How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions?

Garrett B. Goh,* † Charles Siegel, Abhinav Vishnu, † Nathan O. Hodas, ‡ Nathan Baker†

†High Performance Computing Group, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354

‡Data Science and Analytics, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354

* Corresponding Author: Garrett B. Goh
Email: garrett.goh@pnnl.gov

Keyword: Deep learning, deep neural networks, cheminformatics, artificial intelligence, machine intelligence, computer vision, convolutional neural network, computational chemistry, QSAR, QSPR
Abstract

In the last few years, we have seen the rise of deep learning applications in a broad range of chemistry research problems. Recently, we reported on the development of Chemception, a deep convolutional neural network (CNN) architecture for general-purpose small molecule property prediction. In this work, we investigate the effects of systematically removing and adding basic chemical information to the image channels of the 2D images used to train Chemception. By augmenting images with only 3 additional basic chemical information, we demonstrate that Chemception now outperforms contemporary deep learning models trained on more sophisticated chemical representations (molecular fingerprints) for the prediction of toxicity, activity, and solvation free energy, as well as physics-based free energy simulation methods. Thus, our work demonstrates that a firm grasp of first-principles chemical knowledge is not a pre-requisite for deep learning models to accurately predict chemical properties. Lastly, by altering the chemical information content in the images, and examining the resulting performance of Chemception, we also identify two different learning patterns in predicting toxicity/activity as compared to solvation free energy, and these patterns suggest that Chemception is learning about its tasks in the manner that is consistent with established knowledge.
1. **Introduction**

The role of deep learning in transforming computer vision research is significant. Prior to deep learning, computer vision researchers reached a glass ceiling of >20% error rate in image recognition tasks. Within 3 years of using deep neural network (DNN) models, human-level accuracy of under 5% error rate was achieved. Since then, DNN-based models have become the dominant algorithm in computer vision, and this success has been repeated in other technology-related fields as well.\(^1\)\(^2\) In chemistry, DNN-based models were the winning entry in the Merck Kaggle challenge in 2012 and the NIH Tox21 challenge in 2014. Following this trend, several research groups have started using DNN-based models to predict numerous properties, including activity,\(^8\)-\(^11\) toxicity,\(^12\)-\(^13\) reactivity,\(^14\)-\(^16\) solubility,\(^17\) ADMET,\(^18\) docking,\(^19\) and QM-computed energies.\(^20\)-\(^22\) Across the entire chemistry field, DNN-based models have been found to perform as well as or better than previous state-of-the-art models based on traditional machine (ML) learning algorithms.\(^23\)\(^24\)

DNN models are capable of learning representations, which sets it apart from conventional ML algorithms used in chemistry. Representation learning is the process of transforming input data into a feature that can be effectively exploited to identify patterns from data. Historically, computer vision research invested significant effort in designing appropriate features using applied mathematics.\(^25\) However, today, such expert-driven feature engineering research has been replaced by deep learning models that automatically develop their own features, which are not only better than expert-designed features, but it was also the primary reason why DNN models were able to exceed the glass ceilings of that field.\(^1\)\(^2\) In the chemistry context, the analogous process would be to use deep learning to examine chemical structures and to construct features similar to molecular fingerprints and descriptors that can be used to predict more complex chemical properties, and
with minimal assistance from an expert chemist. This approach that leverages representation learning of deep neural networks, is a significant departure from the traditional research paradigm in chemistry. As evidenced by the success of deep learning applications in other domains, this data-driven research paradigm holds the potential for substantially accelerating research progress.

We observed that recent trends in the literature have a subtle shift from using descriptors and fingerprints to train QSAR/QSPR models, to the development of engineered representations of a molecule. Specifically, the Aspuru-Guzik\textsuperscript{26} and Pande groups\textsuperscript{27} have developed deep learning algorithms for analyzing molecular graphs, and using SMILES data (a text-based representation) was also found to be effective.\textsuperscript{28} Recent work by Goh \textit{et. al.}\textsuperscript{29} on the Chemception deep neural network model has also demonstrated that even “unsophisticated” data like 2D images of molecules can be used to develop models that were on average equivalent to contemporary models trained on molecular fingerprints.\textsuperscript{29} Despite the above-mentioned accomplishments, the maximum potential of DNN models are ultimately limited by data. In other words, data-driven models can only be as good as what can be learned from the data. Practically, this poses a problem, as while the amount of chemical data has increased over time, obtaining labeled data (i.e. measurements) still requires experimentation or computation, which is usually an expensive and/or a slow process.

Building on the original Chemception work,\textsuperscript{29} and to mitigate this small labeled data problem in chemistry, we have devised a novel image augmentation method that adds basic chemical information into the image channels of the 2D images used to train Chemception. The hypothesis behind this approach is that if basic chemical information were to be provided to the neural network, it would not need to learn the representations for those features, but instead will be able to direct more of its learning capacity to develop more sophisticated representations that are useful for predicting chemical properties. Our results indicate that only a little extra chemical
information needs to be provided for the Chemception model to outperform both contemporary QSAR/QSPR deep learning models trained on molecular fingerprints for the prediction of toxicity, activity, and solvation free energy, and physics-based free energy simulation methods for computing solvation energies. Thus, our work demonstrates that a firm grasp of first-principles chemical knowledge is not a pre-requisite for deep learning models to accurately predict chemical properties. Then, we systematically altered the chemical information content encoded into the image and observed two distinct learning patterns in Chemception’s approach to predicting toxicity/activity as compared to solvation free energy, and these patterns suggest that Chemception is learning about its tasks in the manner that is consistent with established chemical knowledge.
2. Theory

2.1 Deep Neural Networks

We refer readers to prior publications and reviews\textsuperscript{30-32} that documents the theory of deep neural networks. Here, we will briefly summarize the high-level concepts that are necessary for understanding deep neural networks. Deep learning algorithms are a class of machine learning algorithms used to model and analyze large complex datasets. The basic unit of these neural networks is a “neuron”, which are organized into layers. Each neuron in the network performs a non-linear transformation using rectified linear activation functions (ReLU)\textsuperscript{33} that converts the input data into an output value. A tunable parameter, the “weight” of each neuron’s function is adjusted in the development of the model to minimize the error of the predicted value, and this process is known as “training” the neural network.

