32; 33] (see Appendix for complete details). We refer to the complete data set as `ROCS100k`. We assessed the ability of both fingerprint- and graph-based models to accurately predict pairwise, 3D similarity scores directly from their 2D graph representations. Although thissetup requires computationally-intensive conformer generation and exhaustive similarity scoring, this cost is amortized over the training data. At inference time, predicted Shape- and Color-Tanimoto similarities are directly obtained and circumvent the need for explicit three-dimensional representations.

As illustrated in Table 4, we directly compared our approach to a Morgan fingerprint [34] and dense neural network baseline, with positive performance gains for increasingly large data set sizes. Our graph neural network-based approach based on GINEConv graph layers with independent projection heads for ShapeTanimoto and ColorTanimoto predictions exhibits a clear improvement over hashed fingerprint representations, even with significantly less data. For example, MolCLaSS trained with only 10k examples achieves nearly identical performance to fingerprint-based models trained on 100k examples. Critically, these dense networks learn a non-trivial transformation of the input data. Indeed, as shown in Table 4 (first row), there is nearly no correlation between bulk Tanimoto scores on 2D representations and their 3D Tanimoto scores.

Notably, the MolClASS network can directly predict 3D similarity scores with good accuracy (ShapeTanimoto MAE$= 0.030$, ShapeTanimoto MAE$= 0.035$) at only a fraction of the computational cost. At inference, predicting 3D similarity scores on tens of thousands of molecules takes only seconds, replacing both the need for conformer generation and overlays and representing an improvement of nearly $10^4 - 10^5$ times in speed.

MolCLaSS Representations Capture 3D Shape Similarity. Given the strong performance of MolCLaSS, we investigated the qualities of the learned molecular embeddings. We specifically visualized our hold-out test set of ChEMBL molecules in their graph embedding ($\mathbf{h_i}$) and projection heads ($\mathbf{z_i^s}, \mathbf{z_i^c}$) (Figure 2) against a fingerprint baseline, and colored them based on calculated three-dimensional descriptors [28; 35] of their low-energy conformers, including radius of gyration (Figure 2A) and the first principal moment of inertia (Figure 2B). As shown, both the graph-layers and shape-projection layer provide clear localization based on these 3D properties when compared to Morgan fingerprints. As expected, the color projection head does not exhibit the same localization, as pharmacophore features are less dependent on overall molecular shape.

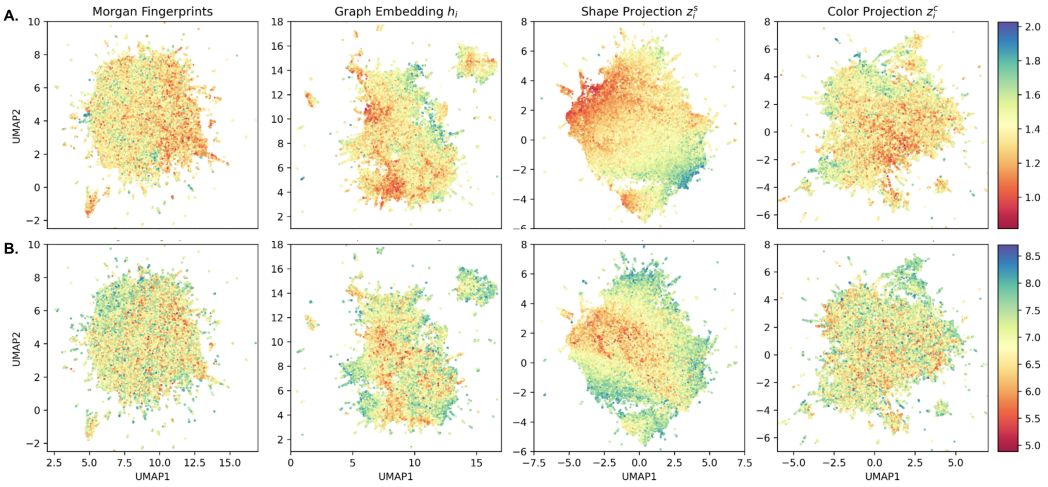

Figure 2: Comparison of MolClaSS latent graph encodings and projection heads to topological fingerprints reveals that MolCLaSS provides meaningful latent organization by three-dimensional features. Latent representations colored by the **A. radius of gyration (log-scale)** and **B. first principal moment of inertia (log-scale)** for a low-energy conformer. The graph embeddings $\mathbf{h}_i$ and shape projection $\mathbf{z}_j^s$ learn more localized structure corresponding to human interpretable features.

