Attention-Based Learning on Molecular Ensembles

Kangway V. Chuang
Department of Pharmaceutical Chemistry
Institute for Neurodegenerative Diseases
University of California, San Francisco
San Francisco, CA 94143
kangway.chuang@ucsf.edu

Michael J. Keiser
Department of Pharmaceutical Chemistry
Institute for Neurodegenerative Diseases
University of California, San Francisco
San Francisco, CA 94143
keiser@keiserlab.org

Abstract

The three-dimensional shape and conformation of small-molecule ligands are critical for biomolecular recognition, yet encoding 3D geometry has not improved ligand-based virtual screening approaches. We describe an end-to-end deep learning approach that operates directly on small-molecule conformational ensembles and identifies key conformational poses of small-molecules. Our networks leverage two levels of representation learning: 1) individual conformers are first encoded as spatial graphs using a graph neural network, and 2) sampled conformational ensembles are represented as sets using an attention mechanism to aggregate over individual instances. We demonstrate the feasibility of this approach on a simple task based on bidentate coordination of biaryl ligands, and show how attention-based pooling can elucidate key conformational poses in tasks based on molecular geometry. This work illustrates how set-based learning approaches may be further developed for small molecule-based virtual screening.

1 Introduction

Molecular shape and geometry are key for highly-specific biophysical recognition. Despite the critical importance of molecular conformation, ligand-based methods that incorporate three-dimensional features have had limited impact on virtual screening and drug discovery [1]. In contrast to structure-based methods, where multiple conformational poses of a molecule can be individually docked and scored against a protein pocket, methods based on ligand similarity present an inherent challenge. Drug-like molecules adopt diverse conformational shapes and orientations, yet for new systems, the relevant conformations are not known *a priori*. Indeed, discovering key conformations and binding modes is often the goal of therapeutic discovery. Furthermore, this ambiguity in molecular representation introduces further challenges as most standard predictive models expect a single input.

The challenges of input representation are highlighted in the context of small-molecule ligand-protein binding (Figure 1). Whereas two-dimensional topological representations have been effective for similarity-based approaches, these fingerprints do not capture the spatial geometry of a small molecule, nor its diverse conformational ensemble (Figure 1A). Furthermore, although low-energy conformations can be generated readily, the lowest energy conformation of a drug often differs significantly from its bioactive pose (Figure 1B). Dietterich et al. [2] formalized this task as the multiple instance learning (MIL) problem, where a set of multiple input instances are mapped to a single output label.

Herein, we report an end-to-end deep multiple-instance learning approach that addresses the challenges of encoding molecular ensembles. We combine the expressive power of graph neural networks for molecular representation with an attention-based network for set aggregation to learn representations of conformational ensembles. We demonstrate on a new molecular dataset how this

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

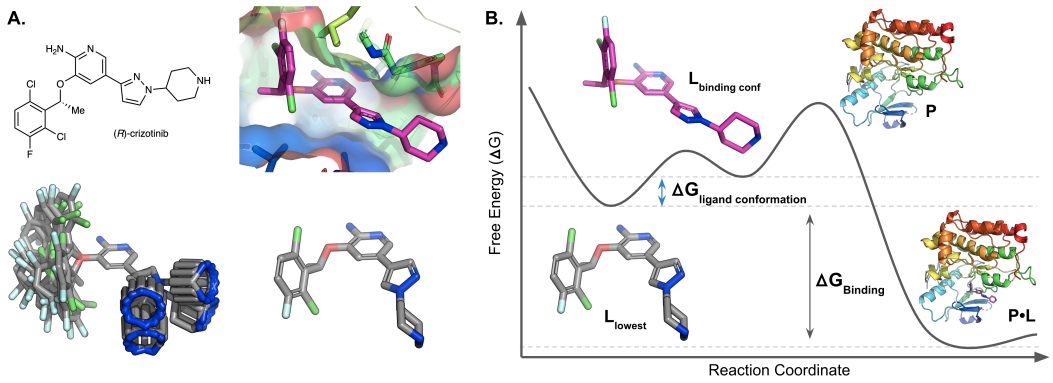

Figure 1: **A.** Two-dimensional topological representations do not capture the inherent three-dimensional shape or dynamics of a molecule. Selecting a single low-energy conformer often does not reflect the salient geometric pose necessary for binding. **B.** A simplified, hypothetical free energy landscape for protein-ligand binding (PDB 5AAB shown). Flexible molecules must traverse higher-energy conformations prior to ligation to the target protein.

approach can be used to simultaneously provide good predictive performance while elucidating key conformational instances on tasks requiring molecular shape.

