Using Hit Expander and QSAR to design and score new ligands *in silico*

J. Plescia, S. Firth-Clark, G. Tedesco

Cresset, Cambridgeshire, UK jessicap@cresset-group.com https://www.cresset-group.com

Introduction

With the rise of computer-aided drug design in drug discovery and development, there is an increased need for combining automated methods to aid in expediting pharmaceutical research.[1] To efficiently combine *in silico* synthesis with ligand- and structure-based drug design, Cresset has developed and released Hit Expander, a tool to enumerate ligand analogs with small structural or atomic changes for library generation.

We herein present a case study using Hit Expander to design new analogs and subsequently predict the activity of each analog against the target protein. Results are cross-validated using QSAR within Flare™. The combination of these tools can give the scientist more confidence in the potential of their results before spending precious resources conducting time-consuming SAR experiments.

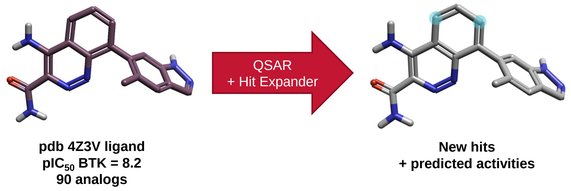


Figure 1: Experiment layout

Method

Dataset

The dataset for this study contains 90 analogs of a Bruton's Tyrosine Kinase (BTK) inhibitor that has been co-crystalized with the target protein, originally taken from a patent.[2] The 90 analogs, ranging in activity from 4 – 9 pIC_{50}, contain variations of both scaffold differences (*e.g.*, indazole to indolinone) and substituent differences (*e.g.*, -H to -OMe). Because Hit Expander focuses on generating small differences *via* functional group additions and atom substitutions, the focus of this experiment is on substituent variations.

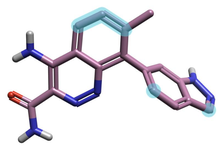


Figure 2: Variations in substitution within the data set

Data preparation

The first step in the study involved confirming that the dataset is amenable to quantitative structure-activity relationship (QSAR) studies in Flare. To do so, we first partitioned the dataset to afford a training set of 63 ligands and a test set of 27(70/30 split) and ensured that an even distribution of activities was present in both data subsets.

Building a QSAR model

With the datasets in hand, we were able to start building the QSAR model. Because the dataset included numerical activity in pIC_{50}, we elected to use the built-in Regression Model builder. The QSAR functionality within Flare includes an automatic option, wherein the program will calculate all available regression models (Gaussian, MLP, *k*-nearest neighbor, Random Forest, SVM) and will choose the best model for the output. The models are built using Cresset 3D descriptors, which use the XED force field to sample the electrostatic potential and volume for each molecule.[3-4] Applying these settings, the dataset was modeled using the SVM regression. The results are shown in Figure 3.

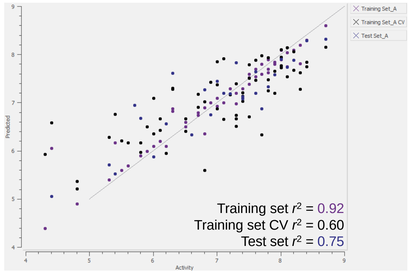


Figure 3: Results of the QSAR (SVM) experiment on the original data set

The regression model output confirmed that the dataset is amenable to QSAR analysis: the correlations for training set cross-validation and test set were 0.60 and 0.75, respectively.

Running Hit Expander

Once we knew the data set was amenable to QSAR analysis, a set of new designs were generated using Hit Expander. Starting from the co-crystallized ligand in the PDB structure, 4Z3V (Figure 1), we ran Hit Expander to automatically generate variants with very small changes. Based on the analogs in the dataset, we chose Me, F, Cl, and OMe. Figure 4 shows the Hit Expander options in Flare.

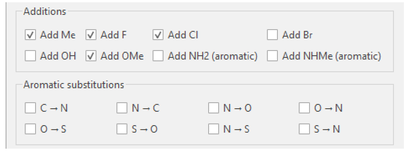


Figure 4: Hit Expander in Flare. A selection of substituent additions or aromatic atom substitutions can be chosen.

We repeated the process with the compound containing the unsubstituted indazole (pIC_{50} = 8.1, Figure 2). To increase the diversity of the newly generated hits, as well as expand the activity range, Hit Expander was also applied to two more inhibitors: compounds **1** (pIC_{50} = 5.8 and **2** (pIC_{50} = 6.8) from the study with lower activities. These two inhibitors were subjected to virtual methylation, generating more analogs with known activities (Figure 5).

