Karlsruher Institut für Technologie

Discovering Structure-Property Relationships with Self-Explaining Graph Neural Networks

Jonas Teufel · AIMAT Group
Institute of Nanotechnology · INT

Why XAI for chemistry?

Why XAI for chemistry?

Does the image show a cat or a dog?

Input Image

AI Model

Computer Vision Expert

Why XAI for chemistry?

Karlsruher Institut für Technologie

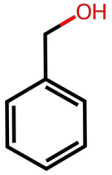
Does the image show a cat or a dog?

Input Image

AI Model

Computer Vision Expert

How soluble is the molecule?



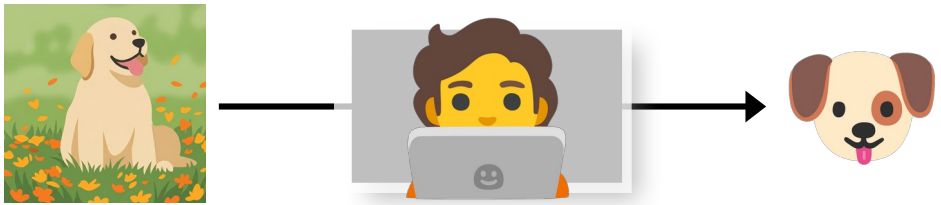
Input Molecule

AI Model

Chemistry Expert

Why XAI for chemistry?

Does the image show a cat or a dog?

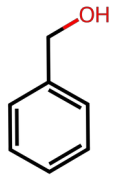


Input Image

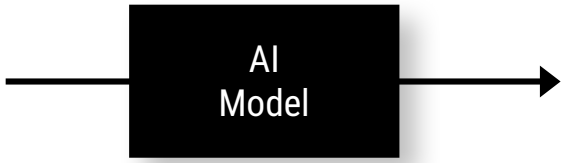
Computer Vision Expert

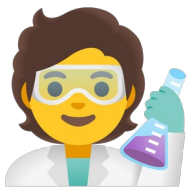
"I could do this myself, but it is easier with a model!"

How soluble is the molecule?



Input Molecule



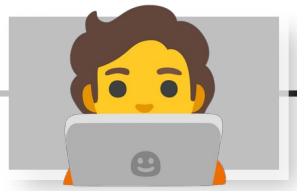


Chemistry Expert

Why XAI for chemistry?

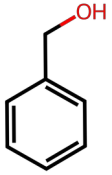
Does the image show a cat or a dog?

Input Image



Computer Vision Expert

How soluble is the molecule?



Input Molecule

Chemistry Expert

“How does the model predict this complex property?”

Why XAI for chemistry?

Does the image show a cat or a dog?

Input Image

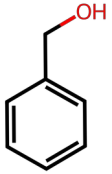
AI Model

Explanation

xAI

Computer Vision Expert

How soluble is the molecule?



Input Molecule

AI Model

Chemistry Expert

Why XAI for chemistry?

Karlsruher Institut für Technologie

Does the image show a cat or a dog?

Input Image

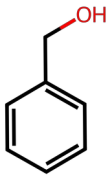
AI Model

Explanation

Computer Vision Expert

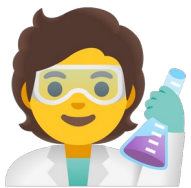
"Does the model work as intended?
Is the model fair?
Are decisions tranparent?"

How soluble is the molecule?



Input Molecule

AI Model



Chemistry Expert

Why XAI for chemistry?

Does the image show a cat or a dog?

Input Image

AI Model

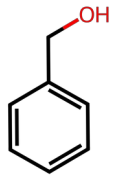
Explanation

xAI

Computer Vision Expert

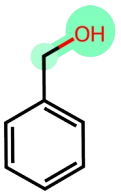
"Does the model work as intended?
Is the model fair?
Are decisions tranparent?"

How soluble is the molecule?



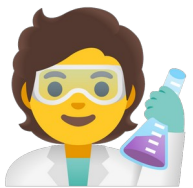
Input Molecule

AI Model



Explanation

xAI



Chemistry Expert

Why XAI for chemistry?

Karlsruher Institut für Technologie

Does the image show a cat or a dog?

Input Image

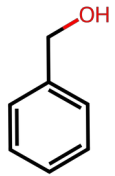
AI Model

Explanation

Computer Vision Expert

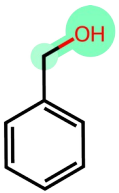
"Does the model work as intended?
Is the model fair?
Are decisions tranparent?"

How soluble is the molecule?

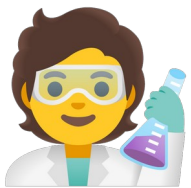


Input Molecule

AI Model



Explanation



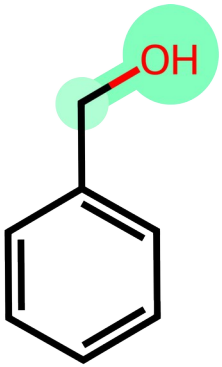
Chemistry Expert

"Now I understand how it works!"

Overview

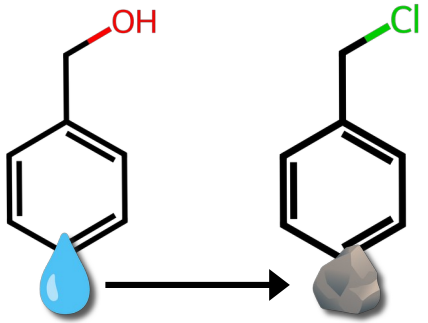
02 Attributional Explanations

Which parts are especially important?



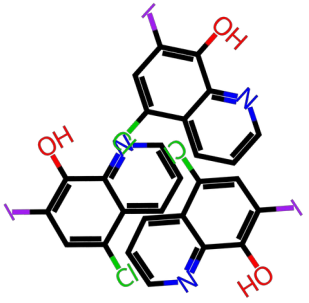
03 Counterfactuals

What would I have to change in the input to change the outcome?



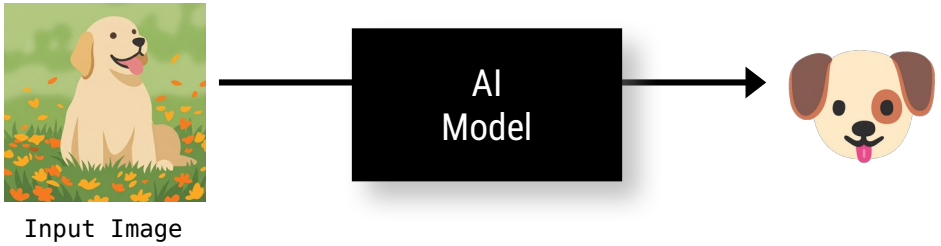
04 Case Study: Aggregation

When do small molecules form clusters and when don't they?



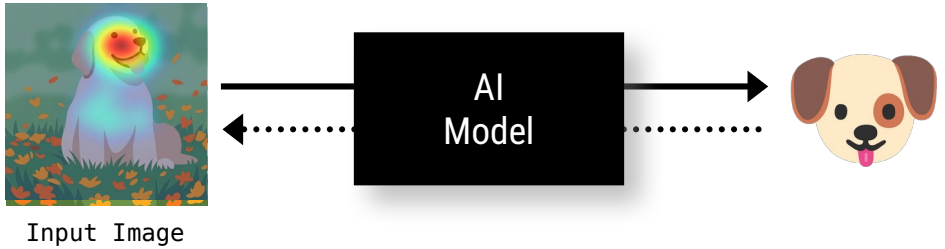
What are attributional explanations?

Attributional Explanations



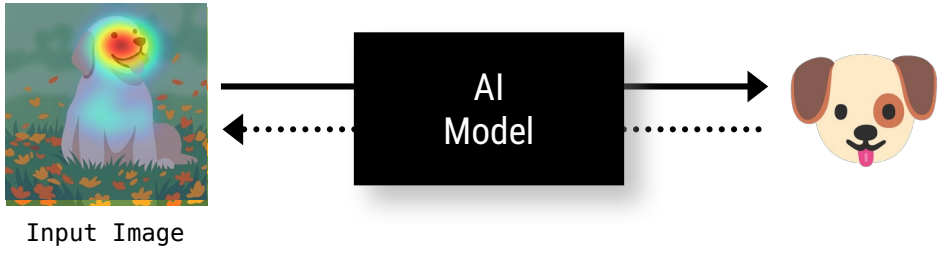
What are attributional explanations?

Attributional Explanations



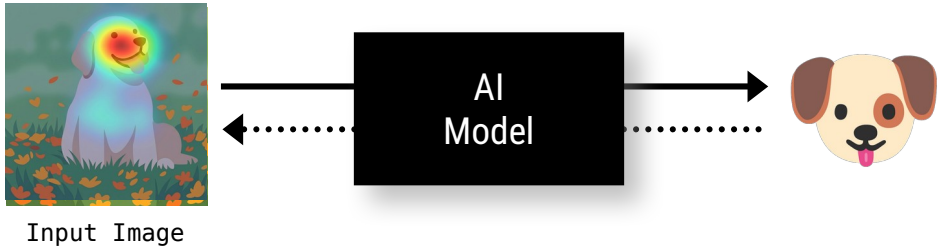
What are attributional explanations?

Attributional Explanations



What are attributional explanations?

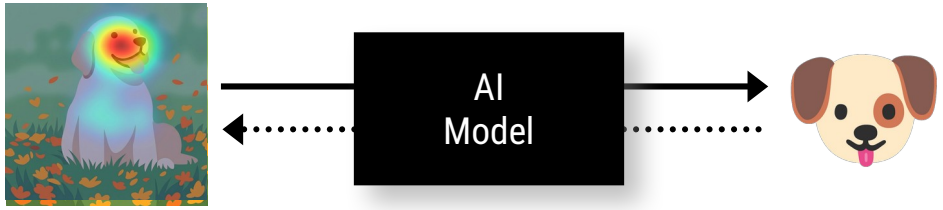
Attributional Explanations

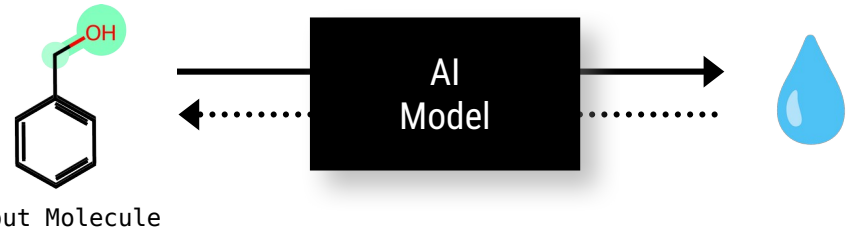


Attributional Explanations highlight parts of the input, the model considers to be especially important!

What are attributional explanations?

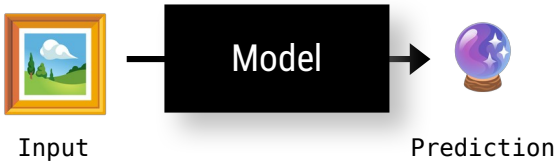
Attributional Explanations





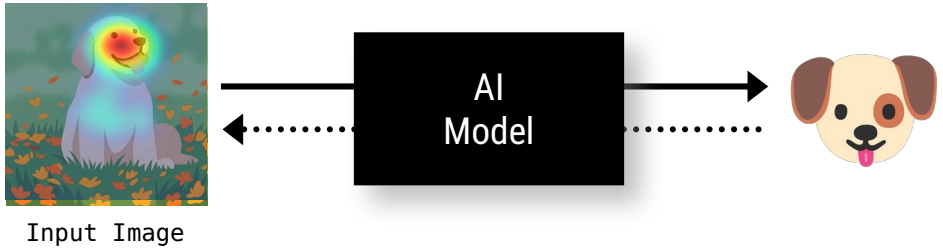
Attributional Explanations highlight parts of the input, the model considers to be especially important!

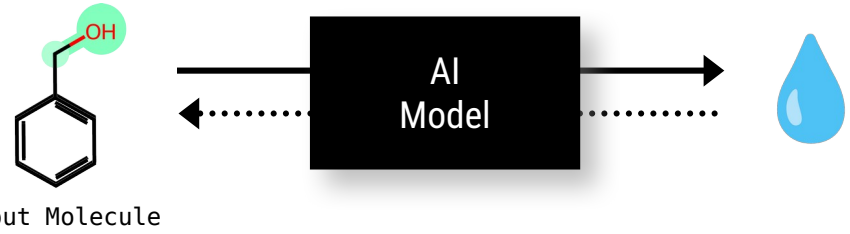
Post-Hoc Explanations



What are attributional explanations?

Attributional Explanations



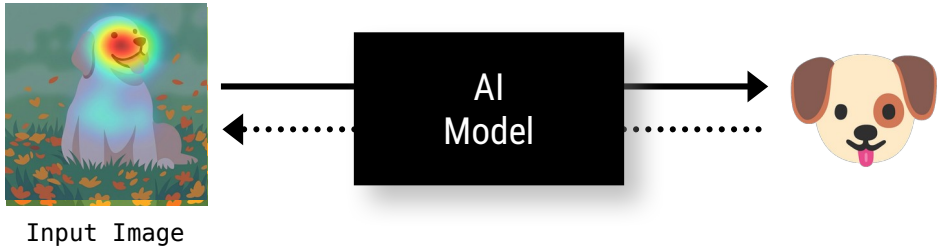


xAI Attributional Explanations highlight parts of the input, the model considers to be especially important!

Post-Hoc Explanations

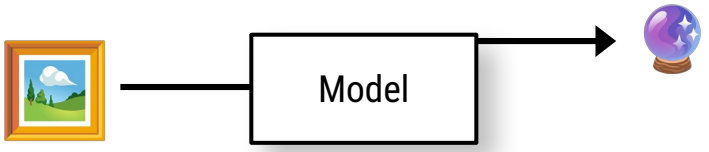
What are attributional explanations?

Attributional Explanations



Post-Hoc Explanations

Self-Explaining Approaches



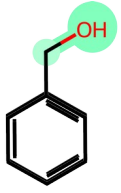
xAI Attributional Explanations highlight parts of the input, the model considers to be especially important!

What are attributional explanations?

Attributional Explanations

Input Image

AI Model



Input Molecule

AI Model

Attributional Explanations highlight parts of the input, the model considers to be especially important!

Post-Hoc Explanations

Input

Model

Prediction

Explainer

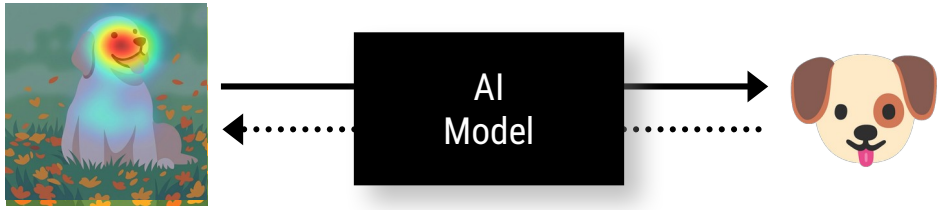
Explanation

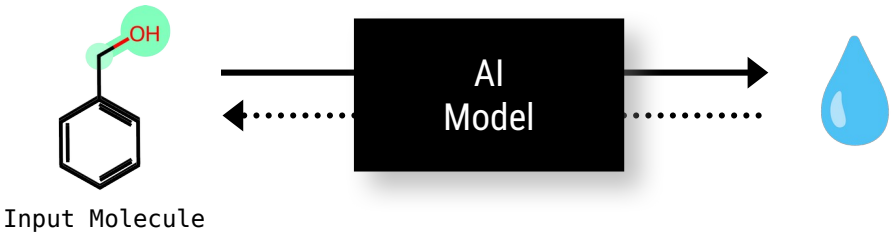
Self-Explaining Approaches

Model

What are attributional explanations?

Attributional Explanations

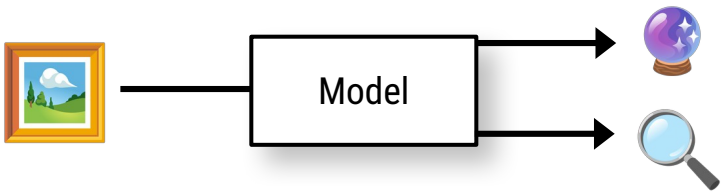




xAI Attributional Explanations highlight parts of the input, the model considers to be especially important!

Post-Hoc Explanations

Self-Explaining Approaches



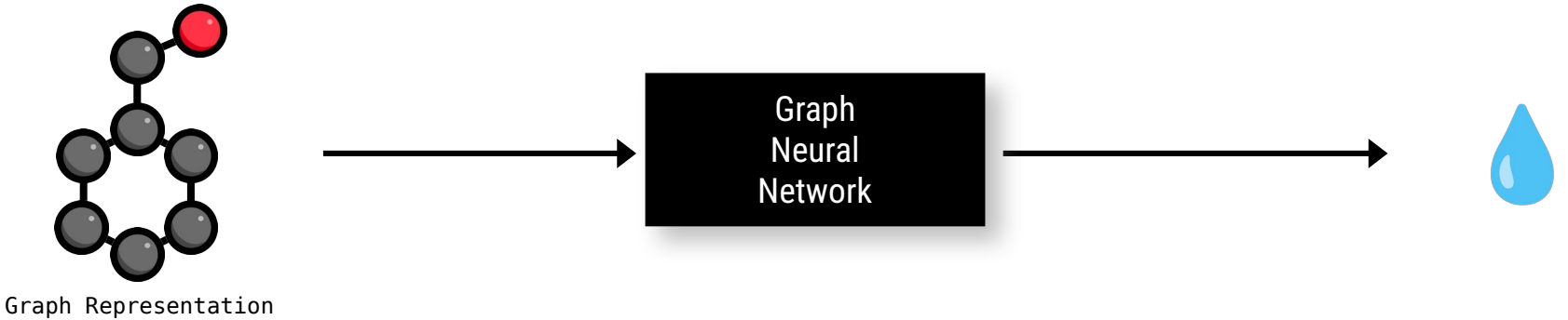
xAI Self-explaining methods provide the prediction and an explanation thereof at the same time

Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22

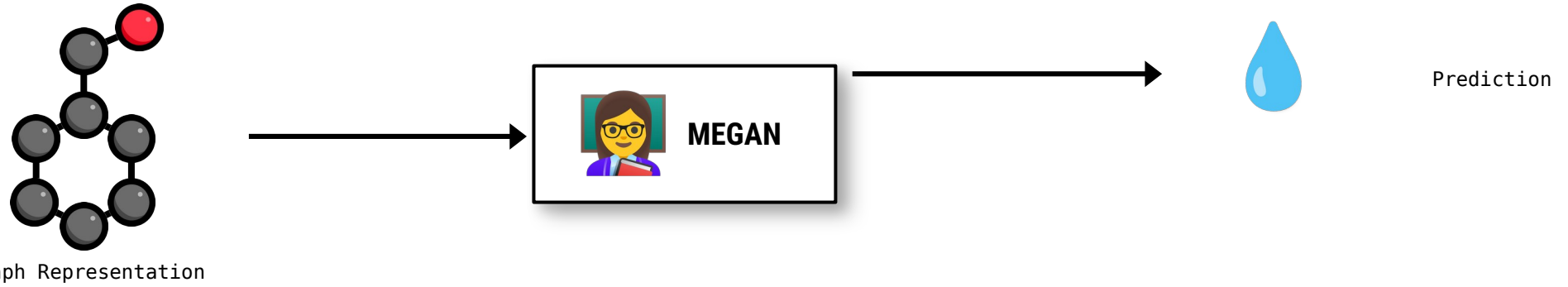
Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22



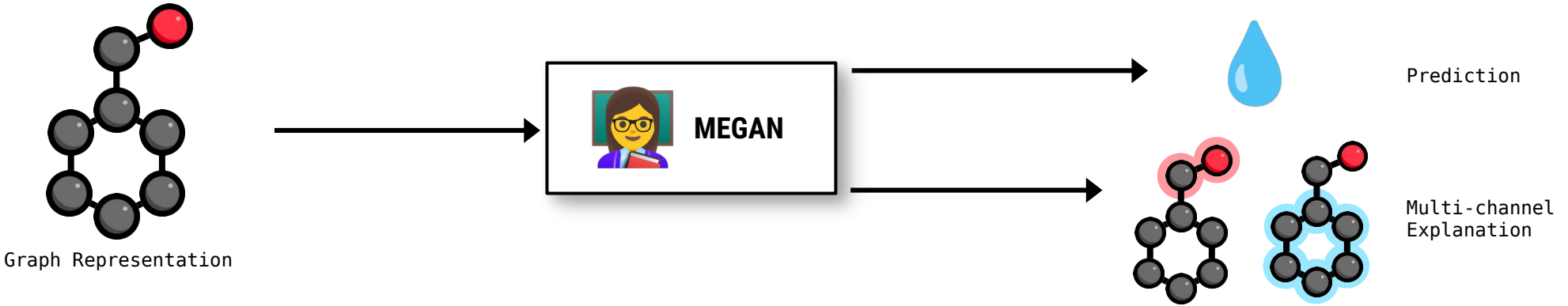
Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22



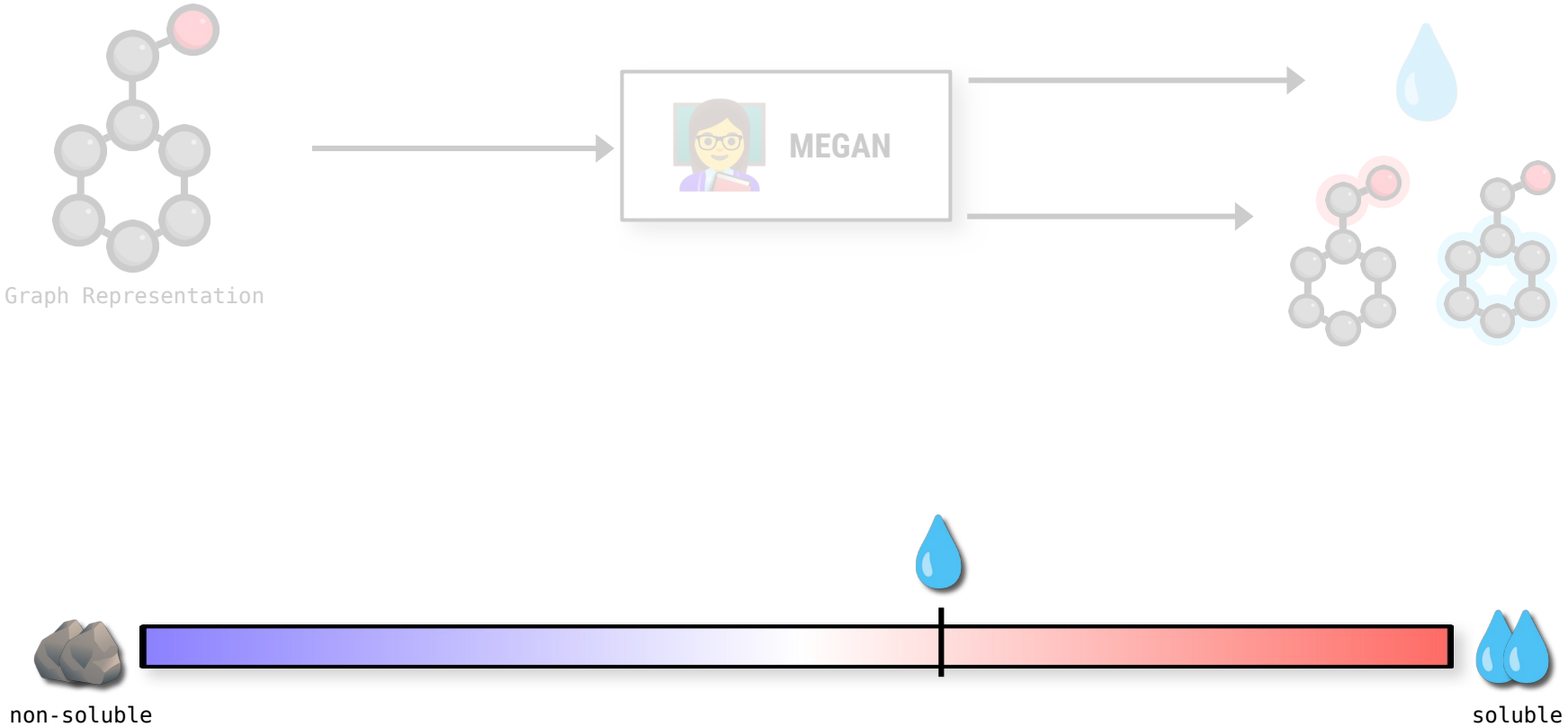
Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22



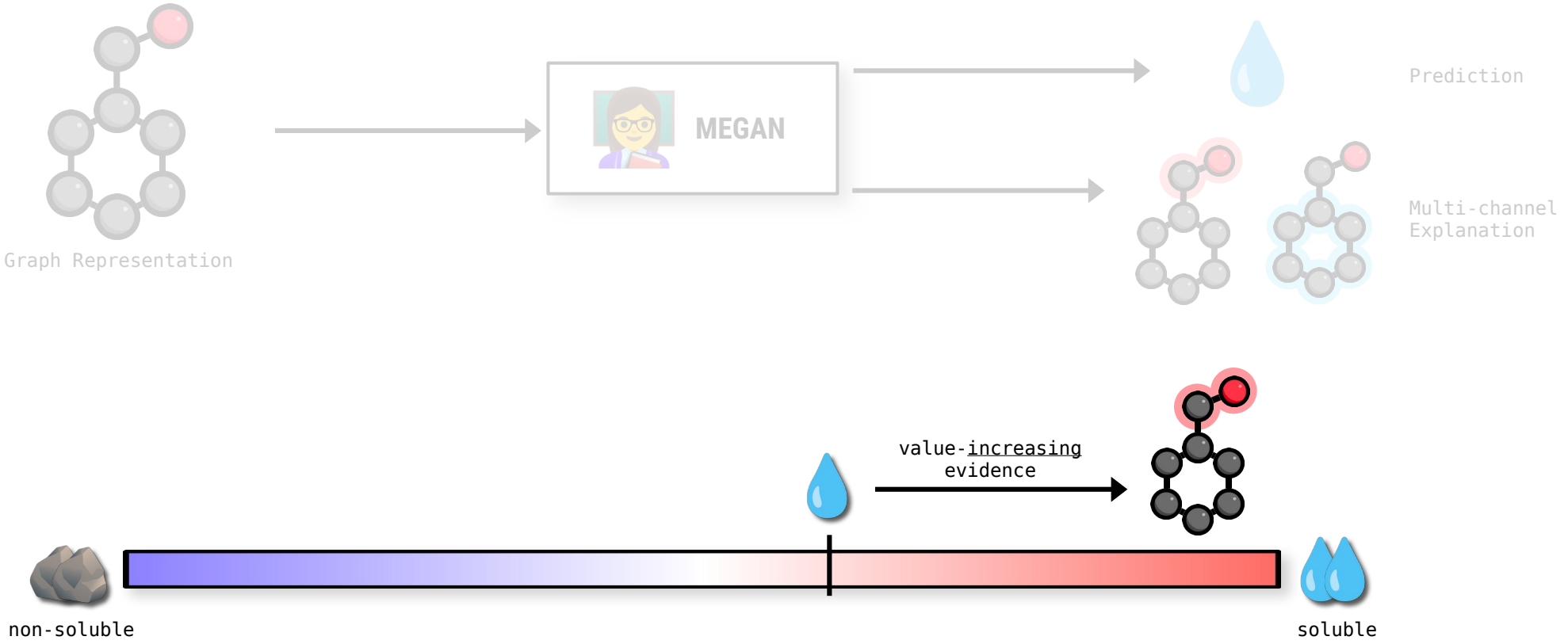
Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22



