

# Explainable AI in Chemistry

**Dr. Geemi Wellawatte**  
**PI: Prof. Philippe Schwaller**  
**EPFL**

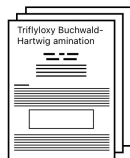
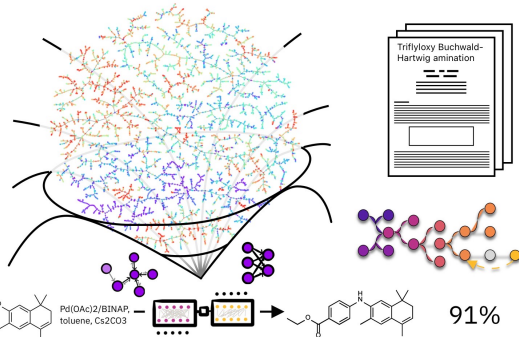


■ **Fact:** AI is advancing the boundaries of scientific research

- Property prediction
- Reaction prediction
- De novo generation
- Molecule optimization

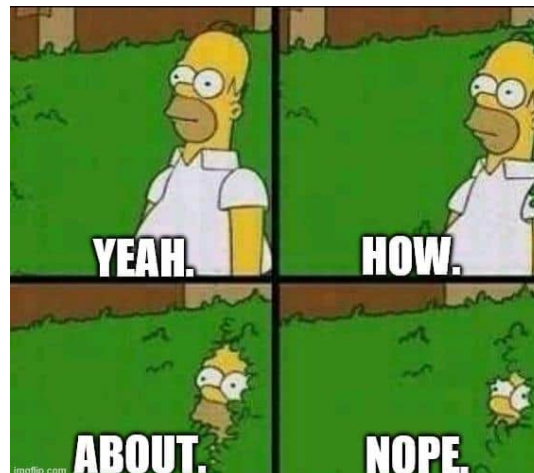
US20030166932A1: General Procedure  
 H A solution of trifluoroacetic acid 3,3,5,5-tetramethyl-1,3-dihydroisophthalen-2-yl ester (Compound 25, 0.41 g, 1.2 mmol), Pd(OAc)<sub>2</sub> (0.027 g, 0.12 mmol), BINAP (0.11 g, 0.18 mmol), Cs<sub>2</sub>CO<sub>3</sub> (0.56 g, 1.72 mmol), ethyl 4-aminobenzoate (0.25 g, 1.5 mmol) and 5 ml of toluene was flushed with argon for 10 min, then stirred at 100°C, in a sealed tube for 48 h. ...

$$\hat{H}\Psi = E\Psi$$



(<https://schwallergroup.github.io/>)

■ **Fact:** Chemists have a tendency to avoid deep learning! Black-box nature is off putting!

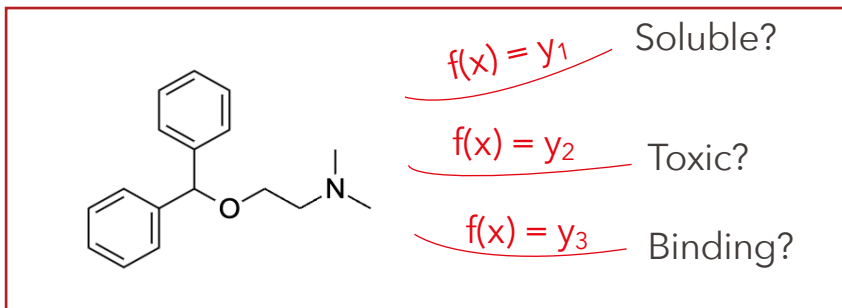


imgflip.com

# Explainable Artificial Intelligence (AI)

# Black-box nature of AI

- Chemists aren't entirely wrong! Neural-networks are not truly transparent.
- A NN is a non-linear function of a linear model. Weights and biases give little insight.



**XAI:** a “hot topic” that explains **WHY** a particular prediction is made. Help to identify data bias and model fairness

## Justifications

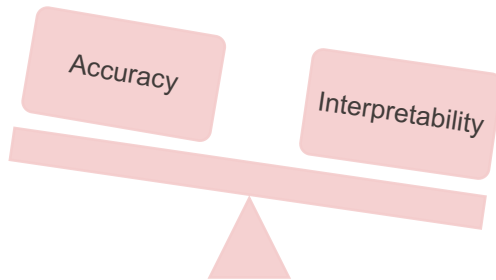
- Quantitative justifications why the model should be trusted.
- “Accuracy” metrics: RMSE, F1 score etc.

## Interpretability

- Human understandability.
- Knowledge that provide insight to a particular problem

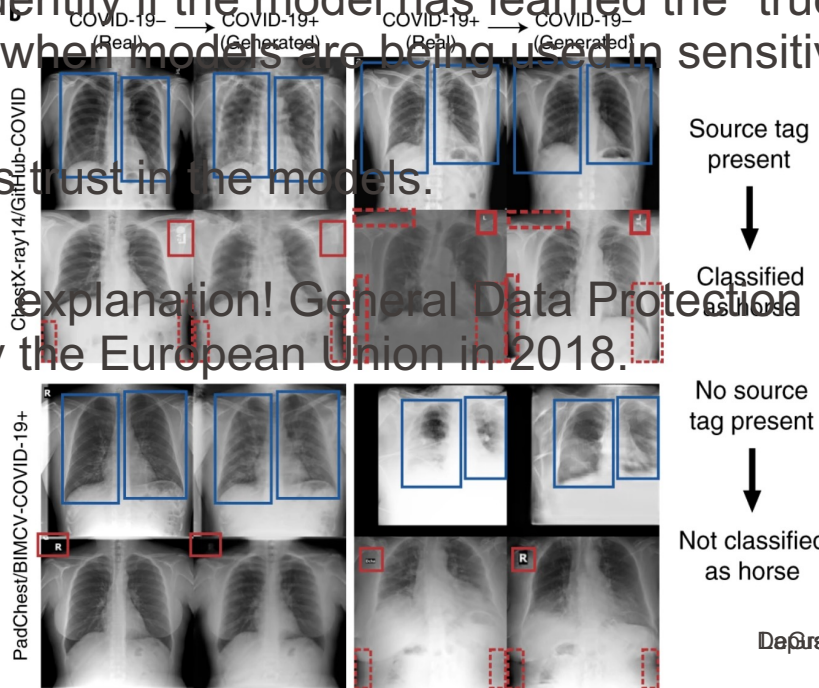
## Explanations

- A description why a prediction was made.
- An active characteristic that clarify the internal decision-making process