A key consequence of the Tanimoto similarity objective (Eq. 1) is that it induces a Euclidean structure over the projected vector space z [36]. Indeed, we found an excellent correlation (Pearson $r = 0.87$) between pairwise Euclidean distance and ShapeTanimoto scores (see Appendix). To further probe the difference in representations we performed a nearest neighbors analysis using the shape projections in MolCLaSS (Figure 3). In our analysis, we consistently find that MolCLaSS preserves molecular shape and size. Scaffolds hops based on ring mutations are found nearby while maintaining excellent overall shape similarity. In contrast, topological fingerprints largely favor substructure matching and exhibits a wider range of molecular shapes. Together, these two studies illustrate how the MolCLaSS framework can capture relevant shape measures through a supervised approach.

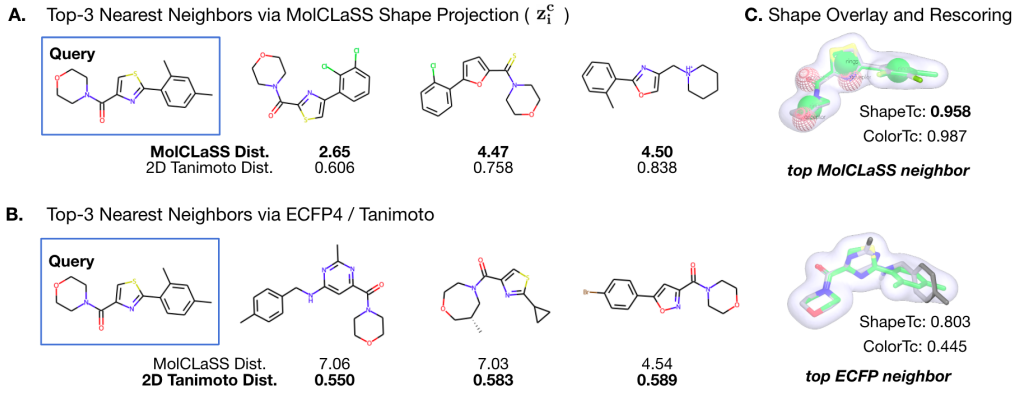

Figure 3: Nearest neighbors analysis using learned latent representations. We analyze the test set using k-nearest neighbors to retrieve similar hits based on a given query (top right). **A.** Nearest neighbors with MolCLaSS (Euclidean distance) preserves scaffold shape and can hop between scaffolds through core mutations (top row). **B.** The same analysis performed with ECFP4 fingerprints and Tanimoto distance (bottom row). Topological fingerprints heavily emphasize subgraph matches. Although diverse hits are found, nearest neighbors have a lower shape similarity. **C.** Reanalysis of top hits by fastROCS demonstrates how MolCLaSS accurately identifies close three-dimensional matches (top right), whereas 2D fingerprints produce lower-quality matches.

5 Conclusions and Future Directions

Our studies above outline a preliminary roadmap for learning three-dimensional representations based on 3D similarities. In contrast to prior work, MolCLaSS learns via supervised pairwise comparisons, and hence is able to relate and differentiate molecules of varying size and shape. We have demonstrated how the MolCLaSS network itself provides a direct and fast approximation method for approximating shape- and 3D pharmacophore-based, and further illustrated how meaningful three-dimensional features are naturally learned through this inductive framework. Our ongoing work seeks to further improve the MolCLaSS framework to improve predictive performance and to explore its application as a pretrained model for broad molecular property prediction tasks.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge members of the Department of Artificial Intelligence and Machine Learning at Genentech Research and Early Development for helpful feedback and support.

