

# Molecular Graph Convolutions: Moving Beyond Fingerprints

Steven Kearnes

Stanford University

kearnes@stanford.edu

Kevin McCloskey

Google Inc.

mccloskey@google.com

Marc Berndl

Google Inc.

marcberndl@google.com

Vijay Pande

Stanford University

pande@stanford.edu

Patrick Riley

Google Inc.

pfr@google.com

## Abstract

Molecular “fingerprints” encoding structural information are the workhorse of cheminformatics and machine learning in drug discovery applications. However, fingerprint representations necessarily emphasize particular aspects of the molecular structure while ignoring others, rather than allowing the model to make data-driven decisions. We describe molecular *graph convolutions*, a machine learning architecture for learning from undirected graphs, specifically small molecules. Graph convolutions use a simple encoding of the molecular graph—atoms, bonds, distances, etc.—which allows the model to take greater advantage of information in the graph structure. Although graph convolutions do not outperform all fingerprint-based methods, they (along with other graph-based methods) represent a new paradigm in ligand-based virtual screening with exciting opportunities for future improvement.

## 1 Introduction

Computer-aided drug design requires representations of molecules that can be related to biological activity or other experimental endpoints. These representations encode structural features, physical properties, or activity in other assays [Todeschini and Consonni, 2009; Petrone et al., 2012]. The recent advent of “deep learning” has enabled the use of very raw representations that are less application-specific when building machine learning models [LeCun et al., 2015]. For instance, image recognition

models that were once based on complex features extracted from images are now trained exclusively on the pixels themselves—deep architectures can “learn” appropriate representations for input data. Consequently, deep learning systems for drug screening or design should benefit from molecular representations that are as complete and general as possible rather than relying on application-specific features or encodings.

First-year chemistry students quickly become familiar with a common representation for small molecules: the molecular graph. Figure 1 gives an example of the molecular graph for ibuprofen, an over-the-counter non-steroidal anti-inflammatory drug. The atoms and bonds between atoms form the nodes and edges, respectively, of the graph. Both atoms and bonds have associated properties, such as atom type and bond order. Although the basic molecular graph representation does not capture the quantum mechanical structure of molecules or necessarily express all of the information that it might suggest to an expert medicinal chemist, its ubiquity in academia and industry makes it a desirable starting point for machine learning on chemical information.

Here we describe molecular *graph convolutions*, a deep learning system using a representation of small molecules as undirected graphs of atoms. Graph convolutions extract meaningful features from simple descriptions of the graph structure—atom and bond properties, and graph distances—to form molecule-level representations that can be used in place of fingerprint descriptors in conventional machine learning applications.

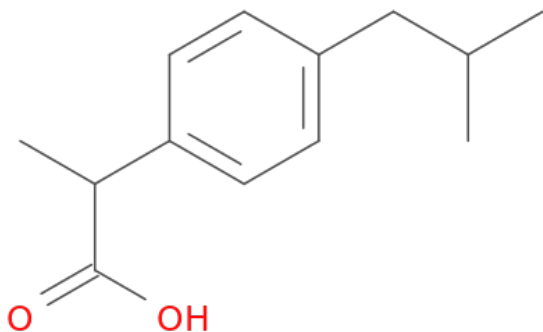


FIGURE 1: Molecular graph for ibuprofen. Unmarked vertices represent carbon atoms, and bond order is indicated by the number of lines used for each edge.

## 2 Related Work

The history of molecular representation is extremely diverse [Todeschini and Consonni, 2009] and a full review is outside the scope of this report. Below we describe examples from several major branches of the field to provide context for our work. Additionally, we review several recent examples of graph-centric approaches in cheminformatics.

Much of cheminformatics is based on so-called “2D” molecular descriptors that attempt to capture relevant structural features derived from the molecular graph. In general, 2D features are computationally inexpensive and easy to interpret and visualize. One of the most common representations in this class is extended-connectivity fingerprints (ECFP), also referred to as circular or Morgan fingerprints [Rogers and Hahn, 2010]. Starting at each heavy atom, a “bag of fragments” is constructed by iteratively expanding outward along bonds (usually the algorithm is terminated after 2–3 steps). Each unique fragment is assigned an integer identifier, which is often hashed into a fixed-length representation or “fingerprint”. Additional descriptors in this class include decompositions of the molecular graph into subtrees or fixed-length paths [OpenEye GraphSim Toolkit], as well as atom pair (AP) descriptors that encode atom types and graph distances (number of intervening bonds) for all pairs of atoms in a molecule [Carhart et al., 1985].

Many representations encode 3D information, with special emphasis on molecular shape and electrostatics as primary drivers of interactions in real-world systems. For example, rapid overlay of chemical structures (ROCS) aligns pairs of pre-generated conformers and calculates shape and chemical (“color”) similarity using Gaussian representations of atoms and

color features defined by a simple force field [Hawkins et al., 2007]. ROCS can also be used to generate alignments for calculation of electrostatic field similarity [Muchmore et al., 2006]. Ultrafast shape recognition (USR) calculates alignment-free 3D similarity by comparing distributions of intramolecular distances [Ballester and Richards, 2007].

The Merck Molecular Activity Challenge [Dahl, 2012] catalyzed interest in deep neural networks trained on fingerprints and other molecular descriptors. In particular, multitask neural networks have produced consistent gains relative to baseline models such as random forest and logistic regression [Dahl et al., 2014; Ma et al., 2015; Mayr et al., 2015; Ramsundar et al., 2015].

Other approaches from both the cheminformatics and the machine learning community directly operate on graphs in a way similar to how we do here. The “molecular graph networks” of Merkwirth and Lengauer [2005] iteratively update a state variable on each atom with learned weights specific to each atom type–bond type pair. Similarly, Micheli [2009] presents a more general formulation of the same concept of iterated local information transfer across edges and applies this method to predicting the boiling point of alkanes.

Scarselli et al. [2009] similarly defines a local operation on the graph. They demonstrate that a fixed point across all the local functions can be found and calculate fixed point solutions for graph nodes as part of each training step. In another vein, Lusci et al. [2013] convert undirected molecular graphs to a directed recursive neural net and take an ensemble over multiple conversions.

Recently, Duvenaud et al. [2015] presented an architecture trying to accomplish many of the same goals as this work. The architecture was based on generalizing the fingerprint computation such that it can be learned via backpropagation. They demonstrate that this architecture improves predictions of solubility and photovoltaic efficiency but not binding affinity.

Bruna et al. [2013] introduce convolutional deep networks on spectral representations of graphs. However, these methods apply when the graph structure is fixed across examples and only the labeling/features on individual nodes varies.

Convolutional networks on non-Euclidean manifolds were described by Masci et al. [2015]. The problem addressed was to describe the shape of the manifold (such as the surface of a human being) in

such a way that the shape descriptor of a particular point was invariant to perturbations such as movement and deformation. They also describe an approach for combining local shape descriptors into a global descriptor and demonstrate its use in a shape classification task.

## 3 Methods

### 3.1 Deep neural networks

Neural networks are directed graphs of simulated "neurons". Each neuron has a set of inputs and computes an output. The neurons in early neural nets were inspired by biological neurons and computed an affine combination of the inputs followed by a non-linear activation function. Mathematically, if the inputs are  $x_1 \dots x_N$ , weights  $w_1 \dots w_N$  and bias  $b$  are parameters, and  $f$  is the activation function, the output is

$$f(b + \sum_i w_i x_i) \quad (1)$$

Popular activation functions include the sigmoid function ( $f(z) = \frac{1}{1+e^{-z}}$ ) and rectified linear unit (ReLU) ( $f(z) = 0$  if  $z \leq 0$  else  $z$ ).

Any mostly differentiable function can be used as the unit of computation for a neuron and in recent years, many other functions have appeared in published networks, including max and sum.

*Convolution* in neural networks refers to using the same parameters (such as the  $w_i$  in Equation 1) for different neurons that are attached to different parts of the input (or previous neurons). In this way, the same operation is computed for many different subsets of the input.

At the "top" of the neural network you have node(s) whose output is the value you are trying to predict (e.g. the probability that this molecule binds to a target or the binding affinity). Many output nodes for different tasks can be added and this is commonly done [Ma et al., 2015; Ramsundar et al., 2015]. In this way, different output tasks can share the computation and model parameters in lower parts of the network before using their own parameters for the final output steps.

The *architecture* of a neural network refers to the choice of the number of neurons, the type of computation each one does (including what learnable parameters they have), which parameters are shared across neurons, and how the output of one neuron is connected to the input of another.

In order to train the network, you first have to choose a *loss function* describing the penalty for the network producing a set of outputs which differ from the outputs in the training example. For example, for regression problems, the L2 distance between the predicted and actual values is commonly used. The objective of training is then to find a set of parameters for the network that minimizes the loss function. Training is done with the well known technique of back-propagation [Rumelhart et al., 1986] and stochastic gradient descent.

### 3.2 Desired invariants of a model

A primary goal of designing a deep learning architecture is to restrict the set of functions that can be learned to ones that match the desired properties from the domain. For example, in image understanding, spatial convolutions force the model to learn functions that are invariant to translation.

For a deep learning architecture taking a molecular graph as input, some arbitrary choice must be made for the order that the various atoms and bonds are presented to the model. Since that choice is arbitrary, we want:

**Property 1** (Order invariance). *The output of the model should be invariant to the order that the atom and bond information is encoded in the input.*

Note that many current procedures for fingerprinting molecules achieve Property 1. We will now gradually construct an architecture which achieves Property 1 while making available a richer space of learnable parameters.

The first basic unit of representation is an *atom layer* which contains an  $n$ -dimensional vector associated with each atom. Therefore the atom layer is a 2 dimensional matrix indexed first by atom. Part of the original input will be encoded in such an atom layer and the details of how we construct the original input vector are discussed in Section 3.5. The next basic unit of representation is a *pair layer* which contains an  $n$ -dimensional vector associated with each pair of atoms. Therefore, the pair layer is a 3 dimensional matrix where the first two dimensions are indexed by atom. Note that the pair input can contain information not just about edges but about any arbitrary pair. Notably, we will encode the graph distance (length of shortest path from one atom to the other) in the input pair layer. The order of the atom indexing for the atom and pair layer inputs must be the same.

We will describe various operations to compute new atom and pair layers with learnable parameters at every step. Notationally, let  $A^x$  be the value of a particular atom layer  $x$  and  $P^y$  be the value of a particular pair layer  $y$ . The inputs that produce those values should be clear from the context.  $A_a^x$  refers to the value of atom  $a$  in atom layer  $x$  and  $P_{(a,b)}^y$  refers to the value of pair  $(a,b)$  in pair layer  $y$ .

In order to achieve Property 1 for the overall architecture, we need a different type of invariance for each atom and pair layer.

**Property 2** (Atom and pair permutation invariance). *The values of an atom layer and pair permute with the original input layer order. More precisely, if the inputs are permuted with a permutation operator  $Q$ , then for all layers  $x, y$ ,  $A^x$  and  $P^y$  are permuted with operator  $Q$  as well.*

In other words, Property 2 means that from a single atom’s (or pair’s) perspective, its value in every layer is invariant to the order of the other atoms (or pairs).

