Model agnostic generation of counterfactual explanations for molecules†

2022-09-02 / JiWung Han

Department of Artificial Intelligence
Korea University

Abstract

◆ **Abstract**

- An outstanding challenge in deep learning in chemistry is its lack of interpretability.
- Leading to neural networks learning spurious correlations that are difficult to notice
- Counterfactuals are a category of explanations that provide a rationale behind a model prediction with satisfying properties like **providing chemical structure insights**.
- Explain any black-box model prediction
- Random forest models, sequence models, and graph neural networks in both classification and regression

Introduction

- **Deep learning**
  - Significant impacts in chemistry because of its ability to regress non-linear relationships between structure and function
  - Despite their empirical accuracy, neural networks are black-box models; they lack interpretability and predictions come without explanation
  - A neural network is a non-linear function and linear model

$$\hat{y} = \vec{w}g(\vec{x}) + b$$

  - Weights and biases give little insights
  - XAI : Explain **WHY** a particular prediction is made

Introduction

◆ **Explainable artificial intelligence**

- Explainable artificial intelligence (XAI) is an emerging field which aims to provide explanations, interpretation, and justification for model predictions

1. **Explanation** – Why a prediction made by a model
2. Interpretability – The degree to which an observer can understand the cause of a decision
   (어떤 모델로 인해서 이러한 결과가 나왔는지)
3. Justification – A description of why a prediction should prediction should be believed
   – Typically relies on estimated model generalization **error**
   (보통 Loss value 로 표현 함)

- Explanation is rare, especially in deep learning where no insight can be gained by inspecting model weights or parameters.

Approaches

◆ **Four major approaches for explaining a prediction from a black-box model**

1. Identifying which features contribute the most
2. Identifying which training data contributes the most
3. Fitting a locally interpretable model around the prediction
4. Providing contrastive or counterfactual points

- Feature importance analysis provides per-feature weights that identify how each feature contributed to the final prediction
  - Can be formulated as SHAP values, a method of computed feature importance weights as a complete explanation $(i.e., \sum w_i = f(x))$

Approaches

◆ **Four major approaches for explaining a prediction from a black-box model**

- Feature importance analysis provides per-feature weights that identify how each feature contributed to the final prediction
  - Can be formulated as SHAP values, a method of computed feature importance weights as a complete explanation $(i.e., \sum w_i = f(x))$ (해당 Feature 에 해당하는 weight 의 값)
  - Effective when working with a sparse set of molecular descriptors, but when working with thousands of descriptors, SMILES or molecular graphs

- Instead, **what changes will result in an alternate outcome?**

Conterfactual?

◆ **Counterfactual (CF) Explanations**

- Counterfactuals are a mature topic in philosophy and mathematics
- Woodward and Hitchcock define a **counterfactual explanation** as one that illustrates **what differences** to an event or instance would **generate a change in an outcome**