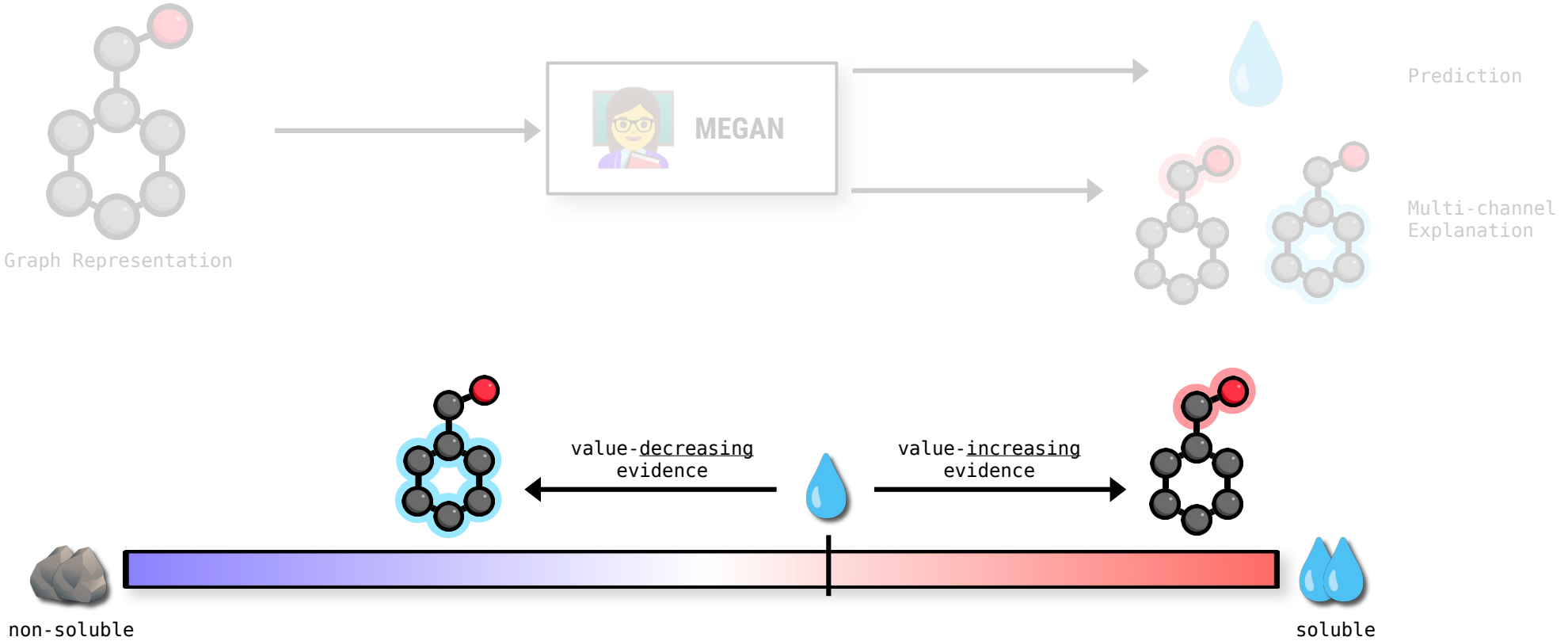
Self-explaining graph neural network

[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22

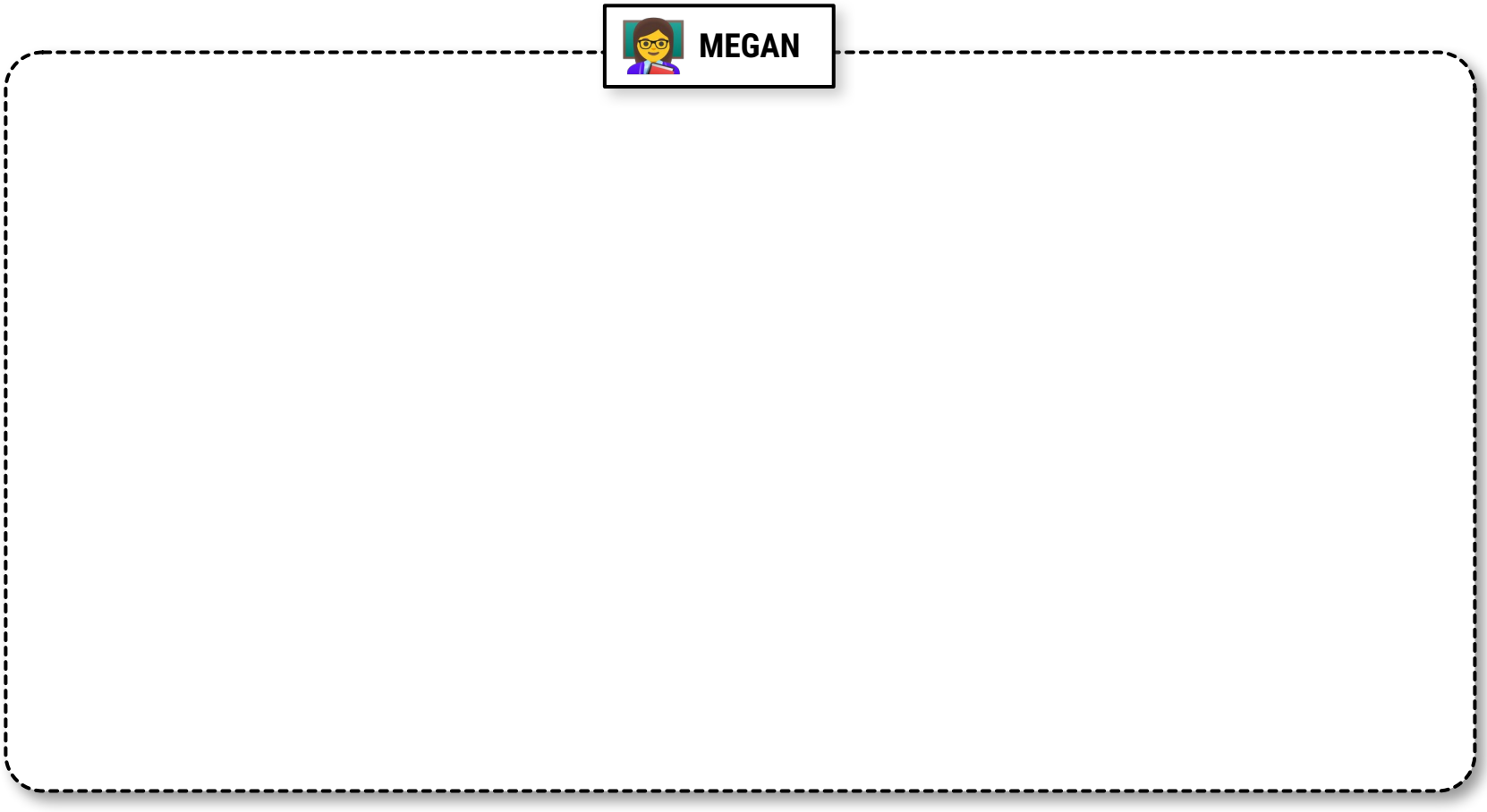


Self-explaining graph neural network

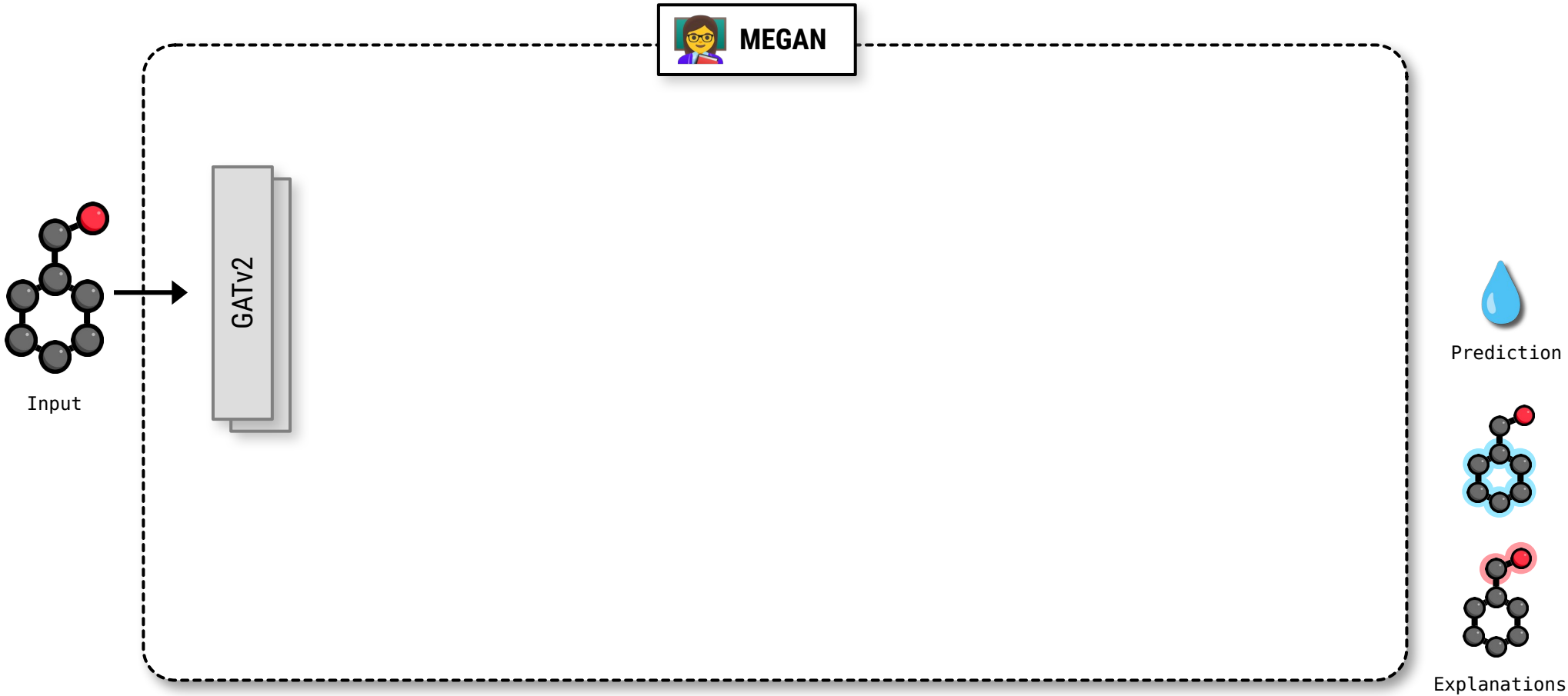
[1] Teufel et al. MEGAN: Multi-Explanation Graph Attention Network. XAI'22



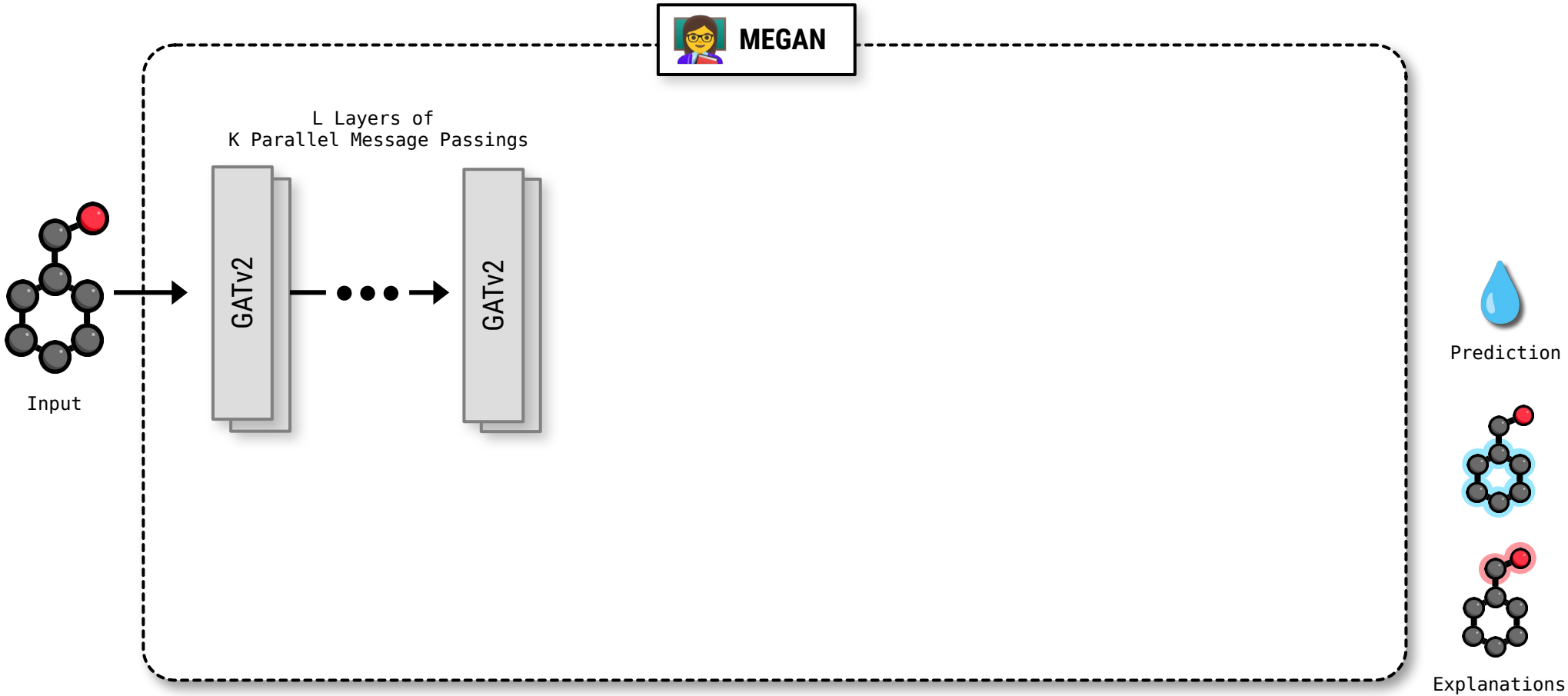
MEGAN: Multi-Explanation Graph Attention Network



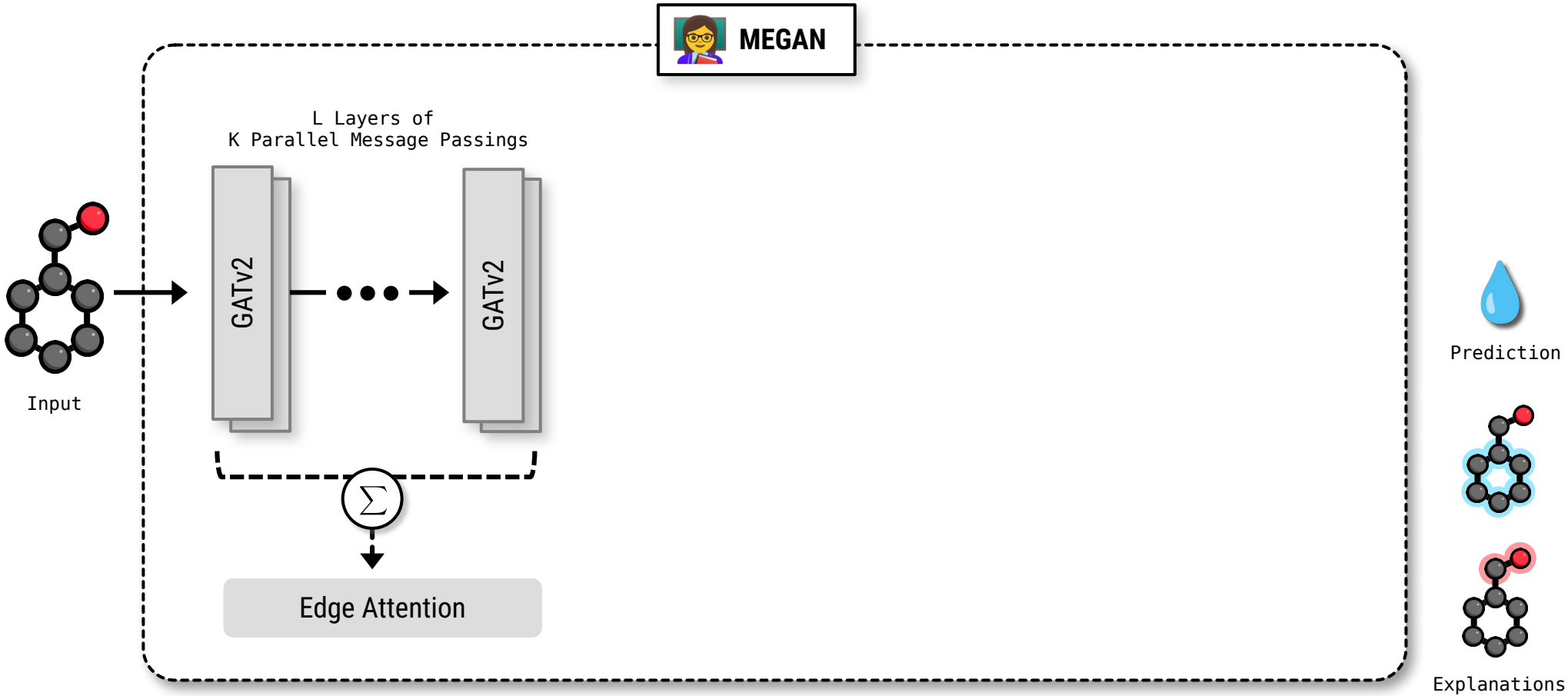
MEGAN: Multi-Explanation Graph Attention Network



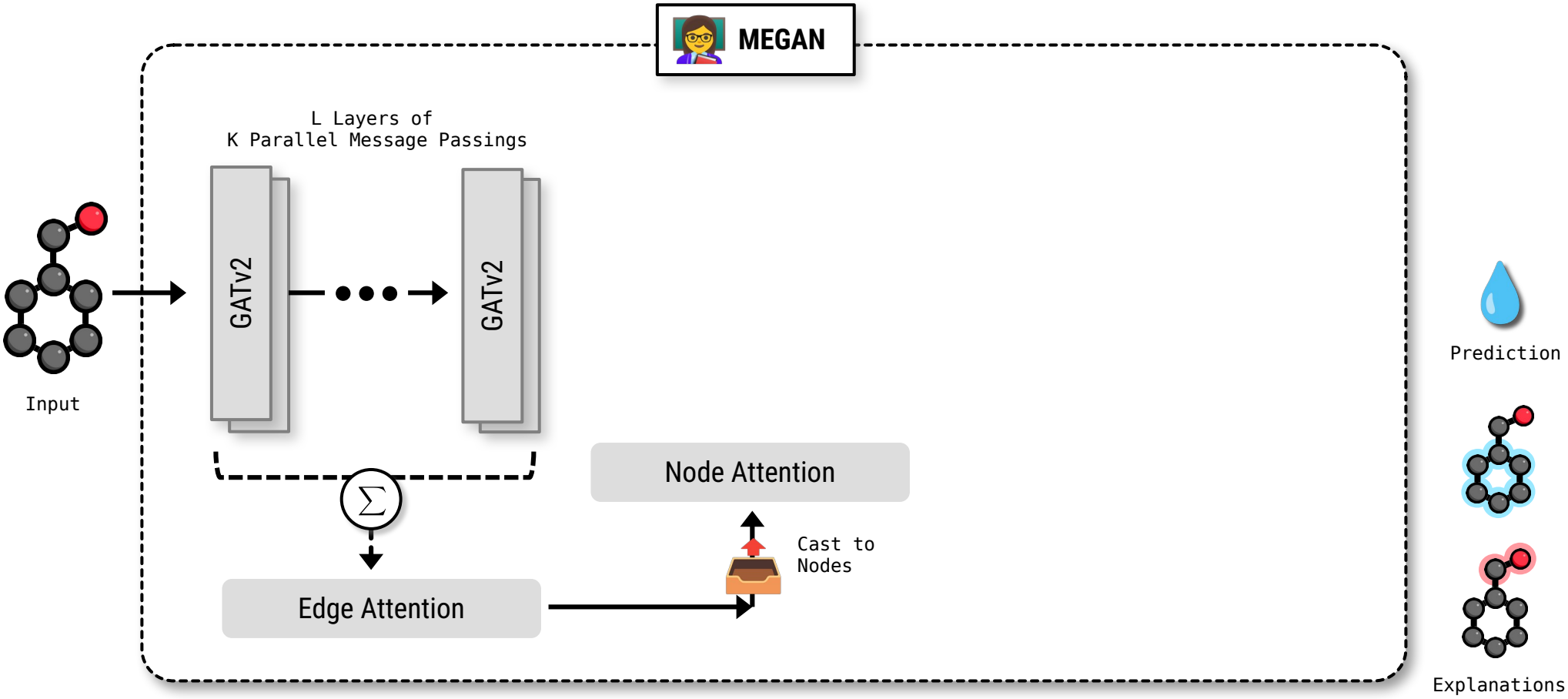
MEGAN: Multi-Explanation Graph Attention Network



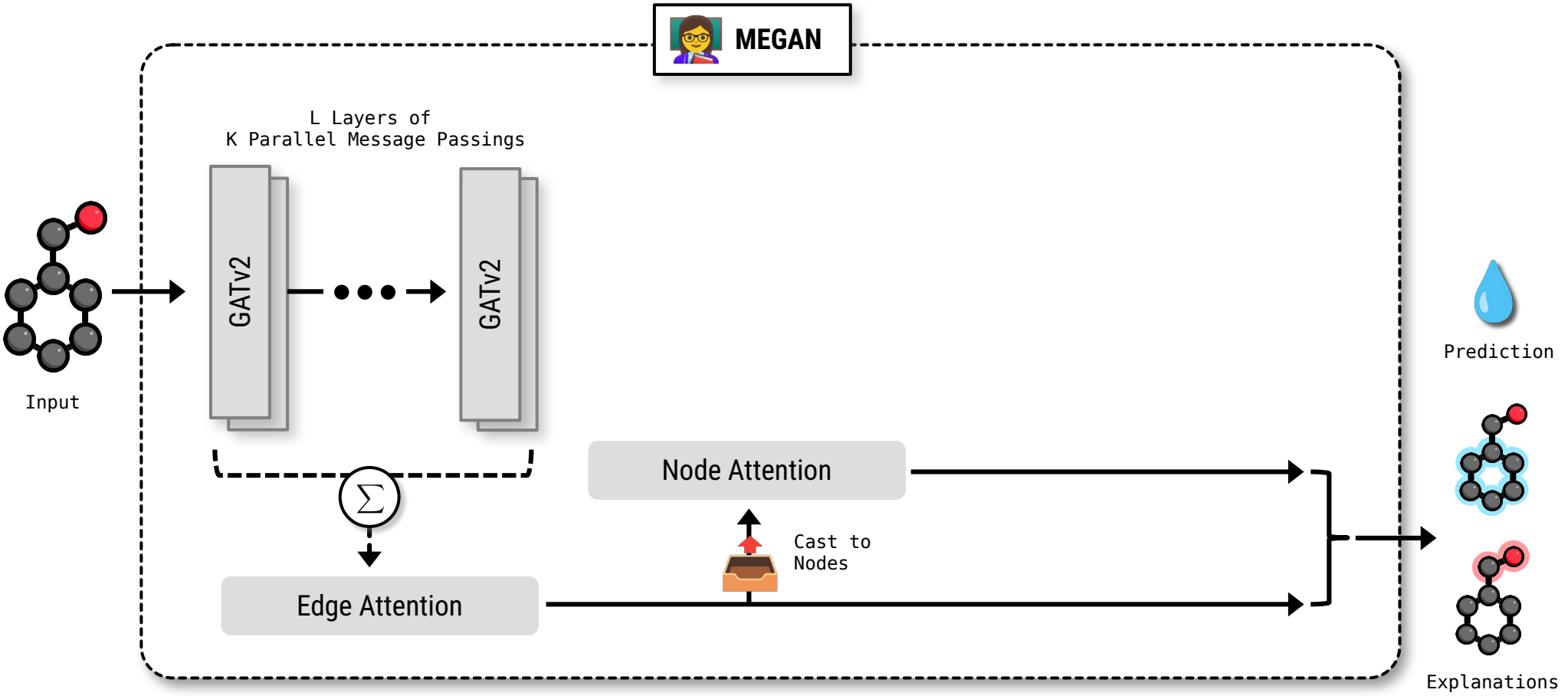
MEGAN: Multi-Explanation Graph Attention Network