Since molecules are undirected graphs, we will also maintain the following:

**Property 3** (Pair order invariance). *For all pair layers  $y$ ,  $P_{(a,b)}^y = P_{(b,a)}^y$*

Property 3 is easy to achieve at the input layer and the operations below will maintain this.

Properties 2 and 3 make it easy to construct a molecule-level representation from an atom or pair such that the molecule-level representation achieves Property 1 (see Section 3.4).

### 3.3 Invariant-preserving operations

We now define a series of operations that maintain the above properties.

Throughout,  $f$  represents an arbitrary function and  $g$  represents an arbitrary *commutative* function ( $g$  returns the same result regardless of the order the arguments are presented). In this work,  $f$  is a learned linear operator with a rectified linear (ReLU) activation function and  $g$  is a sum.

The most trivial operation is to combine one or more layers of the same type by applying the same operation to every atom or pair. Precisely, this means if you have layers  $x_1, x_2, \dots, x_n$  and function  $f$ , you can compute a new atom layer from the previous atom layer ( $A \rightarrow A$ ) as

$$A_a^y = f(A_a^{x_1}, A_a^{x_2}, \dots, A_a^{x_n}) \quad (2)$$

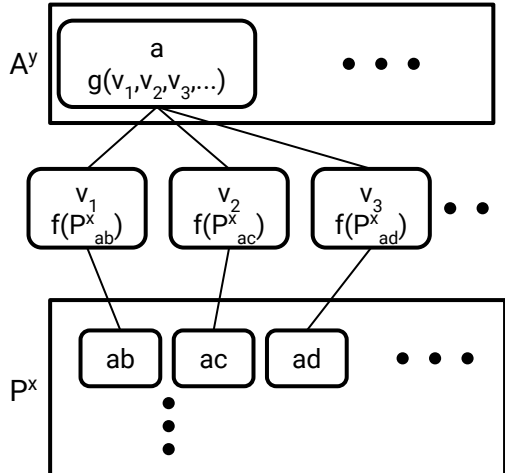


FIGURE 2:  $P \rightarrow A$  operation.  $P^x$  is a matrix containing features for atom pairs  $ab, ac, ad$ , etc. The  $v_i$  are intermediate values obtained by applying  $f$  to features for a given atom pair. Applying  $g$  to the intermediate representations for all atom pairs involving a given atom (e.g.  $a$ ) results in a new atom feature vector for that atom.

or pair layer from the previous pair layer ( $P \rightarrow P$ ) as

$$P_{a,b}^y = f(P_{a,b}^{x_1}, P_{a,b}^{x_2}, \dots, P_{a,b}^{x_n}) \quad (3)$$

Since we apply the same function for every atom/pair, we refer to this as a convolution. All the transformations we develop below will have this convolution nature of applying the same operation to every atom/pair, maintaining Property 2.

When operating on pairs of atoms, instead of putting all pairs through this function, you could select a subset. In Section 4.3.3 we show experiments for restricting the set of pairs to those that are less than some graph distance away.

Next, consider an operation that takes a pair layer  $x$  and constructs an atom layer  $y$  ( $P \rightarrow A$ ). The operation is depicted in Figure 2. Formally:

$$A_a^y = g(f(P_{(a,b)}^x), f(P_{(a,c)}^x), f(P_{(a,d)}^x), \dots) \quad (4)$$

In other words, take all pairs of which  $a$  is a part, run them through  $f$ , and combine them with  $g$ . Note that Property 3 means we can choose an arbitrary one of  $P_{(a,b)}^x$  or  $P_{(b,a)}^x$ .

The most interesting construction is making a pair layer from an atom layer ( $A \rightarrow P$ ). The operation is graphically depicted in Figure 3 and formally as

$$P_{ab}^y = g(f(A_a^x, A_b^x), f(A_b^x, A_a^x)) \quad (5)$$

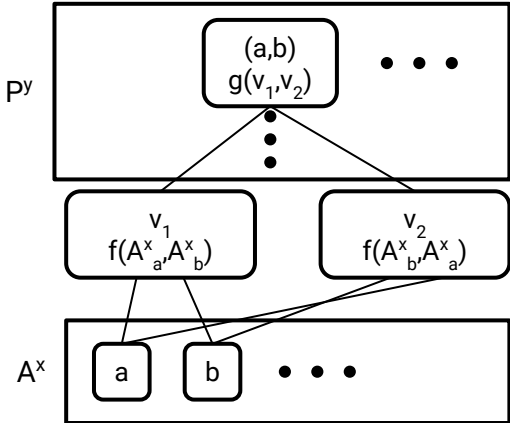


FIGURE 3:  $A \rightarrow P$  operation.  $A^x$  is a matrix containing features for atoms  $a$ ,  $b$ , etc. The  $v_i$  are intermediate values obtained by applying  $f$  to features for a given pair of atoms concatenated in both possible orderings ( $ab$  and  $ba$ ). Applying  $g$  to these intermediate ordered pair features results in an order-independent feature vector for atom pair  $ab$ .

Note that just applying  $g$  to  $A_a^x$  and  $A_b^x$  would maintain Properties 2 and 3 but we use this more complex form. While commutative operators (such as max pooling) are common in neural networks, commutative operators *with learnable parameters* are not common. Therefore, we use  $f$  to give learnable parameters while maintaining the desired properties.

Once we have all the primitive operations on atom and pair layers ( $A \rightarrow A$ ,  $P \rightarrow P$ ,  $P \rightarrow A$ ,  $A \rightarrow P$ ), we can combine these into one module. We call this the Weave module (Figure 4) because the atoms and pair layers cross back and forth to each other. The module can be stacked to an arbitrary depth similar to the Inception module that inspired it [Szegedy et al., 2015]. Deep neural networks with many layers (e.g. for computer vision) learn progressively more general features—combinations of lower-level features—in a hierarchical manner [LeCun et al., 2015]. By analogy, successive Weave modules can produce more informative representations of the original input. Additionally, stacked Weave modules with limited maximum atom pair distance progressively incorporate longer-range information at each layer.

### 3.4 Molecule-level features

The construction of the Weave module maintains Properties 2 and 3. What about overall order invariance (Property 1)? At the end of a stack of Weave

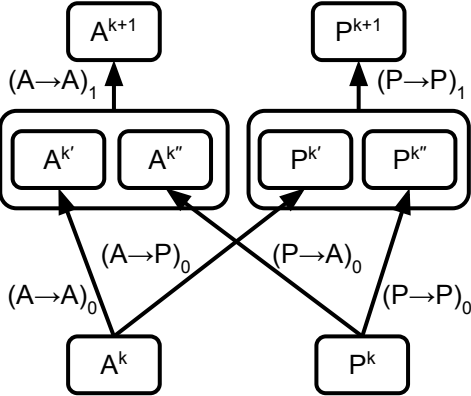


FIGURE 4: Weave module. This module takes matrices  $A^k$  and  $P^k$  (containing atom and pair features, respectively) and combines  $A \rightarrow A$ ,  $P \rightarrow P$ ,  $P \rightarrow A$ , and  $A \rightarrow P$  operations to yield a new set of atom and pair features ( $A^{k+1}$  and  $P^{k+1}$ , respectively). The output atom and pair features can be used as input to a subsequent Weave module, which allows these modules to be stacked in series to an arbitrary depth.

modules we are left with an  $n$ -dimensional vector associated with every atom and an  $m$ -dimensional vector associated with every pair. We need to turn this into a molecule-level representation with some commutative function of these vectors.

In related work [Merkwirth and Lengauer, 2005; Duvenaud et al., 2015; Lusci et al., 2013], a simple unweighted sum is often used to combine order-dependent atom features into order-independent molecule-level features. However, reduction to a single value does not capture the distribution of learned features. We experimented with an alternative approach and created “fuzzy” histograms for each dimension of the feature vector.

A fuzzy histogram is described by a set of *membership functions* that are functions with range  $[0, 1]$  representing the membership of the point in each histogram bin [Zadeh, 1965]. A standard histogram has membership functions which are 1 in the bin and 0 everywhere else. For each point, we normalize so that the total contribution to all bins is 1. The value of a bin in the histogram over all points is just the sum of the normalized contributions for all the points.

Figure 5 gives an example of a fuzzy histogram composed of three Gaussian bins. A histogram is constructed for each dimension of the feature vectors and the concatenation of those histograms is the molecule-level representation.

In this work we used Gaussian membership func-



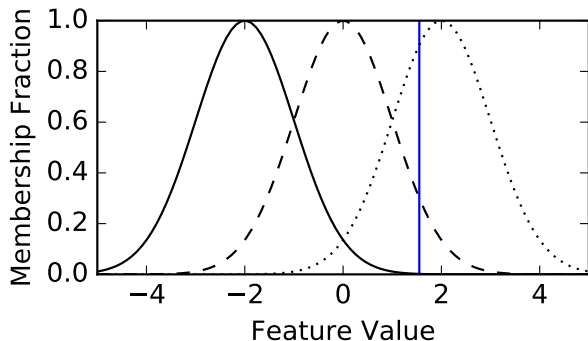


FIGURE 5: Fuzzy histogram with three Gaussian “bins”. Each curve represents the membership function for a different bin, indicating the degree to which a point contributes to that bin. The vertical blue line represents an example point which contributes normalized densities of  $< 0.01$ ,  $\sim 0.25$ , and  $\sim 0.75$  to the bins (from left to right).

tions (which are unnormalized versions of the standard Gaussian PDF) with eleven bins spanning a Gaussian distribution with mean of zero and unit standard deviation, shown in Figure E.1. These bins were chosen somewhat arbitrarily to cover the expected distribution of incoming features and were not optimized further (note that the incoming features were batch normalized; see Section 3.7).

Throughout this paper, we construct the molecule-level features only from the top-level atom features and not the pair features. This is to restrict the total number of feature vectors that must be summarized while still providing information about the entire molecule. Note, however, that the initial and intermediate pair features can influence the final atom features through Weave module operations.

Before the molecule-level featurization, we do one final convolution on the atoms. Since molecule-level featurization can be a major bottleneck in the model, this convolution expands the depth so that each dimension of the atom feature vector contains less information and therefore less information is lost during the molecule-level featurization. On this convolution, we do not use a ReLU activation function to avoid the histogram having many points at zero.

Once you have a molecule-level representation, this becomes a more standard multitask problem. We follow the common approach [Ramsundar et al., 2015; Ma et al., 2015; Mayr et al., 2015] of a small number of fully connected layers on top of the molecule-level features followed by standard softmax classification.

The overall architecture is depicted in Figure 6.

