Ground truth explanation dataset for chemical property prediction on molecular graphs

Authors

Eugen Hruska[1], Liang Zhao[2], Fang Liu[1]

Affiliations

1. Department of Chemistry, Emory University, Atlanta, Georgia, 30322
2. Department of Computer Science, Emory University, Atlanta, Georgia, 30322

corresponding author(s): Fang Liu (fang.liu@emory.edu)

Abstract

Interpretation of chemistry on an atomic scale improves with explainable artificial intelligence (XAI). The parts of the molecule with the most significant influence on the chemical property of interest can be visualized with atomwise and bondwise attributions. Nonetheless, the attributions from different XAI methods regularly disagree substantially, causing uncertainty about which explainability is correct. To determine a ground truth for attributions, we define chemical operations which avoid alchemical steps or approximations and allow extracting one attribution per atom or bond from existing datasets of chemical properties. This general procedure allows generating large datasets of ground truth attributions. The approach allowed us to create a ground truth explanation dataset with more than 5 million data points for the HOMO-LUMO gap chemical property. This open-source dataset of atomistic ground truth explainability may serve as a reference for XAI approaches.

Background & Summary

A wide range of machine learning (ML) approaches allows for explaining the chemistry of molecules, attributing which parts of the molecules are responsible for the chemical property of interest[1-19], and lessening the black box challenge of machine learning[20,21]. Typical explainable ML approaches that provide atomwise attribution include dummy atoms[22], classification of atoms by chemical intuition[23], regression models[24], graph neural network (GNN) attributions[25-28] with gradients[29], perturbations[30], decompositions[31], and surrogates[32]. In contrast, fragment-based explainability approaches generate importance values for groups of atoms or functional groups (subgraphs), e.g., Hammett equation[33], matched molecular pairs[34-36], molecular scaffolds[18,37], and counterfactuals[38]. Due to the ambiguity in molecule fragment identification (e.g., overlapping fragments)[39], this work considers only atomwise (nodes) and bondwise (edges) attributions, where one explainability value (attribution) is generated for each atom or bond.

Although different explainability approaches qualitatively agree with chemical intuition, some differences between the explanations are observed for quantitative structure-activity relationships (QSAR)[22,40] and GNNs[41]. The disagreement problem[42] occurs when explanations for the same chemical property generated by different methods are inconsistent[43], indicating uncertainty in the generated explanations of machine-learned models. Furthermore, inaccurate attribution can lead to unwarranted trust in models[44-46]. Reasons for the disagreement at the atomistic level include incorporating alchemical steps with limited chemical validity[22], noise in the gradient estimation[47,48], and bottlenecks in GNNs[49-51]. Currently, the best-performing method for different explanation tasks varies, and no single method performs well across various tasks[52].

In order to resolve the disagreement problem, we need metrics to evaluate the accuracy and reliability[53] of the explanation. Different metrics can be used[18,53-60] to evaluate reliability. For example, by perturbing the molecules, one can measure the stability[60] of the attribution and the consistency[39] between attributions generated by different approaches. Meanwhile, faithfulness[58] measures change in predictions when perturbing input features for important or unimportant attributions.

Various metrics have also been proposed to quantify the accuracy of explanation values. Some metrics focus on characterizing post hoc explanations, such as Grad-Cam[48], non-zero reference[24], and the assumption of node explanation smoothness[61]. Other metrics focus on the accuracy of the explanations with respect to the ground truth[41,61], which relies on the availability of ground-truth explanation datasets. The domain of XAI for molecular graphs is currently limited by the size of ground-truth datasets[61]. Although there are several multi-million datasets of non-trivial molecular property values (e.g., QM9[62,63], OC20[64], and PCQM4Mv2[65]) suitable for testing the prediction accuracy, no ground-truth explanation dataset of comparable size exists for testing the explanation accuracy. Some existing ground-truth datasets of atomistic explanations on molecular graphs explain trivial properties[66] or are synthetic datasets[30]. Other datasets are based on hundreds of hand-crafted structural subgraph motifs considered ground truth, for example, MUTAG[67] and Ames mutagenicity[68,69]. Further datasets consider other XAI methods as the reference, such as Crippen logP[70].

This paper describes a general method to generate large datasets of accurate atomistic ground truth explanations for molecular graphs. Further, we apply this method to the HOMO-LUMO gap, generating more than 5 million ground truth explanations. The HOMO-LUMO gap is the energy difference between the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) for the ground state of a molecule. Accurate ab initio calculations of the HOMO-LUMO gap are time-consuming, despite being one of the most straightforward properties to calculate in quantum chemistry. The HOMO-LUMO gap can show long-range information propagation[49] across the molecular graph, as exemplified in molecular wires, where the HOMO-LUMO gap shows a distance dependency up to 20 bonds[71]. This dataset aims to fill the void of large datasets for explanation values on molecular graphs and is an opportunity to benchmark different XAI approaches and increase the accuracy and reliability of existing XAI methods. We anticipate the dataset will be useful for the resolution of the disagreement problem in explainable machine learning and for improving the reliability of explanations for chemical properties.