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

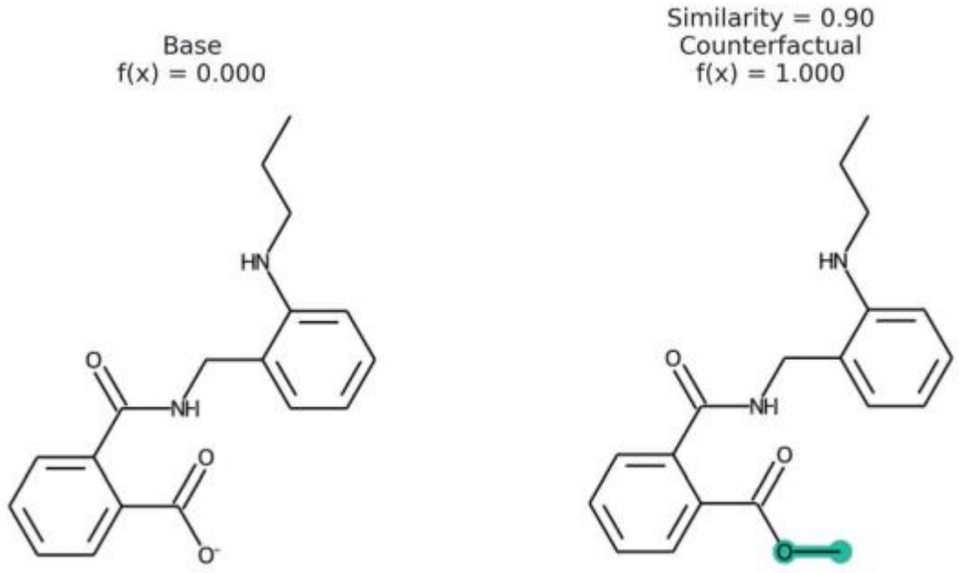
$$minimize\ d(x, x')$$

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

- Intuitive to understand in XAI

Conterfactual?

- **Counterfactual (CF) Explanations**



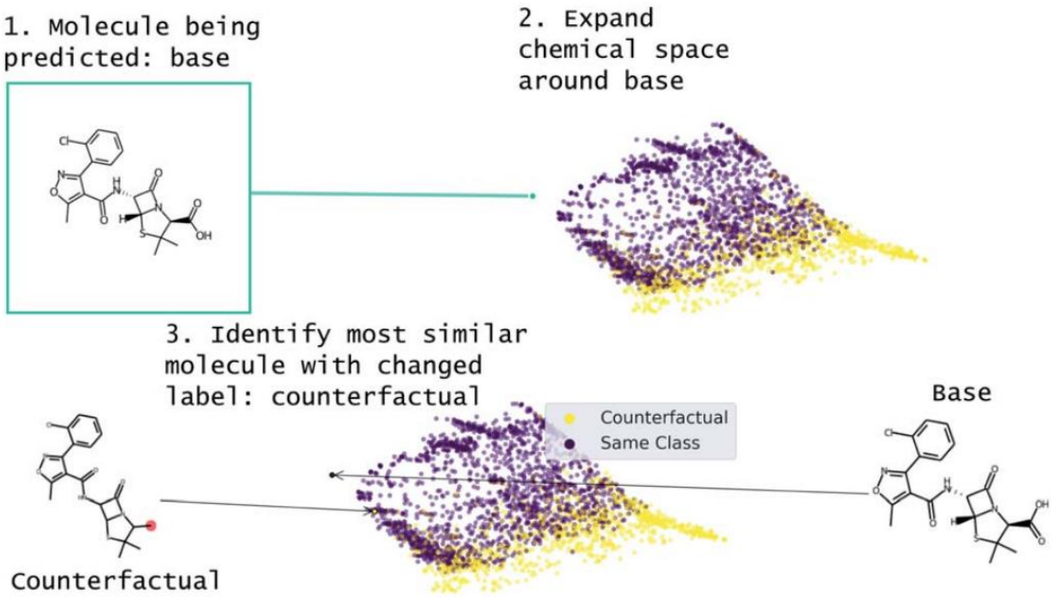
- The molecule on left was predicted to have class of 0, no activity
- With the modification shown in teal(Triethylaluminium), the molecule would be in class 1, active
- This shows that the carboxylic acid is an explanation for lack of activity

Generating CF

- **Generating Molecular counterfactuals**
  - Built on the **S**uperfast **T**raversal, **O**ptimization, **N**ovelty, **E**xploration and **D**iscovery (STONED) method which enables **rapid exploration of chemical space** without a pre-trained generative model or set of reaction rules
  - Expand chemical space around the molecule being predicted (base)
  - **Select a small number of these molecular** counterfactuals with **clustering/Tanimoto similarity**
  - This method works because we represent molecules as **SELF-referencIng Embedded Strings (SELFIES)** and any modification to a SELFIES is **also a valid molecule**

Generating CF

◆ **Generating Molecular counterfactuals**



- The input is a molecule to be predicted
- Chemical space is expanded and clustered
- Counterfactuals are selected from clusters to find succinct explanation of base molecule prediction