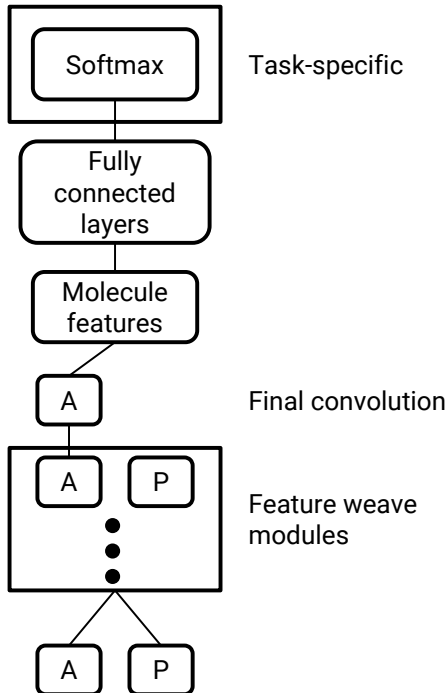


FIGURE 6: Abstract graph convolution architecture. In the current implementation, only the final atom features are used to generate molecule-level features.

Table 1 lists hyperparameters and default values for graph convolution models. In models with multiple Weave modules it is conceivable to vary the convolution depths in a module-specific way. However, the models in this work used the same settings for all Weave modules.

Our current implementation imposes an upper limit on the number of heavy atoms represented in the initial featurization. For molecules that have more than the maximum number of atoms, only a subset of atoms (and therefore atom pairs) are represented in the input encoding. This subset depends on the order in which the atoms are traversed by the featurization code and should be considered arbitrary. In this work we set the maximum number of atoms to 60, and only 814 of the 1 442 718 unique molecules in our datasets (see Section 3.6) exceed this limit.

### 3.5 Input featurization

The initial atom and pair features are summarized in Table 2 and Table 3, respectively. The features are a mix of floating point, integer, and binary values (all encoded as floating point numbers in the network).

TABLE 1: Graph convolution model hyperparameters.

Group	Hyperparameter	Default Value
Input	Maximum number of atoms per molecule	60
	Maximum atom pair graph distance	2
Weave	Number of Weave modules	1
	$(A \rightarrow A)_0$ convolution depth	50
	$(A \rightarrow P)_0$ convolution depth	50
	$(P \rightarrow P)_0$ convolution depth	50
	$(P \rightarrow A)_0$ convolution depth	50
	$(A \rightarrow A)_1$ convolution depth	50
	$(P \rightarrow P)_1$ convolution depth	50
Reduction	Final atom layer convolution depth	128
	Reduction to molecule-level features	Gaussian histogram
Post-reduction	Fully-connected layers (number of units per layer)	2000, 100
Training	Batch size	96
	Learning rate	0.003
	Optimization method	Adagrad

TABLE 2: Atom features.

Feature	Description	Size
Atom type*	H, C, N, O, F, P, S, Cl, Br, I, or metal (one-hot or null).	11
Chirality	R or S (one-hot or null).	2
Formal charge	Integer electronic charge.	1
Partial charge	Calculated partial charge.	1
Ring sizes	For each ring size (3–8), the number of rings that include this atom.	6
Hybridization	sp, sp <sup>2</sup> , or sp <sup>3</sup> (one-hot or null).	3
Hydrogen bonding	Whether this atom is a hydrogen bond donor and/or acceptor (binary values).	2
Aromaticity	Whether this atom is part of an aromatic system.	1
		27

\* Included in the “simple” featurization (see Section 4.2).

TABLE 3: Atom pair features.

Feature	Description	Size
Bond type*	Single, double, triple, or aromatic (one-hot or null).	4
Graph distance*	For each distance (1–7), whether the shortest path between the atoms in the pair is less than or equal to that number of bonds (binary values).	7
Same ring	Whether the atoms in the pair are in the same ring.	1
		12

\* Included in the “simple” featurization (see Section 4.2).

The feature set is intended to be broad, but not necessarily exhaustive, and we recognize that some features can potentially be derived from or correlated to a subset of the others (e.g. atom hybridization can be determined by inspecting the bonds that atom makes). We performed experiments using a "simple" subset of these features in an effort to understand their relative contributions to learning (Section 4.2), but many other questions about specifics of the input featurization are left to future work.

All features were generated with RDKit [Landrum, 2014], including Gasteiger atomic partial charges [Gasteiger and Marsili, 1980]. Although our featurization includes space for hydrogen atoms, we did not use explicit hydrogens in any of our experiments in order to conserve memory and emphasize contributions from heavy atoms.

Other deep learning applications with more "natural" inputs such as computer vision and speech recognition still require some input engineering; for example, adjusting images to a specific size or scale, or transforming audio into the frequency domain. Likewise, the initial values for the atom and pair layers describe these primitives in terms of properties that are often considered by medicinal chemists and other experts in the field, allowing the network to use or ignore them as needed for the task at hand. One of the purposes of this work is to demonstrate that learning can occur with as little preprocessing as possible. Accordingly, we favor simple descriptors that are more or less "obvious".

### 3.6 Datasets

We used a dataset collection nearly identical to the one described by Ramsundar et al. [2015] except for some changes to the data processing pipeline (including the duplicate merging process for the Tox21 dataset) and different cross-validation fold divisions. Briefly, there are 259 datasets divided into four groups indicating their source: PubChem BioAssay [Wang et al., 2012] (PCBA, 128 datasets), the "maximum unbiased validation" datasets constructed by Rohrer and Baumann [Rohrer and Baumann, 2009] (MUV, 17 datasets), the enhanced directory of useful decoys [Mysinger et al., 2012] (DUD-E, 102 datasets), and the training set for the Tox21 challenge (see Mayr et al. [2015]) (Tox21, 12 datasets). The combined dataset contained over 38 M data points and included targets from many different biological classes.

### 3.7 Model training and evaluation

Graph convolution and traditional neural network models were implemented with TensorFlow [Abadi et al., 2015], an open-source library for machine learning. Models were evaluated by the area under the receiver operating characteristic curve (AUC) as recommended by Jain and Nicholls [2008]. We used 5-fold stratified cross-validation, where each fold-specific model used 60% of the data for training, 20% for validation (early stopping/model selection), and 20% as a test set.

Graph convolution models were trained for 10–20 M steps using the Adagrad optimizer [Duchi et al., 2011] with learning rate 0.003 and batch size 96, with periodic checkpointing. All convolution and fully-connected layer outputs were batch normalized [Ioffe and Szegedy, 2015] prior to applying the ReLU non-linearity. Training was parallelized over 96 CPUs (or 96 GPUs in the case of the  $W_4N_2$  model) and required several days for each model. Adding additional Weave modules significantly increased training time. However, models trained on smaller datasets (see Section 3.8) trained much faster.

To establish a baseline, we also trained pyramidal (2000, 100) multitask neural network (PMTNN) [Ramsundar et al., 2015], random forest (RF), and logistic regression (LR) models using Morgan fingerprints with radius 2 (essentially equivalent to ECFP4) generated with RDKit [Landrum, 2014]. As a very simple baseline, we also computed Tanimoto similarity to all training set actives and used the maximum similarity score as the active class probability (MaxSim).

The PMTNN had two hidden layers (with 2000 and 100 units, respectively) with rectified linear activations, and each fold-specific model was trained for 40–50 M steps using the SGD optimizer with batch size 128 and a learning rate of 0.0003, with periodic checkpointing. Additionally, this model used 0.25 dropout [Srivastava et al., 2014], initial weight standard deviations of 0.01 and 0.04 and initial biases of 0.5 and 3.0 in the respective hidden layers. This model did not use batch normalization.

Logistic regression (LR) models were trained with the `LogisticRegressionCV` class in scikit-learn [Pedregosa et al., 2011] using the 'balanced' class weighting strategy and a maximum of either 1000 or 10 000 iterations. Values for the `penalty` and `C` parameters were chosen by grid search with the held-out validation set used for model selection. Random forest (RF) models were trained using the scikit-learn



`RandomForestClassifier` with 100 trees, also with the ‘balanced’ class weighting strategy.

In graph convolution and PMTNN models, active compounds were weighted in the cost function such that the total active weight equalled the total inactive weight within each dataset. Furthermore, these models were evaluated in a task-specific manner by choosing the training checkpoint with the best validation set AUC for each task. We note that some fold-specific models had a small number of tasks were not “converged” in the sense that their validation set AUC scores were still increasing when training was halted, and that the specific tasks that were not converged varied from model to model.

To statistically compare graph convolution and baseline models, we report three values for each dataset group: (1) median 5-fold mean AUC over all datasets, (2) median difference in per-dataset 5-fold mean AUC ( $\Delta$ AUC) relative to the PMTNN baseline, and (3) a 95% Wilson score interval for the sign test statistic relative to the PMTNN baseline. The sign test estimates the probability that a model will achieve a higher 5-fold mean AUC than the PMTNN baseline; models with sign test confidence intervals that do not include 0.5 are considered significantly different in their performance (the median  $\Delta$ AUC can be used as a measure of effect size). To calculate these intervals, we used the `proportion_confint` function in statsmodels [Seabold and Perktold, 2010] version 0.6.1 with `method=‘wilson’` and `alpha=0.05`, counting only non-zero differences in the sign test. We do not report values for the DUD-E dataset group since all models achieved  $> 0.98$  median 5-fold mean AUC.

### 3.8 Comparisons to other methods

In addition to the baseline models described in Section 3.7, there are many other methods that would be interesting to compare to our graph convolution models. In particular, Duvenaud et al. [2015] described “neural fingerprints” (NFP), a related graph-based method. The original publication describing NFP reported mean squared errors (MSE) on datasets for aqueous solubility, drug efficacy, and photovoltaic efficiency. We trained multitask graph convolution models on these datasets using 5-fold cross-validation (note that the published NFP models were single-task).

Additionally, we report results on a dataset used to validate the influence relevance voter (IRV) method of Swamidass et al. [2009], which is a hybrid of neural networks and  $k$ -nearest neighbors. The original pub-

lication reported results for two datasets, HIV and DHFR, but the latter was no longer available from its original source. We trained graph convolution models on the HIV dataset using 10-fold stratified cross-validation. In each cross-validation round, one fold each was used for testing and validation (early stopping), and the remaining folds were used for training. We note that RDKit was only able to process 41 476 of the 42 678 SMILES strings in the HIV dataset.

Although we expect our results on these datasets to provide reasonable comparisons to published data, differences in fold assignments and variations in dataset composition due to featurization failures mean that the comparisons are not perfect.

## 4 Results

### 4.1 Proof of concept

With so many hyperparameters to adjust, we sought to establish a centerpoint from which to investigate specific questions. After several experiments, we settled on a simple model with two Weave modules, a maximum atom pair distance of 2, Gaussian histogram molecule-level reductions, and two fully-connected layers of size 2000 and 100, respectively. Notationally, we refer to this model as  $W_2N_2$ . Table 4 shows the performance of the  $W_2N_2$  model and related models derived from this centerpoint by varying a single hyperparameter. Additionally, Table 4 includes results for several baseline models: MaxSim, logistic regression (LR), random forest (RF), and pyramidal (2000, 100) multitask neural network (PMTNN) models trained on Morgan fingerprints.

Several graph convolution models achieved performance comparable to the baseline PMTNN on the classification tasks in our dataset collection, which is a remarkable result considering the simplicity of our input representation. For example, the centerpoint  $W_2N_2$  model is statistically indistinguishable from the PMTNN for the PCBA, MUV, and Tox21 dataset groups (we do not report results for the DUD-E dataset group because all models achieved extremely high median AUC scores). Additionally, many of the graph convolution models with worse performance than the PMTNN (i.e. sign test confidence intervals excluding 0.5) had very small effective differences as measured by median  $\Delta$ AUC.

TABLE 4: Median 5-fold mean AUC values for reported models. Graph convolution models are labeled as  $W_xN_y$ , where  $x$  and  $y$  denote the number of Weave modules and the maximum atom pair distance, respectively (see the text for descriptions of the simple, sum, and RMS models). All graph convolution models fed into a Pyramidal (2000, 100) MTNN after the molecule-level feature reduction step. MaxSim, logistic regression (LR), random forest (RF), and pyramidal (2000, 100) multitask neural network (PMTNN) baselines used Morgan fingerprints as input. For each model, we report the median  $\Delta$ AUC and the 95% Wilson score interval for a sign test estimating the probability that a given model will outperform the PMTNN baseline (see Section 3.7). Bold values indicate sign test confidence intervals that do not include 0.5.

Model	PCBA ( $n = 128$ )			MUV ( $n = 17$ )			Tox21 ( $n = 12$ )		
	Median AUC	Median $\Delta$ AUC	Sign Test 95% CI	Median AUC	Median $\Delta$ AUC	Sign Test 95% CI	Median AUC	Median $\Delta$ AUC	Sign Test 95% CI
MaxSim	0.754	-0.137	<b>(0.00, 0.04)</b>	0.638	-0.136	<b>(0.01, 0.27)</b>	0.728	-0.129	<b>(0.00, 0.24)</b>
LR	0.846	-0.052	<b>(0.04, 0.14)</b>	0.748	-0.063	(0.13, 0.53)	0.797	-0.060	<b>(0.01, 0.35)</b>
RF	0.802	-0.091	<b>(0.02, 0.09)</b>	0.666	-0.133	<b>(0.00, 0.18)</b>	0.807	-0.043	<b>(0.05, 0.45)</b>
PMTNN	0.906			0.875			0.852		
$W_2N_2$ -simple	0.905	-0.003	<b>(0.27, 0.44)</b>	0.862	0.009	(0.36, 0.78)	0.866	0.001	(0.25, 0.75)
$W_2N_2$ -sum	0.898	-0.011	<b>(0.16, 0.31)</b>	0.818	-0.015	(0.13, 0.53)	0.848	-0.011	(0.09, 0.53)
$W_2N_2$ -RMS	0.901	-0.007	<b>(0.20, 0.36)</b>	0.851	-0.031	<b>(0.10, 0.47)</b>	0.858	-0.006	(0.09, 0.53)
$W_1N_2$	0.905	-0.006	<b>(0.20, 0.35)</b>	0.829	-0.011	(0.22, 0.64)	0.850	-0.008	(0.09, 0.53)
$W_2N_1$	0.907	-0.003	<b>(0.28, 0.45)</b>	0.855	-0.019	(0.13, 0.53)	0.867	-0.001	(0.19, 0.68)
$W_2N_2$	0.909	0.000	(0.40, 0.57)	0.847	-0.004	(0.17, 0.59)	0.862	0.006	(0.32, 0.81)
$W_2N_3$	0.906	-0.001	(0.38, 0.55)	0.838	-0.013	(0.17, 0.59)	0.860	0.000	(0.32, 0.81)
$W_2N_4$	0.908	-0.002	<b>(0.33, 0.49)</b>	0.836	-0.003	(0.17, 0.59)	0.858	0.002	(0.39, 0.86)
$W_2N_\infty$	0.897	-0.008	<b>(0.12, 0.26)</b>	0.841	-0.032	<b>(0.06, 0.41)</b>	0.846	-0.006	(0.14, 0.61)
$W_3N_2$	0.906	0.000	(0.41, 0.59)	0.879	0.000	(0.31, 0.74)	0.859	0.005	(0.47, 0.91)
$W_4N_2$	0.907	-0.001	<b>(0.30, 0.47)</b>	0.856	-0.007	(0.22, 0.64)	0.863	0.006	(0.39, 0.86)

We also trained graph convolution models on some additional datasets in order to compare to the “neural fingerprints” (NFP) of Duvenaud et al. [2015] and the influence relevance voter (IRV) method of Swamidass et al. [2009] (see Section 3.8). Table 5 compares graph convolution models to published results on these datasets under similar cross-validation conditions. Graph convolution results were comparable to published NFP models, with significant improvement on the photovoltaic efficiency task (note that the graph convolution results are from multitask models trained on all three NFP datasets while Duvenaud et al. [2015] report values for single-task models). The 10-fold mean AUC on the HIV dataset was slightly lower than the published IRV value. However, we held out 10% of the data (one fold) in each cross-validation round as a validation set for checkpoint selection, meaning that the graph convolution models were trained with fewer examples than the published IRV models.

## 4.2 Input featurization

As a further proof of concept and to address the importance of the initial featurization, we trained a model using a subset of features that match typical 2D structural diagrams seen in chemistry textbooks: only atom type, bond type, and graph distance are provided to the network. Figure 7 compares a model trained with this “simple” input featurization to the “full” featurization containing all features from Table 2 and Table 3. Both featurizations achieve similar median 5-fold mean AUC scores, suggesting that the additional features in the “full” representation are either mostly ignored during training or can be derived from a simpler representation of the molecular graph. Further work is required to understand the importance of individual features, perhaps with datasets that are sensitive to particular components of the input representation (such as hydrogen bonding or formal charge).

Figure 8 gives examples of how the initial atom features for a single molecule (ibuprofen) evolve as they progress through graph convolution Weave modules. The initial atom and pair feature encodings for the “full” featurization are depicted in Panel A. Comparing the initial atom features to their source molecular graph, the aromatic carbons in the central ring are clearly visible (and nearly identical in the featurization). The pair features are more difficult to interpret visually, and mostly encode graph distance. As the atom features are transformed by the Weave modules

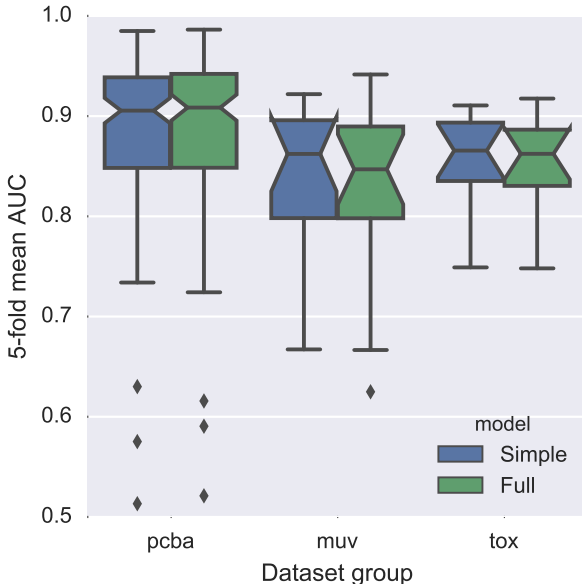


FIGURE 7: Comparison of models with “simple” and “full” input featurizations. The simple featurization only encodes atom type, bond type, and graph distance. The full featurization includes additional features such as aromaticity and hydrogen bonding propensity (see Section 3.4 for more details). Confidence intervals on box plot medians are computed as  $\pm 1.57 \times \text{IQR} / \sqrt{N}$  [McGill et al., 1978].

(Panel B), they become more heterogeneous and reflective of their unique chemical environments. “Simple” features behave similarly, beginning with rather sterile initial values and quickly diverging as neighborhood information is included by Weave module operations (Panel C). Comparison of the “full” and “simple” atom features after the second Weave module shows that both featurizations lead to similarly diverse feature distributions. Figure D.1 and Figure D.2 show similar behavior for pair features.

## 4.3 Hyperparameter sensitivity

### 4.3.1 Number of Weave modules

In relatively “local” models with limited atom pair distance, successive Weave modules update atom features with information from progressively larger regions of the molecule. This suggests that the number of Weave modules is a critical hyperparameter to optimize, analogous to the number of hidden layers in traditional neural networks. Figure 9 compares models with 2–4 Weave modules to a model with a

TABLE 5: Comparison of graph convolution to neural fingerprint (NFP) and influence relevance voter (IRV) models. Section 3.8 provides details for datasets and experimental procedures. Note that the NFP comparisons were performed using multitask graph convolution models, and that graph convolution models for the HIV dataset were trained with fewer examples than IRV since one cross-validation fold was used as a held-out validation set.

Model	Dataset	Metric	Original	Graph Convolution
NFP	Solubility (log M)	MSE	$0.52 \pm 0.07$	$0.46 \pm 0.08$
	Drug efficacy (nM EC <sub>50</sub> )	MSE	$1.16 \pm 0.03$	$1.07 \pm 0.06$
	Photovoltaic efficiency (%)	MSE	$1.43 \pm 0.09$	$1.10 \pm 0.06$
IRV	HIV	AUC	0.845	$0.834 \pm 0.025$

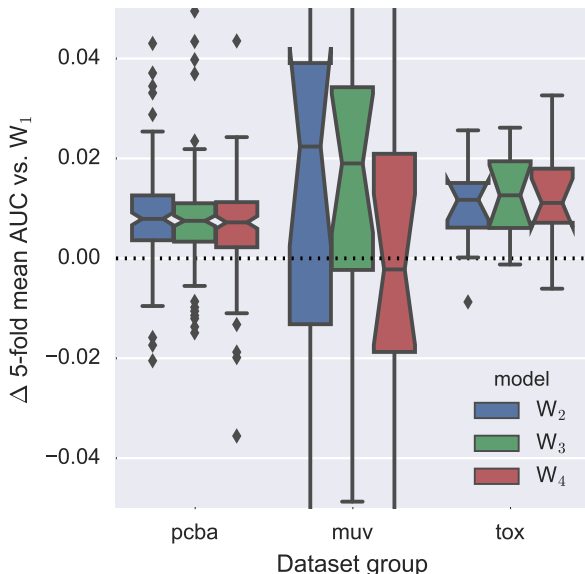


FIGURE 9: Comparison of models with different numbers of Weave modules with a model containing a single Weave module. All models used a maximum atom pair distance of two. The  $y$ -axis is cropped to emphasize differences near zero.

single Weave module. As expected, models with a single Weave layer are outperformed by deeper architectures. For the PCBA and Tox21 datasets, there was not much benefit to using more than two Weave modules (Table 4 and Figure C.1), but using three Weave modules gave the best median AUC for the MUV datasets (in exchange for significantly increased training time).

### 4.3.2 Alternative feature reductions

The reduction of atom features from the final Weave module to an order-invariant, molecule-level representation is a major information bottleneck in graph convolution models. In related work, a simple unweighted sum [Duvenaud et al., 2015; Merkwirth and Lengauer, 2005; Lusci et al., 2013] or root-mean-square (RMS) [Dieleman, March 17, 2015] reduction is used. Using a consistent base architecture with two Weave modules and a maximum atom pair distance of 2, we compared these traditional reduction strategies with our Gaussian histogram approach.

Figure 10 shows that Gaussian histogram models had consistently improved scores relative to sum reductions. RMS reductions were not as robust as Gaussian histograms in terms of per-dataset differences relative to sum reductions, although RMS and Gaussian histogram reductions had similar distributions of absolute AUC values (Figure C.2).

### 4.3.3 Distance-dependent pair features

In Weave modules, atoms are informed about their chemical environment by mixing with pair features in the  $P \rightarrow A$  operation. Recall that during this operation, pair features are combined for pairs that contain a given atom, yielding a new representation for that atom. A critical parameter for this operation is the maximum distance (in bonds) allowed between the atoms of the pairs that are combined. If only adjacent atoms are combined, the resulting atom features will reflect the local chemical environment. As an alternative to increasing the number of Weave modules, longer-range interactions can be captured by increasing the maximum atom pair distance. However, our implementation of the  $P \rightarrow A$  operation uses a simple sum to combine pair features, such that a large amount of information (possibly including every pair

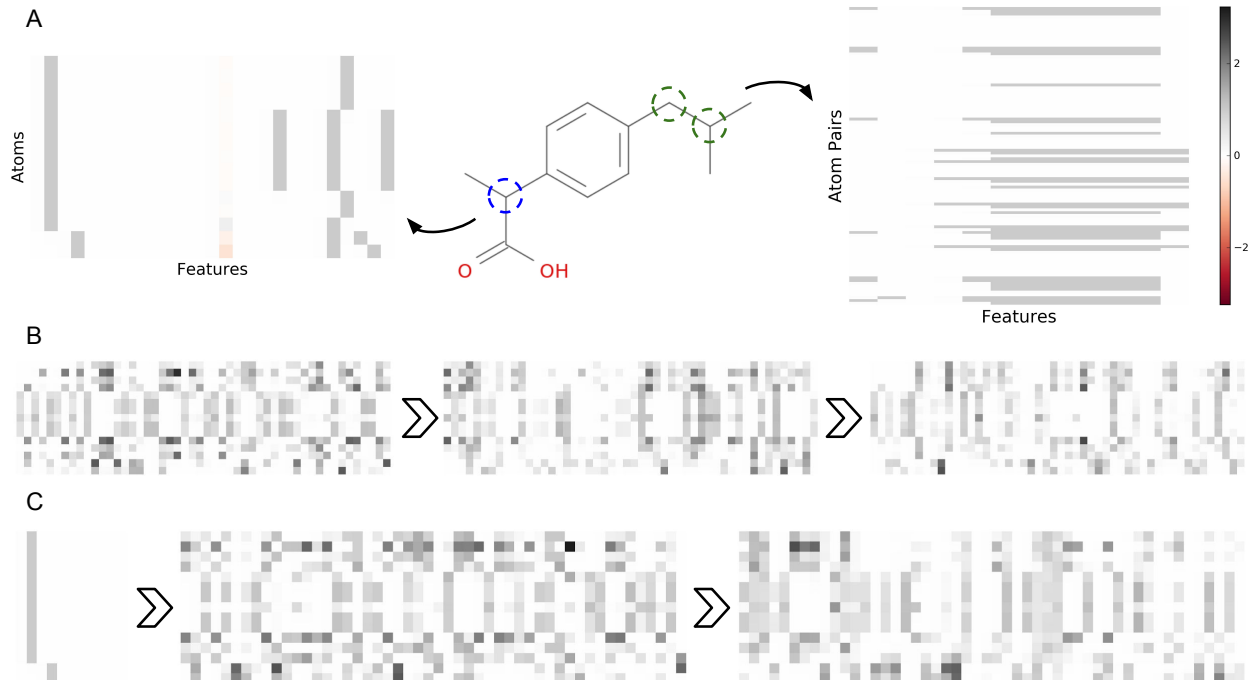


FIGURE 8: Graph convolution feature evolution. Atoms or pairs are displayed on the  $y$ -axis and the dimensions of the feature vectors are on the  $x$ -axis. (A) Conversion of the molecular graph for ibuprofen into atom and (unique) atom pair features. (B) Evolution of atom features after successive Weave modules in a graph convolution model with a  $W_3N_2$  architecture and depth 50 convolutions in Weave modules. (C) Evolution of “simple” atom features (see Section 4.2) starting from initial encoding and progressing through the Weave modules of a  $W_2N_2$  architecture. The color bar applies to all panels.

of atoms in the molecule) is combined in a way that could prevent useful information from being available in later stages of the network.

Figure 11 shows the performance of several models with different maximum pair distances relative to a model that used only adjacent atom pairs ( $N_1$ ). For the PCBA datasets, a maximum distance of 2 ( $N_2$ ) improves performance relative to the  $N_1$  model, and  $N_\infty$  (no maximum distance) is clearly worse. However, the  $N_1$  model achieves the best median AUC score for the MUV and Tox21 datasets (Table 4 and Figure C.3). These results suggest that graph convolution models do not effectively make use of the initial graph distance features to preserve or emphasize distance-dependent information.

To further investigate the effect of distance information in Weave modules, we experimented with models that use distance-specific weights for operations involving pair features in order to maintain distance information explicitly throughout the network. However, results for these models are preliminary and were not included in this report.

## 5 Discussion

*Graph convolutions* are a deep learning architecture for learning directly from undirected graphs. In this work, we emphasize their application to small molecules—undirected graphs of atoms connected by bonds—for virtual screening. Starting from simple descriptions of atoms, bonds between atoms, and pairwise relationships in a molecular graph, we have demonstrated performance that is comparable to state of the art multitask neural networks trained on traditional molecular fingerprint representations, as well as alternative methods including “neural fingerprints” [Duvenaud et al., 2015] and influence relevance voter [Swamidass et al., 2009].

Our experiments with the adjustable parameters in graph convolution models indicate a relatively minor sensitivity to the number of Weave modules and the maximum distance between atom pairs (at least for our datasets). These results suggest that a model with two Weave modules, a maximum atom pair distance of 2, and Gaussian histogram reduc-

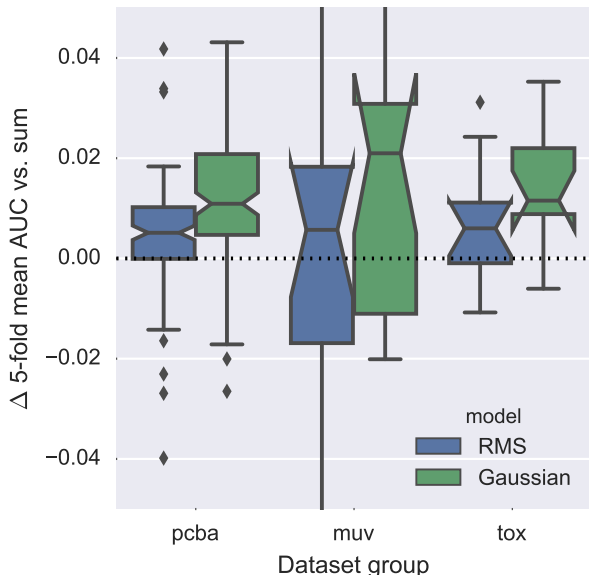


FIGURE 10: Comparison of root-mean-square (RMS) and Gaussian histogram reductions *vs.* sum reduction. The *y*-axis reports difference in 5-fold mean AUC relative to sum reduction. All models used two Weave modules and a maximum atom pair distance of two. The *y*-axis is cropped to emphasize differences near zero.

tions is a good starting point for further optimization. Remarkably, graph convolution models perform well with a “simple” input featurization containing only atom type, bond type, and graph distances—essentially the information available from looking at Figure 1.

Flexibility is a highlight of the graph convolution architecture: because we begin with a representation that encodes the complete molecular graph, graph convolution models are free to use any of the available information for the task at hand. In a sense, every possible molecular “fingerprint” is available to the model. Said another way, graph convolutions and other graph-based approaches purposefully blur the distinction between molecular features and predictive models. As has been pointed out elsewhere [Duvinaud et al., 2015], the ability to use backpropagation to tune parameters at every stage of the network provides greater representational power than traditional descriptors, which are inflexible in the features they encode from the initial representation. Accordingly, it is not appropriate to think of graph-based methods as alternative descriptors; rather, they should be viewed as fully integrated approaches to virtual

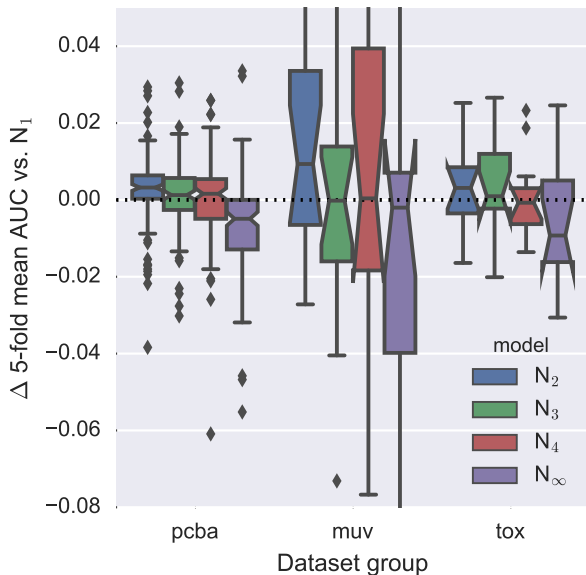


FIGURE 11: Comparison of models with different maximum atom pair distances to a model with a maximum pair distance of one (bonded atoms). All models have two Weave modules. The *y*-axis is cropped to emphasize differences near zero.

screening (although future work could investigate the utility of the learned molecule-level features for additional tasks or other applications such as molecular similarity).

Looking forward, graph convolutions (and related graph-based methods; see Section 2) present a “new hill to climb” in computer-aided drug design and cheminformatics. Although our current graph convolution models do not consistently outperform state-of-the-art fingerprint-based models, we emphasize their flexibility and potential for further optimization and development. In particular, we are aware of several specific opportunities for improvement, including (1) additional optimization of model hyperparameters such as Weave module convolution depths; (2) fine-tuning of architectural decisions, such as the choice of reduction in the  $P \rightarrow A$  operation (currently a sum, but perhaps a Gaussian histogram or distance-dependent function); and (3) improvements in memory usage and training performance, such as not handling all pairs of atoms or implementing more efficient versions of Weave module operations. With these and other optimizations, we expect that graph convolutions could exceed the performance of the best available fingerprint-based methods.



Finally, we note that much (or most) of the information required to represent biological systems and the interactions responsible for small molecule activity is not encapsulated in the molecular graph. Biology takes place in a three-dimensional world, and is sensitive to shape, electrostatics, quantum effects, and other properties that emerge from—but are not necessarily unique to—the molecular graph (see, for example, Nicholls et al. [2010]). Additionally, most small molecules exhibit 3D conformational flexibility that our graph representation does not even attempt to describe. The extension of deep learning methods (including graph convolutions) to three-dimensional biology is an active area of research (e.g. Wallach et al. [2015]) that requires special attention to the added complexities of multiple-instance learning in a relatively small-data regime.