2 Related Work

Multiple-Instance Learning for Molecules Dietterich et al. [2] first formalized the framework of multiple instance learning (MIL) motivated by small-molecule odor prediction. Fu et al. [3] and Zhao et al. [4] investigated support vector machine-based MIL approaches for drug activity prediction by converting embedded sets into instances. Recently, Zankov et al. [5] reported a deep-MIL approach for drug-activity prediction using 3D-derived features and mean-pooling approach for set embedding. Our work differs from these seminal papers in using an end-to-end, differentiable graph neural network to simultaneously learn optimal feature extractors and a set aggregation function.

Attention-Based Pooling for Multiple-Instance Learning Zaheer et al. [6] have characterized the requirements of permutation-invariant functions for set-based representations and illustrate their approach on set statistic estimation and anomaly detection. Our work builds off of Ilse et al. [7] who describe an attention-based, deep-multiple instance learning approach and demonstrate its application to diverse domains including histopathology. Yan et al. [8] extended this method with a dynamic-routing based attention mechanism on similar tasks. Recently, Lee et al. [9] illustrate the strong performance of transformer-based networks for rich representation learning on sets.

Graph Neural Networks for Small Molecules Our instance-level representations are motivated by recent work on graph neural networks for small molecules. Early work on differentiable molecular fingerprints by Duvenaud et al. [10] and Kearnes et al. [11] illustrated performance gains vs standard 2D-topological graphs. The properties of individual conformers can be efficiently estimated by leveraging spatial features as illustrated by Gilmer et al. [12], Schütt et al. [13, 14], and Klicpera et al. [15]. In the context of small molecule bioactivity prediction, PotentialNet by Feinberg et al. [16] and ChemProp by Yang et al. [17] performed well across multiple benchmarks [18].

3 Problem Formulation and Methods

Whereas true biophysical systems are dynamic and dictated by complex enthalpic and entropic contributions, we greatly simplify protein-ligand binding problem to the static recognition of molecular geometry. In this framework, we invoke a variation of the molecular similarity principle [19] and assume that two molecules are similar if they can adopt similar molecule geometries. We further assume that a molecule can be effectively represented by a set of discrete, sampled conformers. As a result, the similarity of two molecules can determined by their sets of of sampled conformers. In this

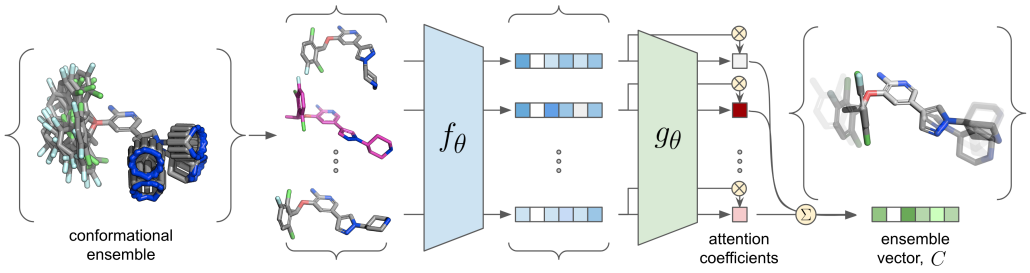

Figure 2: Our network encodes each conformer using a graph neural network, $\mathbf{f}_\theta$, as an instance featurizer, and uses an attention mechanism $\mathbf{g}_\theta$ to aggregate the embeddings from individual conformers. The resulting end-to-end differentiable network can be tied to both classification and regression tasks.

domain, we aim to learn directly on molecular sets while identifying key conformational instances for similar molecules.

