Prime Implicant Explanations for Reaction Feasibility Prediction

Klaus Weinbauer, Tieu-Long Phan, Peter F. Stadler, Thomas Gärtner, Sagar Malhotra

September 15, 2025

AIMLAI at ECMLPKDD 2025

Prime Implicant (PI) Explanations

Prime Implicant Explanations

Prime implicant (PI) explanations were first introduced by Shih et al. in 2018 to obtain symbolic explanations of Bayesian network classifiers.

Prime Implicant Explanations

Prime implicant (PI) explanations were first introduced by Shih et al. in 2018 to obtain symbolic explanations of Bayesian network classifiers.

Definition (PI explanation Shih et al (2018))

Let $f(X)$ be a given decision function. A PI explanation of a decision $f(x)$ is a partial instance z such that

(a) $z \subseteq x$,

(b) $f(x) = f(x')$ for every $x' \supseteq z$, and

(c) no other partial instance $y \subset z$ satisfies (a) and (b).

Prime Implicant Explanations for Graph Classification

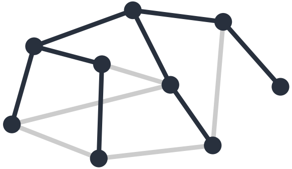
Inspired by general PI explanations.

Explanations are minimally sufficient subgraphs for a decision.

Prime Implicant Explanations for Graph Classification

Inspired by general PI explanations.

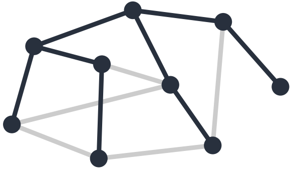
Explanations are minimally sufficient subgraphs for a decision.



Prime Implicant Explanations for Graph Classification

Inspired by general PI explanations.

Explanations are minimally sufficient subgraphs for a decision.



Definition (Subgraph PI explanation (Azzolin et al (2025)))

Let $h : \mathcal{G} \rightarrow \{0, 1\}$ be the binary classification function and $G \in \mathcal{G}$ the graph instance. A PI explanation is a graph Z such that

(a) $Z \subseteq G$,

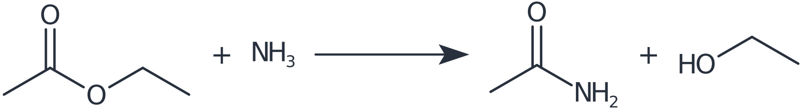
(b) $h(Z') = h(G)$ for all $Z \subseteq Z' \subseteq G$,

(c) and no proper subgraph $Z'' \subset Z$ satisfies (a) and (b).

PI Explanations for Reaction Feasibility Prediction

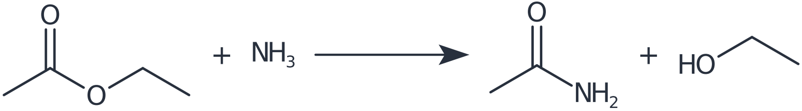
Goal of Prime Implicant Reaction Explanation

Human: Is this reaction feasible?



Goal of Prime Implicant Reaction Explanation

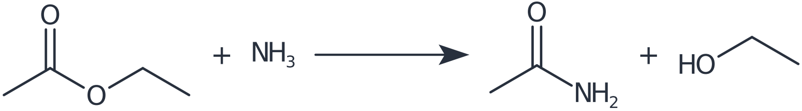
Human: Is this reaction feasible?



Classifier: Yes.

Goal of Prime Implicant Reaction Explanation

Human: Is this reaction feasible?

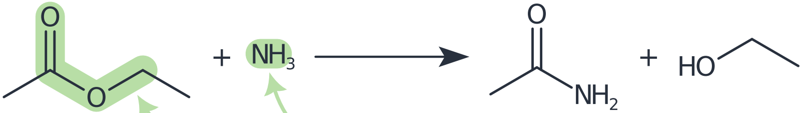


Classifier: Yes.

Human: Why is it feasible?

Goal of Prime Implicant Reaction Explanation

Human: Is this reaction feasible?



