Supplementary Information

1/ Installation and usage of the python script

The script was written in python 2.7.10 but probably runs on any python 2.6 and plus. Not on python 2.3.

Dependencies:

The script requires:

- a recent **RDKit** version
- a recent **scikit-learn** version
- **NumPy**

The training set:

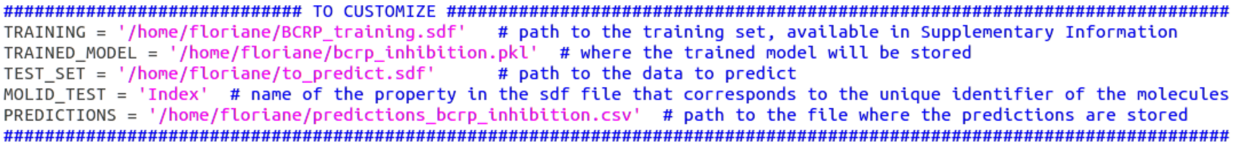
The training set is given as an sd file, named "BCRP_training.sdf". It should contain 978 compounds annotated for their BCRP inhibition (under the section "Activity"). Please refer to "BCRP inhibition: from Data Collection to Ligand-Based Modeling", by Montanari and Ecker, *Molecular Informatics*, 2014 (DOI: 10.1002/minf.201400012) for more details on how the data was collected.

The test set:

We call "test set" the dataset for which you want to obtain a BCRP inhibition prediction. It should be reasonably cleaned (removing salts, standardizing the structure, removing compounds with rare atoms, etc.) since the workflow takes the compounds as-is in the given test set. If possible, it should contain a property that defines a molecule identifier, but it is not compulsory. Any other property will be ignored.

Variables to customize:

The script can be viewed in a regular text editor, Gedit in Ubuntu for example. At the beginning of the script, after the imports, you will see the following:



```
############################# TO CUSTOMIZE ###########################################################################
TRAINING = '/home/floriane/BCRP_training.sdf'   # path to the training set, available in Supplementary Information
TRAINED_MODEL = '/home/floriane/bcrp_inhibition.pkl'  # where the trained model will be stored
TEST_SET = '/home/floriane/to_predict.sdf'      # path to the data to predict
MOLID_TEST = 'Index'  # name of the property in the sdf file that corresponds to the unique identifier of the molecules
PREDICTIONS = '/home/floriane/predictions_bcrp_inhibition.csv'  # path to the file where the predictions are stored
######################################################################################################################
```

This part is the only part you have to edit. For “TRAINING”, replace the pink string by the path to the training set (once you have downloaded it). For “TRAINED_MODEL”, replace the pink string by the path to where you want the model to be stores / where the model is stored (if it has already been trained). For “TEST_SET”, replace the pink string by the path to the sd file containing the compounds you want to predict. For “MOLID_TEST”, replace the pink string by the name of the property that contains the index of the compounds. If the dataset does not contain any index (or molecule identifier), replace the pink string by the word “None” (without quotes). For “PREDICTIONS”, replace the pink string by the path to the file where you want to save the predictions for the test compounds.

How to run the script:

Once all dependencies are installed and the custom variables are properly set, go to the directory where the script is stored and in a terminal write:

```
>> python BCRP_inhibition_model.py
```

2/ Cross-validation and leave-sources-out validation results for the 16 models

Table SI-1:

Learning method	LSO AUC ROC[a]	10-fold CV AUC ROC[b]
MACCS, Naïve Bayes	0.58	0.65
MACCS, logistic regression	0.62	0.83
MACCS, Random Forest	0.65	0.88
MACCS, SVM	0.68	0.88
CDK, Naïve Bayes	0.63	0.67
CDK, logistic regression	0.70	0.85
CDK, Random Forest	0.71	0.87
CDK, SVM	0.71	0.77

ECFP, Naïve Bayes	0.56	0.78
ECFP, logistic regression	0.71	0.90
ECFP, Random Forest	0.65	0.86
ECFP, SVM	0.73	0.90
VolSurf, Naïve Bayes	0.69	0.69
VolSurf, logistic regression	0.72	0.80
VolSurf, Random Forest	0.61	0.77
VolSurf, SVM	0.64	0.73

[a] Area under the ROC curve in the "leave-sources-out" validation setting, average over 166 experiments

