Enhancing Interpretability in Molecular Property Prediction with Contextual Explanations of Molecular Graphical Depictions

Marco Bertolini (Pfizer)

In collaboration with:

Linlin Zhao (Bayer AG)

Djork-Arné Clevert (Pfizer)

Floriane Montanari (Bayer AG)

19th September 2024

Explainability for molecules - challenges

Common deep learning strategies

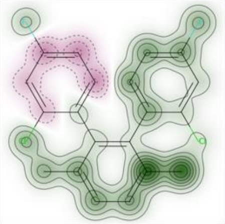
Input type **Explanations**

SMILES based networks

- Attributions are assigned to all input features
- Attributions for structural characters are hard to interpret and visualize

- XAI outputs restrict to attributions for atoms, neglecting input information

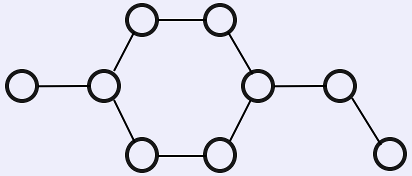
Cc1ccc(C)c(c2ccc(F)cc2Cl)c1-c1ccc(F)cc1Cl

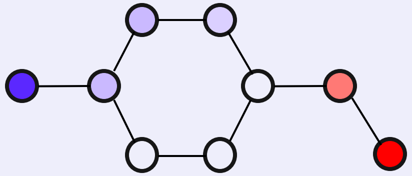


Graph Neural Networks

- Molecules are represented as graphs
- Atoms → Nodes
- Bonds → Edges

- Attributions are assigned to input features, which are atoms
- Measure of how a given atom contribute positevily/negatively to the prediction

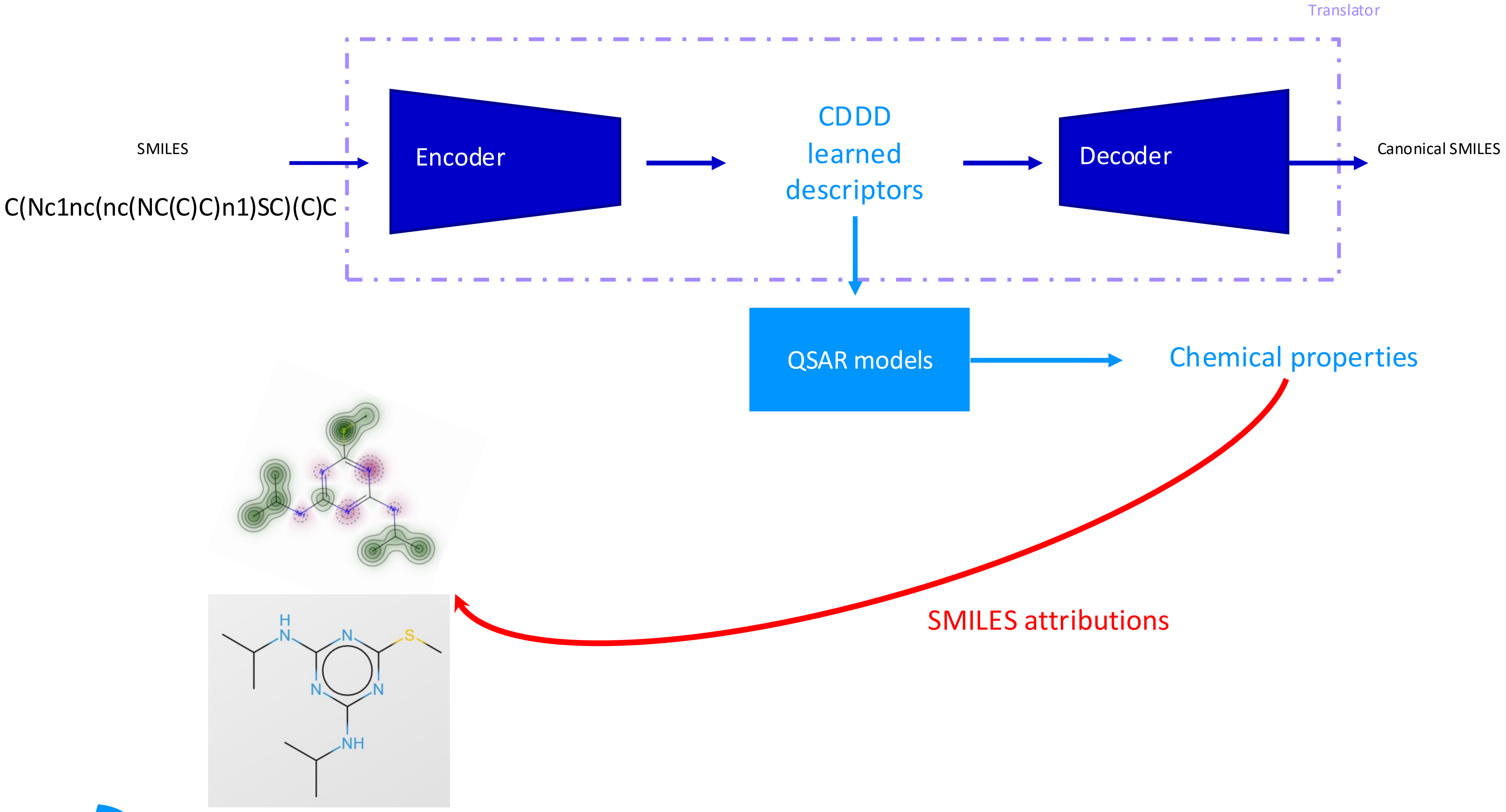




Challenges

1. Lack of global explainability
   - Methods are forced to express explanation in terms of input quantities
   - This exludes more advanced concepts related to global structures (rings, etc.)
2. Lack of symmetry
   - Often the symmetry of a molecule is explicitly broken by the input modality (e.g., SMILES)
   - This is reflected in the XAI attributions
3. Lack of sparsity
   - Often explanations are cluttering and therefore less informative

Models from CDDD space



Translator

SMILES

C(Nc1nc(nc(NC(C)C)n1)SC)(C)C

Encoder

CDDD learned descriptors

Decoder

Canonical SMILES

QSAR models

Chemical properties

SMILES attributions

Winter et al., *"Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations", Chem. Sci., 2019*

Generating SMILES attributions

	O	C	C	(	C	)	C	C
C		1	1		1		1	1
O	1							
(				1				
)						1		

x

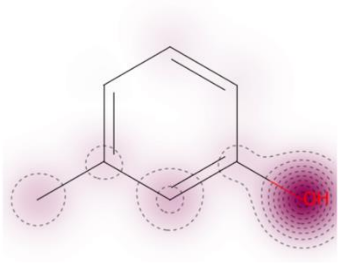
		1		1		1	1
1	■						
			1				
					1		

$\tilde{x}$

$$Attr_C = f\big(g(x)\big) - \frac{1}{n-1}\sum_{i}^{n-1} f(g(\tilde{x}_i))$$

n: vocabulary size

Attributions for each atom

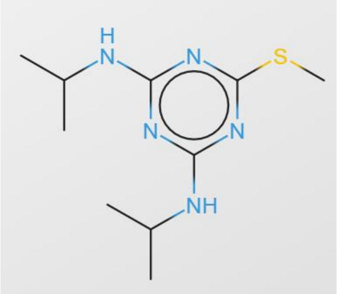


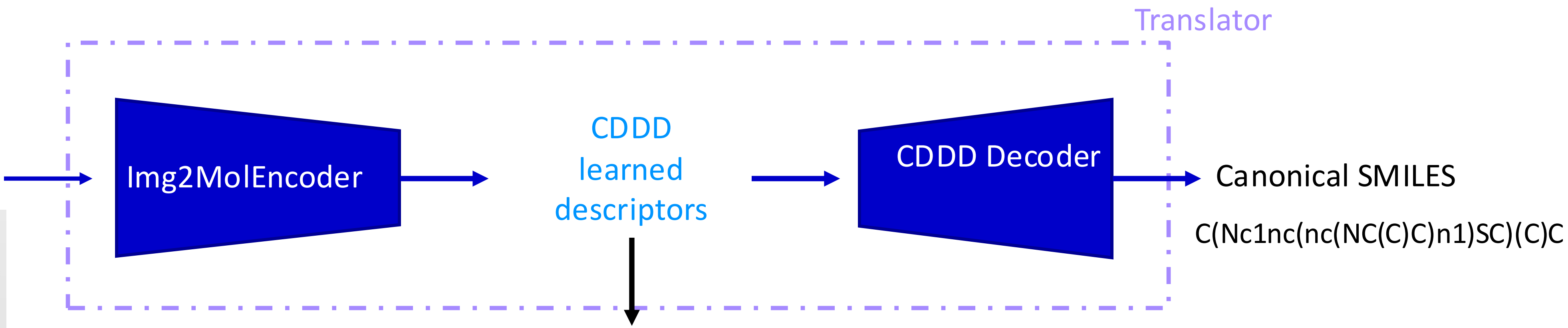
Cc1cccc(O)c1

Zhao et al. *„Modeling bioconcentration factors in fish with explainable deep learning", Artificial Intelligence in the Life Sciences*