Methods

This method generates atomwise and bondwise attributions from any dataset of accurate values of a chemical property of interest for chemically valid molecules. The explanation *e* for the selected chemical property *p* for atom or bond *i* is here defined as the signed difference in chemical properties between the molecule *M* and paired molecule *M*', shown in Eq. 1 and Figures 1-2. The paired molecule *M*' is defined in Eq. 2 by applying the chemical operator $\hat{o}$ on atom or bond *i* in molecule *M*.

$$e(p, M, i) = p(M) - p(M_i{}') \qquad (1)$$

$$M_i{}' = \hat{o}(M, i) \qquad (2)$$

The chemical operator changes only one atom (Figure 1) or bond (Figure 2) to ensure the explanation is defined for exactly one atom or bond. The molecules are represented as molecular graphs with implicit hydrogens, considering chemical validity is sufficient to determine the location and number of hydrogens. The decision trees of chemical operations

shown in Figures 1 and 2 ensure the application of only one chemical operation per atom or bond. For atoms, the decision tree (Figure 1) is as follows: if the atom is a heteroatom (non-carbon), then convert this atom into a carbon atom; else remove the carbon atom if it is a terminal atom (methyl group). For non-terminal carbon atoms, there is no chemical operation to avoid breaking the molecule into several molecules. For bonds, the decision tree (Figure 2) is as follows: if the bond is unsaturated, then saturate this bond; else break the bond if the bond is in a ring. For saturated bonds not in rings, there is no chemical operation to avoid breaking the molecule into several molecules. The decision trees define pairs of molecules that are directional since the chemical operations have a direction. The paired molecule might not be found in the existing dataset. For larger datasets, the odds of finding the paired molecule are higher.

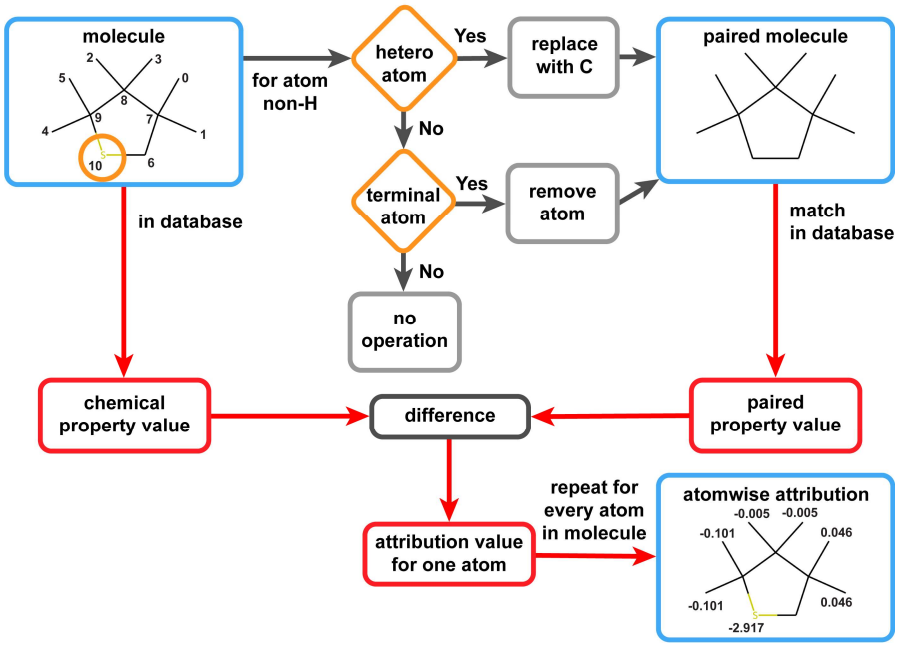

Figure 1 legend: Schema for estimating ground truth atomwise attribution. Only one chemical operation applies for each atom, depending on the decision tree. For a representative molecule, the paired molecule and atomwise ground truth attributions are shown. Symmetrical attributions for neighboring methyl groups are observed.

The different decision trees of chemical operations are the main differences between the approaches for atomwise and bondwise attributions. The explanation values for different atoms (or bonds) in one molecule can be identical, as visible in Figures 1 and 2, when the paired molecules are identical, for example, for neighboring methyl groups.