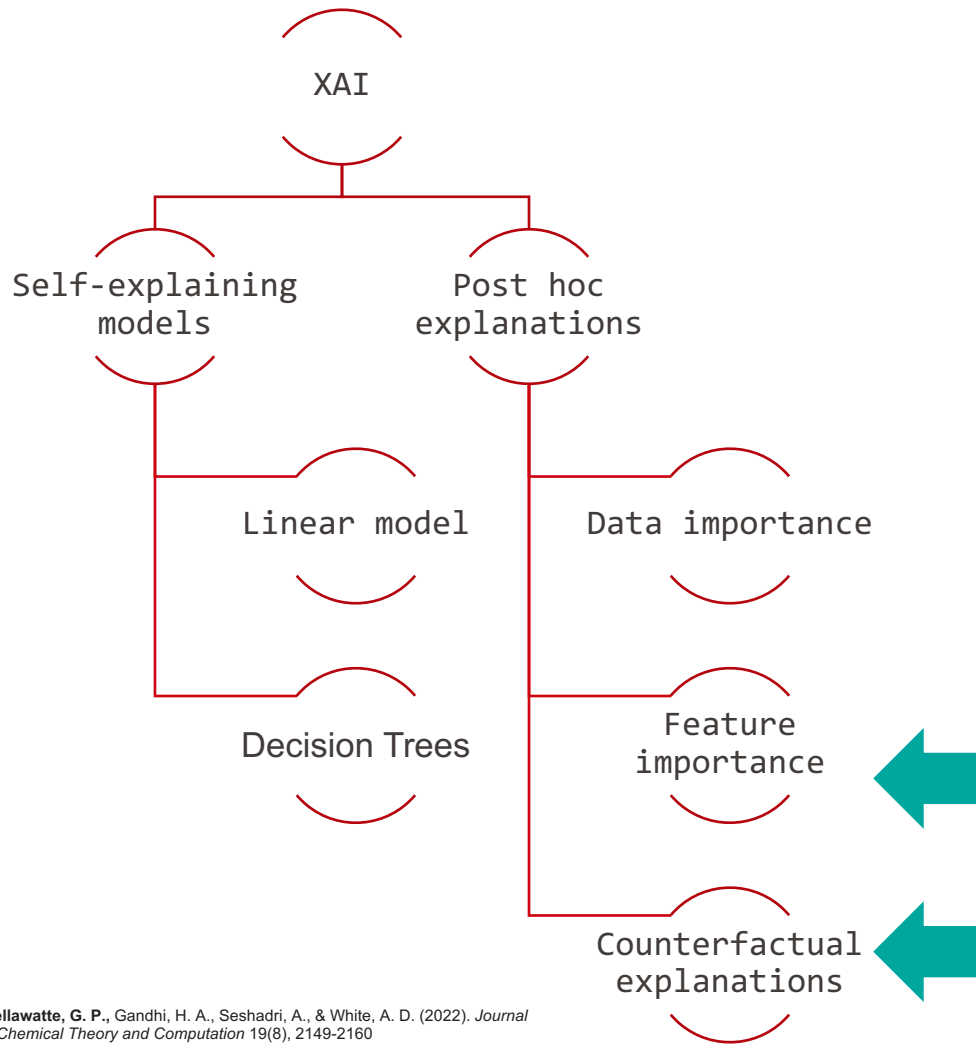


# Why XAI?

- Models can easily learn spurious relationships in data.
- Helps us identify if the model has learned the “true rationale”. Important when models are being used in sensitive applications..
- Establishes trust in the models.
- Right to an explanation! General Data Protection Regulation (GDPR) by the European Union in 2018.



De Gracia et al. *Nature* 2019



# Feature Attribution methods in Chemistry

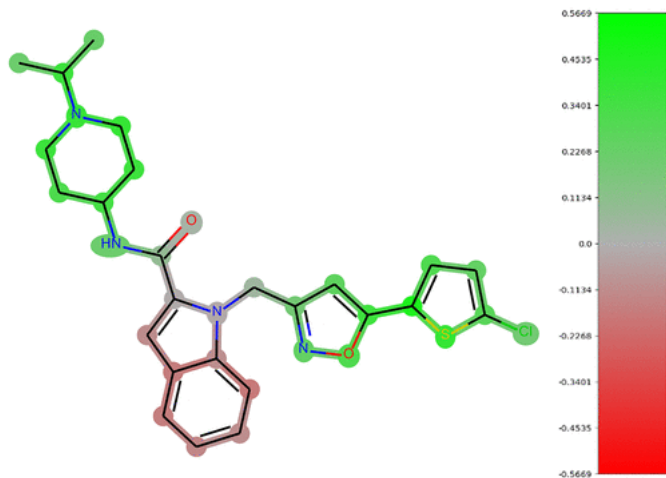


# Common XAI methods in chemistry

- Feature attribution methods assigns each input feature a numerical value.
- Either the input features are perturbed, and their effect on the network output is monitored, or a signal from the network output is back-propagated to the input
- Common approaches: feature based heatmaps, gradient based approaches (eg: CAM, Grad-CAM), Surrogate models, Shapley Additive explanations (SHAP)

# 1. Atom based heatmaps

- Transforms single atoms of an input molecule into dummy atoms and monitoring the change in predicted activity. Relevance w.r.t. to the original molecule.