Comparison on existing work

◆ **Molecular Model Agnostic Counterfactual Explanations**

- GNNExplanier
  - Uses graph edge operations and a relaxed model prediction function to propose counterfactuals and was found to do well on graph datasets
  - **Graph edge operations cannot be used on molecular** structures
    → Because the majority of graph operations will **violate valencies** (원자가)
  - Also requires model gradients with respect to input, which may not be possible for models outside of neural networks
- MMACE
  - Works on descriptors, graphs, SMILES, and SELFIES features
  - Does not require gradients, enabling its use on machine learning methods like random forest classification or support vector machines

Comparison on existing work

- **Molecular Model Agnostic Counterfactual Explanations**
  - Reinforcement learning agent
    - Requires training process
    - Relied on perturbing x using graph transformation operators and reinforcement learning
    - Generate chemically infeasible structures
  - MMACE
    - Does not require training a counterfactual generator because all molecules resulting from STONED are valid compounds

Theory

- **Deep Learning in chemical domain**
  - For chemical applications, input x is typically a representation of a molecule
  - Can be a string (SMILES or SELFIES), a set of chemical descriptors, or a molecular graph
  - Compute chemical descriptors programs
    - **Mordred56**
    - **DRAGON57**
  - can be used to, such as electronegativity or molecular weight, for each molecule

Theory

- **Deep Learning in chemical domain**
  - Molecular graph can consist of a **node feature vector** and an **adjacency matrix**
  - Node feature vector
    - Provides information on the type of atoms (e.g., C, H, O, N) present in the molecule
  - Adjacency matrix
    - Provides information on the edges between each node, or which atoms are bonded together
  - The node feature vector and adjacency matrix → used as a molecular graph **input to a graph neural network model**

Counterfactual x'

◆ **An example closer to the original but with a different outcome**

1. Classification

$$minimize\ d(x, x')$$

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

2. Regression

$$minimize\ d(x, x')$$

$$\text{such that } \left|\hat{f}(x) - \hat{f}(x')\right| \geq \Delta$$

- Distance d → Computed with Tanimoto similarity of ECFP4 molecular fingerprints

  → Considered the "gold standard" in molecular distance measurements

Counterfactual x'

◆ **Optimization**

- The optimization problem could be solved by computing a gradient $\nabla_x \hat{f}(x)$.
- However, there are complexities of computing gradients with respect to x
  - It may be a molecular graph, a SMILES string, or descriptors
  - Propagate derivatives to the molecular structure (분자 구조에서부터 미분을 때리는 것이기 때문에 복잡스럽다)
- Generate chemically infeasible structures

- Our innovation here is to use the **STONED SELFIES method**
  - Rapidly explores local chemical space around a point by exploiting the surjective property of SELFIES
    = **Every SELFIES string is a valid molecule**

Counterfactual x'

- **Optimization**
  - SELFIES(Krenn et al.) to overcome one of the major limitations in SMILES
    → They do not always correspond to valid molecules
  - The STONED protocol consists of string insertion, deletion, and modification steps that can generate thousands of perturbations of x that are valid molecules and close in chemical space
  - This requires no training, is independent of features (e.g., molecular graphs, SMILES, descriptors), and requires no gradients

Method

◆ **Methods**

- Starting Molecule → from encoded into SELFIES & token (amino acids, atoms, residues) deletion, replacement, insertion
- To stay in local in chemical space → limit the number of modifications

→ Total 3,000 modified molecules are generated with at most 2 mutations

→ For substitutions : atoms (B, C, N, O, S, F, Cl, Br, I) – Basic alphabets

→ Molecular fingerprint : ECFP4

→ Similarity metric : Tanimoto Similarity

Method

- **Counterfactual in detail**
  1. STONED : generates a set of molecules around the molecule from base molecule
  2. Apply the optimum condition in eqn (1)
  3. DBSCAN is used to cluster multiple counterfactuals ($\mathcal{E} = 0.15$, 5 samples per cluster)

Blood-brain barrier permeation prediction

◆ **Blood-brain barrier**

CNS 신약개발, 뇌혈관장벽(BBB) 통과해야

최국림 기자 | 승인 2020.10.13 17:52 | 댓글 0 | 좋아요 1개 공유하기

뇌혈관장벽(BBB) 통과 기술, 신약개발에 있어 중요

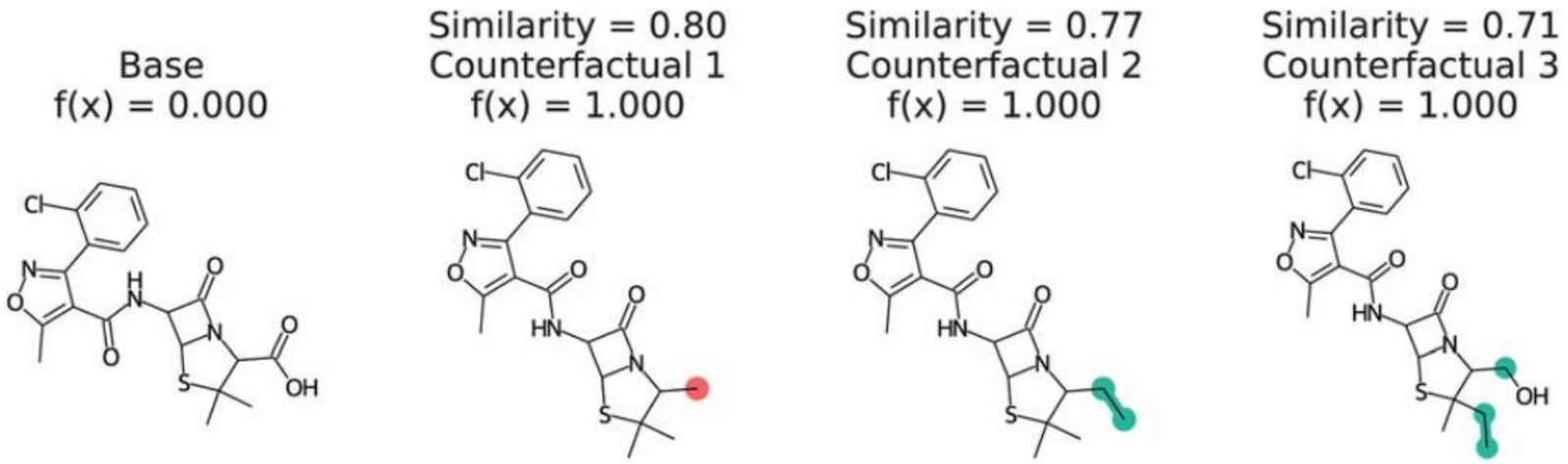
최근 제약업계에서는 뇌혈관장벽(BBB, Blood-brain barrier)이 화두로 떠오르고 있다. 퇴행성 뇌질환나 뇌암 관련 신약 개발에 있어 BBB를 통과하는 기술이 가장 중요한 요인중의 하나로 여겨지고 있는 것이다.

- 뇌 조직을 다른 생체조직과 '구분되게 해주는' 가장 중요한 몸의 장벽
- 색소, 약물, 독물 등 이물질이 뇌조직으로 들어오는 것을 방해하여 뇌를 보호하는 관문
- 뇌세포를 둘러싼 뇌혈관에 전체적으로 분포

Blood-brain barrier permeation prediction

◆ **Blood-brain barrier**

- Binary classification task with a random forest model implemented with Scikit-Learn



- Explanation : The negative example can be made to cross the BBB if the carboxylic group is altered

Small molecule solubility prediction

- **Solubility**
  - Solubility in water plays a critical role in drug design
  - Obtain solubility data from "M. C. Sorkun, A. Khetan and S. Er, Sci. Data, 2019, 6, 1–8." which consists of organic and organometallic molecules
  - Solubility of the molecule in water is measured in log molarity
    (용액 1리터 속에 녹아 있는 용질의 양을 몰로 나타낸 것)

