Machine learning models with distinct Shapley value explanations for chemical compound predictions decouple feature attribution and interpretation

LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Authors: Jannik P. Roth, Prof. Dr. Jürgen Bajorath

B-IT, Department of Life Science Informatics and Data Science,
Lamarr Institute for Machine Learning and Artificial Intelligence
University of Bonn

Compound Activity Prediction

Study outline

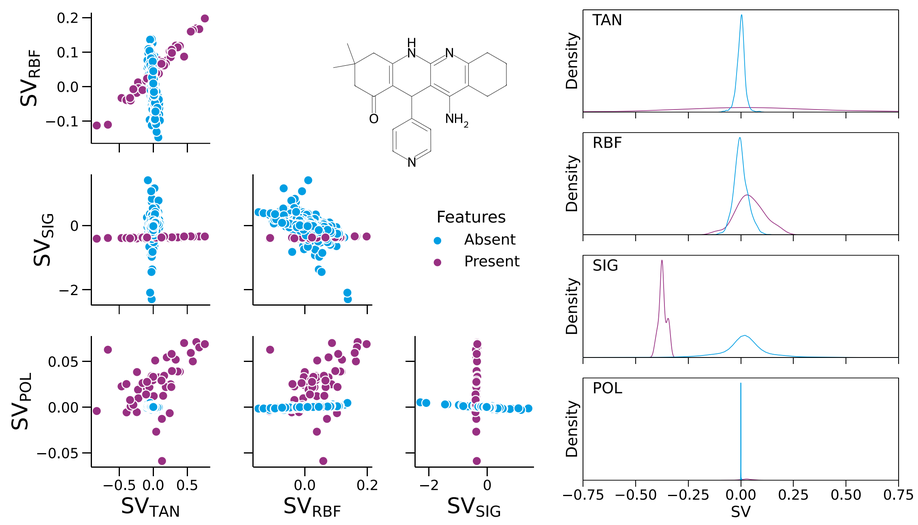
- Derivation of support vector classifiers for compound activity prediction
- Use of four kernels: Tanimoto **TAN**, radial basis function **RBF**, polynomial **POL**, and sigmoid **SIG**
- Molecules represented using **binary** extended connectivity **fingerprints** encoding the presence/absence of molecular features

Performance results

- All kernels perform well on a wide range of activity classes (Accuracy >= 0.95, MCC >= 0.90)
- Prediction quality **not** strongly influenced by choice of kernel

Shapley Values for Explainable Machine Learning

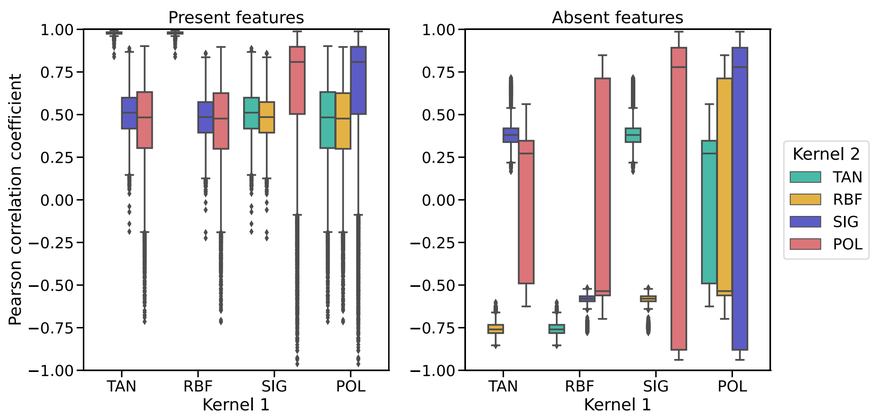
- Shapley value concept from **game theory** quantifies each player's contribution to the outcome of the collaborative game
- Application to machine learning: attributing each feature's contribution to the prediction of a test instance



On the left, scatterplots compare Shapley values (SV) of all features of a test compound for the different kernels (denoted by subscripts). On the right, kernel density estimate plots display Shapley value distributions of all features that were present or absent in the test compound.