1) Harren, Tobias, et al. "Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence." *Journal of Chemical Information and Modeling* 62.3 (2022): 447-462.

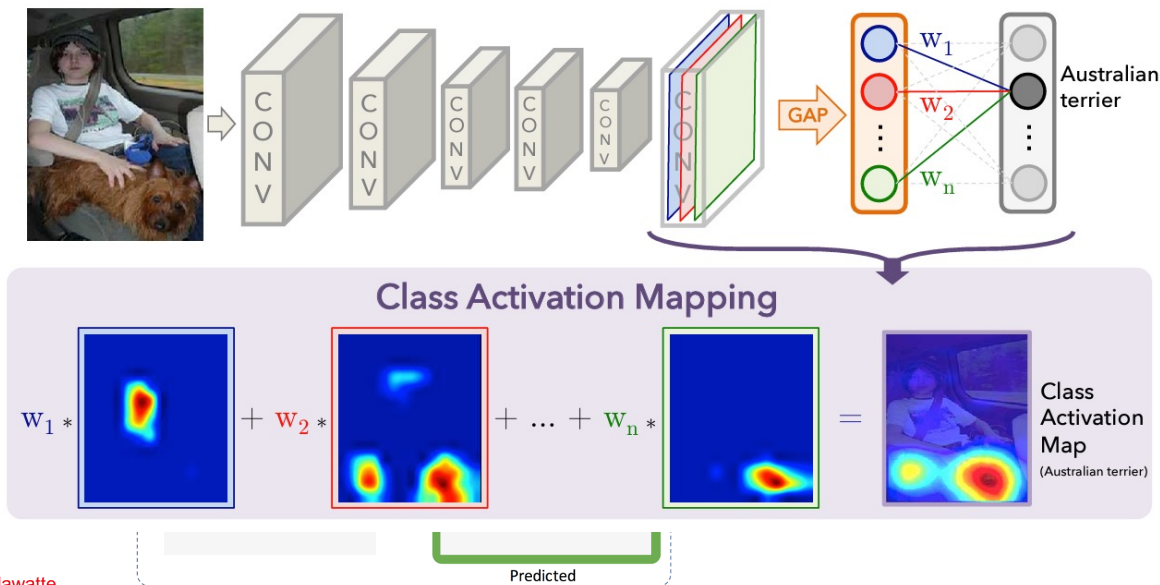
2) Sheridan, Robert P. "Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it?." *Journal of chemical information and modeling* 59.4 (2019): 1324-1337.

3) Rasmussen, Maria H., Diana S. Christensen, and Jan H. Jensen. "Do machines dream of atoms? Crippen's logP as a quantitative molecular benchmark for explainable AI heatmaps." *SciPost Chemistry* 2.1 (2023): 002.

## 2. Gradient based approaches

$$\frac{\Delta f(\hat{x})}{\Delta x_i} \approx \frac{\partial f(\hat{x})}{\partial x_i}$$

- Direct computation of attribution ( $\nabla_x f(x)$ ) suffers from shattered gradient problem. Instead, the gradient can be approximated with different approaches. NOT model agnostic. Eg: CAM, Grad-CAM, integrated gradients.



"Discovering molecular functional groups  
lutional neural networks." arXiv preprint  
(2018)

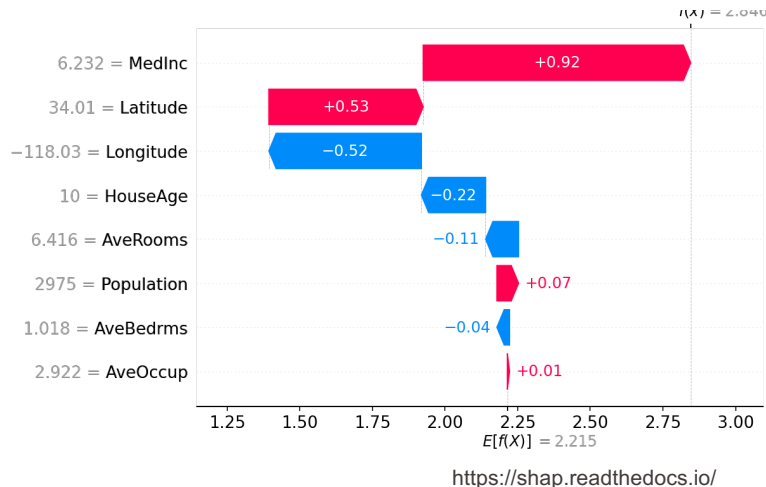
Zhou, Bolei, et al. "Learning deep  
features for discriminative  
localization." *Proceedings of the  
IEEE conference on computer  
vision and pattern recognition.*  
2016.

"1D Gradient-Weighted Class Activation Mapping,  
Process of Convolutional Neural Network-Based  
opy Analysis." *Analytical Chemistry* (2023).

# 3. SHAP values

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

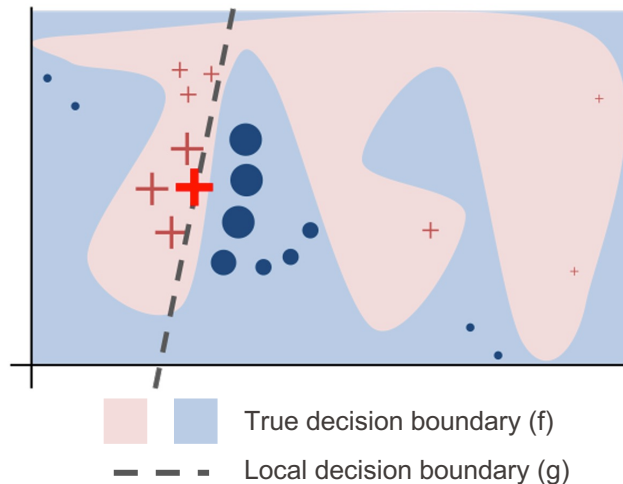


- Most famous XAI approach!
- SHAP values are calculated by comparing a model's predictions with and without a particular feature present. This is done iteratively for each feature and each sample in the dataset.
- Sum of all SHAP values = combined effect of all fts.
- Symmetry (fts with similar contributions have equal weights)
- Additivity of SHAP values show combined contribution

## 4) LIME

Ribeiro, Marco Tullio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



- A local explanation method.
- Learns an interpretable model around a prediction (linear models, decision trees) by creating perturbations around the instance.
- Another commonly used explanation method:
  - Whitmore, Leanne S., Anthe George, and Corey M. Hudson. "Mapping chemical performance on molecular structures using locally interpretable explanations." arXiv preprint arXiv:1611.07443 (2016).
  - Mehdi, Shams, and Pratyush Tiwary. "Thermodynamics of interpretation." arXiv preprint arXiv:2206.13475 (2022).

