Beyond Atoms and Bonds: Contextual Explainability via Molecular Graphical Depictions

Marco Bertolini[a], Linlin Zhao[b], Djork-Arné Clevert[a] & Floriane Montanari[a]

[a] Machine Learning Research, Bayer AG, 13353 Berlin, Germany
[b] Field Solutions, Bayer AG, 40789 Monheim am Rhein, Germany
{marco.bertolini,linlin.zhao1}@bayer.com
{djork-arne.clevert,floriane.montanari}@bayer.com

Abstract

The field of explainable AI applied to molecular property prediction models has often been reduced to deriving atomic contributions. This has impaired the interpretability of such models, as chemists rather think in terms of larger, chemically meaningful structures, which often do not simply reduce to the sum of their atomic constituents. We develop an explanatory framework yielding both local as well as more complex structural attributions. We derive such contextual explanations in pixel space, exploiting the property that a molecule is not merely encoded through a collection of atoms and bonds, as is the case for string- or graph-based approaches. We provide evidence that the proposed explanation method satisfies desirable properties, namely sparsity and invariance with respect to the molecule's symmetries.

1 Introduction

The rapid development of Deep Learning (DL) models for molecular property prediction [1, 2, 3] has increased the need for equally powerful interpretability methods. These are crucial to gain trust in the model, understand its limitations, and support the chemist's knowledge and intuition in the process of property optimization. An ideal explainable AI (XAI) framework for molecular property prediction would assign attributions to both individual atoms and larger substructures. Additionally, it would also be able to provide ideas of modifications that can be made to the structure to overcome a particular issue.

Common modeling strategies involve fully connected networks from pre-computed molecular fingerprints [4] or latent representations and, when enough training data is available, end-to-end training with graph convolutional networks (GCNs) [5, 6]. Explanations for these types of models can take the form of atomic attributions (particularly for GCNs [7, 8]), or feature importance using packages such as SHAP [9]. Many of our in-house models are built upon the CDDD embedding space [10], which poses a challenge for explainability. The CDDD space is the bottleneck layer of a pre-trained autoencoder translating between different SMILES representations of molecules. One approach to explainability consists of assigning the attributions to the original SMILES, i.e., tracing back gradients through the pre-trained encoder. However, the interpretation and visualization of attributions for string characters is challenging [11]. Additionally, the validity of the use of gradients for discrete character inputs can also be questioned [12].

In this work, we propose a novel XAI approach tailored to networks built upon CDDD descriptors. This method, which we refer to as *contextual explainability*, is able to capture both atomic and structural contributions. We rely on explainability of concepts derived in the context of image analysis [13, 14, 15, 16, 17, 18] as well as on Img2Mol, a recently published optical molecular recognition model [19], that is able to translate images of molecules to their CDDD embeddings. We find that early layers in Img2Mol capture basic chemical features like atoms and bonds, while deeper layers learn more complex chemical structures, for instance, rings. By aggregating explanations from all the layers, we show that we can provide sparse and robust explanations that respect molecular symmetry and show both very localized highlights for particular atoms and more global importance for entire substructures.

ELLIS Machine Learning for Molecule Discovery Workshop 2021. https://moleculediscovery.github.io/workshop2021/

