Optimization of Allosteric GCase Correctors Using Shapley-Based Activity Labeling and Scoring (SHAPBALS) Annotations

congruence TX

Christopher Kottke, Jeremy Dupaul-Chicoine, Lee Fader, Marielle Fournel, Indranath Ghosh, Michael-Rock Goldsmith, Stephanie Lavallée, Nathalia Azevedo Portilho, Sina Sarparast, Maximilian Ebert

Abstract

A common challenge encountered during the discovery of pharmacologically active compounds is the multi-parameter optimization (MPO) of compound properties such as toxicity, solubility, or metabolic stability while also maintaining desired potency/efficacy. To address this challenge, we introduce a new technique to assess the contributions of individual atoms in a molecular structure to the associated properties of interest. By evaluating these attributes, regions in a molecular structure associated with undesirable properties as well as desirable properties can be identified. We call this technique the Shapley-Based Activity Labeling and Scoring (SHAPBALS) method which extends upon the Mutual Information Based Activity Labeling and Scoring (MIBALS) method[1] distributed by the Chemical Computing Group (CCG). MIBALS can provide atomic attributions, however it is limited to a specific linear model type which can only operate on binary labels. While SHAPBALS is equivalent to MIBALS when using the same model, it is applicable to any model type (neural network, random forest, etc.) and task (classification / regression). This advancement in explainable molecular property modeling enables practical improvements in MPO where specific motifs in a molecular structure can be identified and modified to enhance ADMET properties. We demonstrate the effectiveness this technique by applying SHAPBALS to optimize hit compounds identified by Revenir™ in our internal discovery of an allosteric activator which rescues functionality of mutant Glucocerebrosidase (GCase), a biological target implicated in Parkinson's Disease.

Drug Discovery Through Revenir™

01 Reveal

02 Dynamic Pocket Hunting

03 Achieve Congruence

04 Hit Validation

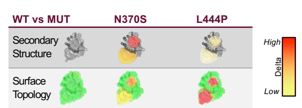
Step 1: Reveal
- Compare WT GCase to pathogenic variants to reveal hypotheses for structural / functional relationships
- Generate Revenir™ biophysical descriptors & fingerprints for conformational ensembles
- Analyze thermodynamic stability, dynamics and surface topological differences between wild-type and mutants

Step 2: Dynamic Pocket Hunting
- Identify binding sites in dynamic ensembles
- Generate binding site pharmacophores from ensembles
- Execute virtual screen; filter and prioritize hits to predict target engagement

Step 3: Achieve Congruence
- Evaluate corrector potential of small molecule hits

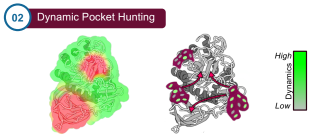
Step 4: Test for experimental hit validation
- Perform GCase trafficking and functional assays

01 Reveal



- Reveal ensemble differences between WT and pathogenic variants to build ML assisted Reversion score for small molecule correction
- Changes in active site and Sap-C site in N370S are consistent with deficient GCase activity
- L444P analysis reveals surface topology changes especially in IgG domain critical for GCase trafficking and stability.

02 Dynamic Pocket Hunting

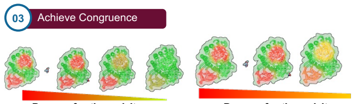


Topology defects in L444P

Residues impacting dynamics

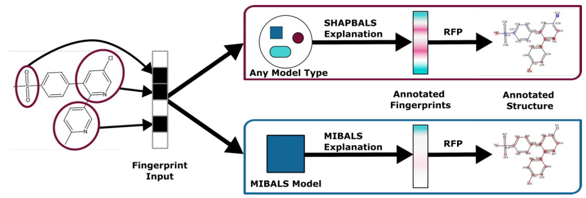
- Revenir™ identifies cryptic and novel pockets in the conformational ensemble of WT and variants which are not found the static structure
- Using an ML-based approach each residue in the ensemble is scored by its contribution to the overall protein dynamics
- Only pockets consisting of residues which alter dynamics are considered allosteric
- Remaining pockets are rank-ordered by their dynamic coupling to problematic regions from Step 1
- Build pharmacophores for top-scoring pockets for vHTS

03 Achieve Congruence



- Biologically functional correctors are compounds displaying properties beyond target engagement
- Generate conformational ensembles for each predicted protein / ligand complex
- Free-energy calculations are combined with the potential of correcting the deficiencies revealed in step 1 to generate the CTX Reversion Score