References

[1] Anthony Nicholls, Georgia B McGaughey, Robert P Sheridan, Andrew C Good, Gregory Warren, Magali Mathieu, Steven W Muchmore, Scott P Brown, J Andrew Grant, James A Haigh, Neysa Nevins, Ajay N Jain, and Brian Kelley. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.*, 53(10):3862–3886, May 2010.

[2] Seth D. Axen, Xi-Ping Huang, Elena L. Cáceres, Leo Gendelev, Bryan L. Roth, and Michael J. Keiser. A simple representation of three-dimensional molecular structure. *Journal of Medicinal Chemistry*, 60(17):7393–7409, 2017. doi: 10.1021/acs.jmedchem.7b00696. URL `https://doi.org/10.1021/acs.jmedchem.7b00696`. PMID: 28731335.

[3] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[4] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf`.

[5] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=B1eWbxStPH`.

[6] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=givsRXsOt9r`.

[7] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1):31–71, January 1997.

[8] Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans. *Multiple Instance Learning: Foundations and Algorithms*. Springer, November 2016.

[9] Kangway V. Chuang and Michael J. Keiser. Attention-based learning on molecular ensembles. *Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io*, 2020. URL `https://arxiv.org/abs/2011.12820`. arXiv:2011.12820 [cs.LG].

[10] Simon Axelrod and Rafael Gomez-Bombarelli. Molecular machine learning with conformer ensembles, 2020. URL `https://arxiv.org/abs/2012.08452`. arXiv:2012.08452 [cs.LG].

[11] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D infomax improves GNNs for molecular property prediction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20479–20502. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

[12] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

[13] J A Grant and B T Pickup. A gaussian description of molecular shape. *J. Phys. Chem.*, 99(11): 3503–3510, March 1995.

[14] Paul C D Hawkins, A Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.*, 50(1):74–82, January 2007.

[15] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf.

[16] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.*, 30(8):595–608, August 2016.

[17] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, August 2019.

[18] Keir Adams, Lagnajit Pattanaik, and Connor W. Coley. Learning 3d representations of molecular chirality with invariance to bond rotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hm2tNDdgaFK.

[19] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

[20] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3fe230348e9a12c13120749e3f9fa4cd-Paper.pdf.

[21] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287, March 2022.

[22] Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *J. Chem. Inf. Model.*, 62(11):2713–2725, June 2022.

[23] Akshay S Raghavendra and Gerald M Maggiora. Molecular basis sets - a general similarity-based approach for representing chemical spaces. *J. Chem. Inf. Model.*, 47(4):1328–1340, July 2007.

[24] Imran S Haque and Vijay S Pande. SCISSORS: a linear-algebraical technique to rapidly approximate chemical similarities. *J. Chem. Inf. Model.*, 50(6):1075–1088, June 2010.

[25] Imran S Haque and Vijay S Pande. Error bounds on the SCISSORS approximation method. *J. Chem. Inf. Model.*, 51(9):2248–2253, September 2011.

[26] Steven M Kearnes, Imran S Haque, and Vijay S Pande. SCISSORS: practical considerations. *J. Chem. Inf. Model.*, 54(1):5–15, January 2014.

[27] M A Johnson and G M Maggiora. *Concepts and applications of molecular similarity*. John Wiley & Sons, New York., 1990.

[28] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. John Wiley & Sons, July 2008.

[29] Liva Ralaivola, Sanjay J Swamidass, Hiroto Saigo, and Pierre Baldi. Graph kernels for chemical informatics. *Neural Netw.*, 18(8):1093–1110, October 2005.

[30] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40 (Database issue):D1100–7, January 2012.

[31] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michal Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, 47(D1): D930–D940, January 2019.

[32] Paul C D Hawkins, A Geoffrey Skillman, Gregory L Warren, Benjamin A Ellingson, and Matthew T Stahl. Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.*, 50(4):572–584, April 2010.

[33] Paul C D Hawkins and Anthony Nicholls. Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J. Chem. Inf. Model.*, 52(11):2919–2936, November 2012.