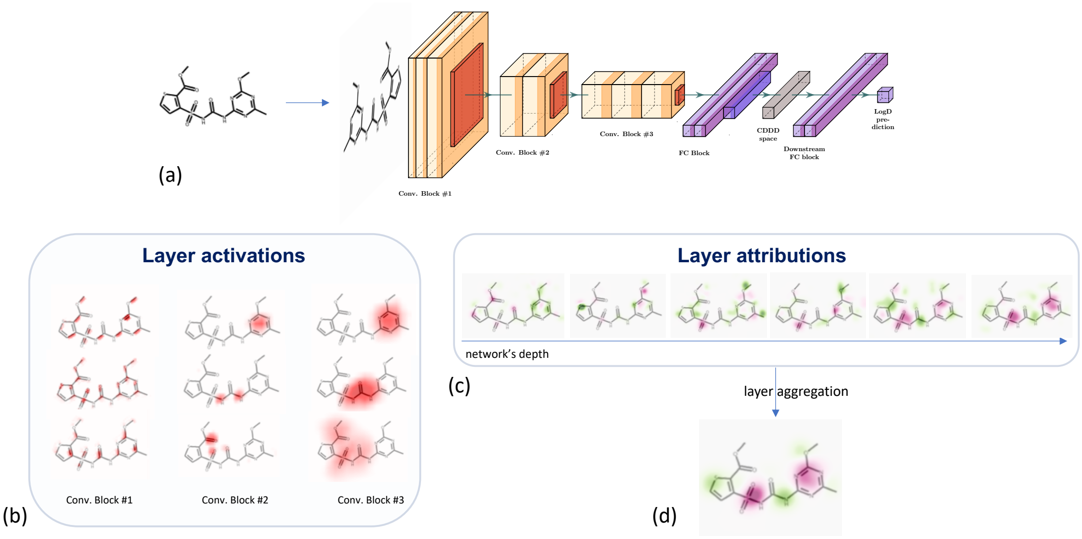

Figure 1: Summary of the contextual explanation framework for molecular property predictions: (a) Architecture of the image-based QSAR model; (b) channel layer activations learned by Img2Mol; (c) layer attributions; (d) contextual explanation obtained by aggregating over the various layer attributions. Green (pink) overlay indicates positive (negative) contribution towards the prediction.

2 Setup

Our explainability framework relies on the recently proposed Img2Mol model. It consists of a convolutional neural network whose task is to map molecular graphical depictions to their CDDD embeddings. The CDDD space $\mathcal{C} = [-1, 1]^{512}$ is constructed as the bottleneck layer of a Seq2Seq-autoencoder network trained to translate several million chemically-equivalent SMILES representations of molecules and defines a continuous molecular descriptor, which can be utilized as a powerful input for training downstream tasks. Figure 1a depicts the structure of the Img2Mol encoder. Img2Mol is trained on over ten million unique canonical SMILES and establishes the new state-of-the-art performance in reconstruction accuracy. The training objective consists in minimizing the distance in CDDD space between the Img2Mol embeddings and the embeddings obtained through the encoder from [10]. The reconstruction from CDDD to SMILES to evaluate the model's performance occurs through the pre-trained decoder from [10]. For further details concerning the model architecture, as well as the training procedure and the model performance, we refer the reader to [19].

We trained a quantitative structure-property relationship (QSAR) model to predict the lipophilicity of small molecules. The dataset consists of $\sim$63000 molecules with measured values in an in-house logD assay. Specifically, the downstream model is a multilayer perceptron (MLP) with two hidden layers and has been trained on the molecules' CDDD embeddings. The model performance is excellent with a cluster cross-validation coefficient of determination (R^2) score of 0.902. Upon testing on an independent dataset of 62 molecules, whose endpoints have been reviewed and curated from the Pesticide Properties Database [20], the final model led to an R^2 score of 0.914. All the XAI experiments and examples presented in this work are obtained from this final model, where the CDDD embeddings are generated through the Img2Mol encoder network. All the used input molecules are obtained from public data.

3 Methods

Our strategy is based on the fact that 1) deep layers in neural networks learn high-level concepts and 2) for pure convolutional networks, the value of each "superpixel" is determined by its receptive field in input space [21]. We combine these two properties by tracing back attributions to pixel space through the Img2Mol encoder instead of the original CDDD encoder. We remark that this is possible since both encoders map the respective inputs to the CDDD space. Explicitly, let $\Lambda : \mathcal{C} \to \mathbb{R}$ be the QSAR downstream model and Img2Mol : $\mathcal{M} \to \mathcal{C}$, where $\mathcal{M} \simeq [0, 255]^{224\times 224}$ is the input space consisting of images of 224×224 pixels. We then construct the network $\Phi = \text{Img2Mol} \circ \Lambda : \mathcal{M} \xrightarrow{\psi_p} \mathcal{M}_p \xrightarrow{\xi_p} \mathcal{C} \xrightarrow{\Lambda} \mathbb{R}$ by concatenating the Img2Mol encoder with the logD downstream network described in the previous section. Here, $\mathcal{M}_p$ is the output space of the p^{th} convolutional layer in the network. $\mathcal{M}_p$ has dimension $k_p \times k_p \times C_p$, where C_p is the number of channels in the p^{th} layer, and k_p is the size length of the embedding in terms of superpixels. Thus, our contextual explanations are obtained via the network Φ applied to a graphical depiction of the sample molecule.

