Improving Molecular Graph Neural Network Explainability with Orthonormalization and Induced Sparsity

Ryan Henderson [1]   **Djork-Arné Clevert** [1]   **Floriane Montanari** [1]

Abstract

Rationalizing which parts of a molecule drive the predictions of a molecular graph convolutional neural network (GCNN) can be difficult. To help, we propose two simple regularization techniques to apply during the training of GCNNs: Batch Representation Orthonormalization (BRO) and Gini regularization. BRO, inspired by molecular orbital theory, encourages graph convolution operations to generate orthonormal node embeddings. Gini regularization is applied to the weights of the output layer and constrains the number of dimensions the model can use to make predictions. We show that Gini and BRO regularization can improve the accuracy of state-of-the-art GCNN attribution methods on artificial benchmark datasets. In a real-world setting, we demonstrate that medicinal chemists significantly prefer explanations extracted from regularized models. While we only study these regularizers in the context of GCNNs, both can be applied to other types of neural networks.

1. Introduction

Graph convolutional neural networks (GCNNs) have shown particular promise in predicting properties of small molecules. These properties can be biological activity against protein targets (Yang et al., 2020; Sakai et al., 2021; Nguyen et al., 2020), more general pharmacokinetics and physicochemical properties (Montanari et al., 2019; Peng et al., 2020; Feinberg et al., 2020) or toxicity (Ma et al., 2020). Using accurate *in silico* predictions of such properties to prioritize the synthesis of compounds can lead to tremendous time and cost savings during drug discovery projects.

[1]Digital Technologies, Bayer AG, Berlin, Germany. Correspondence to: Ryan Henderson <ryan.henderson@bayer.com>, Floriane Montanari <floriane.montanari@bayer.com>.

However, the opaque nature of artificial neural networks has long been a stumbling block for wider adoption. In particular, the difficulty in extracting a straightforward rationalization for a molecular prediction in terms of atomic or fragment contributions limits the adoption of machine learning models among medicinal chemists.

What makes a good rationalization? In the case of explaining molecular properties, the answer is often site attribution. A chemist, when presented with an image of a molecule, may infer properties like solubility and melting point from a few specific atoms and fragments rather than the molecule as a whole. Explaining the predictions in terms of atomic contributions helps build trust in deep learning for molecular properties.

2. Our Contribution

In a GCNN, particular combinations of dimensions of the learned representations can be visually mapped onto the molecule after any convolution. Existing attribution methods for GCNNs make use of that fact. If these dimensions are correlated or if a particular prediction makes use of very many dimensions, the resulting explanations may be confusing and overwhelming for the end user. To mitigate these problems, we introduce two new methods to improve the interpretability of GCNN predictions: Batch Representation Orthonormalization (BRO) and Gini regularization. Both are available in the Pytorch Geometric library (Fey & Lenssen, 2019): https://pytorch-geometric.readthedocs.io/en/latest/modules/nn.html#functional. An overview of our approach is shown in Figure 1.

2.1. Batch Representation Orthonormalization (BRO)

Our work is loosely inspired by linear combination of atomic orbitals (LCAO) theory (Albright et al., 2013). In LCAO theory, a complete set of molecular orbitals ψ is built from the superposition of an orthonormal basis set of atomic electronic wavefunctions: $\psi_j = \sum_i^N c_{ij} \phi_i$ where ϕ_i are the electronic atomic basis functions. The molecular orbitals must also be mutually orthonormal–that is, the coefficients c_{ij} must be normalized so that $\psi_j^* \psi_j = 1$ and $\psi_j^* \psi_{j'} = 0$.
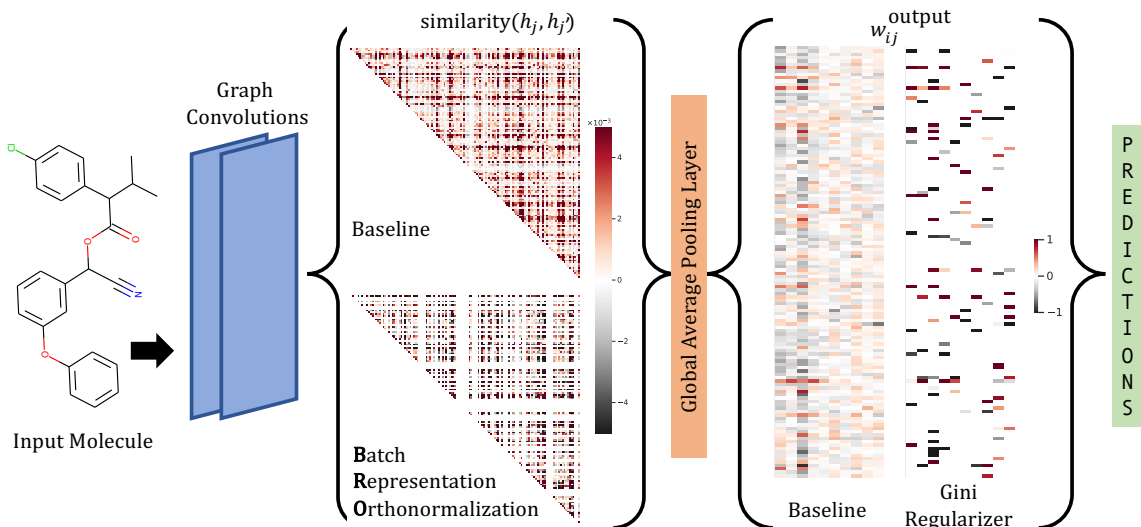
Figure 1. **Batch Representation Orthnormalization and Gini Regularization** From left to right: the input molecule passes through a number of graph convolutional operators, generating new graph embeddings. n embeddings of arbitrary dimension (in this figure: 128) are generated, where n is the number of nodes or atoms in the graph. During training, a regularization loss will be computed on the final embedding, $\mathbf{H}$, according to the BRO regularizer in Equation 1. The contrast between the cosine similarities among all pairs of node representations h_j for a model trained without (Baseline) and with the BRO regularizer is shown: note that in the BRO case, many more overlaps are pushed to zero. The node representations are then aggregated by a global average pooling layer. A final linear layer transforms these aggregations into predictions. The multi-task model depicted in the figure has 10 outputs. The effects of constraining the weights of this layer according to the Gini regularizer (Equation 2) is shown on the right.