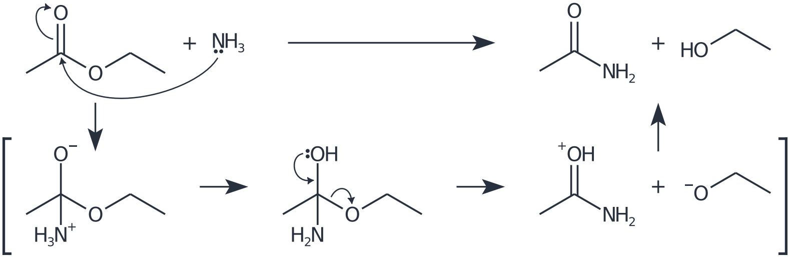
Classifier: Yes.

Human: Why is it feasible?

Explanation Method: **Because of this substructure.**

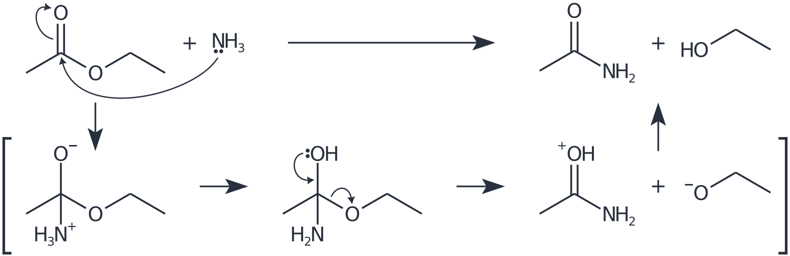
Imaginary Transition State (ITS) Graph

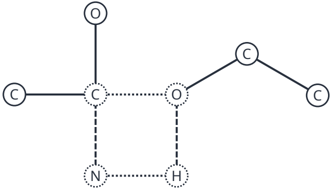
Fujita (1986)



Imaginary Transition State (ITS) Graph

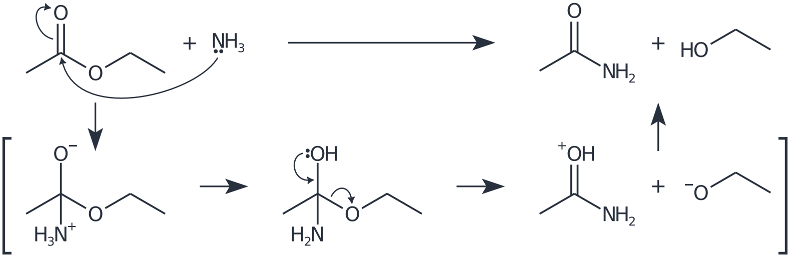
Fujita (1986)

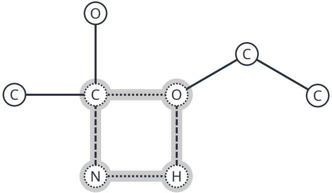




Imaginary Transition State (ITS) Graph

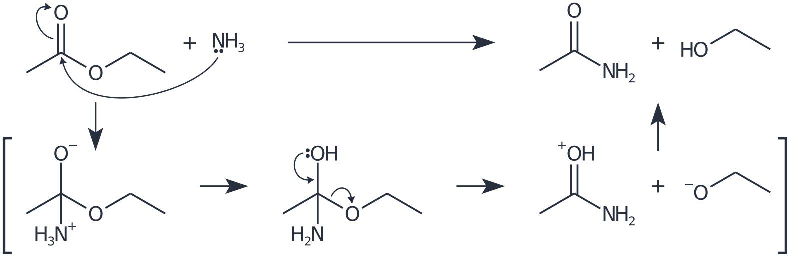
Fujita (1986)

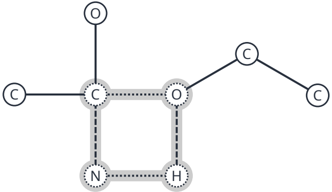




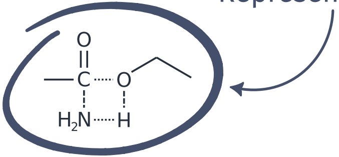
Imaginary Transition State (ITS) Graph

Fujita (1986)





Compact Representation



Domain Specific Constraints

The relevant ITS graph is

connected

and each ITS graph contains a necessary reaction center $R \subseteq G$.