During the training process, it is necessary to determine how to assign error attribution and make corrections to the network’s weights by working backward originating from the predicted output, a process known formally as “backpropagation”. Typically, a gradient descent algorithm is used to find the minimum in the error surface caused by each respective neuron. Conceptually, gradient descent is similar to the steepest descent algorithm used in classical molecular dynamics simulation. However, instead of iteratively minimizing an energy function and updating atomic coordinates for each step, a loss function of the DNN is iteratively minimized and the weights of the neurons are updated each step. Steps in the training process are also known as “iterations”, and the data in the training set may be iterated over multiple times, with a complete pass over the data being called an “epoch.”
2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a special type of neural network developed to handle image data. As illustrated in Figure S1a, a CNN is constructed from convolutional layers. Unlike the typical (multi-layer perceptron) neural networks discussed in the preceding paragraphs, instead of having every neuron in each layer connected to every neuron in the previous layer, each neuron in a convolutional layer (or “filter” as it is more commonly termed in the CNN literature) only receives input from a small, spatially contiguous window on the output of the previous layer. Furthermore, there are filters that receive windows across the entire input with shared weights, and this allows them to detect the same feature in different locations in an image. Thus, convolutional layers preserve spatial structure in data, such as the relative positions of pixels in images, or in the context of our images, the relative position of atoms and functional groups.
3. Methods

3.1 Dataset Description

Table 1: Summary of datasets used to evaluate the performance of Chemception

| Dataset | Property                  | Task                          | Size   |
|---------|---------------------------|-------------------------------|--------|
| Tox21   | Physiological: Toxicity   | Multi-task binary classification | 8014   |
| HIV     | Biochemical: Activity     | Single-task binary classification | 41,193 |
| FreeSolv| Physical: Free energy of solvation | Single-task regression      | 643    |

Following an identical approach as the original Chemception paper, we obtained several publically available datasets (Table 1) from the MoleculeNet benchmark database to evaluate the performance of Chemception. The first dataset, Tox21 is a collection of 12 different toxicity measurements for 8014 compounds, classified as either “toxic” or “non-toxic”. The second dataset, HIV, is a collection of the measured inhibition of HIV replication in vitro for 41,913 compounds, classified as either “active” or “inactive”. The last dataset, FreeSolv, has 643 compounds with the measured hydration free energies spanning the range from –25.5 to 3.4 kcal/mol. This dataset also included alchemical free energy calculations obtained from first-principles molecular dynamics simulations, and these were used as target comparisons for physics-based models.

3.2 Dataset Preprocessing

The dataset pre-processing steps are identical to that reported previously. We used a 5-fold cross validation protocol for training and evaluated the performance and early stopping criterion of the model using the validation set. We also included the performance on a separate test set as an indicator of generalizability. Specifically, for the Tox21 and HIV dataset, 1/6th of the database was separated out to form the test set, and for the Freesolv dataset, 1/10th of the database was used to form the test set. The remaining 5/6th or 9/10th of the dataset was then used in the random 5-fold cross validation approach for training Chemception. We also oversampled the
minority class to address the class imbalance in the dataset. This was achieved by computing the imbalance ratio and appending additional data from the minority class by that imbalance ratio, and this oversampling step was performed after stratification.

3.3 Data Preparation and Augmentation

Figure 1: After a SMILES to structure conversion, the 2D images are discretized and used to train a deep neural network. In this work, different image representations were used. Augmented images included additional atom and bond-specific chemical information in the image channels, and reduced information images had simplified information content encoded into a single image channel.

A summary of the end-to-end workflow of Chemception is illustrated in Figure 1. Briefly, we used open-source cheminformatics software, including OpenBabel$^{40}$, Pybel$^{41}$ and RDKit$^{42}$ through the Cinfony interface,$^{43}$ to convert SMILES strings to their respective 2D molecular structures, mapped the coordinates into a discretized image, which was used to train Chemception
in a supervised fashion. Specifically, we used RDKit\cite{2009_schombert} to convert the SMILES strings\cite{2002_smiles,2006_smiles} to their corresponding 2D molecular structure. The resulting coordinates of each molecule were then mapped onto an 80 x 80 grid, where each pixel had a resolution of 0.5 Å as illustrated in Figure 1. Under the current data preparation workflow, molecules that are longer than 40 Å (i.e. cannot fit into an 80 x 80 image) are not included in the final dataset. In this work, a number of augmented image formats and reduced information image formats were developed, and their details are outlined in the Results section.

During the training of Chemception, we performed additional real-time data augmentation to the image using the ImageDataGenerator function in the Keras API as reported previously.\cite{2021_cemencia} Briefly, each image was randomly rotated between 0 to 180 degrees before being parsed into Chemception. No additional operations were performed.