## Acknowledgments

We thank Bharath Ramsundar, Brian Goldman, and Robert McGibbon for helpful discussion. We also acknowledge Manjunath Kudlur, Derek Murray, and Rajat Monga for assistance with TensorFlow. S.K. was supported by internships at Google Inc. and Vertex Pharmaceuticals Inc. Additionally, we acknowledge use of the Stanford BioX3 cluster supported by NIH S10 Shared Instrumentation Grant 1S10RR02664701. S.K. and V.P. also acknowledge support from from NIH 5U19AI109662-02.

## References

- Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*, 2015.
- Pedro J Ballester and W Graham Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of computational chemistry*, 28(10):1711–1723, 2007.
- Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.
- Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.
- George Dahl. Deep learning how I did it: Merck 1st place interview. *Online article available from <http://blog.kaggle.com/2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview>*, 2012.
- George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*, 2014.
- Sander Dieleman. Classifying plankton with deep neural networks. March 17, 2015. URL <http://benanne.github.io/2015/03/17/plankton.html>.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.
- David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, pages 2215–2223, 2015.
- Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.
- Paul CD Hawkins, A Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. *Journal of medicinal chemistry*, 50(1):74–82, 2007.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Ajay N Jain and Anthony Nicholls. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design*, 22(3-4):133–139, 2008.
- Greg Landrum. RDKit: Open-source cheminformatics; <http://www.rdkit.org>, 2014. URL <http://www.rdkit.org>.

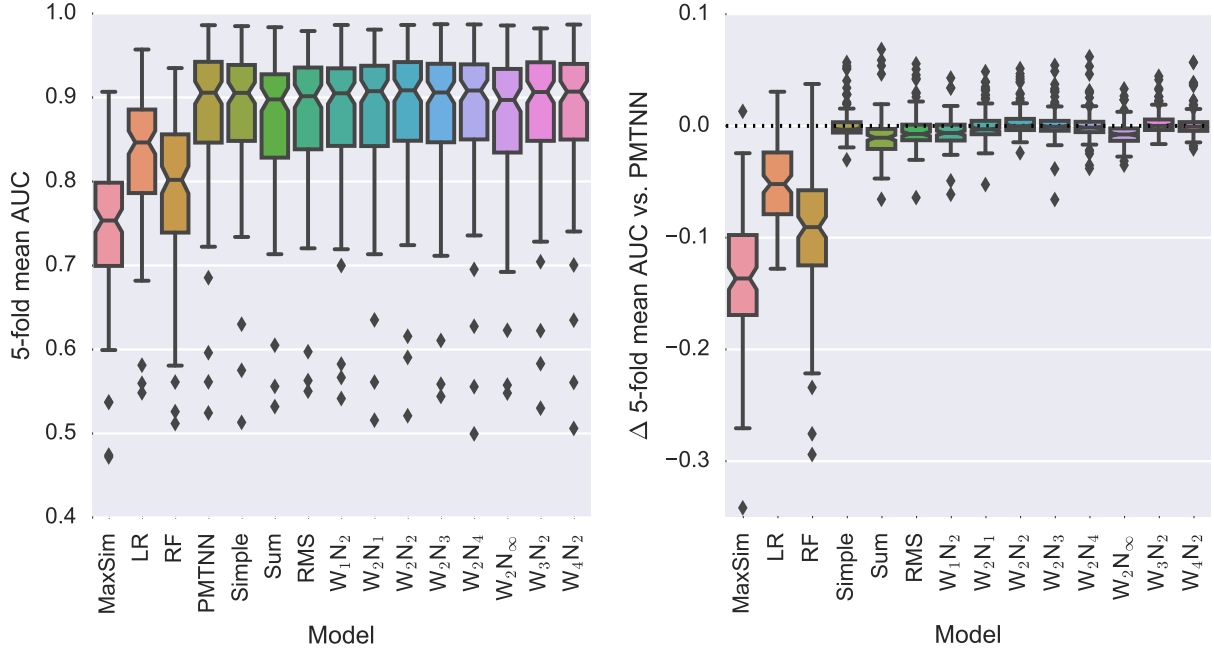
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling*, 53(7):1563–1575, 2013.
- Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2):263–274, 2015.
- Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 37–45, 2015.
- Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2015.
- Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.
- Christian Merkwirth and Thomas Lengauer. Automatic generation of complementary descriptors with molecular graph networks. *Journal of chemical information and modeling*, 45(5):1159–1168, 2005.
- Alessio Micheli. Neural network for graphs: A contextual constructive approach. *Neural Networks, IEEE Transactions on*, 20(3):498–511, 2009.
- Steven W Muchmore, Andrew J Souers, and Irini Akritopoulou-Zanze. The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chemical biology & drug design*, 67(2):174–176, 2006.
- Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.
- 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, et al. Molecular shape and medicinal chemistry: a perspective. *Journal of medicinal chemistry*, 53(10):3862–3886, 2010.
- OpenEye GraphSim Toolkit. URL <http://www.eyesopen.com>. OpenEye Scientific Software, Santa Fe, NM.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Paula M Petrone, Benjamin Simms, Florian Nigsch, Eugen Lounkine, Peter Kutchukian, Allen Cornett, Zhan Deng, John W Davies, Jeremy L Jenkins, and Meir Glick. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS chemical biology*, 7(8):1399–1409, 2012.
- Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- Sebastian G Rohrer and Knut Baumann. Maximum unbiased validation (MUV) data sets for virtual screening based on pubchem bioactivity data. *Journal of chemical information and modeling*, 49(2):169–184, 2009.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.
- Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *Neural Networks, IEEE Transactions on*, 20(1):61–80, 2009.
- Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61, 2010.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov.

- Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- S Joshua Swamidass, Chloé-Agathe Azencott, Ting-Wan Lin, Hugo Gramajo, Shiou-Chuan Tsai, and Pierre Baldi. Influence relevance voting: an accurate and interpretable virtual high throughput screening method. *Journal of chemical information and modeling*, 49(4):756–766, 2009.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabbinovich. Going deeper with convolutions. In *CVPR 2015*, 2015. URL <http://arxiv.org/abs/1409.4842>.
- Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*, volume 41. John Wiley & Sons, 2009.
- Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.
- Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A Shoemaker, et al. PubChem’s BioAssay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.
- Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

# APPENDIX

## A Appendix: Model comparison

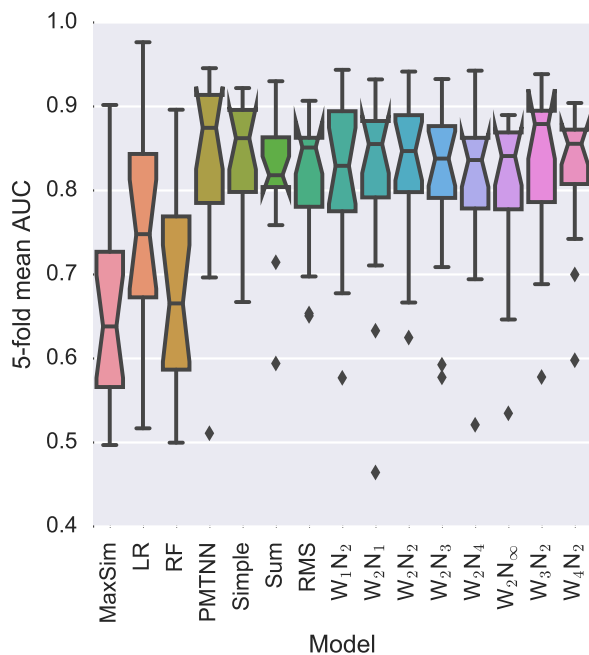
The following figures are box plot representations of the data summarized in Table 4, organized by dataset group. We provide (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against the pyramidal (2000, 100) multitask neural network (PMTNN) baseline model. The difference box plots are visual analogs of the sign test confidence intervals reported in Table 4. Note, however, that the confidence intervals on box plot medians (calculated as  $\pm 1.57 \times \text{IQR} / \sqrt{N}$  [McGill et al., 1978]) do not necessarily correspond to the sign test confidence intervals.



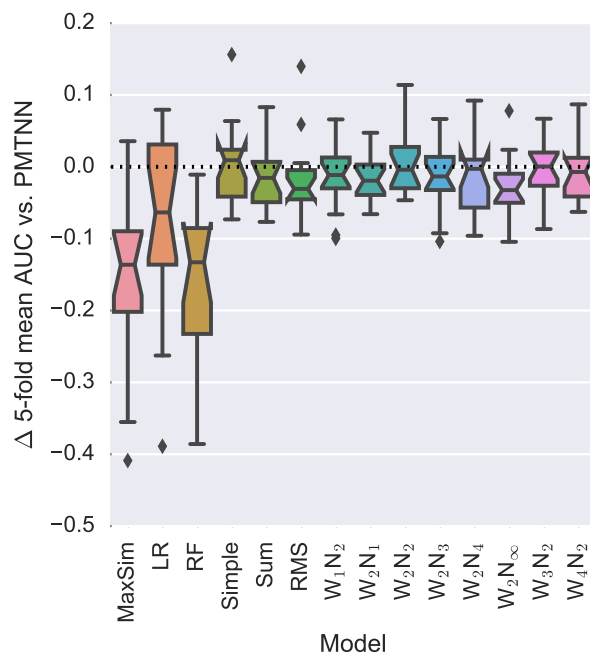
(A) Full box plot.

(B) Difference box plot vs. PMTNN.

FIGURE A.1: Model performance on PCBA datasets.

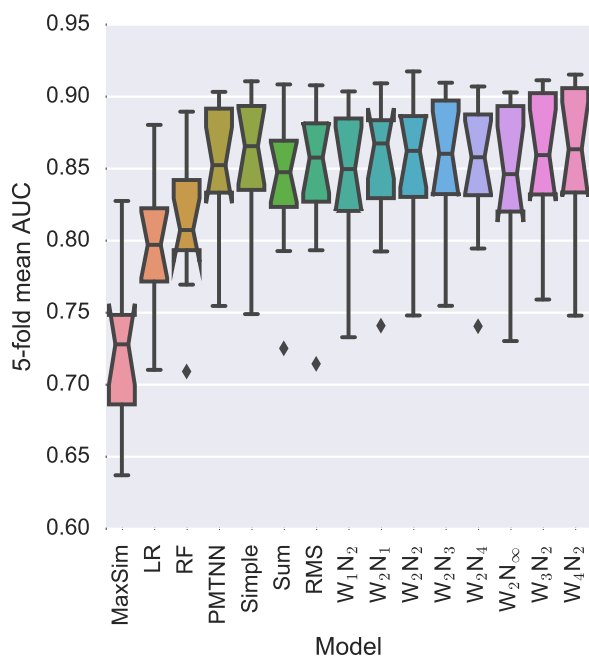


(A) Full box plot.