Pictorially, one can think of LCAO as building up the molecular orbitals by sequentially making bonding and anti-bonding combinations of symmetry-related atoms until the full molecular orbital set is constructed. GCNNs can also be conceptualized as building up a full molecular picture from atomic sites. Instead of combining symmetry-related sites, GCNNs combine the information from increasingly distant atomic neighbors on the molecule. This correspondence is not only superficial: the final molecular orbitals obtained by this symmetry mixing procedure can be described by the eigenvectors of the adjacency matrix. The contrast of these two principles is depicted in Figure 2.

In the LCAO picture, orthonormality is guaranteed by construction. The representations $\mathbf{H}$ created by a graph convolution (h_{ij} for the ith atom and jth molecular embedding; h_j is a vector that can be mapped directly onto the molecule) have no such guarantee. In this work, we introduce an orthonormality constraint in the training process to partially recover this property. For a molecule with representation $\mathbf{H}$, the molecular regularization loss is defined by:

$$\mathcal{L}_{\text{BRO}}^{\text{mol}} = \frac{\lambda}{2} \, ||\mathbf{H}\mathbf{H}^T - \mathbf{I}||_2 \qquad (1)$$

Where $||\cdot||_2$ indicates the vector 2-norm and $\mathbf{I}$ is the identity matrix. λ is a hyperparameter. Since the forward pass of a GCNN will typically use mini-batching, the regularization loss must be aggregated over all graphs in the batch. Computing so many normalizations is computationally taxing, so in this work we limit BRO application to the outputs of the final graph convolution. We observe that training typically takes four times as long. The effects of this constraint on the node embeddings for a single molecule are shown in the center of Figure 1.

2.2. Gini Regularization

Multitask neural networks (MTNNs) and multitask graph convolutional neural networks (MT-GCNNs) are of special interest for modeling molecular properties. Often, a user is interested in multiple properties of a molecule. Since many physicochemical properties are related, it makes intuitive sense that internal representations could be reused for different tasks.

A typical MT-GCNN architecture will follow the graph convolutional layers with a global graph aggregation layer, which aggregates the node-level outputs into graph-level
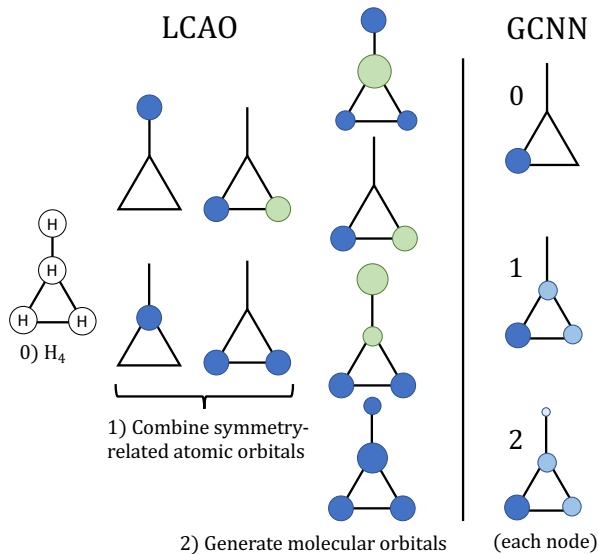
Figure 2. On the left we illustrate the LCAO principle for a hypothetical four-hydrogen molecule (σ-bonding only). The colors represent the sign of the atomic orbital: neighboring atomic sites with different signs are said to be anti-bonding, with correspondingly higher energy. Step 0 shows the topology of the molecule. Step 1 shows combining atomic orbitals which are symmetry-related in bonding and anti-bonding combinations. Step 2 shows a combination of the intermediate orbitals from Step 1 into the final molecular orbitals. The atomic contributions in the final molecular orbitals are sized according to the eigenvectors of the adjacency matrix, and they are ranked according to eigenvalue: most bonding (bottom) to most anti-bonding (top). The right side of the image shows an analogous process for GCNNs, receiving information from increasingly distant (represented by a lighter shade) neighbors for each convolutional step. This happens simultaneously for each node.

outputs for graph property prediction. A sequential, fully-connected neural network then makes graph level predictions based on the graph-level outputs from the global aggregation layer. The simplest example of this architecture, which we will use in this paper, is a global average pooling (GAP) layer followed by a single fully-connected layer.

We wish to constrain the weights of the final fully-connected layer to be sparse. We reason that this should have the two-fold effect of reducing the number of node representations h_j that are relevant to a specific prediction and revealing which representations are shared among tasks. The usual ℓ_1 or ℓ_2 regularization is not appropriate here, as penalizing the magnitude of the weights directly damages the performance of regression metrics.

Instead, we use a regularizer inspired by the Gini coeffi-

cient (Gini, 1912).[1] The Gini coefficient was invented in the context of economics as a straightforward way to compare income or wealth inequality across different countries (Dorfman, 1979). A high Gini coefficient implies a inequal wealth distribution, while a low one indicates a more equitable distribution.

For our model, we want the representations to have high inequality: each prediction should be dominated by as few representations as possible. We select each row w_i of the weights of the final fully-connected layer which is responsible for predicting task i from the n outputs of the GAP layer and compute:

$$\mathcal{L}_{\text{Gini}}^i = \sum_{j}^{n} \sum_{j'}^{n} \frac{|w_{ij} - w_{ij'}|}{(2n^2 - n)\bar{w}_i} \tag{2}$$

where j ranges over all weights in the row. $\bar{w}_i$ is the mean weight value for the row. The Gini coefficient $\mathcal{L}_{\text{Gini}}^i$ ranges from zero to one: zero if all w are equal and one if one w_i is non-zero and the rest zero. Since weights of a linear transform are not necessarily restricted to be non-negative, we will always use $|w|$ rather than w for computing the Gini regularization during training.

We take the mean $g = \frac{1}{n} \sum_i^n \mathcal{L}_{\text{Gini}}^i$ over all tasks. The training loss becomes L/g^m where L is the multi-task regression loss, and m is a hyperparameter to tune the effect of the Gini regularization. The effects of the Gini regularization on a trained model are shown on the right side of Figure 1.

3. Related Work

Our work is inspired by convolutional neural network (CNN) approaches. In particular, the work of Zhang et al. 2018 seeks to align individual convolutional filters to particular concepts, constraining categorical and spatial entropy. Like our approach, this method requires no annotations. Forcing orthonormality in the learned representations is conceptually similar to disentanglement or finding a more interpretable "basis set": this approach is explored for CNNs in Zhou et al. 2019.

Orthogonality or orthormality is often used as a constraint in training neural networks, however this is usually applied to the weights of learnable parameters (Huang et al., 2018). Recurrent neural networks in particular stand to benefit from orthonormalization as a means to address exploding or vanishing gradients (Arjovsky et al., 2016; Jing et al., 2017; Vorontsov et al., 2017). The exact form of the BRO layer was derived from Xie et al. 2017, although here again

---

[1]We first documented this approach in our workshop paper Henderson et al. 2020.

applied to learnable parameters rather than learned representations.

In the context of GCNNs, methods such as Graph Information Bottleneck (Wu et al., 2020), Deep Graph Infomax (Velikovi et al., 2018), and Infograph (Sun et al., 2019) seek to constrain representations according to information-theoretic principles.

Many attribution methods for neural networks have been proposed (Zhou et al., 2016; Selvaraju et al., 2017; Shrikumar et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017) and later adapted to GCNNs (Xie & Lu, 2019; Ying et al., 2019; Pope et al., 2019; Baldassarre & Azizpour, 2019). In this work we focus on Class Activation Maps (CAM) (Zhou et al., 2016). CAM is a simple yet powerful method connecting network outputs and the learned node representations. In GCNNs where the node feature aggregation is done by a GAP layer, CAM corresponds to the scalar product of individual node features by the weights of the output layer:

$$\text{Attribution}(\text{atom}_i) = w_j^\mathsf{T} \cdot h_i \tag{3}$$

where h_i corresponds to the learned feature vector for atom i and w_j corresponds to the row of the output weight matrix of the network corresponding to the jth task. If in the single-task case, $w_j = \mathbf{W}$. We expect our contributions to specifically help with CAM-like attribution methods: the Gini constraint reduces the number of non-zero elements in $\mathbf{W}$ and the BRO orthonormalization ensures that the different dimensions of the learned atom representations $\mathbf{H}$ are independent from each other.

4. Experiments and Results

We wish to show the effect of BRO and Gini regularization on model prediction explainability. To illustrate this, we show the effect of our regularizers on attribution maps on benchmark datasets for which an exact attribution score can be calculated.

Next, we train new models on proprietary assay data to predict physicochemical endpoints. We generate attribution maps from models trained with and without our constraints and survey experts on which attribution they prefer.

In all experiments, no hyperparmeter tuning was performed: the BRO and Gini regularizers are added to the baseline models with the their respective hyperparameters fixed ($\lambda = 0.001$ for BRO regularization and $m = 5.0$ for Gini regularization). We chose values for parameters λ and m qualitatively: we sought the largest values for each that did not degrade evaluation metrics too much (see Section 4.3).

4.1. Attribution Benchmarks

Recently, Sanchez-Lengeling et al. 2020 proposed three artificial tasks to evaluate attribution methods on molecular GCNNs. These simple tasks have an associated ground truth explanation that can be used to score any attribution method's output. For instance, if we trained a model to predict whether a molecule contains a benzene ring or not, a perfect attribution would highlight only the atoms in the benzene ring and no others. We followed their protocol and metrics to benchmark the effect of our proposed regularizers on the attribution performance.

We focus on CAM and CAM-derived attribution methods, including GradCAM (Selvaraju et al., 2017) either on all convolutional layers or only the last one. This is because the authors found CAM to systematically outperform other attribution methods. Since our regularization method has not been implemented yet on edge features, we focus on the benchmark graph architectures that use node features exclusively (graph convolutional network (GCN) and graph attention network (GAT) in Sanchez-Lengeling et al. 2020). We also introduce a variation on CAM that we call TopRep. TopRep applies the CAM Equation 3 using only the weight of the output layer that has the highest magnitude: that is, w_j of Equation 3 retains the entry with the largest absolute value, and all others are set to zero.

4.1.1. DATASET AND NETWORK ARCHITECTURE

Sanchez-Lengeling et al. 2020 present three attribution tasks: "Benzene", "Amine-Ether-Benzene" and "Crippen-LogP." The Benzene task is as described above. The Amine-Ether-Benzene task tests if each of the Amine, Ether, and Benzene fragments are present in the molecule (logical AND). Both of these attribution tasks are scored using AU-ROC (Bradley, 1997). The underlying classification task is also scored with AUROC. The CrippenLogP task is a regression task, with both attribution and regression scored with Pearson correlation. We discuss this task in detail at the end of this section.

For the GCN and GAT models, we use the hyperparemeters given in Sanchez-Lengeling and modify their code only insofar as necessary to add the BRO and Gini regularizers. Eighty trials of each combination are run. Our modifications to the benchmarking code are available at https://github.com/bayer-science-for-a-better-life/graph-attribution.

4.1.2. RESULTS

Figure 3 show the detailed distributions of attribution scores for the Benzene, Amine-Ether-Benzene, and CrippenLogP tasks ($n = 80$ for every model/task combination). For many combinations, BRO or Gini regularization markedly

improve the attribution AUROC score. Often, the combination of BRO and Gini yields better results than either regularizer alone. This effect can be most strongly seen in the Benzene GAT combination (bottom row). This bolsters our reasoning that it takes both sparsity *and* orthogonal resprespresentations to generate a good attribution. While there are some task/model combinations for which none of our constraints improve attribution, no combination's attribution AUROC score suffers significantly from the Gini+BRO regularizers.

TopRep also gives competitive results, even for the baseline configuration. This is surprising in the baseline case, because TopRep is implicitly throwing out a lot of information if the output weight matrix has not been made sparse through, for example, Gini regularization.

Table 1 summarizes our results on all three benchmark tasks, including CrippenLogP.

Table 1. Performance of the attribution methods on the three benchmark tasks (mean for 80 trials per combination). For classification tasks, we report the attribution AUROC score and for regression the Pearson correlation. Best value in each column in bold.

Attr.	Const.	Benzene		AmEthBenz		CrippenLogP	
		GCN	GAT	GCN	GAT	GCN	GAT
Random	None	0.614	0.613	0.500	0.495	-0.086	-0.093
	BRO	0.613	0.613	0.507	0.503	-0.092	-0.093
	Gini	0.614	0.613	0.498	0.500	-0.091	-0.088
	both	0.612	0.615	0.499	0.500	-0.091	-0.091
Grad	None	0.678	0.670	0.475	0.476	0.130	0.071
CAM	BRO	0.673	0.659	0.471	0.500	0.141	0.048
(last)	Gini	0.718	0.697	0.520	0.536	0.176	0.045
	both	0.715	0.716	0.537	0.503	0.182	0.041
Grad	None	0.692	0.628	0.488	0.448	0.181	0.023
CAM	BRO	0.691	0.632	0.482	0.477	0.194	0.003
(all)	Gini	0.733	0.687	0.523	0.505	0.192	-0.002
	both	0.732	0.698	0.530	0.474	0.201	-0.022
CAM	None	0.989	0.984	0.621	0.691	0.268	0.221
	BRO	0.989	0.977	0.619	0.701	0.274	0.219
	Gini	0.990	**0.989**	**0.653**	**0.703**	0.278	0.208
	both	0.990	**0.989**	0.646	0.699	**0.279**	**0.248**
TopRep	None	**0.991**	0.979	0.556	0.595	0.205	0.153
	BRO	0.982	0.967	0.486	0.619	0.214	0.131
	Gini	0.984	0.988	0.510	0.558	0.208	0.148
	both	**0.991**	0.987	0.555	0.593	0.204	0.155

We notice that the CAM-related attribution methods work poorly for the regression task CrippenLogP. This was already observed in Sanchez-Lengeling et al. 2020, where only the integrated gradients approach (Sundararajan et al., 2017) seemed to perform slightly better than other attribution methods.

We interpret this finding in the construction of the benchmark task itself. The authors used a dataset with experimentally determined solubility (Delaney, 2004) and consider as ground truth explanation the atom contribution for the water-octanol partition coefficient as calculated by the Crippen LogP empirical model (Wildman & Crippen, 1999). Although logP and solubility are correlated, we assume that
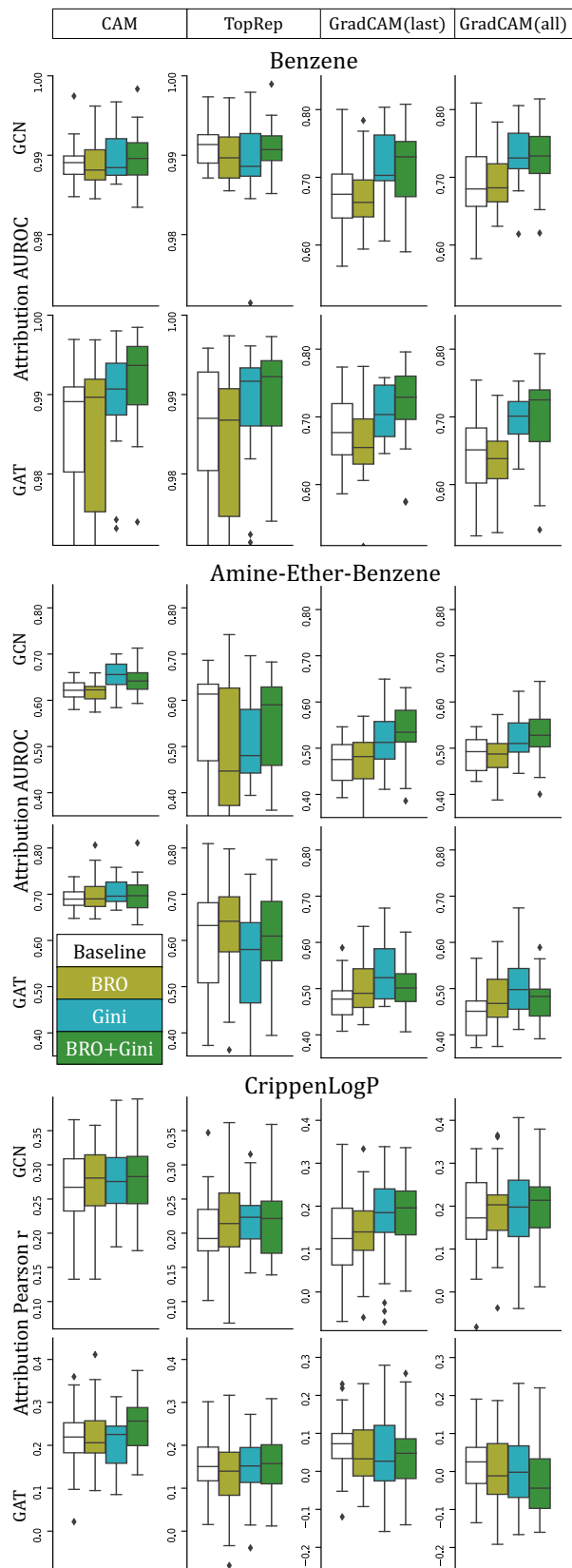

Figure 3. Boxplots of attribution AUROC scores for all tasks. Both GCN and GAT models were considered.

logP is not always a good explanation for a more complex endpoint like solubility, and therefore it should not be surprising if the built model's explanations do not align very well with the logP atom contributions. A better task might be to build a model for logP and use the Crippen LogP attributions as ground truth explanations. This experiment was not performed here to stay in the framework of the benchmark.

4.2. Physicochemical endpoints

The results on the constructed benchmark dataset in the previous section encouraged us to evaluate our methods against expert opinion: a much costlier undertaking.

It is impossible to reliably and quickly calculate properties like solubility or binding affinity *ab initio* from the chemical structure of a compound, let alone attribute the prediction to specific sites or functional groups. When screening drug candidates, a medicinal chemist may synthesize many similar molecules with the desired activity while simultaneously trying to minimize or maximize some other property.

Thus, being able to guess which modifications may alter these properties is extremely helpful. Here, we seek to mimic the chemists' intuition, and see if the BRO and Gini constraints may produce models that generate attributions that more closely match their instincts.

4.2.1. DATASET

The data used for building the multitask GCNN model is a dataset of measured physicochemical properties for small molecules from 10 different assays. Extensive discussion of the preparation and characteristics of the dataset can be found in our previous work Montanari et al. 2019. There are a total of 537,443 compounds with assays covering properties such as solubility, lipophilicity (logD at two different pHs), melting point, human serum albumin binding and membrane affinity; 79% of the compounds are measured for only one endpoint, 11% for two, 9% for 3, and 1% four or more.

All ten endpoints along with their frequencies and the correlation between pairs is shown in Figure 4. We also show the cosine similarities of the rows of the trained output weights corresponding to each endpoint. The similarities of both the baseline and Gini-constrained model mimic the underlying correlation between the measurements, but the Gini-constrained version is sparser. This may imply it has learned a more specific relationship between the endpoints than merely the data distribution.

Seven cross-validation folds were generated using k-means clustering on the ECFC6 fingerprints (Rogers & Hahn, 2010).
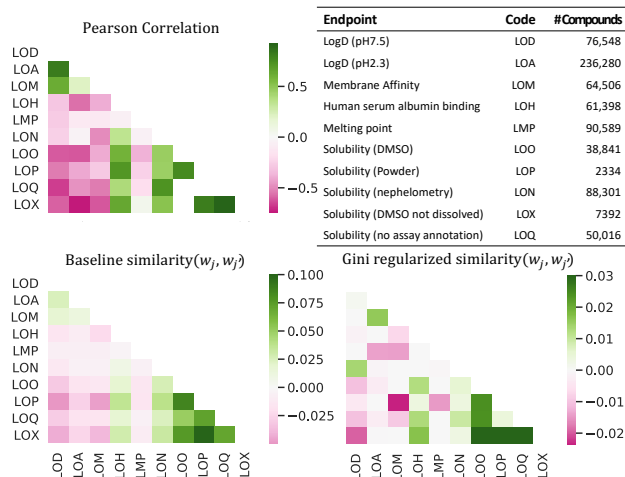

Figure 4. The first row shows the Pearson correlation between each pair of endpoints, along with a table giving a description and count of each. The second row shows the cosine similarity of the rows of an output weight matrix for an unconstrained and Gini-constrained model.

4.2.2. NETWORK ARCHITECTURE

The model, initially built in Tensorflow with the DeepChem library (Ramsundar et al., 2019) was converted to Pytorch Geometric for easier access and customization (Fey & Lenssen, 2019). We tried as much as possible to follow the DeepChem graph convolution mechanism which is an implementation of the Duvenaud et al. 2015 algorithm.

Input node features are computed with DeepChem and consist of 75 atomic properties: number of radical electrons, whether the atom is aromatic, and one-hot encoding on each of the atom type, degree, formal charge, and hybridization. Two convolutional layers are then applied, then a GAP layer going from individual node features to a global graph encoding. A hyperbolic tangent function is applied to the graph encoding before the final linear layer.

Our models were trained using the ADAM optimizer (Kingma & Ba, 2017) for forty epochs with exponential learning rate decay with a base of 0.97 decaying every 1000 steps. No hyperparemeter search is performed: models with either or both of BRO and Gini regularization use the same hyperparameters as the baseline models. Model code is available in the Supplementary Material.

4.2.3. SURVEY OF MEDICINAL CHEMISTS

To test our methods in a realistic setting, we prepared a survey to gather opinions from human experts. We collected public compounds with reported measured solubility (Sorkun et al., 2019), logD (Alelyunas et al., 2010; Low

et al., 2016) or melting point (Williams et al., 2015). These compounds were chosen for their problematic properties: they were either very insoluble, having a high logD, or a low melting point.

We then used the multitask GCN built on physicochemical data (in both the constrained and non constrained version) to predict the solubility, logD or melting point of those molecules. Because the data is collected from public sources, potentially many different assays were used to measure the different endpoints of interest. Therefore, an exact match of the predictive model (built on proprietary established assay data) with the publicly reported values is not expected. Since we focused on problematic molecules, we compared predictions and reported experimental values qualitatively only. For example, insoluble molecules with a predicted solubility over 30 mg/L were discarded. In general, molecules with grossly inaccurate predictions by the models were excluded from the survey. In total, 10 molecules were kept for solubility, 8 for logD and 6 for melting point. Attribution methods were then applied on each predicted molecule: either CAM (our baseline, using the unconstrained model), $CAM_{BRO+Gini}$ (CAM applied to the Gini-sparsified and BRO-disentangled model), $TopRep_{BRO+Gini}$ (CAM using only the top weight from the constrained model output), or a random attribution map (any node representation generated by the constrained model, selected randomly).

The users always had the possibility to skip a question or select a negative answer ("No answer convinces me"). The attribution maps were plotted on the molecule structure using the RDKit's SimilarityMap functionality (Landrum, 2006). An example question as shown to the medicinal chemists is shown in Figure 5. 15 medicinal chemists answered the 24 questions and the results are shown in Figure 6. The participants were all employed at Bayer on two different sites in Germany, and were at different stages of their career. They had no previous explicit experience with such specific molecules or tasks although the molecules are known drugs and the tasks are tasks they reason about on a daily basis.

$CAM_{BRO+Gini}$ and $TopRep_{BRO+Gini}$ are the most voted attribution methods among the 24 questions, with $CAM_{BRO+Gini}$ being favored for logD and $TopRep_{BRO+Gini}$ for solubility and melting point. Across all endpoints, they are both picked more often than would be expected at random (binomial test, p-value=0.01 and p-value < 0.001 respectively). Surprisingly, Baseline CAM cannot be distinguished from the Random attribution result (binomial test, p-value=0.07).