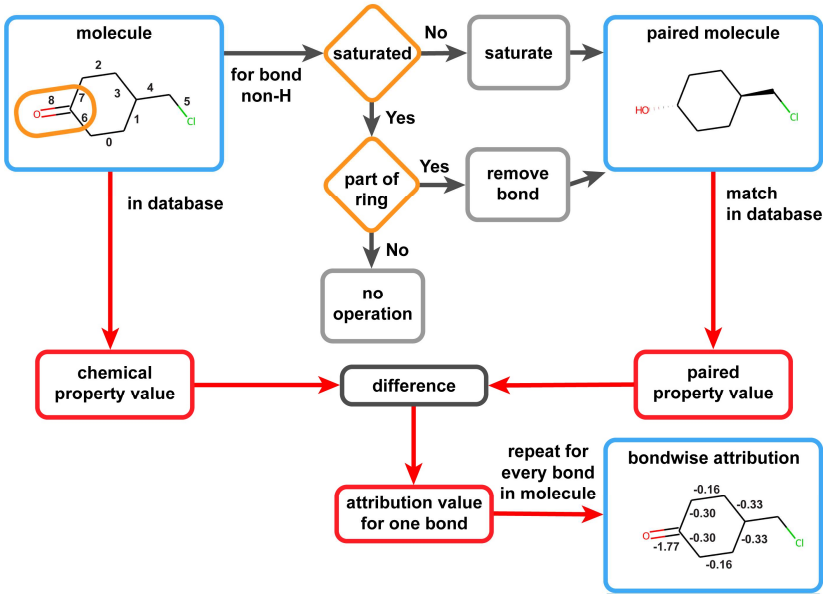

Figure 2 legend: Schema for estimating ground truth bondwise attribution. As presented in the decision tree, only one chemical operation applies for each bond. The paired molecule and bondwise ground truth explanations are shown for a representative molecule.

Data Records

Applying the above-described method to the PCQM4Mv2 dataset[65], we generated a dataset of ground truth explanations for the HOMO-LUMO gap molecular property. This publicly available dataset on Figshare[72] contains two csv files, one for atomwise and one for bondwise attributions. Table 1 summarizes the format of both csv files: each record of the explanation (a row of the csv) has five columns corresponding to the molecule index, the operation index (which of the four chemical operations was applied), the index of the atom or bond where the chemical operation was applied, the paired molecule index, and explanation values. The unit of the explanation values depends on the investigated chemical property; for the HOMO-LUMO gap, the unit is electronvolt (eV).

csv column	1	2	3	4	5
column name	molecule index	operation index	atom index or bond index	paired molecule index	explanation value
example data	3	0	13	2213220	-0.2013642494

Table 1 legend: Format of the csv files containing the ground truth explainability values.

The chemical operation indices are defined as follows: index 0 is the atomwise conversion of a heteroatom into a carbon atom, index 1 is the atomwise removal of a terminal carbon atom, index 2 is the bondwise saturation, and index 3 is the bondwise removal of a saturated bond in a ring. The order of atom and bond indices for individual molecules has to be preserved to correspond to the correct atoms and bonds. The README and python script on Figshare[72] describe the order of atom and bond indices used during dataset generation. The python script can be used to extract ground truth explanations from any other chemical dataset similar to

PCQM4Mv2 or for different chemical properties, as illustrated in Figures 1 and 2. The generated ground truth explanation dataset does not consider the chirality, charge, or radicals of the molecules.

Technical Validation

The generated ground truth atomic explanation dataset includes 3,619,970 atomwise and 1,555,262 bondwise explanation values. An explanation value was not generated for every atom, or bond, due to the option of no chemical operation (see Figures 1 and 2) or if the paired molecule was not found in the dataset. The distribution of the number of explanation values in each molecule is shown in the left panel of Figure 3. On average, each molecule with explanation values contains 1.82 atomwise and 1.98 bondwise values. The fraction of molecules with more than five explainability values per molecule is small. The distribution of explanation values (right panel of Figure 3) peaks at a value of 0 and does not show a dependency on the value of the chemical property, here the HOMO-LUMO gap.

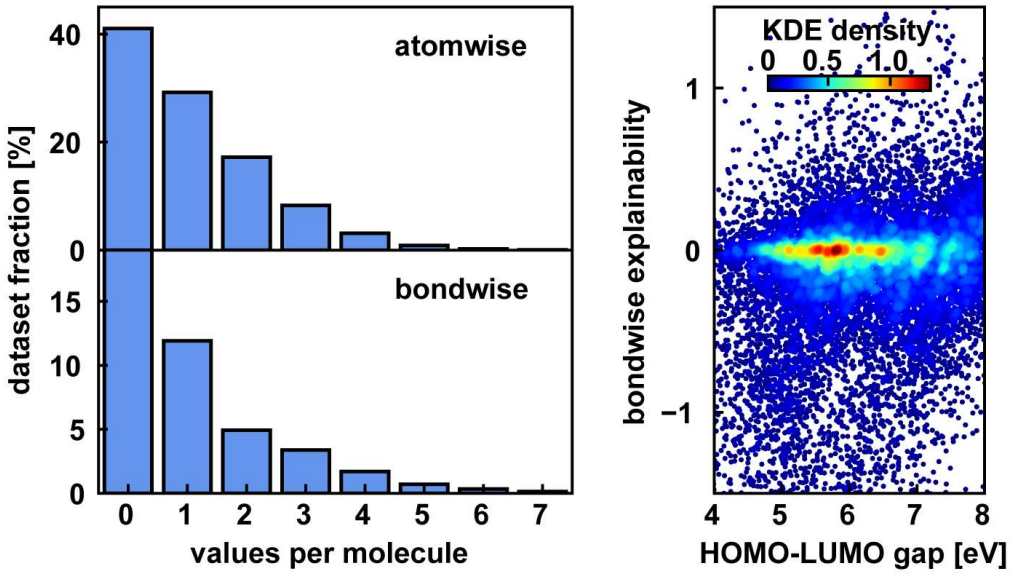

Figure 3 legend: Distribution of explanation values in the dataset. On the left is the distribution of the number of data points for each molecule. The distribution of explainability vs. HOMO-LUMO gap value is shown on the right as kernel density estimation (KDE).

The distribution of explanation values depends strongly on the atom type, as shown in Figure 4. The distribution for most atom types is non-gaussian, asymmetric, and not centered at zero value due to the directional nature of the decision tree and chemical operations. For carbon atoms, the distribution of explanation values is substantially narrower than for other atom types and peaks at zero value, and the reason may be that only terminal carbons are eligible in this atomwise attribution definition. For boron, sulfur, and phosphorus, flatter and broader histograms are observed with negative average atomwise explanation values, indicating the diverse roles these elements play in impacting the HOMO-LUMO gap of different molecules and an overall tendency to decrease the gap. For germanium, arsenic, and selenium, the

histogram is sparse due to the limited number of molecules in the PCQM4Mv2 dataset. A heavy-tailed distribution for silicon, phosphorus, and sulfur could be caused by the range of possible oxidation numbers for these elements.

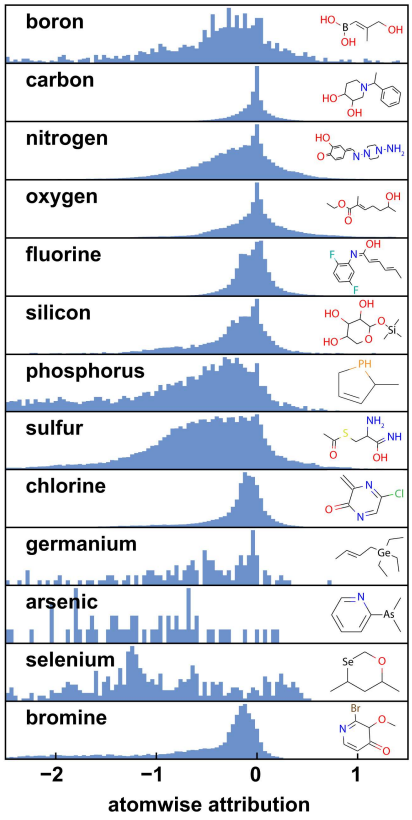

Figure 4 legend: Histograms of atomwise attribution for different atom types. The average atomwise explanation values for different atom types differ. For each atom type, an example molecule is shown.

The bondwise attribution depends on whether the bond is part of a ring and on the size of the ring, as shown in Figure 5. The average explanation value of bonds outside of rings is more negative (with a larger amplitude) than the average explanation value of bonds in rings, indicating that unsaturated bonds outside of rings impact the HOMO-LUMO gap more than bonds in rings. The difference can be explained by the different chemical operations, $\hat{o}$, applied to find paired molecules, $M_i{}'$, as defined in Eq. (2). For bonds outside of a ring, the valid chemical operation (changing to a single bond) only applies to unsaturated bonds. Hence, the statistics always reflect the impact on HOMO-LUMO gaps by unsaturated bonds, which correspond to degenerate molecular orbitals and smaller HOMO-LUMO gaps, leading to negative explanation values. In contrast, for bonds in rings, $\hat{o}$ applies in two situations: saturating an unsaturated bond in a ring (a big impact on the HOMO-LUMO gap) and removing a saturated bond in a ring (less impact on the HOMO-LUMO gap). The average impact of these two situations of bonds in rings leads to a smaller amplitude of the explanation value. Another observation is that the average bondwise attribution becomes more negative as the ring size increases. A possible explanation is that very small rings (3- or 4- membered rings) are more likely to be saturated, so removing a bond in the ring is likely to have less impact on the HOMO-LUMO gap.

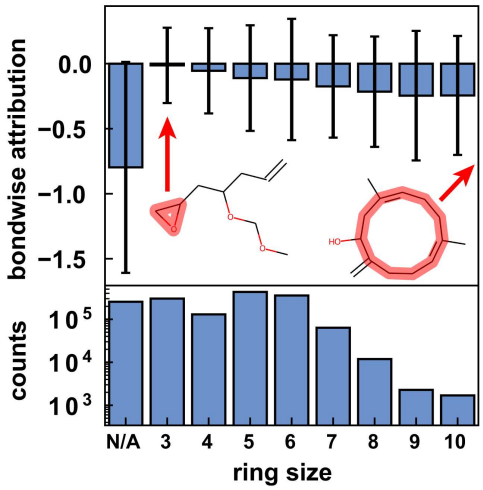

Figure 5 legend: Dependence of bondwise attribution on ring size. Two example molecules with rings of sizes 3 and 11 are shown.

Neighboring atoms (or bonds) in a molecule can have similar attributions. This assumption of smoothness[61] for neighboring attributions can be validated with the ground truth dataset. Here, we analyze our explanation dataset to investigate how the similarity of the explanation values for a pair of atoms (or bonds) in the same molecule depends on their closeness. For each pair of atoms (or bonds), the closeness is measured by their topological distance, the shortest path (number of edges) between them in the molecular graph. Then the similarity of pairs of explanations at a certain topological distance is measured in two ways. The first approach is the topological autocovariance[73-75] $\mathrm{AC_L}$ between atoms (or bonds) at a topological distance L. The topological autocovariance (Eq. 3) is the conditional expectation E of the product of the mean-subtracted explanation values for two atoms (or bonds) *i* and *j* in a molecule *M*, where the topological distance d_{ij} is L.

$$\mathrm{AC_L} = E\big([e(p,M,i) - \bar{e}] \cdot [e(p,M,j) - \bar{e}] \big| d_{ij} = L\big) \qquad (3)$$

Here, $e(p,M,i)$ denotes the explanation for atom *i* in molecule *M* about property p as defined in Eq. (1); $\bar{e}$ is the mean of explanation values for all atoms in all molecules in the dataset. The second approach is the statistics of the absolute difference of explanation values between each pair of atoms (or bonds) in the same molecule. The results for both approaches are visualized in Figure 6. Only molecules with at least one topological distance of length six are included in the atomwise analysis to avoid bias from comparing small and large molecules. With this criterion, 503,508 molecules are included in this figure, with 3,887,724 atomwise explanation pairs and 996,081 bondwise explanation pairs. We observe the autocovariance decreases strongly after a distance of two bonds for atomwise explanation and a distance of one bond for bondwise explanation values, supporting the assumption of smoothness for neighboring attributions. Similarly, the absolute difference for bondwise attribution increases almost linearly with the distance up to a distance of four bonds, also confirming the abovementioned assumption.

In contrast, for atomwise attribution, a substantial average absolute difference is observed even for neighboring atoms, but a lower absolute difference is present for a distance of two bonds. The ubiquitous methyl groups can explain this counterintuitive behavior in our dataset since adjacent carbon atoms (non-terminal) are not eligible for explanations, as demonstrated by a low count of atomwise pairs at a distance of one bond in Figure 6. The frequent

occurrence of neighboring methyl groups is also reflected in Figure 1, which explains the lower average absolute difference at a distance of two bonds since neighboring methyl groups have a topological distance of 2 and identical explanation values due to symmetry.

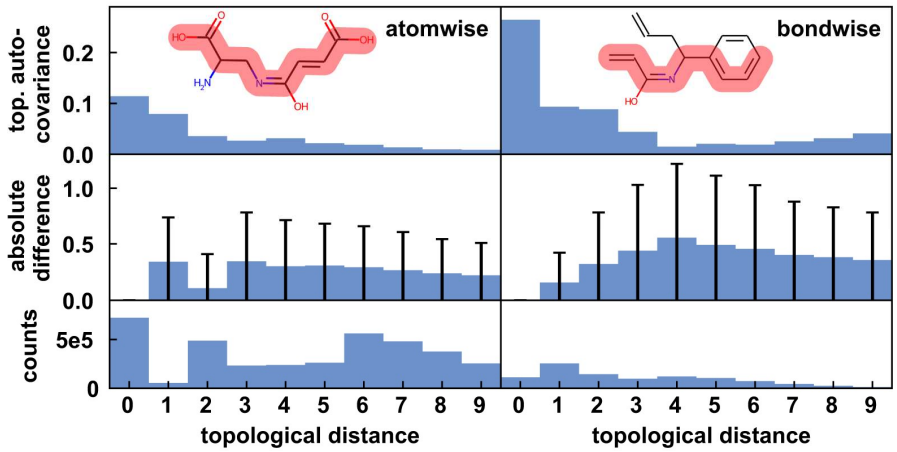

Figure 6 legend: Dependence of the similarity between two explanation values in the same molecule with respect to the topological distance. The topological autocovariance, absolute difference, and counts for atomwise (left) and bondwise (right) attributions are shown. Examples of longer topological distance paths are highlighted in red.

Code Availability

The python script used to generate ground truth atomwise, and bondwise attribution from the PCQM4Mv2 dataset is available on Figshare[72].

Acknowledgments

Start-up funds provided by Emory University supported this work. Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund for support of this research through grant number PRF #65858-DNI4.

Author contributions

All authors contributed to the research.

Competing interests

The authors declare no competing financial interests.

References

1 Humer, C. *et al.* Cheminformatics model explorer (cime): Exploratory analysis of chemical model explanations. *J. Cheminf.* **14**, 21 (2022). https://doi.org:10.1186/s13321-022-00600-z

2 Sobieraj, M. & Setny, P. Granger causality analysis of chignolin folding. *J. Chem. Theory Comput.* **18**, 1936-1944 (2022). https://doi.org:10.1021/acs.jctc.1c00945

3 Urbina, F. *et al.* Uv-advisor: Attention-based recurrent neural networks to predict uv-vis spectra. *Anal. Chem.* **93**, 16076-16085 (2021). https://doi.org:10.1021/acs.analchem.1c03741

4 Verma, S., Rivera, M., Scanlon, D. O. & Walsh, A. Machine learned calibrations to high-throughput molecular excited state calculations. *J. Chem. Phys.* **156**, 134116 (2022). https://doi.org:10.1063/5.0084535

5 Maser, M. R. *et al.* Multilabel classification models for the prediction of cross-coupling reaction conditions. *Journal of Chemical Information and Modeling* (2021). https://doi.org:10.1021/acs.jcim.0c01234

6 Harren, T., Matter, H., Hessler, G., Rarey, M. & Grebner, C. Interpretation of structure-activity relationships in real-world drug design data sets using explainable artificial intelligence. *J. Chem. Inf. Model.* **62**, 447-462 (2022). https://doi.org:10.1021/acs.jcim.1c01263

7 Gandhi, H. A. & White, A. D. Explaining molecular properties with natural language. *Preprint* (2022). https://doi.org:10.26434/chemrxiv-2022-v5p6m-v3

8 van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *Preprint* (2022). https://doi.org:10.26434/chemrxiv-2022-mfq52-v3

9 Galuzzi, B. G. *et al.* Machine learning for efficient prediction of protein redox potential: The flavoproteins case. *J. Chem. Inf. Model.* **62**, 4748-4759 (2022). https://doi.org:10.1021/acs.jcim.2c00858

10 Kanamaru, Y. & Matsui, T. Factor analysis of error in oxidation potential calculation: A machine learning study. *J. Comput. Chem.* **43**, 1504-1512 (2022). https://doi.org:10.1002/jcc.26953

11 Sharma, B. *et al.* Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations. *Preprint* (2022). https://doi.org:10.48550/arXiv.2204.06614

12 Moreira-Filho, J. T. *et al.* Beetoxai: An artificial intelligence-based web app to assess acute toxicity of chemicals to honey bees. *Artif. Intell. Life Sci.* **1**, 100013 (2021). https://doi.org:10.1016/j.ailsci.2021.100013

13 Lee, K. *et al.* Combating small-molecule aggregation with machine learning. *Cell Rep. Phys. Sci.* **2**, 100573 (2021). https://doi.org:10.1016/j.xcrp.2021.100573
14 Bender, A. *et al.* Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* (2022). https://doi.org:10.1038/s41570-022-00391-9
15 Poelking, C. *et al.* Meaningful machine learning models and machine-learned pharmacophores from fragment screening campaigns. *Preprint* (2022). https://doi.org:10.48550/arxiv.2204.06348
16 Tinkov, O. *et al.* The influence of structural patterns on acute aquatic toxicity of organic compounds. *Mol. Inf.* **40**, e2000209 (2021). https://doi.org:10.1002/minf.202000209
17 Low, K., Coote, M. L. & Izgorodina, E. I. Explainable solvation free energy prediction combining graph neural networks with chemical intuition. *J. Chem. Inf. Model.* (2022). https://doi.org:10.1021/acs.jcim.2c01013
18 Amara, K., Rodriguez-Perez, R. & Jiménez Luna, J. A substructure-aware loss for feature attribution in drug discovery. *Preprint* (2022). https://doi.org:10.26434/chemrxiv-2022-qxq56-v2
19 Teufel, J., Torresi, L., Reiser, P. & Friederich, P. Megan: Multi-explanation graph attention network. *Preprint* (2022). https://doi.org:10.48550/arxiv.2211.13236
20 Li, J. & Lopez, S. A. A look inside the black box of machine learning photodynamics simulations. *Acc. Chem. Res.* (2022). https://doi.org:10.1021/acs.accounts.2c00288
21 Omidvar, N. *et al.* Interpretable machine learning of chemical bonding at solid surfaces. *J. Phys. Chem. Lett.* **12**, 11476-11487 (2021). https://doi.org:10.1021/acs.jpclett.1c03291
22 Sheridan, R. P. Interpretation of qsar models by coloring atoms according to changes in predicted activity: How robust is it? *J. Chem. Inf. Model.* **59**, 1324-1337 (2019). https://doi.org:10.1021/acs.jcim.8b00825
23 Wildman, S. A. & Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868-873 (1999). https://doi.org:10.1021/ci990307l
24 Letzgus, S. *et al.* Toward explainable ai for regression models. *Preprint* (2021). https://doi.org:10.48550/arxiv.2112.11407
25 Yuan, H., Yu, H., Gui, S. & Ji, S. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans. Pattern. Anal. Mach. Intell.* **PP** (2022). https://doi.org:10.1109/TPAMI.2022.3204236
26 Mastropietro, A., Pasculli, G., Feldmann, C., Rodríguez-Pérez, R. & Bajorath, J. Edgeshaper: Bond-centric shapley value-based explanation method for graph neural networks. *iScience* **25**, 105043 (2022). https://doi.org:10.1016/j.isci.2022.105043

27 Müller, P., Faber, L., Martinkus, K. & Wattenhofer, R. Dt+ gnn: A fully explainable graph neural network using decision trees. (2022). https://doi.org:10.48550/arXiv.2205.13234
28 Giunchiglia, V., Shukla, C. V., Gonzalez, G. & Agarwal, C. Towards training gnns using explanation directed message passing. *Preprint* (2022). https://doi.org:10.48550/arxiv.2211.16731
29 Baldassarre, F. & Azizpour, H. Explainability techniques for graph convolutional networks. *Preprint* (2019). https://doi.org:10.48550/arxiv.1905.13686
30 Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *Adv. neural inf. process. syst.* **32**, 9240-9251 (2019). https://doi.org:10.48550/arxiv.1903.03894
31 Schnake, T. *et al.* Higher-order explanations of graph neural networks via relevant walks. *IEEE Trans. Pattern. Anal. Mach. Intell.* **44**, 7581-7596 (2022). https://doi.org:10.1109/TPAMI.2021.3115452
32 Huang, Q., Yamada, M., Tian, Y., Singh, D. & Chang, Y. Graphlime: Local interpretable model explanations for graph neural networks. *IEEE Trans. Knowl. Data. Eng.*, 1-6 (2022). https://doi.org:10.1109/TKDE.2022.3187455
33 Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **59**, 96-103 (1937). https://doi.org:10.1021/ja01280a022
34 Hussain, J. & Rea, C. Computationally efficient algorithm to identify matched molecular pairs (mmps) in large data sets. *J. Chem. Inf. Model.* **50**, 339-348 (2010). https://doi.org:10.1021/ci900450m
35 Dalke, A., Hert, J. & Kramer, C. Mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.* **58**, 902-910 (2018). https://doi.org:10.1021/acs.jcim.8b00173
36 Naveja, J. J. & Vogt, M. Automatic identification of analogue series from large compound data sets: Methods and applications. *Molecules* **26** (2021). https://doi.org:10.3390/molecules26175291
37 Hu, Y., Stumpfe, D. & Bajorath, J. Computational exploration of molecular scaffolds in medicinal chemistry. *J. Med. Chem.* **59**, 4062-4076 (2016). https://doi.org:10.1021/acs.jmedchem.5b01746
38 Wellawatte, G. P., Seshadri, A. & White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.* **13**, 3697-3705 (2022). https://doi.org:10.1039/d1sc05259d
39 Jiménez-Luna, J., Skalic, M. & Weskamp, N. Benchmarking molecular feature attribution methods with activity cliffs. *J. Chem. Inf. Model.* **62**, 274-283 (2022). https://doi.org:10.1021/acs.jcim.1c01163
40 Matveieva, M. & Polishchuk, P. Benchmarks for interpretation of qsar models. *J. Cheminf.* **13**, 41 (2021). https://doi.org:10.1186/s13321-021-00519-x

