
Explainability Techniques for Graph Convolutional Networks

Federico Baldassarre¹ Hossein Azizpour¹

Abstract

Graph Networks are used to make decisions in potentially complex scenarios but it is usually not obvious how or why they made them. In this work, we study the explainability of Graph Network decisions using two main classes of techniques, gradient-based and decomposition-based, on a toy dataset and a chemistry task. Our study sets the ground for future development as well as application to real-world problems.

1. Introduction

Many concepts from chemistry, life science, and physics are naturally represented in the graph domain. Most Machine Learning (ML) methods, however, were devised to operate on Euclidean space. One of the most successful ML techniques, Deep Learning, has been recently generalized to operate on a graph domain. These Graph Networks (GN) have achieved remarkable performances in various applications thanks to their consistency to the native data representation, with examples from biochemistry (Duvenaud et al., 2015; Kearnes et al., 2016; Fout et al., 2017; Zitnik et al., 2018), physics (Battaglia et al., 2016; Chang et al., 2017; Gilmer et al., 2017; Watters et al., 2017; Sanchez-Gonzalez et al., 2018), visual recognition (Qi et al., 2017; 2018; Narasimhan et al., 2018), and natural language processing (Bastings et al., 2017; Beck et al., 2018).

ML algorithms become increasingly trustworthy to humans when the basis for their decisions can be explained in human terms. Interpretability is also useful for diagnosing biases, designing datasets and gaining insight on governing laws. While several methods have been developed for standard deep networks, there is a lack of study for applicability of such methods on GNs; that is the focus of this work.

In this work, we assume the general form of GNs as defined in (Battaglia et al., 2018). Regarding explanation algorithms,

¹Division of Robotics, Perception, and Learning, School of Electrical Engineering and Computer Science, KTH (Royal Institute of Technology), Stockholm, Sweden. Correspondence to: Federico Baldassarre <fedbal@kth.se>, Hossein Azizpour <azizpour@kth.se>.

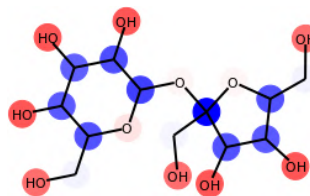


Figure 1: **Explanation for the solubility of sucrose:** the prediction is decomposed into positive (red) and negative (blue) contributions attributed to the atoms using Layer-wise Relevance Propagation.

we consider two main classes: a) gradient-based such as Sensitivity Analysis (Baehrens et al., 2010; Simonyan et al., 2014) and Guided Backpropagation (Springenberg et al., 2015), b) decomposition-based such as Layer-wise Relevance Propagation (Bach et al., 2015) and Taylor decomposition (Montavon et al., 2017). We base the discussions on a toy dataset and a chemistry task. We hope this work will set the ground for this important topic and suggest future directions and applications.

The contributions of this work can be summarized as:

1. to the best of our knowledge, this is the first work to focus on explainability techniques for GN
2. we highlight and identify the challenges and future directions for explaining GN
3. we compare two main classes of explainability methods on graph-based tasks predictions

Our PyTorch (Paszke et al., 2017) implementation of GNs equipped with different explanation algorithms is available at github.com/baldassarreFe/graph-network-explainability.

2. Related works

This work is closely related to Graph Networks and explanation techniques developed for standard networks.

Graph Networks Graphs can be embedded in Euclidean spaces using Neural Networks (Perozzi et al., 2014; Hamilton et al., 2017; Kipf & Welling, 2016). This representation preserves the relational structure of the graph while enjoying the properties of Euclidean space. The embedding can then be processed further, e.g. with more interpretable linear models (Duvenaud et al., 2015). On the other hand, DL algorithms that operate end-to-end in the graph domain can leverage the structure to predict at vertex (Hu et al., 2018),

edge (Qi et al., 2018), or global (Zitnik et al., 2018) levels. Several variants of GN have been proposed starting from the early work of (Scarselli et al., 2009). It was further extended, by gating (Li et al., 2016), convolutions in spectral (Bruna et al., 2014) and spatial (Kipf & Welling, 2017) domains, skip connections (Rahimi et al., 2018) and recently attention (Velickovic et al., 2018). In this work, we focus on the GN as described in (Battaglia et al., 2018), which tries to generalize all variants while remaining conceptually simple.

Model interpretation and explanation of predictions ML methods are ubiquitous in several industries, with deep networks achieving impressive performances. Where humans are involved, the expectation of high levels of safety, reliability, and fairness arises. Thanks to this demand, several techniques have been developed for increasing the transparency of the most successful models. At high level these techniques can be divided into those that attempt to *interpret the model as a whole* (Simonyan et al., 2014; Nguyen et al., 2016) and those that try to *explain individual predictions* made by a model (Sung, 1998; Bach et al., 2015)¹.

This paper focuses on the latter, for which several techniques have been developed on standard deep networks. These include works that attempt at variation-based analysis (Baehrens et al., 2010; Simonyan et al., 2014; Springenberg et al., 2015; Ribeiro et al., 2016; Smilkov et al., 2017; Bordes et al., 2018) and output decomposition (Montavon et al., 2017; Shrikumar et al., 2017; Kindermans et al., 2018). (Sundararajan et al., 2017) combines the two principles and invert the decisions (Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015; Carlsson et al., 2017). Furthermore, (Dhurandhar et al., 2018) uses contrastive explanations, (Zhang et al., 2018) identifies minimal changes in the input to get the desired prediction. In this work, we evaluate variation- and decomposition-based techniques in the context of GNs.

Explanation for GNs To the best of our knowledge, exploring explanation techniques for GNs has not been the focus of any prior work. In (Duvenaud et al., 2015), GNs are used to learn molecular fingerprints and predict their chemical properties. Chemically grounded insights are then extracted from the model by heuristic inspection of their results². In our experiments, we show that GNs trained end-to-end on those problems achieve similar performance while making it possible to *explain individual predictions*.

3. Method

3.1. Graph Network

GN as described in (Battaglia et al., 2018) use a message-passing algorithm that aggregates local information similar to convolutions in CNNs. Graphs can contain features on

edges $E = \{e_k\}$, nodes $V = \{v_i\}$ and graph-level u . At every layer of computation, the graph is updated using three update functions ϕ and three aggregation functions ρ :

$$\begin{aligned} e'_k &= \phi^e(e_k, v_{r_k}, v_{s_k}, u) & \bar{e}'_i &= \rho^{e \rightarrow v}(E'_i) \\ v'_i &= \phi^v(\bar{e}'_i, v_i, u) & \bar{e}' &= \rho^{e \rightarrow u}(E') \\ u' &= \phi^u(\bar{e}', \bar{v}', u) & \bar{v}' &= \rho^{v \rightarrow u}(V') \end{aligned} \quad (1)$$

where r_k and s_k represent the sender and receiver nodes of the k -th edge, and the sets E'_i , E' , V' represent the edges incident to node i , all edges updated by ϕ^e and all nodes updated by ϕ^v . Each processing layer leaves the structure of the graph unchanged, updating only the features of the graph and not its topology. The mapping $f : (E, V, u) \rightarrow y$, can represent a single quantity of interest (e.g. the solubility of a molecule) or a graph with individual predictions for nodes and edges. In this work, all ϕ are linear transformations followed by ReLU activations, and all ρ are sum/mean/max pooling operations.

3.2. Explainability

Sensitivity Analysis (SA) produces local explanations for the prediction of a differentiable function f using the squared norm of its gradient w.r.t. the inputs x (Gevrey et al., 2003): $S(x) \propto \|\nabla_x f\|^2$. The saliency map S produced with this method describes the extent to which variations in the input would produce a change in the output.

Guided Backpropagation (GBP) also constructs a saliency map using gradients (Springenberg et al., 2015). Different from SA, negative gradients are clipped during backpropagation, which concentrates the explanation on the features that have an excitatory effect on the output.

Layer-wise Relevance Propagation (LRP) produces relevance maps by decomposing the output signal of every transformation into a combination of its inputs. For certain configuration of weights and activations, LRP can be interpreted as a repeated Taylor decomposition (Montavon et al., 2017) that preserves the amount of total relevance R across layers: $\sum R^{(x)} = \dots = \sum R^{(\ell)} = \dots = f(x)$. (Bach et al., 2015) introduces two rules for LRP, the $\alpha\beta$ -rule and the ϵ -stabilized rule, both discussed in Appendix A.2. We chose the latter for its robustness and simplicity.

Different to the previous two methods, LRP identifies which features of the input contribute the most to the final prediction, rather than focusing on its variation. Furthermore, it is capable of handling positive and negative relevance, allowing for a deeper analysis of the contributing factors.

Autograd-based implementation All three methods rely on a backward pass through the network to propagate gradients/relevance from the output and accumulating it to the input. Since the computational graph of a GN can become complex and non-sequential, we take advantage of PyTorch’s capability to track operations and implement these algorithms on top of its *autograd* module.

¹For insight on such grouping refer to (Gilpin et al., 2018)

²as described in the authors’ rebuttal: (NIPS reviews, 2015)

4. Experiments

To evaluate different explainability methods on GNs, we consider a toy graph problem and a chemistry regression task. Task-specific comments can be found in this section, followed by a more general discussion in Section 5.

4.1. Infection

In this toy problem, the input graph represents a group of individuals who are either *sick* or *healthy*, as well as *immune* to a certain disease. Between people are directed edges, representing the relationships they have, characterized as *virtual* or not. The disease spreads according to a simple rule: a sick node infects the neighbors to which it is connected through a non-virtual edge, unless the target node is immune. The objective is to predict the state of every node of the graph after one step of the spread, and then evaluate the correctness of the explanations produced with Sensitivity Analysis, Guided Backpropagation and Layer-wise Relevance Propagation against the logical infection dynamics. Details about the dataset, the network and the training procedure are in Appendix B.1.

When tasked to explain the prediction for a single node, all three techniques identify the relevant nodes/edges of the input graph (Fig. 2). We note however that the explanations produced by variation-based methods tend to diverge from how a human would intuitively describe the process in terms of cause and effect, while LRP results are more natural. Appendix C.1 provides a detailed case-based visualization of the explanations, down to the individual features.

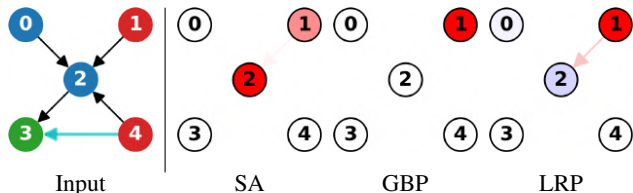


Figure 2: **Explaining why node 2 becomes infected.**

SA places high relevance on the node itself (if 2 was more sick at the beginning, it would be more infected at the end). GBP correctly identifies node 1 as a source of infection, but very small importance is given to the edge. LRP decomposes the prediction into a negative contribution (blue, node 2 is not sick), and two positive contributions (red, node 1 is sick and $1 \rightarrow 2$ is not virtual). Node 4 is ignored due to max pooling.

4.2. Solubility

We train a GN to predict the aqueous solubility of organic compounds from their molecular graph as in (Duvenaud et al., 2015). Our multi-layer GN matches their performances while remaining simple. Details about the dataset, the network and the training procedure are in Appendix B.1.

When explaining the predictions of the network, LRP attributes positive and negative relevance to features that are

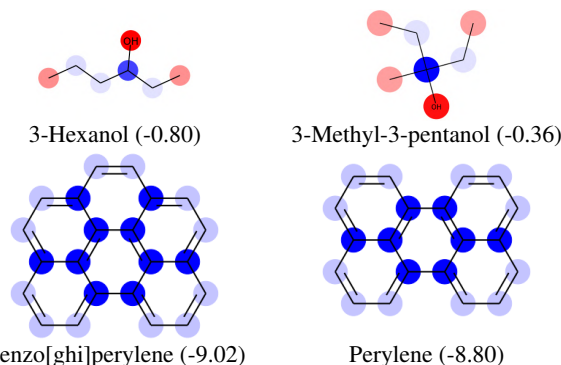


Figure 3: **Predicted solubility (log mol/L) and its explanation produced with LRP.** Positive relevance (red) are on R-OH groups, indicating their positive contribution to the predicted value. Negative relevance (blue) can be found on central carbons and non-polar aromatic rings, indicating they advocate towards lower solubility values. See the appendix for a breakdown of the explanation onto the individual features.

known to correlate with solubility, such as the presence of R-OH groups on the outside of the molecule and features that typically indicate low solubility such as repeated non-polar aromatic rings (Fig. 3). Similar observations are made in (Duvenaud et al., 2015), although by manual inspection of high-scoring predictions.

Note that LRP was originally introduced to explain classification predictions, but here it is adopted for a regression task. A discussion on how to interpret these explanations can be found in section C.2 of the appendix.

5. Discussion

Recent methods for explainability have been developed within image or text domains (Bach et al., 2015; Springenberg et al., 2015; Ribeiro et al., 2016). With the experiments presented in this work, we intend to highlight some key differences of the graph domain that require special consideration to produce meaningful explanations.

5.1. The role of connections

Images can be seen as graphs with a regular grid topology and whose features are attributed only to nodes. In this context, an explanation can take the form of a heatmap over the image, highlighting relevant pixels and, implicitly, their local connections. For graphs with irregular connectivity, edges acquire a more preeminent role that can be missed when using image-based explanation techniques. For example, in a graph where edge features are not present or are all identical (not informative), neither gradients nor relevance would be propagated back to these connections, even though the presence itself of a connection between two nodes is a source of information that should be taken into account for explanations. We propose to take advantage of the structure-

preserving property of graph convolution and aggregate explanations at multiple steps of message-passing, arguing that the importance of connections should emerge from the intermediate steps (Fig. 4).

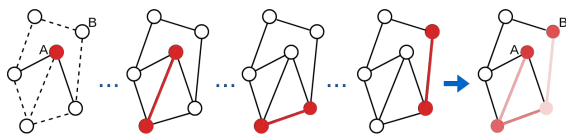


Figure 4: **Node A is responsible for a prediction on B.** Even if the input graph does not have features on the edges (represented as dashed lines), the relevant path $A \rightarrow B$ is identified by aggregating the relevance at multiple steps of graph convolution.

5.2. Pooling

Architectural choice In standard NN, pooling operations are commonly used to aggregate features. In message-passing GNs, pooling is used to aggregate edge and node features at a local and global level, while not modifying the topology of the network (Eq. 1). The choice of pooling function in GN is closely related to the learning problem, e.g. sum pooling is best for counting at a global level, while max pooling can be used for identifying local properties.

Explanations The choice of aggregation also influences the explanations obtained for a prediction. Sum and mean pooling propagate gradients/relevance to all their inputs, possibly identifying all sources of signal. Max pooling, instead, considers only one of its inputs and disregards others, regardless of their magnitude, which can lead to incomplete explanations (e.g. multiple neighboring sick nodes could explain an infection). To counter this, LRP proposes to approximate max pooling with L_p -norm during relevance propagation, but this approach can over-disperse relevance to unimportant inputs. We suggest that the backward pass through max pooling should be approached as a search that only redistributes relevance to those inputs that result in a similar prediction if chosen as maxima (Fig. 5).

5.3. Heterogeneous Graph Features

Images are usually represented as a matrix of continuous RGB pixel values, while graphs are often employed for domains that require a mixed encoding of continuous, binary and categorical features that are *semantically meaningful* (Fout et al., 2017; Kearnes et al., 2016; Sanchez-Gonzalez et al., 2018). Thus, rather than aggregating the explanation at the node/edge level, it can be of higher interest to evaluate the importance of individual features. For this reason, visualizations based on graph heatmaps might be insufficient. We suggest a more detailed visualization in Appendix C.

5.4. Perturbation-based evaluation

Images and graphs can be considered as points in very high-dimensional spaces, belonging to complex and structured

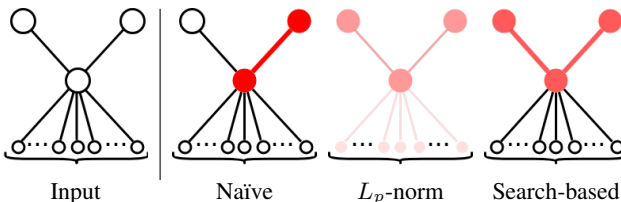


Figure 5: **Propagation rules for max pooling.** The two top nodes are important, the N bottom ones are not. Relevance can be naïvely propagated from the central node to only one of the top nodes. Approximating max pooling as a L_p -norm would result in a more complete explanation, but for large N relevance could disperse to unimportant nodes. A search-based method identifies and propagates relevance only to the relevant nodes.

manifolds (Tenenbaum et al., 2000). The commonly used representation of an image introduces an elevated degree of redundancy, so that changing the value of a single pixel minimally affects the content and meaning associated with the image. Under this observation, an explanation can be quantitatively evaluated by progressively “graying out” pixels in order of importance and measuring how it affects the prediction (Bach et al., 2015). On the other hand, graph representations tend to be less redundant and the structure of the graph is a constituent part of its identity, therefore small alterations of nodes/edges can drastically alter the meaning of the graph. In our chemistry problem, for example, replacing an atom or bond would fundamentally change a molecule or invalidate it. As a viable strategy, one could rely on domain-specific knowledge to perform such changes while remaining semantically close to the original. Alternatively, one can learn a bijective grounding of graphs onto a manifold with meaningful neighborhood to conduct such an evaluation. In Appendix C.2 we present a hand-crafted example of progressively eliminating atoms from a molecule in order of importance using domain knowledge.

6. Conclusion

As an expository paper, in this work, we introduced and focused on an important problem with impactful applications: we analyzed the major existing explanation techniques in the context of Graph Networks. We further conducted a case-based analysis on two simple but complementary tasks as well as some important high-level discussions on design choices when explaining a GNs decision. Finally, we provided an implementation of five different explanation techniques using PyTorch autograd which can be readily used for any definition of GN. In tandem with the high-level technical novelty, we hope these contributions open up fruitful discussions at the workshop and pave the road for future development of specific techniques for GN explanation for real-world applications.

Acknowledgement. Federico Baldassarre is partially funded by Swedish Research Council project 2017-04609

References

- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10(7):e0130140, 2015.
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.
- Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 1957–1967, 2017. URL <https://aclanthology.info/papers/D17-1209/d17-1209>.
- Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pp. 4502–4510, 2016.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- Beck, D., Haffari, G., and Cohn, T. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 273–283, 2018.
- Bordes, F., Berthier, T., Di Jorio, L., Vincent, P., and Bengio, Y. Iteratively unveiling new regions of interest in deep learning models. 2018.
- Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and locally connected networks on graphs. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL <http://arxiv.org/abs/1312.6203>.
- Carlsson, S., Azizpour, H., Razavian, A. S., Sullivan, J., and Smith, K. The preimage of rectifier network activities. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. URL https://openreview.net/forum?id=rk7YG_4Yg.
- Chang, M., Ullman, T., Torralba, A., and Tenenbaum, J. B. A compositional object-based approach to learning physical dynamics. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=Bkab5dqxe>.
- Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pp. 592–603, 2018.
- Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.
- Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. Protein interface prediction using graph convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 6530–6539, 2017.
- Gevrey, M., Dimopoulos, I., and Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264, 2003.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR.org, 2017.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. IEEE, 2018.
- Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pp. 1024–1034, 2017.
- Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018.
- Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.

- Kindermans, P., Schütt, K. T., Alber, M., Müller, K., Erhan, D., Kim, B., and Dähne, S. Learning how to explain neural networks: Patternnet and patternattribution. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=Hkn7CBaTW>.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kipf, T. N. and Welling, M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL <https://openreview.net/forum?id=SJU4ayYgl>.
- Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. S. Gated graph sequence neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.05493>.
- Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- Narasimhan, M., Lazebnik, S., and Schwing, A. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*, pp. 2654–2665, 2018.
- Nguyen, A. M., Yosinski, J., and Clune, J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *CoRR*, abs/1602.03616, 2016. URL <http://arxiv.org/abs/1602.03616>.
- NIPS reviews. Convolutional networks on graphs for learning molecular fingerprints. https://media.nips.cc/nipsbooks/nipspapers/paper_files/nips28/reviews/1321.html, 2015. Accessed: 2019-04-26.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710. ACM, 2014.
- Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–417, 2018.
- Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R. 3d graph neural networks for rgb-d semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5199–5208, 2017.
- Rahimi, A., Cohn, T., and Baldwin, T. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2009–2019, 2018.
- Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.
- Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., and Battaglia, P. Graph networks as learnable physics engines for inference and control. In *International Conference on Machine Learning*, pp. 4467–4476, 2018.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.
- Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153. JMLR. org, 2017.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL <http://arxiv.org/abs/1312.6034>.

- Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL <http://arxiv.org/abs/1706.03825>.
- Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6806>.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.
- Sung, A. Ranking importance of input parameters of neural networks. *Expert Systems with Applications*, 15(3-4): 405–411, 1998.
- Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL <https://openreview.net/forum?id=rJXMpikCZ>.
- Watters, N., Zoran, D., Weber, T., Battaglia, P., Pascanu, R., and Tacchetti, A. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pp. 4539–4547, 2017.
- Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.
- Zhang, X., Solar-Lezama, A., and Singh, R. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems*, pp. 4874–4885, 2018.
- Zitnik, M., Agrawal, M., and Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457, 2018.

A. Explainability techniques

A.1. Explanations for individual features

Explanations for image-based tasks usually aggregate the importance of the input features at the pixel level, e.g. by taking an average over the individual color channels. This is done under the reasonable assumption that spatial locations are the smallest unit of input that can still be interpreted by humans. The tasks considered in this work make use of node/edge features that are heterogeneous and individually interpretable. Therefore, we choose to present the explanations at the feature level, rather than aggregating at node or edge level. Furthermore, we observed that the sign of the gradients produced with Sensitivity Analysis can provide additional context to the explanation. For this reason, the visualizations in this appendix will make use of the gradient "as is" and not of its squared norm.

Overall, we observe that explanations produced by variation-based methods tend to diverge from how a human would intuitively describe the process in terms of causes and effects. Decomposition-based methods result instead in more natural explanations. We posit that the decomposition of the output signal makes LRP more suitable for the categorical distribution of the relevant features on both nodes and edges.

A.2. Layer-wise Relevance Propagation rules

Layer-wise Relevance Propagation (LRP) is a signal decomposition method introduced in (Bach et al., 2015), where the authors mainly propose two rules.

The former, known as $\alpha\beta$ -rule:

$$R_i^{(\ell)} = \sum_j \left(\alpha \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+ + b_j^+} + \beta \frac{z_{ij}^-}{\sum_{i'} z_{i'j}^- + b_j^-} \right) R_i^{(\ell+1)},$$

where $\alpha + \beta = 1$, x is the input of the layer, w are its weights, and $z_{ij} = x_i^{(\ell)} w_{ij}$.

The latter, known as ϵ -stabilized rule:

$$R_i^{(\ell)} = \sum_j \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+ + b_j^+ + \epsilon} R_i^{(\ell+1)},$$

where ϵ is a small number to avoid division by zero.

We found the former with $\beta \neq 0$ to be quite unstable in the presence of zeros in the input or in the weights, a situation that occurs often when using one-hot encoding of categorical features and L1 regularization for the weights. Therefore, despite the $\alpha\beta$ -rule should allow for more flexibility in tuning the ratio of positive and negative relevance, we chose the simpler ϵ -stabilized rule with $\epsilon = 10^{-16}$.

A.3. LRP for regression

Layer-wise Relevance Propagation was initially developed as an explanation technique for classification tasks. In the

context of our Solubility experiment, we extend its application to a regression task. Since the prediction target is now a continuous variable, the explanations produced by LRP can be interpreted as "How much does this feature of this atom/bond contribute, positively or negatively, to the final predicted value?".

Also note, that due to the use of bias terms in our networks, the conservation property of LRP does not hold in full. Some relevance, in fact, will inevitably be attributed to the biases, that are internal parameters of the model and therefore not interpretable.

B. Experiment details

B.1. Infection

FEATURE REPRESENTATION

The feature vectors $e_k \in [-1, +1]^2$ and $n_i \in [-1, +1]^4$ encode edge and node features respectively. Both include uninformative features that the network should learn to ignore and which should be attributed no importance by the explanation techniques (Fig. 6). Notably, binary features are encoded as $\{-1, +1\}$ rather than $\{0, 1\}$, while this does not affect variation-based models (SA and GPB), it facilitates the propagation of relevance to the input when LRP is used. The synthetic dataset used for training contains 100,000 with 30 or fewer nodes generated with the Barabási-Albert algorithm. The datasets used for validation and testing contain graphs of up to 60 nodes and different percentages of sick and immune nodes.

ARCHITECTURE AND TRAINING

The network used for the Infection task makes use of a single layer of graph processing as in Eq. 1, without graph-level features. The update functions for the edges and the nodes are shallow multi-layer perceptrons, with ReLU activations and we use sum/max pooling to aggregate the edges incident to a node. We use the Adam optimizer (Kingma & Ba, 2014) to minimize the binary cross-entropy between the per-node predictions and the ground truth.

Multiple choices of hyperparameters such as learning rate, number of neurons in the hidden layers and L1 regularization yield similar outcomes. Both sum and max pooling perform well, but the former fails in some corner cases (Fig. 17).

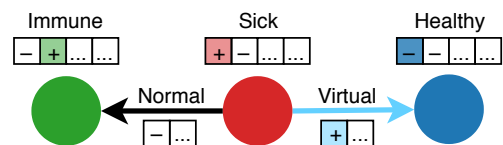


Figure 6: **Graph features for the Infection problem.**

Nodes encode whether they are sick or healthy, immune or at risk, plus two uninformative features. Edges encode whether they are virtual or not, plus a single uninformative feature.

B.2. Solubility

DATASET AND FEATURES

The Solubility dataset is the same as (Duvenaud et al., 2015), consisting of around 1000 organic molecules represented as SMILES strings and their measured solubility in water. The molecules are represented as graphs having atoms as nodes (with degree, number of hydrogens, implicit valence and type as features) and bonds as edges (with their type, whether they are conjugate and whether they are in a ring as features).

ARCHITECTURE AND TRAINING

As optimization objective we use the mean squared error between the measured log-solubility and the global features u of the output graph of a multi-layer GN, where each layer performs updates the graph as in Eq. 1. Using multiple layers of graph convolution allows the network to aggregate information at progressively larger scales, starting from the local neighborhood and extending to wider groups of atoms. Dropout is applied at the output of every linear transformation, as a technique to counteract overfitting.

We tested multiple combinations of hyperparameters and obtain results comparable to (Duvenaud et al., 2015) using 3-5 hidden graph layers with a dimensionality of either 64, 128 or 256 and sum/mean pooling for all aggregation operations. Max pooling performed much worse, probably due to the nature of the task.

C. Additional results

C.1. Infection

EXAMPLE PREDICTION ON A MEDIUM-SIZED GRAPH

In the following pages we present an in-depth comparison between the three explainability methods we experimented with. We consider a graph with multiple sources of infection and immune nodes. The network, that uses max pooling to aggregate information from the incoming edges, correctly predicts the state of every node after one step of infection propagation (Fig. 8). In the figures that follow, we present a visualization of the explanations produced for three nodes of the graph: one that becomes infected (Fig. 9), one that receives no infection from its neighbors (Fig. 10) and one that is immune (Fig. 11). Refer to the captions for observations specific to every example.

AGGREGATION: MAX VS. SUM COMPARISON

It then follows an overview of explanations obtained for smaller graphs. For every input graph, we show two predictions: one made by a GN that uses max pooling and one made by a GN that uses sum pooling. The predictions are followed by a visualization of the explanations pro-

duced with Sensitivity Analysis, Guided Backpropagation and Layer-wise Relevance Propagation (in this order). For each explainability method, we represent the values of the gradient/relevance as a heatmap over the individual features of every node/edge, as well as with a graphical representation of the graph. Refer to the captions for observations specific to every example.

C.2. Solubility

PROGRESSIVE ALTERATION OF A MOLECULE

As mentioned in the discussion (Sec. 5), choosing to model molecules as graphs yields a non-redundant and highly structured representation. As a consequence, the ability to slightly alter a molecule by performing small steps in Euclidean space is lost. This makes it hard to verify that explanations correspond to how the trained network predicts solubility. In fact, it is not possible to automatically alter a molecule according to the importance of its atoms/bonds and still obtain a valid molecule. In this case, it is necessary to apply domain-specific knowledge and identify which changes are viable in the space of valid molecules. In Figure 7 we show a trivial example where we use LRP to identify important atoms/bonds of a molecule and progressively remove them to reduce the predicted solubility.

SOLUBILITY EXPLANATIONS

In figures 18 and 19 we present a visualization of the explanation for the predicted solubility of glucose (moderately soluble) and 4-hexylresorcinol (moderately insoluble). The explanations are produced by applying LRP and propagating positive and negative relevance to the individual features of the atoms/bonds.

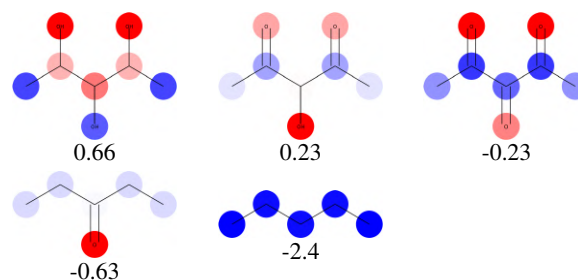


Figure 7: **Progressively removing molecular elements that are important for solubility.** The top-left molecule is rendered progressively less soluble by a) turning single C-O bonds into double C=O bonds and b) removing oxygen atoms altogether (left-to-right, top-to-bottom). Such modifications result in chemically valid molecules and are taken according to the importance attributed via LRP (red = positive relevance). The predicted solubility in log mol/L is reported under each molecule.

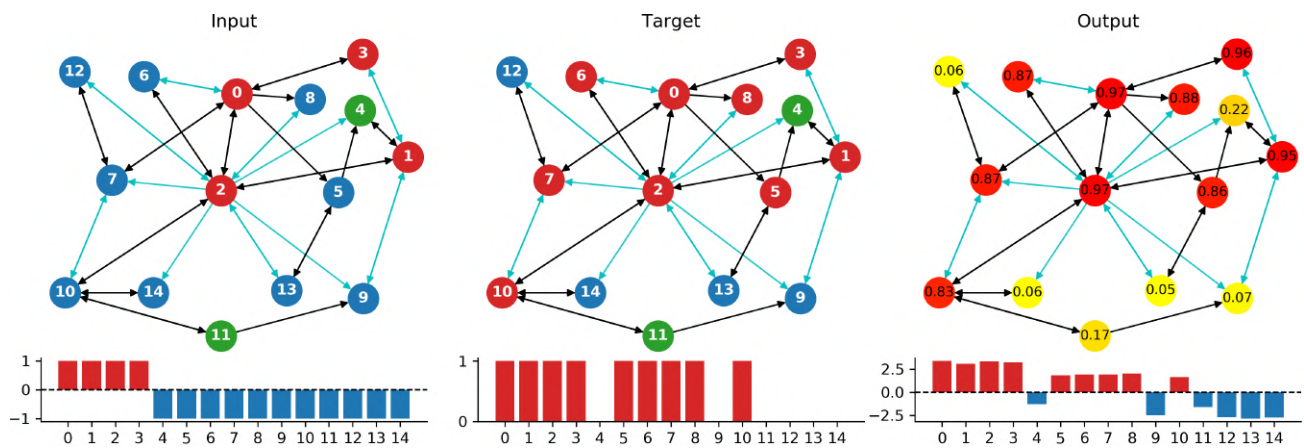


Figure 8: **Graph 1:** nodes 0, 1, 2 and 3 are initially sick; nodes 4 and 11 are immune; the others are healthy. After one propagation step, the infection reaches nodes 5, 6, 7 and 10. The network predicts the correct label for every node of the graph, following the spread of the infection along non-virtual edges to non-immune nodes. The figures that follow are a visualization of the explanations produced for nodes: 10, 13, 4

Explainability Techniques for Graph Convolutional Networks

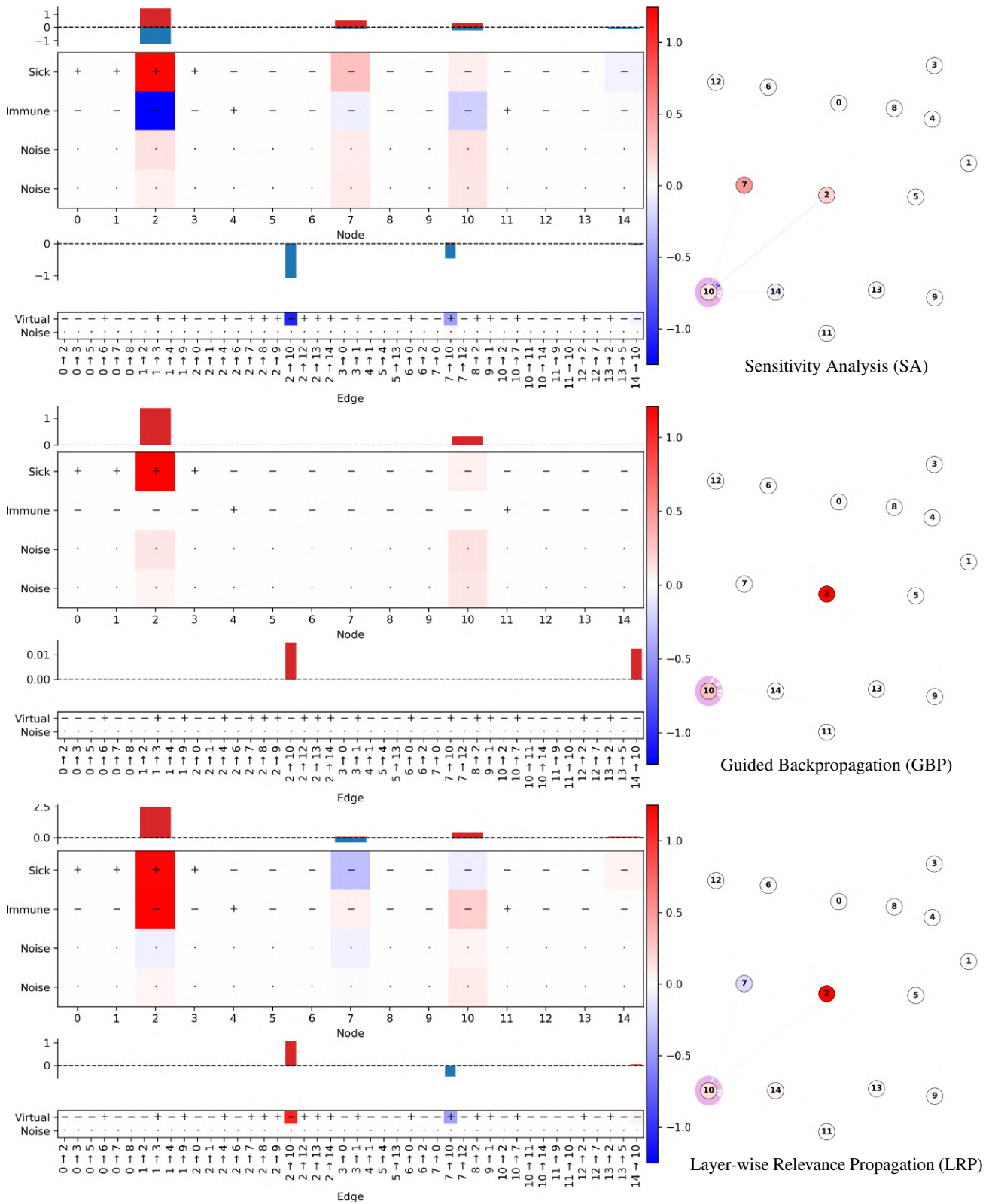


Figure 9: **Graph 1 - Explanation for node 10.** Node 10 is initially healthy and becomes infected due to the incoming edge from node 2. SA suggests that if node 2 was more sick and the edge $2 \rightarrow 10$ was less virtual, then the prediction for node 10 would be even higher. GBP yield a similar explanation, but limited to the positive gradients. LRP decomposes the positive prediction into a positive contribution from the fact that node 2 is sick and that the edge $2 \rightarrow 10$ is not virtual, and a negative contribution from the fact that node 7 is healthy and its connection to node 10 is virtual.