[b] Area under the ROC curve in 10-fold cross-validation

3/ Characterization of the PLB985 cells stably expressing BCRP

Materials and methods

Western blot analysis

Cells were harvested and lysed in lysis buffer (50 mM Tris pH8, 120 mM NaCl, 1 mM EDTA, 2% Triton X-100) containing protease inhibitors (Complete Protease Inhibitor Cocktail Tablets, Roche Diagnostics, Indianapolis, IN). Cell debris was pelleted by centrifugation (1000 g, 5 min, 4°C), and supernatant, containing 20 µg protein/sample, was mixed with sample buffer (reaching final concentrations of 8% glycerol, 0.8% SDS, 0.01% bromophenol blue, 5% 2-mercaptoethanol). Samples were separated on 8% SDS polyacrylamide gel, and then electrophoretically transferred onto a nitrocellulose blotting membrane (GE Healthcare Life Sciences, Freiburg, Germany). The membranes were blocked with 5% BSA in TBS (25 mM Tris, 140 mM NaCl, 2.5 mM KCl, pH 7.4) and incubated with BXP-21 mouse anti-BCRP (Santa Cruz Biotechnology, CA, USA) or β-Actin (D6A8) Rabbit mAb (Cell Signaling Technology MA, USA) antibodies diluted 1:1000, overnight at 4°C. IRDye 800CW goat anti-mouse IgG and IRDye 680 goat anti-rabbit IgG (LI-COR Biotechnology, Homburg, Germany), diluted 1:10000, were used as secondary antibodies and were added for 45 min incubation at room temperature. All antibodies were diluted in 5% BSA in TBS-T (TBS with 0.1% Tween 20). Fluorescence was detected on LI-COR Odyssey® CLx Imager (LI-COR Biotechnology) at 800 nm and 700 nm, respectively.

Results

To confirm expression of BCRP in PLB985 cells stably expressing BCRP, Western blot analyses were performed (Figure SI-1A) showing BCRP protein expression only in the BCRP expressing PLB985 cells but not in the parental cell line. Functionality of BCRP in overexpressing PLB985 cells was verified by the steady state mitoxantrone accumulation assay (Figure SI-1B). As the efflux of mitoxantrone is mediated by BCRP, BCRP overexpressing PLB985 cells show significantly decreased mitoxantrone accumulation compared to the parental cell line, confirming functionality of BCRP in overexpressing PLB985 cells. Addition of Ko143, a known BCRP inhibitor, to BCRP-overexpressing PLB985 cells increased accumulated mitoxantrone levels similar to that observed in parental PLB985 cells (Figure SI-1B). On the contrary, mitoxantrone accumulation in parental PLB985 cells was not affected by Ko143, confirming the absence of any endogenous mitoxantrone transporter sensitive to Ko143. Furthermore, IC_{50} measurements for Ko143 (Figure SI-1C) were conducted in BCRP-overexpressing PLB985 cells, giving an IC_{50} value of 9.8 ± 1.6 nM, which is in accordance to the previous published IC_{50} value of Weiss, J. et al. (10 nM; Weiss, J., Rose, J., Storch, C. H., Ketabi-Kiyanvash, N., Sauer, A., Haefeli, W. E., & Efferth, T. (2007). Modulation of human BCRP (ABCG2) activity by anti-HIV drugs. The Journal of Antimicrobial Chemotherapy, 59(2), 238–45).