# Counterfactual Explanations for Molecular Models

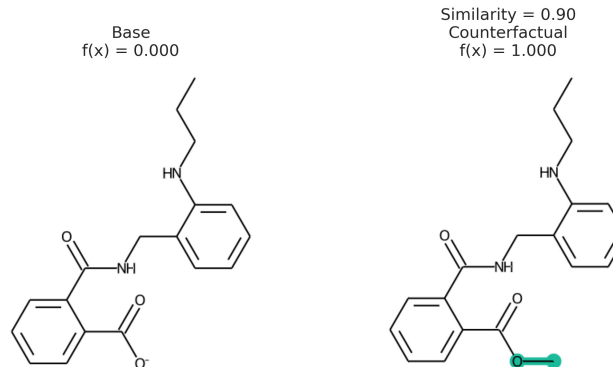
- An example closer to the original but with a different outcome

Eg: “Your paper would be better cited if you had a shorter title”

minimize  $d(x, x')$

such that  $\hat{f}(x) \neq \hat{f}(x')$

- Counterfactual explanations can capture causal and non-causal relations.
- CFs provide **local explanations** that are intuitive to understand in XAI.



Shows that the carboxylic acid group can explain for lack of activity

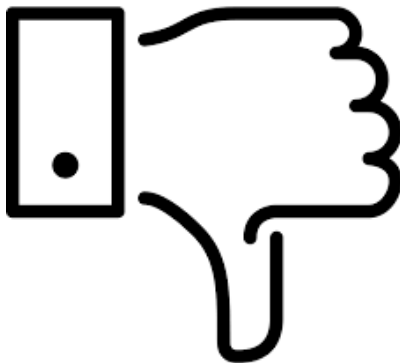
Wellawatte, G. P., Seshadri, A., & White, A. D. (2022). *Chemical science*, 13(13), 3697-3705.

# CF generation is challenging

Highly dependent  
on the model  
architecture

Includes a  
generative model  
and a prediction  
model

May require new  
data, re-training  
of models

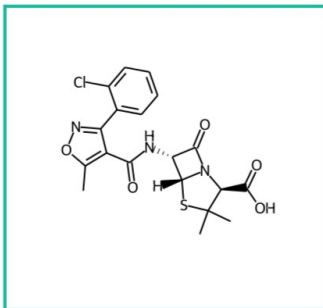




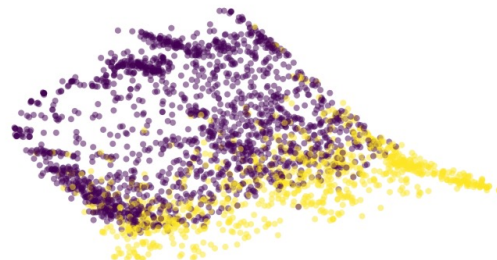
# EPFL Molecular Model Agnostic Counterfactual Explanations

- MMACE algorithm requires no gradients, training, or additional data to create explanations.
- Independent of the model architecture used for classification and regression tasks.
- MMACE looks at changes to molecules which affect activity. We use a generative algorithm.
- Built on [STONED-SELFIES](#) method (Nigam et al., 2021) to sample a local chemical space given a starting molecule. Surjective property of representation (Krenn et al., 2020).

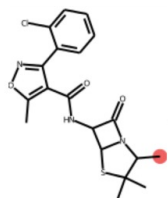
1. Molecule being predicted: base



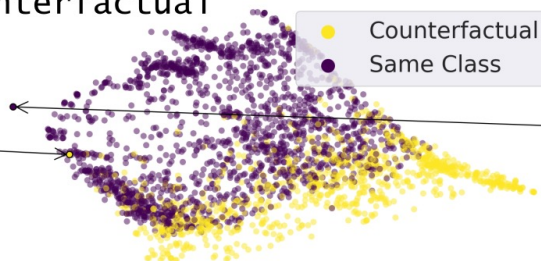
2. Expand chemical space around base



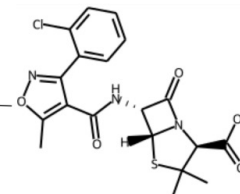
3. Identify most similar molecule with changed label: counterfactual