Explainability Techniques for Graph Convolutional Networks

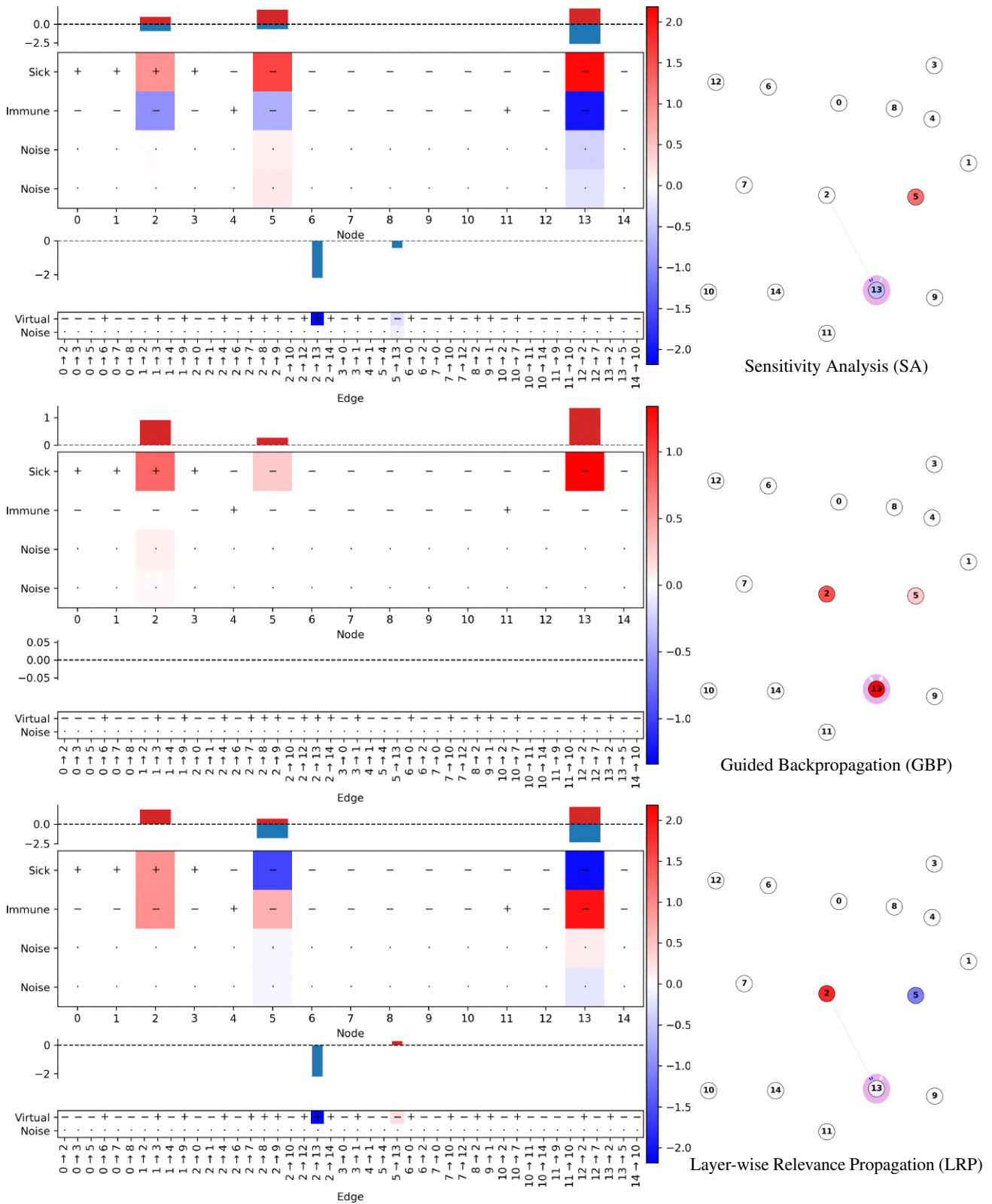


Figure 10: **Graph 1 - Explanation for node 13.** Node 13 remains healthy after one propagation step and the network correctly predicts a negative output (before activation). SA explanation suggests that for the prediction to be higher, the edge 2 → 13 should be less virtual, or banally that node 13 should be more sick and less immune. GBP produces a similar explanation, but only captures the positive gradient on the "sick" feature of node 13 itself. For LRP the fact that node 13 is initially healthy is a negative contribution to the output, together with the fact that its neighbor 5 is healthy as well; a positive contribution comes from the fact that node 2 is sick, which is counteracted by the negative contribution of the edge 2 → 13.

Explainability Techniques for Graph Convolutional Networks

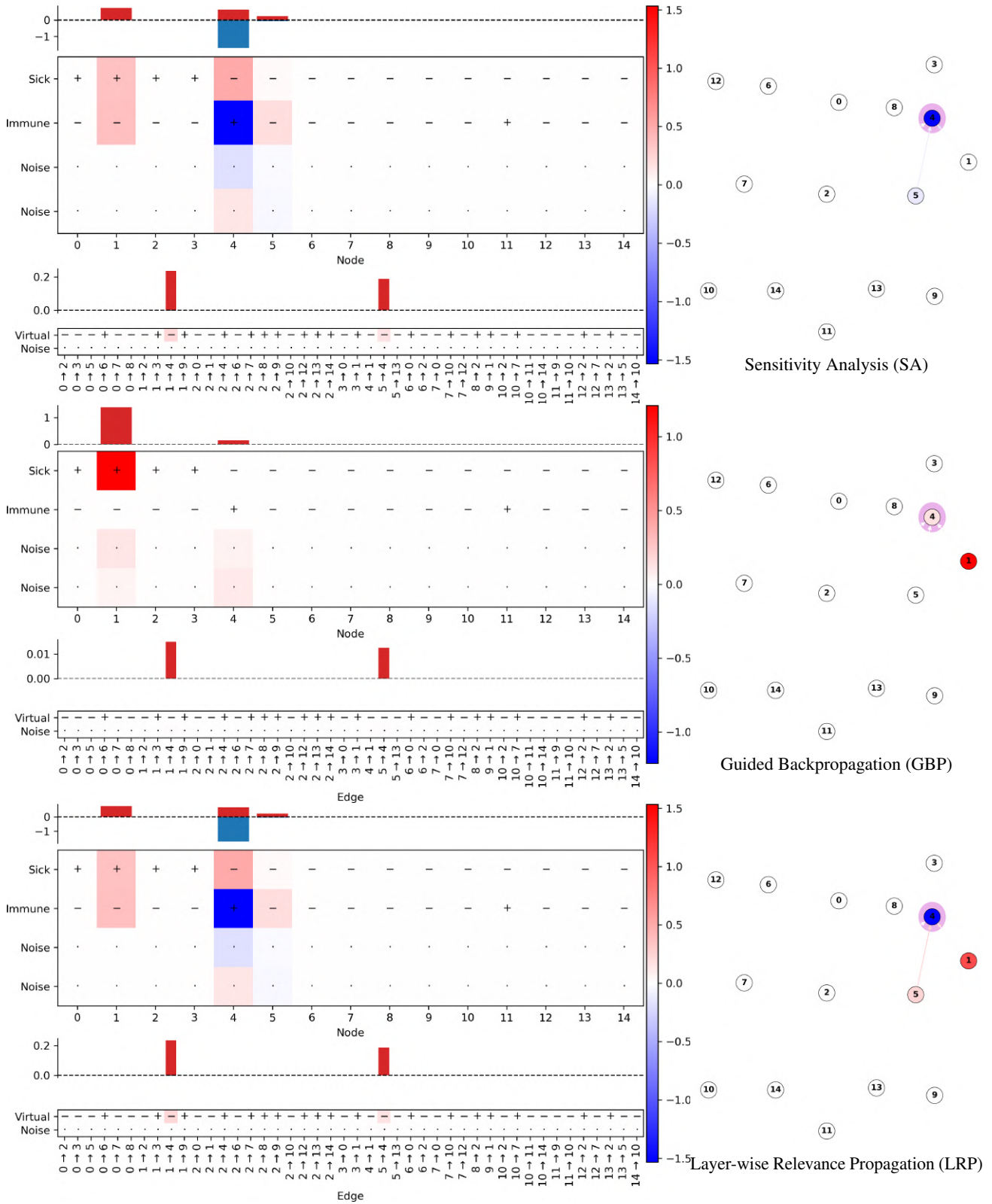


Figure 11: **Graph 1 - Explanation for node 4.** The node is immune and therefore remains healthy after one step of infection propagation. SA places a negative gradient on the "immune" feature of node 4, meaning that if it wasn't immune it would become infected. GBP focuses on node 1, which is a sick neighbor of node 4. LRP decompose most of the negative prediction into a negative contribution from the node itself, weakly counteracted by positive contributions from its neighbors.

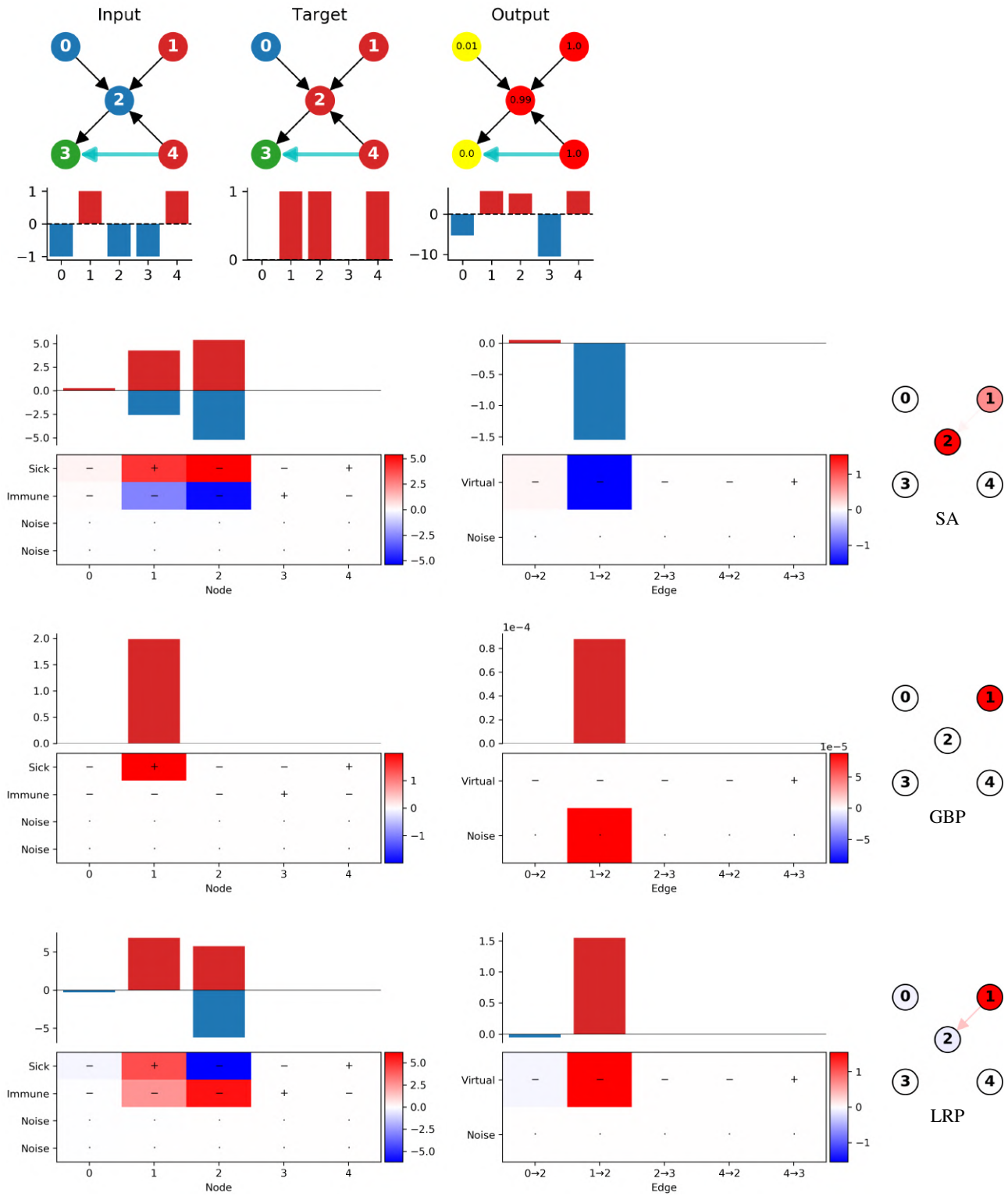


Figure 12: **Graph 2 - Max pooling.** The healthy central node, on which the explanation is focused, has two neighbors that are sick. All methods are only able to identify one neighbor as a possible explanation for the network’s prediction, due to the fact that max pooling is used to aggregate incoming features. This example is similar to Fig. 5 in the main text.

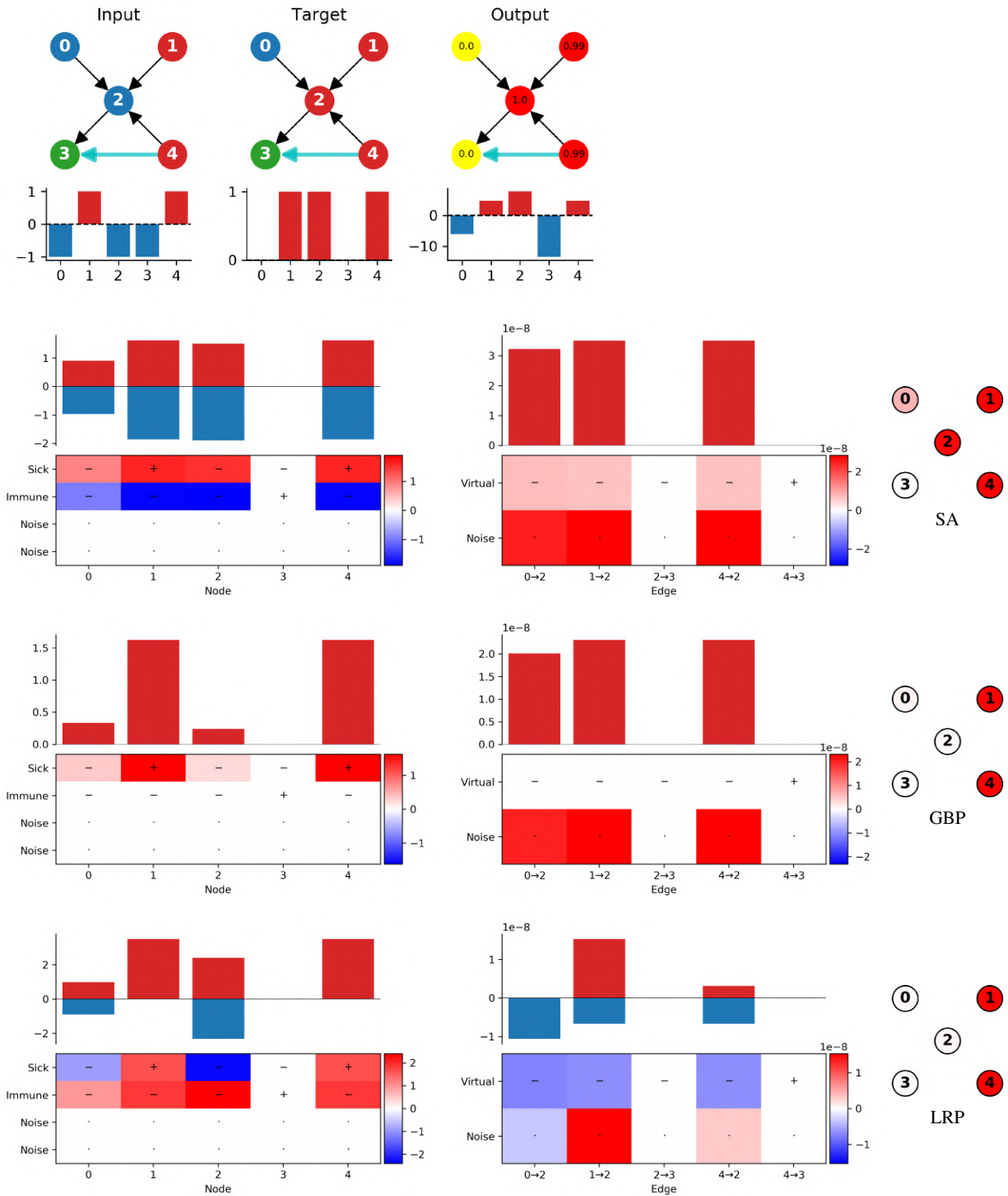


Figure 13: **Graph 2 - Sum pooling.** Compared to the max pooling version, the network that uses sum pooling produces explanations that include all possible sources of infection. However, sum pooling might result in incorrect predictions, as demonstrated in Fig. 17. This example is similar to Fig. 5 in the main text.

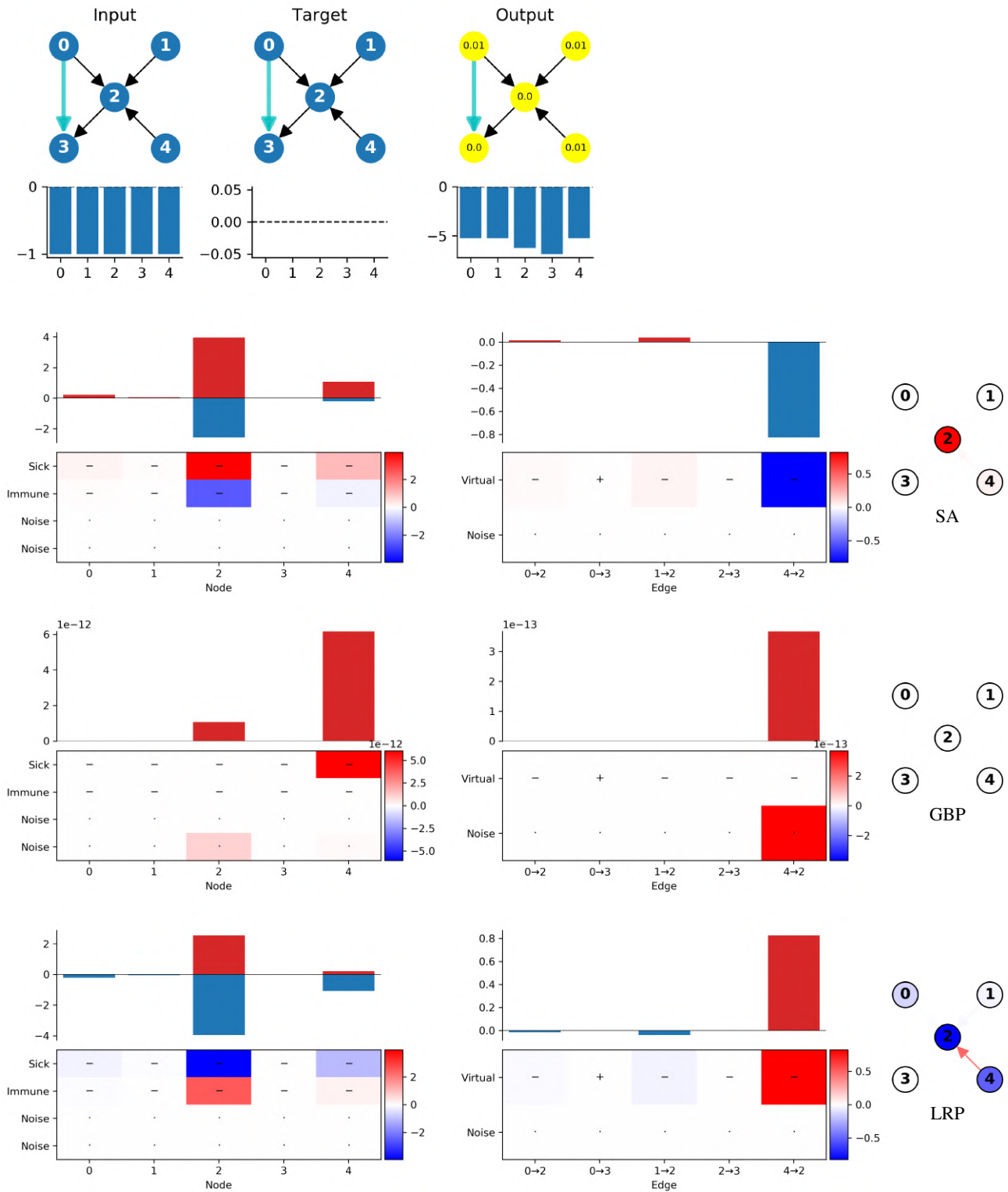


Figure 14: **Graph 3 - Max pooling.** All nodes are healthy, the explanation focuses on node 2, trying to explain why the network predicts it will remain healthy. SA assigns a positive gradient to the "sick" feature of node 2 itself, indicating that if the node was more sick the prediction would also shift towards mode positive values, i.e. sick. LRP decomposes most of the negative prediction onto the "sick" feature of node 2 itself, which is similar to SA, but more intuitive from a user perspective.

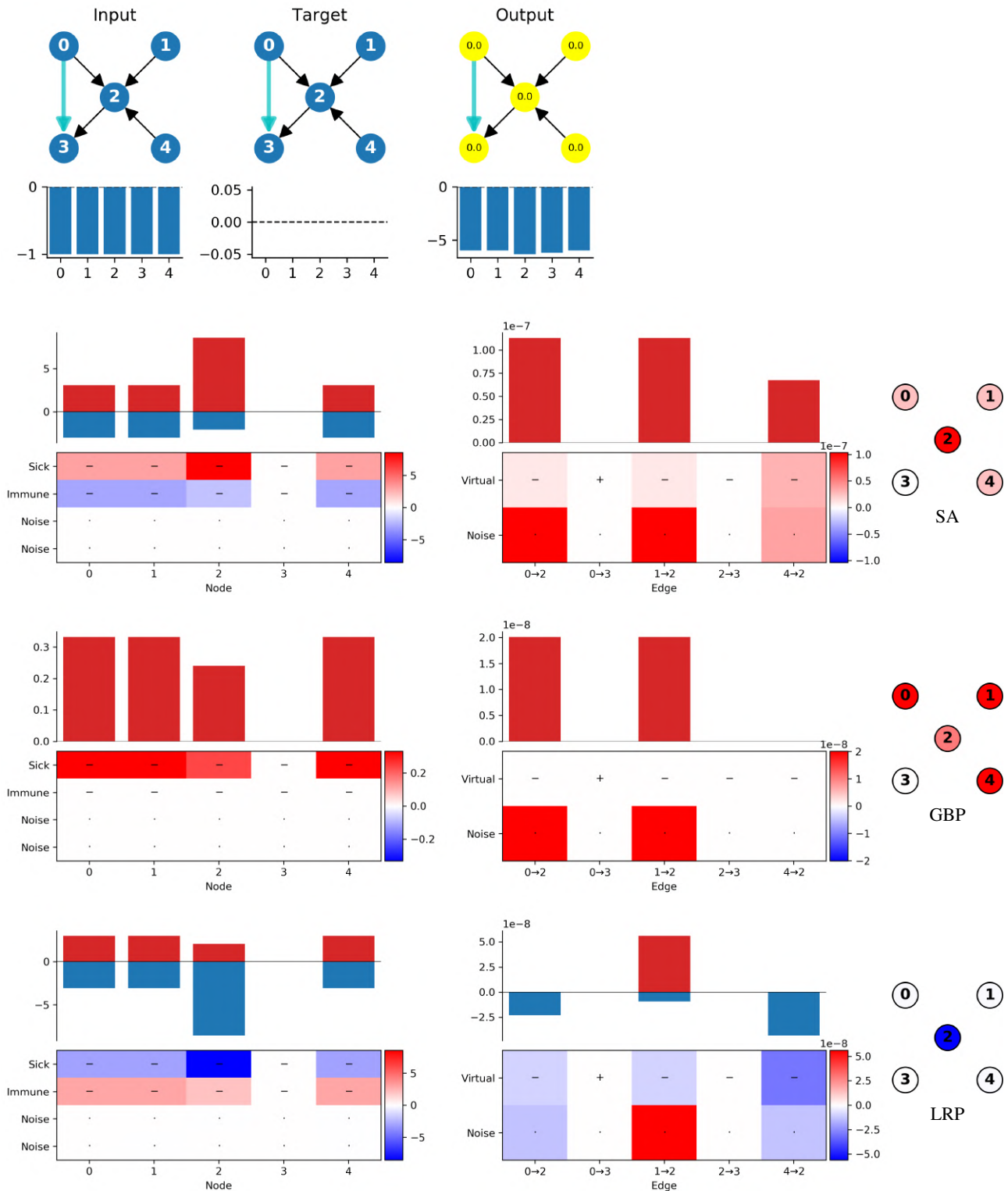


Figure 15: **Graph 3 - Sum pooling.** All nodes are healthy and the network outputs the correct prediction. The explanation focuses on node 2. Both SA and GBP focus on what would need to happen for the prediction to change, e.g. the surrounding nodes should be more sick and less immune for the central node to become infected. LRP instead decomposes the negative prediction into several negative contributions, the largest of which is attributed to the central node's feature of not being sick.

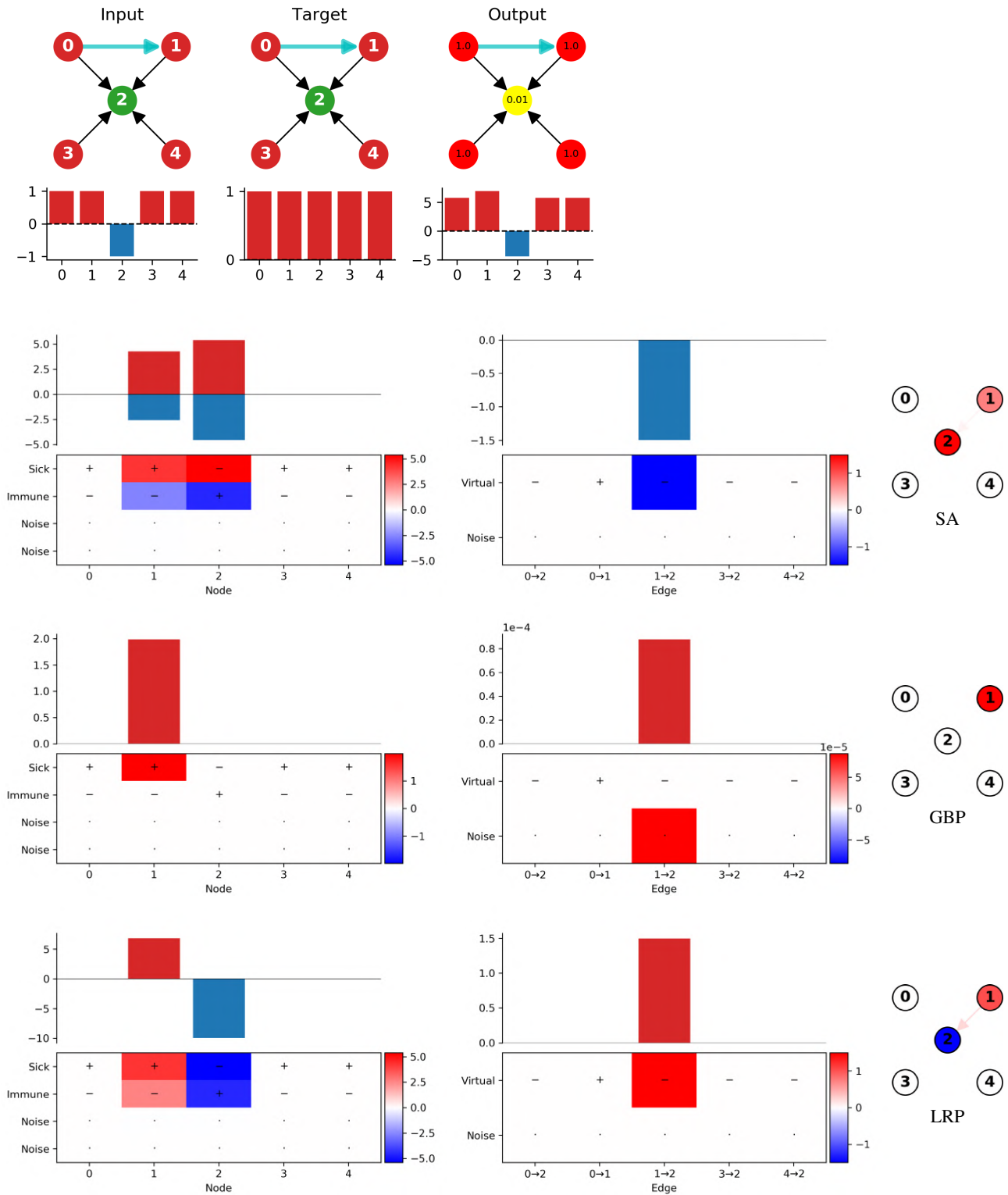


Figure 16: **Graph 4 - Max pooling.** Node 2 is immune and surrounded by sick nodes, the network is able to produce the correct answer, since max pooling is invariant w.r.t. the number of edges connected to node 2. Both SA and LRP identify the feature of being immune as important, while GBP does not consider it as its gradient would be negative.