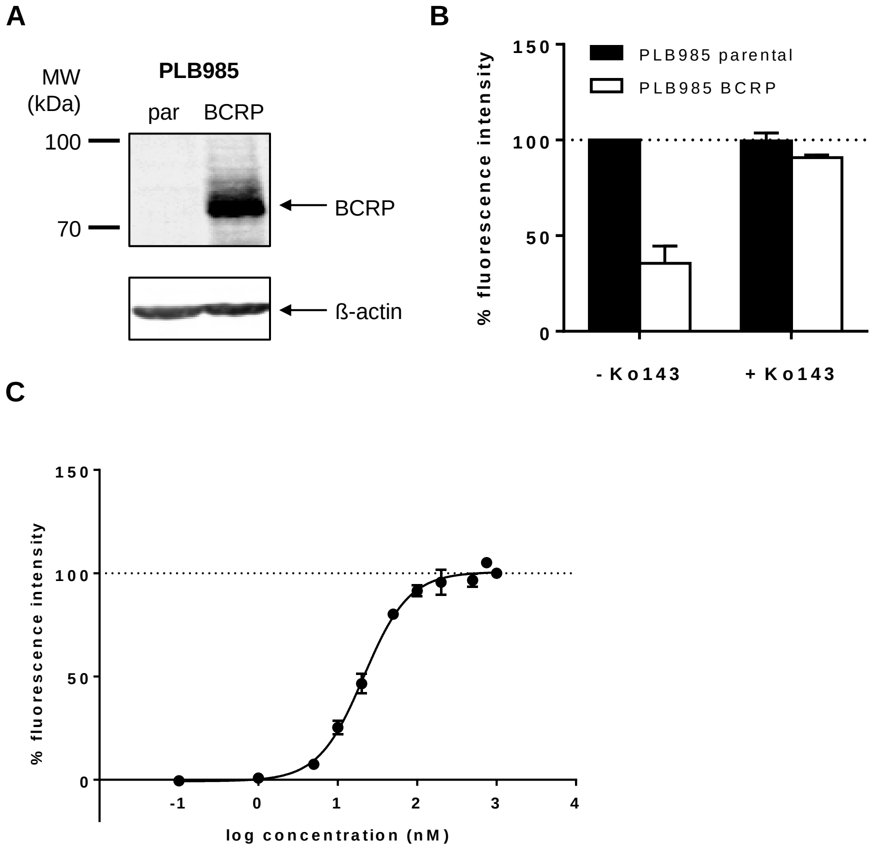


Figure SI-1: Characterization of BCRP-overexpressing PLB985 cells.

A. Expression analysis of BCRP in parental (par) and BCRP stably expressing PLB985 cells at the protein level using Western blot. To dissociate the BCRP complex into monomers cell lysates were treated with 5% 2-mercaptoethanol. The BCRP monomer band is detected at 72 kDa. ß-actin served as a loading control. The figure shows a representative Western blot of two independent experiments.

B. Verification of BCRP function in BCRP overexpressing PLB985 cells compared to parental PLB985 cells. Steady state accumulation of mitoxantrone (7 µM) was measured in the absence (-Ko143) or presence of 1 µM Ko143 (+Ko143) as described in the materials and methods section of the original article, except fluorescence intensity measurement, which was done on BD FACSCalibur flow cytometer (Becton Dickinson, San Jose, CA, USA). Data show the mean percentage fluorescence intensity after subtracting the background fluorescence of unstained cells and subsequent normalization to parental cells without Ko143 treatment, which was set to 100%, and ± SD of 2 independent experiments. Each experiment was performed in technical duplicates.

C. $IC_{50apparent}$ measurement of Ko143. Steady state accumulation of mitoxantrone (7 µM) in the absence and presence of 11 different concentrations of Ko143 ranging from 0.1 to 1000 nM was measured as described in the materials and methods section of the original article, except fluorescence intensity measurement, which was done on BD FACSCalibur flow cytometer (Becton Dickinson, San Jose, CA, USA). Data given here show the mean percentage fluorescence intensity after subtracting the background fluorescence of unstained cells and the fluorescence of the DMSO control and subsequent normalization to the fluorescence intensity at the highest Ko143 concentration, which was set as 100%, ± SD of 3 independent experiments.

4/ Sensitivity analysis of the logistic regression model

The sensitivity analysis performed here relates to the uncertainty of the source-by-source threshold assignment that was chosen and described in Montanari and Ecker, *Molecular Informatics*, 2014. The model built depends on the input *X* (feature matrix) and *y* (labels vector). *Y* in turn depends on the thresholds initially applied. In this analysis, we try to simulate the effect of changing the thresholds on the output model.

Methods

The sensitivity analysis is performed with the following restraints: the number of compounds in the training set will not vary (which concretely means that we will ignore potential label discrepancies arising from the new thresholds for compounds measured in several sources) and the model settings will not vary (which concretely means that we will ignore the fact that for some *y* the ideal model may not be a logistic regression).
To evaluate the effect of changing the thresholds and class assignments, we look at the DrugBank screen results.

For each source in the training set, a range of possible thresholds was defined that would be sensible in the context of the experiment and reported end-point. Table SI-2 reports these ranges.