In standard ligand-based machine learning, each molecule serves as an input, X, with an accompanying label y. A multiple-instance learning approach [20] extends this supervised paradigm: each molecule is represented as a set of K conformers, $X = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_K}\}$, where $\mathbf{x_K}$ denotes a single sampled conformer, with only a single set-level label Y. Although instances in the set are responsible for the overall set label, instance-level labels are not known *a priori* (i.e. y_k corresponding to x_k are hidden or unknown). Importantly, molecules vary substantially in both size and conformational flexibility. Any representation of the set must therefore satisfy both permutation invariance (i.e. exchangeability of its elements) and handle variable-length set sizes.

In this work, we describe an embedding-based approach with two levels of representation learning that consists of: 1) an instance featurizer, $f_\theta(x)$, that can expressively embed each conformer into a set of K graph embeddings, $H = \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_K}\}$ and 2) a set-level aggregation function, $g_\theta(x)$, to pool instances into a fixed-length embedding (Figure 2). Here, we represent all conformers as three-dimensional graphs, and our instance featurizer $f_\theta(x)$ is a graph neural network. We specifically use the edge-conditioned neural network as described by Gilmer et al. [12] and Simonovsky and Komodakis [21], with the corresponding message and update functions for each iteration, t:

$$\mathbf{m_i}^{t+1} = \sum_{j \in N(i)} \mathbf{A}(\mathbf{e_{ij}}) \cdot \mathbf{h_j}^t \quad \text{with} \quad \mathbf{h_i}^{t+1} = \text{GRU}(\mathbf{m_i}^{t+1}, \mathbf{h_i}^t) \tag{1}$$

The message from each graph neighbor $\mathbf{m_i}^{t+1}$ depends on the corresponding distance and edge-type $\mathbf{e_{ij}}$ through the learned weight matrix $\mathbf{A}$. Although any symmetric and differentiable function can serve as our set aggregator, $g_\theta(x)$, we opt to use an attention-based aggregation function [22; 7] for simplicity and interpretability. For a set of K graph embeddings H:

$$\mathbf{c} = \sum_{k=1}^{K} \alpha_k \mathbf{h_k}, \quad \text{with} \quad \alpha_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \quad \text{and} \quad z_j = \mathbf{w^T} \tanh(\mathbf{V h_j^T}). \tag{2}$$

The resulting context vector $\mathbf{c}$ is calculated as the expectation over each conformer instance, and represents the entire conformational ensemble. The parameters of the attention-layer, and the graph neural network are simultaneously optimized via backpropagation, and hence are trained specifically for the prediction task. From one perspective, the attention network learns a conformer-level energy function that is normalized across a softmax distribution, mimicking the classic Boltzmann distribution for molecular ensembles.

4 Experiments and Results

As a proof-of-concept model system, we constructed a small synthetic dataset of biaryl ligands to model protein-ligand binding (BIPY-MIL). We specifically analyze bidentate coordination modes, in which a *s-cis* conformation of the 1,2-diamine is necessary for binding (see Appendix A). Each

Table 1: Classification results on the BIPY-MIL dataset.

Model	n	Acc.	AUROC	AUPRC	Top-1	Top-5	Top-10
RF + ECFP4	100	0.901	0.962	0.917	–	–	–
RF + ECFP4	500	**0.960**	**0.980**	**0.967**	–	–	
GNN + Attention	100	0.886	0.956	0.918	0.628	0.840	**0.935**
GNN + Attention	500	0.928	0.976	0.957	**0.807**	**0.885**	0.923
Lowest Energy Pose	–	–	–	-	0.038	0.045	0.352

molecule in the dataset is considered positive if *at least one conformer* bears a 1,2-diamine substructure with a dihedral angle of 0 degrees. Note that the network is only given a set-level label. The key conformer instances are never revealed to the network and are withheld solely for evaluation.

We trained our models in a binary classification setting, and found comparable performance with random forest baselines known to work well in low data domains (100 and 500 training examples, Table 1). Here, the simple fingerprints derived from the molecular graph (ECFP4) alone can predict the set level; however, these 2D-baselines do not offer an interpretable method to identify key instances within each set.

To understand the interpretability of this approach, we analyzed the attention coefficients for each conformer of each set, and used the rank order of the coefficients to compare against the hidden ground truth instance labels. Our model attributes its highest attention coefficient to a key instance in 80.7% (Top-1) of all positive test cases. As a comparison, the lowest-energy pose only predicts the key instance in 3.8% of cases, showing how encoding only a single lowest-energy conformer can create inject misleading bias. As depicted in Figure 3A, the attention coefficients aptly identify the key instance from a positive set (**1**), with high coefficients for conformers that are similar, but slightly out of plane. Figure 3B shows a true negative example. The network simultaneously predicts a correct set label (**0**), and although there is no key instance in the set, the attention coefficients remain highest for the two conformers with the smallest N-C-C-N dihedral angles.

5 Conclusions and Future Directions

Our studies demonstrate the potential to learn on conformational ensembles for simple molecular systems. We have specifically illustrated how a simple attention mechanism can automatically retrieve key, driving instances from a large set of possible conformations using an embedding-based, multiple instance learning approach. These preliminary results lay a promising foundation toward learning

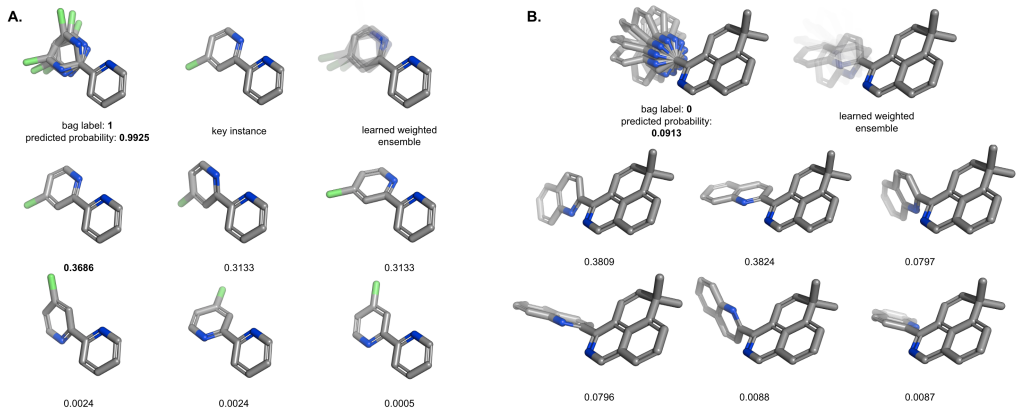

Figure 3: Two examples from the BIPY-MIL dataset and their corresponding attention. **A.** A true positive example with high classification confidence contains a key instance where the N-C-C-N motif adopt a coplanar *s-cis* conformation. **B.** A true negative example that does not contain the coplanar pattern, but contains two close instances. In both cases, the attention coefficients are significant for conformations most similar to the key instance (the instance that triggers the bag level).

more challenging molecular tasks, and we anticipate that additional adjustments in neural network architecture and attention mechanisms will enable this approach for complex tasks in small-molecule property and activity prediction. Our ongoing work is focused on further developing this approach and its applications toward real-world drug discovery.

Acknowledgments and Disclosure of Funding

We thank Laura Gunsalus for insightful discussion on graph neural networks and attention mechanisms. We also thank Will Connell, Dr. Jessica McKinley, and members of the Keiser laboratory for manuscript feedback and suggestions. OpenEye Scentific is gratefully acknowledged for an academic license to the OpenEye Toolkits, including OMEGA used for all conformer generation used in this paper. We are grateful to the Arnold and Mabel Beckman Foundation for generous support of this work through an Arnold O. Beckman Postdoctoral Fellowship in the Chemical Sciences to K. V. C., and the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation, for Grant 2018-191905 to M.J.K.

References

[1] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jurgen Bajorath. Molecular similarity in medicinal chemistry. *J. Med. Chem.*, 57(8):3186–3204, 2013.

[2] 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.