Domain Specific Constraints

The relevant ITS graph is

connected

and each ITS graph contains a necessary reaction center $R \subseteq G$. Hence, the explanations are

rooted

at the reaction center R. (González Laffitte et al (2024))

Domain Specific Constraints

The relevant ITS graph is

connected

and each ITS graph contains a necessary reaction center $R \subseteq G$. Hence, the explanations are

rooted

at the reaction center R. (González Laffitte et al (2024))

Still exponential running time but solvable for small instances.

Prime Implicant Reaction Explanation

Definition (PI reaction explanation)

Let $h : \mathcal{G} \rightarrow \{0, 1\}$ be a reaction feasibility classifier, and $G \in \mathcal{G}$ be an instance from the class of connected ITS graphs with R denoting its reaction center. A PI reaction explanation is a graph Z such that

(a) $R \subseteq Z \subseteq G$,

(b) $h(Z') = h(G)$ for all $Z \subseteq Z' \subseteq G$,

(c) Z is connected,

(d) and no proper subgraph $Z'' \subset Z$ satisfies (a) to (c).

Computing PI Reaction Explanations

Extension Construction

Finding PI Explanations

Computing PI Reaction Explanations

Extension Construction

Algorithm adapted from Alokshiya et al. (Alokshiya et al (2019)) based on reverse search (Avis and Fukuda (1996))

Finding PI Explanations

Computing PI Reaction Explanations

Extension Construction

Algorithm adapted from
Alokshiya et al. (Alokshiya et al (2019))
based on reverse search (Avis and Fukuda (1996))

Extensions represented as DAG
⇒ Nodes are subgraphs
⇒ Edges are subgraph relations

Finding PI Explanations

Computing PI Reaction Explanations

Extension Construction

Algorithm adapted from
Alokshiya et al. (Alokshiya et al (2019))
based on reverse search (Avis and Fukuda (1996))

Extensions represented as DAG
⇒ Nodes are subgraphs
⇒ Edges are subgraph relations

Partial order (lattice) induced by subgraph relations.

Hasse diagram of possible extensions.

Finding PI Explanations

Computing PI Reaction Explanations

Extension Construction

Algorithm adapted from Alokshiya et al. (Alokshiya et al (2019)) based on reverse search (Avis and Fukuda (1996))

Extensions represented as DAG
⇒ Nodes are subgraphs
⇒ Edges are subgraph relations

Partial order (lattice) induced by subgraph relations.

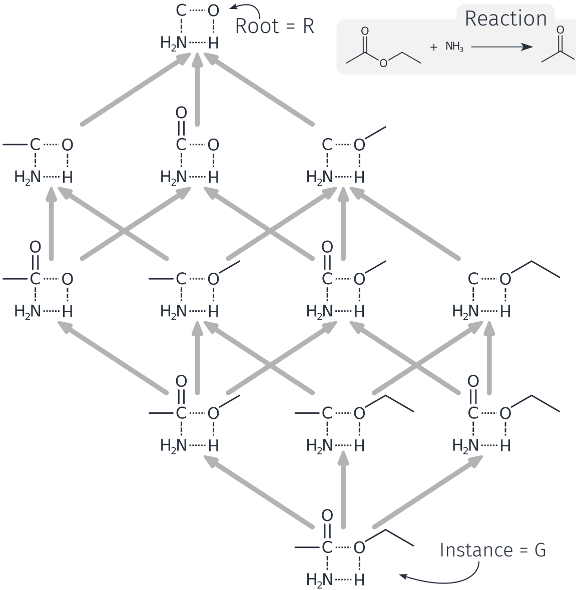
Hasse diagram of possible extensions.