41 Sanchez-Lengeling, B., Wei, J. & Lee, B. Evaluating attribution for graph neural networks. *Adv. neural inf. process. syst.* (2020).
42 Krishna, S. *et al.* The disagreement problem in explainable machine learning: A practitioner's perspective. *Preprint* (2022). https://doi.org:10.48550/arxiv.2202.01602
43 Gilpin, L. H., Paley, A. R., Alam, M. A., Spurlock, S. & Hammond, K. J. "Explanation" is not a technical term: The problem of ambiguity in xai. *Preprint* (2022). https://doi.org:10.48550/arxiv.2207.00007
44 Slack, D., Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society*, 180-186 (2020). https://doi.org:10.1145/3375627.3375830
45 DeGrave, A. J., Janizek, J. D. & Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* (2021). https://doi.org:10.1038/s42256-021-00338-7
46 Rozenblit, L. & Keil, F. The misunderstood limits of folk science: An illusion of explanatory depth. *Cogn. Sci.* **26**, 521-562 (2002). https://doi.org:10.1207/s15516709cog2605_1
47 Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: Removing noise by adding noise. *Preprint* (2017). https://doi.org:10.48550/arxiv.1706.03825
48 Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618-626 (2017). https://doi.org:10.1109/ICCV.2017.74
49 Alon, U. & Yahav, E. On the bottleneck of graph neural networks and its practical implications. *Preprint* (2020). https://doi.org:10.48550/arxiv.2006.05205
50 Chen, D. *et al.* Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *AAAI* **34**, 3438-3445 (2020). https://doi.org:10.1609/aaai.v34i04.5747
51 Dwivedi, V. P. *et al.* Long range graph benchmark. *Preprint* (2022). https://doi.org:10.48550/arxiv.2206.08164
52 Han, T., Srinivas, S. & Lakkaraju, H. Which explanation should i choose? A function approximation perspective to characterizing post hoc explanations. *Preprint* (2022). https://doi.org:10.48550/arxiv.2206.01254
53 Agarwal, C. & Zitnik, M. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. *… Conference on Artificial …* (2022).
54 Henderson, R., Clevert, D. A. & Montanari, F. Improving molecular graph neural network explainability with orthonormalization and induced sparsity. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139*, 4203-4213 (2021).

55 Dai, E. *et al.* A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Preprint* (2022). https://doi.org:10.48550/arxiv.2204.08570

56 Agarwal, C., Queen, O., Lakkaraju, H. & Zitnik, M. Evaluating explainability for graph neural networks. *Preprint* (2022). https://doi.org:10.48550/arxiv.2208.09339

57 Henderson, R. & Clevert, D. A. Improving molecular graph neural network explainability with orthonormalization and induced sparsity. *… Conference on Machine …* (2021).

58 Agarwal, C. *et al.* Openxai: Towards a transparent evaluation of model explanations. *Preprint* (2022). https://doi.org:10.48550/arxiv.2206.11104

59 Slack, D., Hilgard, A. & Singh, S. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural …* (2021).

60 Funke, T., Khosla, M., Rathee, M. & Anand, A. Zorro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Trans. Knowl. Data. Eng.*, 1-12 (2022). https://doi.org:10.1109/TKDE.2022.3201170

61 Gao, Y. *et al.* Gnes: Learning to explain graph neural networks. *2021 IEEE International Conference on Data Mining (ICDM)*, 131-140 (2021). https://doi.org:10.1109/ICDM51629.2021.00023

62 Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data.* **1**, 140022 (2014). https://doi.org:10.1038/sdata.2014.22

63 Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *figshare* (2014). https://doi.org:10.6084/m9.figshare.978904

64 Chanussot, L. *et al.* Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catal.*, 6059-6072 (2021). https://doi.org:10.1021/acscatal.0c04525

65 Hu, W. *et al.* Ogb-lsc: A large-scale challenge for machine learning on graphs. *Preprint* (2021). https://doi.org:10.48550/arxiv.2103.09430

66 Rao, J., Zheng, S., Lu, Y. & Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Preprint*, 100628 (2022). https://doi.org:10.1016/j.patter.2022.100628

67 Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J. & Hansch, C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.* **34**, 786-797 (1991). https://doi.org:10.1021/jm00106a046

68 Sushko, I., Salmina, E., Potemkin, V. A., Poda, G. & Tetko, I. V. Toxalerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* **52**, 2310-2316 (2012). https://doi.org:10.1021/ci300245q

69 Hansen, K. *et al.* Benchmark data set for in silico prediction of ames mutagenicity. *J. Chem. Inf. Model.* **49**, 2077-2081 (2009). https://doi.org:10.1021/ci900161g

70 Rasmussen, M. H., Christensen, D. S. & Jensen, J. H. Do machines dream of atoms? A quantitative molecular benchmark for explainable ai heatmaps. *Preprint* (2022). https://doi.org:10.26434/chemrxiv-2022-gnq3w

71 Valdiviezo, J., Rocha, P., Polakovsky, A. & Palma, J. L. Nonexponential length dependence of molecular conductance in acene-based molecular wires. *ACS Sensors* **6**, 477-484 (2021). https://doi.org:10.1021/acssensors.0c02049

72 Hruska, E., Zhao, L. & Liu, F. Ground truth explanation dataset for chemical property prediction on molecular graphs. *figshare* (2022).

73 Morea, G. & Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *New J. Chem.* (1980).

74 Fernandez, M., Abreu, J. I., Shi, H. & Barnard, A. S. Machine learning prediction of the energy gap of graphene nanoflakes using topological autocorrelation vectors. *ACS Comb. Sci.* **18**, 661-664 (2016). https://doi.org:10.1021/acscombsci.6b00094

75 Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: Feature selection for machine learning and structure-property relationships. *J. Phys. Chem. A* **121**, 8939-8954 (2017). https://doi.org:10.1021/acs.jpca.7b08750