This real-life experiment shows that, with a predictive model that is in practice useful for medicinal chemists, the regularization constraints (Gini and BRO) lead to significantly preferred attribution maps. We also note that the logD end-
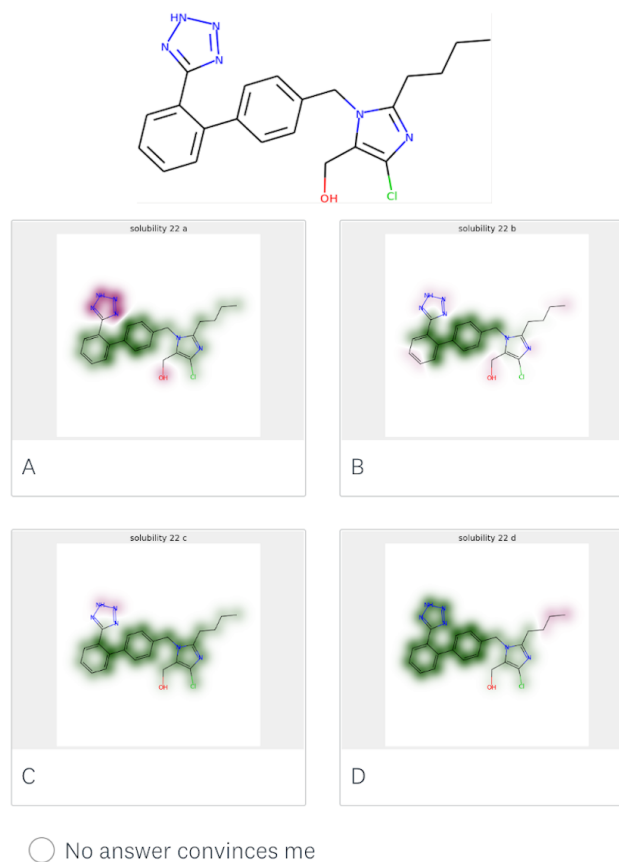

Figure 5. An example question from the survey of medicinal chemists: "Which atoms drive the low solubility of the molecule? (Green: lower solubility / Magenta: Higher solubility." This shows one of the query molecules for solubility taken from (Sorkun et al., 2019) with various site attributions for solubility (LOO task). Answers were randomly ordered for each question, blind to both us and the medicinal chemists. In this case we have a) $CAM_{BRO+Gini}$ b) random c) CAM (baseline) d) $TopRep_{BRO+Gini}$.

point is the easiest for chemists to rationalize and led to the most imbalanced votes in favor of $CAM_{BRO+Gini}$ (coefficient of variation $\sigma/\mu = 0.95$). Solubility and melting point are less straightforward, and, accordingly, chemists tend to disagree more in their voting ($\sigma/\mu = 0.86$ and 0.59 respectively).

All survey questions with attribution maps, including aggregated responses and answer key, are available in the Supplementary Material.

4.3. Model Performance

The Gini and BRO regularizations have a slight negative impact on the classification or regression metrics of the models. The impact on the AUROC scores for the "Benzene"
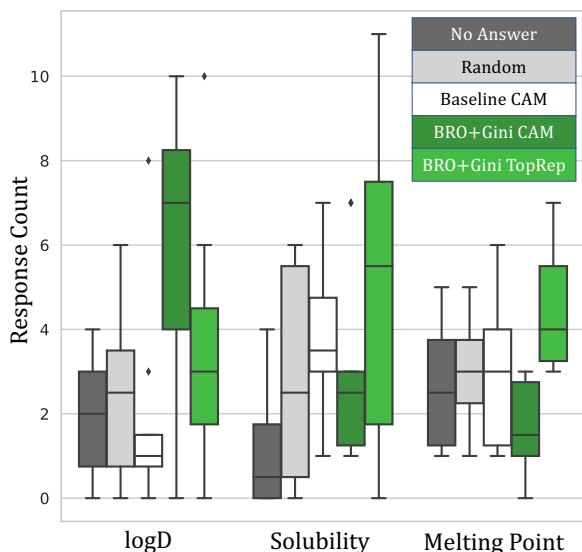
Figure 6. Results of survey of medicinal chemists. Not all participants answered all questions. The plot shows the distribution of responses for each option for each endpoint. For example: for logD questions, a median of 7 respondents (ranging from 0 to 10) picked CAM$_{BRO+Gini}$, with 50% of the logD questions receiving between 4 to 8 votes for CAM$_{BRO+Gini}$.

and "Amine-Ether-Benzene" classification tasks and the Pearson correlation for the "CrippenLogP" regression task described in Section 4.1 are illustrated in Figure 7.

As with the Attribution AUROC, it is nearly impossible to improve the classification AUROC of the Benzene task: all models are extremely close to perfect classification. Even in this regime, it is clear that the BRO layer negatively effects performance. In the Amine-Ether-Benzene task, the differences are not significant. This may be because both BRO and Gini regularization were designed with multitask models in mind, and the Amine-Ether-Benzene is really a multitask problem in disguise. Finally, adding the BRO and Gini constraints to the CrippenLogP task does not drastically alter model performance, despite doing little to improve interpretability.

The various endpoints of the multitask models trained on physicochemical data described in Section 4.2 are affected unevenly. The cross-validation R^2 scores are shown in Table 2. The solubility endpoints LOP and LOX are strongly affected. These are the two least represented assays in the dataset, with 2,334 and 7,392 out of 537,443 compounds, respectively.

It is possible that hyperparameter tuning could recover these performance losses. Anecdotally, we note that adding additional graph convolutional layers seems to help. This
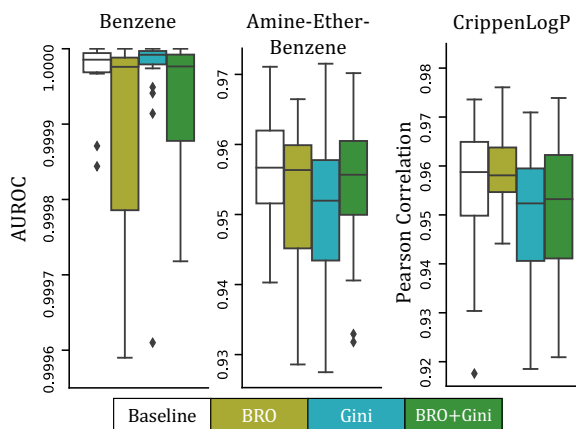

Figure 7. Effect of regularization techniques on the evaluation metric for the Benzene, Amine-Ether-Benzene, and CrippenLogP tasks.