Table SI-2: Possible values taken by the new thresholds in the sensitivity analysis for each source

Source	**End point unit**	**Initial threshold**	**Lowest allowed**	**Highest allowed**
Pick_2008	pIC_{50} (M)	4	3	6
Pick_2010	IC_{50} (µM)	25	5	35
Pick_2011	IC_{50} (µM)	10	5	25

Ahmed-Belkacem_2005	IC$_{50}$ (μM)	10	5	25
Ahmed-Belkacem_2007	% inhibition	50	40	70
acridones_Boumendjel_2007	IC$_{50}$ (μM)	15	5	30
Boumendjel_2005	fluorescence intensity	200	150	300
Matsson_2007	fold increase	3	2	5
Saito_2006	% inhibition	20	15	70
cdkinhib_An_2008	IC$_{50}$ (μM)	25	5	35
Katayama_2007	RI$_{50}^{-1}$ (μM^{-1})	0.1	0.01	6
Loevezijn_2001	fluorescence intensity	150	90	210
Juvale_2012	IC$_{50}$ (μM)	15	5	30
phenylquinazolines_Juvale_2012	IC$_{50}$ (μM)	10	5	25
Jin_2006	IC$_{50}$ (μM)	10	5	25
Cramer_2007	EC$_{50}$ (μM)	10	5	25
Imai_2004	degree of resistance	10	5	15
Sugimoto_2003	reversal index	1.15	1.1	2
flavonoids_Zhang_2004	Substrate accumulation	300	200	400
flavonoids_Zhang_2005	EC$_{50}$ (μM)	15	5	30
Colabufo_2008	EC$_{50}$ (μM)	10	5	25
Colabufo_2008_ext	IC$_{50}$ (μM)	10	5	25
Holland_2007	fluorescence intensity	25	15	35
Ivnitski-Steele_2008	percent inhibition	50	40	70
Ivnitski-Steele_2010	rank score	52	40	70
Njus_2010	toxic dose / reversal index	10	5	30
Xiao-Ling_2008	IC$_{50}$ (μM)	10	5	25
Zembruski_2011	PubChem annotation	None	None	None
Hacker_2009	IC$_{50}$ (μM)	15	5	30
Arnaud_2010	percent inhibition	50	40	70
Jimenez-Alonso_2008	PubChem annotation	None	None	None
Kuhnle_2009	IC$_{50}$ (μM)	15	5	30
Ochoa-Puentes_2011	IC$_{50}$ (μM)	15	5	30
Ali-Versiani_2011	IC$_{50}$ (μM)	10	5	25
Bokesch_2010	IC$_{50}$ (μM)	18	5	30
Takada_2010	IC$_{50}$ (μM)	20	10	35
Feng_2008	IC$_{50}$ (μM)	5	3	15
Feng_2009	IC$_{50}$ (μM)	5	3	15
Giannini_2008	IC$_{50}$ (μM)	15	5	30

Mao_2004	IC$_{50}$ (µM)	15	5	30
Pan_2013	IC$_{50}$ (µM)	15	5	30
Marighetti_2013	IC$_{50}$ (µM)	15	5	30
Wang_2008	IC$_{50}$ (µM)	25	10	40
Matsson_2009	percent inhibition	50	40	70
Patel_2011	IC$_{50}$ (µM)	25	10	40
Weiss_2007	IC$_{50}$ (µM)	30	15	45
Curtis_2007	IC$_{50}$ (µM)	10	5	25

The sensitivity analysis was then performed by repeating 2000 times the following experiment:

A random number of sources (between 2 and 30) are picked for threshold change. For each of these sources, the threshold is randomly chosen within the range proposed in Table SI-2. The new thresholds are then used to derive the vector *y* of training labels. The logistic model described in the Methods section of the manuscript is then rebuilt using these new training data. The DrugBank set is then passed through the new model, and scores are kept.

Results

The impact of changing the thresholds in the training set was evaluated on the DrugBank screen results. First, the ranking of the compounds obtained in each experiment was compared with the initial ranking by means of Spearman correlation coefficient. The distribution of Spearman coefficients is shown in Figure SI-2.