3.4 Convolutional Neural Network Design

We used the recently developed Chemception deep neural network architecture in this work.\cite{2021_cemencia} Chemception is a modified neural network that is optimized for handling sparse images of 2D representations of chemical structures, which incorporates the two major advances in CNN research: inception modules\cite{2015댕어} and residual learning.\cite{2016_he} Chemception’s high-level architectural design includes 6 segments. As illustrated in Figure S1b, this includes the stem layer, followed by a series of alternating Inception-Resnet and Reduction blocks, before reaching a global average pooling layer that leads directly to the final output layer. As illustrated in Figure S2, the inception and reduction blocks in Chemception shares a common reference design block. In each reference design block, there is a reference layer from which the number of convolutional filters of other layers are calculated. For classification problems (Tox21, HIV), a softmax layer was used as the output layer, and for regression problems (FreeSolv), a linear layer was used. For the results of
this paper, we only report the performance of the baseline Chemception T1_F32 model and the optimized Chemception T3_F16 model. For additional architectural design details, we refer our readers to the original Chemception paper.29

3.5 Chemception Training Protocol

Chemception was trained using a Tensorflow backend47 with GPU acceleration using NVIDIA CuDNN libraries.48 The network was created and executed using the Keras 1.2 functional API interface.49 Chemception was trained using a two-stage protocol. In the first stage, we use the RMSprop algorithm50 for 50 epochs using the standard settings recommended (learning rate = 10^{-3}, \rho = 0.9, \varepsilon = 10^{-8}). A second fine-tuning stage was then performed, using the stochastic gradient descent (SGD) algorithm with momentum for another 50 epochs, using an initial learning rate of 10^{-3} with an exponential learning rate decay mapped using the following function:

\[ lr = lr_{ini} \times \gamma^{epoch} \]

where \( lr \) denotes the current learning rate, \( lr_{ini} \) denotes the initial learning rate, \( \gamma \) is an empirical scaling factor that is set to 0.92, and epoch denotes the current epoch. We used a batch size of 32, and also included an early stopping protocol to reduce overfitting. This was done by monitoring the loss of the validation set, and if there was no improvement in the validation loss after 25 epochs, the last best model as evaluated by the validation loss was saved, and this was used for the second fine tuning stage of training or saved as the final model.

Lastly, for the Tox21 and HIV dataset, the evaluation metric reported in our paper is the area under the ROC-curve (AUC). For the FreeSolv dataset, the evaluation metric is RMSE. The reported results in the paper are the mean value and standard deviation of the performance metric, obtained from the 5 runs in the 5-fold cross validation protocol.
4. Results and Discussion

4.1 Design of Augmented and Reduced Chemical Image Representations

In the original Chemception paper, we used single-channel greyscale images to depict 2D drawings of various molecules. A number corresponding to the atomic number of the element was used as a unique identifier and to distinguish atoms from empty space, which was encoded as “0”. In addition, bonds were identified as “2”, which would normally correspond to Helium, but since this element was not present it served as a unique identifier for representing bonds. Apart from this data, no additional information was encoded in the channels of the image.

Figure 2: Illustration of the structure of a 4-channel CMYK image. Instead of encoding color information into the image channels, atom and bond-specific chemical information are encoded in this augmented image representation.

In computer vision research using deep neural networks, colored images are typically used in the training data. Instead of having a single channel like in a greyscale image, colored images may have 3 or 4 channels. As illustrated in Figure 2, an image with N channels may be perceived as a collection of pixels in the x and y-dimension (i.e. the image), with an additional z-dimension
that contains $N$ elements carrying the channel information. A colored image, therefore, encodes information relevant to color in its channel, using a 3-channel RGB or 4-channel CMYK color model. Adopting this principle of encoding additional information into image channels, for our work, we have encoded chemistry specific information into a 4-channel image format to create a set of augmented image representation. We anticipate that the addition of appropriate atom and/or bond-specific chemistry information would thus accelerate Chemception training, as this information can be used directly by the deep neural network, and it does not need to learn additional representations to describe the information that is already provided in the channels. We also hypothesize that benefits of training on augmented images will be useful in the small labeled data regime, where there may be insufficient data for the neural network to formulate optimal representations and features. However, it should be emphasized that the additional information encoded will only be helpful, so as long as they are relevant to the task to be predicted – either through direct correlation or through the formation of higher-level representations that correlates with the task.

In this work, our goal is to develop a general-purpose format for encoding chemical images using the following design principles: (i) it should be quick to compute and scalable, and (ii) it should only encode basic chemical information. The rationale for the first principle is obvious, as we would want this framework to be scalable to millions of compounds for future big data chemistry research. Keeping scalability considerations in mind, and compatibility with existing image analysis tools, we therefore limited the maximum number of channels to 4, which is analogous to 4-channel CMYK images. In addition, since the information encoded would need to be computed for every new molecule, and with speed considerations in mind, we have also excluded chemical information that is computed from QM methods. For the second design
principle, we hypothesize that it would be more effective and generalizable to provide the network with the building blocks to construct more complex features, and thus we favor the selection of basic chemical information.

Based on these two design principles, we selected several computable properties for encoding, which includes: bond order, hybridization, valency, and partial charge. Technically, information pertaining to bond order can be inferred from the general topology of the 2D structure, but this representation requires sufficient data for deep neural networks to learn, and therefore there may still be benefits in encoding it directly. For example, C=C bonds which are more frequently observed in a typical small molecule database may be more easily learned, but more exotic bonds such as those found in a nitro (NO$_3^-$) group may not be learned optimally. Hybridization in this context refers to the type of hybridization (sp, sp$^2$, sp$^3$, etc.) of the atom, and valence refers to the number of explicit connections an atom is bonded to. When combined together in the appropriate context, an atom’s hybridization and valence can lead to inferences about bond order, lone pairs, and other related basic information. Lastly, partial charge was computed by the methods developed by Gasteiger et. al.$^{51}$ which represents one of the more basic electrostatics descriptions of an atom that can be computed quickly, from which more sophisticated representations may be developed by the neural network. Using these basic computable properties, several permutations of augmented image representations were evaluated. As summarized in Table 2, the schema for EngA, EngB, and EngC provides a different channel for atomic identity (encoded as atomic number) and bond identity (encoded as bond order) information. EngD compresses both atomic and bond identity information into a single channel (as per the original standard image used in Chemception), and adds all 3 atom-specific information into the remaining channels.
Next, we also explored alternative single-channel image representations to infer the nature and type of chemical knowledge Chemception has used or learned from its training data. For this approach, we start with the default single-channel standard image used in the original Chemception paper (denoted as Standard), and systematically reduced information content from this channel. In the first level of simplification, RedB, all atoms are now identical and assigned “1”, while bonds are assigned “2”, which effectively gives an image of atoms and bond information only, where the identity of atoms and bonds are not distinguishable. In RedA, bonds are removed from the images and only atoms remain. Such an image representation may be analogous to image data obtained from X-ray crystallography diffraction images of small molecules where only atomic positions can be obtained.