Model & Prediction

- **GRU / RNN**
  - Predict solubility of a given molecule using a gated recurrent unit (GRU)
  - A standard approach in natural language programming tasks because of their ability to handle
  - long sequences and model long-range correlations
  - Commonly used in chemistry applications with SMILES sequences

  - For the regression task

$$minimize\ d(x, x')$$

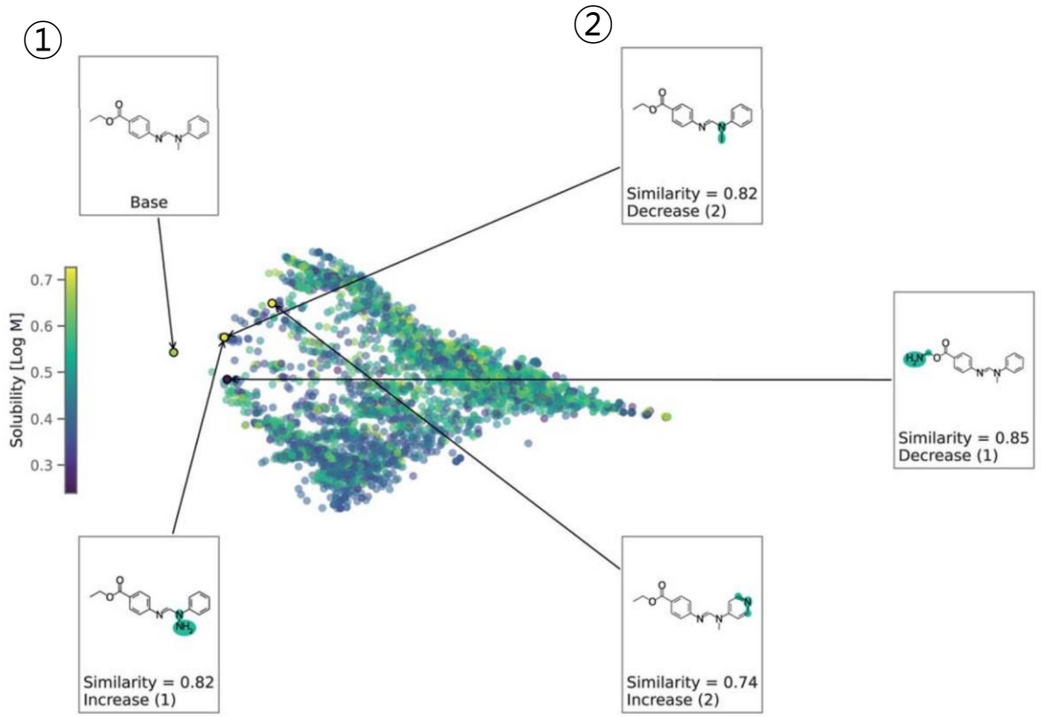
$$\text{such that } \left|\hat{f}(x) - \hat{f}(x')\right| \geq \Delta$$