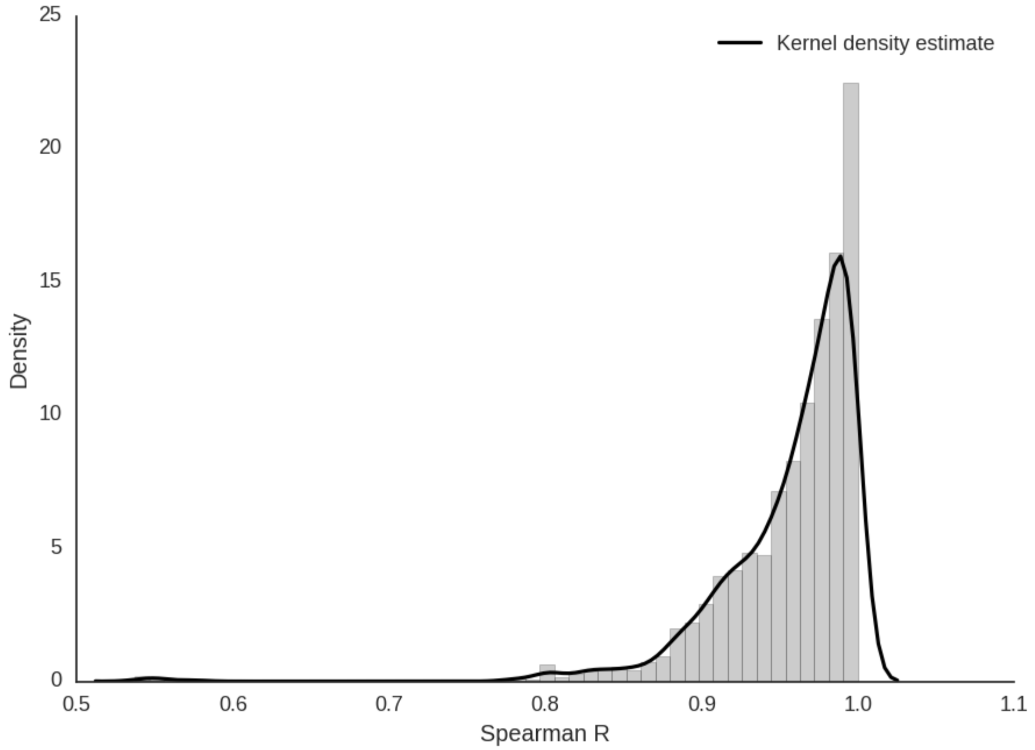


Figure SI-2: Distribution of Spearman coefficient correlations for the 2000 rankings of DrugBank in the sensitivity analysis compared with the initial ranking.

Most of the experiments led to a Spearman R over 0.9, which means that the obtained rankings are very close to the original ranking.

Next, the impact of the thresholds in the training set on the scores obtained for the 10 compounds that were selected for testing is shown in Figure SI-3.