SHAPBALS – Method



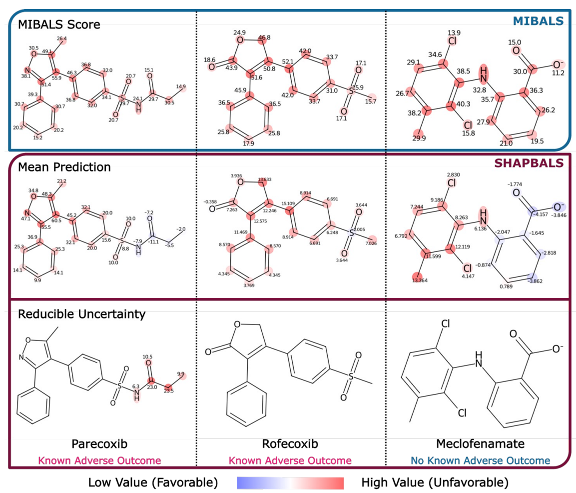
SHAPBALS uses SHAP [2] to attribute changes in the presence of input bits to changes in model output. The computed SHAP values can be used as a 'drop-in' replacement of the per-bit scores used by MIBALS.

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

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 Monte-Carlo methods to make this calculation tractable.

SHAPBALS – Results

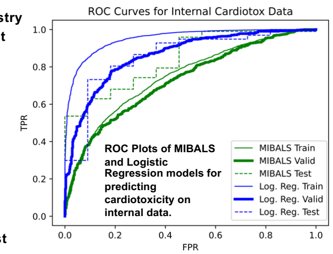
To showcase SHAPBALS annotations on non-proprietary structures, MIBALS and Gaussian Process (GP) models were created from the Side-Effect Resource (SIDER) database [3] to predict if a compound is associated with cardiotoxic / stroke adverse outcomes. Molecular representations are concatenated RDKit [4]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

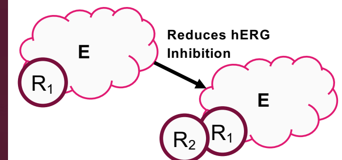


Comparison of MIBALS (Linear Model) and SHAPBALS (Gaussian Process) RFP outputs applied to COX2 inhibitors. Each row corresponds to a model output. *In the last row, only atoms which increased the GP uncertainty were labeled.*

SHAPBALS Enabled Compound Optimization

- Revenir™ + traditional medicinal chemistry produced Cmpd. A which was potent but also had significant hERG inhibition.
- SHAPBALS annotation identified the molecular component driving hERG.
- Replacement of this hERG component reduced hERG, but also significantly reduced potency.
- Virtual fragment enumeration + ML property prediction delivered Cmpd. B improving efficacy and hERG for modest potency reduction.
- Subsequent efforts optimized Cmpd. B to obtain C and D.
- SHAPBALS enables use of highly expressive models which are able to internalize experimental SARs and successfully transfer them to other compounds in the series (A-D).





Compound	EC50 (μM)	Emax	hERG inh. @ 10uM
Experimental			
A	6.5	108%	>90%
B	9.2	118%	60%
C	8.8	173%	17%
D	4.0	131%	7%
E	5.2	177%	78%
E-R2	2.7	149%	41%
ML – Predicted			
A-R2	17.8	112%	81%
B-R2	18.8	103%	34%
C-R2	13.7	114%	10%
D-R2	13.1	125%	5%

Conclusions

The SHAPBALS method improves upon on MIBALS, a previously known method for identifying and labeling the individual atoms which participate in SARs for molecular property modeling. SHAPBALS serves as a broadly applicable, model-agnostic technique which can project predictions back onto an input structure to provide a structural explanation for a prediction. We demonstrate the utility of this technique by showcasing its use in our internal project to develop a compound which provides functional rescue of GCase mutants. By enabling explainability for more accurate and generalizable models, SHAPBALS is a valuable tool for medicinal chemists and reduces the number of design cycles in molecular optimization by identifying specific motifs responsible for desired and undesired properties. This enables the rational design of new iterations of compounds by reducing the need for exploratory modifications necessary to identify liability sites while also suggesting avenues to preserved desired activities in a drug candidate. SHAPBALS is a powerful new tool which allows for accurate, atom-level property modeling for the efficient and rational design of drug-like molecules.

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
MPO – Multi-Parameter Optimization
RFP – Reverse Fingerprint
GP – Gaussian Process
SAR – Structure-Activity Relationship

Congruence Therapeutics Inc.
congruencetx.com