Model & Prediction

◆ **Chemical space for solubility prediction RNN Model**

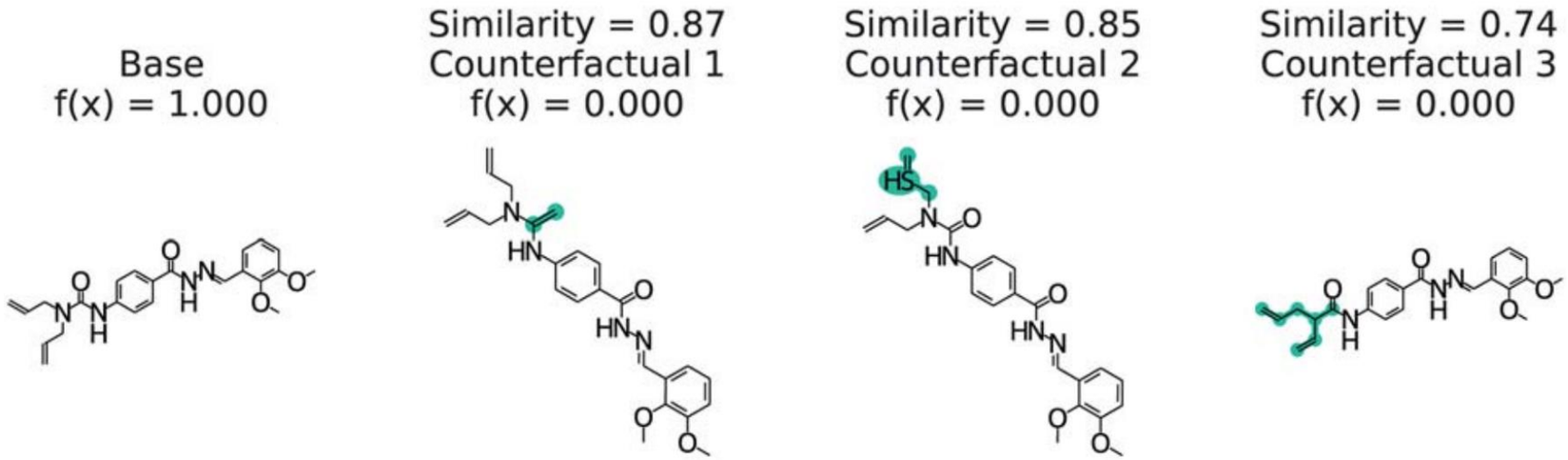
① Starting with base molecule

② Generate Counterfactual



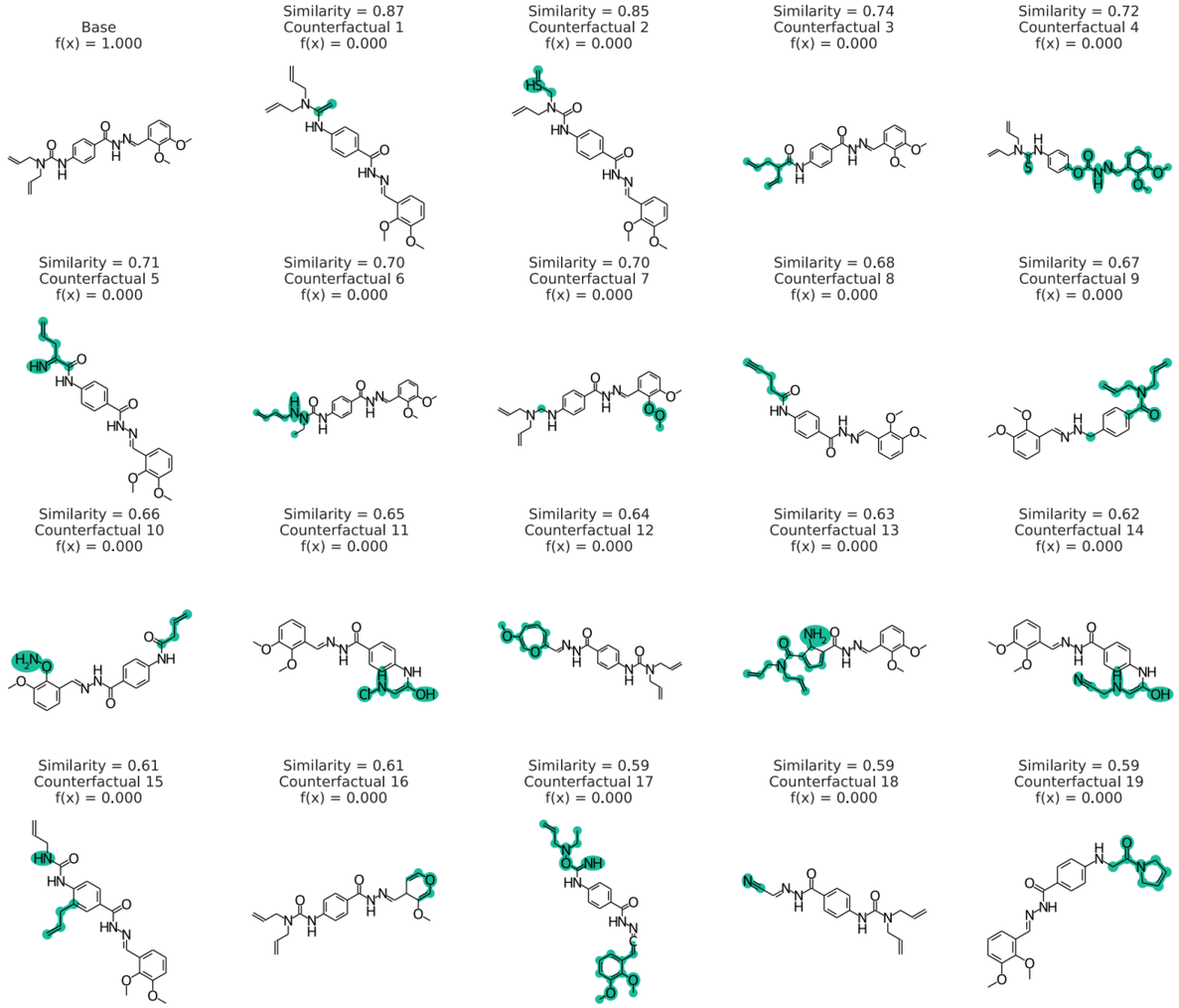
HIV Inhibiting prediction

- **HIV Inhibiting**
  - Binary classification task with a GCN implemented
  - Kaggle competition (prepared by Drug Therapeutics Program (DTP) - 40,000 compounds
  - Normalized molecular graphs from SMILES



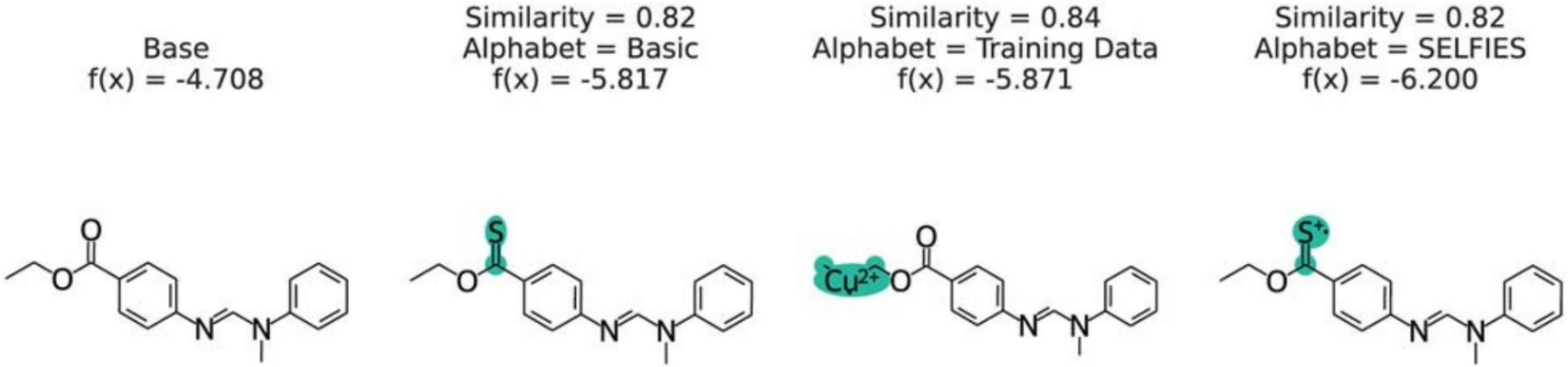
HIV Inhibiting prediction

- **Counterfactuals**



Parameters

- **Importance of parameters**
  - The number of molecules to sample,
  - the number of mutations,
  - the choice of alphabet



- The basic alphabet provides a balance of intuitive counterfactuals and enough tokens to explore chemical space

Conclusions

Conclusions

- This work proposes a universal explainer for any black-box model without requiring training data and regardless of model type
- This is based on counterfactuals, which are interpretable explanations composed of molecular structures
- To illustrate the model-agnostic nature of MMACE we tested our method on three different model types and three datasets

Q & A

Thank You!