

Supplemental Information

Open Source Bayesian Models: I. Application to ADME/Tox and Drug Discovery Datasets

Alex M. Clark^{1*}, Krishna Dole², Anna Coulon-Spector², Andrew McNutt², George Grass³,
Joel S. Freundlich^{4,5}, Robert C. Reynolds⁶ and Sean Ekins^{2,7*}

¹ Molecular Materials Informatics, Inc., 1900 St. Jacques #302, Montreal H3J 2S1,
Quebec, Canada

² Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA
94010, USA

³ G2 Research, Inc., PO Box 1242, Tahoe City, CA 96145

⁴Center for Emerging & Re-emerging Pathogens, Division of Infectious Diseases,
Department of Medicine, Rutgers University-New Jersey Medical School, Newark, New
Jersey 07103, United States

⁵Department of Pharmacology & Physiology, Rutgers University-New Jersey Medical
School, Newark, New Jersey 07103, United States

⁶University of Alabama at Birmingham, College of Arts and Sciences, Department of
Chemistry, 1530 3rd Avenue South, Birmingham, AL 35294-1240, USA.

⁷ Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526,
USA

*Corresponding authors: E-mail: aclark@molmatinf.com and E-mail:
ekinssean@yahoo.com Phone (215) 687-1320

Running title: Open Source Bayesian Models I.

Supplementary information 1. A Description how to use CDD Models and illustrating its intuitive design

The following briefly describes CDD Models as implemented at the time of writing. This may change as the software is developed. The "build model" icon is in the search results toolbar after a search is performed.



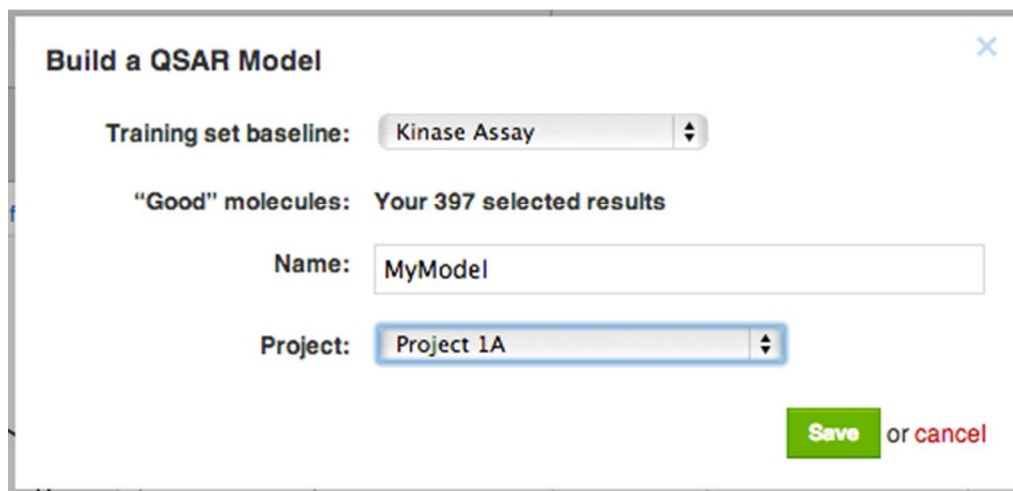
The model is built on two collections of molecules: The "good" molecules- the ones you'd like the model to consider as positives, and the "baseline" molecules, which are a large number of other types of molecules. Good molecules and baseline molecules do not necessarily mean active or inactive molecules in an assay, but the good set should include structures you wish to train the model to recognize amongst a large set of different kinds of structures represented by the baseline set.

Training set baseline molecules may come from the same protocol as the "good" set, a different protocol in the project, or a protocol in a CDD Public data-set. All of the molecules in the selected protocol will become part of the training baseline. This set should ideally contain many different kinds of molecules, including some of the "good" ones.

When a search is performed, all of the molecules in the search results will be considered part of the "good" set by default. For example, you can perform a search for hits of your assay with $IC_{50} < 10 \mu M$. You can also cherry-pick the good molecules from

the search results by using the selection tool on the left of every structure: molecules are all selected by default, and you can click on the check-box to unselect it.

Fill in the model form



Build a QSAR Model

Training set baseline: Kinase Assay

"Good" molecules: Your 397 selected results

Name: MyModel

Project: Project 1A

Save or cancel

The model will also need a name, and a project where this model will be created.