**Table 2:** Summary of the schema of various image representations evaluated

| Representation | Channels | Description                                                   |
|----------------|----------|---------------------------------------------------------------|
| Standard       | 1        | Atomic number + bond (“2” for bond)                          |
| RedA           | 1        | “1” for atom                                                 |
| RedB           | 1        | “1” for atom, “2” for bond                                   |
| EngA           | 4        | Atomic Number, Bond Order, Partial Charge, Hybridization       |
| EngB           | 4        | Atomic Number, Bond Order, Partial Charge, Valence            |
| EngC           | 4        | Atomic Number, Bond Order, Valence, Hybridization             |
| EngD           | 4        | Atomic number + bond (“2” for bond), Partial Charge, Valence, Hybridization |
| Noise          | 1        | Random scattering of “1” in null “0” space                   |
| Truth          | 1        | “1” for active, “0” for inactive, only for atom/bond pixels  |
| Scrambled      | 1        | Unique random numbers for each bond/atom pixel              |
Lastly, we devised a Noise representation, which is a random dispersion of noise values (“1”) in an empty (“0”) image. This representation was intentionally constructed to be not-a-molecule, thus serving as a negative control to determine if Chemception was making predictions from noise. In a similar fashion, we constructed a Truth representation to serve as positive control. In this representation, we systematically assigned “1” for every atom and bond pixel if the molecule’s property is active and “0” for every atom and bond pixel if the molecule’s property is inactive, and the empty (“0”) background is retained. Based on the results of our work, it also motivated the development of additional representations. Namely, the Scrambled representation was developed to test for the importance of chemical periodicity. In this representation, each atom and bond is assigned a random but unique number, ensuring that they can still be uniquely identified, but information about chemical periodicity which may be inferred from atomic number information is not available.

4.2 Tox21 Results

![Graph](image)

**Figure 3:** Chemception performance on Tox21 toxicity prediction when trained on various reduced and augmented image representations

We trained the baseline Chemception T1_F32 model and optimized Chemception T3_F16 model on the Tox21 dataset using both reduced and augmented images and the results are
summarized in **Figure 3** and **Table S1**. For brevity, we only included the mean AUC metrics across all 12 measurements in the Tox21 database. Our previous best model achieved a validation/test AUC of 0.768/0.773 using the standard image representation. Augmenting image channels with basic chemical information improved the performance of validation AUC to the range of 0.776 to 0.796. The best performance achieved was a validation/test AUC of 0.796/0.798 for the optimized Chemception model and 0.791/0.796 for the baseline Chemception model using the EngA image representation. We also observed that other engineered representation (EngB, EngC, EngD) achieved comparable performance, albeit in the 0.78 to 0.79 range. All engineered representations were better than the standard representation which indicates that the additional information provided was used by the model to improve its performance.

Next, we examine the performance of the images that had reduced chemical information content. The RedA image representation may be considered as a precursor to a proper molecular drawing, as it only has positions of atoms, and has no bonds. However, one has to recall that a bond is an artificial construct introduced to denote the linkages between various atoms, and a key role it plays in chemistry is to make it easier for chemists to use the notion of a bond to formulate more advanced concepts. When trained on a set of RedA images, both Chemception T1_F32 and Chemception T3_F16 achieved similar validation/test AUC of 0.717/0.713 and 0.716/0.716 respectively. The fact that the AUC is not at 0.5 indicates the neural network has learned some chemistry to distinguish between toxic and non-toxic molecules just from the relative positions of generic indistinguishable atoms. As we add more chemical information to the RedB image representation, where bonds are explicitly indicated in the image, we observed a marginal improvement to a validation/test AUC 0.724/0.731 and 0.726/0.726 for each respective model.
This result indicates that explicit knowledge of bonds was apparently not the most important requirement for determining molecular toxicity.

Lastly, we examine the performance of the control set of images. We note that the images used in training Chemception are extremely sparse compared to natural images used in typical computer vision research; this means that typically less than 10% of the image has usable information (since the other 90% is empty), and it is likely that representations learned by the neural network would have to identify relevant data in a sub-1% portion of the image. Due to this peculiar image characteristic, a significant concern is that the neural network learned representations might not be robust, as a random scattering of pixels may not look that different from a proper drawing of a molecule. In the context of our work, less robust learning would mean that the model’s performance when trained on RedA/RedB images may be a consequence of learning from noise. Based on the results of the noise image, we observed that the validation/test AUC achieved was 0.5/0.5. An AUC of 0.5 means that the model has no predictive power. At the opposite spectrum, training Chemception on the truth images resulted in a validation/test AUC achieved of 1.0/1.0. Both sets of control experiments verify that Chemception is not learning “something from nothing”, and also confirms the ability of the neural network to extract out relevant information, even from an extremely sparse image. Based on these observations, we infer that Chemception is likely using the chemical structure as presented in the images in predicting molecular toxicity.

4.3 HIV and FreeSolv Results

Having demonstrated that Chemception, when trained on augmented images, can lead to increased model performance, we now address the question of whether this effect can be generalized to other chemical properties. In this section, we examine model performance on
predicting HIV activity, a biochemical property, and free energy of solvation, a physical property. As these are separate types of chemical properties, the additional information encoded may not necessarily provide a consistent level of performance improvement.