[3] Gang Fu, Xiaofei Nan, Haining Liu, Ronak Y Patel, Pankaj R Daga, Yixin Chen, Dawn E Wilkins, and Robert J Doerksen. Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinformatics*, 13 Suppl 15:S3, September 2012.

[4] Zhendong Zhao, Gang Fu, Sheng Liu, Khaled M Elokely, Robert J Doerksen, Yixin Chen, and Dawn E Wilkins. Drug activity prediction using multiple-instance learning via joint instance and feature selection. *BMC Bioinformatics*, 14 Suppl 14:S16, October 2013.

[5] Dmitry V Zankov, Maxim D Shevelev, Alexandra V Nikonenko, Pavel G Polishchuk, Asima I Rakhimbekova, and Timur I Madzhidov. Multi-instance learning for Structure-Activity modeling for molecular properties. In *Analysis of Images, Social Networks and Texts*, pages 62–71. Springer International Publishing, 2020.

[6] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017.

[7] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden*, pages 10–15, 2018.

[8] Yongluan Yan, Xinggang Wang, Xiaojie Guo, Jiemin Fang, Wenyu Liu, and Junzhou Huang. Deep multi-instance learning with dynamic pooling. In Jun Zhu and Ichiro Takeuchi, editors, *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pages 662–677. PMLR, 2018.

[9] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based Permutation-Invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753, Long Beach, California, USA, 2019. PMLR.

[10] 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 D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232. Curran Associates, Inc., 2015.

[11] 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.

[12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1263–1272, Sydney, NSW, Australia, 2017. JMLR.org.

[13] 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 V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 991–1001. Curran Associates, Inc., 2017.

[14] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8: 13890, January 2017.

[15] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. arXiv:2003.03123 [cs.LG], March 2020.

[16] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. PotentialNet for molecular property prediction. *ACS Cent. Sci.*, 4(11):1520–1530, November 2018.

[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] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, 9(2):513–530, January 2018.

[19] M A Johnson and G M Maggiora. *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, New York, 1990.

[20] 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.

[21] Martin Simonovsky and Nikos Komodakis. Dynamic Edge-Conditioned filters in convolutional neural networks on graphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Colin Raffel and Daniel P W Ellis. Feed-Forward networks with attention can solve some Long-Term memory problems. arXiv:1512.08756 [cs.LG], 2016.

[23] Greg Landrum. RDKit: Open-source cheminformatics. *https://rdkit.org*, 2006.

[24] 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.

[25] 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 dAlch'e Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019.

[26] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch geometric. arXiv:1903.02428 [cs.LG], March 2019.

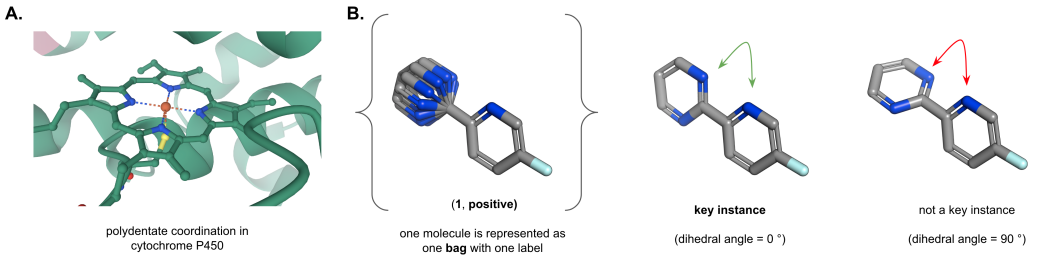

Figure 4: The BIPY MIL Task. **A.** Metal coordination requires polydentate ligands. **B.** We use a bipyridine scaffold to mimic cases where the coplanar nitrogens are able to coordinate a metal center, and use that as our key motif.

Table 2: Detailed Statistics of the BIPY-MIL Dataset (n = 1157).

Property	min.	max.	mean	std.
Sampled Conformers	1	30	13.8	7.4
Heavy Atom Count	10	32	17.8	3.35
Molecular Weight	130.0	507.9	250.5	53.0
Rotatable Bonds	0	1	0.99	0.09

6 Appendix A: Dataset Creation and Details

All molecules were analyzed and processed using the RDKit [23] and conformers generated using OpenEye OMEGA [24].