Table 2. Effect of BRO and Gini regularizers on predictive performance for all physicochemical endpoints. Mean of scaffold-split cross-validation R^2 score and standard deviation in parenthesis.

	Baseline	BRO	Gini	BRO + Gini
LOD	0.85 (0.03)	0.82 (0.03)	0.83 (0.03)	0.81 (0.04)
LOA	0.82 (0.04)	0.80 (0.04)	0.82 (0.04)	0.79 (0.04)
LOM	0.56 (0.13)	0.53 (0.13)	0.53 (0.12)	0.51 (0.13)
LOH	0.50 (0.07)	0.48 (0.09)	0.47 (0.09)	0.44 (0.10)
LMP	0.44 (0.08)	0.45 (0.08)	0.44 (0.08)	0.43 (0.09)
LON	0.51 (0.09)	0.48 (0.09)	0.49 (0.08)	0.45 (0.10)
LOO	0.50 (0.20)	0.46 (0.24)	0.47 (0.20)	0.45 (0.23)
LOP	0.47 (0.15)	0.44 (0.19)	0.30 (0.40)	0.34 (0.34)
LOQ	0.56 (0.11)	0.52 (0.14)	0.54 (0.11)	0.50 (0.15)
LOX	0.52 (0.06)	0.48 (0.07)	0.44 (0.09)	0.43 (0.09)

can, however, dilute the explanations which become further "de-localized" with each convolution.

5. Conclusion

We describe and implement two new regularization techniques: Batch Representation Orthonormalization and Gini regularization. We show qualitatively that these techniques disentangle node representations and force models to make predictions using fewer of them.

We demonstrate that models trained with these constraints generate better site attributions on a benchmark dataset for many attribution methods. Further, we show that for models trained on assay data of interest to medicinal chemists, human experts significantly prefer attribution maps generated from models trained with the constraints introduced in this paper.

While we give an exact definition for the BRO constraint in Equation 1, the choice of the vector 2-norm is arbitrary. The

novelty of our approach is normalizing the node representations per graph, and it is likely that other norms may yield better results in other circumstances. Similarly, while the Gini coefficient has desirable properties for this application, other measures of dispersion could also be explored.

Finally, our methods could be extended to graph architectures which exploit edge features. Future work may explore the separate or joint orthonormalization of node and edge features.

Acknowledgements

References

Albright, T. A., Burdett, J. K., and Whangbo, M.-H. *Orbital Interactions in Chemistry: Albright/Orbital Interactions in Chemistry*. John Wiley & Sons, Inc., Hoboken, NJ, USA, April 2013. ISBN 978-1-118-55840-9 978-0-471-08039-8. doi: 10.1002/9781118558409. URL http://doi.wiley.com/10.1002/9781118558409.

Alelyunas, Y. W., Pelosi-Kilby, L., Turcotte, P., Kary, M.-B., and Spreen, R. C. A high throughput dried DMSO LogD lipophilicity measurement based on 96-well shake-flask and atmospheric pressure photoionization mass spectrometry detection. *Journal of Chromatography A*, 1217(12):1950–1955, March 2010. ISSN 0021-9673. doi: 10.1016/j.chroma.2010.01.071. URL http://www.sciencedirect.com/science/article/pii/S0021967310001299.

Arjovsky, M., Shah, A., and Bengio, Y. Unitary Evolution Recurrent Neural Networks. In *International Conference on Machine Learning*, pp. 1120–1128. PMLR, June 2016. URL http://proceedings.mlr.press/v48/arjovsky16.html. ISSN: 1938-7228.

Baldassarre, F. and Azizpour, H. Explainability Techniques for Graph Convolutional Networks. *arXiv:1905.13686 [cs, stat]*, May 2019. URL http://arxiv.org/abs/1905.13686. arXiv: 1905.13686.

Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. ISSN 0031-3203. doi: 10.1016/S0031-3203(96)00142-2. URL http://www.sciencedirect.com/science/article/pii/S0031320396001422.

Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, May 2004. ISSN 0095-2338. doi: 10.1021/ci034243x. URL https://doi.org/10.1021/ci034243x. Publisher: American Chemical Society.

Dorfman, R. A Formula for the Gini Coefficient. *The Review of Economics and Statistics*, 61(1):146–149, 1979. ISSN 0034-6535. doi: 10.2307/1924845. URL https://www.jstor.org/stable/1924845. Publisher: The MIT Press.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*, 28:2224–2232, 2015. URL https://papers.nips.cc/paper/2015/hash/f9be311e65d81a9ad8150a60844bb94c-Abstract.html.

Feinberg, E. N., Joshi, E., Pande, V. S., and Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *Journal of Medicinal Chemistry*, 63(16):8835–8848, August 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b02187. URL https://doi.org/10.1021/acs.jmedchem.9b02187. Publisher: American Chemical Society.

Fey, M. and Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *arXiv:1903.02428 [cs, stat]*, April 2019. URL http://arxiv.org/abs/1903.02428. arXiv: 1903.02428.

Gini, C. *Variabilit e Mutuabilit. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna, 1912.

Henderson, R., Clevert, D.-A., and Montanari, F. Gini in a Bottleneck: Sparse Molecular Representations for Graph Convolutional Neural Networks. In *Machine Learning for Molecules Workshop at NeurIPS 2020*, November 2020. URL http://arxiv.org/abs/2010.04535. arXiv: 2010.04535.

Huang, L., Liu, X., Lang, B., Yu, A. W., Wang, Y., and Li, B. Orthogonal Weight Normalization: Solution to Optimization over Multiple Dependent Stiefel Manifolds in Deep Neural Networks. In *AAAI Conference on Artificial Intelligence*, pp. 8, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17072.

Jing, L., Shen, Y., Dubcek, T., Peurifoy, J., Skirlo, S., LeCun, Y., Tegmark, M., and Soljai, M. Tunable Efficient Unitary Neural Networks (EUNN) and their

application to RNNs. In *International Conference on Machine Learning*, pp. 1733–1741. PMLR, July 2017. URL http://proceedings.mlr.press/v70/jing17a.html. ISSN: 2640-3498.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Landrum, G. RDKit: Open-source cheminformatics, 2006. URL http://www.rdkit.org/.