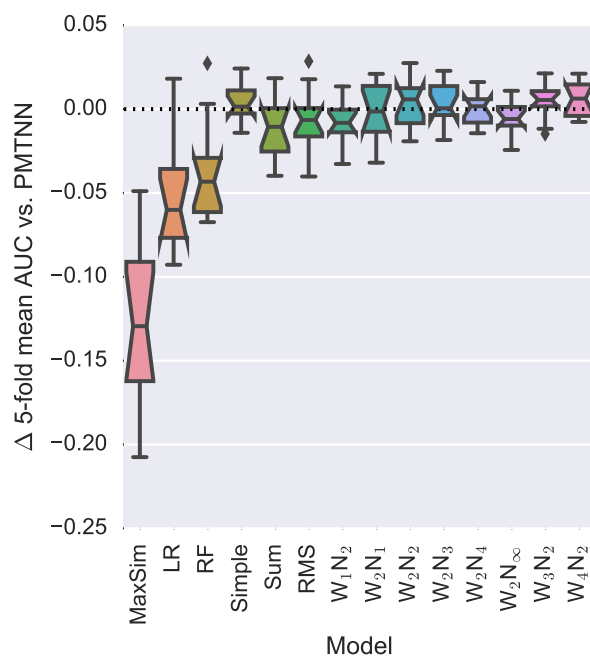


(B) Difference box plot vs. PMTNN.

FIGURE A.2: Model performance on MUV datasets.



(A) Full box plot.

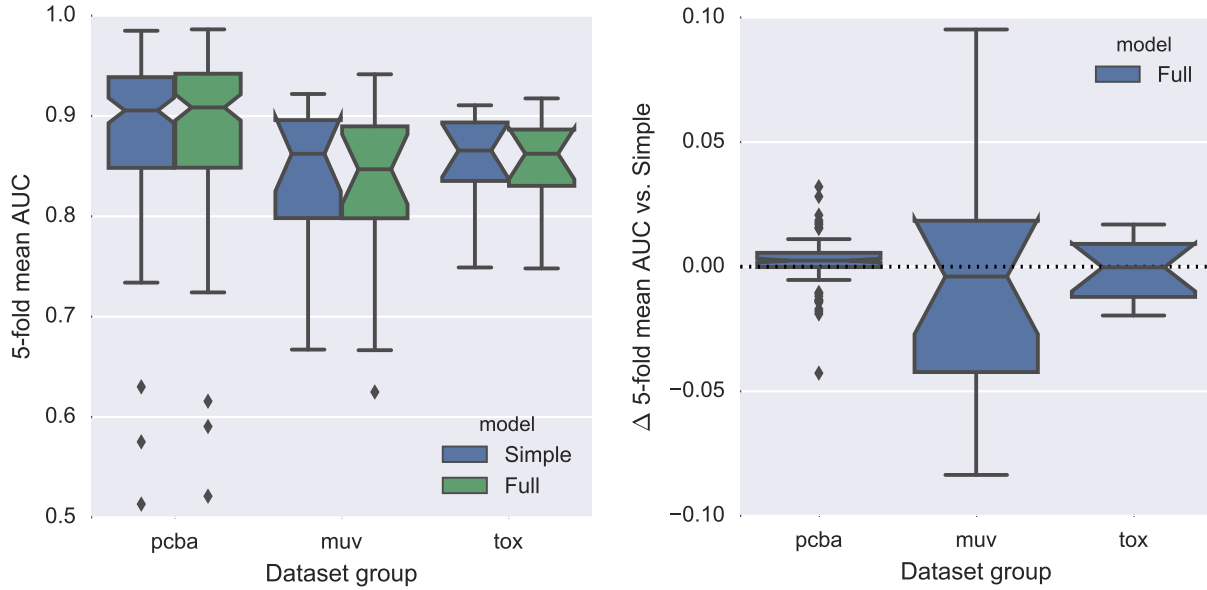


(B) Difference box plot vs. PMTNN.

FIGURE A.3: Model performance on Tox21 datasets.

## B Appendix: Input featurization

For each of the experiments described in Section 4.2, we provide figures showing (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against a baseline model (without any  $y$ -axis cropping).



(A) Full box plot.

(B) Difference box plot vs. "simple" featurization.

FIGURE B.1: Comparison of models with "simple" and "full" input featurizations.



## C Appendix: Hyperparameter sensitivity

For each of the experiments described in Section 4.3, we provide figures showing (a) box plots for absolute 5-fold mean AUC scores for each model and (b) difference box plots showing differences in 5-fold mean AUC scores against a baseline model (without any  $y$ -axis cropping).

### C.1 Number of Weave modules

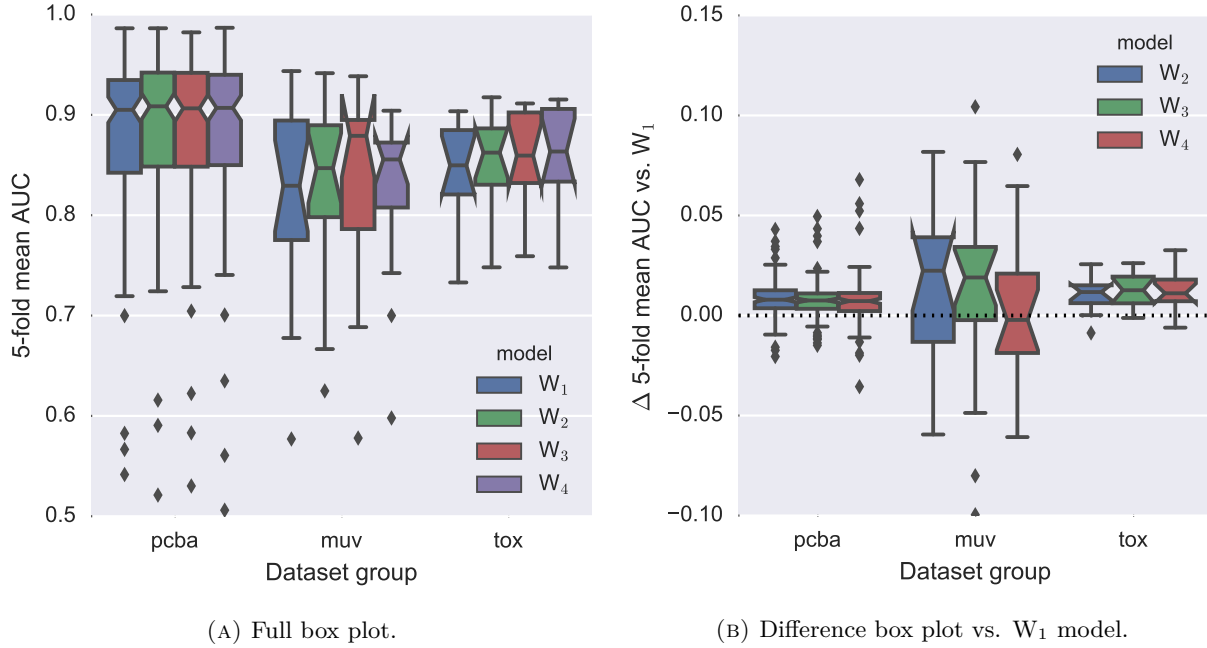


FIGURE C.1: Comparison of models with different numbers of Weave modules.

## C.2 Alternative feature reductions

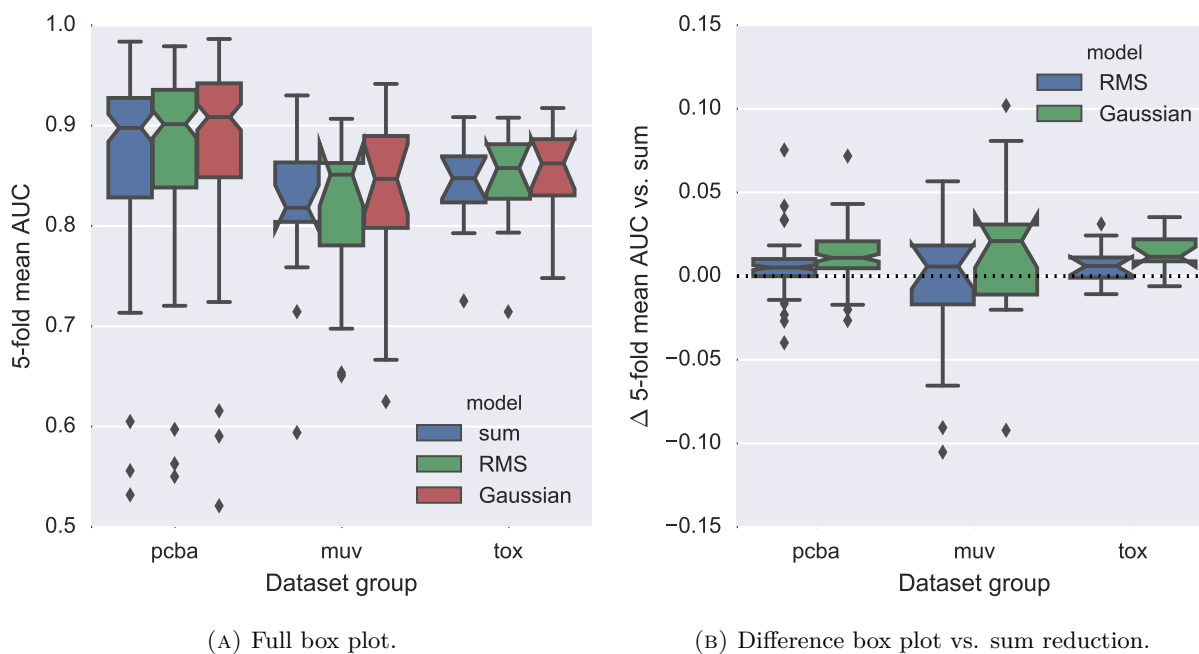


FIGURE C.2: Comparison of models with different feature reduction methods.

## C.3 Distance-dependent pair features

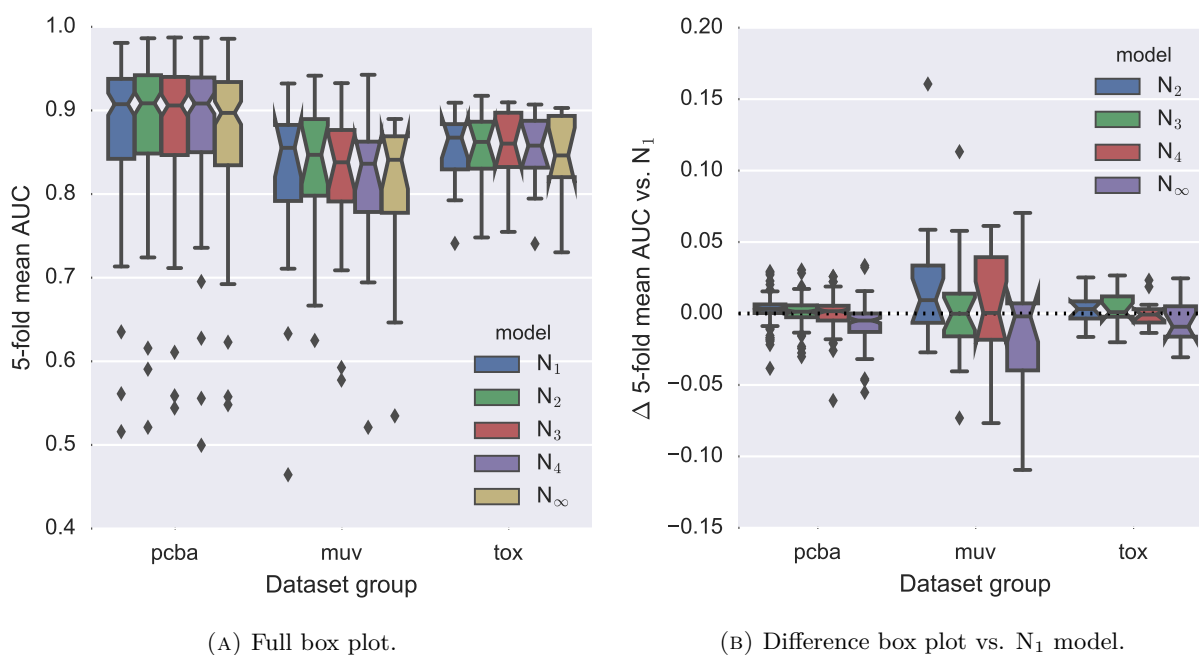


FIGURE C.3: Comparison of models with different maximum atom pair distances.