Counterfactual



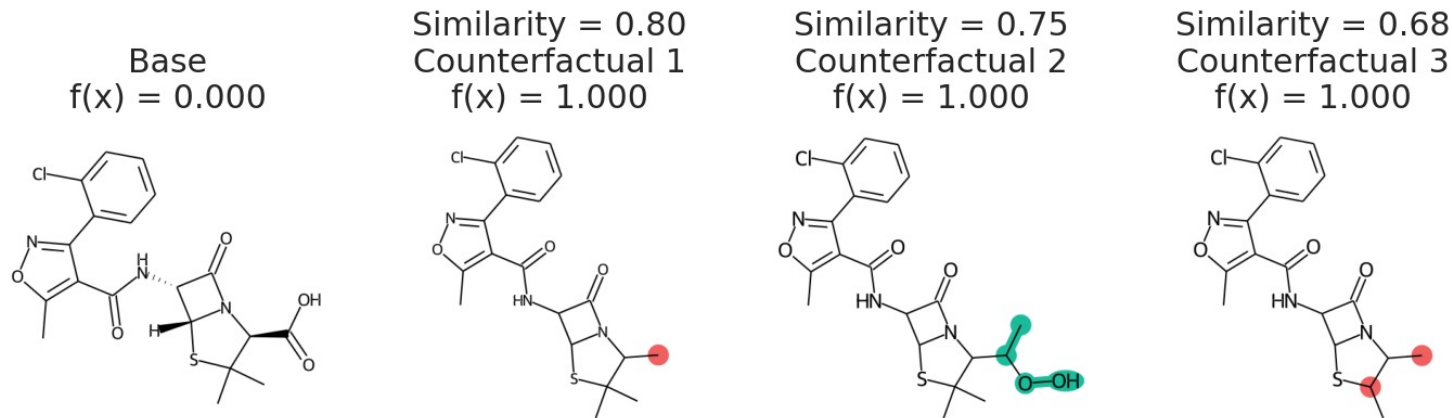
Base



# Application: RF model for blood-brain barrier permeation prediction

- BBB permeation is a thoroughly studied question in drug discovery.
- Perceived as a binary classification task with a random forest model implemented with Scikit-Learn.
- ROC-AUC of our model was computed as 0.91 - comparable to 0.95-0.98 benchmark models in literature.

**GOAL:** Has my trained model learned chemistry? How can I use CFs to explain the model behavior?

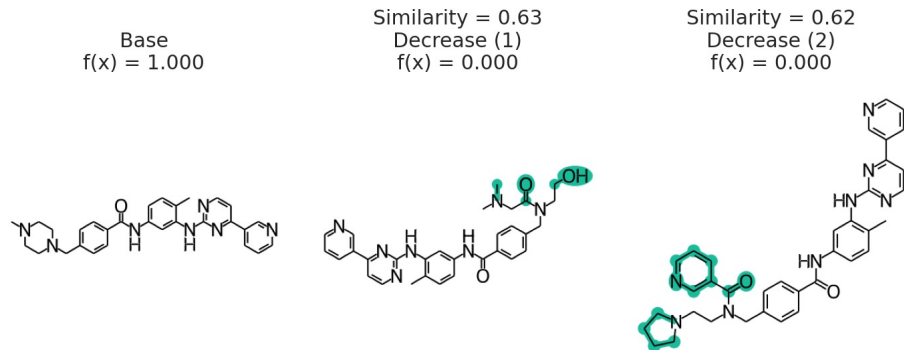


**Explanation:** The negative example can be made to cross the blood brain barrier if the carboxylic group is altered.

**Experimental observations:** hydrophobic interactions and surface area govern BBB permeation (Boobier S, *et al.*, *Nat Commun.* 2020)

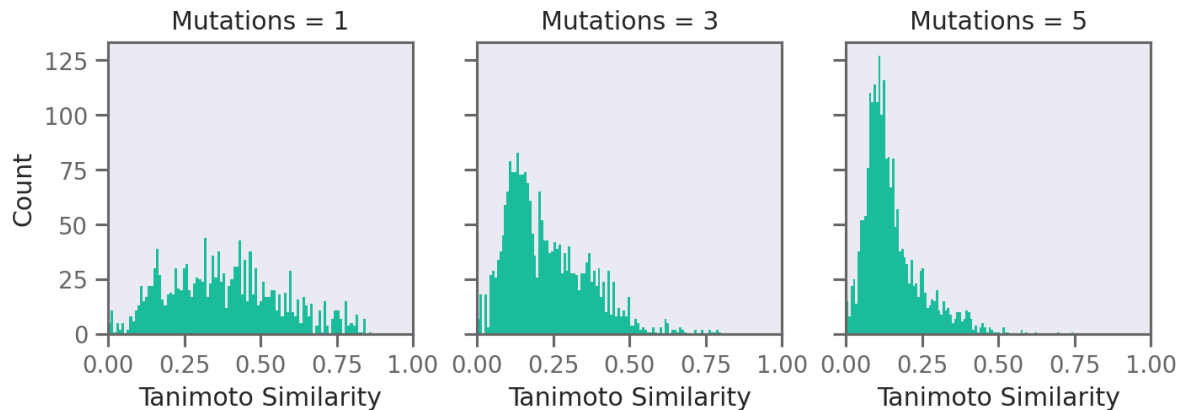
# Validity VS Stability of generated molecules

- STONED algorithm generates valid molecules, but experimental stability is not guaranteed.
- Chemed method: query PubChem database.
- Or users can query their own databases too.



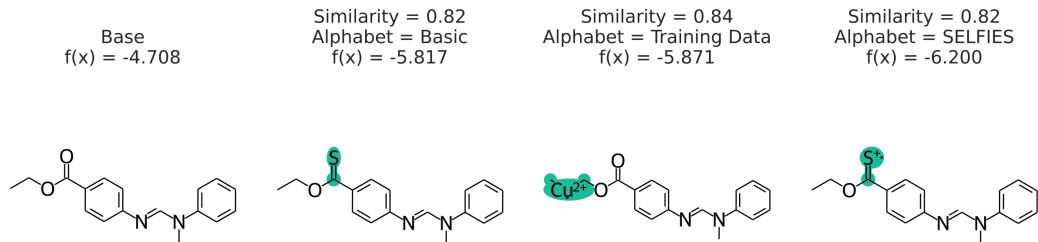
**Explanation: tertiary amine of the pyridine plays a vital role in blood-brain barrier permeation**

# Impact of MMACE parameters



**Similarity in the chemical space decreases with the number of allowed mutations.**

**The alphabet can be used to control the generated chemical space.**



# MMACE algorithm is open source for use

exmol 1.0.0

```
pip install exmol
```



<https://github.com/ur-whitelab/exmol>



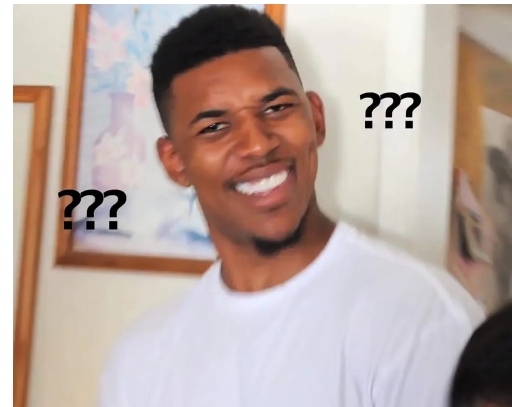
- Counterfactual explanations are intuitive and actionable – help to uncover rationale behind predictions in molecular models.
- MMACE is an easy to implement, computationally inexpensive, model independent algorithm.

