Open-Source Cheminformatics and Machine Learning

X-FP – eXplainable FingerPrints

20th September 2023

RDKit UGM 2023 - Mainz

MAX PLANCK INSTITUTE OF MOLECULAR PHYSIOLOGY

Aishvarya Tandon[1,2] and Marcel Baltruschat[3]

sanofi

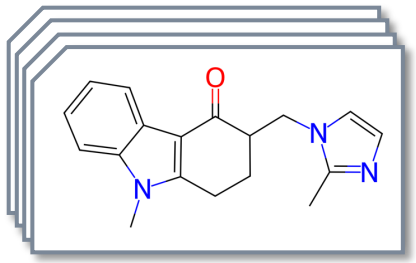
1. Johannes Gutenberg University, Mainz
2. Max Planck Institute of Molecular Physiology, Dortmund
3. Sanofi-Aventis Deutschland GmbH, Frankfurt a.M.

Disclaimer

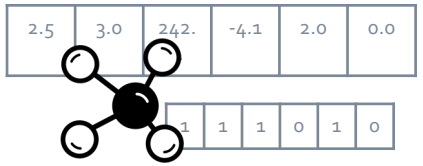
The content of this presentation, the results shown, and the research behind them have been generated exclusively outside Sanofi and are therefore not related to any work at Sanofi or any results of Sanofi's internal research and processes.

What it is all about?

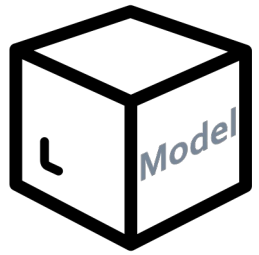
What happens here?



Molecule Dataset



- Molecular Descriptors
- Molecular Fingerprints
- Graphs
- …



- Support Vector Machines
- Random Forest
- Multilayer-Perceptrons
- Graph Neural Networks
- …

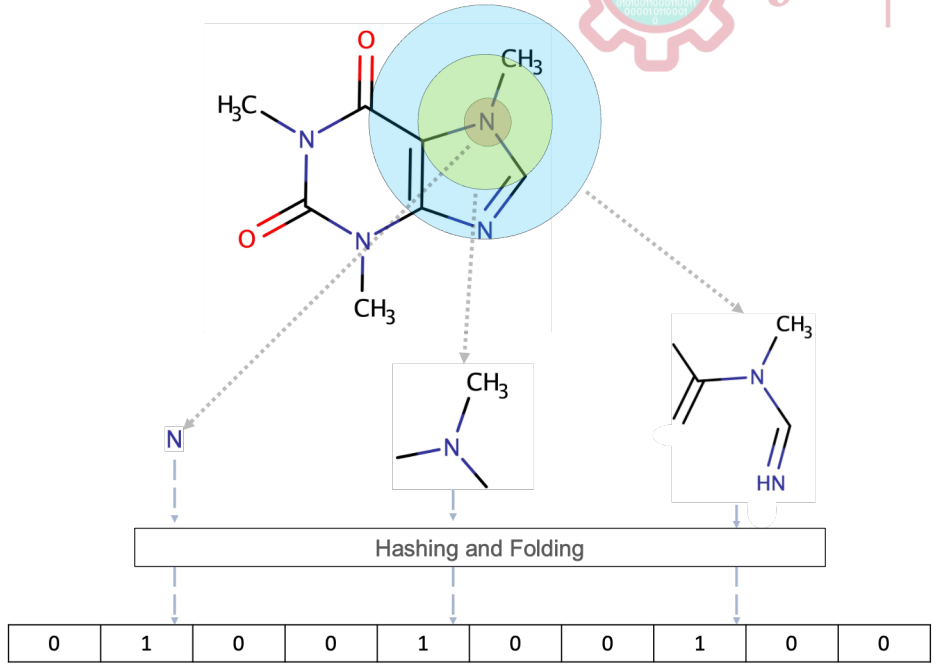
Prediction of molecular property/activity

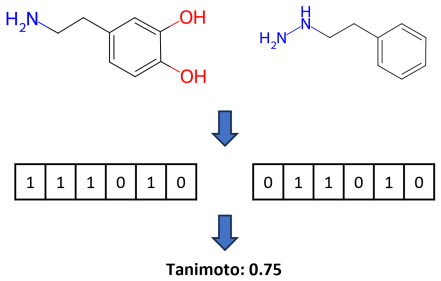
What information does the model actually get?

Morgan Fingerprints: Introduction

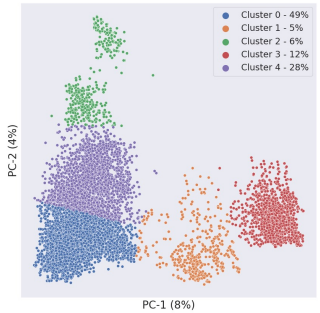
- Circular fingerprint that takes the neighborhood of each atom within different radii into account
- Considers element, total degree, total number of Hs, formal charge, isotope and ring membership
- Very similar to ECFP[1] implementation in Pipeline Pilot

  → Morgan2 roughly equals ECFP4
- Presence of unique substructure environments are encoded into every bit
- Is usually folded to a fixed bit vector length → Possible Collisions