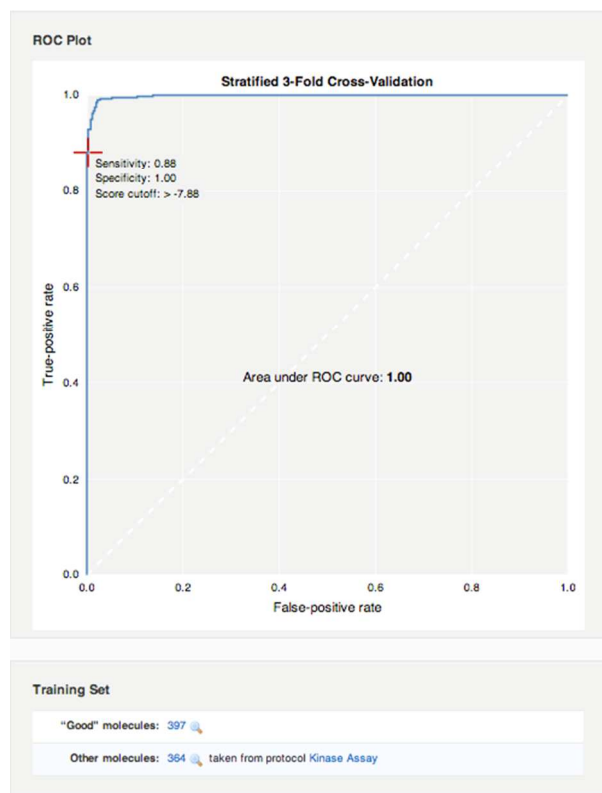
Project selection is important, because once the model is created, all of the molecules in this project will automatically be scored. When the model is saved, you a green status bar with a hyperlinked model name appears. The user can click on the link to analyze results.

Successfully saved your QSAR model as ["MyModel"](#)

The model is created as a special kind of protocol in the project selected. The new protocol is created automatically when a model is saved. The protocol will not be

editable, and will have a "Machine-Learning model" category, 3 readout definitions, and will have one automatically created run for scoring all molecules in the selected project.

ROC curves are graphic representations of the relationship existing between the sensitivity and the specificity of a statistical test. It is generated by plotting the fraction of true positives (sensitivity) out of the total actual positives versus the fraction of false positives out of the total actual negatives (specificity). The user can gauge the effectiveness of the model by examining the curve: the best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The diagonal line represents random guessing. Points above the diagonal represent good classification results (better than random), points below the line poor results (worse than random).



Right below the ROC curve, the user can review the molecule sets that went into model creation.

The protocol also contains readout definitions to store the predicted ranking of molecules that have not been screened yet. As a model is created, it is assigned to a particular project. The model is applied to every molecule belonging to that project, and the results can be seen as a run of the machine learning model protocol. The user can click on the magnifying glass to view the results.

Project 1A				
Run date	Lab	Person	Conditions	Molecules
2014-05-09		Terry Jones		2383

The “score” assigned to a molecule by a model is a relative predictor of the likelihood of the molecule being a true positive according to the model: the higher the score, the higher the likelihood. To get an idea of the range of scores, the user can sort the score column by clicking on the header in the search results table. The user can click again to sort from highest number to lowest. This can help filter the molecules to show only high values.

The image shows a search filter interface with the following elements:


- A dropdown menu labeled "In" with a downward arrow.
- A text input field containing "MyModel" with a downward arrow.
- A green arrow icon followed by the text "with run".
- A dropdown menu labeled "specific run" with a downward arrow.
- A dropdown menu labeled "2014-05-09 (Terry Jones)" with a downward arrow.
- A green arrow icon followed by the text "with readout definition".
- A dropdown menu labeled "Score" with a downward arrow.
- A dropdown menu labeled ">" with a downward arrow.
- A text input field containing "10.0".


“Applicability” is the fraction of structural features shared with the training set of molecules.

“Maximum Similarity is the Tanimoto/Jaccard similarity to any of the "good" molecules in the training set. This value is independent of the Bayesian model, and provides a way for to perform a similarity search using all active compounds at once.

To apply the model to other untested molecules in another project in the user’sCDD Vault, the protocol can be shared with that project. The underlying training set data used in the creation of the model will not be disclosed to the destination project if sharing permissions do not allow it. Nevertheless, all of the molecules belonging to the destination project will be automatically scored, and a new run with results will be added to the model protocol. The process may take several minutes, depending on the number of molecules to score, but when the yellow status bar at the top of the page disappears, all of the molecules will have been scored.

The user can then look on the "Run data" tab of the model protocol, and click the magnifying glass icon as before.

Run date	Lab	Person	Conditions	Molecules
2014-05-09		Terry Jones		417 

Run date	Lab	Person	Conditions	Molecules
2014-05-09		Terry Jones		2383 

The current version of the modeling functionality does not automatically score all new compounds which are added to a project after the model was last applied. To apply the model to new molecules, you will need to manually un-share and re-share the model with the current project. Here are the general steps:

1. Make sure the model is shared with more than one project. You risk deleting the entire model if it's only shared with the one project you're interested in- you will delete this sharing in step 2.
2. Once the model is shared with several projects, the user can go to the Run data tab of the model protocol, and delete the old run for your project. The user should

click on the Run date, go to the Run Details tab, and "Delete this run" in the side-bar.

3. Back on the Protocol Details form the user should - un-share the model with the project of interest. Finally the user should go to "Manage project access" click in the side-bar, and remove access from the project.
4. Add back the access to this project just removed in step 3. This will cause the model to run against all the compounds in the project, and a new run of the modeling protocol to be created from scratch.

'Collections' are named lists of compounds that facilitate compound library management, and complex query building. Collections can also be used to create an extensive training set of inactives. A collection can include both privately stored and publically shared compounds, thus expanding the range of molecules that can be used to train a model to include the entire public data space of CDD if needed.

Figure S1. An example of a large whole cell screen against *Mycobacterium tuberculosis* dataset (MLSMR) used to generate a bayesian model in CDD Models.