# XAI beyond model interpretability



# But...

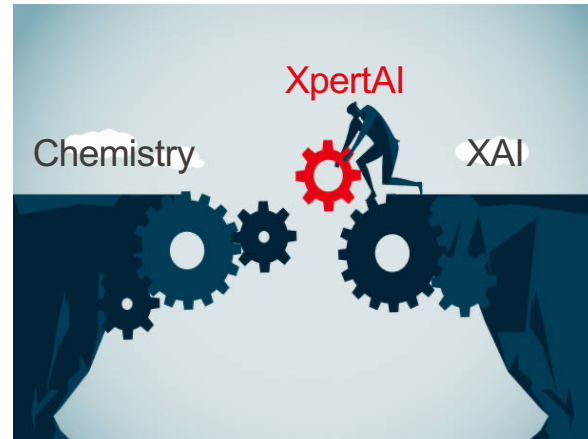
- The explanations are not inherently interpretable?
- XAI methods are developed for technical users. Eg: computer vision.
- Requires special attention when using in chemistry.



- **Can we make XAI interpretable?**
- **Are there other purposes of XAI?**

# Interpretability & natural language

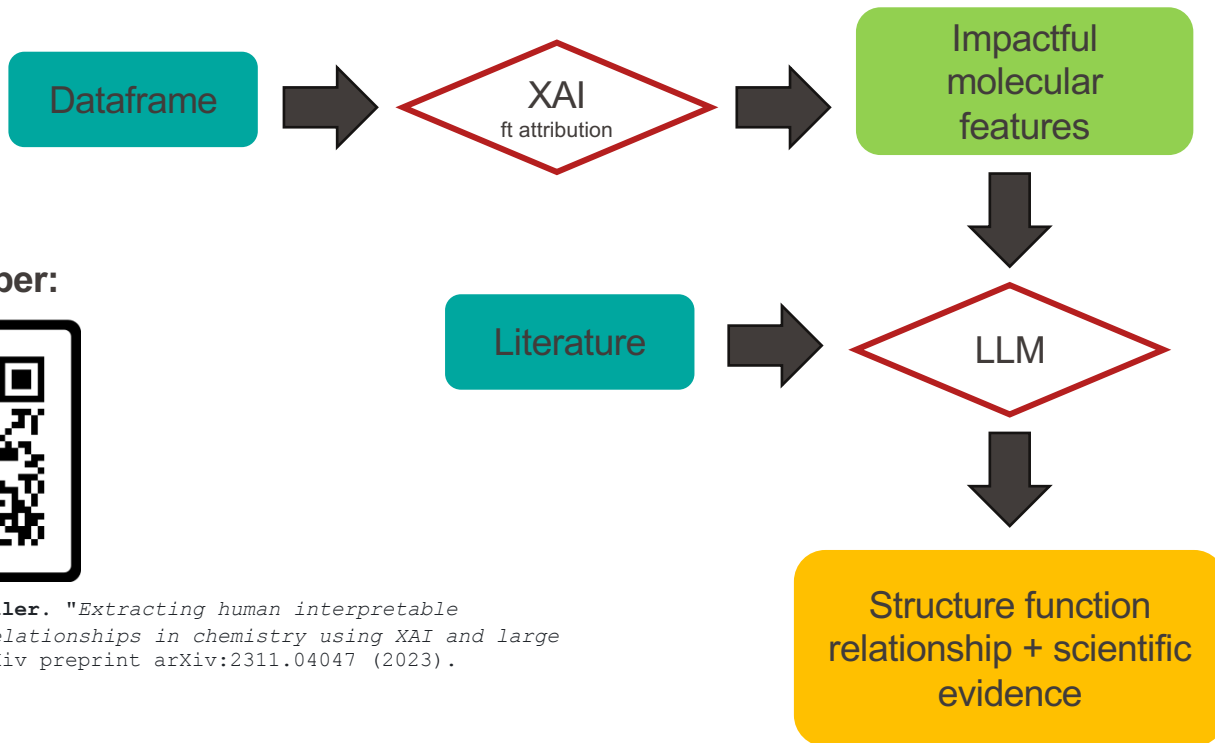
- Natural language is the default when it comes to interpretability.
- LLMs are generative models which can predict an output sequence given an input sequence.
- LLMs in isolation can be limited.
- LLMs + XAI is a powerful combination!



✦ ✦ ✦ XpertAI ✦ ✦ ✦

A tool that combines XAI with LLMs to generate  
**intelligent** explanations!

# From raw data to human interpretable structure-property relationships



Link to paper:



Wellawatte and Schwaller. "Extracting human interpretable structure-property relationships in chemistry using XAI and large language models." arXiv preprint arXiv:2311.04047 (2023).

# Why XpertAI?

	Interpretable	Targeted explanations	Literature evidence	Accessible to non-technical users
XAI	✓	✓	x	x
LLMs	✓	x	x	✓
LLMs + Literature	✓	x	✓	✓
XpertAI	✓	✓	✓	✓

- ✓ Establishes connections between black-box models, XAI tools, and literature.
- ✓ Delivers precise natural language explanations (NLEs) tailored to specific datasets.
- ✓ Identifies crucial features within the dataset and draws on scientific evidence to articulate structure-property relationships.