Figure 1b shows a few channels activations, further grouped by the corresponding convolutional block. This example supports our intuition: while filters in early layers reduce to node, angle, and edge detectors, filters in deeper layers are activated by larger sub-structures in the molecule, e.g., rings and functional groups. It is then natural to use these layer attributions as a chemically meaningful feature basis for our explanations. Thus, we compute feature attributions values for each convolutional layer of the network Φ, choosing gradients to measure feature importance. Explicitly, for each convolutional layer p we compute superpixel attributions as

$$a_p(\mathbf{x}) = \sum_{c_p=1}^{C_p} \frac{\partial(\xi_{p,c_p} \circ \Lambda)(\mathbf{x})}{\partial\psi_{p,c_p}(\mathbf{x})} \times \psi_{p,c_p}(\mathbf{x}) \; , \tag{1}$$

where the sum is over the channel dimension. The above formula formalizes our intuition: for a given convolutional layer p, each channel output activation $\psi_{p,c_p}(\mathbf{x})$ is weighted by its contribution $\partial(\xi_{p,c_p} \circ \Lambda)(\mathbf{x})/\partial\psi_{p,c_p}(\mathbf{x})$ to the endpoint prediction. The attribution method (1) is known as activation×gradient, which is a natural extension of input×gradient [22] to obtain layer-wise attributions. Our implementation of (1) is based on the `captum` package [23]. Figure 1c depicts some layer-wise explanations. We notice that attributions for early layers, as expected, focus on simple geometric features like atoms and bonds, in contrast to attributions for deeper layers, where explanations involve entire functional groups.

Finally, as we assume that an exhaustive explanation would involve a combination of both local and structural features, we propose a simple procedure to extract a single explanation from the layer attribution maps. Namely, we aggregate the maps (1) over all the convolutional layers to obtain a unique network-wide attribution map

$$a(\mathbf{x}) = \sum_{p \in \{\text{conv.layers}\}} a_p(\mathbf{x}) \; . \tag{2}$$

The above equation determines the weighting of the various local and structural components, as determined by the relative value of the different layer attributions, resulting in the final contextual attribution map. Figure 1d illustrates an example of the result of such an aggregation strategy.

4 Experiments and Properties of Contextual Explanations

In this section, we turn to examine some of the desired properties that our contextual explanations (2) possess.

Contextual explanations and sparsity. The example in Figure 2a illustrates the defining characteristic of our approach. The attribution heat map consists of both atomic and structural features. Specifically, the model assigns positive contributions (green overlay) to the outmost `Cl` atom and methyl group, while it assigns negative contributions (pink overlay) to the central `N` atom and the triazine ring. These assignments are in alignment with a medicinal chemist's intuition about logD contributions.

The aggregation procedure (2) has, in addition, a denoising effect. As can be seen in Figure 1d, the aggregated map is more sparse than the individual layer attribution maps, as it concentrates only on the most important features contributing to the prediction. Sparsity is a desirable property for an explanation, as feature cluttering impairs the interpretability of the model.

Invariance with respect to molecule's symmetries. An important property for interpretability is that the explanations respect the symmetries of the input molecule. Among the CDDD-based methods, those based on SMILES will fail to produce invariant explanations, as the SMILES string representations explicitly break the molecule's symmetries. In what follows, we provide evidence that our contextual explanations, instead, tend to be invariant under such symmetries. Explicitly, let $\mathcal{T}$ be the symmetry group of a molecule's graphical depiction, that is, the group of image transformations that leave the chemical content invariant: given a molecule image $\mathbf{x}$ and a transformation $T \in \mathcal{T}$, then $\mathbf{x}' = T(\mathbf{x})$ corresponds to the same molecule. To quantify the invariance of our contextual explanations with respect to a symmetry group $\mathcal{T}$, we define the symmetry score for the transformation $T \in \mathcal{T}$ as

$$s_T(\mathbf{x}) = \frac{1}{2}\overline{|\widehat{a}(T(\mathbf{x})) - T(\widehat{a}(\mathbf{x}))|} \; , \tag{3}$$

where $\widehat{a}$ is obtained from (2) upon normalization to the range $[-1, 1]$, and the overline denotes the average in pixel space. The score measures the average absolute difference between two attribution maps, and