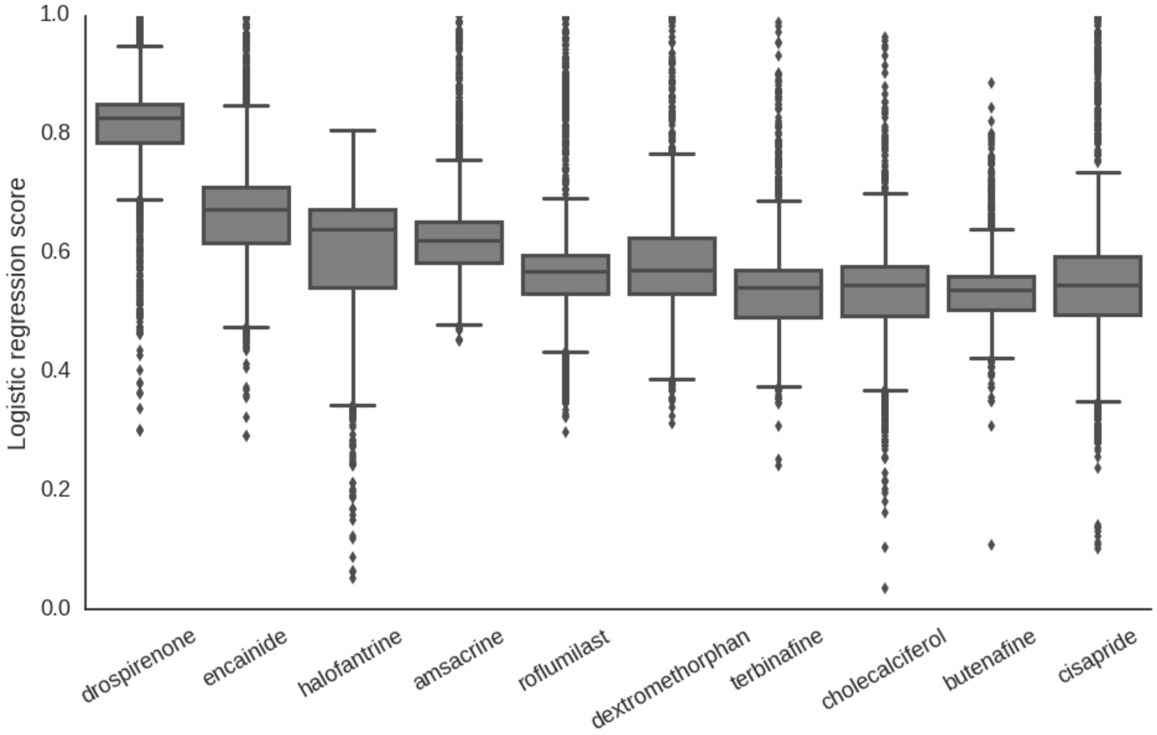


Figure SI-3: Boxplot of the scores obtained by the 10 tested compounds across the 2000 experiments of the sensitivity analysis.

The scores varied quite a bit for some experiments, but the overall population of scores is centered on the original score for each compound (see Table 2 in the main manuscript).

Finally, as a proof that the threshold ranges chosen actually had an effect on the training labels, we propose in Figure SI-4 a distribution of the proportions of labels affected by changes in the 2000 experiments. We see that in most cases almost half of the labels are actually affected.

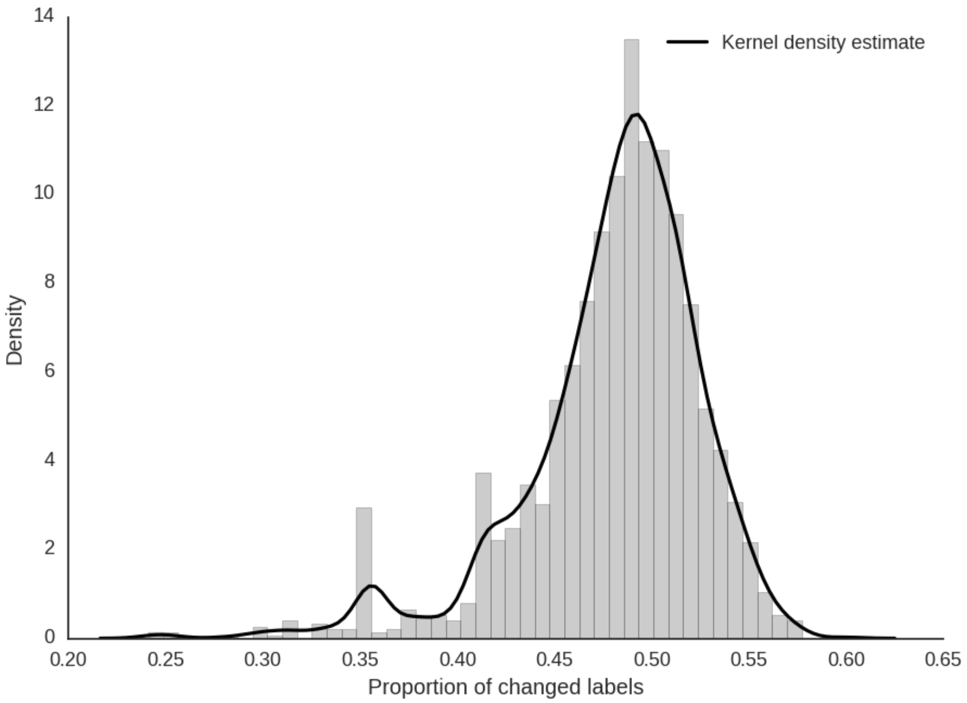


Figure SI-4: Distribution of proportion of labels affected by changes across the 2000 experiments of the sensitivity analysis.