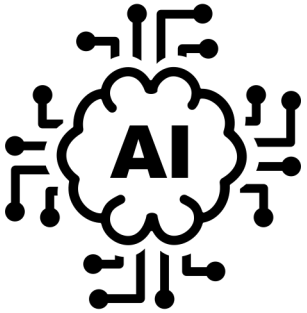




Similarity Calculation



Visualization of Chemical Space



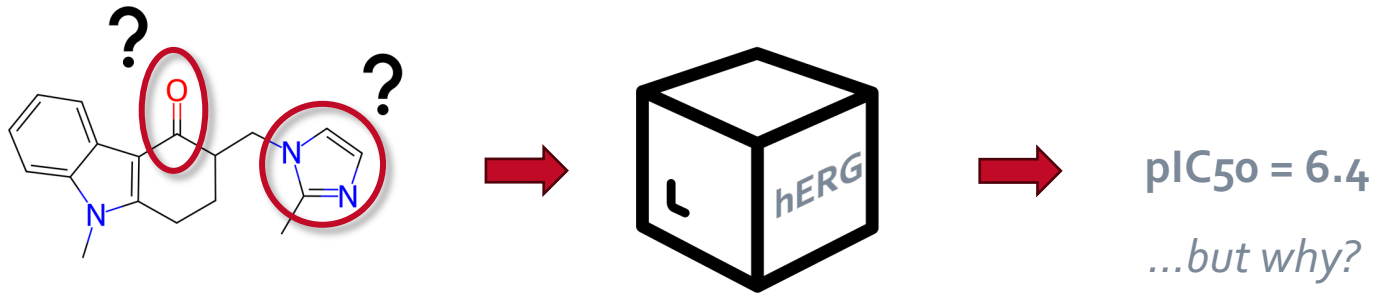
Machine Learning

→ Bits encode different substructures

→ **Which are the important features for the given problem?**

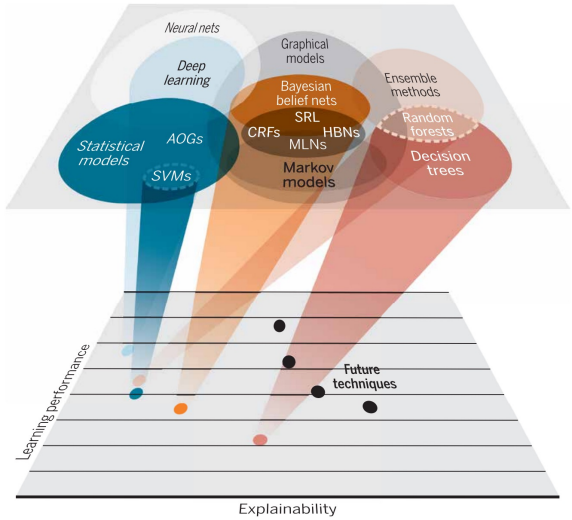
1. Rogers and Hahn, "Extended-Connectivity Fingerprints" *J. Chem. Inf. Model.*, *50*(5), 742–754. DOI: 10.1021/ci100050t