## D Appendix: Atom pair feature evolution

Figure 8 showed the evolution of atom features at different stages of a graph convolution model (after subsequent Weave modules). The following figures show the evolution of atom pair features from the same models, using both the “full” and “simple” input featurization. As in Figure 8, the initial pair features describe ibuprofen. Most of the initial featurization describes the graph distance between the atoms in the pair (see Table 3). There are many blank rows since pairs separated by more than the maximum atom pair distance are masked. Note that only unique pairs are represented (i.e.  $(a, b)$  but not  $(b, a)$ ). As the pair features move through the graph convolution network, it can be seen that similar initial featurizations diverge as a consequence of Weave module operations.

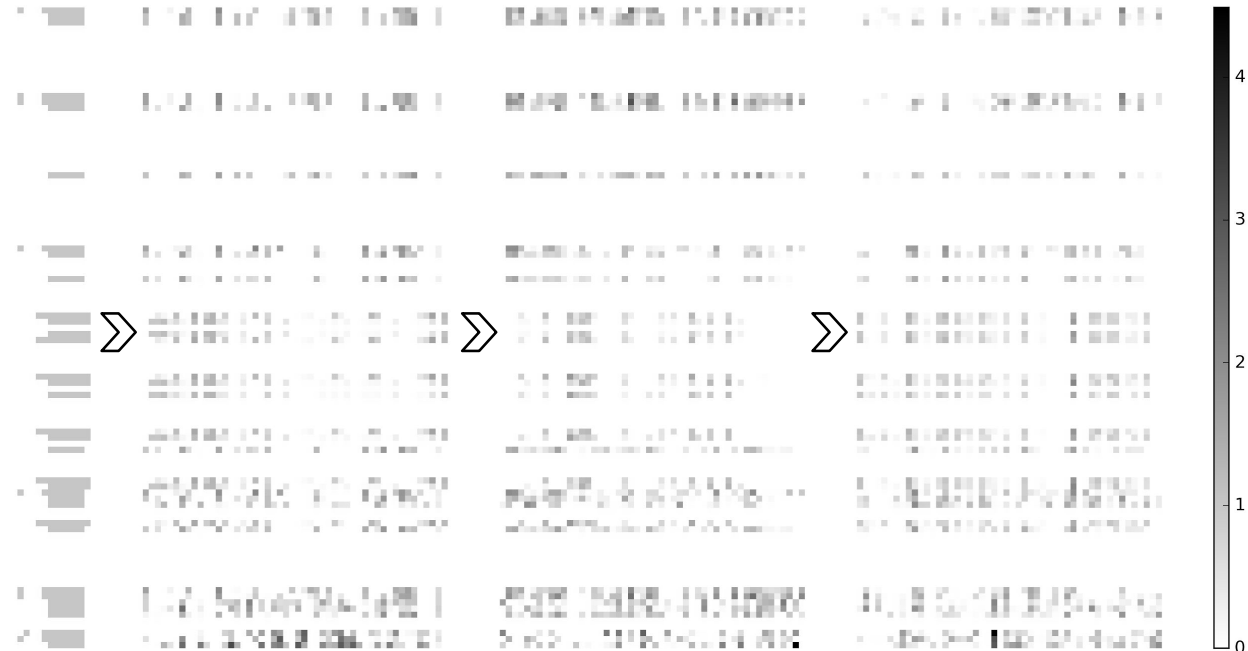


FIGURE D.1: Graph convolution atom pair feature evolution using the “full” featurization in a  $W_3N_2$  architecture. Unique atom pairs are on the  $y$ -axis (one atom pair per row). Initial pair features are shown on the left, with whitespace separating subsequent Weave module outputs.

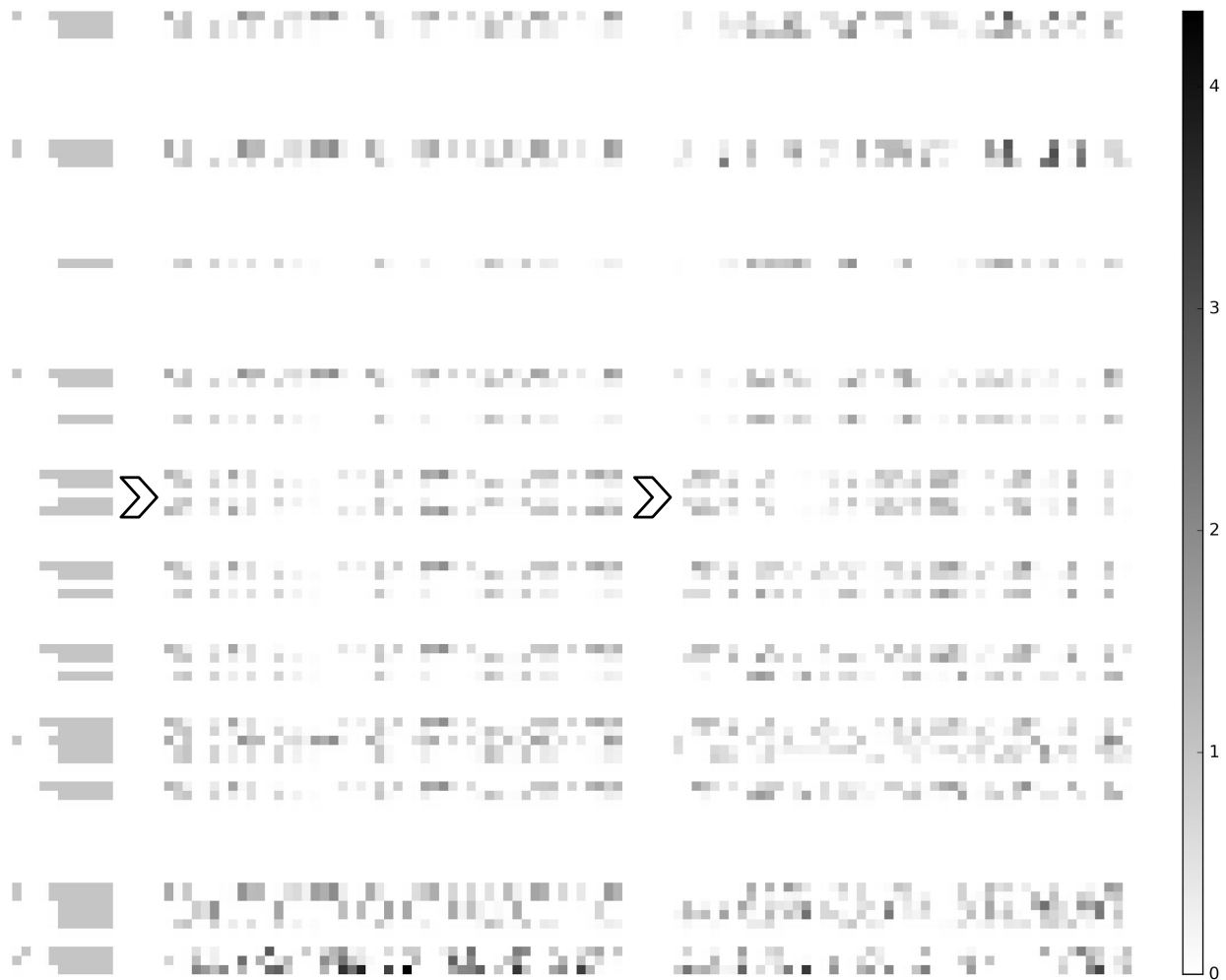


FIGURE D.2: Graph convolution atom pair feature evolution using the “simple” featurization in a  $W_2N_2$  architecture. Unique atom pairs are on the  $y$ -axis (one atom pair per row). Initial pair features are shown on the left, with whitespace separating subsequent Weave module outputs.

## E Appendix: Gaussian histogram membership functions

TABLE 6: Gaussian membership functions.

Mean	Variance
−1.645	0.080
−1.080	0.029
−0.739	0.018
−0.468	0.014
−0.228	0.013
0.000	0.013
0.228	0.013
0.468	0.014
0.739	0.018
1.080	0.029
1.645	0.080

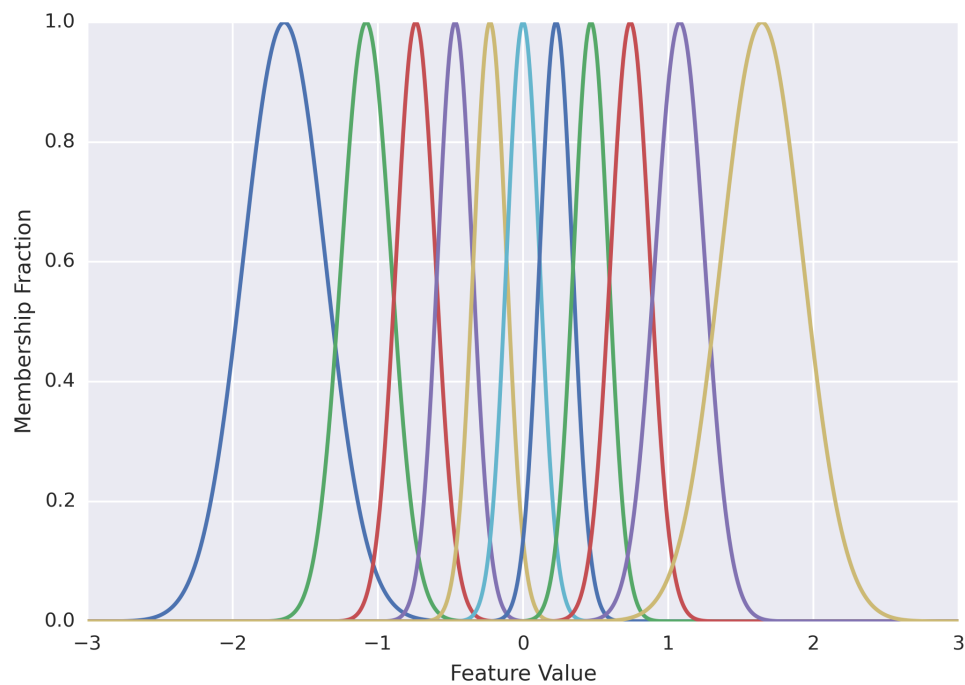


FIGURE E.1: Visualization of the Gaussian membership functions.

## References

Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.