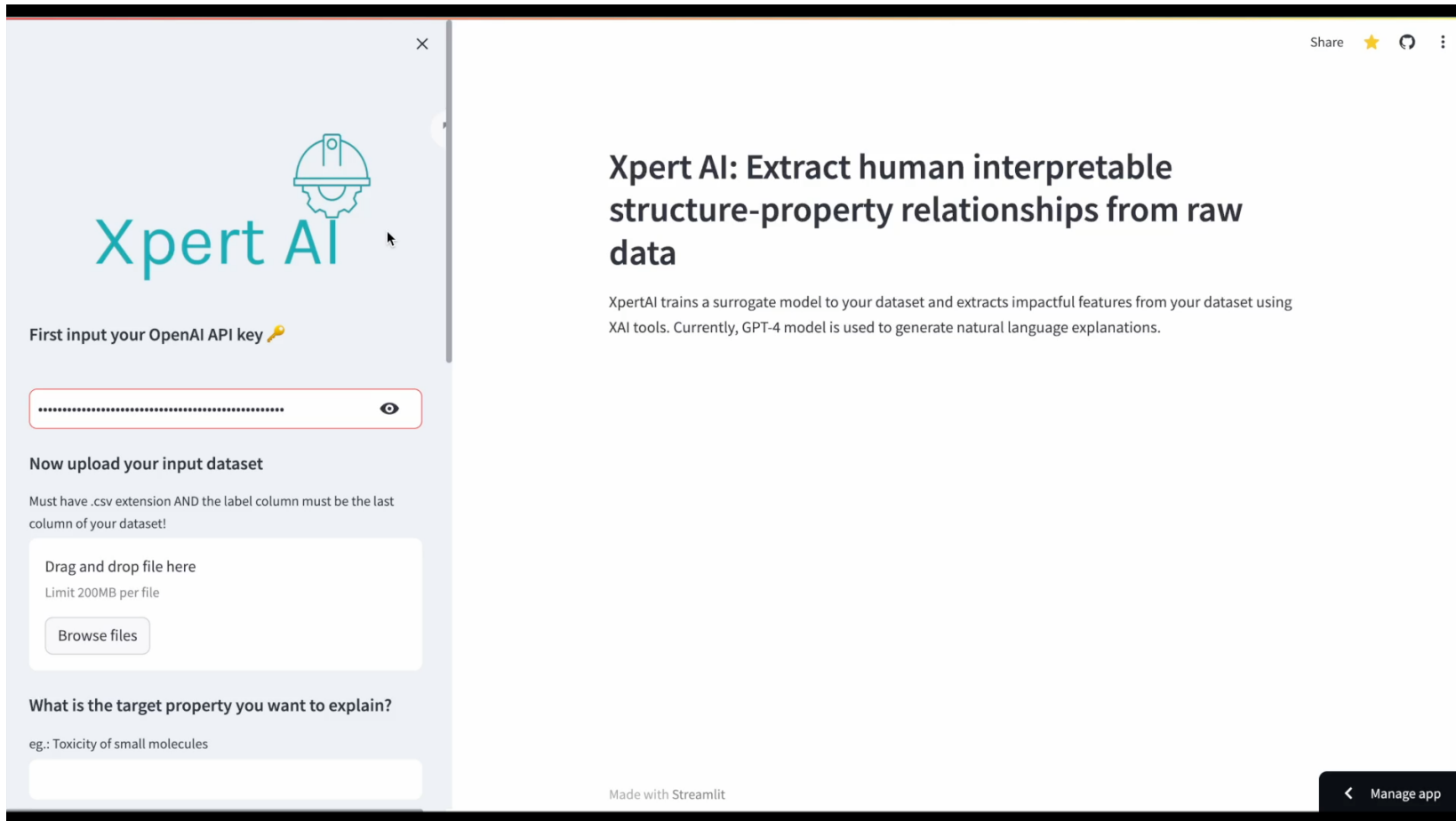
- We conducted 5 case studies
  1. Presence of open metal sites in metal-organic frameworks (classification)
  2. Pore limiting diameter in metal-organic frameworks (regression)
  3. Toxicity of small molecules (classification)
  4. Solubility of small molecules (regression)
  5. Flammability of organic molecules (regression) – **negative example**



**Xpert AI: Extract human interpretable structure-property relationships from raw data**


## Case study 1: Presence of open metal sites in MOFs

- The features identified by the XAI analysis that affect the presence of open metal sites in MOFs are the **Fraction of Metals, Solid Density, and Average Cationic Radius**. Additional features that may be correlated with the presence of open metal sites in MOFs include the **type of metal ions used, the coordination environment, and the presence of structural defects** (Kökçam-Demir et al., 2020; Jianwei Ren et al., 2017).
- The Fraction of Metals in a MOF can influence the presence of open metal sites. A higher fraction of metals may lead to more open metal sites, as these sites are typically part of the metal node or metal secondary building unit (Kökçam-Demir et al., 2020).
- The Solid Density of a MOF can also impact the presence of open metal sites. A lower crystal density is associated with a higher hydrogen-storage capacity, which suggests a correlation with the presence of open metal sites (Qingyuan Yang, Chongli Zhong, 2006).
- The Average Cationic Radius can affect the strength of interaction between the metal ions and hydrogen, which can influence the presence and effectiveness of open metal sites (Wei Zhou, Hui Wu, Taner Yildirim, 2008).
- In summary, the presence of open metal sites in MOFs is influenced by a variety of factors, including the Fraction of Metals, Solid Density, and Average Cationic Radius, as well as the type of metal ions used, the coordination environment, and the presence of structural defects. By manipulating these features, it may be possible to alter the presence of open metal sites in MOFs and thereby optimize their performance for various applications.
- **References:**
  - Kökçam-Demir, Anna Goldman, Leili Esrafilı, Maniya Gharib, Ali Morsali, Oliver Weingart, Christoph Janiak, (2020). Coordinatively unsaturated metal sites (open metal sites) in metal-organic frameworks: design and applications.
  - ...