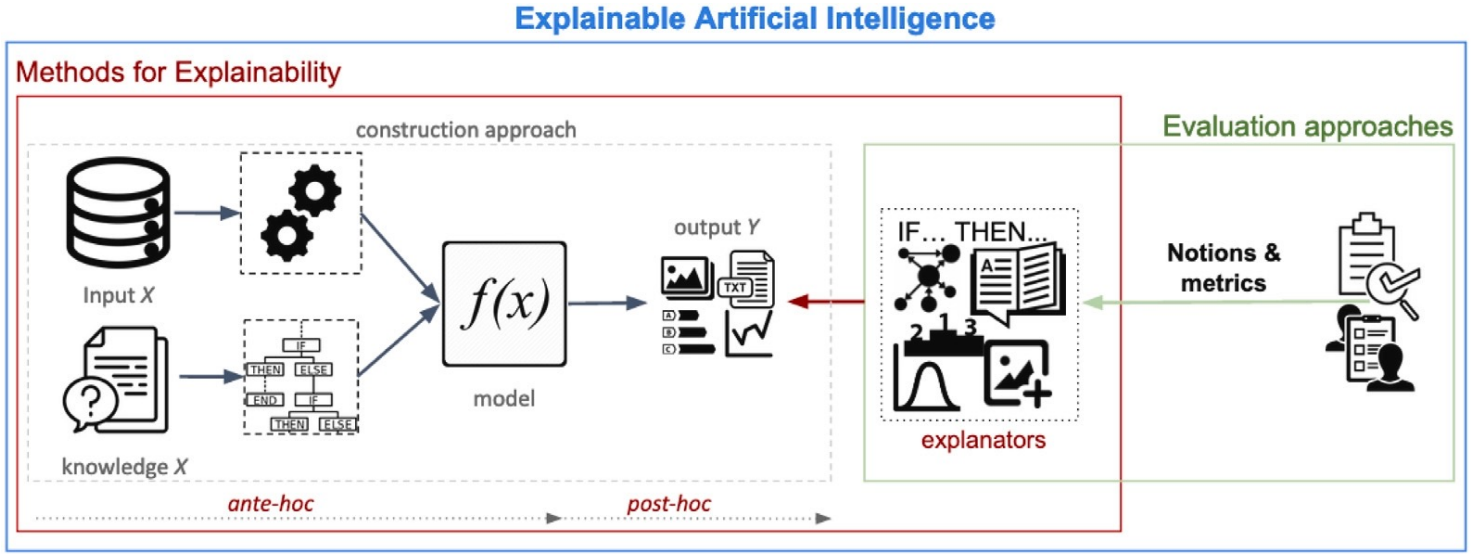
Explainable AI

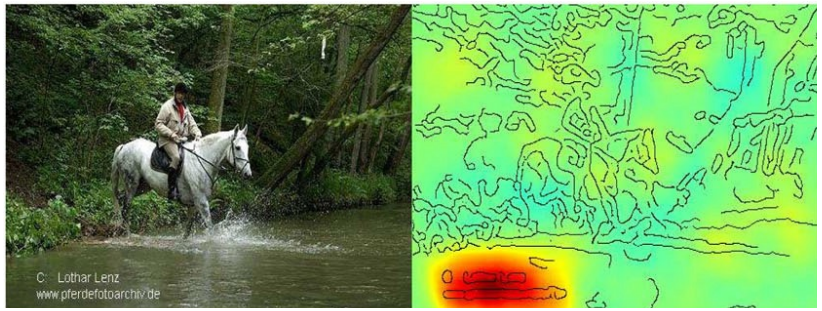


pIC50 = 6.4

...but why?



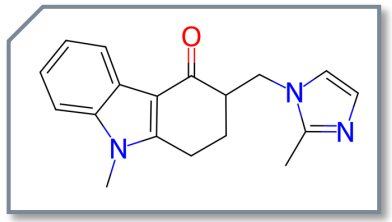




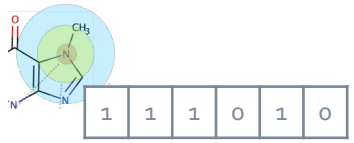
→ It is important to understand why a model makes a prediction

1. Gunning *et al.*, "XAI—Explainable artificial intelligence", *Sci. Robot.* 4, eaay7120. DOI: 10.1126/scirobotics.aay7120
2. Vilone and Longo, " Notions of explainability and evaluation approaches for explainable artificial intelligence", *Information Fusion*, *76*, 89–106. DOI: 10.1016/j.inffus.2021.05.009
3. Lapuschkin *et al.*, "Unmasking Clever Hans predictors and assessing what machines really learn", *Nat Commun* **10**, 1096 (2019). DOI: 10.1038/s41467-019-08987-4

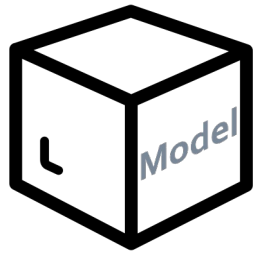
Using MF for XAI is challenging



1 Molecule

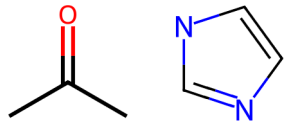


Morgan FP



? Why?

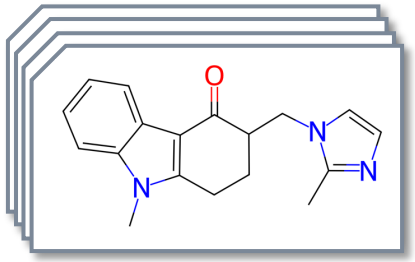
Prediction



Substructure *A* and *B*

Known Feature Importance
→ Bit *X* is most important

? Are these really important for the model?



Many molecules

Morgan FPs

Substructures Bit *X*

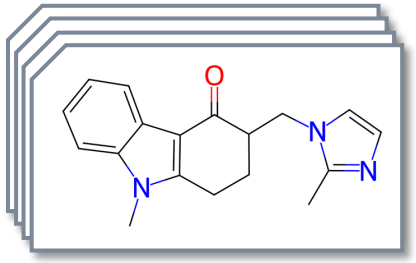
Substructure	Count
A	5
B	15
C	870
D	110

→ Most relevant substructure for the model might be C

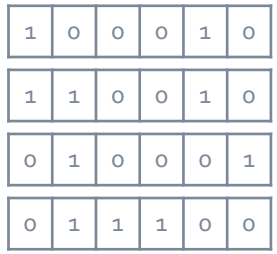
Introducing X-FP

- X-FP → eXplainable FingerPrints
- An open-source, model-agnostic Python software to perform explainable machine learning on Morgan Fingerprint based models.
- Features
  - Extract and visualize Morgan FP substructures in context of a complete dataset
  - Export substructures as SMARTS pattern
  - Examine importance of fingerprint bits via different methods, e.g. SHAP
  - Generate PDF reports about the most important bits and their substructures and distributions
- GitHub: https://github.com/czodrowskilab/x-fp

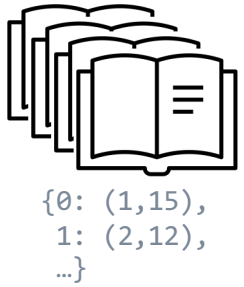
X-FP: Substructure Calculation



Molecules

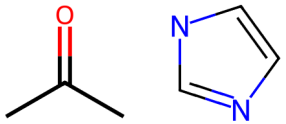


Fingerprints



„bitInfo“ dictionaries

Extract all substructures of all molecules



…

Convert to unique canonical SMARTS patterns[1]