We constructed a toy dataset inspired by the bidentate coordination of substituted pyridines to a single attachment point (Figure 4). In this task, the relevant binding mode requires the 1,2-diamine motif to adopt a dihedral angle of $0°$. Despite the enthalpic benefit of bidentate bonding, the *s-cis* conformation of bipyridine is typically not the lowest energy conformer. We enumerated a small synthetic library of biaryl ligands varying in substitution pattern and conformational rigidity. For each small-molecule, we used OpenEye OMEGA to generate up to 30 distinct conformations.

As labels, we analyzed each conformational ensemble for a *s-cis* diamine motif, dictated by a dihedral angle of $< 1°$. Each molecule (set of conformers) is assigned a a **0/1** binary label if at least one conformer in the set is able to adopt a coplanar bipyridine. The final dataset consists of 1,157 molecules, 15,959 conformers, with 398 positives and 759 negatives. A random sampling of representative examples are shown in Appendix Figure 5, and additional summary statistics described in Appendix Table 2.

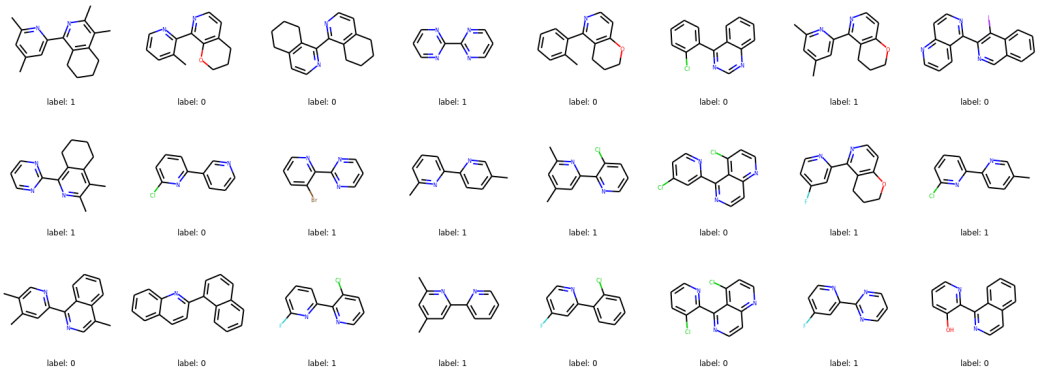

Figure 5: A random sample of molecules used in the BIPY dataset and their associated labels.

Table 3: Embedding-based neural network model architecture

Layer	Description
Graph Featurizer	NNConv, iterations = 3, $h_{dim} = 16$
Set-Aggregator	Feed-Forward Attention $h_{dim} = 128 \rightarrow 1$
Fully-Connected Layer	$h_{dim} = 16 \rightarrow 1$ + Sigmoid

7 Appendix B: Neural Network Architectures and Random Forest Baselines

We implemented all experiments in Python using PyTorch 1.5 [25] and PyTorch Geometric [26].

Of the 1,157 examples above, we randomly split our data into a training set (500), validation set (200), and test set (457). We trained our networks as various subsets of the training set (n = 100, 500) for the data shown in the table above.

We trained all neural networks up to 200 epochs using the Adam optimizer (learning rate = 1×10^{-3} to 1×10^{-4}) and a batch size ranging from 1 – 16, using the early stopping criterion based on the validation set described above. We found that a small batch size typically improved training speeds, but the increased variability occasionally led to stalled training. The model architecture and hidden dimensions are specified in Appendex Table 3. All networks use three layers of graph featurization followed by a single attention aggregation step. The entire network is trained on binary labels ($\mathbf{0}/\mathbf{1}$) for each set of conformers.

All random forest classifiers were trained using scikit-learn, using ensembles of 100 trees. The biaryl ligands were represented using RDKit's Morgan Fingerprints (ECFP4), with the radius set to 2 and a bit vector length of 128.