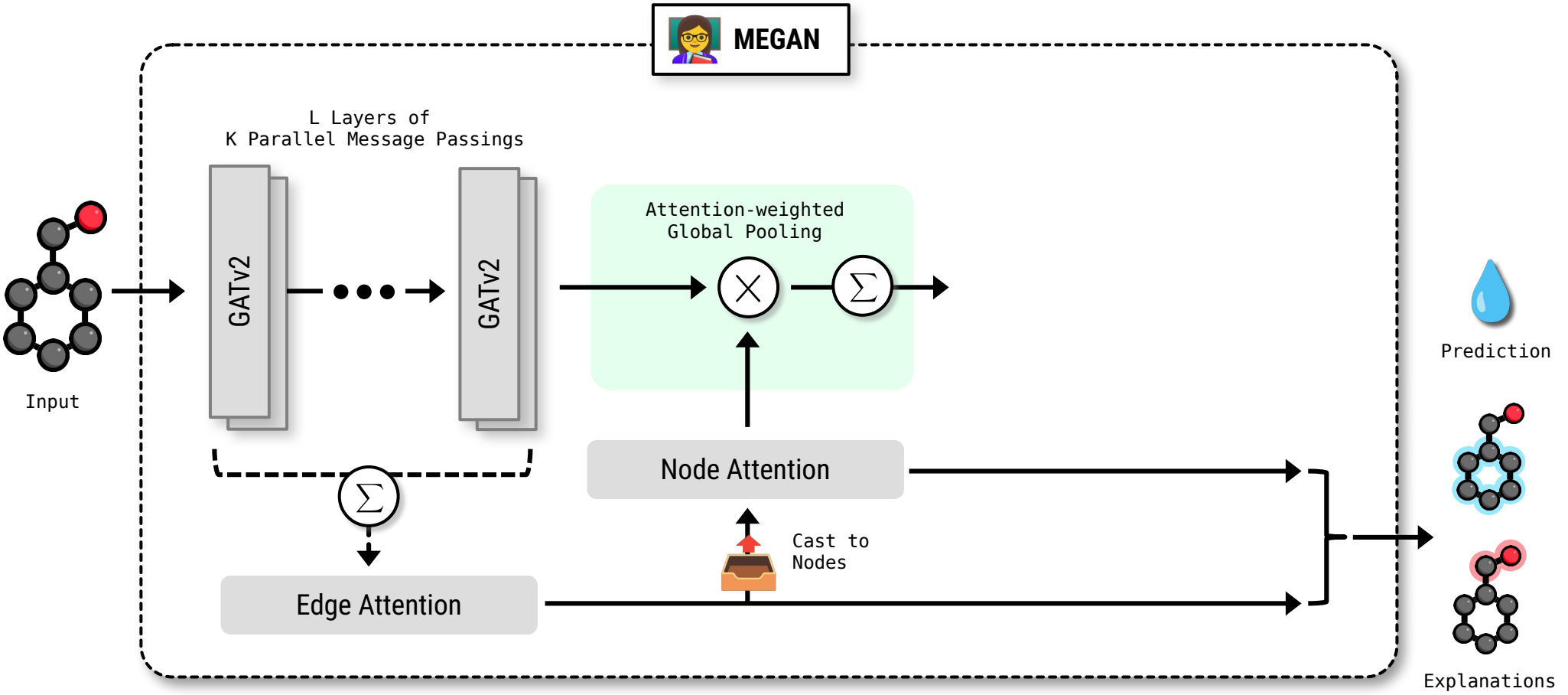
MEGAN: Multi-Explanation Graph Attention Network



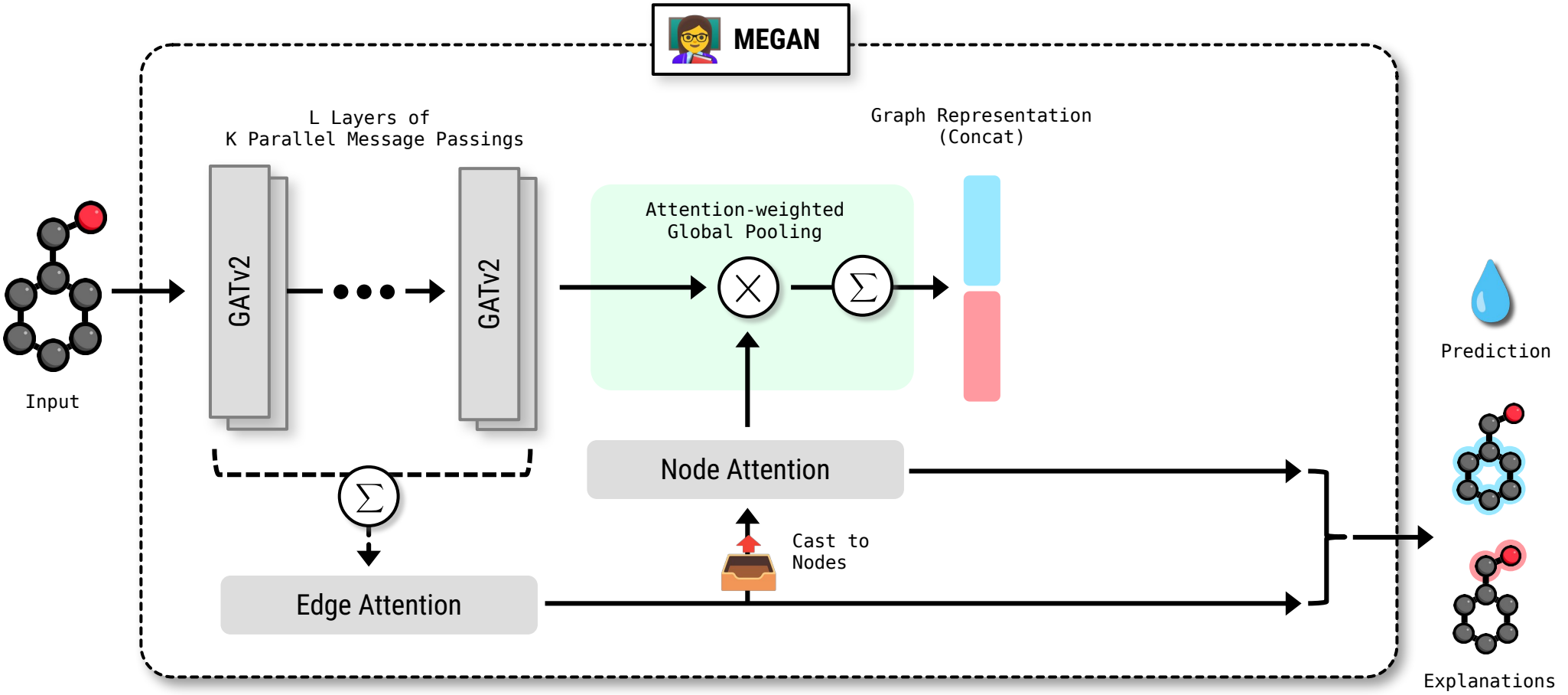
MEGAN: Multi-Explanation Graph Attention Network



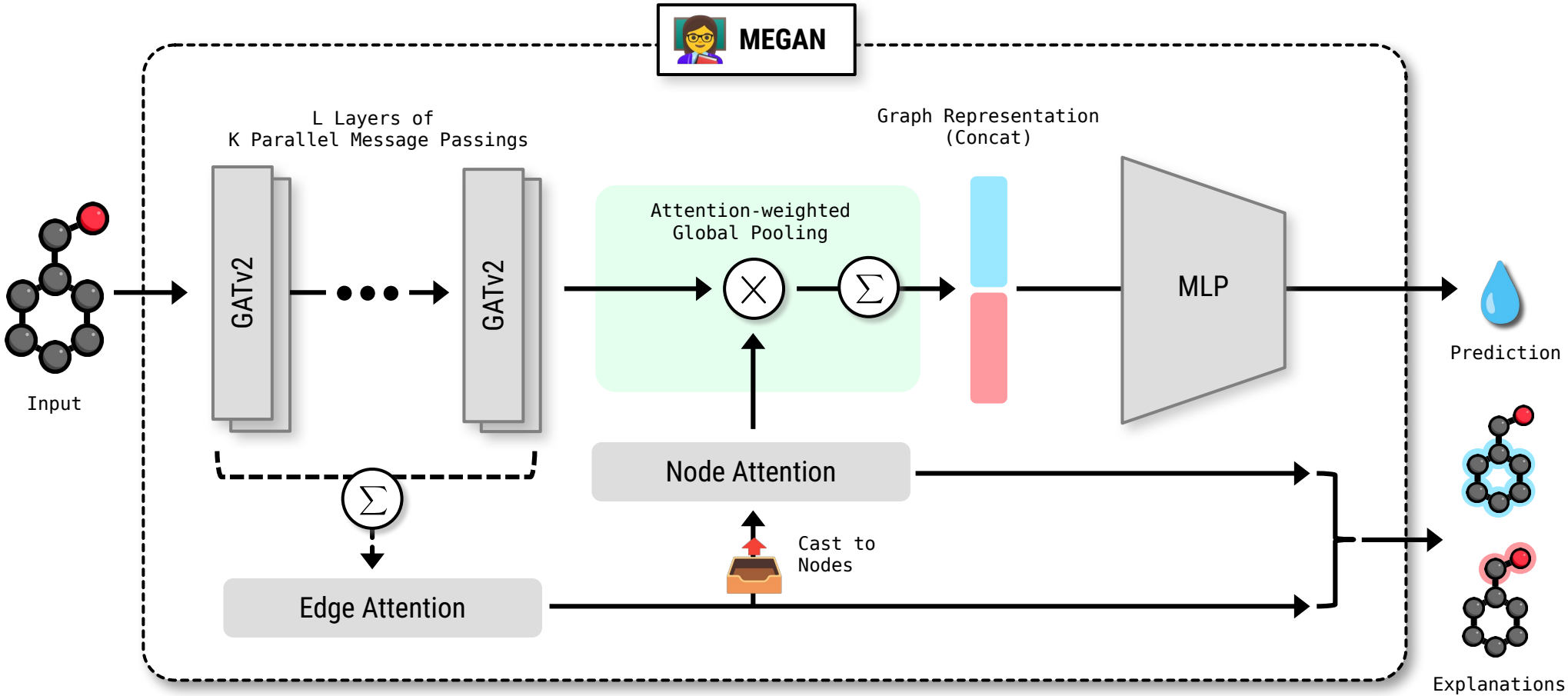
MEGAN: Multi-Explanation Graph Attention Network



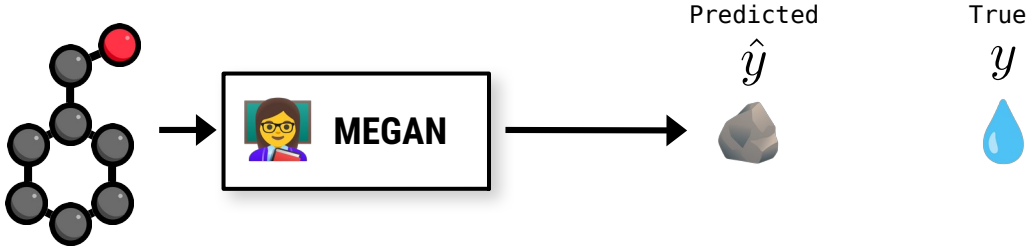
MEGAN: Multi-Explanation Graph Attention Network



MEGAN: Multi-Explanation Graph Attention Network

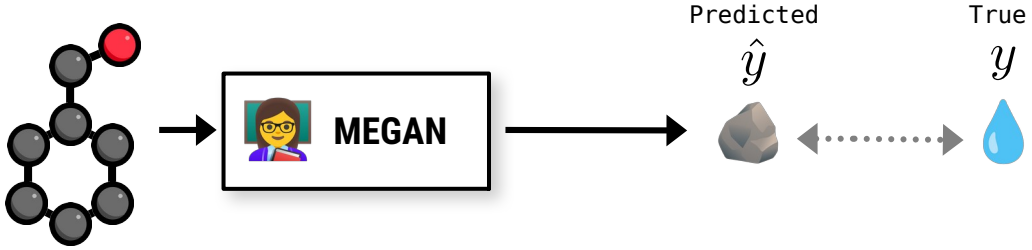


MEGAN explanation co-training objective



Training Objective

MEGAN explanation co-training objective

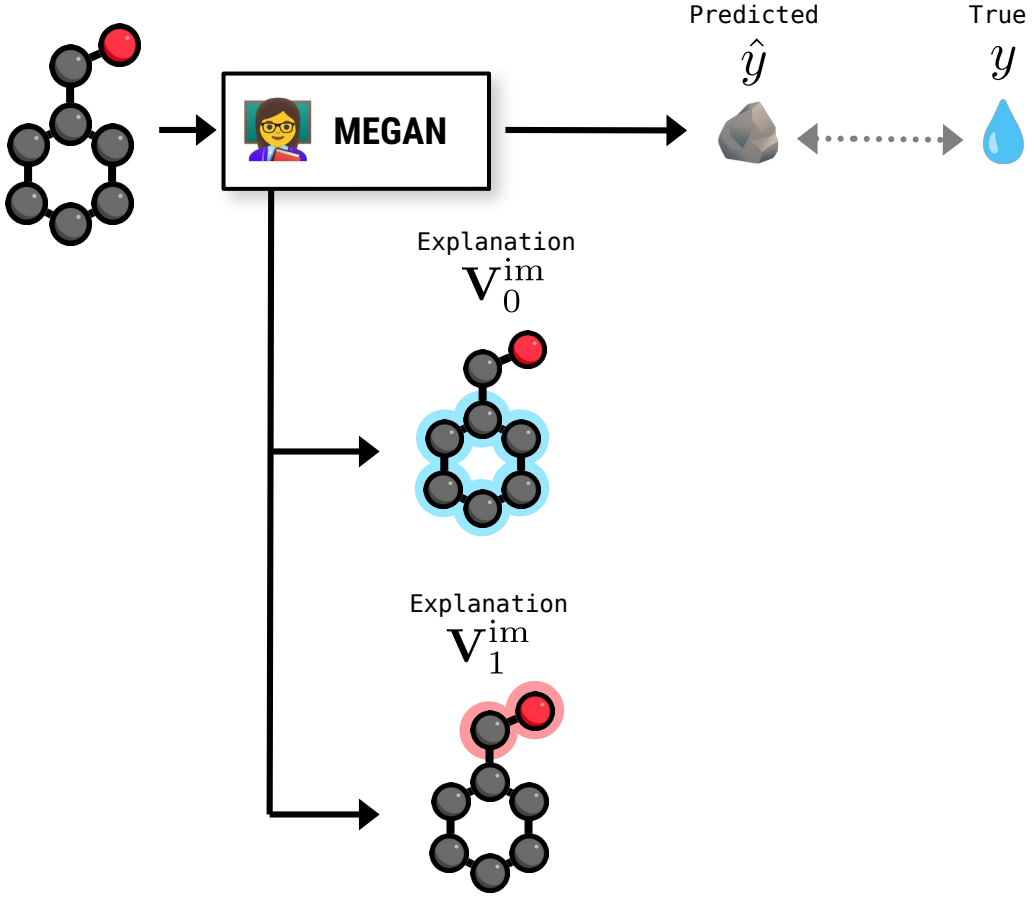


Training Objective

Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

MEGAN explanation co-training objective

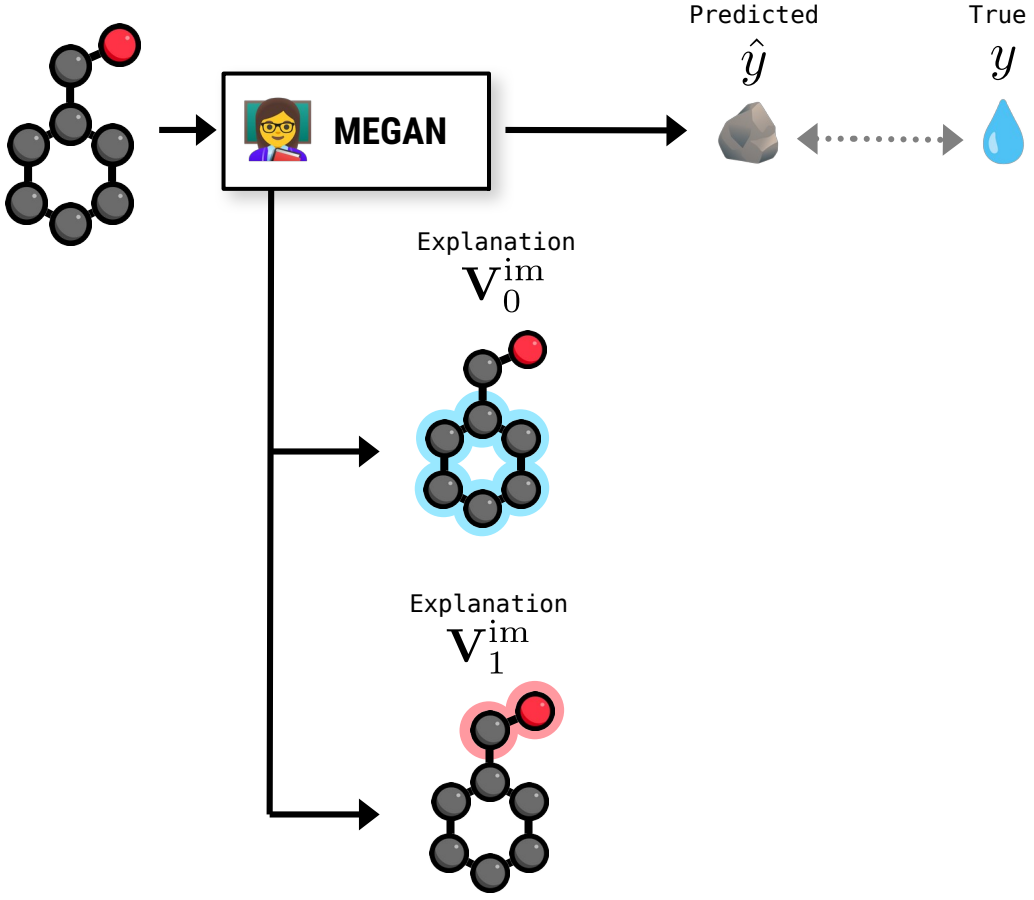


Training Objective

Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

MEGAN explanation co-training objective



Training Objective

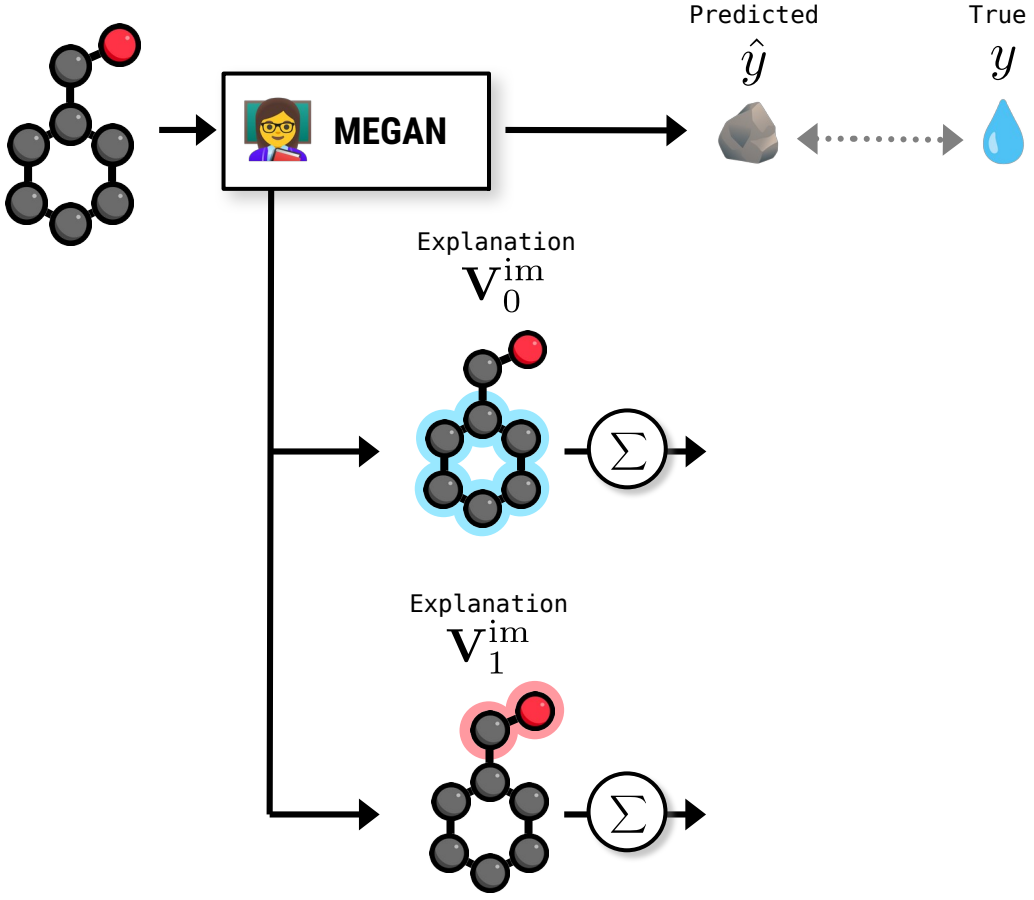
Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

MEGAN explanation co-training objective



Training Objective

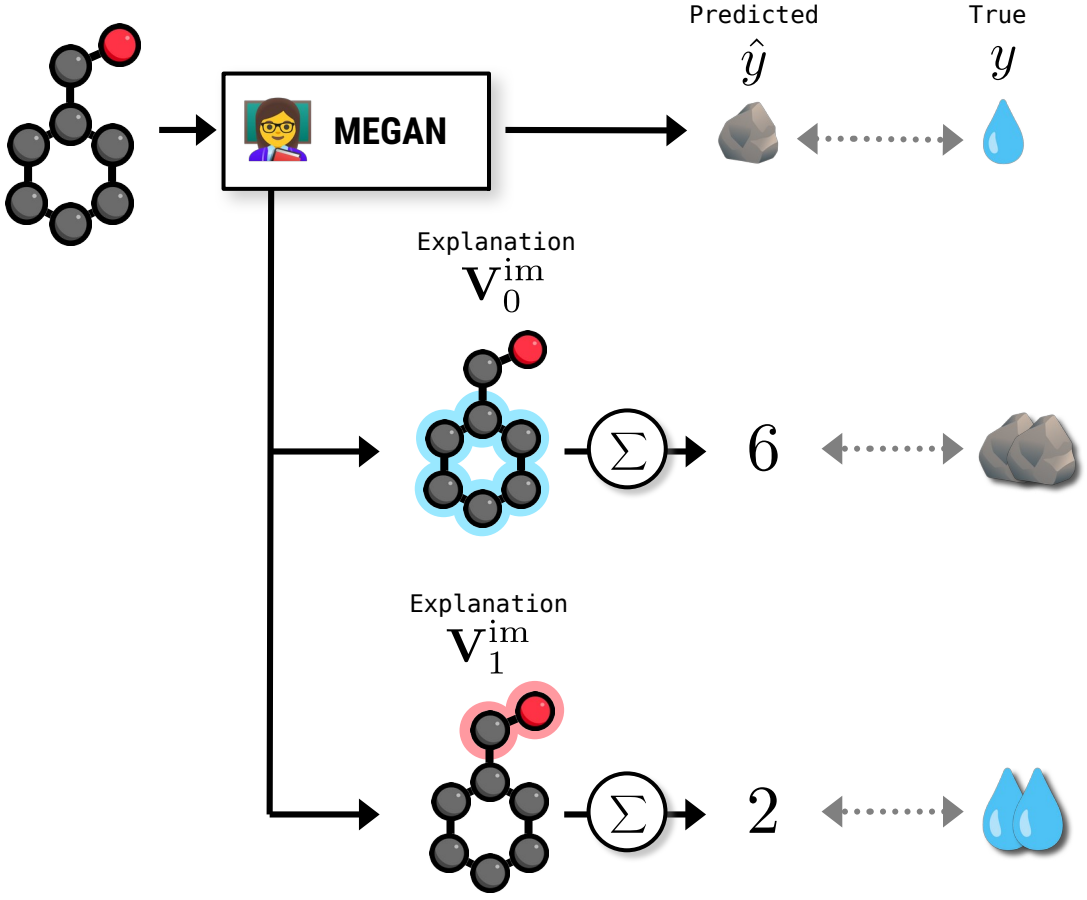
Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

MEGAN explanation co-training objective



Training Objective

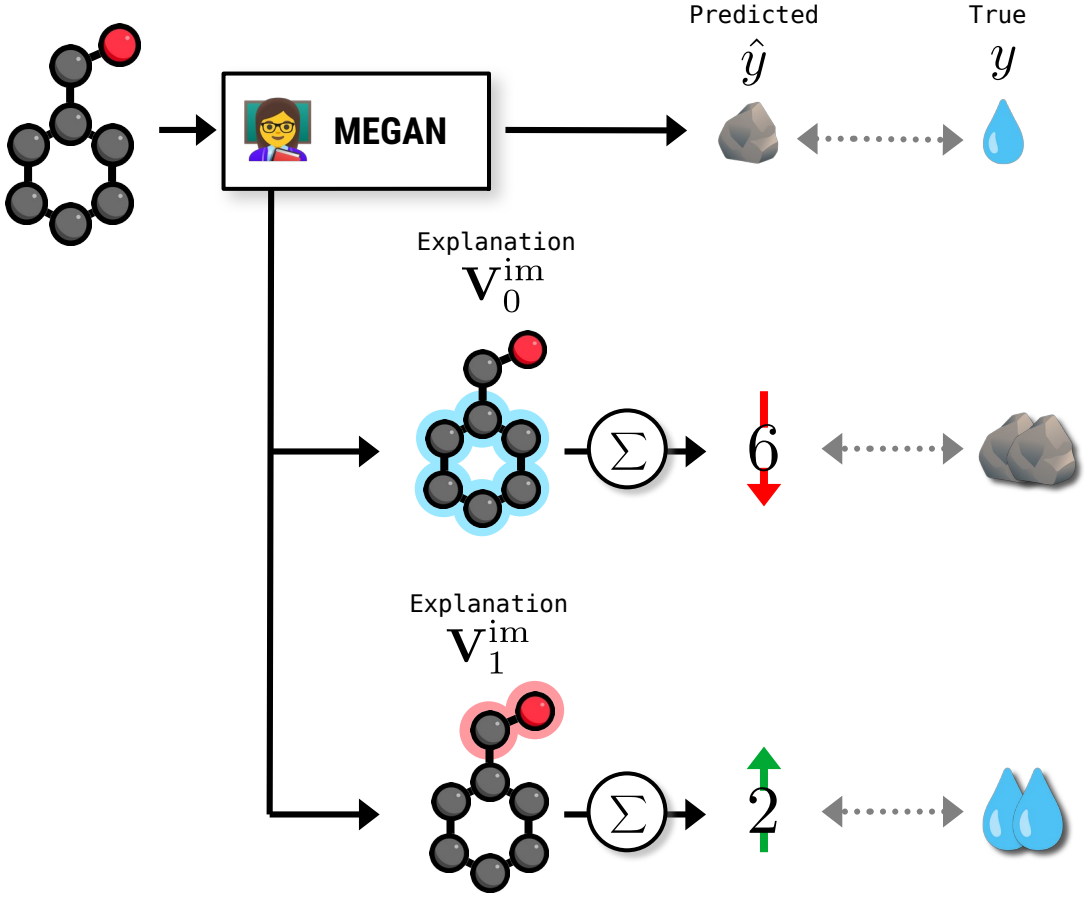
Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

MEGAN explanation co-training objective



Training Objective

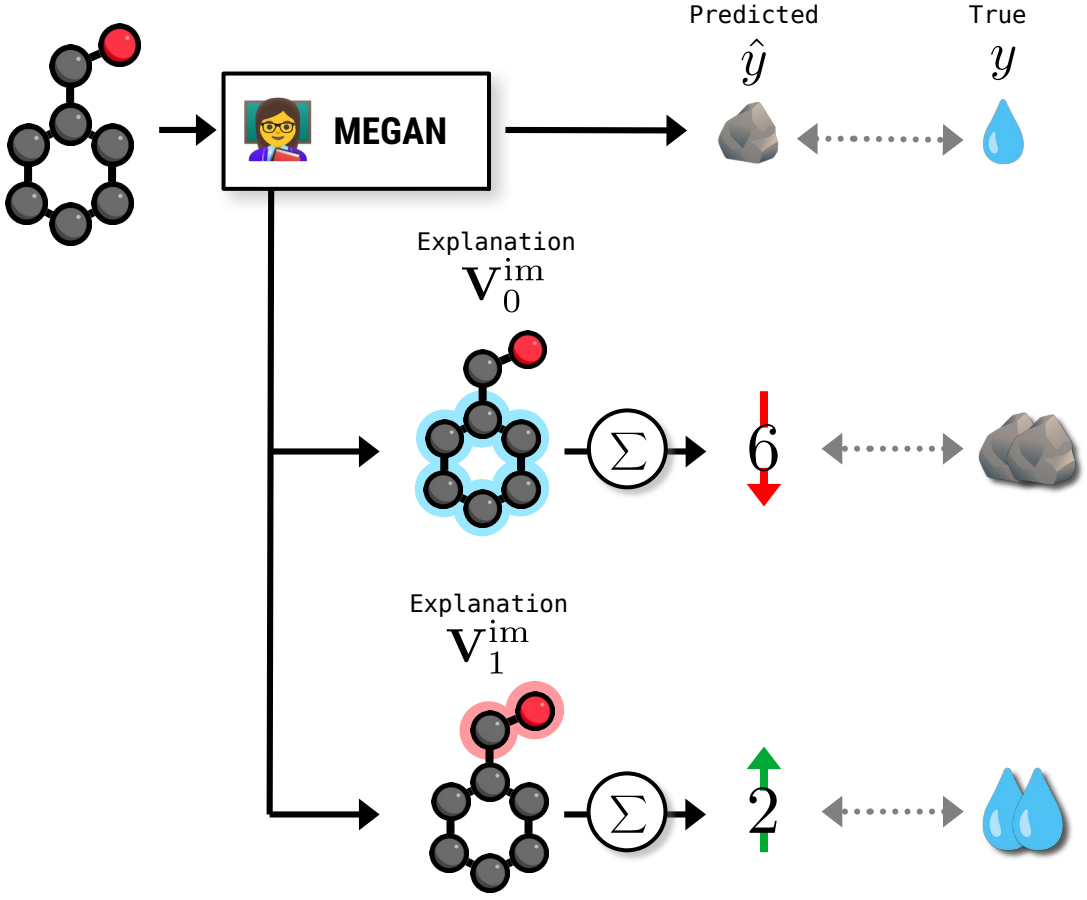
Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

MEGAN explanation co-training objective



Training Objective

Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

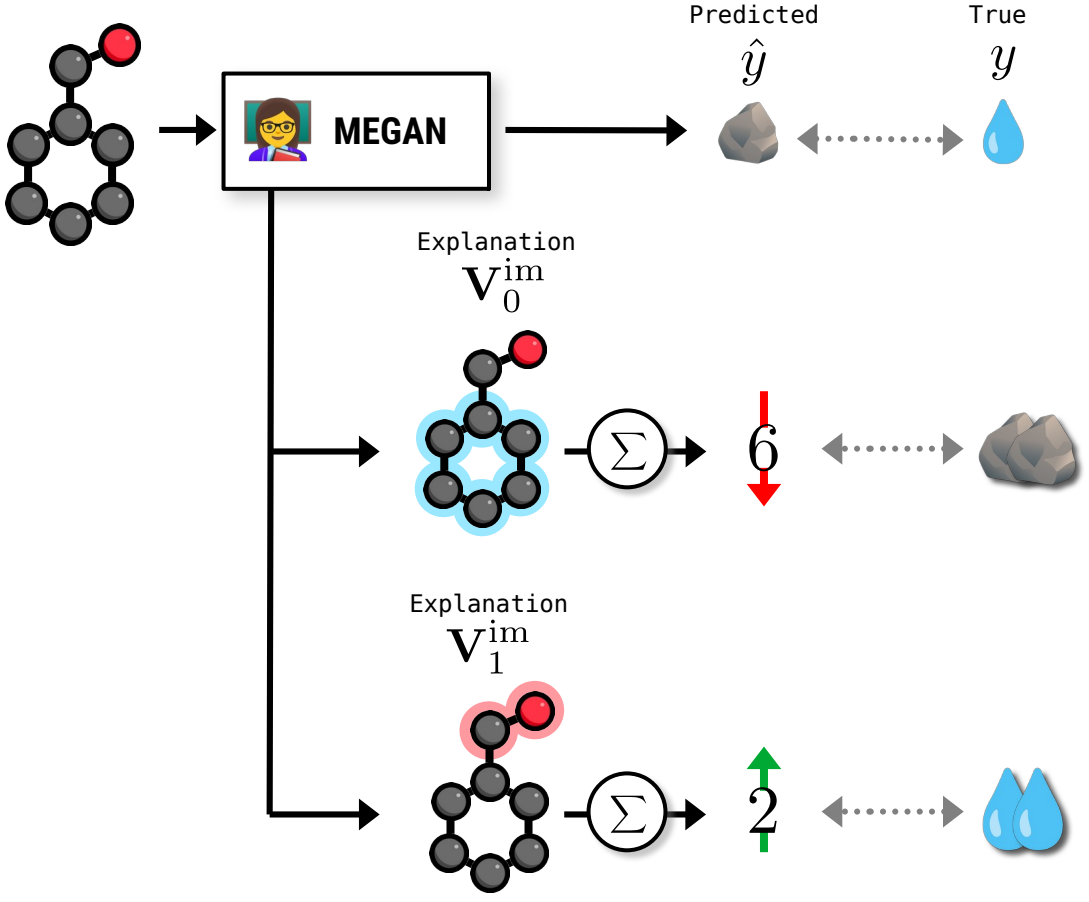
Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

Explanation Co-Training

$$+\lambda_2 \cdot (\text{BCE}(\sigma(|\mathbf{V}_0^{\text{im}}|), y < \delta) + \text{BCE}(\sigma(|\mathbf{V}_1^{\text{im}}|), y > \delta))$$

MEGAN explanation co-training objective



Training Objective

Main Prediction Loss

$$\mathcal{L} = |y - \hat{y}|_2^2$$

Sparsity Regularization

$$+\lambda_1 \cdot |\mathbf{V}^{\text{im}}|_1$$

Explanation Co-Training

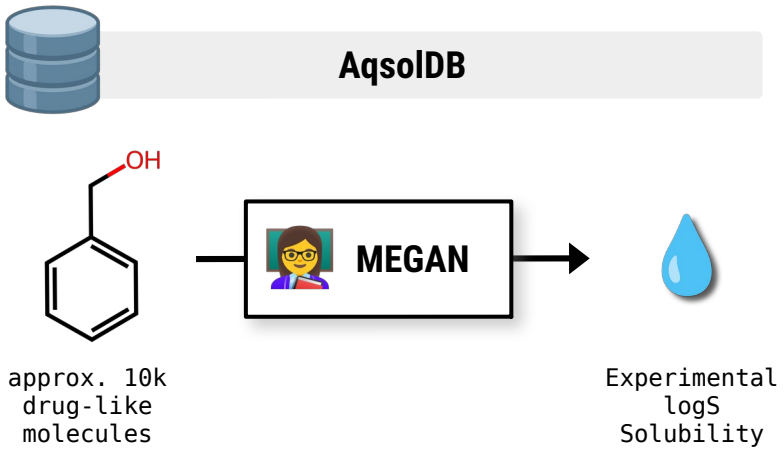
$$+\lambda_2 \cdot (\text{BCE}(\sigma(|\mathbf{V}_0^{\text{im}}|), y < \delta) + \text{BCE}(\sigma(|\mathbf{V}_1^{\text{im}}|), y > \delta))$$

Encourage explanations by themselves to approximately solve the main prediction objective

Example MEGAN explanations

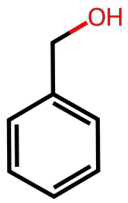
AqsolDB

Example MEGAN explanations



Example MEGAN explanations

AqsolDB

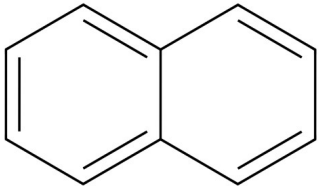


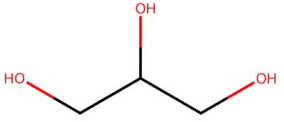
approx. 10k drug-like molecules

MEGAN

Experimental logS Solubility

Decreasing

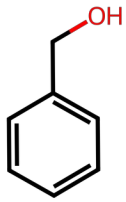




Increasing

Example MEGAN explanations

AqsolDB

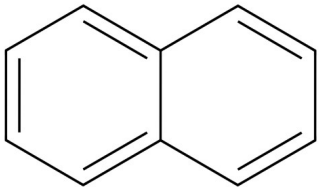


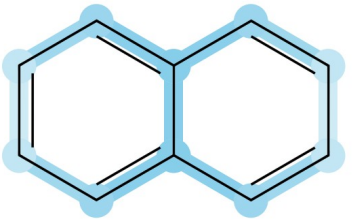
approx. 10k drug-like molecules

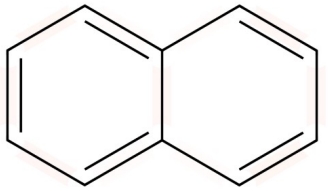
Experimental logS Solubility

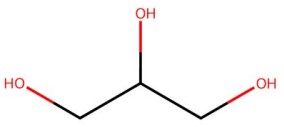
Decreasing

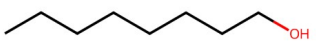
Increasing





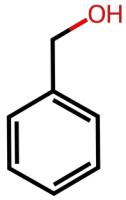






Example MEGAN explanations

AqsolDB

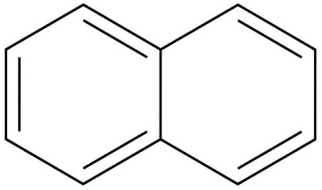


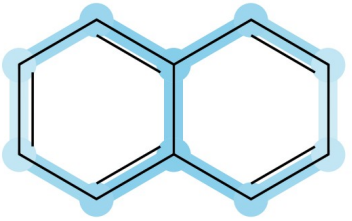
approx. 10k drug-like molecules

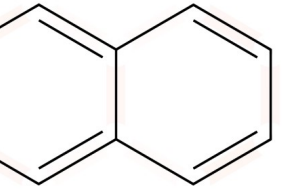
Experimental logS Solubility

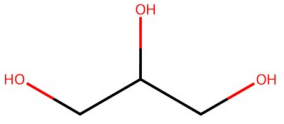
Decreasing

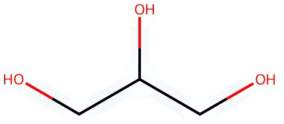
Increasing

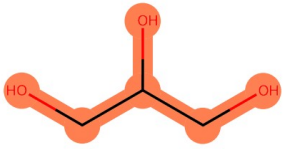






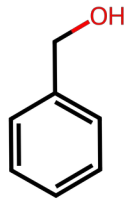






Example MEGAN explanations

AqsolDB

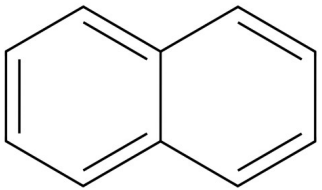


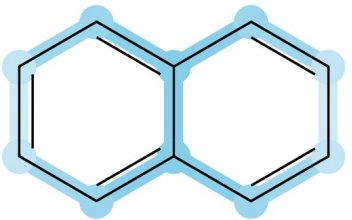
approx. 10k drug-like molecules

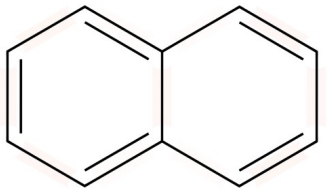
Experimental logS Solubility

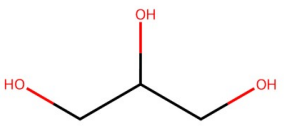
Decreasing

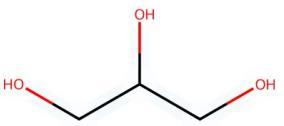
Increasing

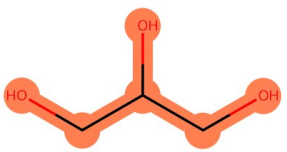


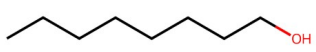








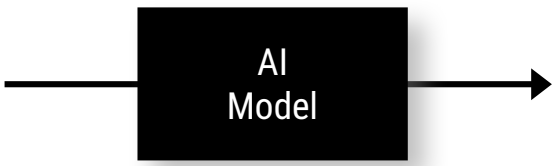




What are counterfactual explanations?

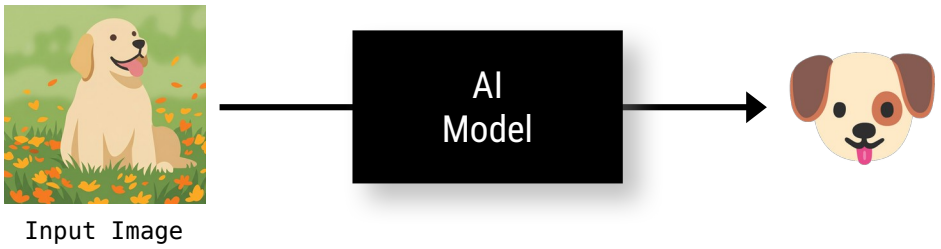
What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

Input Image



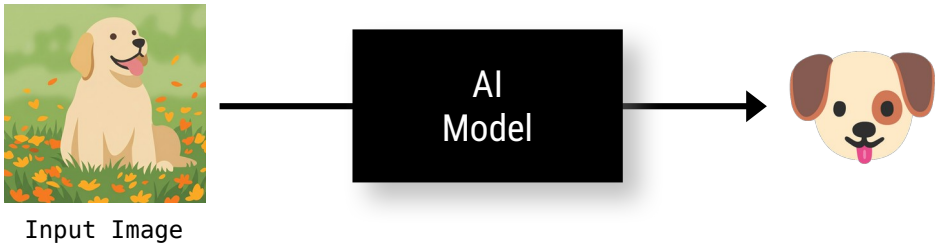
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?



What are counterfactual explanations?

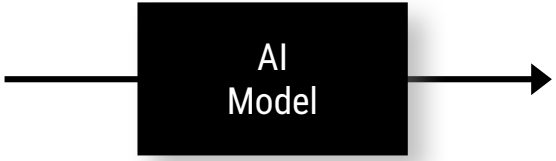
What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?



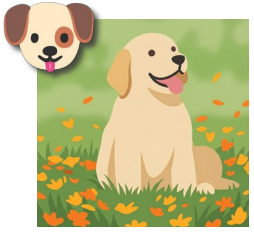
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

Input Image



Original Input



Counterfactual Samples

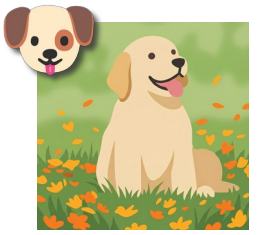
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

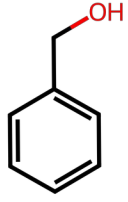
Input Image

AI Model

Original Input

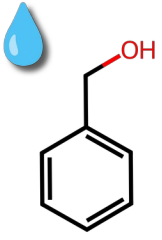


Counterfactual Samples



Input Molecule

AI Model



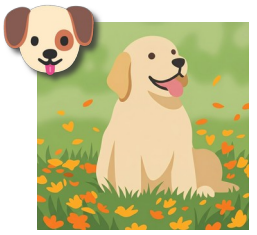
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

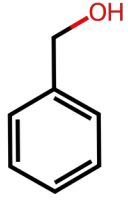
Input Image

AI Model

Original Input

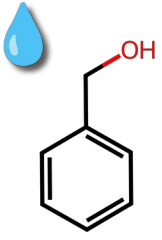


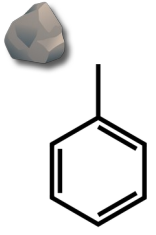
Counterfactual Samples



Input Molecule

AI Model





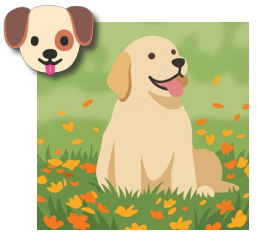
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

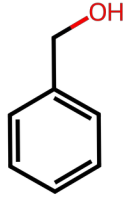
Input Image

AI Model

Original Input

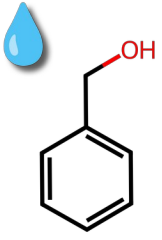


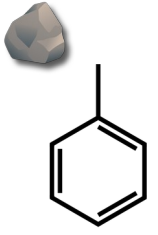
Counterfactual Samples

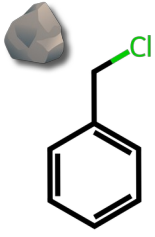


Input Molecule

AI Model







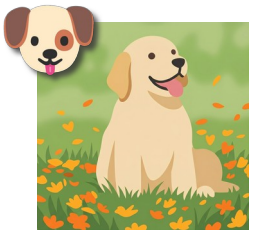
What are counterfactual explanations?

What is the *minimal change* to the *input*, which causes the *maximal deviation* of the *prediction outcome*?

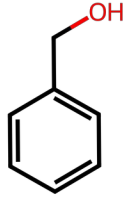
Input Image

AI Model

Original Input

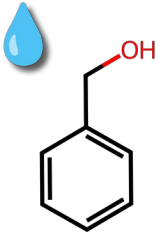


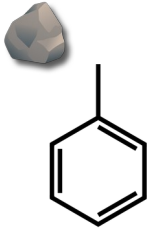
Counterfactual Samples

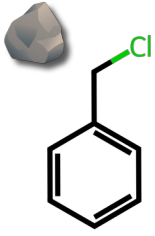


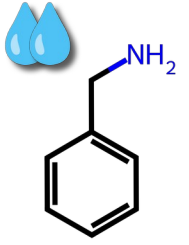
Input Molecule

AI Model

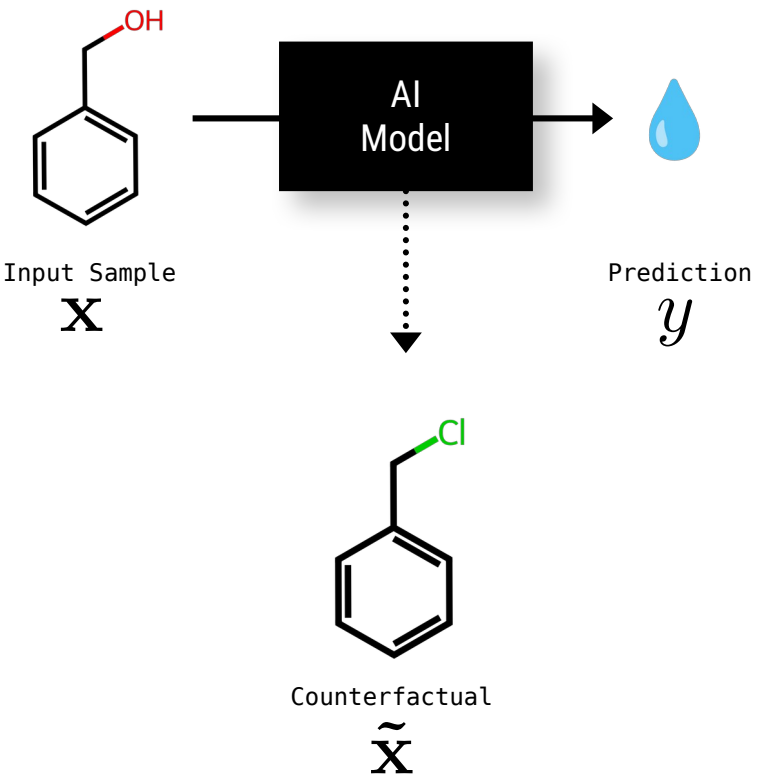




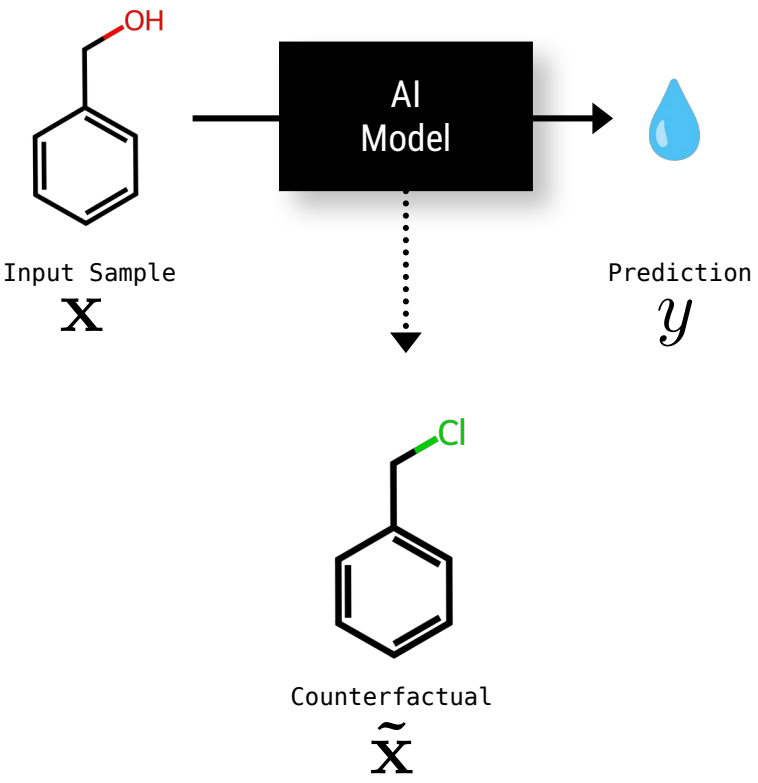




The counterfactual optimization problem



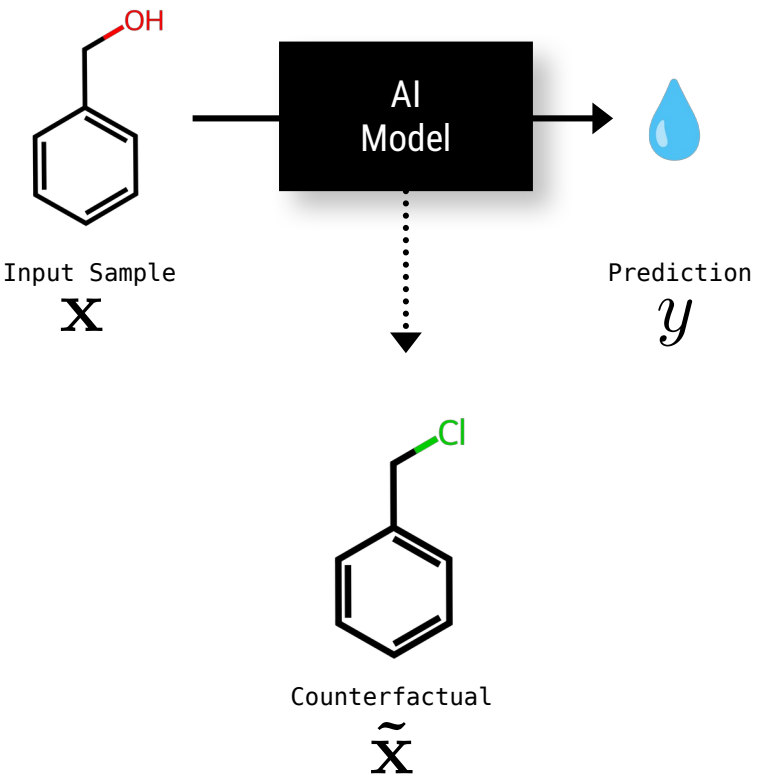
The counterfactual optimization problem



Minimize structural distance to original sample

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}'} \left(d(\mathbf{x}, \mathbf{x}') - \lambda \cdot |y - y'| \right)$$

The counterfactual optimization problem

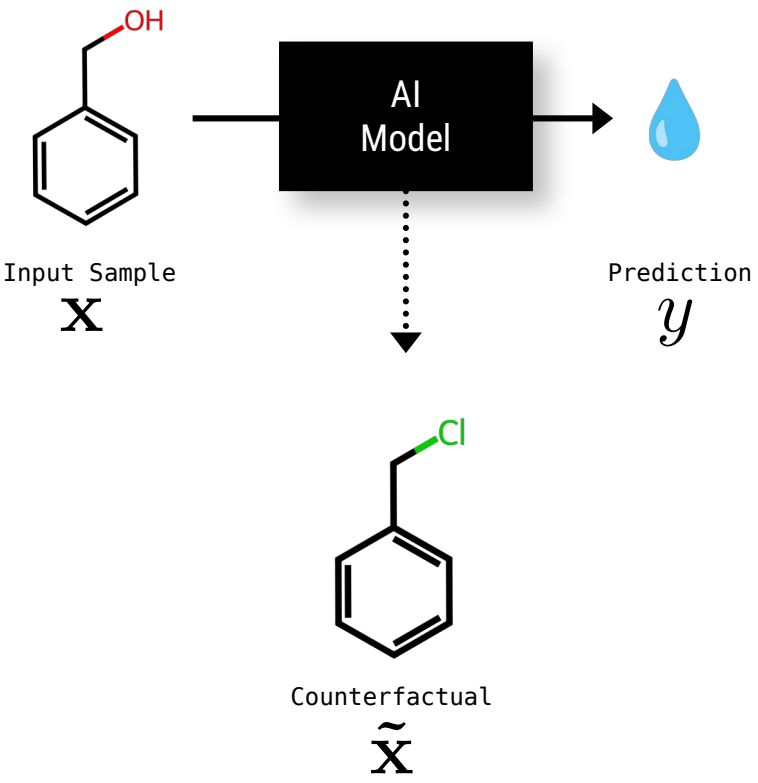


Minimize structural distance to original sample

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}'} \left(d(\mathbf{x}, \mathbf{x}') - \lambda \cdot |y - y'| \right)$$

Maximize the prediction deviation

The counterfactual optimization problem



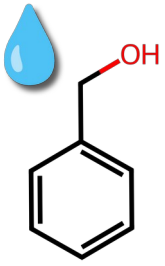
Minimize structural distance to original sample

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}'} \left(d(\mathbf{x}, \mathbf{x}') - \lambda \cdot |y - y'| \right)$$

Maximize the prediction deviation

How to find such counterfactals?

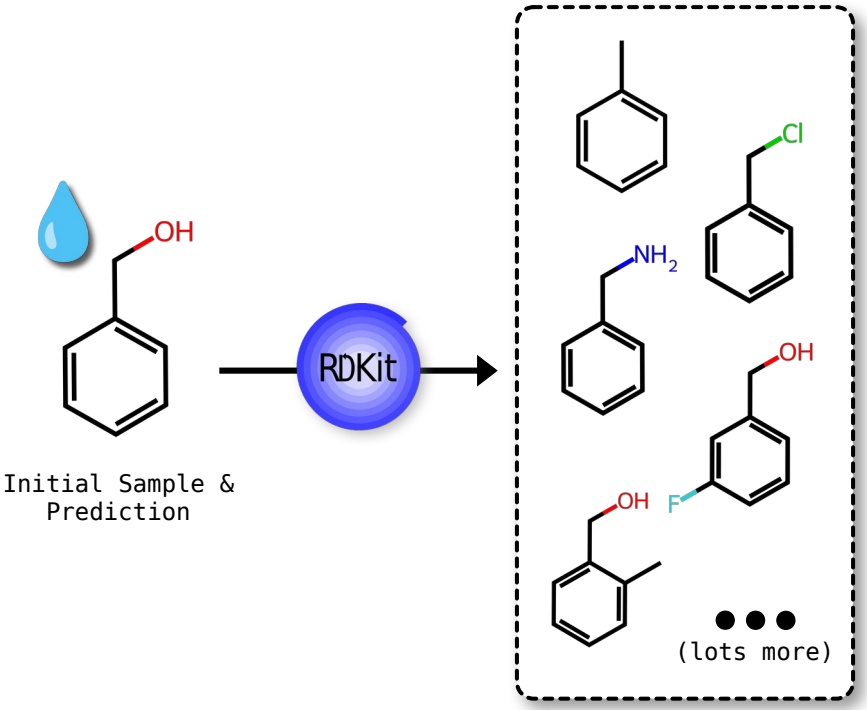
Creating counterfactuals through total enumeration



Initial Sample &
Prediction

Creating counterfactuals through total enumeration

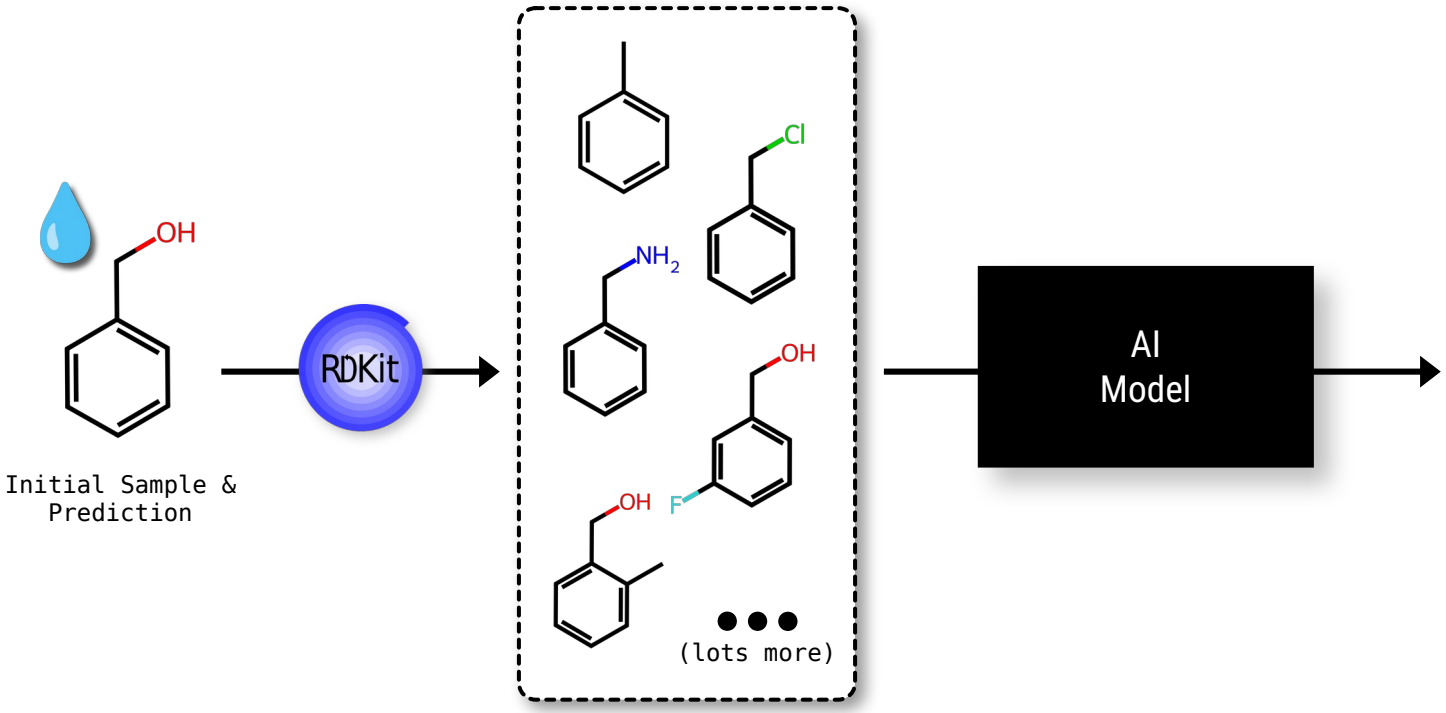
I Obtain all chemically feasible single-edit modifications



Creating counterfactuals through total enumeration

I Obtain all chemically feasible single-edit modifications

II Query the model for predictions

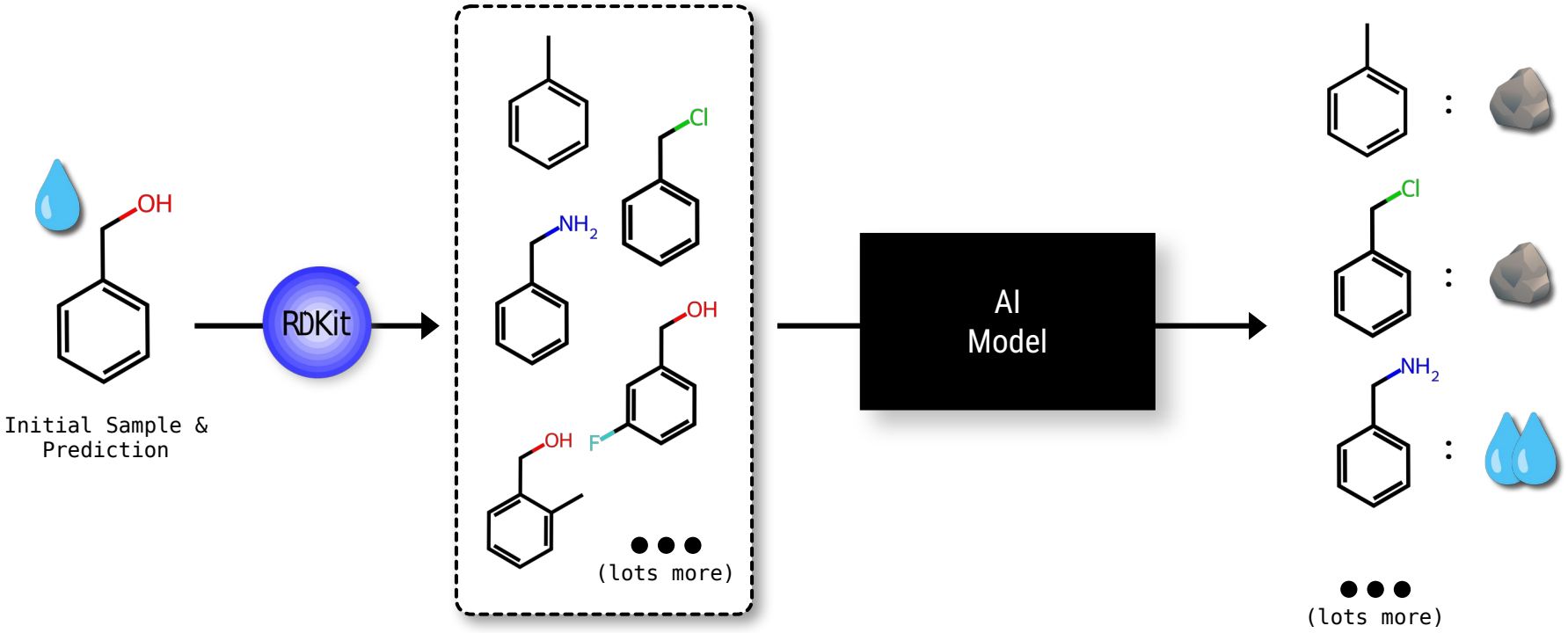


Creating counterfactuals through total enumeration

I Obtain all chemically feasible single-edit modifications

II Query the model for predictions

III Rank by deviation and select top-k as counterfactuals

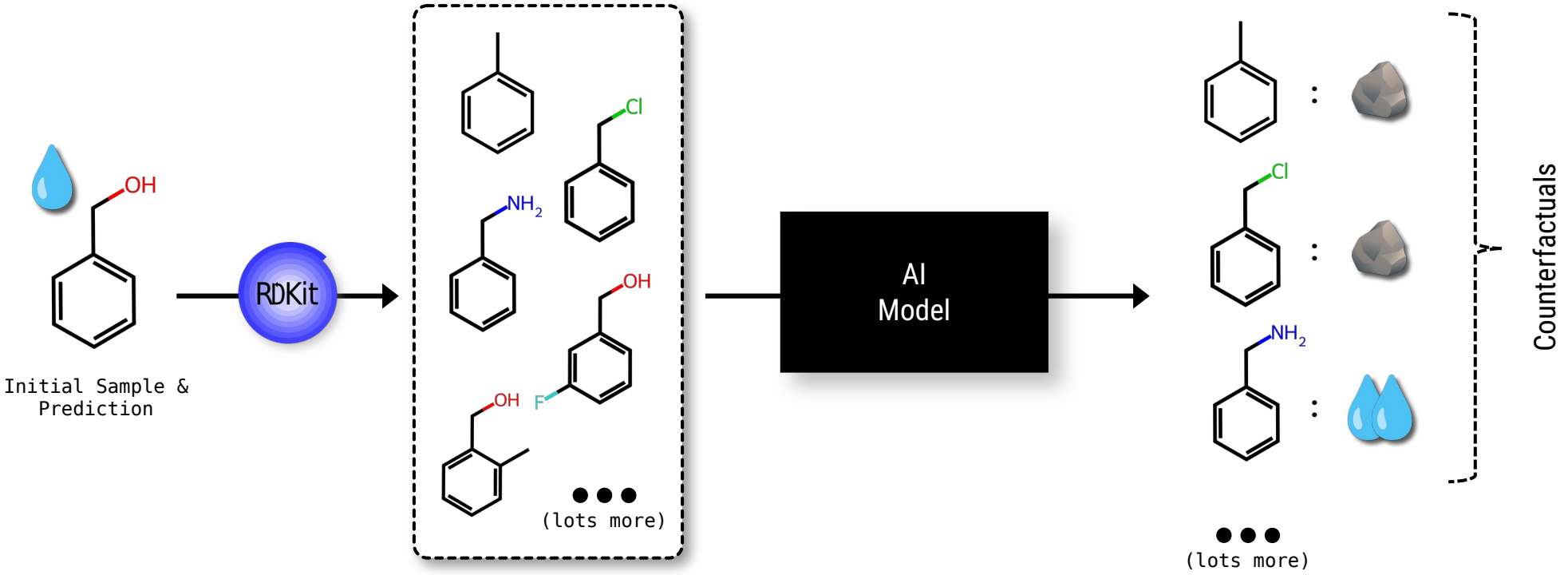


Creating counterfactuals through total enumeration

I Obtain all chemically feasible single-edit modifications

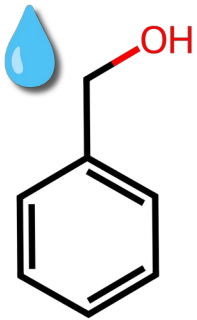
II Query the model for predictions

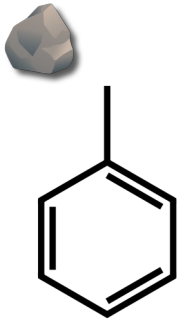
III Rank by deviation and select top-k as counterfactuals

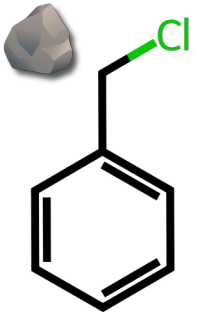


Improving counterfactuals with uncertainty quantification

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



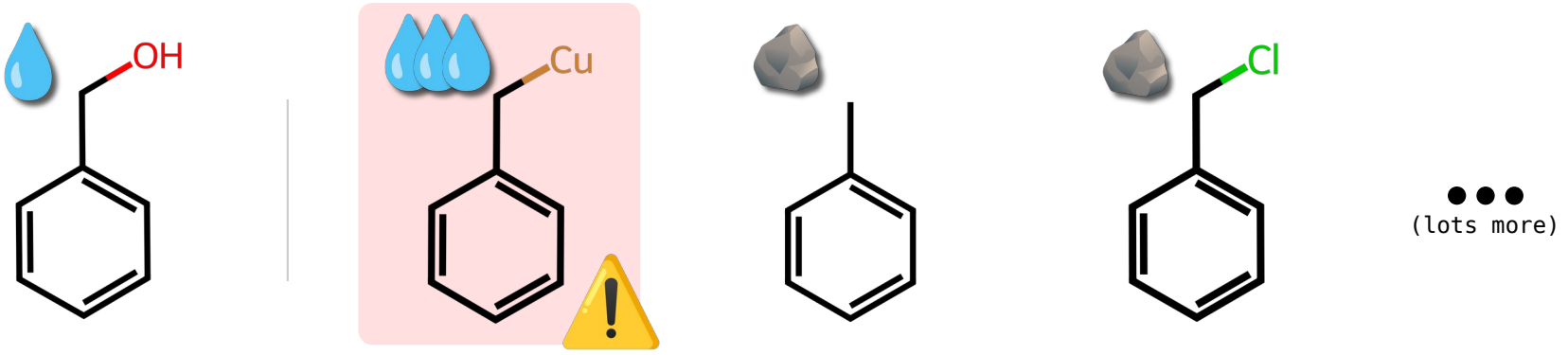




(lots more)

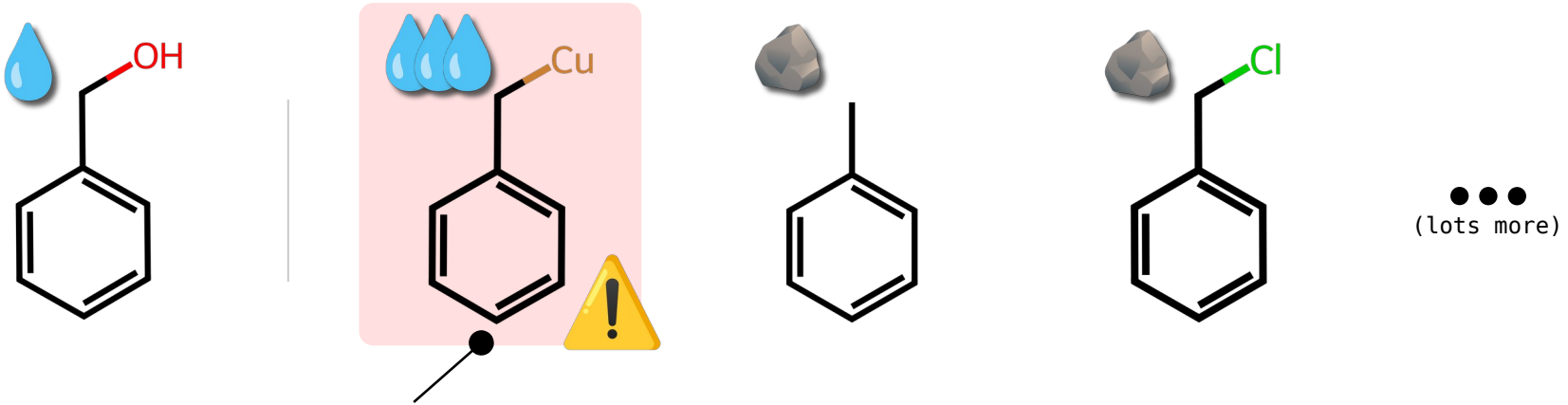
Improving counterfactuals with uncertainty quantification

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



Improving counterfactuals with uncertainty quantification

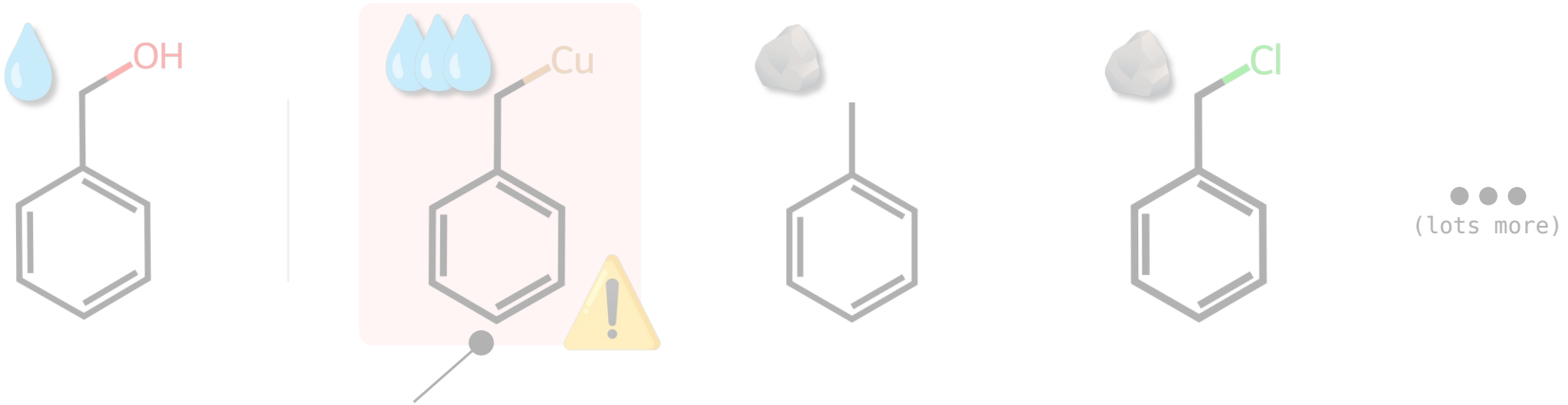
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



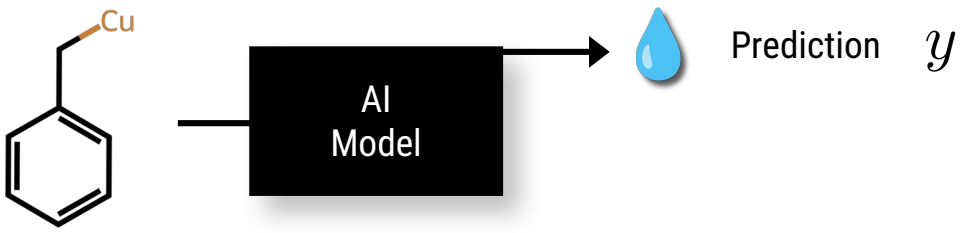
Particularly when generating counterfactual candidates, it is easy to end up with out-of-distribution samples!

Improving counterfactuals with uncertainty quantification

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

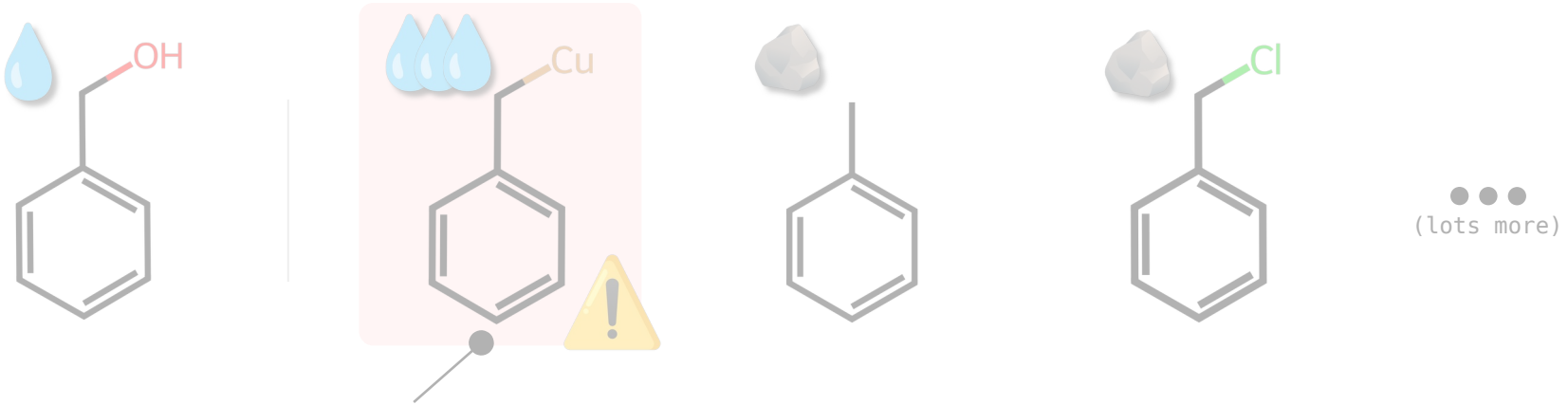


Particularly when generating counterfactual candidates, it is easy to end up with out-of-distribution samples!

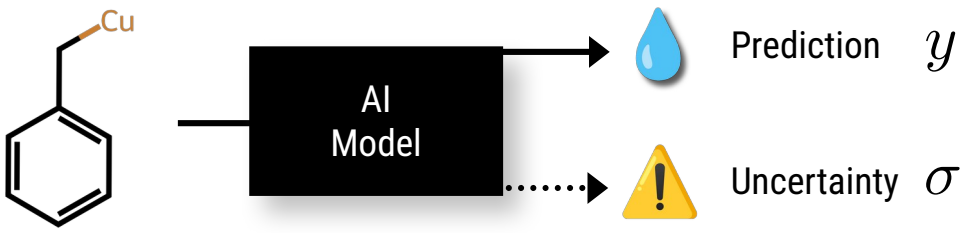


Improving counterfactuals with uncertainty quantification

[2] Teufel et al. Improving Counterfactual Truthfulness [...] with Uncertainty Quantification. XAI'24

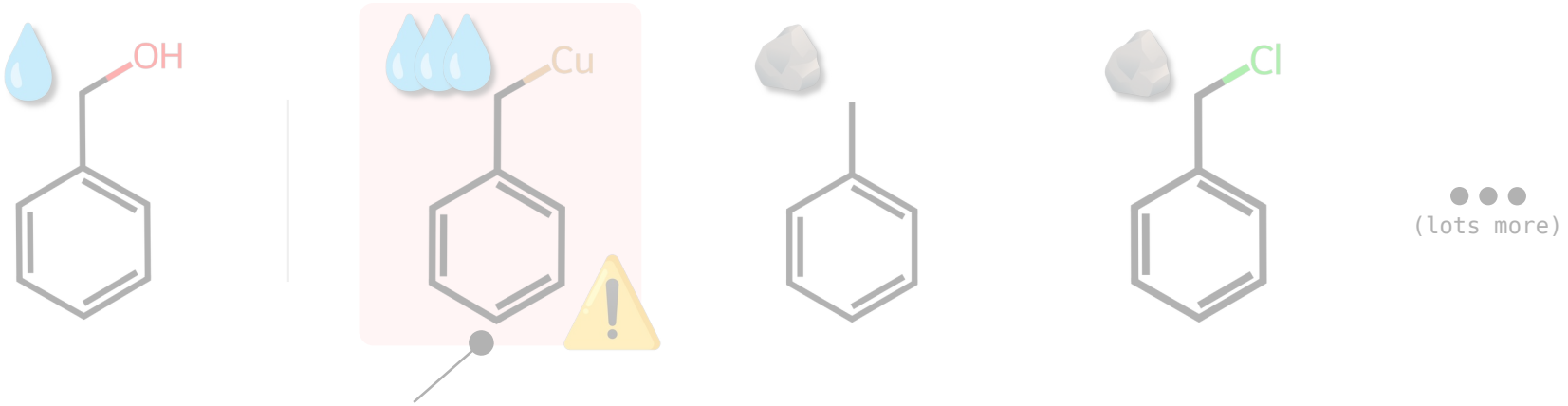


Particularly when generating counterfactual candidates, it is easy to end up with out-of-distribution samples!

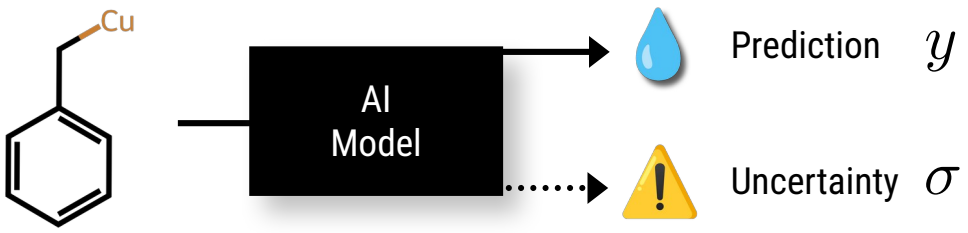


Improving counterfactuals with uncertainty quantification

[2] Teufel et al. Improving Counterfactual Truthfulness [...] with Uncertainty Quantification. XAI'24



Particularly when generating counterfactual candidates, it is easy to end up with out-of-distribution samples!

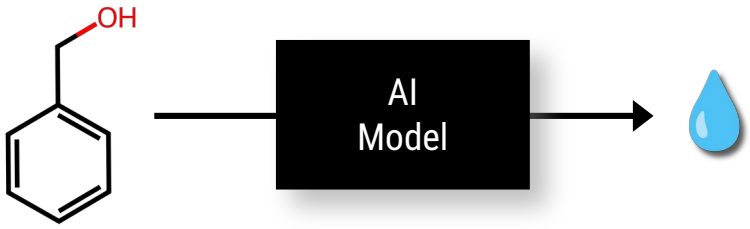


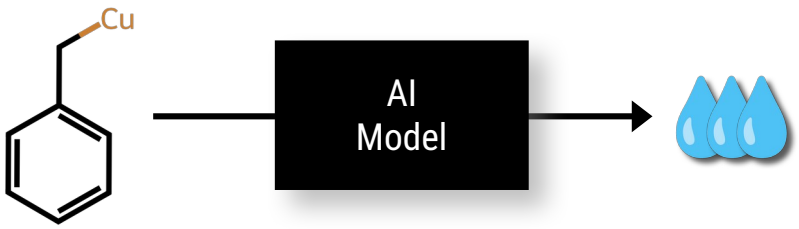
As an additional objective In counterfactual optimization we can minimize the uncertainty of the candidates

$$\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}'} \left(d(\mathbf{x}, \mathbf{x}') - \lambda_1 \cdot |y - y'| + \lambda_2 \cdot \sigma' \right)$$

Improving counterfactuals with uncertainty quantification II

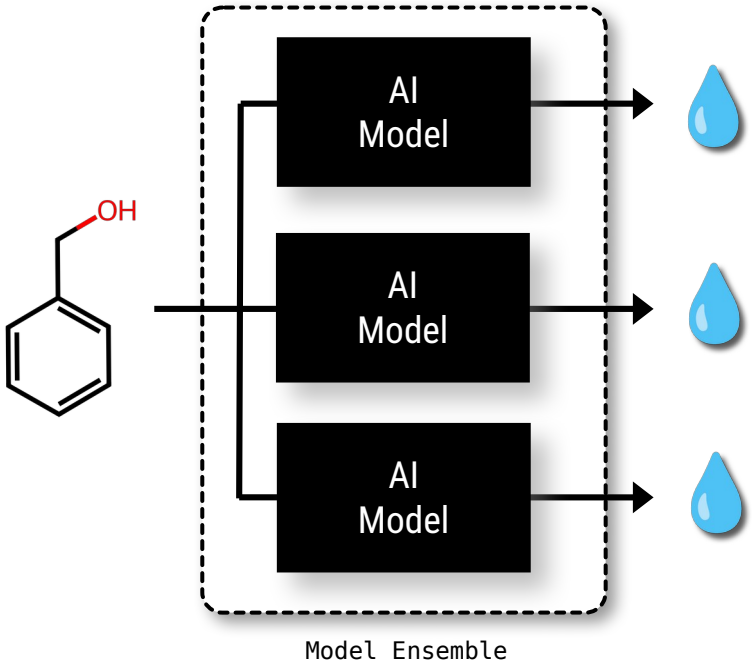
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

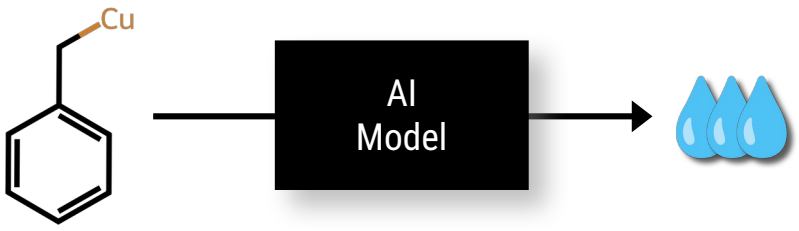




Improving counterfactuals with uncertainty quantification II

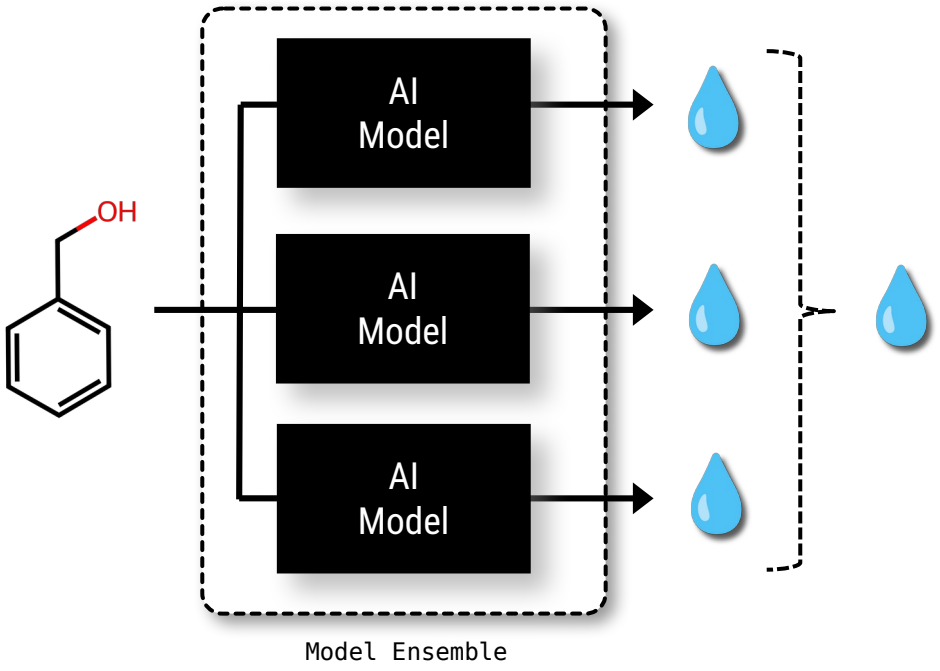
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

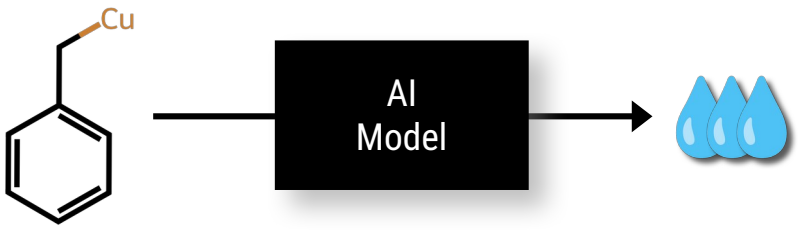




Improving counterfactuals with uncertainty quantification II

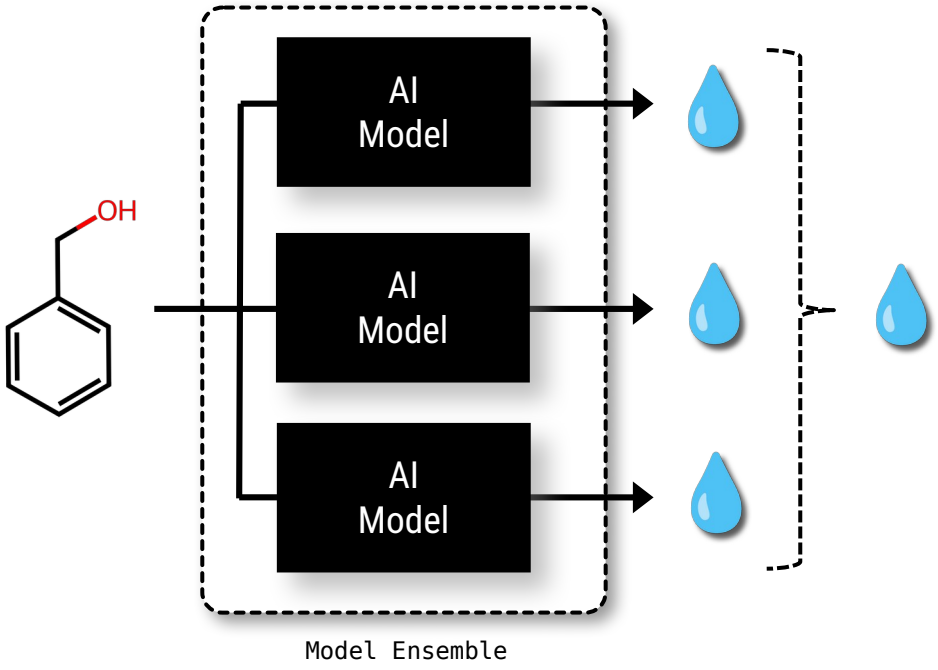
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



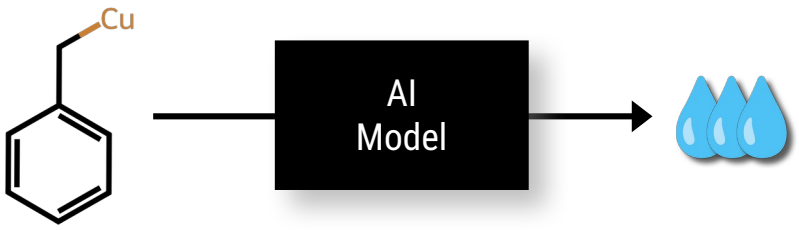


Improving counterfactuals with uncertainty quantification II

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

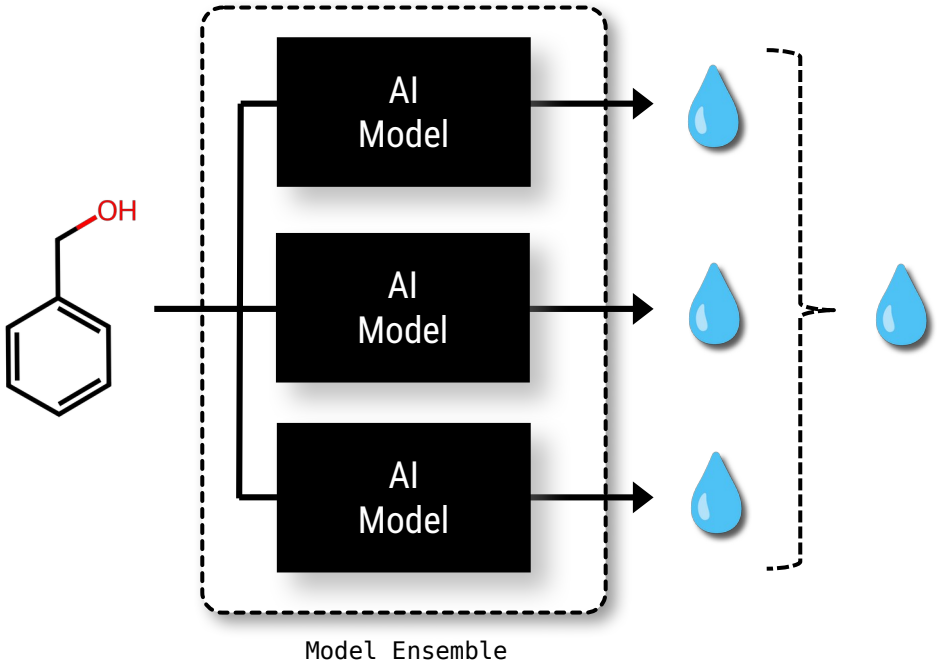


✅ Low Disagreement → Low Uncertainty

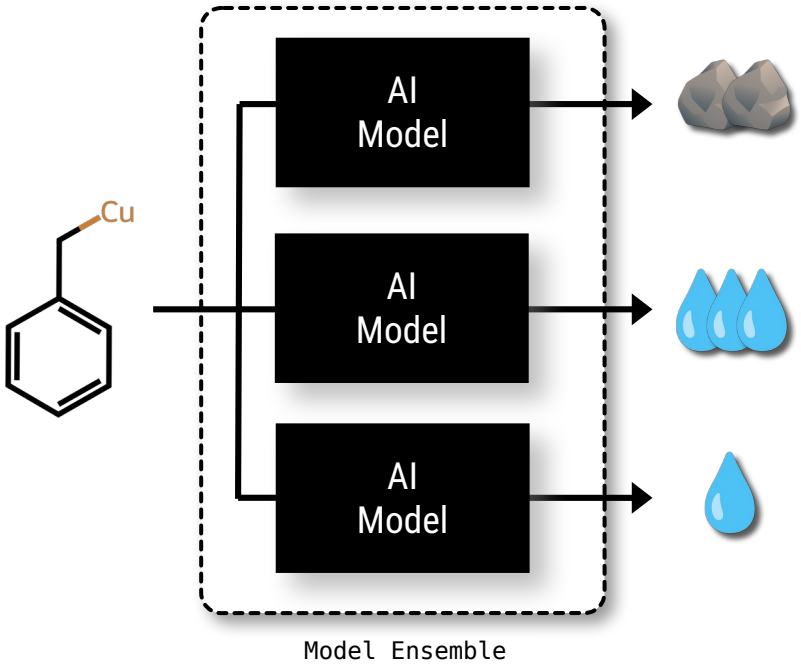


Improving counterfactuals with uncertainty quantification II

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

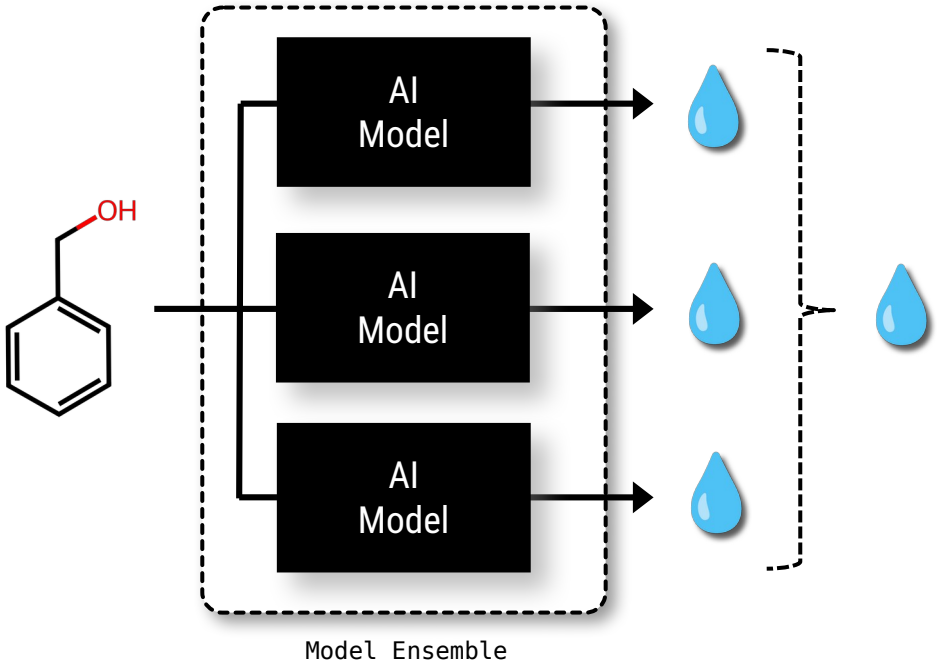


✅ Low Disagreement → Low Uncertainty

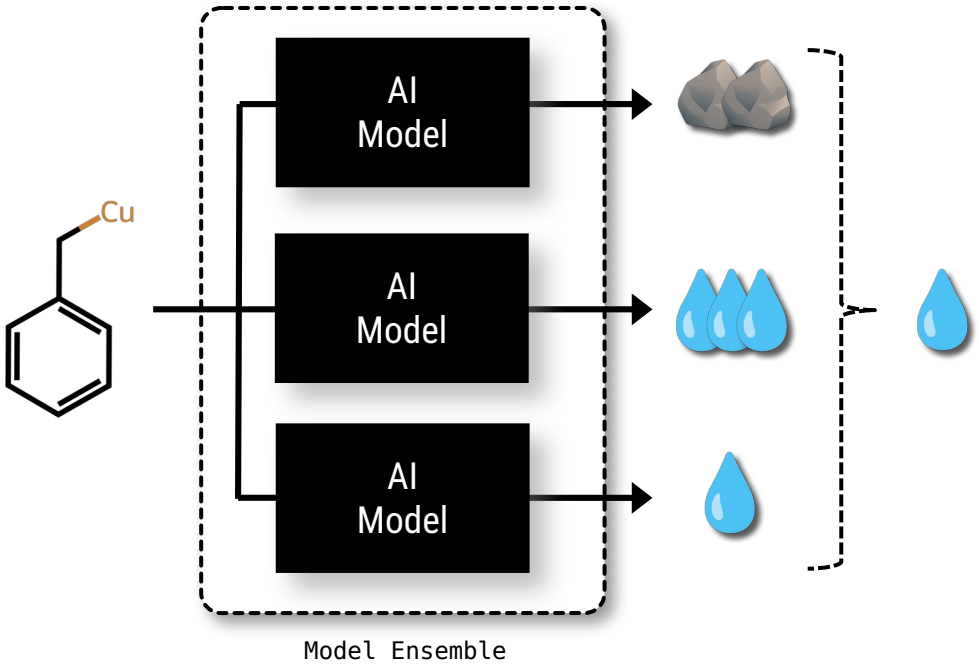


Improving counterfactuals with uncertainty quantification II

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

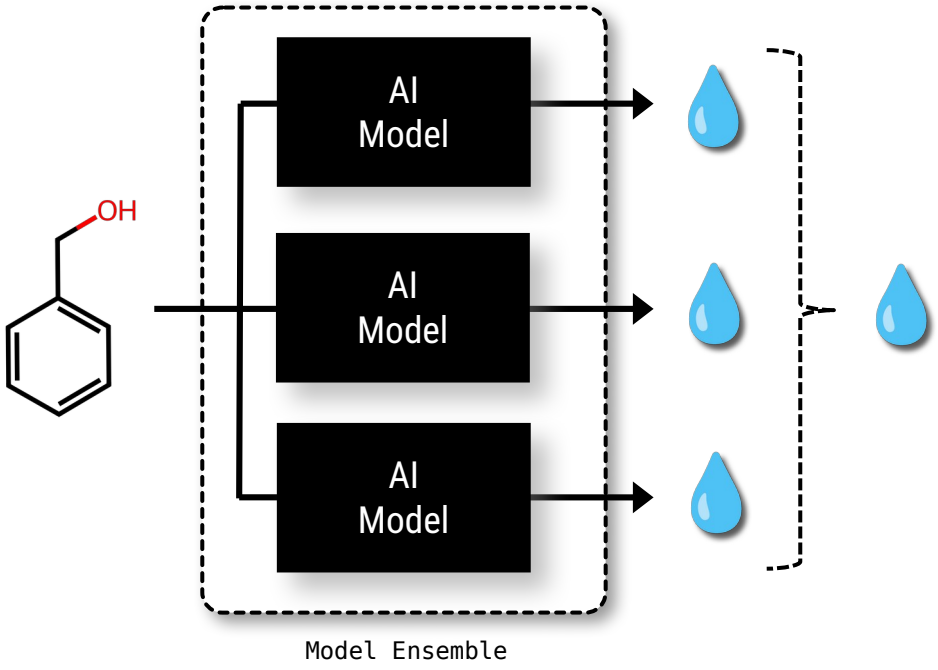


✅ Low Disagreement → Low Uncertainty

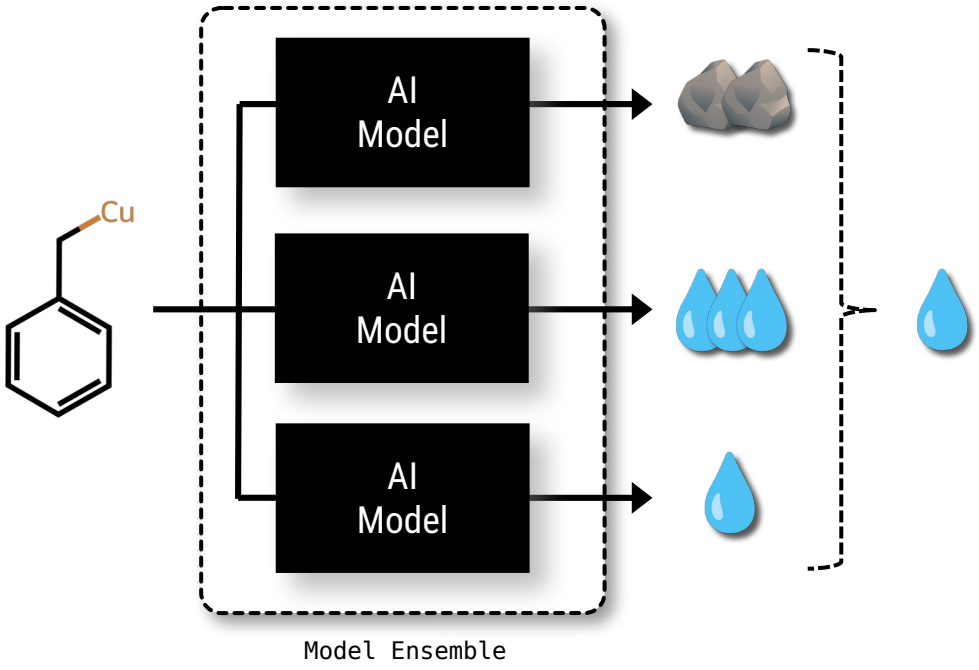


Improving counterfactuals with uncertainty quantification II

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



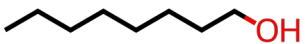
✅ Low Disagreement → Low Uncertainty

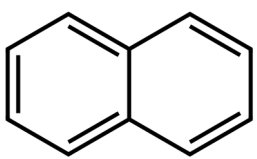


⚠️ High Disagreement → High Uncertainty

Improving counterfactuals with uncertainty quantification III

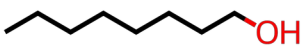
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

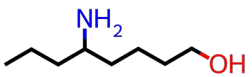


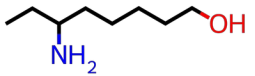


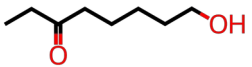
Improving counterfactuals with uncertainty quantification III

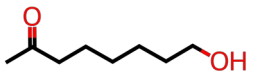
[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24

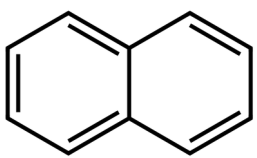






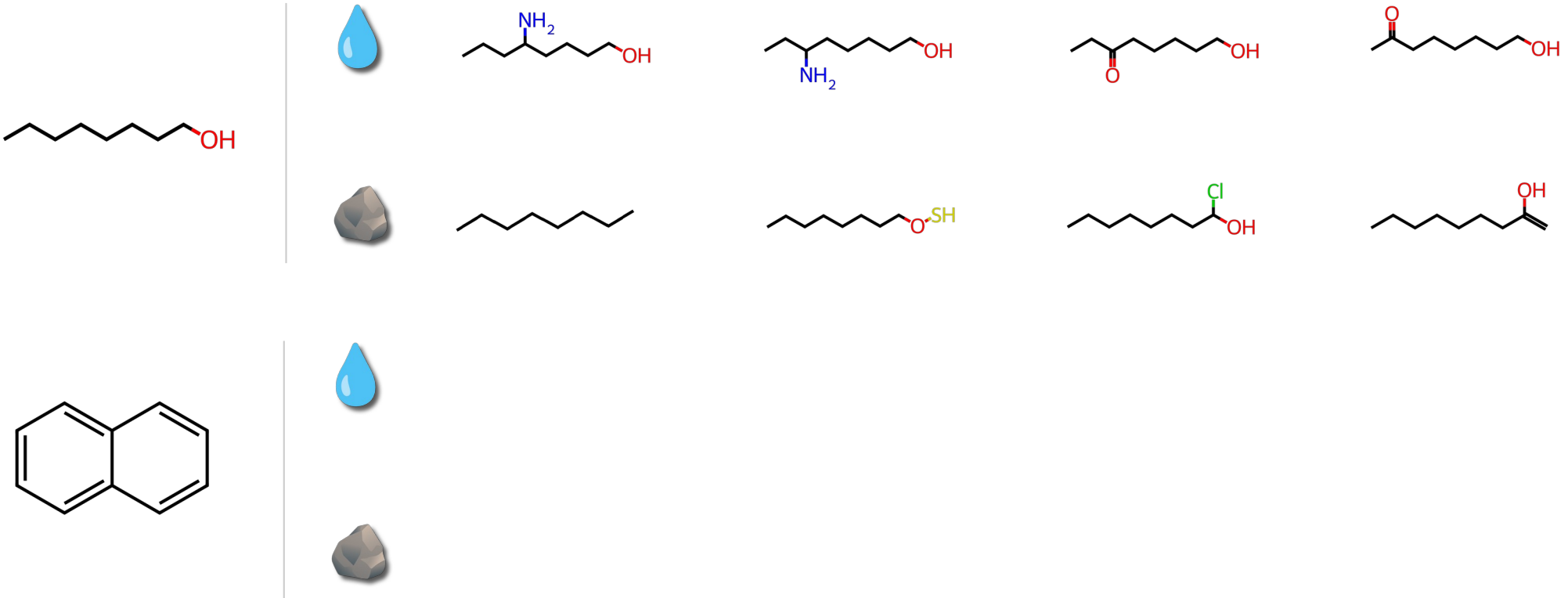






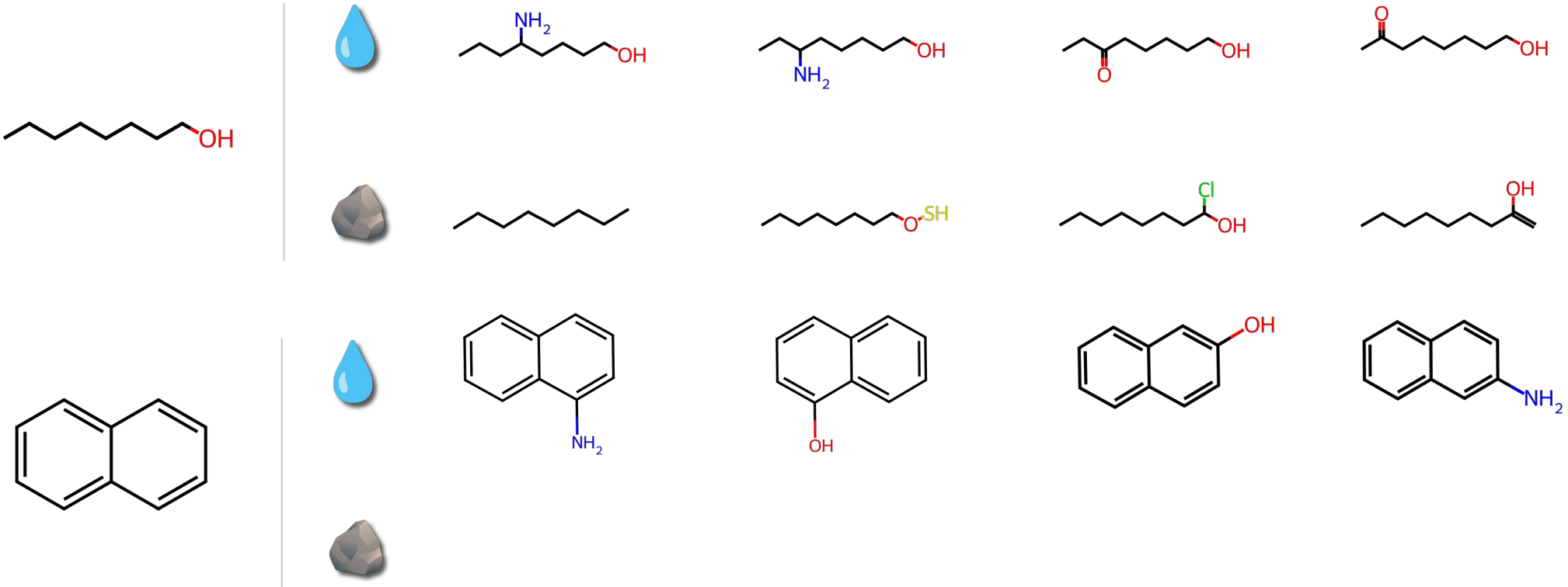
Improving counterfactuals with uncertainty quantification III

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



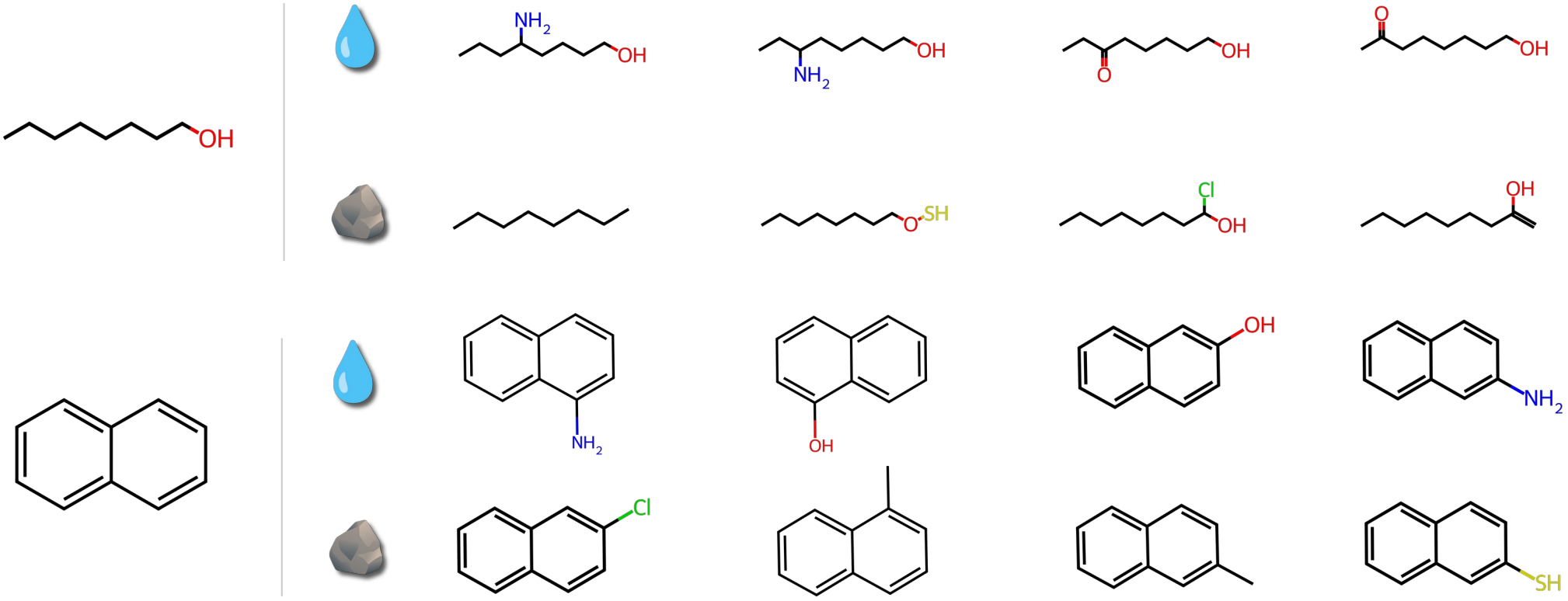
Improving counterfactuals with uncertainty quantification III

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



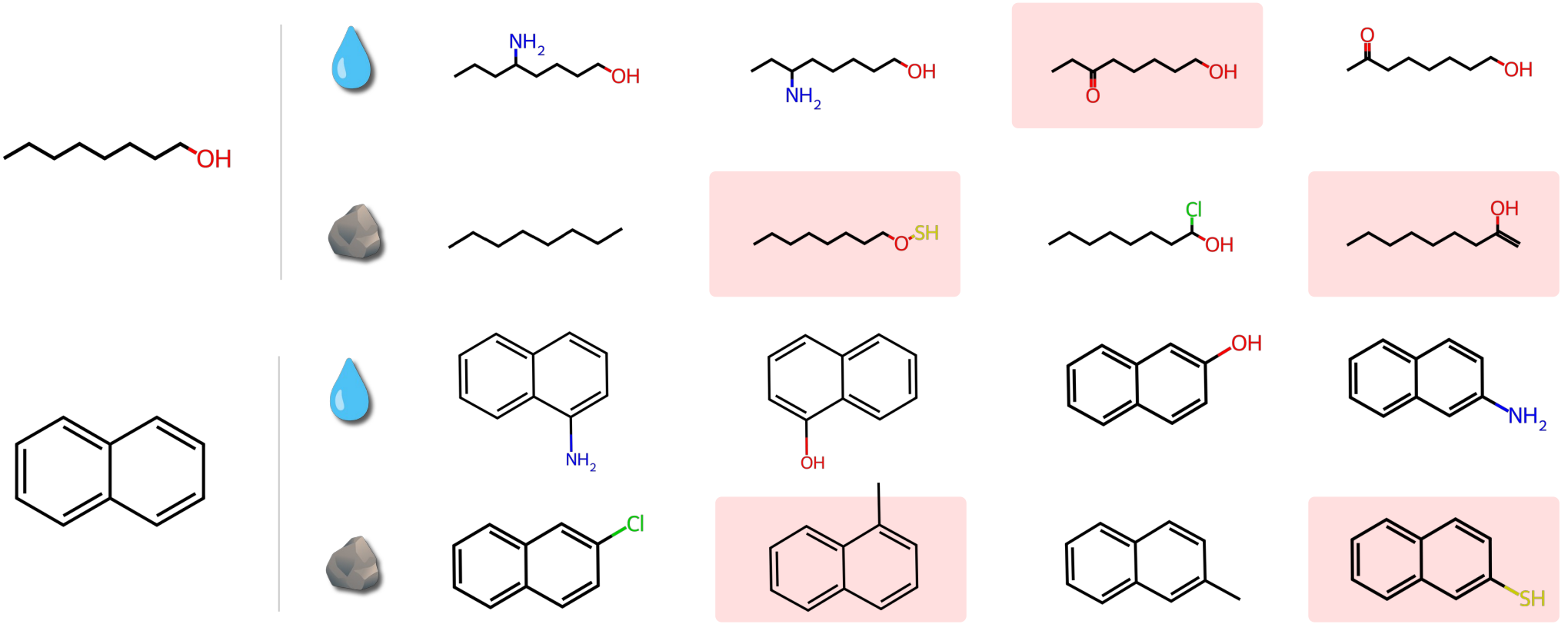
Improving counterfactuals with uncertainty quantification III

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



Improving counterfactuals with uncertainty quantification III

[2] Teufel et al. Improving Counterfactual Truthfulness […] with Uncertainty Quantification. XAI'24



Predicting small molecule aggregation

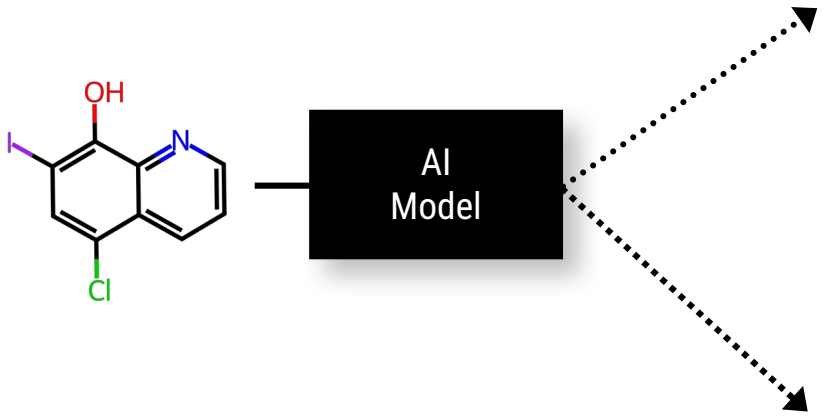
[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

DAVIS RESEARCH GROUP

University of Manitoba

Predicting small molecule aggregation

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

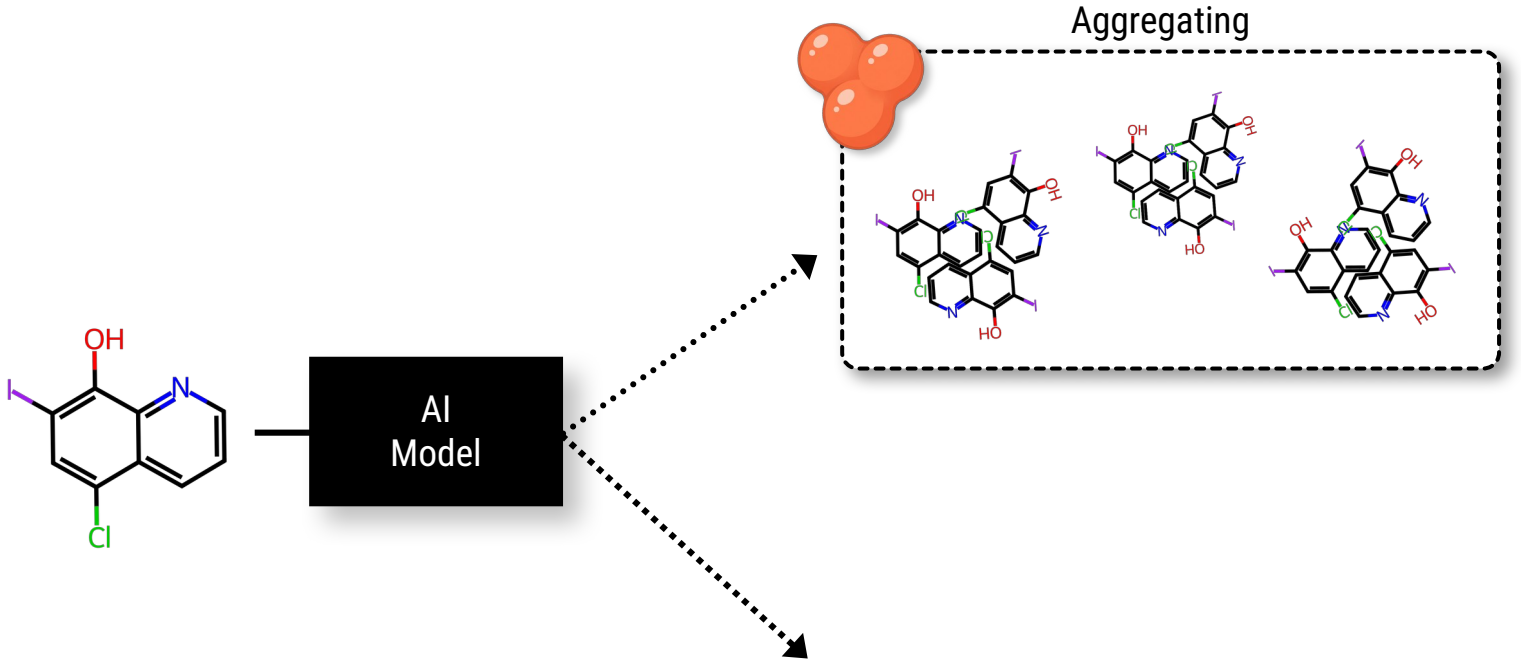


DAVIS RESEARCH GROUP

University of Manitoba

Predicting small molecule aggregation

[3] Sturm et al. Mitigating molecular aggregation in drug discovery [...]. Angew. Chem. 2025

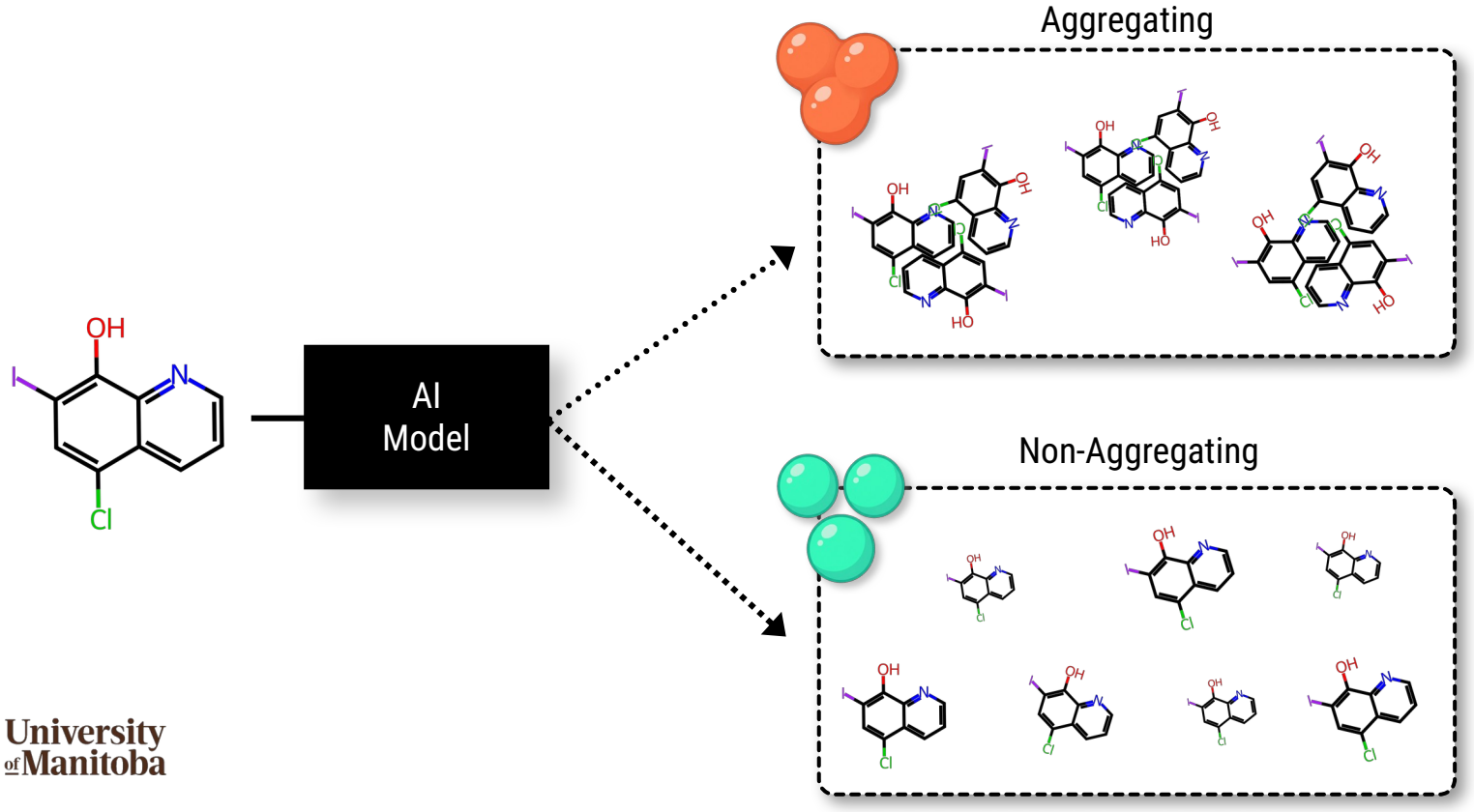


DAVIS RESEARCH GROUP

University of Manitoba

Predicting small molecule aggregation

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025



DAVIS RESEARCH GROUP

University of Manitoba

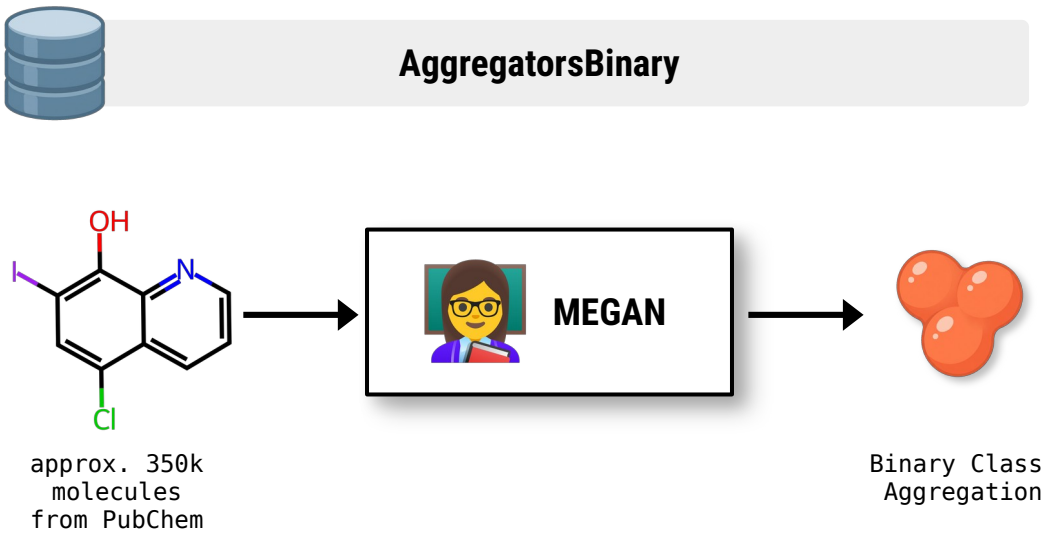
Predicting small molecule aggregation II

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

AggregatorsBinary

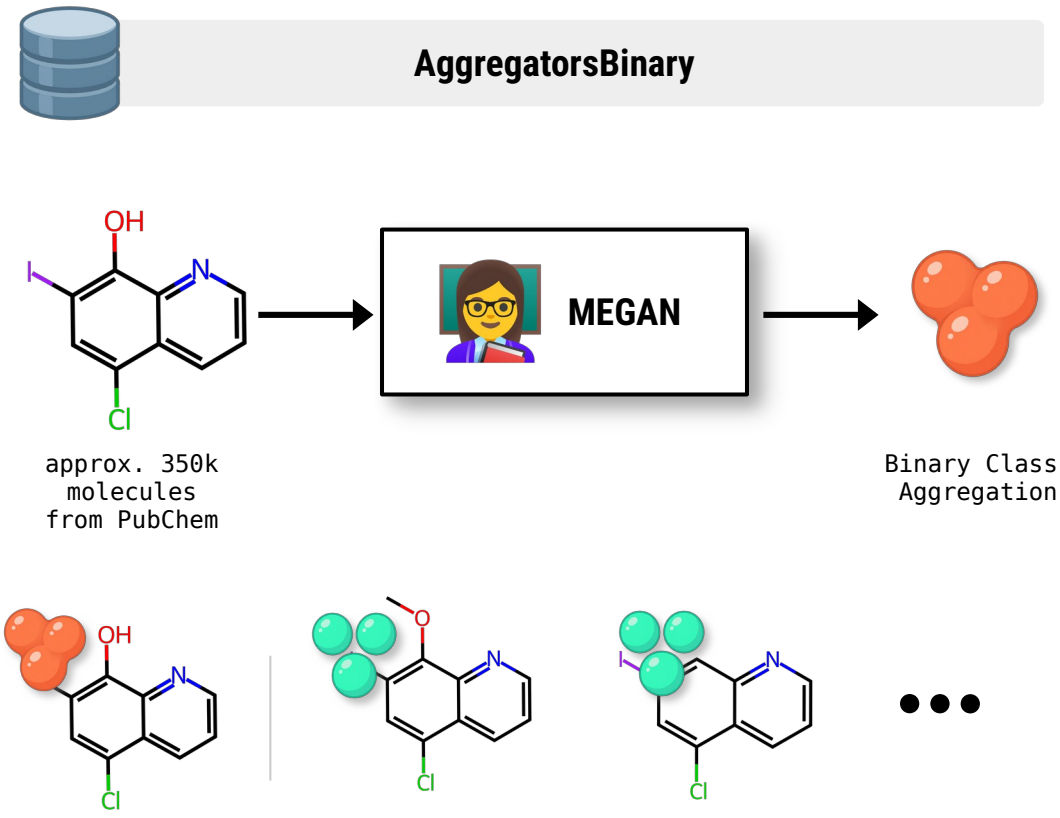
Predicting small molecule aggregation II

[3] Sturm et al. Mitigating molecular aggregation in drug discovery [...]. Angew. Chem. 2025



Predicting small molecule aggregation II

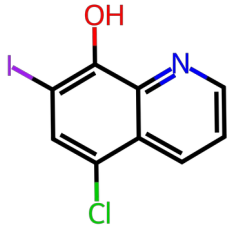
[3] Sturm et al. Mitigating molecular aggregation in drug discovery [...]. Angew. Chem. 2025



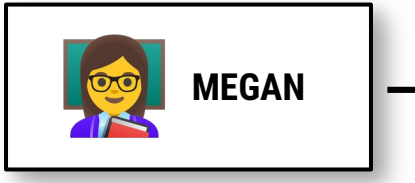
Predicting small molecule aggregation II

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

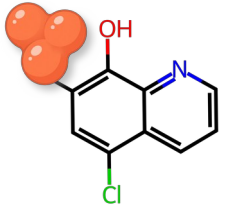
AggregatorsBinary

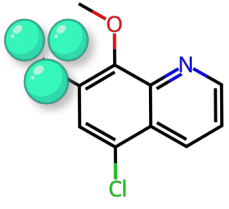


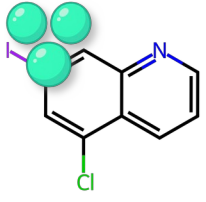
approx. 350k molecules from PubChem



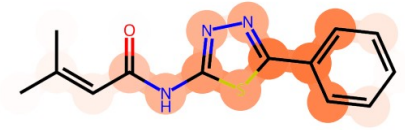
Binary Class: Aggregation

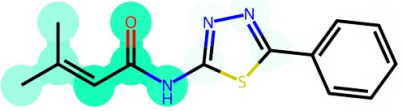






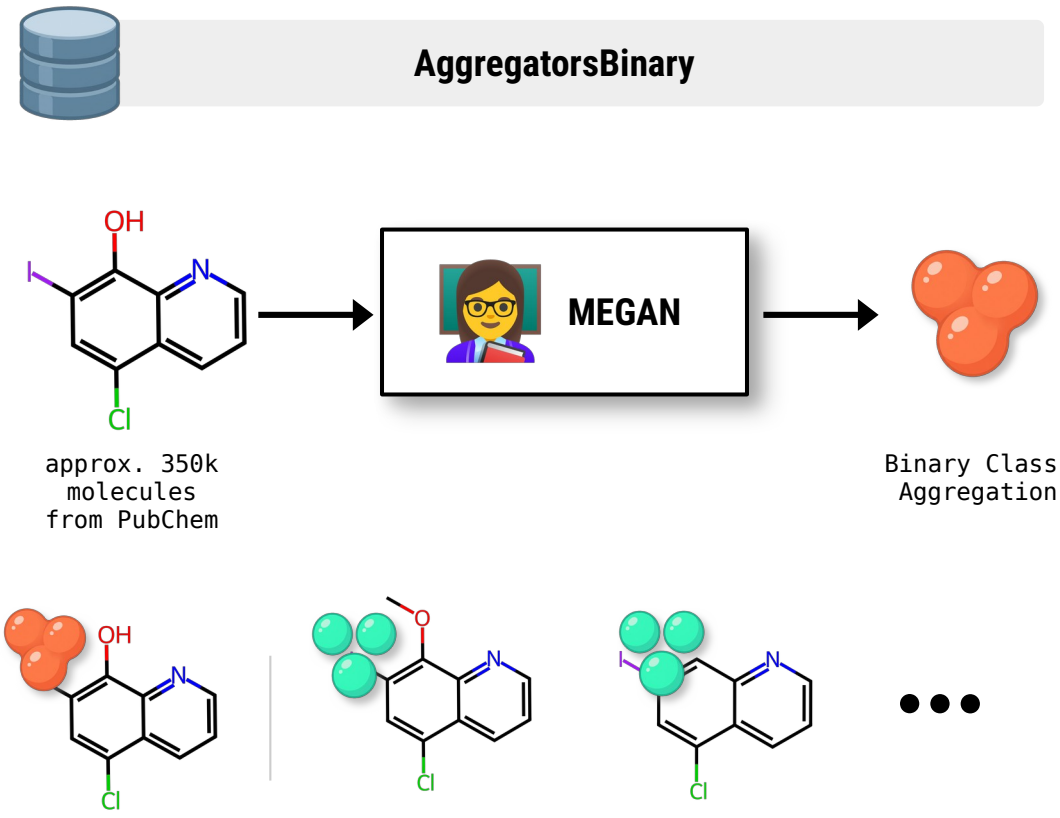
●●●

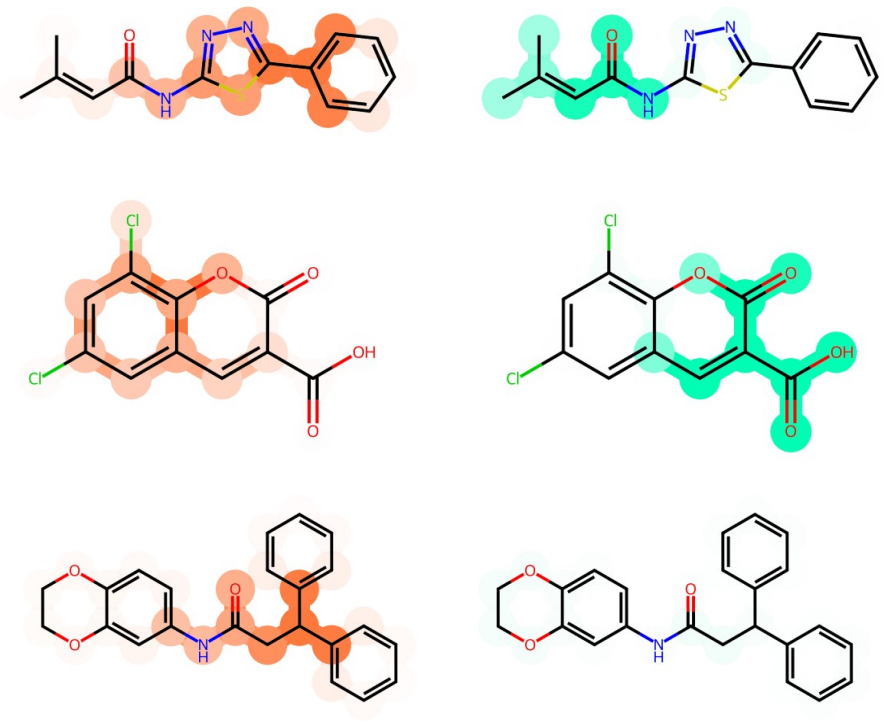




Predicting small molecule aggregation II

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

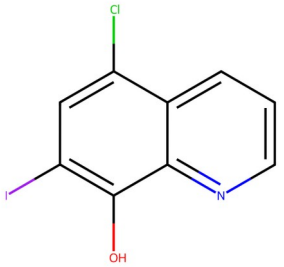




Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

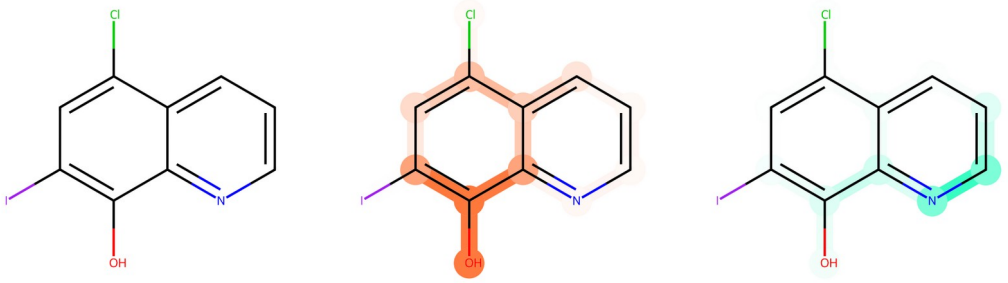
Case Study: **Clioquinol**



Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

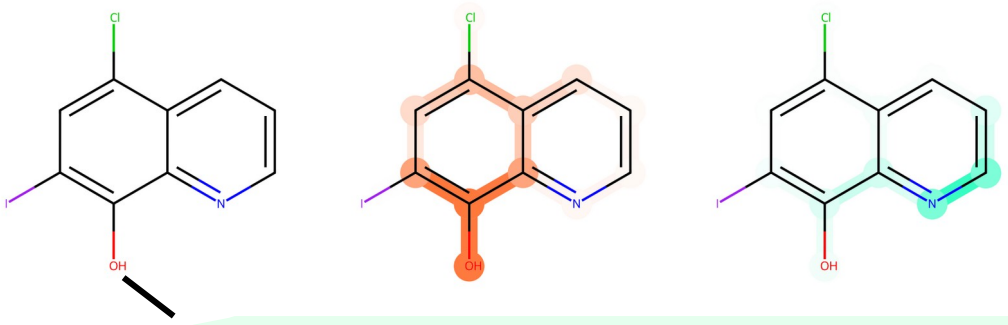
Case Study: **Clioquinol**



Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

Case Study: **Clioquinol**

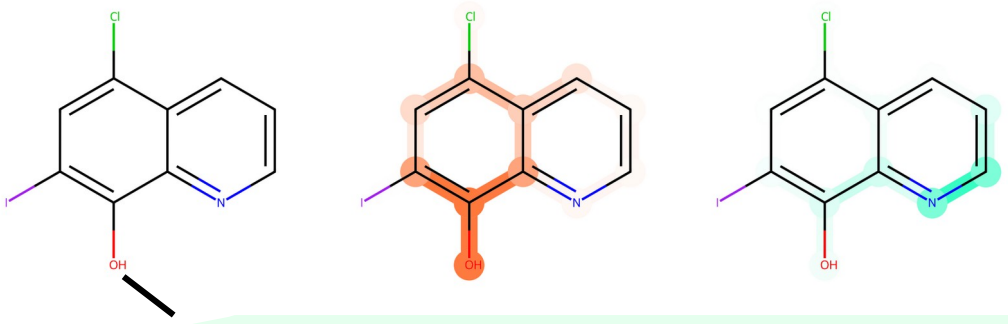


Counterfactual: Methyl-Clioquionol (Non-Agg)

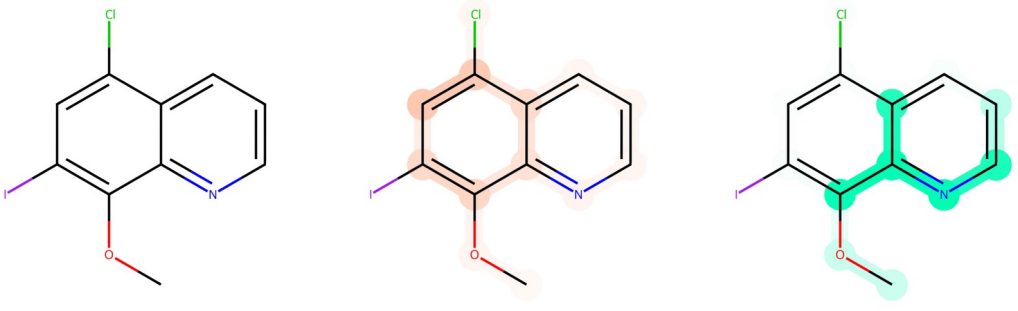
Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

Case Study: **Clioquinol**



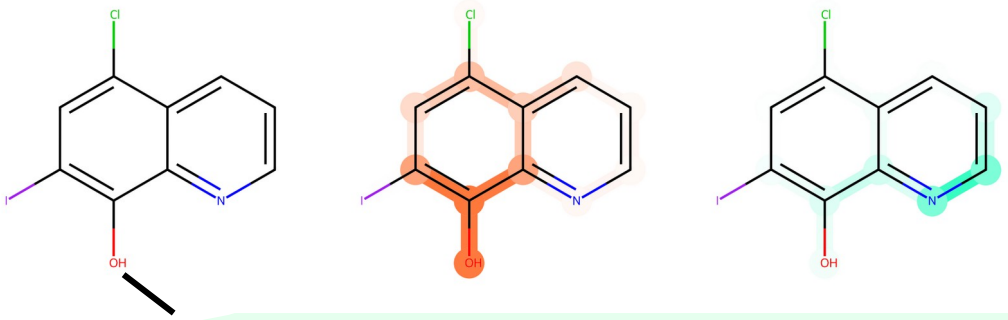
Counterfactual: Methyl-Clioquionol (Non-Agg)



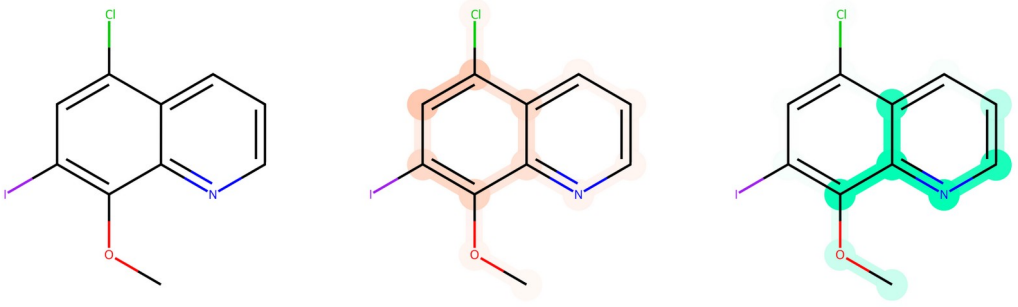
Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

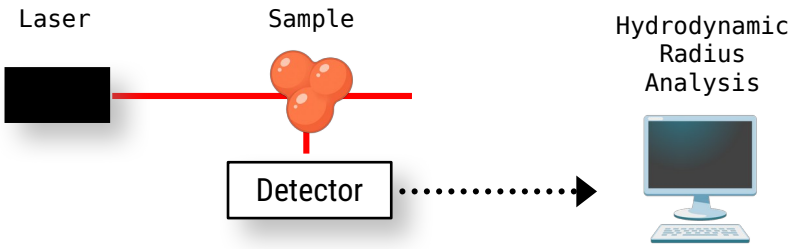
Case Study: **Clioquinol**



Counterfactual: Methyl-Clioquionol (Non-Agg)



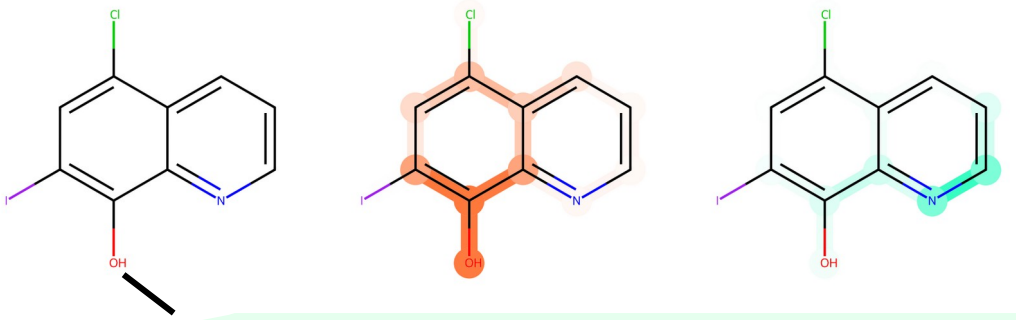
Dynamic Light Scattering (DLS) Verification



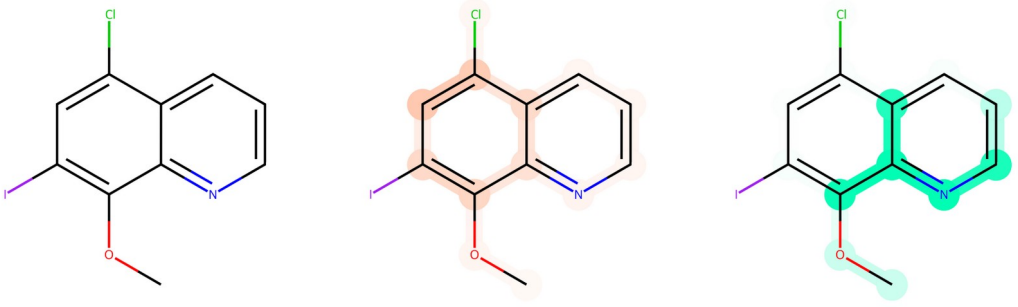
Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

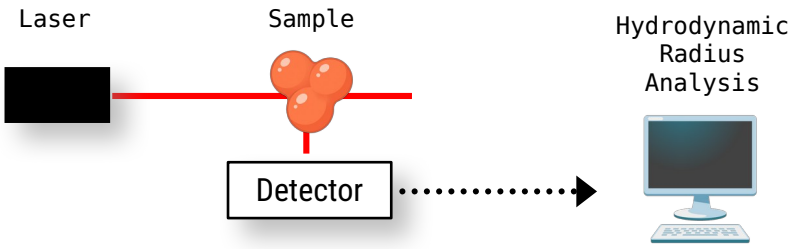
Case Study: **Clioquinol**

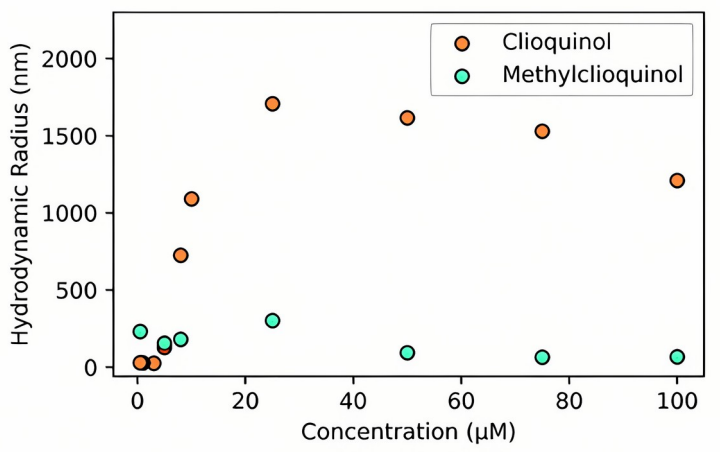


Counterfactual: Methyl-Clioquionol (Non-Agg)



Dynamic Light Scattering (DLS) Verification

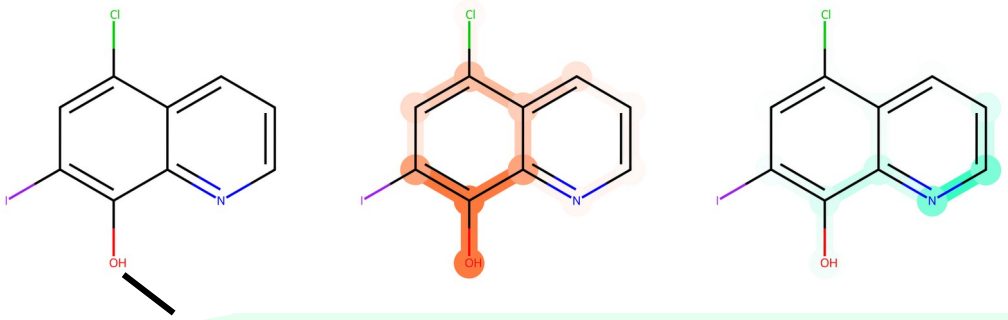




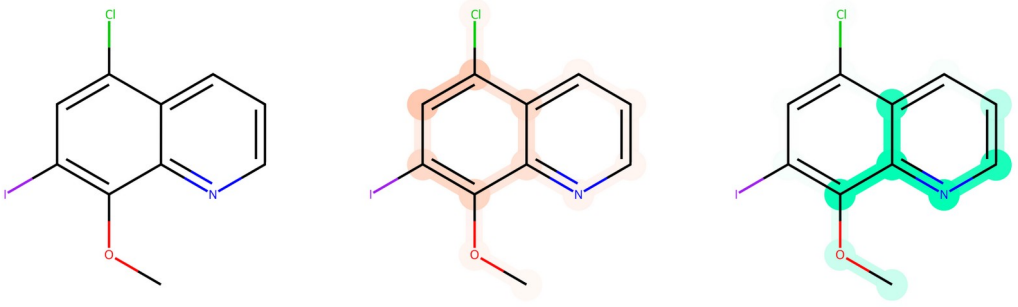
Predicting small molecule aggregation III

[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025

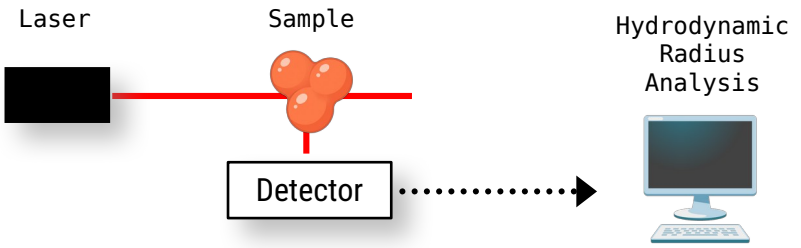
Case Study: **Clioquinol**

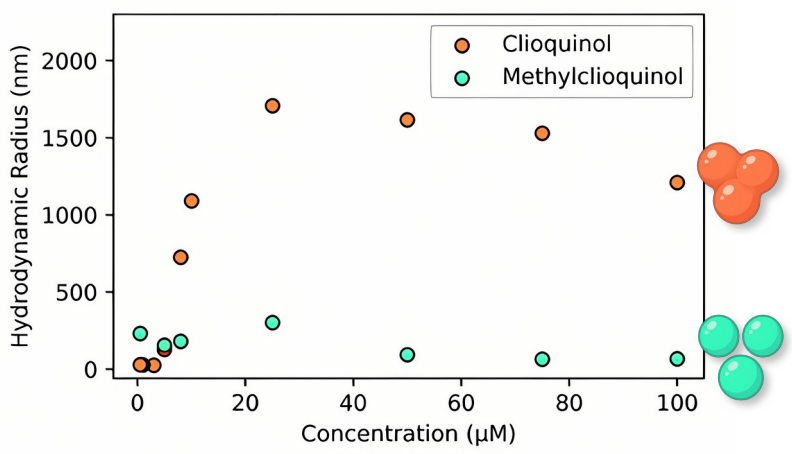


Counterfactual: Methyl-Clioquionol (Non-Agg)



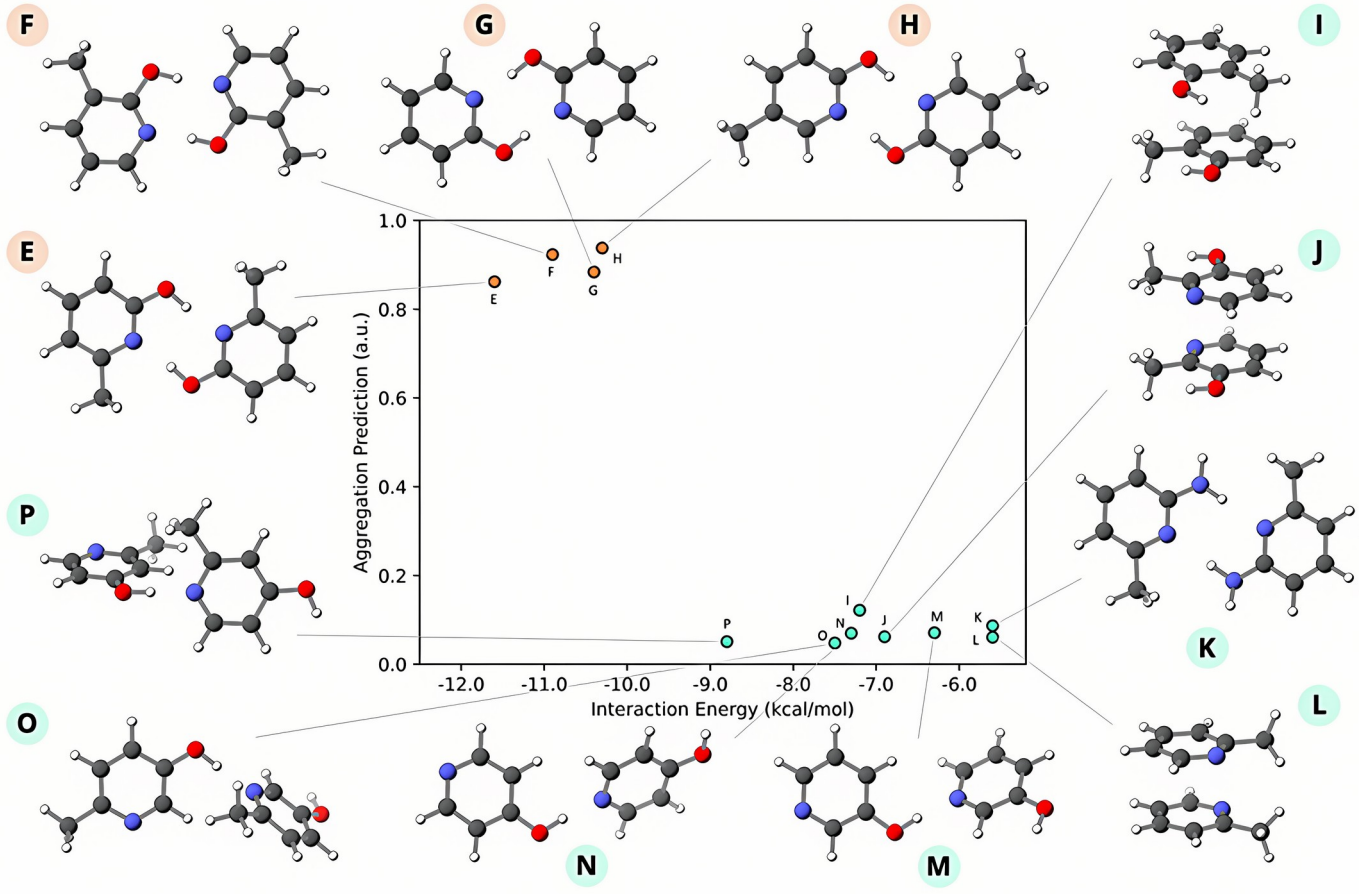
Dynamic Light Scattering (DLS) Verification



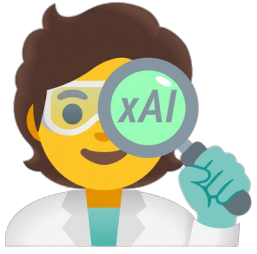


Predicting small molecule aggregation IV

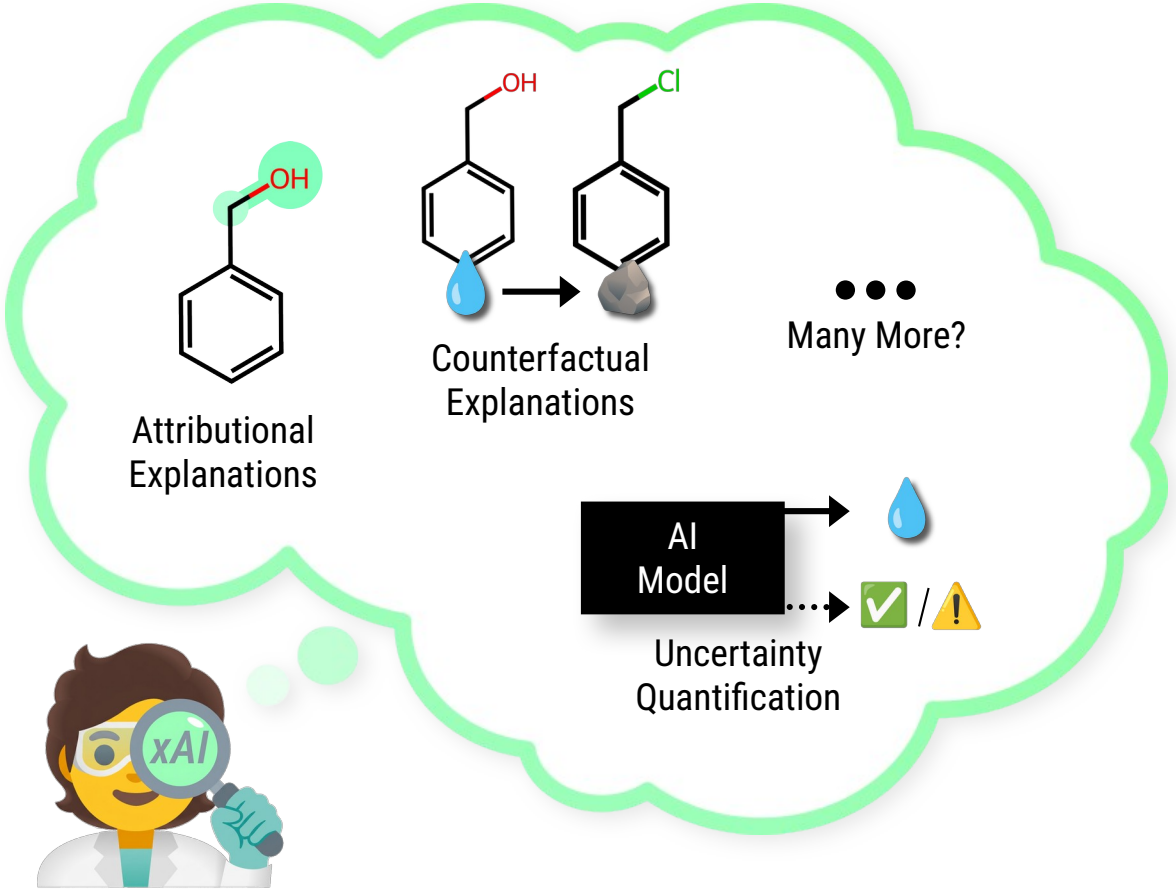
[3] Sturm et al. Mitigating molecular aggregation in drug discovery […]. Angew. Chem. 2025



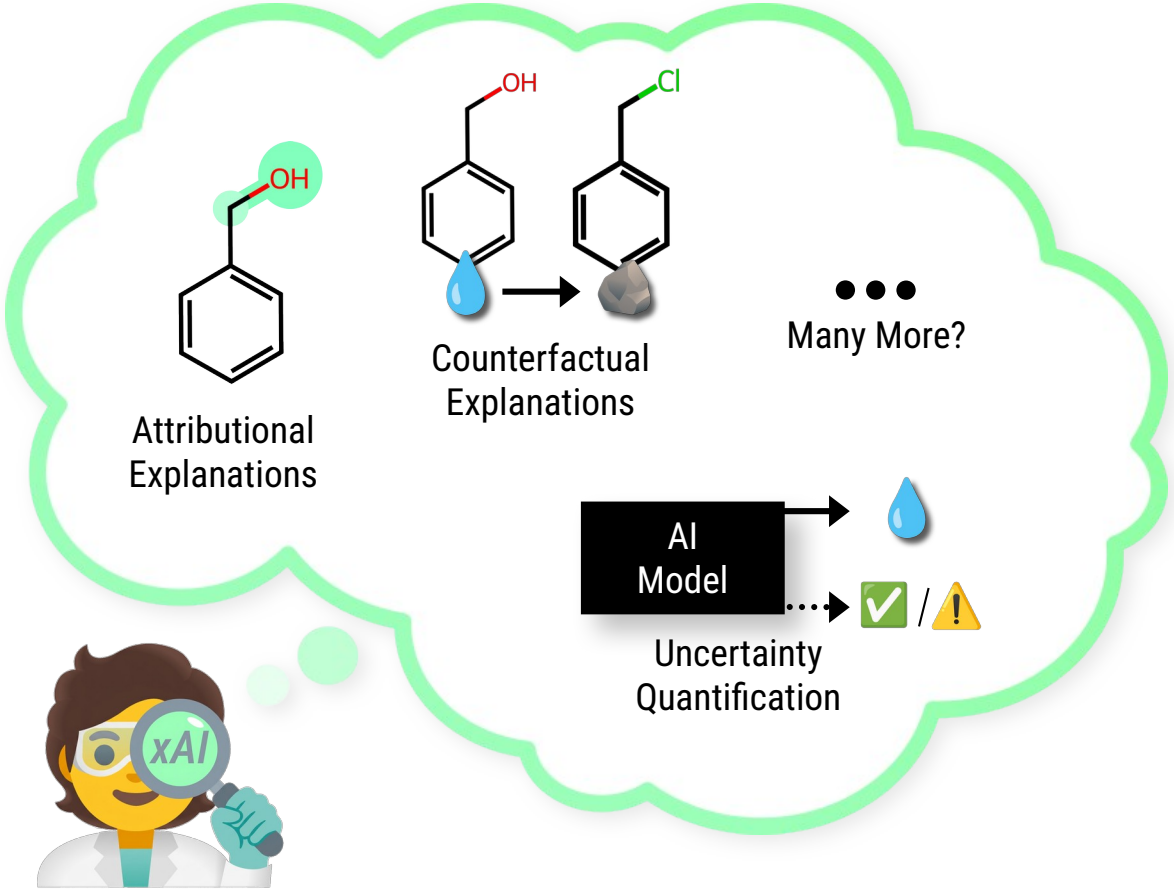
Using XAI to understand complex properties



Using XAI to understand complex properties



Using XAI to understand complex properties



Pitfalls

- Out-of-distribution effects (e.g. Counterfactuals)
- Explainer hyperparameter sensitivity
- Dataset biases
- …

Thanks for your attention!
Questions?