Low, Y. W. I., Blasco, F., and Vachaspati, P. Optimised method to estimate octanol water distribution coefficient (logD) in a high throughput format. *European Journal of Pharmaceutical Sciences*, 92:110–116, September 2016. ISSN 0928-0987. doi: 10.1016/j.ejps.2016.06.024. URL http://www.sciencedirect.com/science/article/pii/S092809871630238X.

Ma, H., An, W., Wang, Y., Sun, H., Huang, R., and Huang, J. Deep Graph Learning with Property Augmentation for Predicting Drug-Induced Liver Injury. *Chemical Research in Toxicology*, December 2020. ISSN 0893-228X. doi: 10.1021/acs.chemrestox.0c00322. URL https://doi.org/10.1021/acs.chemrestox.0c00322. Publisher: American Chemical Society.

Montanari, F., Kuhnke, L., Ter Laak, A., and Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules*, 25(1):44, December 2019. ISSN 1420-3049. doi: 10.3390/molecules25010044. URL https://www.mdpi.com/1420-3049/25/1/44.

Nguyen, C. Q., Kreatsoulas, C., and Branson, K. M. Meta-Learning GNN Initializations for Low-Resource Molecular Property Prediction. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, June 2020. URL https://openreview.net/forum?id=MQ_t7LRvsW.

Peng, Y., Lin, Y., Jing, X.-Y., Zhang, H., Huang, Y., and Luo, G. S. Enhanced Graph Isomorphism Network for Molecular ADMET Properties Prediction. *IEEE Access*, 8:168344–168360, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3022850.

Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. Explainability Methods for Graph Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10772–10781, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Pope_Explainability_Methods_for_Graph_Convolutional_Neural_Networks_CVPR_2019_paper.html.

Ramsundar, B., Eastman, P., Walters, P., and Pande, V. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery and more*. O'Reilly Media, Sebastopol, CA, first edition edition, 2019. ISBN 978-1-4920-3983-9. OCLC: on1051083869.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-960X. doi: 10.1021/ci100050t.

Sakai, M., Nagayasu, K., Shibui, N., Andoh, C., Takayama, K., Shirakawa, H., and Kaneko, S. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Scientific Reports*, 11(1): 525, January 2021. ISSN 2045-2322. doi: 10.1038/s41598-020-80113-7. URL https://www.nature.com/articles/s41598-020-80113-7. Number: 1 Publisher: Nature Publishing Group.

Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P., Qian, W. W., McCloskey, K., Colwell, L., and Wiltschko, A. Evaluating Attribution for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://proceedings.neurips.cc//paper_files/paper/2020/hash/417fbbf2e9d5a28a855a11894b2e795a-Abstract.html.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, October 2017. doi: 10.1109/ICCV.2017.74. ISSN: 2380-7504.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, July 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html. ISSN: 2640-3498.

Smilkov, D., Thorat, N., Kim, B., Vigas, F., and Wattenberg, M. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825 [cs, stat]*, June 2017. URL http://arxiv.org/abs/1706.03825. arXiv: 1706.03825.

Sorkun, M. C., Khetan, A., and Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6(1): 143, August 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0151-1. URL https://www.nature.

com/articles/s41597-019-0151-1. Number: 1 Publisher: Nature Publishing Group.

Sun, F.-Y., Hoffman, J., Verma, V., and Tang, J. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *International Conference on Learning Representations*, September 2019. URL https://openreview.net/forum?id=r1lfF2NYvH.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328, Sydney, NSW, Australia, August 2017. JMLR.org.

Velikovi, P., Fedus, W., Hamilton, W. L., Li, P., Bengio, Y., and Hjelm, R. D. Deep Graph Infomax. In *International Conference on Learning Representations*, September 2018. URL https://openreview.net/forum?id=rklz9iAcKQ.

Vorontsov, E., Trabelsi, C., Kadoury, S., and Pal, C. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning*, pp. 3570–3578. PMLR, July 2017. URL http://proceedings.mlr.press/v70/vorontsov17a.html. ISSN: 2640-3498.

Wildman, S. A. and Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39 (5):868–873, September 1999. ISSN 0095-2338. doi: 10.1021/ci990307l. URL https://doi.org/10.1021/ci990307l. Publisher: American Chemical Society.

Williams, A., Lowe, D., and Tetko, I. Melting Point and Pyrolysis Point Data for Tens of Thousands of Chemicals. 2015. doi: 10.6084/m9.figshare.2007426.v1. URL https://figshare.com/articles/dataset/Melting_Point_and_Pyrolysis_Point_Data_for_Tens_of_Thousands_of_Chemicals/2007426.

Wu, T., Ren, H., Li, P., and Leskovec, J. Graph Information Bottleneck. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://proceedings.neurips.cc//paper_files/paper/2020/hash/ebc2aa04e75e3caabda543a1317160c0-Abstract.html.

Xie, D., Xiong, J., and Pu, S. All You Need is Beyond a Good Init: Exploring Better Solution for Training Extremely Deep Convolutional Neural Networks with Orthonormality and Modulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5075–5084, July 2017. doi: 10.1109/CVPR.2017.539. ISSN: 1063-6919.

Xie, S. and Lu, M. Interpreting and Understanding Graph Convolutional Neural Network using Gradient-based Attribution Method. *arXiv:1903.03768 [cs]*, April 2019. URL http://arxiv.org/abs/1903.03768. arXiv: 1903.03768.

Yang, S., Lee, K. H., and Ryu, S. A comprehensive study on the prediction reliability of graph neural networks for virtual screening. *arXiv:2003.07611 [cs, stat]*, March 2020. URL http://arxiv.org/abs/2003.07611. arXiv: 2003.07611.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in Neural Information Processing Systems*, 32:9244–9255, 2019. URL https://papers.nips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.

Zhang, Q., Wu, Y. N., and Zhu, S. Interpretable Convolutional Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, June 2018. doi: 10.1109/CVPR.2018.00920. ISSN: 2575-7075.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, June 2016. doi: 10.1109/CVPR.2016.319. ISSN: 1063-6919.

Zhou, B., Bau, D., Oliva, A., and Torralba, A. Interpreting Deep Visual Representations via Network Dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2131–2145, September 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2858759.