Img2Mol

Image





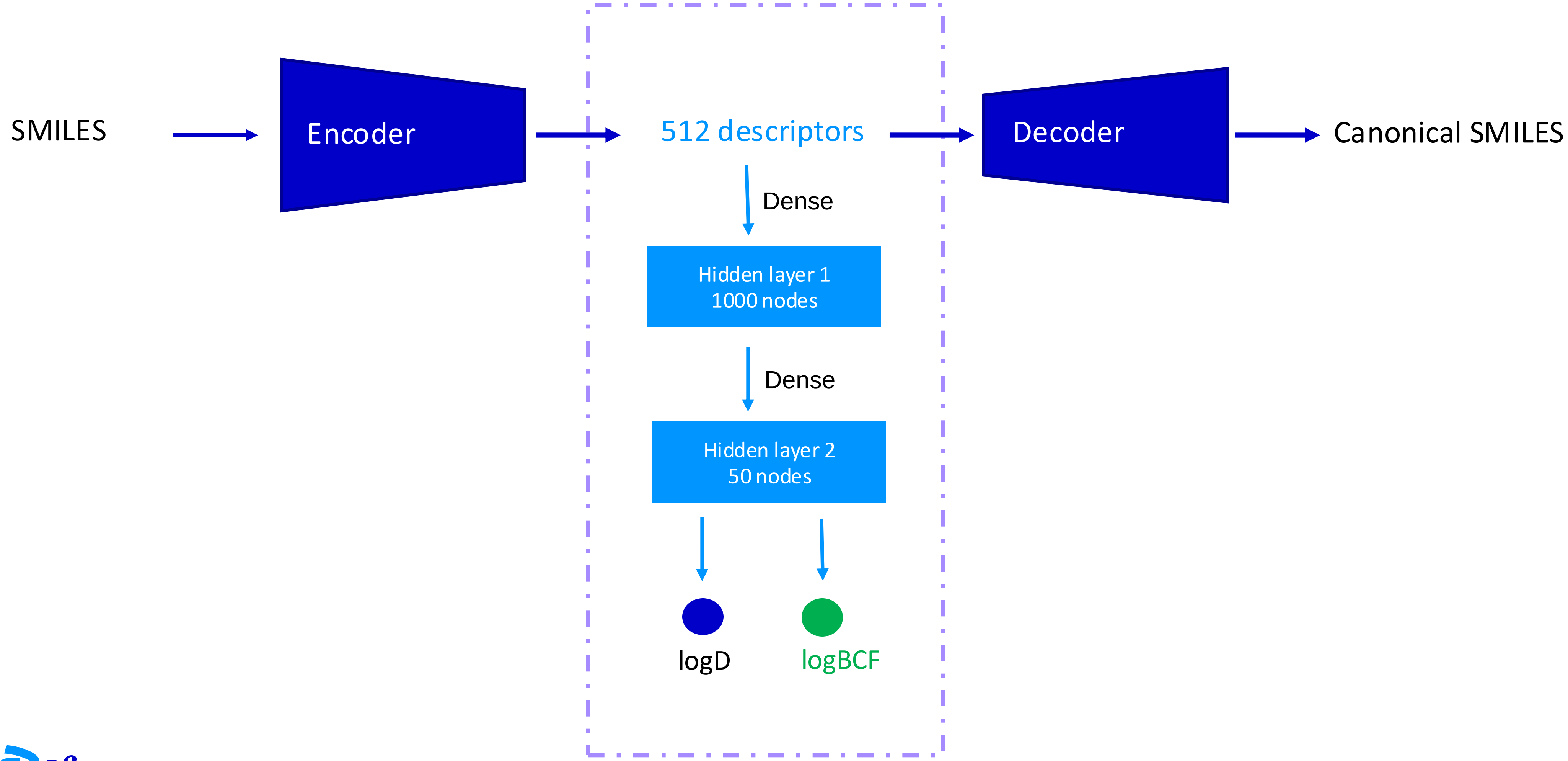
C(Nc1nc(nc(NC(C)C)n1)SC)(C)C

Encoder trained to map images embeddings to their corresponding CDDDs.

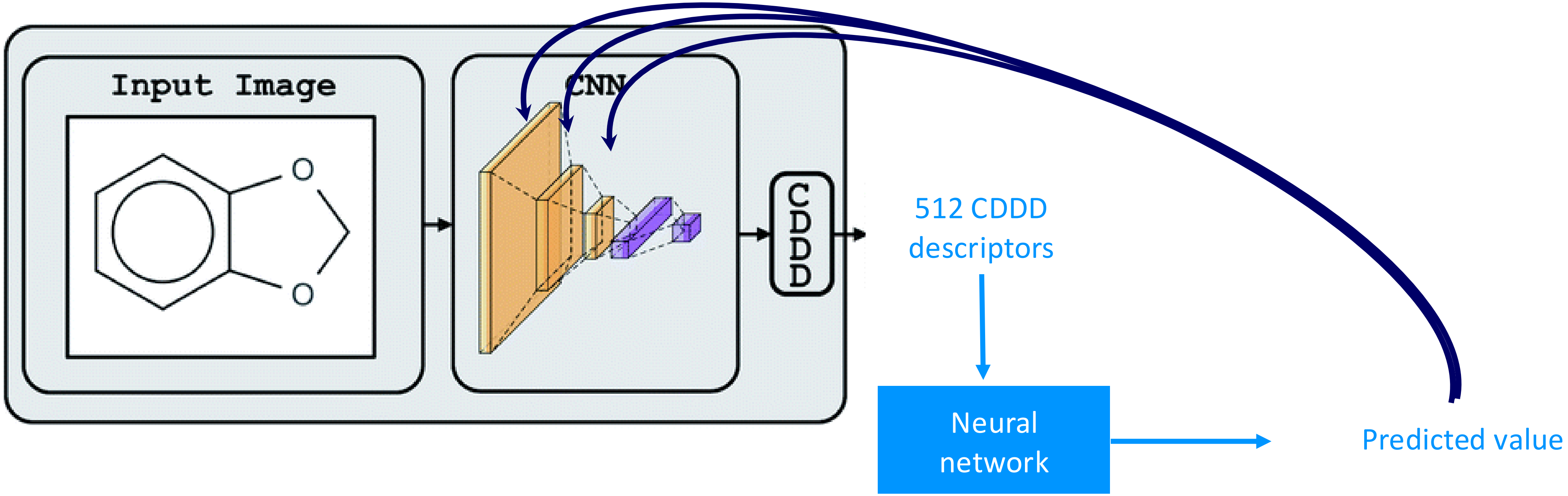
$$\mathrm{Loss} = (\mathrm{cddd}_{\mathrm{true}} - \mathrm{cddd}_{\mathrm{pred}})^2$$

Clevert et al., 2021, Img2Mol – Accurate SMILES recognition from Molecular Graphical Depictions, Chemical Science, 2022