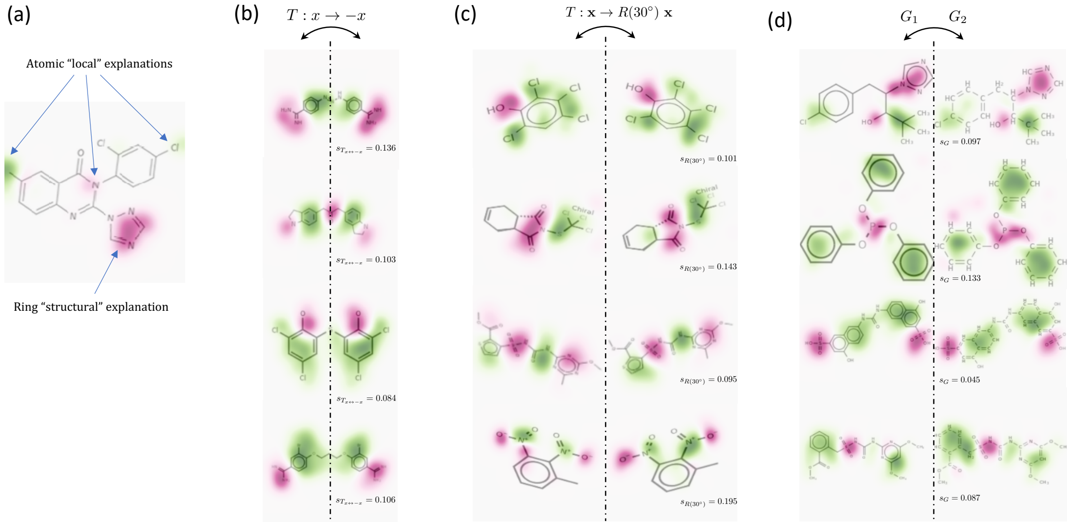
(a)
Atomic "local" explanations
Ring "structural" explanation
(b)
$T : x \to -x$
$s_{T_{x\leftrightarrow -x}} = 0.136$
$s_{T_{x\leftrightarrow -x}} = 0.103$
$s_{T_{x\leftrightarrow -x}} = 0.084$
$s_{T_{x\leftrightarrow -x}} = 0.106$
(c)
$T : \mathbf{x} \to R(30^\circ)\ \mathbf{x}$
$s_{R(30^\circ)} = 0.101$
$s_{R(30^\circ)} = 0.143$
$s_{R(30^\circ)} = 0.095$
$s_{R(30^\circ)} = 0.195$
(d)
G_1
G_2
$s_G = 0.097$
$s_G = 0.133$
$s_G = 0.045$
$s_G = 0.087$

Figure 2: Properties of contextual explainability for molecular property prediction: (a) explanations are sparse and incorporate both local and structural features; explanations tend to preserve the molecule's depiction symmetry under (b) reflections and (c) rotations; (d) explanations are robust with respect to different pictorial representations. Green (pink) overlay indicates positive (negative) contribution towards the prediction.

thus provides a quantitative measure of the correlation between two explanations. In performing this average, we only include normalized attributions $\widehat{a}(\mathbf{x})$ in absolute value above a given threshold (0.05), to avoid the score to depend on the amount of white space in the picture. The score is normalized such that it takes value between 0 (which occurs when the transformation commutes with the attribution maps, $\widehat{a}T = T\widehat{a}$) and 1 (which occurs when $\widehat{a}T = -T\widehat{a}$ and $\widehat{a} = \pm 1$). As a reference, if $\widehat{a} = U(-1, 1)$, the uniform distribution in the interval $[-1, 1]$, then $\mathbb{E}[s_T] = 1/2$.

We compute the score (3) for two transformations, namely reflection across the vertical axis $T = T_{x\leftrightarrow -x}$, and rotation of a 30° angle in the plane of the image $T = R(30^\circ)$. For reflections we report a value of $\mathbb{E}[s_{T_{x\leftrightarrow -x}}] = 0.135 \pm 0.003$, computed by averaging scores for 21 images of molecules exhibiting such symmetry. This value indicates that the symmetry is well captured by our explanations, as can be seen in Figure 2b. For rotations we instead report an average score over 121 molecule images of $\mathbb{E}[s_{R(30^\circ)}] = 0.169 \pm 0.004$, which again indicates that upon rotations, the attribution maps show a high consistency. We report in Figure 2b-c some examples of such transformations, the respective explanations and the associated scores. Such examples provide a visual intuition that for the achieved values of the score, the symmetries are well-respected by our explanations. We note that these tend to be less sparse than the original contextual explanations due to the normalization we introduced in (3).