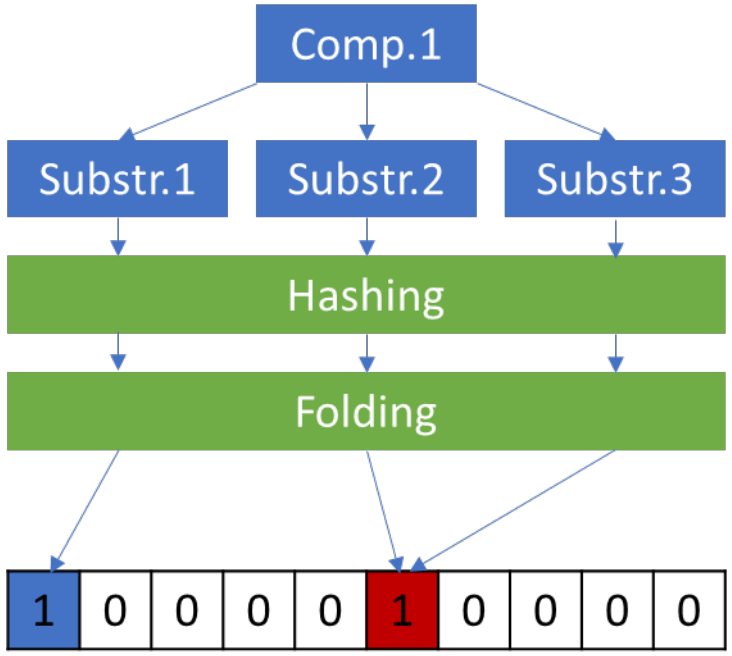
Important RDKit Functions

```
Chem.GetMorganFingerprintAsBitVect()

Chem.FindAtomEnvironmentOfRadiusN()

Chem.MolFragmentToSmiles()
```

1. Greg Landrum, “Looking at statistics for bits set by Morgan fingerprints” (03.02.2016), Weblink: https://rdkit.blogspot.com/2016/02/morgan-fingerprint-bit-statistics.html (accessed 10.06.2023)

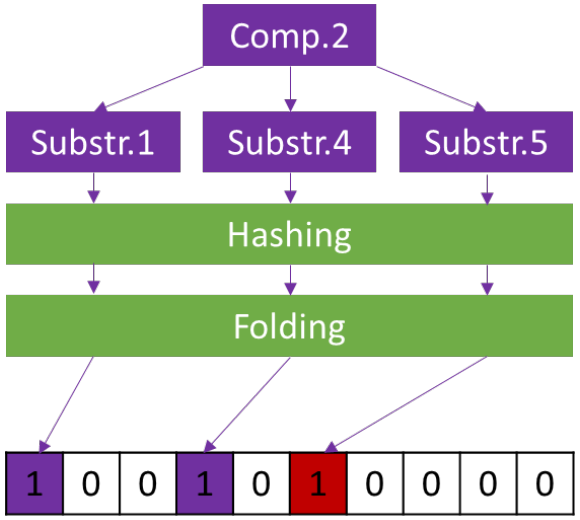
Bit collisions

- Local bit collision

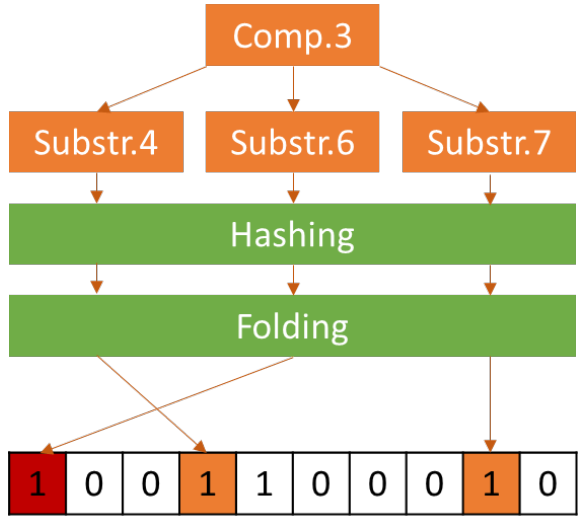


Bit 0 encodes substructure 1
Bit 5 encodes substructures 2 and 3

- Global bit collision



Bit 0 encodes substructure 1
Bit 3 encodes substructure 4
Bit 5 encodes substructure 5



Bit 0 encodes substructure 6
Bit 3 encodes substructure 4
Bit 8 encodes substructure 7

Globally:

- Bit 0 encodes substructures 1 and 6
- Bit 3 encodes substructure 4
- Bit 5 encodes substructures 2, 3 and 5
- Bit 8 encodes 7

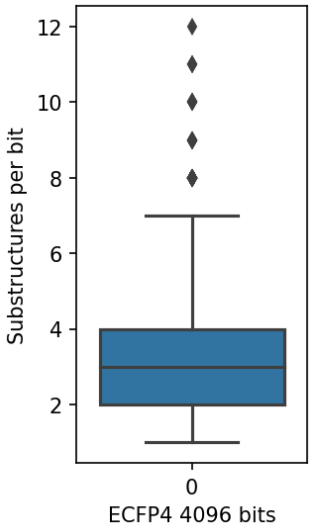
Bit collisions and substructure frequency

Local bit collision

- Rare
  - depends on number of bits

Global bit collision

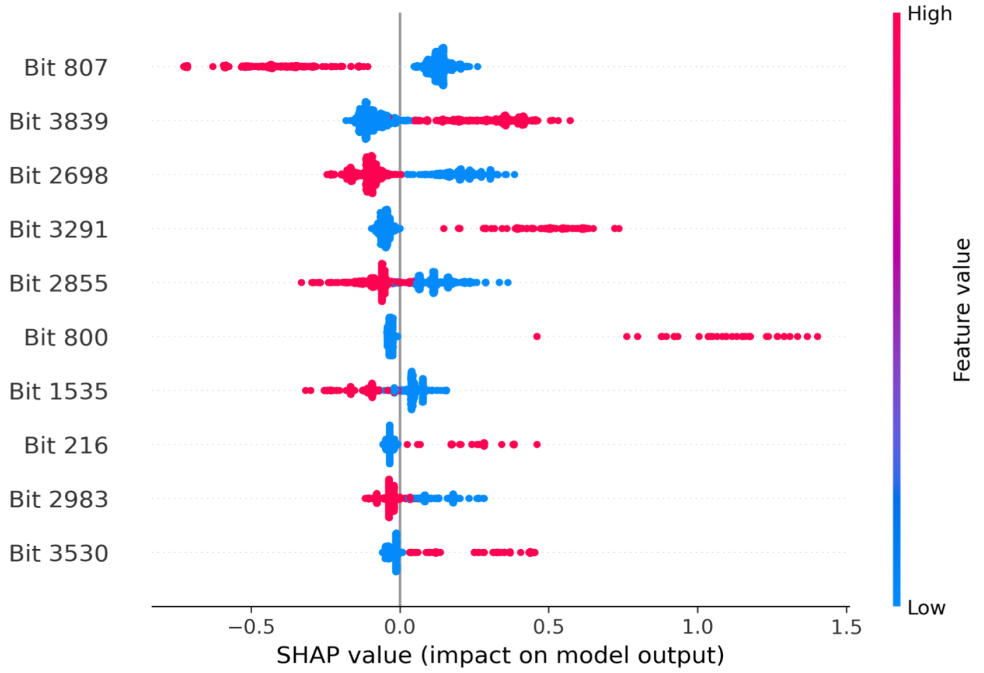
- Very common
  - depends on number of bits and dataset size
    - "Bit 5 is important, but which substructure encoded is the most important one?"
    - Solution: substructure frequency



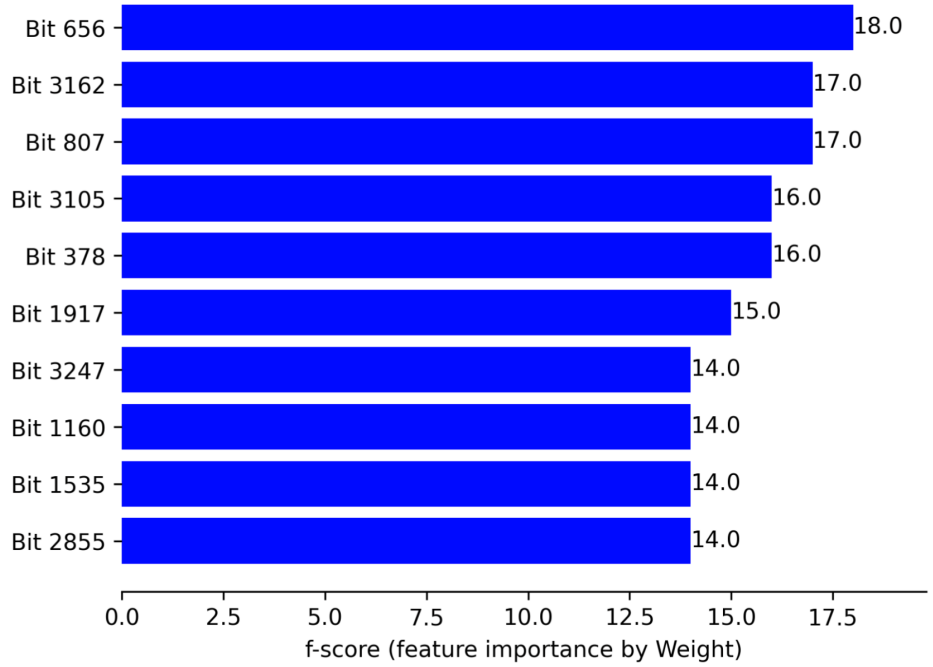
X-FP offers "Substructures per bit" plot to quickly inspect global bit collision

X-FP: Feature importance methods of choice

- Different feature importance methods – different ways to calculate feature importance
- X-FP's modularity allows usage of any suitable method



SHAP Summary Plot for the Top 10 Morgan Fingerprint Bits



XGBoost's Weight Feature Importance Plot for the Top 10 Morgan Fingerprint Bits

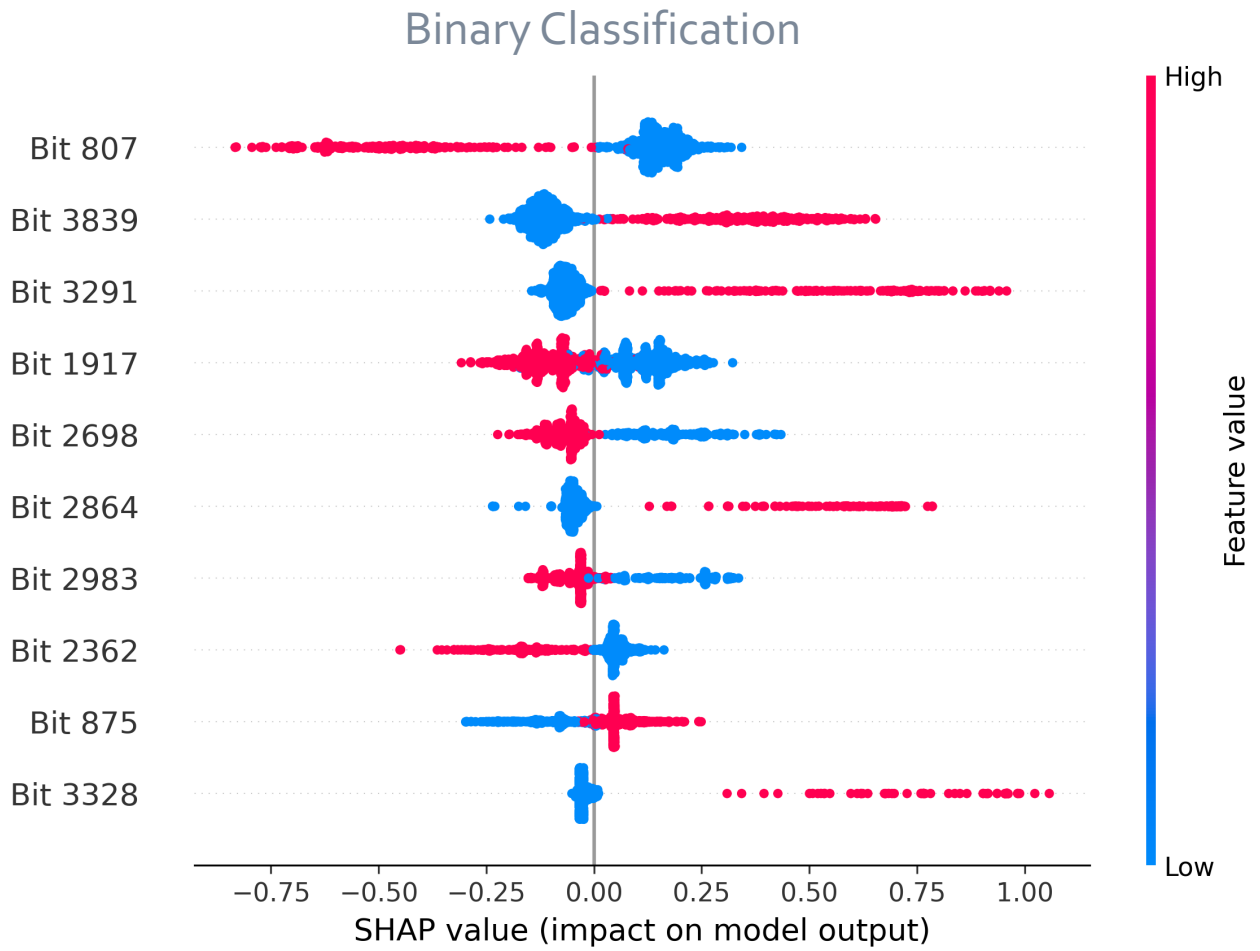
SHAP TreeExplainer and XGBoost's feature importance methods are already integrated in X-FP

X-FP: SHAP TreeExplainer

- SHAP TreeExplainer is a feature importance method for tree-based models [1]
- Based on SHAP (SHapley Additive exPlanations) – a Shapley game-theory based approach for explainable machine learning [2]

1. Lundberg et al., "From local explanations to global understanding with explainable AI for trees", *Nat Mach Intell* 2, 2020, 56–67, DOI: 10.1038/s42256-019-0138-9
2. Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems* 30, 2017, 4765-4774c

X-FP: SHAP TreeExplainer



SHAP Summary Plot for the Top 10 Morgan Fingerprint Bits

X-FP: Intuitive report generation

- Allows flexibility for developers to add document manager of choice
- The default implementation is generation of an intuitive PDF report, using FPDF2[1] Python library
- This PDF report is easy to understand by both – cheminformaticians and chemists

1. Weblink: https://github.com/py-pdf/fpdf2/ (accessed: 3 September 2023)

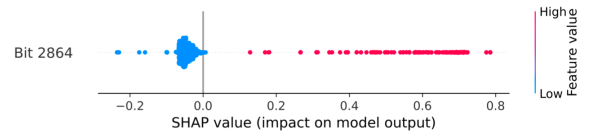
X-FP: Intuitive report generation

X-FP

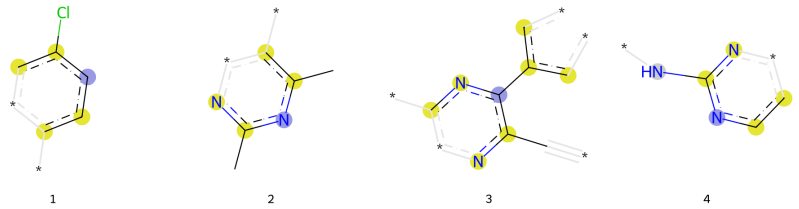
Bit Analysis

Bit 2864

SHAP Summary Plot for the Bit 2864:



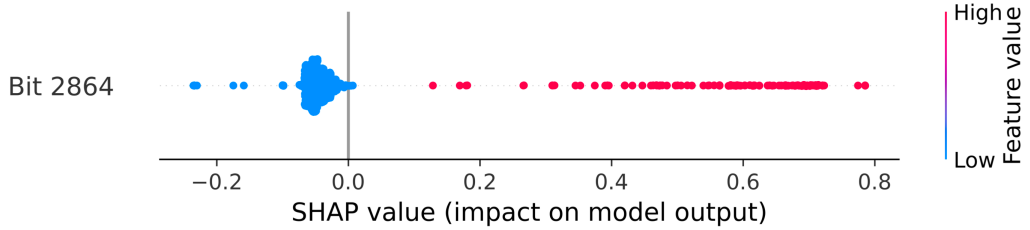
Substructures present:



Substructures frequency table (total compounds: 1216):

Serial	Substructures (as SMARTS)	Total occurrences	Unique occurrences
1	[cH;R]1:[cH;R]:[c;R](~*)~*~[cH;R]:[c;R]:1-[Cl]	117	105
2	[n;R]1:[c;R](:[n;R]~*~[c;R](:[c;R]:1-[CH3])~*)-[CH3]	3	3
3	[c;R]1(:[n;R]:[c;R](~*)~*~[n;R]:[c;R]:1-[C]~*)-[c;R](:[cH;R]~*):[cH;R]~*	1	1
4	[n;R]1:[cH;R]:[cH;R]~*~[n;R]:[c;R]:1-[NH;R]~*	1	1

SHAP Summary Plot for the Bit 2864:



- Bit's feature importance plot
- Here, SHAP TreeExplainer Summary Plot is used
- This plot can be interpreted as that when this bit is on, the model outputs are impacted towards the active class in the binary classification

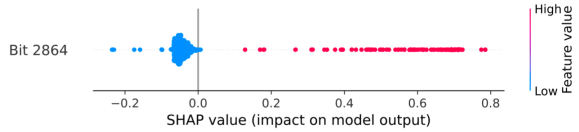
X-FP: Intuitive report generation

X-FP

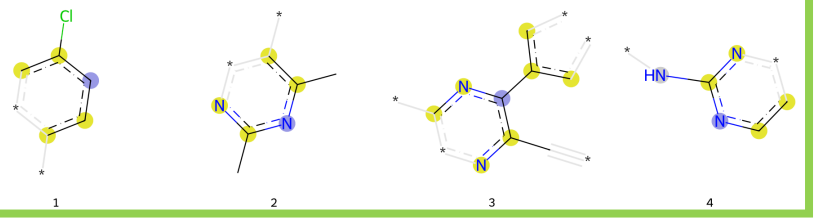
Bit Analysis

Bit 2864

SHAP Summary Plot for the Bit 2864:



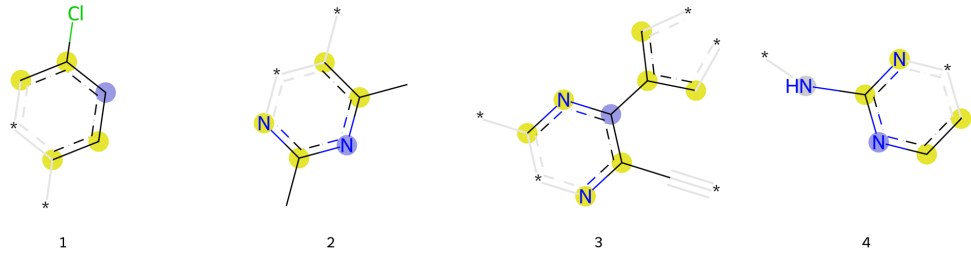
Substructures present:



Substructures frequency table (total compounds: 1216):

Serial	Substructures (as SMARTS)	Total occurrences	Unique occurrences
1	[cH;R]1:[cH;R]:[c;R](~*)~*~[cH;R]:[c;R]:1-[Cl]	117	105
2	[n;R]1:[c;R](:[n;R]~*~[c;R](:[c;R]:1-[CH3])~*)-[CH3]	3	3
3	[c;R]1(:[n;R]:[c;R](~*)~*~[n;R]:[c;R]:1-[C]~*)-[c;R](:[cH;R]~*):[cH;R]~*	1	1
4	[n;R]1:[cH;R]:[cH;R]~*~[n;R]:[c;R]:1-[NH;R]~*	1	1

Substructures present:



- Here we see the substructures encoded by this bit present in the query dataset
- The bit's importance can now be intuitively connected with the substructure information

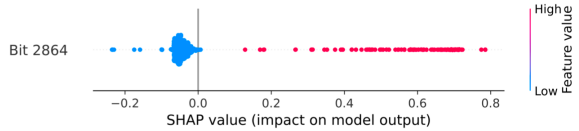
X-FP: Intuitive report generation

X-FP

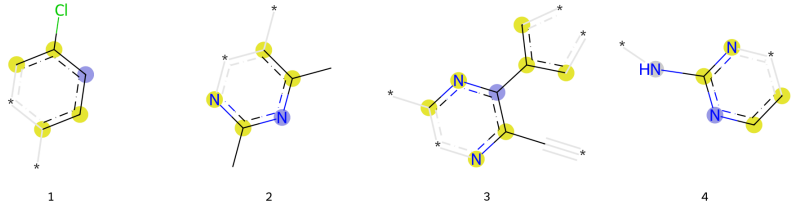
Bit Analysis

Bit 2864

SHAP Summary Plot for the Bit 2864:



Substructures present:



Substructures frequency table (total compounds: 1216):

Serial	Substructures (as SMARTS)	Total occurrences	Unique occurrences
1	[cH;R]1:[cH;R]:[c;R](~*)~*~[cH;R]:[c;R]:1-[Cl]	117	105
2	[n;R]1:[c;R](:[n;R]~*~[c;R](:[c;R]:1-[CH3])~*)-[CH3]	3	3
3	[c;R]1(:[n;R]:[c;R](~*)~*~[n;R]:[c;R]:1-[C]~*)-[c;R](:[cH;R]~*):[cH;R]~*	1	1
4	[n;R]1:[cH;R]:[cH;R]~*~[n;R]:[c;R]:1-[NH;R]~*	1	1

Substructures frequency table (total compounds: 1216):

Serial	Substructures (as SMARTS)	Total occurrences	Unique occurrences
1	[cH;R]1:[cH;R]:[c;R](~*)~*~[cH;R]:[c;R]:1-[Cl]	117	105
2	[n;R]1:[c;R](:[n;R]~*~[c;R](:[c;R]:1-[CH3])~*)-[CH3]	3	3
3	[c;R]1(:[n;R]:[c;R](~*)~*~[n;R]:[c;R]:1-[C]~*)-[c;R](:[cH;R]~*):[cH;R]~*	1	1
4	[n;R]1:[cH;R]:[cH;R]~*~[n;R]:[c;R]:1-[NH;R]~*	1	1

- **Substructures (as SMARTS)**: Accurate SMARTS of the substructure encoded
- **Total Occurrences**: Total number of times the substructure is present in the dataset
- **Unique Occurrences**: Total number of compounds having the substructure at least once in the dataset

X-FP: Validation case study

Binary classification of activity of several well-studied targets in the field of drug discovery

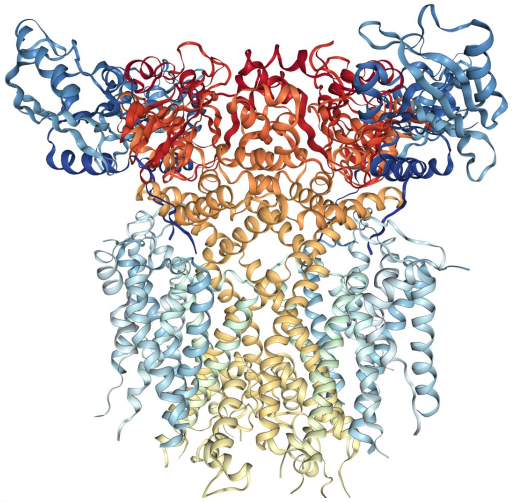
Per target:

- Publicly available data acquisition from ChEMBL
- XGBoost binary classifiers
- X-FP on test set of the best performing model
- Comparison of top substructures and their impacts on activity class obtained by X-FP with the literature

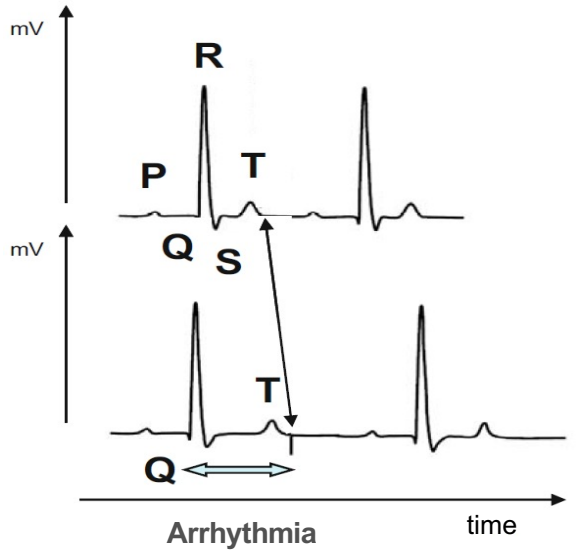
Research by: Hanna Rieger

X-FP: Validation case study - hERG

- Human ether-à-go-go-related gene (hERG)
- Codes for a protein important for the function of heart muscles
- Well-established toxicity target[1,2]
- In drug discovery, small molecules predicted to cause this toxicity are de-prioritized



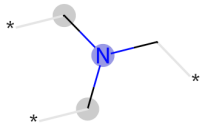
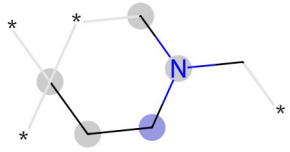
[3]



Q: Start
T: End
of the expulsion phase of the blood pumped out of the heart

1. P. Czodrowski, "hERG Me Out", *J. Chem. Inf. Model.* 2013, 53, 9, 2240-2251. DOI: 10.1021/ci400308z
2. C. Jamieson et al., "Medicinal Chemistry of hERG Optimizations", *J. Med. Chem.* 2006, 49, 17, 5029–5046. DOI: 10.1021/jm060379l
3. Cryo-EM structure of the hERG related K+ channel. PDB ID: 5VA2. Weblink: https://www.rcsb.org/structure/5VA2. Last accessed: 7 June 2023

X-FP: Validation case study - hERG

Bit	Key substructure	Model output impacted towards (active class: causes hERG inhibition, inactive class: doesn't causes hERG inhibition)	Reference
807	Hydroxyl 	Inactive	Zhang et al. *J. Chem. Inf. Model.* 2022, 62, 8, 1830–1839 DOI: acs.jcim.2c00256
3839	Tertiary amine 	Active	Jamieson et al. *J. Med. Chem.* 2006, 49, 17, 5029–5046 DOI: 10.1021/jm060379l
2864	Arylchloride 	Active	Zhang et al. *Toxicol Res (Camb)*. 2016, 5(2), 570-582 DOI: 10.1039/c5tx00294j
3328	Piperidine-derivate 	Active	Jamieson et al. *J. Med. Chem.* 2006, 49, 17, 5029–5046 DOI: 10.1021/jm060379l

Outlook

- Introduction and validation of X-FP
- Available on GitHub; manuscript soon

- Recommendation: X-FP on different subsets

- Next:
  - More feature importance methods
  - More fingerprints
  - Machine readable outputs

Thanks for your attention

Acknowledgment

- Greg and his blogs
- Our colleagues at:
  - CzodrowskiLab
  - Sanofi
  - Chemical biology department, MPI Dortmund

Disclosures

Marcel Baltruschat is Sanofi employee and may hold shares and/or stock options in the company.
Aishvarya Tandon has nothing to disclose.