Model use case



Overall concept

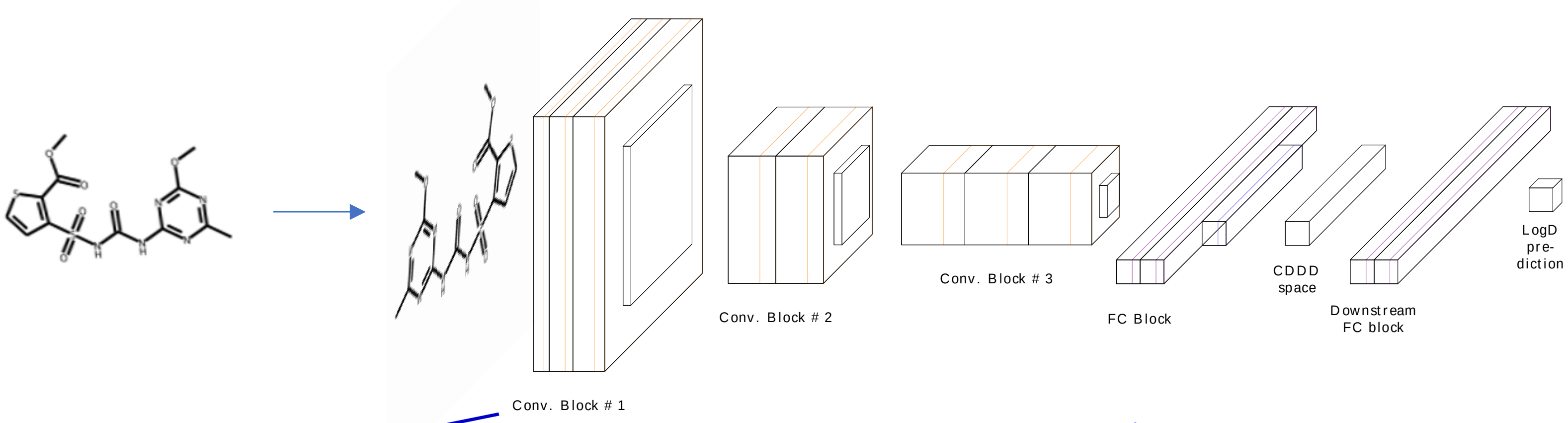


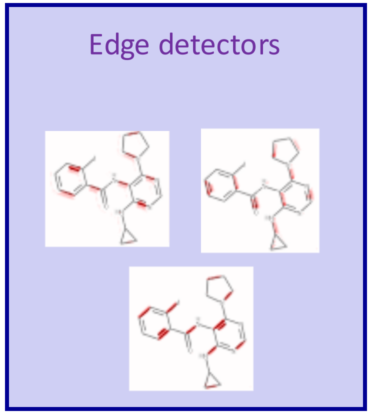
512 CDDD descriptors

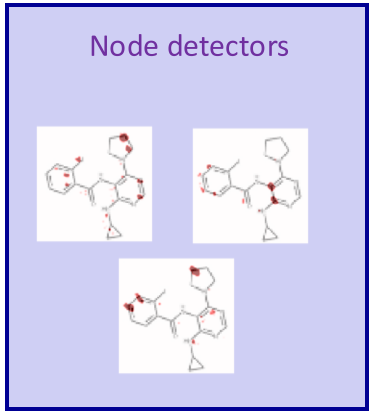
Neural network

Predicted value

Img2mol learns local and global concepts

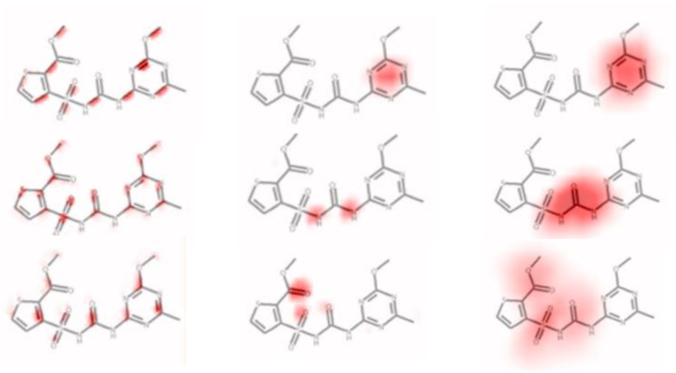


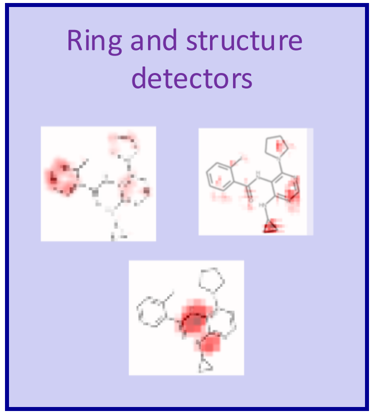




- Shallow layers learn simple geometric features, e.g., edges and nodes
- These also correspond to basic chemical concepts

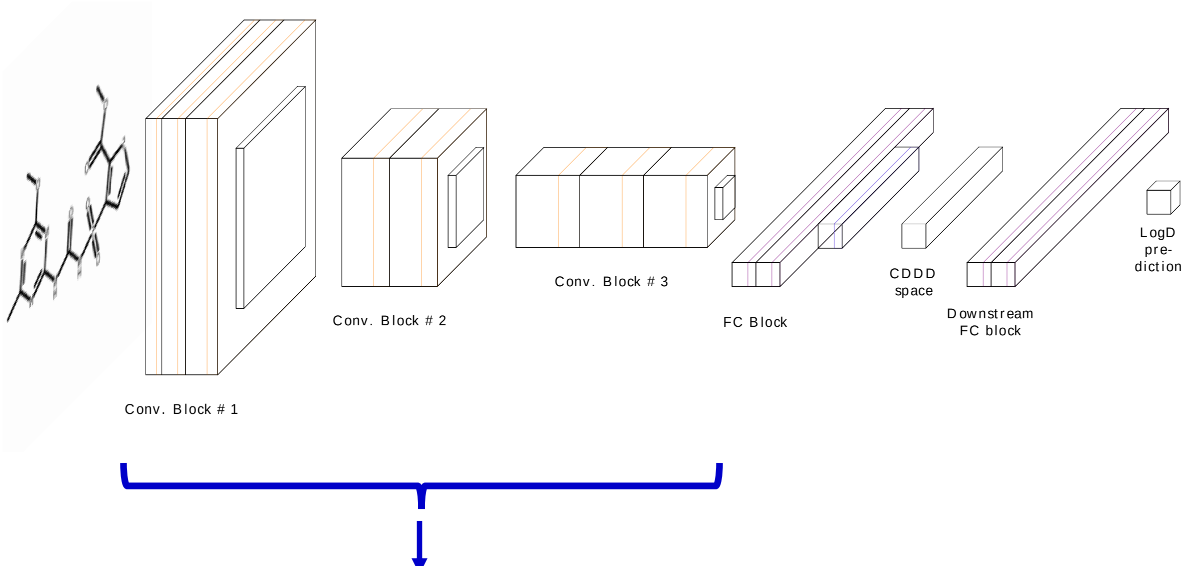
Layer activations



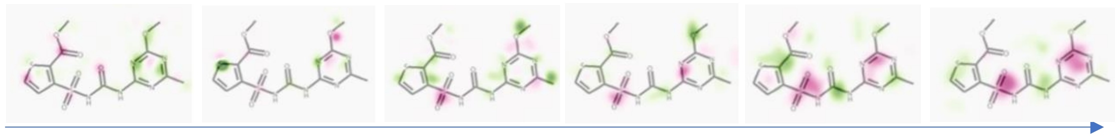


- Deeper layers learn more advanced geometric features, e.g., rings
- Such layers also learn high level concepts, which translate to chemical substructures

Local and global explanations

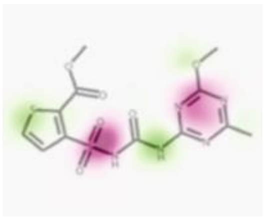


Layer attributions



network's depth

layer aggregation



STRATEGY

1. Explain downstream prediction (logD)

$$\Phi = \text{Img2Mol} \circ \Lambda : \mathcal{M} \xrightarrow{\psi_p} \mathcal{M}_p \xrightarrow{\xi_p} \mathcal{C} \xrightarrow{\Lambda} \mathbb{R}$$

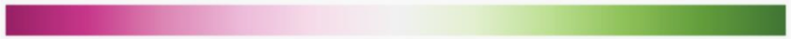
2. Compute attributions for each convolutional layer

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

Gradients restricted at layer l

Activation restricted at layer l

Measure of importance contribution of layer-learned features to the prediction



Negative contribution to the prediction

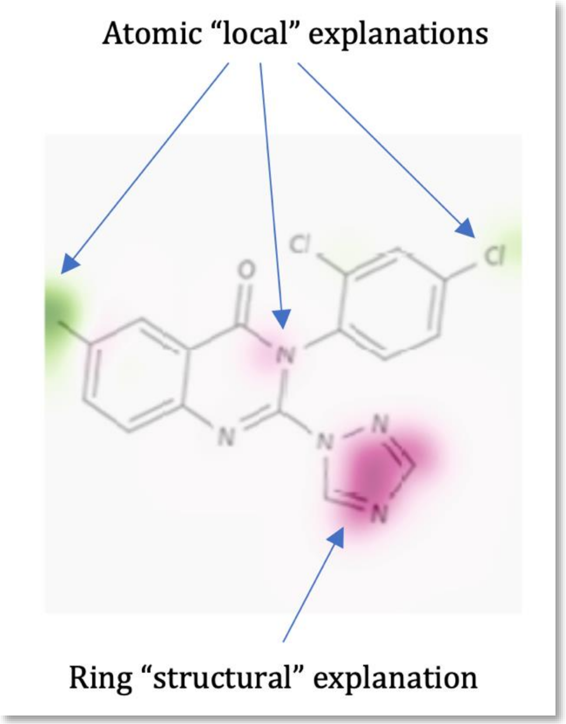
Positive contribution to the prediction

3. Aggregate over layers

$$a(X) = \sum_l a_l(X)$$

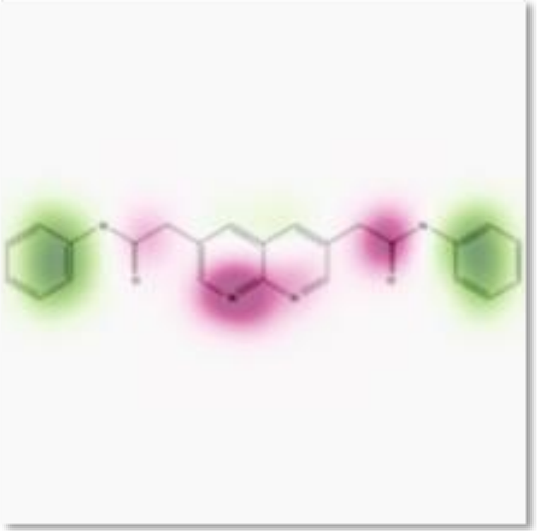
- ➢ Automatic weighting between layers' contribution
- ➢ No need to ad-hoc restrict to local or global features

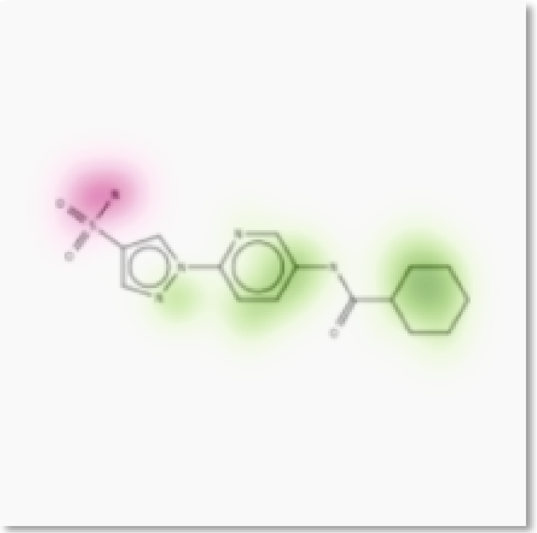
Examples



Atomic “local” explanations

Ring “structural” explanation





Invariance with respect to symmetries

SYMMETRY SCORE

- Let $\mathcal{T}$ be the symmetry group of a molecule's graphical depiction
  - $x, T(x)$ correspond to same molecule for all $T \in \mathcal{T}$

$$\begin{tikzcd}
\mathbf{x} \arrow[r, "a"] \arrow[d, "T"'] & a(\mathbf{x}) \arrow[d, "T"] \\
T(\mathbf{x}) \arrow[r, "a"] & a'
\end{tikzcd}$$

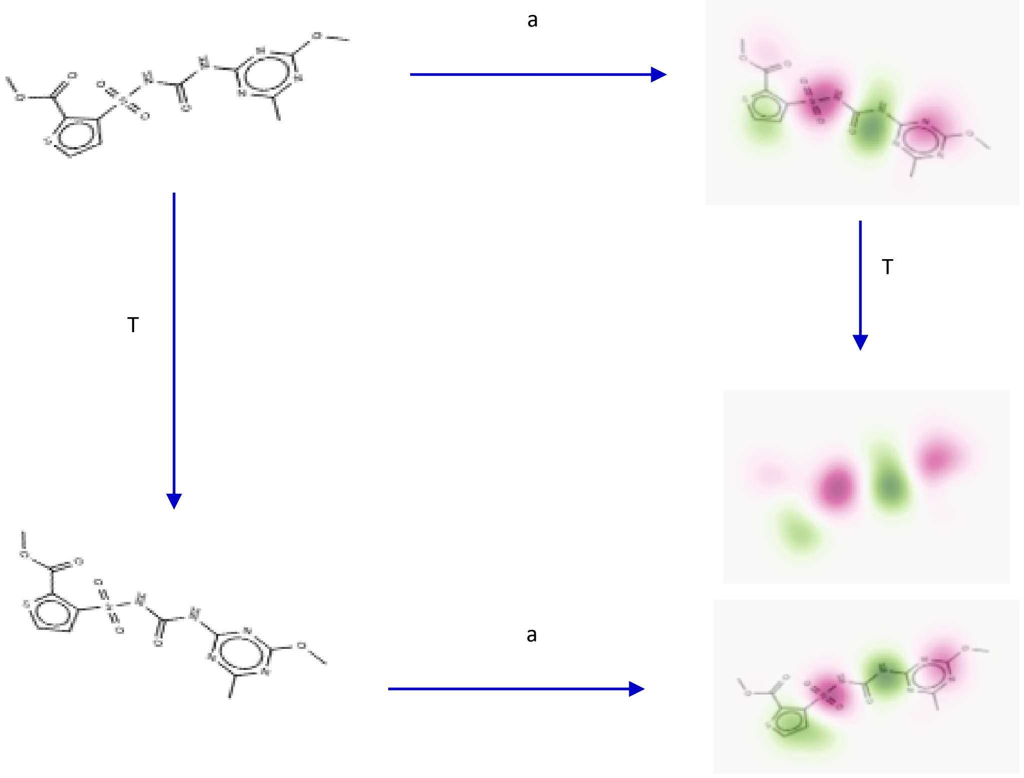
- Symmetry score for transformation T

$$s_T(\mathbf{x}) = \frac{1}{2}\overline{|\hat{a}(T(\mathbf{x})) - T(\hat{a}(\mathbf{x}))|}$$

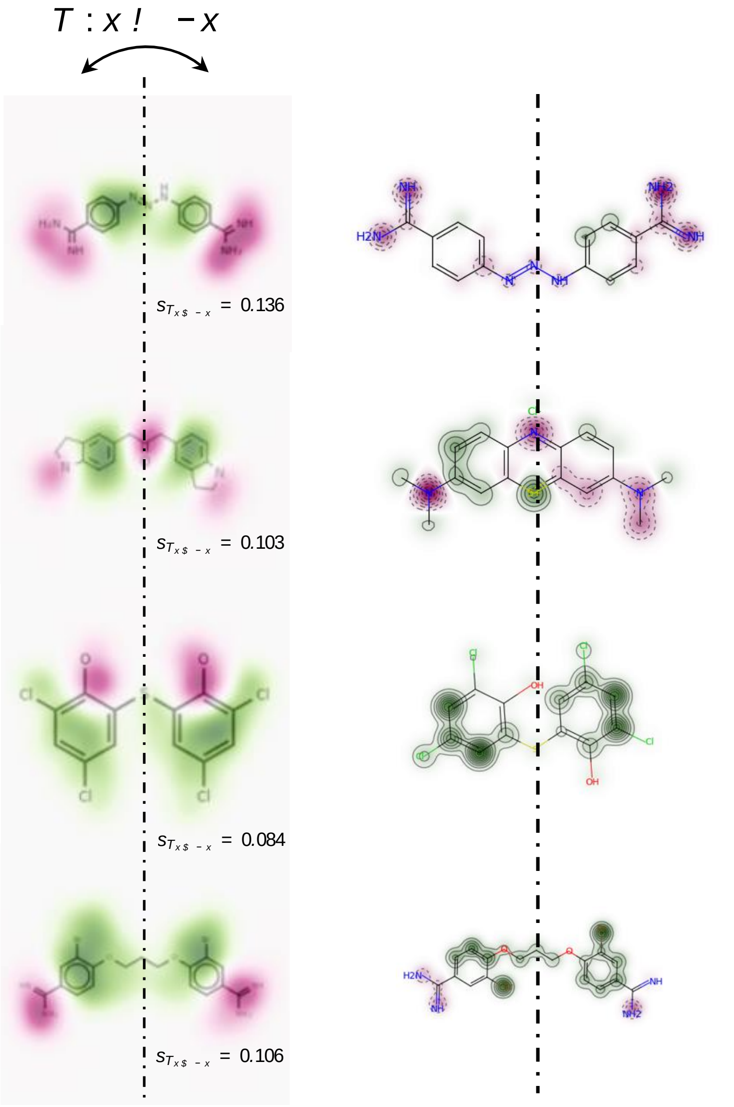
- A is normalized between [-1,1]

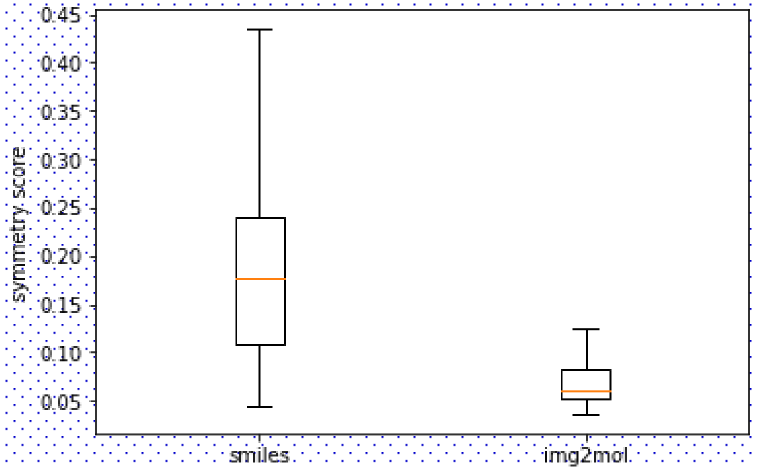
- $s_T(\mathbf{x}) = 0 \quad \Longleftrightarrow \quad \hat{a}T = T\hat{a}$

Example: Rotations



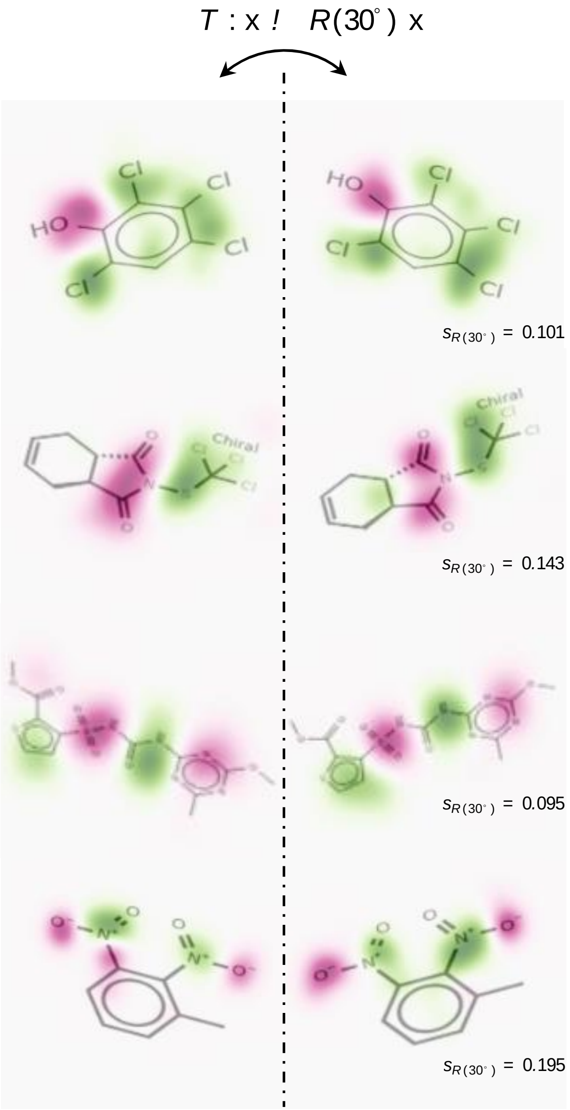
Molecule symmetries: reflection

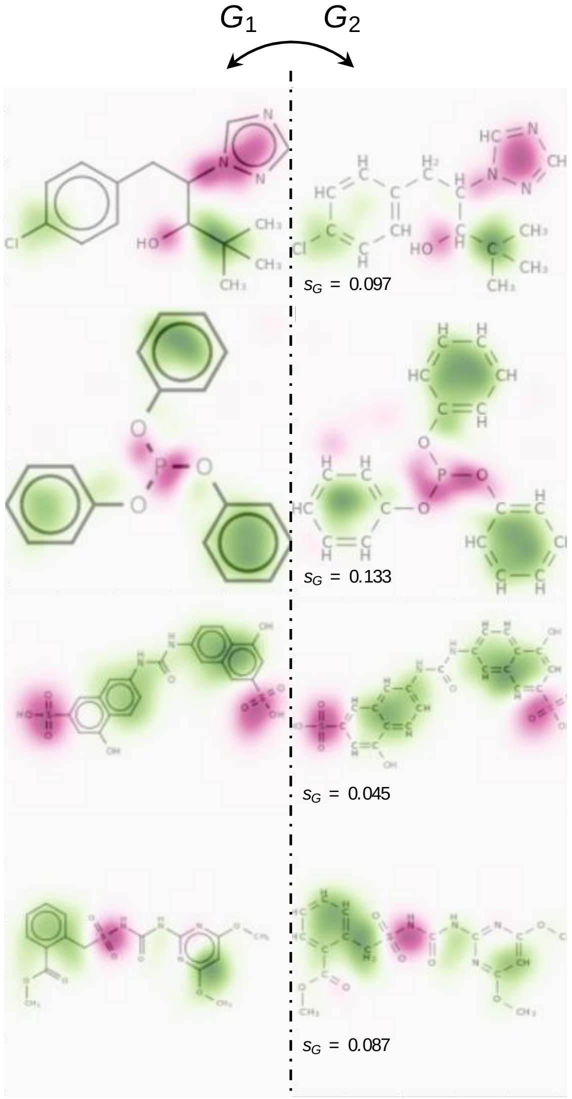




- The symmetry is well captured by our explanations
- In average, our contextual explanations respect the symmetry to a higher degree than the smiles-based explanations

Depiction’s symmetries


$T : x\ !\ \ R(30^{\circ})\ x$
$s_{R(30^{\circ})} = 0.101$
$s_{R(30^{\circ})} = 0.143$
$s_{R(30^{\circ})} = 0.095$
$s_{R(30^{\circ})} = 0.195$

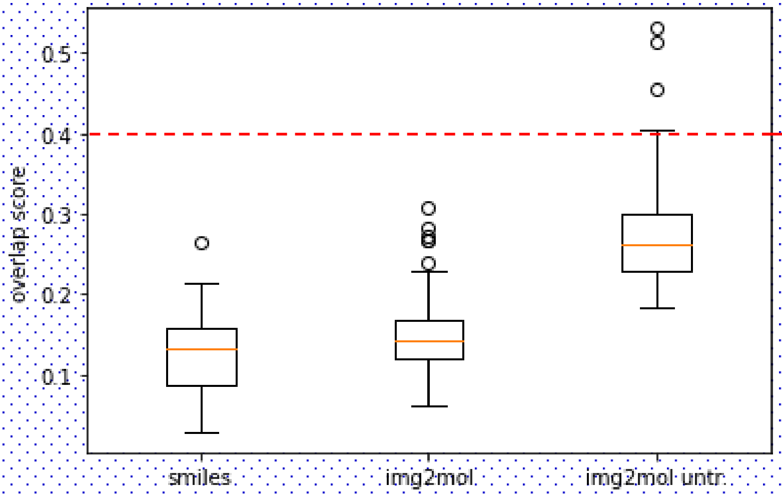


Ground truth – benzene task

- We train a simple downstream model based on CDDD to recognize whether a molecule containes benzine rings
- We compare the attributions with the ground truth
- We define an overlap score as follows

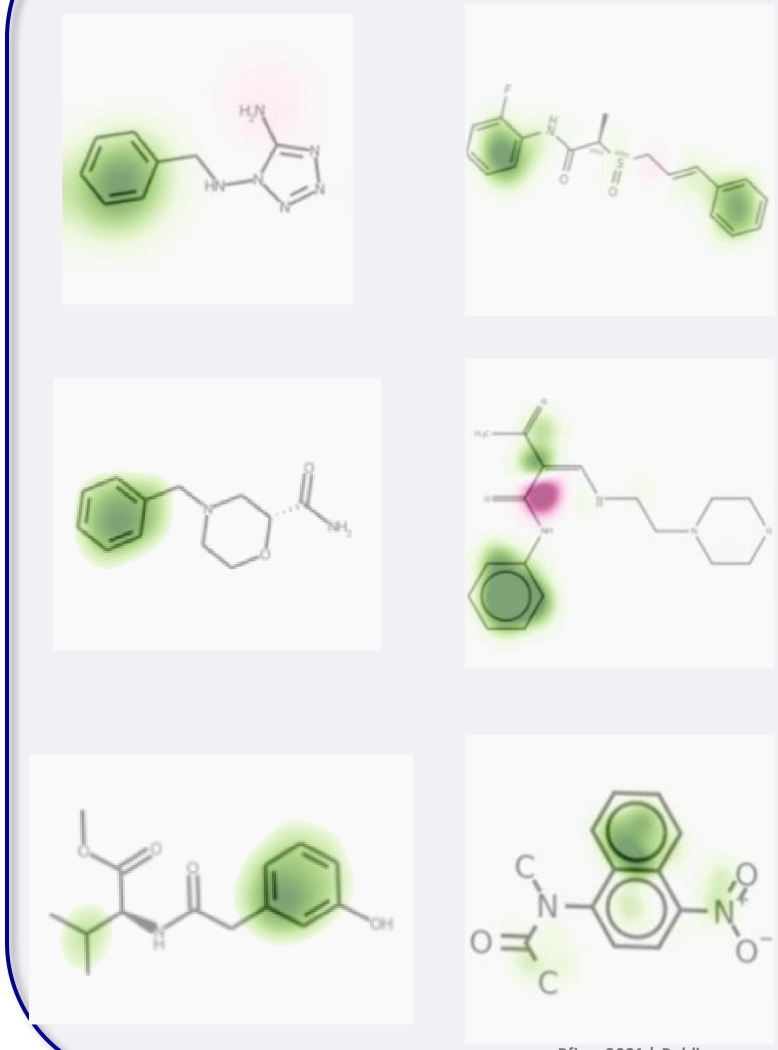
$$S_O(\mathbf{x}) = \frac{1}{2}\overline{|\widehat{a}(\mathbf{x}) - g(\mathbf{x})|}$$

Ground truth



Random maps

Examples

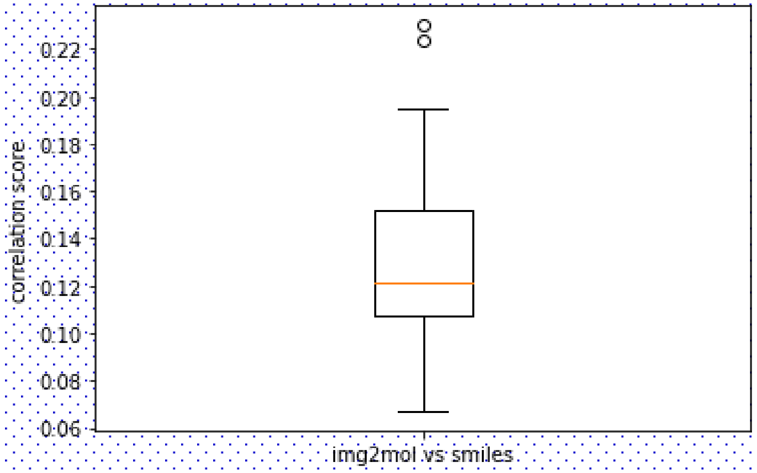


Correlation with SMILES explanations

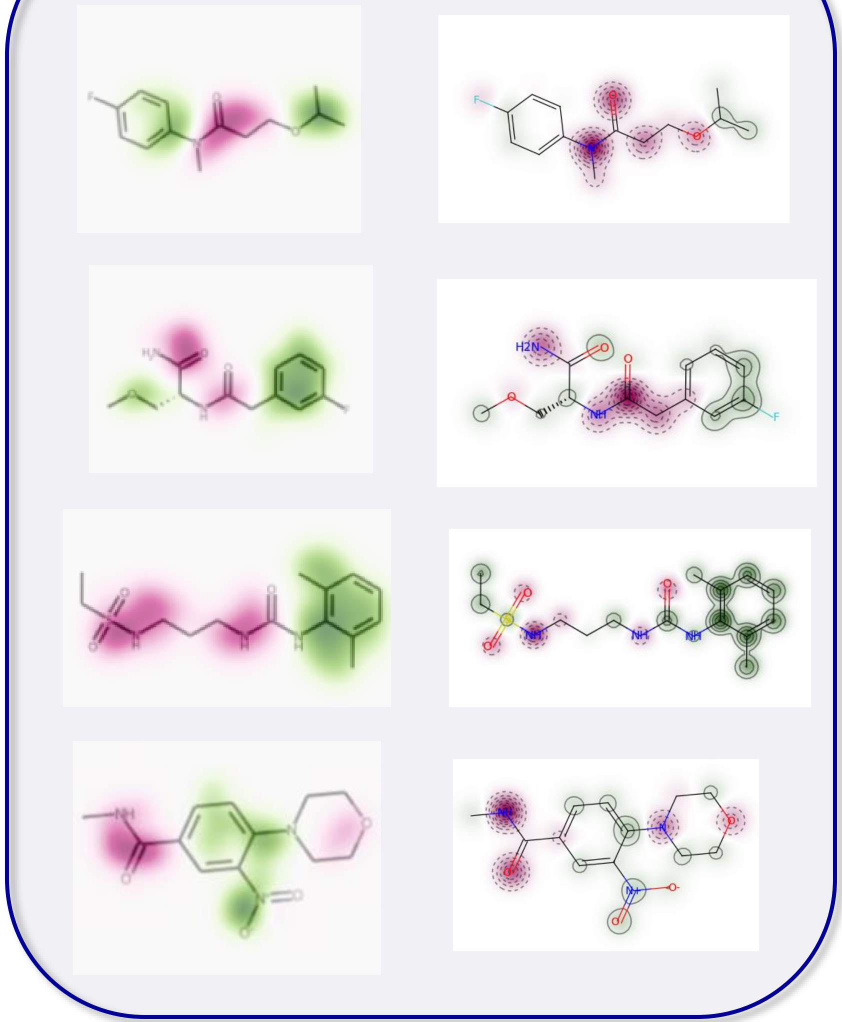
- Both CDDD-based explanation should correlate
- Quantifying the correlation helps establishing the robustness of the explanations
- We define an overlap score as follows

$$S_O(\mathbf{x}) = \frac{1}{2}\overline{|\widehat{a}(\mathbf{x}) - g(\mathbf{x})|}$$

SMILES explanations



Examples



Thank you for your attention!