SHAPBALS: Improving Molecular Mapping and Interpretability of Machine Learning for Property Prediction

congruence TX

Christopher Kottke, Michael-Rock Goldsmith, Maximilian Ebert

Abstract

The SHAPley Based Activity Labeling and Scoring (SHAPBALS) method is a technique for enhancing the interpretability of machine-learning (ML)-based approaches for predicting properties of biomolecular systems. The method is a generalized extension of the Mutual Information Based Activity Labeling and Scoring (MIBALS) method developed by The Chemical Computing Group (CCG). Both SHAPBALS and MIBALS project outputs of ML models onto input chemical structures at inference using reverse fingerprinting to explain which components of a molecule contribute positively or negatively to the output predictions of the model. SHAPBALS is a general-purpose black-box explainer approach to models which take binary and/or count fingerprints as input and is agnostic to the model itself. This technique addresses multiple needs on the interface of quantitative structure-property relationship (QSPR) modeling and medicinal chemistry by enhancing trust in models by opening the black-box that models usually exist within to allow practitioners to better understand what components of an input structure influence a particular model output.

MIBALS – Background

Techniques for explaining the rationale behind model predictions is an important, yet often overlooked aspect of the field of QSPR for drug-like molecule property prediction. In this context, the cost of experiments can be significant, so the ability to understand the factors which lead to a model prediction is vital to being able to take actionable steps based on such predictions. MIBALS [1] is one such technique which uses a mutual-information-based scoring function to identify how binary input features tend to correlate with binary outputs. MIBALS then uses reverse fingerprinting (RFP), a technique to associate molecular fingerprint bits to the input structure components which contribute to each bit, to project these learned scores onto the input molecule.

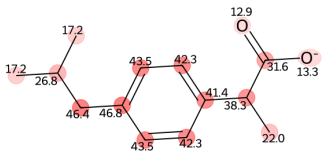


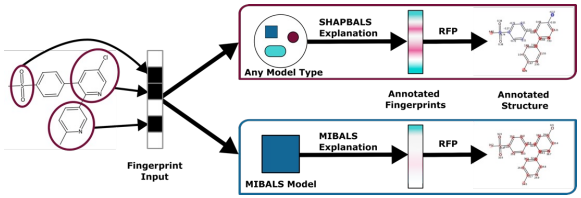
Figure 1. **Sample MIBALS projection of a cardiotoxic/stroke adverse outcome model applied to ibuprofen. Atoms with larger positive scores participate in fingerprint bits which are present more frequently in molecules with known cardiotoxic/stroke adverse outcomes.**

$$S_a = \sum_k \left(\mathbb{1}(a \in k) \cdot \frac{S_{x\,k}}{n_k \cdot \sum_{a'} \mathbb{1}(a' \in k)} \right)$$

Equation 1. **Formula for attributing per-bit scores (S_{xk}) to each atom in the input structure. For each bit (k), we identify if the atom of interest participates in that bit. If so, we add the score of that bit divided by the number of other atoms in that bit and the number of times that bit feature appears in the structure. This results in an additive explanation where the sum of the bit scores equals the sum of the atom scores.**

MIBALS is an advantageous approach to molecular property modeling as the lack of hyperparameters mean that little-to-no ML experience is necessary to create useful models. This comes at a cost however, since MIBALS can only be used in datasets with binary inputs and outputs. Further, it is restricted to a basic mutual information-based naïve Bayes model which has a tendency to underfit to the data. To extend MIBALS to include count fingerprint inputs, continuous outputs, and any model type, we introduce the SHAPBALS method.

SHAPBALS – Method



SHAPBALS uses SHAP [2] to attribute changes in the presence of input bits to changes in model output. The computed SHAP values can be used as a 'drop-in' replacement of the per-bit scores used by MIBALS in *Equation 1*.

$$S_k(f,x) = \mathrm{E}_{\hat{x} \subseteq D}[f(\hat{x}_{+k}) - f(\hat{x}_{-k})]$$

Equation 2. **Equation for computing the Shapley value of a dimension, k, centered at a point, x, for a general function, f. In words, the Shapley value is equal to the average difference in function values for all sets of counterfactual inputs, $\hat{x}$, where +k denotes that $\hat{x}$ takes the value of x at bit k, and -k denotes that $\hat{x}$ takes the values of $\hat{x}$ at bit k. SHAP uses a Monte-Carlo approximation to make this calculation tractable.**

SHAPBALS – Results

MIBALS and Gaussian Process (GP) models were created from the Side-Effect Resource (SIDER) database [3] to predict if a compound is associated with cardiotoxic / stroke adverse outcomes. Molecular representations are concatenated RDKit [4] bit fingerprints. COX2 inhibitors were removed from the training set to allow for performance comparison.

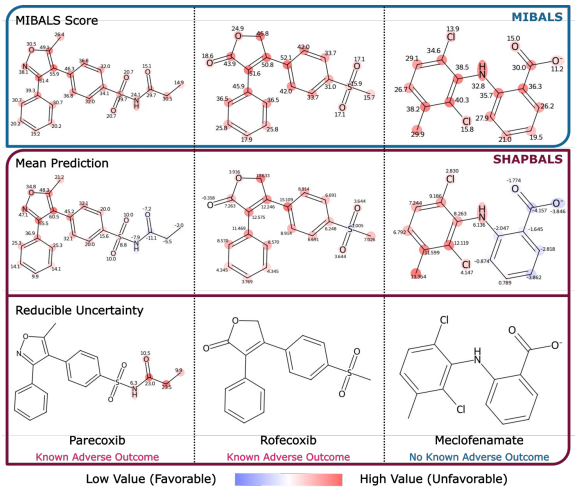


Figure 2. **Comparison of MIBALS (Linear Model) and SHAPBALS (Gaussian Process) RFP outputs applied to COX2 inhibitors. Each row corresponds to a model output. In the last row, only atoms which increased the GP uncertainty were labeled.**

SHAPBALS / MIBALS Comparison

MIBALS

- \+ Parameter-free model type
- \+ Robust to overfitting
- \- Restricted to Naïve Bayes
- \- Restricted to bit-based I/O
- \- Models tend to underfit

SHAPBALS

- \+ Independent of model-type
- \+ Continuous label compatible
- \+ Full generalization of MIBALS
- \+ Usable with model distillation
- \- Models should be properly regularized

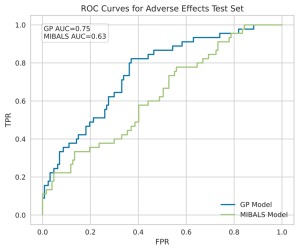


Figure 3. **ROC Plots of the MIBALS and GP model for predicting cardiotoxic and stroke-based adverse outcomes on the independent test set. MIBALS requires use of a Naïve Bayes model while SHAPBALS is compatible with any model type.**

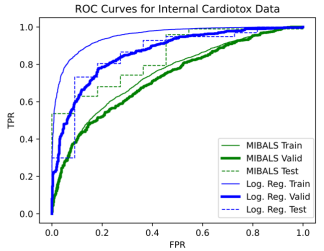


Figure 4. **ROC Plots of MIBALS and Logistic Regression models for predicting cardiotoxicity on internal data. Both models shown are explainable with SHAPBALS, however MIBALS can only be used with the MIBALS model.**

Conclusions

The SHAPBALS method improves upon on MIBALS, a previously known method for identifying and labeling the individual atoms which participate in structure-property relationships in molecular property modeling. Unlike MIBALS, SHAPBALS serves as a broadly applicable, model-agnostic technique which can project predictions from both bit and count fingerprints back onto the original structure. By improving upon the explainability of these more accurate and generalizable models, SHAPBALS provides a pathway to reduce the number of design cycles in molecular optimization by identifying the specific motifs responsible for potential liabilities. This enables the rational design of new iterations of compounds by reducing the need for random modifications to identify liability sites while also maintaining the desired activity of a drug candidate.

References

[1] C. Williams and S. Schreyer, "Reverse fingerprinting and mutual information-based activity labeling and scoring (MIBALS)," Comb Chem High Throughput Screen, pp. 424-39, 2009.

[2] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NIPS'17, pp. 4768 - 4777, 2017.

[3] M. Kuhn, I. Letunic, L. J. Jensen and P. Bork, "The SIDER database of drugs and side effects," Nucleic Acids Research, vol. 44, no. D1, p. D1075–D1079, 2016.

[4] "RDKit: Open-source cheminformatics," [Online]. Available: https://www.rdkit.org.

Abbreviations:
SHAPBALS – Shapley Based Activity Labeling and Scoring
MIBALS – Mutual Information Based Activity Labeling and Scoring
ML – Machine Learning
QSPR – Quantitative Structure Property Relationship
RFP – Reverse Fingerprint
GP – Gaussian Process

Acknowledgements:
Thanks to Behrgen Smith and Sina Sarparast for their review of this poster