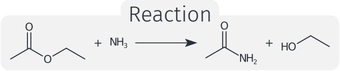
Finding PI Explanations

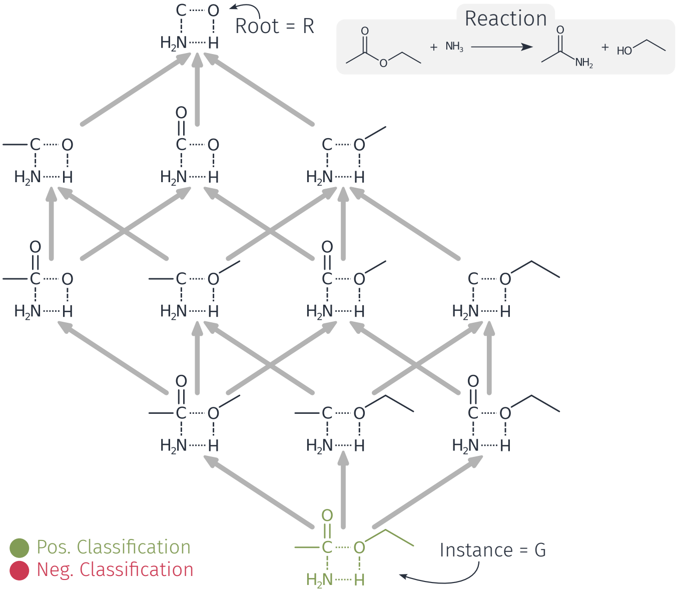
Operates on extension DAG.

Queries classifier with selected extensions.

Annotates and prunes the extension DAG until all PI explanations are found.

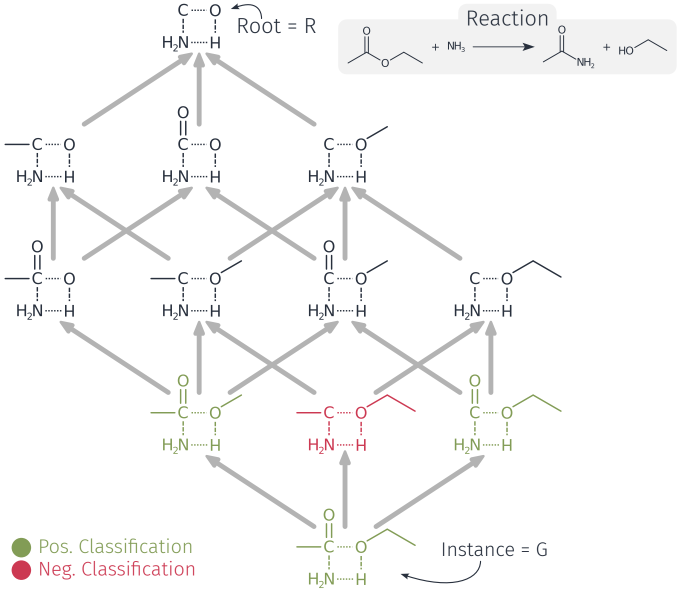

Root = R
Instance = G

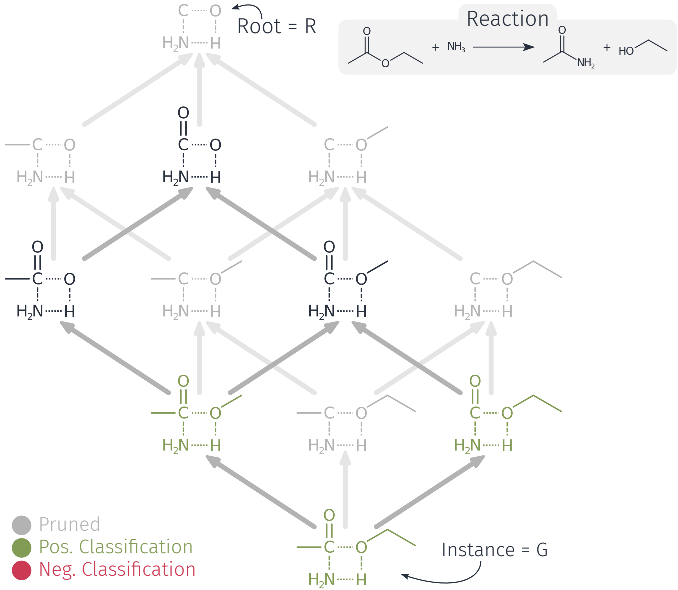

Reaction

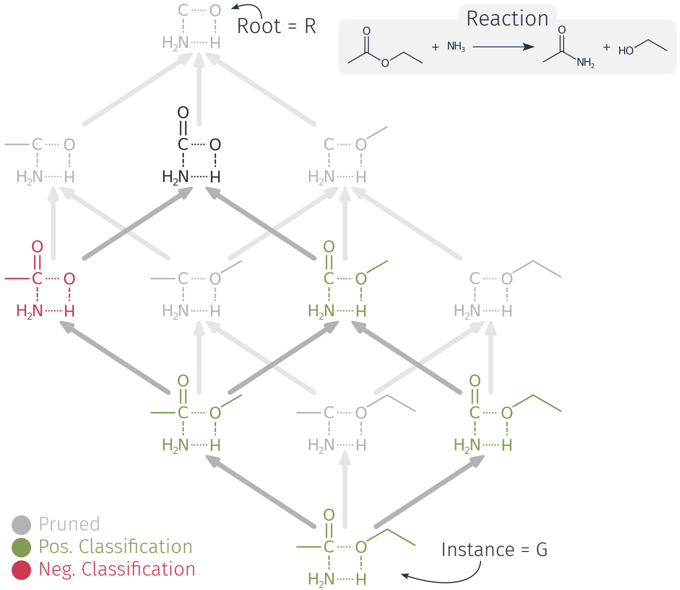


🟢 Pos. Classification

🔴 Neg. Classification


Root = R
Reaction
+ NH3
+ HO
Instance = G
● Pos. Classification
● Neg. Classification


Root = R
Reaction
+ NH3
+ HO
Instance = G
● Pruned
● Pos. Classification
● Neg. Classification



Root = R

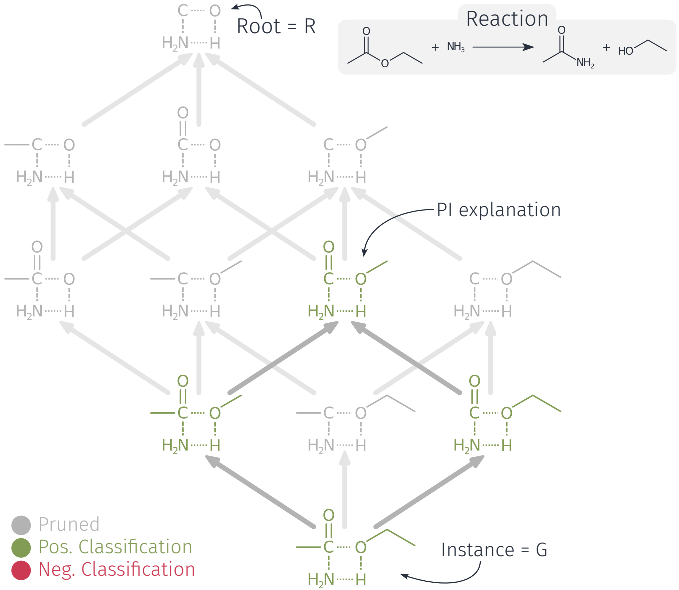
Reaction

Instance = G

● Pruned

● Pos. Classification

● Neg. Classification


Root = R
Reaction
PI explanation
Instance = G
● Pruned
● Pos. Classification
● Neg. Classification

Experimental Evaluation

Do PI reaction explanations capture what a chemist would consider the structural cause of the reaction?

Experimental Evaluation

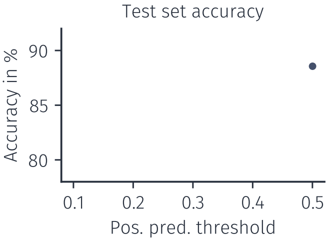
Do PI reaction explanations capture what a chemist would consider the structural cause of the reaction?

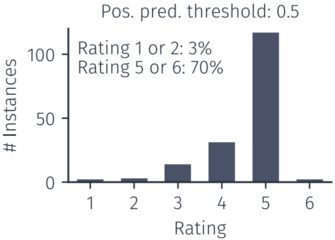
Yes

Are PI reaction explanations readily interpretable by chemists?

No

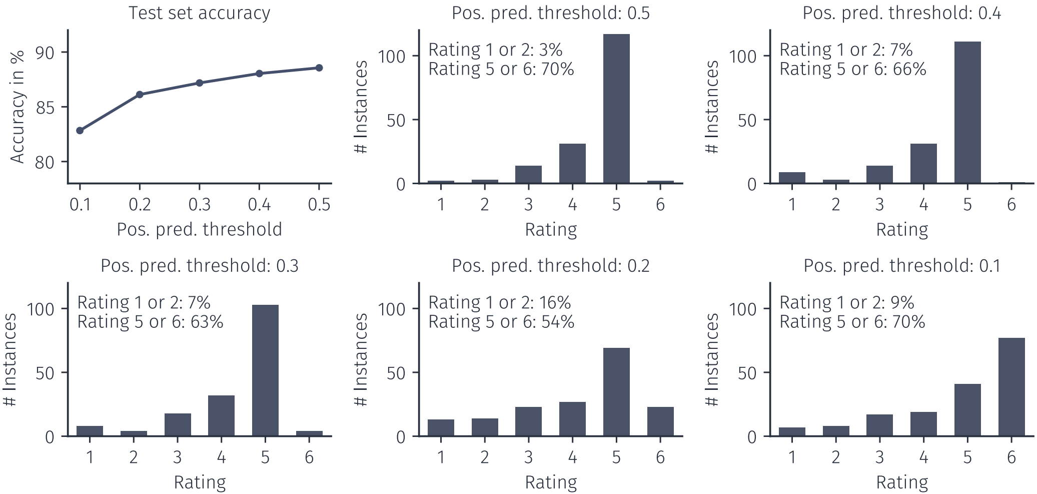
Experimental Results


Test set accuracy


Pos. pred. threshold: 0.5

Classifier: GIN, 5 layers of size 32, dropout 0.04, max pool, lr 0.003   Dataset: 6094 train / 1524 test   Test Acc=86.1±2.1 AUROC=93.2±1.4 (10 runs)

Experimental Results



Classifier: GIN, 5 layers of size 32, dropout 0.04, max pool, lr 0.003   Dataset: 6094 train / 1524 test   Test Acc=86.1±2.1 AUROC=93.2±1.4 (10 runs)

Summary, Limitations, and Further Directions

PI reaction explanations contain a chemist’s notion of cause, but are generally not human interpretable.

Provide valuable insights into model decisions.

Summary, Limitations, and Further Directions

PI reaction explanations contain a chemist’s notion of cause,
but are generally not human interpretable.

Provide valuable insights into model decisions.

Computational intractability of the presented method.

Lack of benchmarks for reaction feasibility explanations.

Which capabilities of PI reaction explanations remain to be explored?

Questions?

Klaus Weinbauer

Tieu-Long Phan

Peter F. Stadler

Thomas Gärtner

Sagar Malhotra

Contact:

Klaus Weinbauer
klaus.weinbauer@tuwien.ac.at

TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

UNIVERSITÄT LEIPZIG

TACsy Training Alliance for Computational systems chemistry

Funded by the European Union

MARIE CURIE ACTIONS

References

Mohammed Alokshiya, Saeed Salem, Fidaa Abed (2019) A linear delay algorithm for enumerating all connected induced subgraphs. BMC Bioinformatics 20(12):319, DOI 10.1186/s12859-019-2837-y, URL https://doi.org/10.1186/s12859-019-2837-y

David Avis, Komei Fukuda (1996) Reverse search for enumeration. Discrete Applied Mathematics 65(1):21–46, DOI https://doi.org/10.1016/0166-218X(95)00026-N, URL https://www.sciencedirect.com/science/article/pii/0166218X9500026N, first International Colloquium on Graphs and Optimization

Steve Azzolin, Sagar Malhotra, Andrea Passerini, Stefano Teso (2025) Beyond topological self-explainable gnns: A formal explainability perspective. URL https://arxiv.org/abs/2502.02719, 2502.02719

Shinsaku Fujita (1986) Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. Journal of Chemical Information and Computer Sciences 26

Marcos E González Laffitte, Klaus Weinbauer, Tieu-Long Phan, Nora Beier, Nico Domschke, Christoph Flamm, Thomas Gatter, Daniel Merkle, Peter F Stadler (2024) Partial imaginary transition state (its) graphs: A formal framework for research and analysis of atom-to-atom maps of unbalanced chemical reactions and their completions. Symmetry 16(9), DOI 10.3390/sym16091217, URL https://www.mdpi.com/2073-8994/16/9/1217

Andy Shih, Arthur Choi, Adnan Darwiche (2018) A symbolic approach to explaining bayesian network classifiers. arXiv preprint arXiv:180503364