**Figure 4:** Chemception performance on HIV activity prediction when trained on various reduced and augmented image representations

The results of the HIV activity prediction is summarized in Figure 4 and Table S2. Our previous best model achieved a validation/test AUC of 0.745/0.744, which was achieved using the standard image representation. Augmenting image channels with basic chemical information improved the performance of validation AUC to the range of 0.744 to 0.762. The best performance achieved was a validation/test AUC of 0.762/0.777 for the baseline Chemception model and 0.757/0.773 for the optimized Chemception model using the EngC image representation. Similar performance metrics were observed for other engineered representations (EngA, EngB, EngD) that achieved a comparable performance in the 0.75 range. In addition, the best representation for toxicity prediction (EngA) was not the best representation for activity prediction (EngC), although we observed that the difference in AUC metrics is minimal. As before, using augmented images provided consistently better results than using standard images, but the performance improvement was not as much as in the toxicity predictions. Our results of reducing chemical information for
activity prediction also parallels that observed for toxicity prediction. As one would expect, the noise image representation had a validation/test AUC of 0.5/0.5, and increasing information content from RedA to RedB saw a gradual improvement from 0.660/0.672 to 0.698/0.730 for the baseline Chemception model, and a similar trend was observed for the optimized Chemception model.

![Image: Chemception performance on free energy of solvation prediction when trained on various reduced and augmented image representations](image)

**Figure 5:** Chemception performance on free energy of solvation prediction when trained on various reduced and augmented image representations

The last dataset we will examine is the FreeSolv dataset, which is a much smaller dataset (600) on the prediction of solvation free energies. Our previous work on Chemception trained on standard images achieved a validation/test RMSE of 1.51/1.75 kcal/mol. As summarized in **Figure 5** and **Table S3**, augmenting the image with chemical information substantially improved the model’s performance. The best performance achieved was a validation/test RMSE of 1.16/1.27 for the baseline Chemception model and 1.21/1.24 kcal/mol for the optimized Chemception model using the EngD image representation. In addition, the results from EngA are similar to that achieved with EngD with RMSE metrics in the 1.1 to 1.2 kcal/mol range, but we noted EngB and EngC did comparably worse with RMSE metrics in the 1.3 to 1.4 kcal/mol range.
Due to the nonlinear distributed representations a neural network develops, and the difficulty in interpreting such “black box” algorithms, it is not possible to state with certainty the reasons for this observed performance difference in the prediction of solvation free energies, as all 4 augmented image representations have performed comparably to one another in both toxicity and activity predictions. A cursory examination indicates that EngA and EngD both had partial charge and hybridization information added, but EngB and EngC do not. Using chemistry intuition, it is not unexpected that partial charge would be relevant information to predict solvation free energies as it could serve as a proxy for the electrostatic nature of the atom and its local environment.

Next, we examine the images with reduced chemical information. The noise image representation attains a validation/test RMSE of 3.89/3.30 kcal/mol. From our earlier observations of Chemception models trained on noise images in toxicity and activity, this has resulted in a model of no predictive power, and as such the above-specified RMSE value may be a reasonable zero baseline in the context of solvation free energy predictions. Images with reduced chemical information were much less accurate in predicting solvation free energies. All of the reduced image representations (RedA, RedB) achieved RMSE metrics of 3.1 to 3.3 kcal/mol, which when compared to the 1.51/1.75 kcal/mol accuracy of standard images, are significantly worse and closer to the null results obtained from noise images. This performance trend is unlike that observed for the toxicity and activity predictions, where the reduced information images had accuracy that was more similar to standard images than noise images.
4.4 Augmented Chemception Performance Against Contemporary Models

Our results indicate that all of the 4 augmented image representations developed provided consistent performance improvement to Chemception’s accuracy relative to standard images. In addition, this improvement appears to be consistent across the 3 different datasets evaluated (although the extent of improvements does vary), which suggests generalizability of the augmented image representations.

In this section, we will compare the performance of augmented Chemception against other deep learning QSAR/QSPR models reported in the literature, both in terms of performance metrics as well as the amount of chemical information that is used and/or provided as input data. We used the results reported from the MoleculeNet benchmark released by Pande and co-workers. Specifically, we compare to the random data splitting performance of the multilayer perceptron (MLP) deep neural networks, that is trained using engineered features in the form of ECFP fingerprints as inputs. In addition, we have also included the results from a novel convolutional graph algorithm (ConvGraph). Instead of providing explicit chemistry features, the ConvGraph algorithm uses an engineered representation to depict the molecule, where the molecular structure is encoded as a graph, and the nodes are annotated with atom-specific information such as atomic identity, partial charge, and hybridization. Conceptually, the ConvGraph algorithm and augmented images reported in this work are similar, in the sense that both methods use an engineered representation of the data as opposed to explicit chemistry features such as molecular descriptors or fingerprints. The main difference is that our method encodes the molecule in an image format, whereas ConvGraph encodes it as a graph. In terms of pre-requisite chemistry knowledge, both engineered representations require substantially less than molecular fingerprints or descriptors, as only basic information such as hybridization and partial charge are provided as input data. However, the graph representation used in ConvGraph implicitly provides additional
information about the molecule’s topology by virtue of its data format, whereas in an image, a molecule’s topology needs to be learned through representations developed by the neural network. Based on this assessment, we order the level of chemical information in the following ascending order: reduced information images, standard images, augmented images, molecular graphs, molecular fingerprints, molecular descriptors.

**Figure 6:** Chemception performance relative to contemporary machine and deep learning models, and a novel convolutional graph algorithm for Tox21 toxicity prediction.