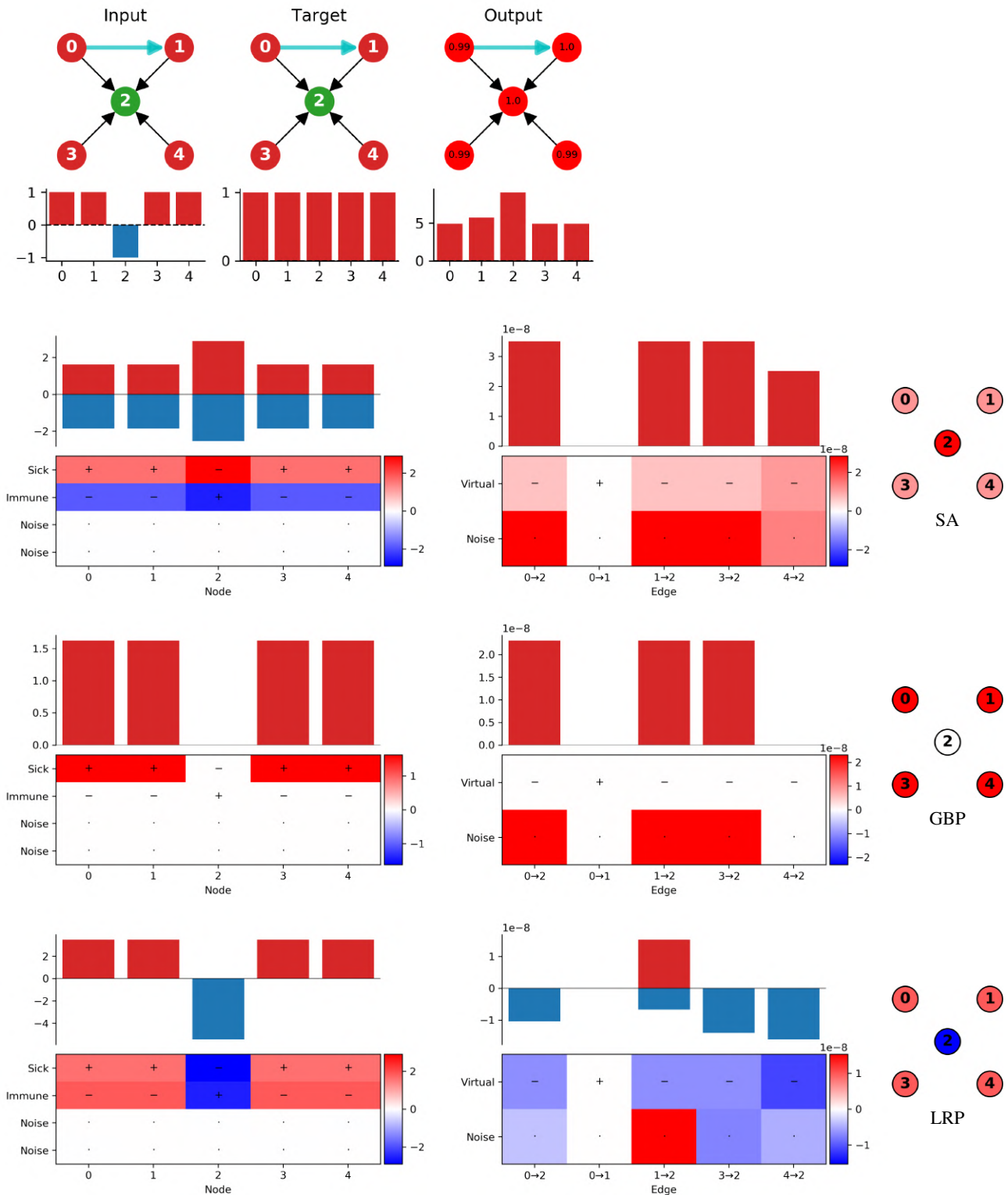


Figure 17: **Graph 4 - Sum pooling.** In this case, the network trained with sum pooling fails to produce the correct prediction. In contrast to the image domain, where the number of pixels in a convolutional neighborhood is fixed, GNs aggregate information over a potentially very large neighborhood, where even small contributions can add up and produce a wrong answer.

Explainability Techniques for Graph Convolutional Networks

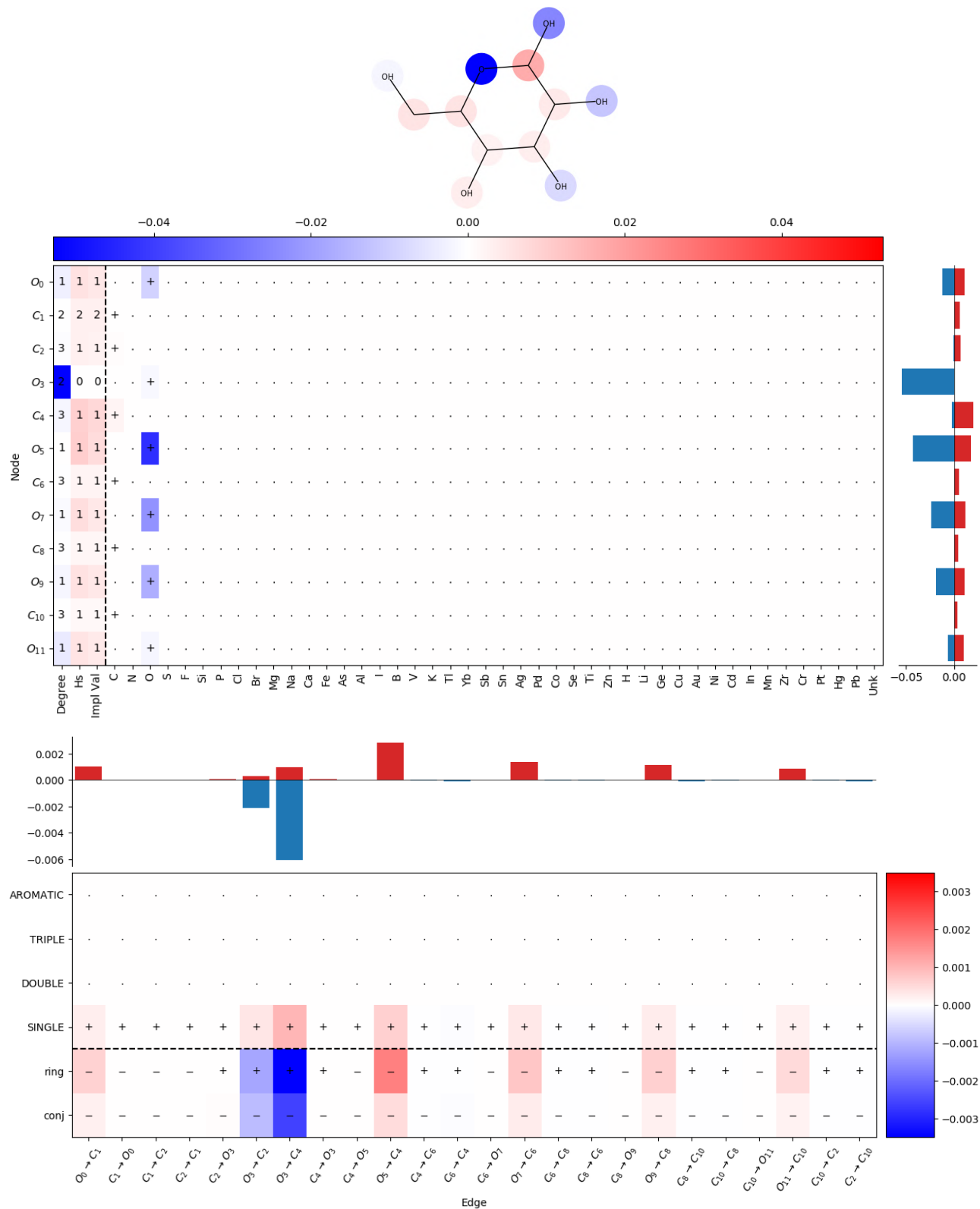


Figure 18: **Solubility of glucose:** predicted 0.74 log mol/L, measured 0.67 log mol/L, explanation over the atom and bond features using Layer-wise Relevance Propagation. Note how the main negative contribution is assigned to the oxygen atom that renders the ring heterocyclic.

Explainability Techniques for Graph Convolutional Networks

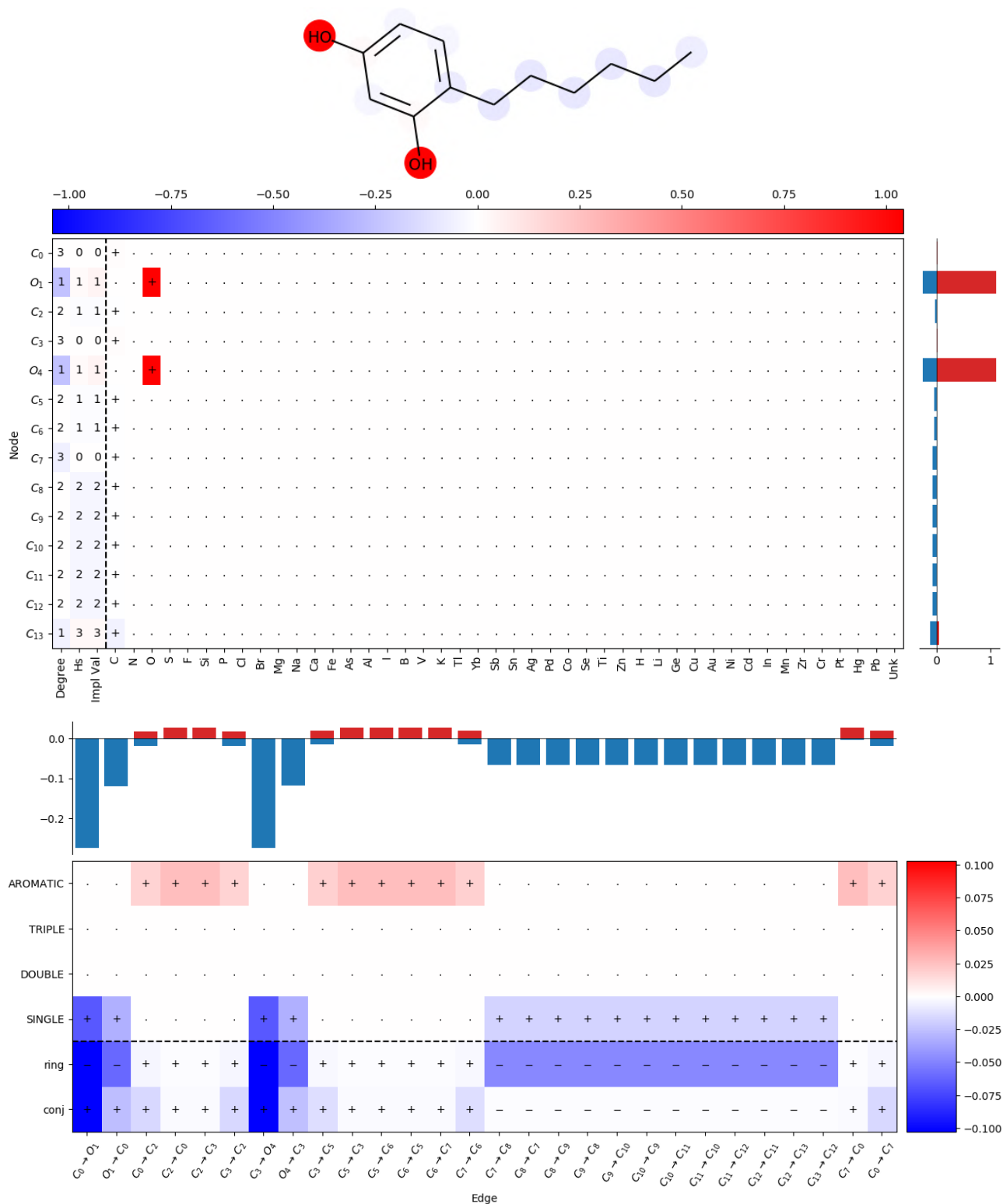


Figure 19: **Solubility of 4-hexylresorcinol:** predicted -2.59 log mol/L, measured -2.54log mol/L, explanation over the atom and bond features using Layer-wise Relevance Propagation. The explanation attributes negative relevance to the long hydrocarbon chain and positive relevance to the OH groups linked to the aromatic ring, which is in agreement with common knowledge in organic chemistry.