Share ★ ↻ ⋮

## Xpert AI

First input your OpenAI API key 

Now upload your input dataset

Must have .csv extension AND the label column must be the last column of your dataset!

Drag and drop file here  
Limit 200MB per file

Browse files

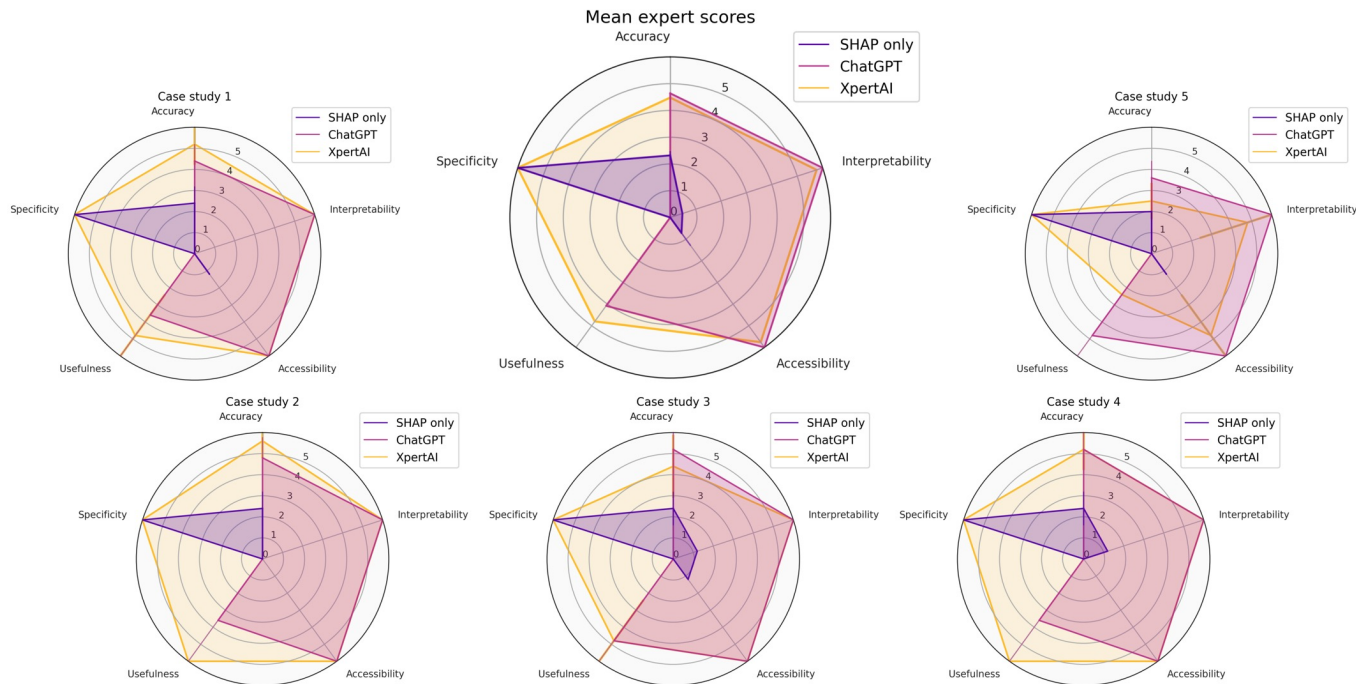
What is the target property you want to explain?  
eg.: Toxicity of small molecules

Made with Streamlit

Manage app

# Evaluations (only one result is shown here)

- 5 human experts were asked to evaluate explanations from **XpertAI**, **ChatGPT**, and **SHAP** plots for each case study based on *accuracy*, *interpretability*, *accessibility*, *usefulness* in research, and *specificity* to given data.





# Integrating Open-LLMs in XpertAI

LLM	Size	Accurately describes each feature and how it is related to the target	Accurately describes how the target can be altered w.r.t each feature	Lists and explains additional features	Accuracy of generated references	Average RougeL score
Llama2 <sup>[77]</sup>	3.8 GB	0	1	2	0.3	0.52±0.05
mixtral:8x7b-instruct-v0.1-q5_0 <sup>[78]</sup>	32 GB	4	5	5	1.25	0.49±0.04
Phi:2.7b <sup>[79]</sup>	1.6 GB	1	0	3	0	0.38±0.06
starling-lm:7b-alpha <sup>[80]</sup>	4.1 GB	5	2	4	1.6	0.46±0.02
GPT-4 <sup>[31]</sup> (default in XpertAI)	N/A	5	5	5	5	0.64±0.05

# Wrapping up

- XAI helps us uncover rationale for model predictions.
- XAI must be a commonplace practice, specially in cases where AI is used for decision making, sensitive application.
- Can be used to uncover input-output relationships.
- XpertAI = Raw data + XAI + LLM + Scientific Evidence =  
🌟🌟🌟🌟 Natural language structure-function relationships 🌟🌟🌟🌟
- Leveraging on advantages of XAI and LLMs to provide task specific, intelligent explanations. Reduces hallucinations!
- We can be optimistic about using open-source LLMs in place of proprietary models.

# Questions?



Funding Acknowledgement  
**SusEcoCCUS**

- We also asked **Claude AI** assistant to compare and score two explanations from XpertAI and ChatGPT.
- Claude rates the explanations from XpertAI higher than ChatGPT's for 4/5 tasks.

"Explanation A directly discusses the specific features identified by the XAI analysis and provides concrete examples of how changing those features affect the target property, indicating high relevance for research. Explanation B provides a more general background on how molecular structure influences the target properties. While still relevant, it does not directly address the specific features called out in the XAI analysis."

- However, Claude rates XpertAI's explanation in case study 5 less than ChatGPT's. A similar observation was made from the expert evaluations.

# Variation among generated explanations computed with RougeL metric