In **Figure 6** and **Table S4 to S6**, we summarize the best Chemception results trained on standard and augmented images. The best performing contemporary model is the multi-task MLP DNN with a validation/test AUC of 0.777/0.799, and this underperforms a single-task Chemception trained on augmented images that achieve a validation/test AUC of 0.796/0.798. There are several mitigating factors that should make it easier for MLP DNN model to outperform Chemception. First, the MLP DNN model was trained using engineered features (ECFP fingerprints), while Chemception even with augmented images only adds a few basic chemical information. Second, the MLP DNN model was trained in a multi-task setting, which has been previously reported to outperform single-task models.8,11 Despite these handicaps, the better performance of Chemception provides evidence that leveraging the representation learning ability of deep learning models will enhance and possibly reduce the necessity of manual feature
engineering of explicit chemistry features. When compared to convolutional graphs which achieved a validation/test AUC of 0.811/0.848, we observed that Chemception trails in performance. Next, we examine the performance on the HIV dataset. The original Chemception model with a validation/test AUC of 0.745/0.744 was already outperforming contemporary deep learning (MLP) models with a validation/test AUC of 0.715/0.732. Using augmented images did improve the performance to validation/test AUC of 0.762/0.777, but it still underperforms relative to the performance of ConvGraph which achieved a validation/test AUC of 0.809/0.798.

Lastly, we examine the results of the FreeSolv dataset where we observed the largest improvement in using augmented images. Unlike the toxicity and activity classification tasks, Chemception with augmented images which attain a validation/test RMSE of 1.17/1.22 kcal/mol, is the best performing model across all methods reported to date. This includes ConvGraph that achieved RMSE of ~1.3 kcal/mol and physics-based model obtained from free energy calculations that achieved RMSE of 1.5 kcal/mol. At its current level of performance, Chemception is approaching the gold standard ~1.0 kcal/mol chemical accuracy. Recent work on using deep learning models have also reported similar success in predicting other physical properties, particularly those that can be computed from QM-calculations,²¹,²² and achieving ~1.0 kcal/mol chemical accuracy is within the feasibility of deep neural networks, within the chemical space it was trained on. Furthermore, we note that current deep learning models for quantum chemistry properties require substantial pre-requisite chemical knowledge, as physical equations or restraints inspired by first-principles knowledge are used to constrain the search space of the neural network. In contrast, Chemception operates at a chemistry-agnostic level; it receives no guidance on the underlying equations that govern the properties it predicts, and yet the level of accuracy it achieves in predicting physical (free energy of solvation) properties is similar to that of the more
sophisticated deep learning models. Collectively, our results and earlier work by other groups,\textsuperscript{21,22} provides further validation that deep neural networks of various implementations are becoming a practically viable option for calculating physical properties, a problem that typically lies in the realm of first-principles simulations.

4.5 Inferring the Learning Patterns of a Deep Neural Network

Deep learning, like many other machine learning algorithms, is a “black box” algorithm, which has limited interpretability. It is not possible to determine what exactly a deep neural network learns as it is trained to predict toxicity, activity, and solvation free energies. Nevertheless, in our work, we have systematically removed and added information into the image channels, and together with the numbers reported by Pande and co-workers, we have observed a consistent pattern in the results that may help form a hypothesis on \textit{what} or \textit{how} Chemception learns. In this last section, we will make inferences from the learning patterns of Chemception.

We begin by summarizing several key observations. First, we note that for toxicity and activity prediction, the reduced image representation provide less predictive accuracy, but still resulted in Chemception models that were fairly predictive of their respective tasks. In contrast, for solvation free energy prediction, the reduced image representation resulted in an accuracy that is close to that when trained on noise images. The key difference between reduced images and standard/augmented images is the inclusion of atomic number information. From the atomic number, it is possible for the neural network to form representations related to periodic trends, which is a key concept in chemistry. Therefore, based on the difference in model accuracy improvement with increasing chemical information, toxicity/activity vs. free energy of solvation predictions appear to follow two distinct learning patterns.
The current approach in toxicity research involves the identification of toxic functional groups (toxicophore) as a proxy to predict molecular toxicity.\textsuperscript{52-54} For activity, current SAR studies are premised on finding the appropriate functional groups on the molecule that can interact with the binding pocket of the protein in an optimally aligned fashion.\textsuperscript{55,56} From the observed trends in Chemception accuracy on toxicity/activity modeling, it is evident that just position of atoms is sufficient to construct a model that is more accurate than a null model. At the RedA/B level, the only representations that one can reasonably learn are patterns of how atoms are arranged. Learning such representations would be analogous to the identification of functional groups. Therefore, this suggests that Chemception is operating primarily as a functional group identifier, and this approach is consistent with our existing understanding of toxicity and activity, through the identification of toxicophores and pharmacophores.

We now contrast this to solvation free energies, which unlike toxicity/activity prediction, is a physical property that can be computed from physics-based simulation methods. We note that the reduced image representations are almost as inaccurate as a model trained on noise images, and a significant improvement in accuracy is noted in standard images, where atomic number information is first introduced. This suggests that a functional group identifier learning mode is not sufficient to predict solvation free energies and atomic number is a key component. In addition, atomic number may serve as a proxy for related chemical information, such as the number of electrons (particularly for neutral compounds), atomic mass (assuming a consistent ratio of neutrons to protons, which holds approximately true for the lighter elements that are predominantly represented in the FreeSolv training set), and energy (via Einstein’s mass-energy equivalence). Contemporary physics-based simulation methods for calculating free energies incorporate some of the above-mentioned information. Therefore, we suggest that the significant improvement in
Chemception accuracy with the introduction of atomic mass information is an indicator that it is learning in a manner similar to established physics-based models.

To validate this hypothesis on the importance of atomic number information, we developed another image representation, “Scrambled” that contains almost the same amount of information as the standard images used. In this representation, atoms can be uniquely identified, but information about their periodicity is lost. Chemception models trained on scrambled images (see Table S1 to S3) achieved a validation/test AUC of ~0.74 for the Tox21 dataset and ~0.72 on the HIV dataset. Compared to the results of using standard images (AUC of ~0.77 for Tox21, ~0.74 for HIV) relative to noise images (AUC of ~0.5 for Tox21 and HIV), we observed that scrambling the atomic identities had little effect, which is expected if one is primarily learning to identify specific patterns of pixels in the image (i.e. functional groups). This result is in contrast to the performance of scrambled images on free energy of solvation prediction, which achieved a validation/test RMSE of 3.22/3.28 kcal/mol. Compared to the results of using standard images (RMSE of 1.51/1.71 kcal/mol) relative to noise images (RMSE of ~3.8/3.3 kcal/mol), the huge gap in performance indicates that Chemception is likely to be using the atomic number information in its calculation of solvation free energies.

Next, we identify another trend by comparing Chemception against the results from ConvGraph, where Chemception systematically underperforms in activity and toxicity prediction but outperforms in solvation free energy prediction. This is despite the fact that almost the same type of chemical information is being provided. For augmented Chemception, explicit information was added to the channels on information pertaining to hybridization, valance, bond-order, partial charge and atomic number. For convolutional graphs, the nodes were annotated with information pertaining to atom type (encoded as a one-hot), bond-type and graph distance. Given the molecular
graph format, valence can also be inferred directly from the connectivity to each atom. In addition, in the full featurization used by Pande and co-workers, it also encoded other properties into the graph such as chirality, formal charge, partial charge, ring size, hybridization, hydrogen bonding and aromaticity, and some of these properties were included in the channels of our augmented image representation.

Our earlier hypothesis that Chemception operates as a functional group identifier for activity and toxicity prediction, explains the apparent underperformance relative to the ConvGraph algorithm. This is because when the molecule is represented as a graph, its topology is implicit to how the data is being presented. In contrast, for an image, representations will need to be developed to understand molecular topology. Therefore, it would be easier for graph-based neural networks to identify topological patterns in the data, which will facilitate functional group identification. Next, we observed that the ConvGraph algorithm encoded atomic identity as a one-hot vector, as opposed to our work that used atomic number directly. This means that periodicity will be difficult to infer. Our earlier hypothesis that atomic number enables Chemception to learn similarly to physics-based models, explains the apparent underperformance of the ConvGraph algorithm relative to Chemception. A logical prediction of this hypothesis is, if the ConvGraph could be amended to include atomic number information instead of one-hot encoded atomic identity, it should boost the resulting model’s accuracy in predicting solvation free energies.
4.6 How Much Chemistry Does a Neural Network Need to Know?

In this final section, we address the underlying question behind this work: How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions? Based on our findings, only basic chemical information is necessary for a deep neural network to predict complex chemical properties like toxicity, activity, and solvation free energy. Sophisticated chemistry knowledge relevant to the task, such as providing toxicophores identification for toxicity prediction, is evidently not a pre-requisite for deep neural networks to achieve accurate predictions. In addition, we have identified that atomic number information is critical for predicting solvation free energies, and potentially other physical properties that can be computed from first-principles simulation methods. By using augmented image representations with no more than 3 additional basic chemical information like valence, hybridization, and partial charge has allowed Chemception to consistently outperform contemporary deep learning models (MLP) trained on more advanced engineered features like molecular fingerprints. We also observed the high efficiency at which Chemception uses relevant information, as only less than 1% of the input data was altered between image representations.

Lastly, our research into varying chemical information content and observing the resulting model’s accuracy, and comparisons between Chemception and ConvGraph have indicated that Chemception may be learning about its tasks in the manner that is consistent with established knowledge about toxicity/activity and solvation free energy. Our work suggests that using deep neural networks can be a viable first approach for a novel chemistry problem, and could serve as to provide a baseline model from which to work on. The corollary to this statement is that, if a well-trained neural network provided with sufficient data cannot achieve sufficient accuracy, it is likely that the data provided is not entirely relevant for the task being predicted.
5. Conclusion

In conclusion, we have developed a series of augmented image representations for use in training Chemception, a novel deep convolutional neural network that predicts general chemical properties using only image data of 2D drawings of molecules. By providing additional basic chemical information into the image channels, such as partial charge and hybridization, it has improved the accuracy of Chemception, where it now achieves consistent outperformance against contemporary deep learning models that are trained on explicit chemistry features, such as ECFP fingerprints. Specifically, Chemception achieved a validation/test AUC of 0.796/0.798 for Tox21 toxicity prediction, 0.762/0.777 for HIV activity prediction and a validation/test RMSE of 1.17/1.22 kcal/mol for free energy of solvation prediction. Furthermore, by altering the chemical information content provided in the images, and comparing our results to a convolutional graph algorithm that uses a similar data representation, we were able to identify two distinct learning patterns in Chemception that appears to be consistent with established chemistry knowledge. Specifically, we infer that Chemception learns toxicity and activity primarily as a functional group identifier, which is analogous to current research in identifying functional groups responsible for toxicity (i.e. toxicophores) or functional groups that form complementary interactions with the binding pocket of a protein for predicting activity (i.e. pharmacophores). For solvation free energy, we show that atomic number is necessary for Chemception to achieve predictive accuracy, and this importance of chemical periodicity parallels the underlying principles behind current physics-based simulations. In addition, our work also demonstrates the robustness of Chemception to noise and the high efficiency at which Chemception can extract relevant data even from a sparse representation. Our findings suggest that Chemception will be a valuable tool in the future of deep learning assisted computational chemistry research in a data-driven research paradigm, where it
can be used as a first-pass approach to organize and order information in big data chemistry research problems.
Acknowledgement

This research was funded by the Pacific Northwest National Laboratory (PNNL), Laboratory Directed Research Development (LDRD) Program for the Linus Pauling Distinguished Postdoctoral Fellowship. Additional funding was provided through PNNL LDRD Deep Learning for Scientific Discovery Initiative, and DOE ASCR grant 69696. Computing resources were provided by the XSEDE supercomputing grant CHE170015 and PNNL PIC computing resources.
References

(1) He, K.; Zhang, X.; Ren, S.; Sun, J. arXiv:1502.01852 2015.
(2) Ioffe, S.; Szegedy, C. arXiv:1502.03167 2015.
(3) Baldi, P.; Sadowski, P.; Whiteson, D. Nat. Commun. 2014, 5, 4308.
(4) Charles Siegel; Jeff Daily; Vishnu, A. arXiv:1610.00790 2016.
(5) Nikhil Mukund; Sheelu Abraham; Shivaraj Kandhasamy; Sanjnit Mitra; Philip, N. S. arXiv:1609.07259 2017.
(6) Chicco, D.; Sadowski, P.; Baldi, P. Proc. of the 5th ACM Conf. on Bioinf. Comput. Biol. and Health Inf. 2014, 533.
(7) Weihua Guo; You Xu; Feng, X. arXiv:1705.03094 2017.
(8) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. arXiv:1406.1231 2014.
(9) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. J. Chem. Inf. Model. 2015, 55, 263.
(10) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. arXiv:1502.02072 2015.
(11) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Ceulemans, H.; Wegner, J.; Hochreiter, S. Conference Neural Information Processing Systems Foundation (NIPS 2014) 2014.
(12) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. Front. Env. Sci. 2016, 3, 1.
(13) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. J. Chem. Inf. Model. 2015, 55, 2085.
(14) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. ACS Cent. Sci. 2016, 2, 529.
(15) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. ACS Cent. Sci. 2015, 1, 168.
(16) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Chem. Res. Toxicol. 2015, 28, 797.
(17) Lusci, A.; Pollastri, G.; Baldi, P. J. Chem. Inf. Model. 2013, 53, 1563.
(18) Kearnes, S.; Goldman, B.; Pande, V. arXiv:1606.08793 2016.
(19) Wallach, I.; Dzamba, M.; Heifets, A. arXiv:1510.02855 2016.
(20) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. New J. Phys. 2013, 15, 095003.
(21) Smith, J. S.; Isayev, O.; Roitberg, A. E. Chem Sci 2017, 8, 3192.
(22) Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A. Nat Commun 2017, 8, 13890.
(23) Goh, G. B.; Hodas, N. O.; Vishnu, A. Journal of computational chemistry 2017, 38, 1291.
(24) Gawehn, E.; Hiss, J. A.; Schneider, G. Mol. Inf. 2016, 35, 3.
(25) Lowe, D. G. Proceedings of the International Conference on Computer Vision 1999, 1150.
(26) David Duvenaud; Dougal Maclaurin; Jorge Aguilera-Iparraguirre; Rafael Gómez-Bombarelli; Timothy Hirzel; Alán Aspuru-Guzik; Adams, R. P. arXiv:1509.09292 2015.
(27) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Journal of computer-aided molecular design 2016, 30, 595.
(28) Rafael Gómez-Bombarelli; David Duvenaud; José Miguel Hernández-Lobato; Jorge Aguilera-Iparraguirre; Timothy D. Hirzel; Ryan P. Adams; Aspuru-Guzik, A. arXiv:1610.02415 2016.
(29) Garrett B. Goh; Charles Siegel; Abhinav Vishnu; Hodas, N. O.; Baker, N. manuscript in preparation 2017.
(30) Bengio, Y.; Courville, A.; Vincent, P. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798.
(31) Schmidhuber, J. Neural Netw. 2015, 61, 85.
(32) Arel, I.; Rose, D. C.; Karnowski, T. P. IEEE Comput. Intell. M. 2010, 5, 13.
(33) Glorot, X.; Bordes, A.; Bengio, Y. Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics (AISTATS) 2011.
(34) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. arXiv:1409.4842 2014.
(35) He, K.; Zhang, X.; Ren, S.; Sun, J. arXiv:1512.03385 2015.
(36) Zhenqin Wu; Bharath Ramsundar; Evan N. Feinberg; Joseph Gomes; Caleb Geniesse; Aneesh S. Pappu; Karl Leswing; Pande, V. arXiv:1703.00564 2017.
(37) Huang, R.; Sakamuru, S.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Houck, K. A.; Casey, W.; Hsieh, J. H.; Shockley, K. R.; Ceger, P.; Fostel, J.; Witt, K. L.; Tong, W.; Rotroff, D. M.; Zhao, T.; Shinn, P.; Simeonov, A.; Dix, D. J.; Austin, C. P.; Kavlock, R. J.; Tice, R. R.; Xia, M. Sci. Rep. 2014, 4, 5664.
(38) Data, A. A. S.
(39) Mobley, D. L.; Guthrie, J. P. Journal of computer-aided molecular design 2014, 28, 711.
(40) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. J Cheminform 2011, 3, 33.
(41) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Chem Cent J 2008, 2, 5.
(42) RDKit, O.-S. C.
(43) O'Boyle, N. M.; Hutchison, G. R. Chem Cent J 2008, 2, 24.
(44) Weininger, D. J Chem Inf Comp Sci 1988, 28, 31.
(45) Weininger, D.; Weininger, A.; Weininger, J. L. J Chem Inf Comp Sci 1989, 29, 97.
(46) Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. arXiv:1602.07261 2016.
(47) Martin Abadi, e. a. arXiv:1605.08695 2016.
(48) Sharan Chetlur; Cliff Woolley; Philippe Vandermersch; Jonathan Cohen; John Tran; Bryan Catanzaro; Shelhamer, E. arXiv:1410.0759 2014.
(49) Chollet, F. https://github.com/fchollet/keras 2015.
(50) Tieleman, T.; Hinton, G. COURSERA: Neural Networks for Machine Learning 2012.
(51) Gasteiger, J.; Marsili, M. Tetrahedron 1980, 36, 3219.
(52) Williams, D. P.; Naisbitt, D. J. Curr Opin Drug Di De 2002, 5, 104.
(53) Schultz, T. W.; Cronin, M. T. D.; Walker, J. D.; Aptula, A. O. J Mol Struc-Theochem 2003, 622, 1.
(54) Muster, W. G.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Muller, L.; Pahler, A. Drug Discov Today 2008, 13, 303.
(55) Mason, J. S.; Good, A. C.; Martin, E. J. Curr Pharm Design 2001, 7, 567.
(56) Yang, S. Y. Drug Discov Today 2010, 15, 444.