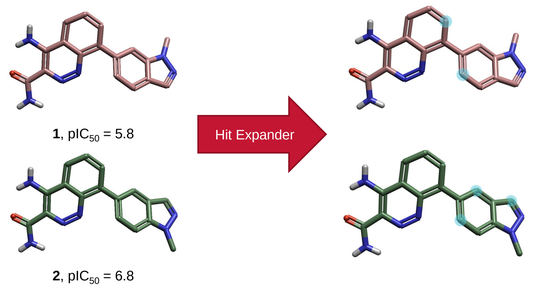


Figure 5. Hit Expander on compounds **1** and **2** (differing in substitution of the indazole), adding Me to various positions.

Results

Observing the new ligands

The Hit Expander function enumerated 72 total analogs from reference compounds in Figure 1, Figure 2, and Figure 5. The selected functional group additions were added to the aliphatic carbon and unsubstituted aromatic carbons. Nineteen of the 72 compounds were duplicates of ligands in the original training and test sets.

Rebuilding the QSAR model

In order to minimize the bias of the new predictions, we removed the 19 duplicate ligands from the original training and testing sets and rebuilt the QSAR model (Figure 6).

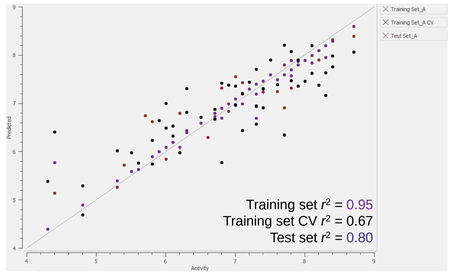


Figure 6: QSAR model (SVM) rebuilt with modified datasets

As indicated in Figure 6, the change in the QSAR model is very minor and therefore still reliable. With the new model, the newly generated ligands can be directly scored to obtain their predicted activities. Hit Expander automatically outputs results that are aligned to their reference ligands with quick minimizations. However, in order to score the new hits, the ligands must first be assigned Cresset XED shape and electrostatic descriptors.[3-4] These descriptors, summarized into *field point* spheres, were calculated and are shown in Figure 7.

We can see the tight clustering of common functional groups, *e.g.*, positive electrostatic field points on the primary aniline, as well as negative electrostatic field points corresponding to the exocyclic amide oxygen, and hydrophobic field points corresponding to the chlorines.

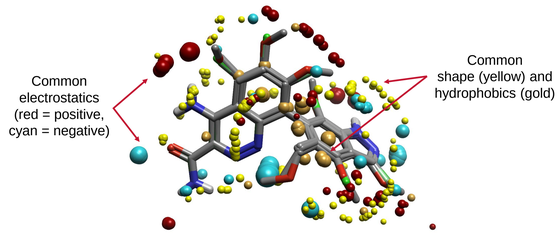


Figure 7. Hit Expander matches aligned to the reference ligand. Clusters of specific field points highlighted.

Scoring the newly generated hits

The predicted activities for the new analogs are promising, with 54% of them having $pIC_{50} \geq 7.5$. The predicted activities of the 19 known ligands are plotted against their known activities in Figure 8. With a correlation of 0.62, the QSAR predictive model is reliable.

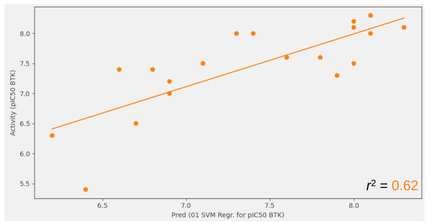
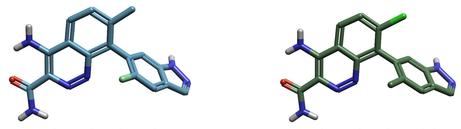


Figure 8: Known vs predicted activity for Hit Expander matches.

With this predictive model in hand, there are several new potentially active analogs enumerated by Hit Expander. Below are two compounds with predicted pIC_{50} > 8.



Variant #6, pIC_{50} = 8.1 **Variant #9**, pIC_{50} = 8.2

Conclusion

Using a combination of Flare functions, including QSAR, XED descriptors, and new analog generation with Hit Expander, we are able to build a reliable model to reasonably predict the activity of known ligands in the dataset, as well as predict the activity of unknowns. This workflow can be similarly applied to ligand-based drug design projects wherein the protein structural information is unavailable.

References

1. Brogi, S. *et al. Front Chem* **2020**, 8, 612.
2. US patent 2015/0038510
3. Vinter, J.G. *J. Comp. Aided. Mol. Des.* **1994**, 8, 653-668.
4. Cheeseright, T. *J. Chem. Inf. Model.* **2006**, 46, 665-676

© Cresset®