[34] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50 (5):742–754, May 2010.

[35] Greg Landrum. RDKit: Open-source cheminformatics, 2006. URL `https://rdkit.org`.

[36] Matthew B A McDermott, Brendan Yap, Peter Szolovits, and Marinka Zitnik. Structure inducing Pre-Training. arXiv:2103.10334 [cs.LG], March 2021. URL `https://arxiv.org/abs/2103.10334`.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

[38] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch geometric. arXiv:1903.02428 [cs.LG], March 2019. URL `https://arxiv.org/abs/1903.02428`.

Appendix

All experiments were performed using Python and standard numerical libraries. For cheminformatics analysis, all molecules were processed using either OpenEye Applications and Toolkits or the open-source cheminformatics library RDKit [35]. We implemented all experiments in Python using PyTorch `1.11` [37] and PyTorch Geometric [38].

Appendix A: Molecular Dataset Properties and Creation

All data were downloaded directly from ChEMBL31 release of the ChEMBL database [30; 31]. Molecules were filtered based on OpenEye Filter using the drug-likeness that limits. The resulting detailed statistics of the molecules and some of their properties are shown in Table 2, with a representative sample of molecules shown in Figure 4.

OMEGA Conformer Generation

For each molecule, we generate a conformational ensemble using OpenEye Applications (`2022.1.1`) using OMEGA (`v.4.2.0`) [32; 33]. Conformational ensembles were generated with the optimized default `fastROCS` settings with multiprocessing: `maxconfs=10, ewindow=15, flipper=False, mpi=128` which been shown to accurately recapitulate binding poses. Molecules with ambiguous or undefined stereochemistry were dropped during the conformer generation process.

OpenEye fastROCS Scoring

We use OpenEye's GPU-accelerated fastROCS Toolkit (`v.2.2.2.1`) to calculate ShapeTanimoto and ColorTanimoto scores. For each conformer database, we generate an all-by-all similarity matrix that scores every conformer of every molecule against all conformers of the rest of the database, saving the maximum score between two molecules. In practice, generation of the ROCS100k training dataset generates 4.95 billion ShapeTanimoto and ColorTanimoto scores, each, and requires 7 hours on 3 NVIDIA 3090 RTX GPUs with 80 processes.

Computational Complexity of All-by-All Conformer ROCS Exhaustive similarity comparisons between n molecules scales at $\mathcal{O}(n^2)$. As an upper limit, for k conformers are generated per molecule, generation of the all-conformer by all-conformer similarity matrix scales at $\mathcal{O}(k^2n^2)$. For symmetric similarity measures like ColorTanimoto and ShapeTanimoto, $\mathcal{O}((n)(n-1)/2)$ comparisons are required with no self-comparisons. Similarity, for k sampled conformers we require $\mathcal{O}((kn)(kn-1)/2)$ pairwise comparisons.

Table 2: Detailed Statistics of the ROCS100k Training Dataset (n = 100,000).

Property	min.	max.	mean	median	std.
Sampled Conformers (k)	1	10	9.48	10	1.78
Heavy Atom Count	15	35	23.88	24	4.71
Molecular Weight	220.20	598.03	337.73	334.42	66.38
Rotatable Bonds	0	11	4.51	4	1.99
Aromatic Rings	0	12	3.14	3	1.01
H-Bond Donors	0	5	1.44	1	0.98
H-Bond Acceptors	0	12	4.34	4	1.72
Heteroatoms	2	14	6.17	6	1.84

Appendix B: Network Architecture and Training Details

All studies were trained using the generated ROCS100k dataset. The size dependence studies illustrated in 4 using 10k and 50k examples use subsets of the full 100k molecules. A separate validation set of 8,548 molecules from ChEMBL31 were additionally used for model tuning and selection. All results in Table 4 are reported on an independent, random test set of 49,621 molecules from ChEMBL31.

Table 3: Detailed Statistics of the ROCS50k Test Dataset (n = 49,621).

Property	min.	max.	mean	median	std.
Sampled Conformers (k)	1	10	9.53	10	1.70
Heavy Atom Count	15	35	23.66	23	4.77
Molecular Weight	220.17	557.52	335.4	331.42	67.56
Rotatable Bonds	0	10	4.42	4	1.96
Aromatic Rings	0	8	3.08	3	1.00
H-Bond Donors	0	5	1.31	1	0.97
H-Bond Acceptors	0	12	4.62	5	1.70
Heteroatoms	2	14	6.37	6	1.85

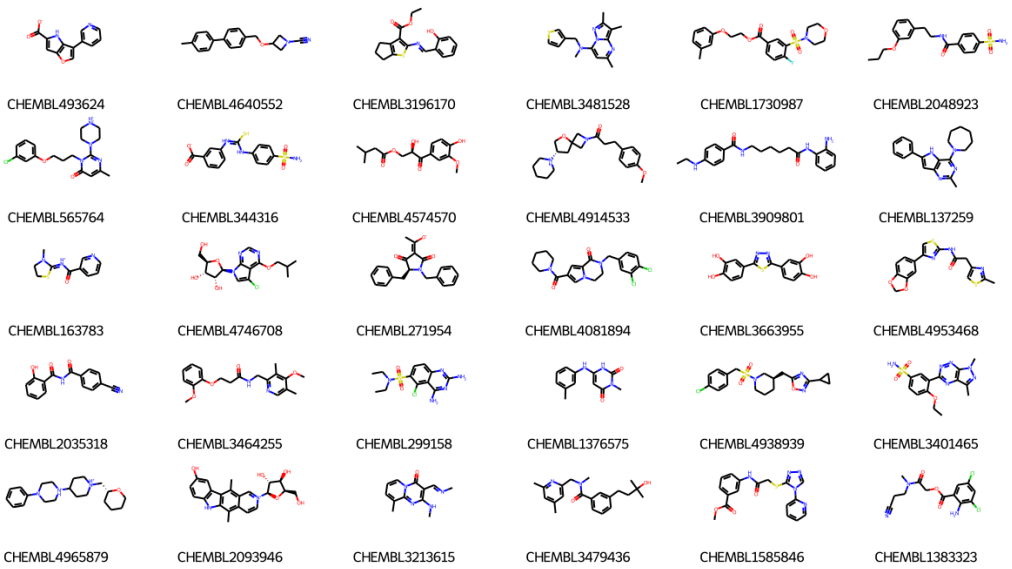

Figure 4: Random selection of 50 examples from ChEMBL31 for generating the ROCS100k dataset.

All neural networks were trained using the Adam optimizer (learning rate = 1×10^{-3} to 1×10^{-4}) and a batch size of 2048 for up to 4,000 epochs, using the early stopping criterion based on the validation set described above. The model architecture and hidden dimensions are specified in Appendix Table 4. All networks use five graph encoding layers with two project heads (single-hidden layer MLPs). The entire network is trained on Shape- and Color-Tanimoto targets bounded within [0,1], using mean squared error as the loss criterion and trained to early stopping.

Neural Network Performance

Our summary of overall performance results and metrics for different model architectures are shown in Table 4. Below, we include scatter plots of our best model trained on the full ROCS100k dataset in Figure 5.

Correlation of Euclidean Distance with ShapeTanimoto and ColorTanimoto

The project heads g_θ ultimately learn a meaningful Euclidean distance. As shown in the plots below, molecules closer in the embedded latent space also tend to have a much higher Shape and ColorTanimoto score (Figure 6).

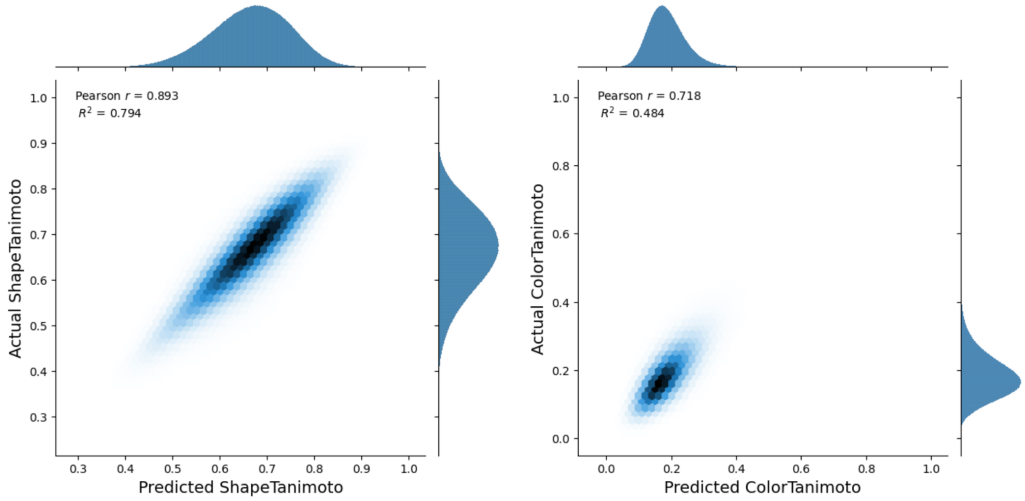

Figure 5: Performance of MolCLaSS for predicting ShapeTanimoto and ColorTanimoto trained on the full `ROCS100k` data set, corresponding to the final entry in Table 4. results plotted are on the full 49,621 test set molecules.

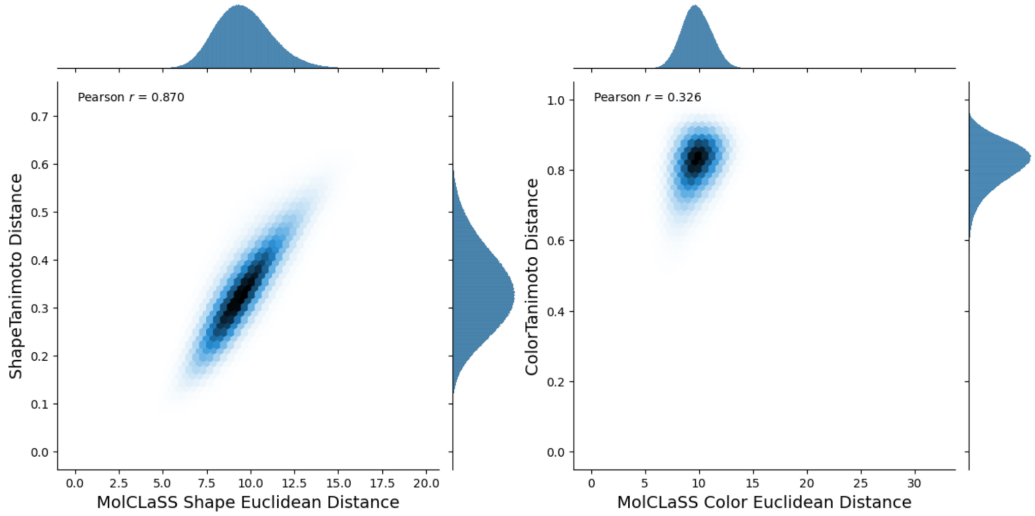

Figure 6: Plots comparing distance in ShapeTanimoto an ColorTanimoto, vs Euclidean distance for the Shape- and Color-projection heads. MolClaSS learns a meaningful structural embedding space where similar shapes are closer together in their latent representation (left), with a more modest correlation observed for ColorTanimoto (right). Correlations plotted for the unseen 49,621 molecules in the test set.

Table 4: Neural Network Model Architectures

Model & Module	Layer & Description	h_{dim} Sizes
MolCLaSS		
Graph Encoder f_θ	5 x [GINEConv + BatchNorm w/ ReLU]	$5 \times (512 \rightarrow 1024 \rightarrow 512)$
Projection Head g_θ	MLP + ReLU	$512 \rightarrow 1024 \rightarrow 256$
ECFP4 + FF-NN		
Encoder f_θ	MLP w/ ReLU	$2048 \rightarrow 2048 \rightarrow 512$
Projection Head g_θ	MLP + ReLU	$512 \rightarrow 1024 \rightarrow 256$