Robustness with respect to the graphical depiction. There are several standards for graphically representing a molecule structure. We provide evidence that our contextual explanations are robust with respect to different graphical representations by slightly modifying the score (3). Let $\mathbf{G}_1(\mathrm{m}), \mathbf{G}_2(\mathrm{m})$ be two different graphical depictions of the same molecule m, then the score $s_G(\mathrm{m}) = \frac{1}{2}\overline{|\widehat{a}(\mathbf{G}_1(\mathrm{m})) - \widehat{a}(\mathbf{G}_2(\mathrm{m}))|}$ measures the average absolute difference between two attribution maps obtained from the two different graphical methods. We computed the score across a set of 121 molecules, and we obtained an average value of $\mathbb{E}[s_G] = 0.148 \pm 0.003$, which reveals a high level of agreement between explanations obtained from different graphical representations. Figure 2d shows some examples of such pictorial representations with their respective contextual explanations.

5 Conclusions

This work introduced an approach to explaining molecular property predictions based on molecules' graphical depictions, which we named contextual explainability. Our method is able to capture both basic (like atoms and bonds) as well as more complex structures (like rings and chemical groups), yielding explanations that are more aligned with chemists' intuition. We provided evidence that our contextual explanations possess several desirable properties: the attributions tend to be sparse, are robust with respect to the chosen graphical representation, and respect the symmetry of the input image. It would be interesting to explore our explanation framework in the context of property optimization: explanations in pixel space have the advantage that the model can explain a prediction not exclusively in terms of what is present in the given molecule, but also in terms of what is missing. The explanations could then be employed to provide suggestions of structure modifications for optimizing the given molecular property.

References

[1] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” vol. 70, pp. 1263–1272, 06–11 Aug 2017.

[2] Z. Wu, B. Ramsundar, E. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, “Moleculenet: a benchmark for molecular machine learning,” *Chem. Sci.*, vol. 9, pp. 513–530, 2018.

[3] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, “Analyzing learned molecular representations for property prediction,” *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370–3388, 2019. PMID: 31361484.

[4] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010. PMID: 20426451.

[5] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, “Convolutional networks on graphs for learning molecular fingerprints,” in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 2224–2232, Curran Associates, Inc., 2015.

[6] F. Montanari, L. Kuhnke, A. Ter Laak, and D.-A. Clevert, “Modeling physico-chemical admet endpoints with multitask graph convolutional networks,” *Molecules*, vol. 25, no. 1, 2020.

[7] R. Henderson, D.-A. Clevert, and F. Montanari, “Improving molecular graph neural network explainability with orthonormalization and induced sparsity,” in *Proceedings of the 38th International Conference on Machine Learning*, pp. –, 2021.

[8] S. Xie and M. Lu, “Interpreting and understanding graph convolutional neural network using gradient-based attribution method,” 2019.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[10] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert, “Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations,” *Chem. Sci.*, vol. 10, pp. 1692–1701, 2019.

[11] P. Karpov, G. Godin, and I. V. Tetko, “Transformer-cnn: Swiss knife for qsar modeling and interpretation,” *Journal of Cheminformatics*, vol. 12, 2020.

[12] H. Akita, K. Nakago, T. Komatsu, Y. Sugawara, S.-i. Maeda, Y. Baba, and H. Kashima, “Bayesgrad: Explaining predictions of graph convolutional networks,” in *International Conference on Neural Information Processing*, pp. 81–92, Springer, 2018.

[13] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” *arXiv preprint arXiv:1610.01644*, 2016.

[14] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, “On the expressive power of deep neural networks,” in *international conference on machine learning*, pp. 2847–2854, PMLR, 2017.

[15] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[17] J. Engel, M. Hoffman, and A. Roberts, “Latent constraints: Learning to generate conditionally from unconditional generative models,” *arXiv preprint arXiv:1711.05772*, 2017.

[18] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *International conference on machine learning*, pp. 2668–2677, PMLR, 2018.

[19] D.-A. Clevert, T. Le, R. Winter, and F. Montanari, “Img2mol – accurate smiles recognition from molecular graphical depictions,” *Chem. Sci.*, vol. -, pp. –, 2021.

[20] K. A. Lewis, J. Tzilivakis, D. J. Warner, and A. Green, “An international database for pesticide risk assessments and management,” *Human and Ecological Risk Assessment: An International Journal*, vol. 22, no. 4, pp. 1050–1064, 2016.

[21] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4905–4913, 2016.

[22] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” 2017.

[23] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” 2020.