- Substantial **different importance attribution** between kernels
- Analysis needs to be carried out **separately for present and absent features**

Correlation of Shapley Values

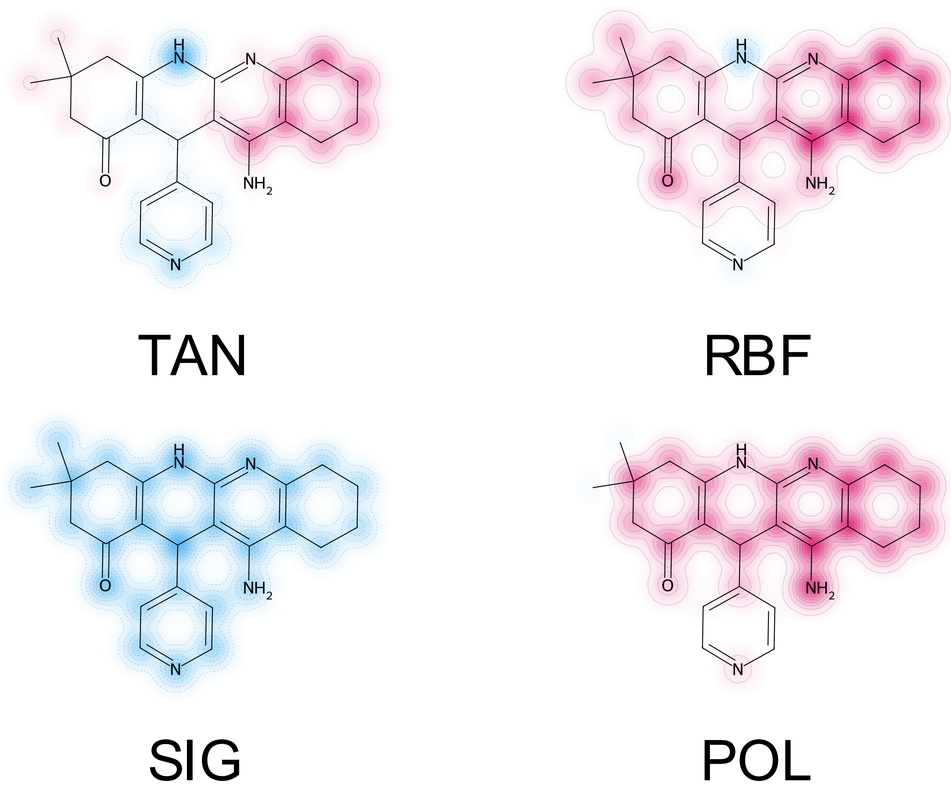


Boxplots show the distribution of Pearson correlation coefficient values between Shapley values for all kernel pairs and test compounds. On the left and right, correlation statistics are reported for present and absent features, respectively.

- Only **limited correlation** for kernel pairs
- Correlation **substantially differs between present and absent features**, even for the same kernel pair

Shapley Value Mapping and Interpretation

- Structural fingerprints enable **atom-based mapping of importance values of structural features present in test compounds**
- For each atom, the contributions from all features are summed
- Mappings can be **used as explanation to aid in** decision-making



Shapley values of molecular features mapped onto a correctly predicted test compound for all kernels. The contributions were transformed into color-coded contours such that increasing atom-based values corresponded to increasing color density. Pink and blue coloring indicates positive and negative contributions to the prediction of activity, respectively.

- Interpretation of mapped features **vastly different across kernels**
- Different kernels lead to **contrary contributions** for the same chemical substructure
- Large differences in feature importance **hinder the chemical interpretation** greatly

Summary

- Comparison of Shapley values reveals **differences between explanations**
- Different kernels show **different attribution behaviors** and only **limited feature value correlation**
- **Feature mapping** on test compounds delineates **different substructures**
- **Kernel-dependent differences in feature attribution lead to inconsistent explanations of SVM predictions**

Outlook for explainable machine learning

- **Success of machine learning** in interdisciplinary research settings greatly **depends on transparency of predictions**
- Feature attribution techniques **do not always produce interpretable explanations**
- **Feature attribution and other numerical approaches for explaining predictions should best enable intuitive and consistent interpretations**

References and further reading

Publication

Repository

Main publication:

1. Jannik P. Roth, and Jürgen Bajorath. "Machine learning models with distinct Shapley value explanations decouple feature attribution and interpretation for chemical compound predictions." *Cell Reports Physical Science* 5.8 (2024).

Further Reading:

1. Chirsitan Feldmann, and Jürgen Bajorath. "Calculation of exact Shapley values for support vector machines with Tanimoto kernel enables model interpretation." *iScience* 25.9 (2022).
2. Andrea Mastropietro, Christian Feldmann, and Jürgen Bajorath. "Calculation of exact Shapley values for explaining support vector machine models using the radial basis function kernel." *Scientific Reports* 13.1 (2023): 19561.
3. Andrea Mastropietro, and Jürgen Bajorath. "Protocol to explain support vector machine predictions via exact Shapley value computation." *STAR Protocols* 5.2 (2024): 103010.

Partner institutions:

technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Institutionally funded by:

Bundesministerium für Bildung und Forschung
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen