JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

eCounterscreening: Using QSAR Predictions to Prioritize Testing for Off-Target Activities and Setting the Balance between Benefit and Risk

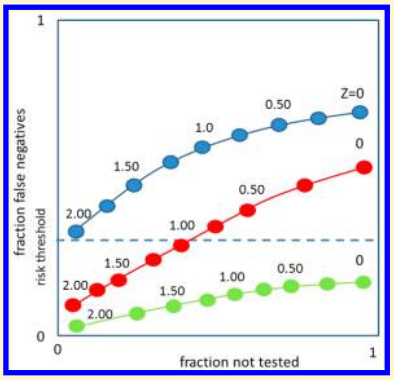
Robert P. Sheridan,*[,†] Daniel R. McMasters,[‡] Johannes H. Voigt,[†,∥] and Mary Jo Wildey[§]

[†]Structural Chemistry, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065, United States

[‡]Structural Chemistry, Merck Research Laboratories, 33 Avenue Louis Pasteur, Boston, Massachusetts 02115, United States

[§]In Vitro Pharmacology, Merck Research Laboratories, 2000 Galloping Hill Road, Kenilworth, New Jersey 07033, United States

ABSTRACT: During drug development, compounds are tested against counterscreens, a panel of off-target activities that would be undesirable for a drug to have. Testing every compound against every counterscreen is generally too costly in terms of time and money, and we need to find a rational way of prioritizing counterscreen testing. Here we present the eCounterscreening paradigm, wherein predictions from QSAR models for counterscreen activity are used to generate a recommendation as to whether a specific compound in a specific project should be tested against a specific counterscreen. The rules behind the recommendations, which can be summarized in a risk−benefit plot specific for a counterscreen/project combination, are based on a previously assembled database of prospective QSAR predictions. The recommendations require two user-defined cutoffs: the level of activity in a specific counterscreen that is considered undesirable and the level of risk the chemist is willing to accept that an undesired counterscreen activity will go undetected. We demonstrate in a simulated prospective experiment that eCounterscreening can be used to postpone a large fraction of counterscreen testing and still have an acceptably low risk of undetected counterscreen activity.



INTRODUCTION

In the pharmaceutical industry, compounds must be active on the target on which they are intended to act but also lack activity on a number of targets for which activity would be undesirable. The collection of assays for those latter targets are known as "counterscreens". Strong activity on any counterscreen might be enough to keep a compound from advancing further because of potential toxicity or unexpected changes in drug metabolism. For example, most pharmaceutical companies screen for hERG ion channel activity, since binding to that channel can cause heart arrhythmias.[1,2] Inhibition of other ion channels such as $Ca_v1.2$ and $Na_v1.5$ can also result in cardiac toxicity.[3] Similarly, inhibiting cytochrome P450s could cause drug−drug interactions.[4,5] There is also concern about mechanism-specific inhibition of cytochrome P450 3A4[6] and the induction of 3A4 by the pregnane X receptor and other nuclear hormone receptors.[7−9]

An important question is when to submit a molecule to a counterscreen. One approach, which was practiced at Merck, was for a chemist working on a therapeutic project to choose a set of counterscreens to be run simultaneously on each molecule at the time of registration. Not every molecule was tested on every counterscreen, but typically several counterscreens were chosen per molecule. This approach has the advantage that problematic counterscreen activity would be detected very early. However, much of the time the counterscreen results provided little new information, and this system was very costly.

The methodology in this paper is meant to support a different philosophy where unnecessary counterscreening is reduced: For each compound, we should test first on those counterscreens for which the compound is likely to be active; showing a single problematic activity early could stop further futile development of the compound, i.e., it could "fail early". On the other hand, one could postpone testing on counterscreens for which the compound was unlikely to show activity. Ultimately, all compounds that are very far along in development would be tested on every counterscreen, but the number of such compounds is very small compared with the number of compounds registered. If one follows this philosophy, however, one must accept some risk that a compound will have some problematic counterscreen activity that is discovered only after the compound has advanced.

In this paper we present a paradigm called "eCounterscreening" to reduce counterscreening by a substantial amount while keeping the risk of missing some important counterscreen activity to a manageable level. To do this we use predictions from QSAR models of the counterscreens to estimate the risk that a given molecule in a given project will be active on that counterscreen.

Received: November 4, 2014
Published: December 31, 2014

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231−238

METHODS

Our approach follows the scheme shown in Figure 1.

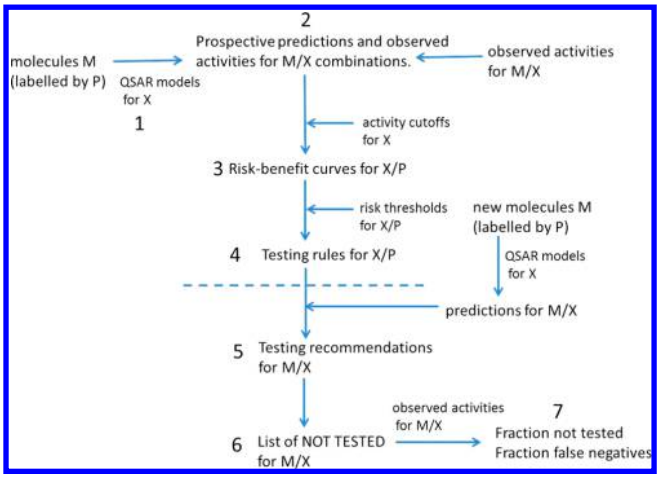


Figure 1. Scheme for the eCounterscreening method.

QSAR Models for Counterscreens (Section 1 in Figure 1). At Merck there are nine assays currently selected as "global counterscreens", i.e., counterscreens that apply to most projects. These are listed in Table 1. Let X denote a counterscreen assay from Table 1 (e.g., HERG), and let P represent an individual therapeutic project (e.g., HIV integrase). At Merck we maintain QSAR models of each X, which are updated every few months using new data. These models are "global" in the sense that each model is built on all of the molecules tested for X, not just molecules tested on a given project P. The number of molecules in each model can be quite large: ~24 000 molecules in the case of 2C8 but >300 000 molecules in the case of HERG.

We have shown that a single global QSAR model for X suffices to predict compounds for all projects P and that there is no advantage in having a separate X model for each P in terms of accurately predicting the X activity for P compounds.[10] However, experience has shown that the accuracy of prediction of the model for X varies among P and even among series within P, and it is necessary to have separate rules for each X/P combination in terms of which molecules should be counterscreened.

In-house models for X are usually built as regressions using the random forest method[11] and the descriptors AP[12] and DP,[13] perhaps with the addition of the MOE_2D descriptors from the MOE modeling package.[14] The details of the modeling are unimportant for our purposes; the eCounterscreening paradigm is useful as long as the models for X are reasonably predictive. In addition to predicted activities, we are able to calculate for each molecule M predicted for X a "prediction uncertainty" (i.e., a 1 standard deviation "error bar" on the prediction, assuming a Gaussian distribution) using domain applicability metrics. Details for calculating these are provided in ref 15. Again, the specific method for calculating the prediction uncertainty is not important. Both the prediction and prediction uncertainty are needed for eCounterscreening.

Database of Prospective Predictions (Section 2 in Figure 1). We keep a historic database of prospective predictions and prediction uncertainties plus observed activities for a number of molecules M on each model X, where each M is assigned a project P upon registration. By "prospective" we mean that the prediction is done on a version of the QSAR model for X created prior to M being measured. This is important, since once M is incorporated into a random forest model for X, its prediction will always be unrealistically accurate. We have accumulated prospective predictions for ~133 000 compounds registered since July 2011, although not all of the compounds have observed activities for all X. The overall accuracies of the models in distinguishing "active" from "inactive" in prospective prediction, as measured by Cohen's κ and percent correctly predicted, are listed in Table 1. By Cohen's κ, the predictive ability is moderate (with chance level being 0 and perfection being 1.0). On the other hand, since the prospective data sets are quite unbalanced (i.e., except for TDI, almost all compounds are predicted and observed to be inactive), the percent correct can be quite high.

Risk–Benefit Curve (Section 3 in Figure 1). For each X/P combination we need to generate a "risk–benefit curve", which shows the level of risk one must accept for a given amount of counterscreening saved. We generate the curve by using the database of prospective predictions and simulating what happens when compounds are "not tested" when predicted to be inactive for X and the prediction has a certain reliability. In particular, we are interested in what fraction of those "not tested" compounds will turn out to be active on X despite the prediction. Here it should be noted that we are using the phrase "not tested" to indicate that we are treating the observed activity of M as not being known for the purposes of the simulation, whereas in reality both the predicted and observed activities of M are in the database. It is assumed that molecules predicted to be active will

Table 1. Counterscreens

name	description	activity threshold	molecules in the model by the end of 2013	descriptors in the model	Cohen's κ (percent correct)[a]
2C8	inhibition of CYP 2C8 in $-\log(IC_{50})$	10 μM	24000	AP, DP	0.49 (81)
2C9	inhibition of CYP 2C9 in $-\log(IC_{50})$	10 μM	180000	AP, DP	0.46 (85)
2D6	inhibition of CYP 2D6 in $-\log(IC_{50})$	10 μM	180000	AP, DP	0.35 (95)
3A4	inhibition of CYP 3A4 in $-\log(IC_{50})$	10 μM	180000	AP, DP	0.44 (92)
HERG	inhibition of the hERG channel in $-\log(IC_{50})$	10 μM	314000	AP, DP	0.51 (78)
CAV	inhibition of the $Ca_v1.2$ channel in $-\log(IC_{50})$	10 μM	108000	AP, DP	0.56 (79)
NAV	inhibition of the $Na_v1.5$ channel in $-\log(IC_{50})$	10 μM	102000	AP, DP	0.51 (82)
TDI	time-dependent (also called mechanism-dependent) inhibition of CYP 3A4 as the log of the ratio of IC_{50} values with and without NADP	ratio of 4	25000	AP, DP	0.43 (74)
PXR	enhancement of 3A4 induction on the pregnane X receptor relative to rifampicin, expressed as a percentage.	40	180000	AP, DP, MOE_2D	0.50 (80)

[a]These two values are measures of the overall accuracy in distinguishing "active" from "inactive".

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231–238

never be "not tested" and as such do not contribute to the risk–benefit curve.

First, we need to define for each X an activity cutoff below which a compound is considered "inactive," i.e., where the counterscreen activity is low enough not to be a concern. For example, for HERG the threshold would be an IC_{50} of 10 μM. Table 1 shows the default cutoffs for Merck's counterscreens. For this analysis, we assume that the cutoff for X applies for all P. We also assume that the same activity cutoff applies equally well to predicted and observed activities; we can make this assumption because we adjust all of the models in such a way that the predicted and observed activities of the training-set molecules are numerically matched.

Given that we have a prediction and a prediction uncertainty for M on X, we define a metric z that represents the reliability of a prediction of inactivity for X:

$$z(\mathrm{M}, \mathrm{X}) = \frac{\text{activity cutoff}(\mathrm{X}) - \text{prediction}(\mathrm{M}, \mathrm{X})}{\text{prediction uncertainty}(\mathrm{M}, \mathrm{X})}$$

A more positive prediction(M, X) implies more activity on X, which we wish to avoid. When $z(\mathrm{M}, \mathrm{X}) = 0$, the prediction is exactly on the activity cutoff. A higher $z(\mathrm{M}, \mathrm{X})$ means that M is more likely to be inactive on X. A series of thresholds Z (0, 0.25, 0.5, 0.75, 1, etc.) is set for the minimum tolerable $z(\mathrm{M}, \mathrm{X})$, i.e., a minimum level of reliability. For $Z = 0$, all compounds predicted to be inactive are "not tested" on X. For $Z = 1$, compounds with $z(\mathrm{M}, \mathrm{X}) \geq 1$ would be "not tested."

A diagram of the analysis of such data is shown in Figure 2 (top). Each circle represents the prospective prediction of a compound M in project P for assay X. All of the molecules have a prediction uncertainty, and we show these for a few examples (narrow horizontal bars). The "fraction not tested" is the ratio of the number of compounds in area B (i.e., predicted inactive) with $z(\mathrm{M}, \mathrm{X}) \geq Z$ (i.e., with a sufficient reliability) to the total number of compounds. This represents a "benefit" in the sense that we avoid testing this fraction of compounds.

Figure 2 (bottom) shows the same diagram, this time divided into two sections by the activity cutoff in the observed activity axis. Molecules in area C represent "not tested" molecules that are predicted to be inactive with a given reliability but turn out to be active in reality. The ratio of the number of "not tested" molecules that turn out active to the total number of "not tested" compounds is called the "fraction false negatives." This represents "risk" in the sense that these compounds have an undesirable X activity that is not realized since we did not test them. It is useful to define the minimum number of "not tested" compounds one must have before one can sensibly calculate the "fraction false negatives". Here we use 50.

One can graph "fraction false negatives" vs "fraction not tested" to make the risk–benefit curve for each X/P combination. It should be remembered that only compounds "not tested" contribute to the curve. In Figure 3 we show curves for three example X's for the same P. Along each curve are several values of Z. The general trend of the curves is to go to the lower left as Z increases. As we raise Z, a smaller fraction of molecules are "not tested" (hence movement leftward), but since the predictions of inactivity are more reliable, the fraction of false negatives gets smaller (hence the movement downward). For each X/P combination, a user can pick the threshold of acceptable risk (the horizontal dashed line), i.e., the fraction of false negatives that is tolerable. A reasonable value might be 0.1.

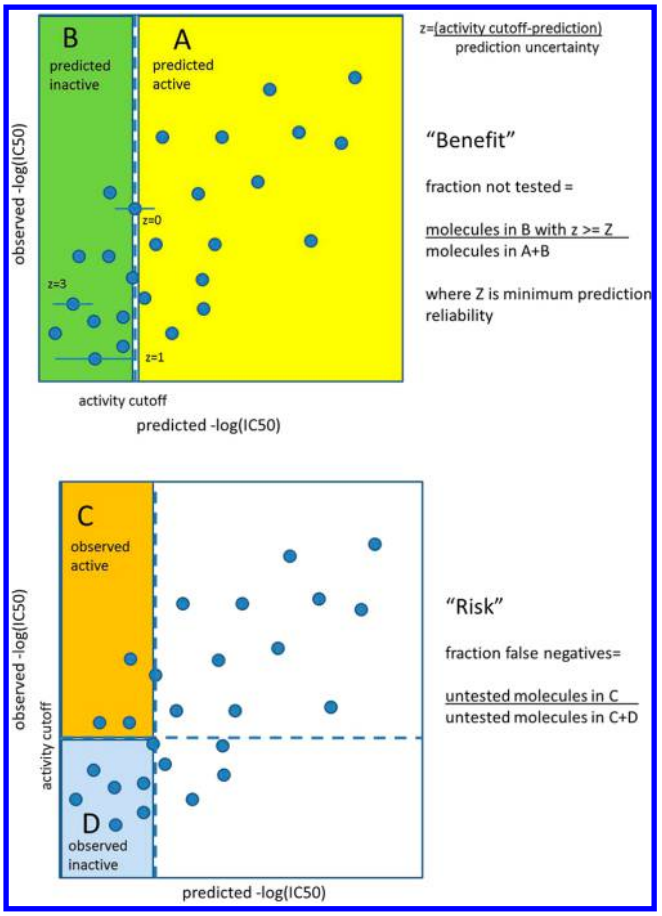


Figure 2. Scheme by which "fraction not tested" and "fraction false negatives" are calculated. Each symbol represents one molecule being predicted. In each plot, higher predicted activity is toward the right and higher observed activity is toward the top.

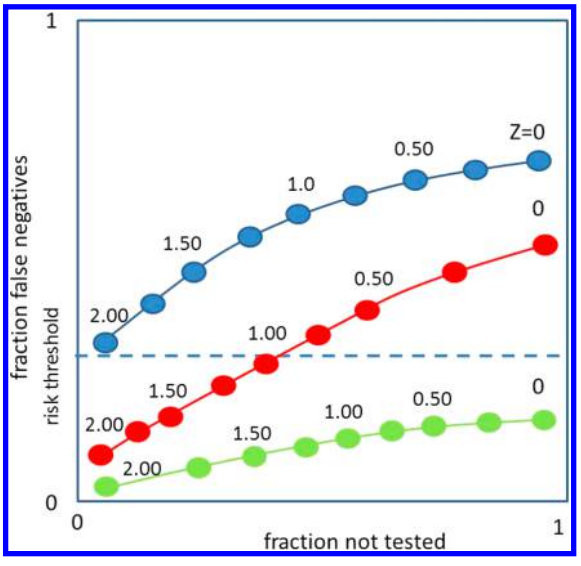


Figure 3. Idealized risk–benefit curves. Each curve represents a different X/P combination, for example, different counterscreens for a given project. Each symbol represents a set of compounds "not tested" at a given level of Z (which are labeled as numbers). The horizontal line represents the user-set risk threshold of having the molecules in the set be active despite being predicted as inactive.

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231–238

Generating Testing Rules from the Risk−Benefit Curve (Section 4 in Figure 1). Given the risk threshold, to maximize the number that can be "not tested", we can look for the lowest Z on the curve that is below the risk threshold. This includes only those values of Z for which there are at least 50 "not tested" molecules. We can assign this value to $R(X, P)$, which serves a testing "rule" for the X/P combination. In our idealized examples in Figure 3, we see that the green curve is always below the threshold; thus, the lowest Z value below the threshold is 0, so $R(X, P) = 0$. This means we can "not test" any M with $z(M, X) \geq 0$, i.e., all M that are predicted to be inactive. For the blue curve, there is no value of Z below the threshold, so $R(X, P)$ is assigned an arbitrarily high number, say $R(X, P) = 99$. There is effectively no M with a $z(M, X)$ so high, so every M must be tested. For the red curve, the lowest Z below the threshold is 1, so $R(X, P) = 1$; thus, only M with $z(M, X) \geq 1$, i.e., with a sufficiently reliable prediction of inactivity, would be "not tested".

Generating Testing Recommendations for New Molecules (Section 5 in Figure 1). Given the set of rules, i.e., a list of $R(X, P)$ generated above from historical data, one can recommend whether a new M should be tested for X. The assumption is made that the accuracy of prediction historically observed for X/P will be similar for the new X/P compounds. M is registered with a project label P, so we know which $R(X, P)$ to use. The prediction and a prediction uncertainty (and therefore a $z(M, X)$) for M can be generated from the latest global model for X, which we assume does not include M. These are the possible recommendations in order of decreasing priority:

1. M has already been tested for X. Recommendation: NOT TEST
2. M is predicted to be active on X. Recommendation: TEST
3. M is predicted to be inactive on X, but $z(M, X) < R(X, P)$. Recommendation: TEST
4. M is predicted to be inactive on X, and $z(M, X) \geq R(X, P)$. Recommendation: NOT TEST

Here we are using "NOT TEST" for testing being deferred on the basis of preexisting rules, as opposed to "not test", which was applied to compounds during generation of the rules.

In practice, the presumption is that a chemist will be able to override any of these recommendations for special cases (a previous measurement needs repeating, a compound is interesting enough that testing results are required sooner than expected, etc.)

RESULTS

We have 292 199 M/X combinations for which M was tested for X, where M was registered between July 2011 (when prospective predictions were first calculated) and the end of December 2013. From this data set, a risk−benefit curve was calculated for each X/P combination for nine counterscreens and 59 projects. The activity cutoffs in Table 1 were used to generated those curves (sections 1−3 in Figure 1). We also have a set of 39 585 M/X combinations for which M was tested for X, where M was registered between January and July 2014. This set was used to perform a prospective test of the method.

***z*(M, X) and the Probability of Being Inactive.** Figure 4 shows the probability for a compound registered between January and July 2014 to be observed as inactive given its $z(M, X)$, i.e., the reliability of a prediction of "inactive". Plots for the combined data and also for three individual counterscreen assays are presented. As expected, at $z(M, X) = 0$ the probability is 50%, and the probability approaches 100% as $z(M, X)$ increases and approaches 0% as $z(M, X)$ becomes more negative. The probability distribution is sigmoidal, which is in line with the assumption that the prediction uncertainty is Gaussian in distribution and similar for every X. This validates the idea of using $z(M, X)$ as a metric for "reliability." The reason HERG has $z(M, X)$ within ±1.5 is that the prediction uncertainties for HERG tend to be among the largest.

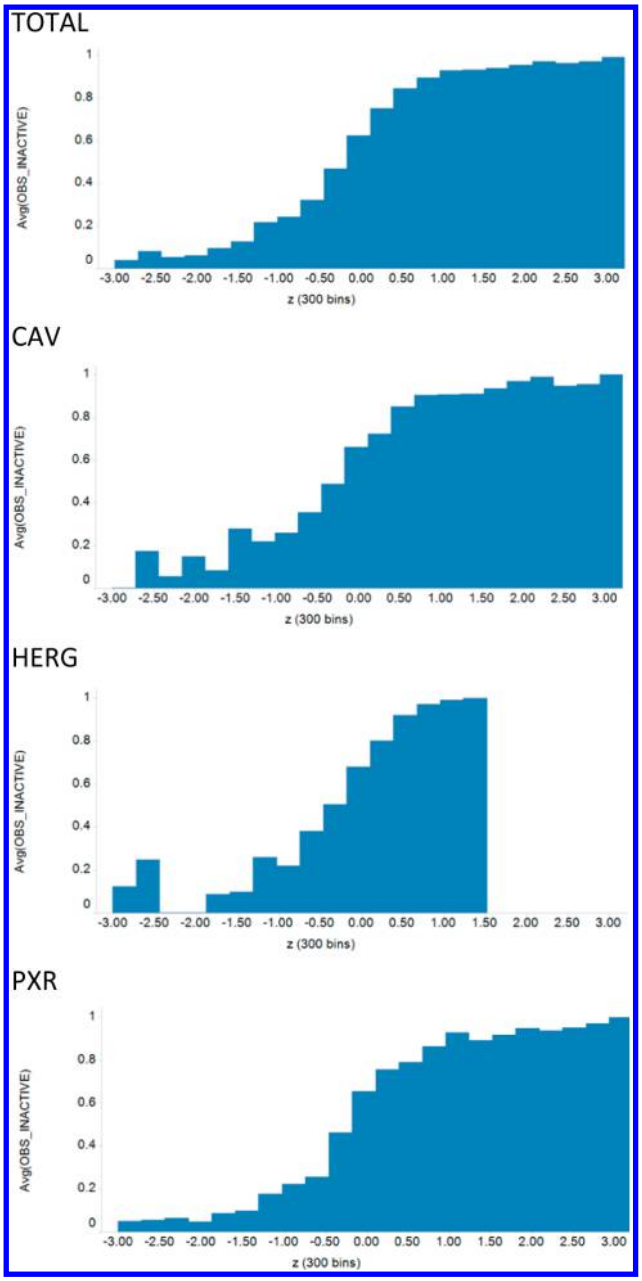


Figure 4. Probabilities of being inactive vs $z(M, X)$ for all predictions of M registered between January and July 2014 for three example counterscreens.

Example Risk−Benefit Curves. Figure 5 shows the risk−benefit curves for 2D6 and HERG, as examples, for all of the projects (distinguished by colors). We see that the risk for any given X varies among projects. 2D6 is an example of a counterscreen that seldom presents much risk: the risk−benefit curves for almost all of the projects have a very low fraction false negatives. The one project for which the curve is well above the risk threshold (here shown at 0.1) has another cytochrome P450

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231−238

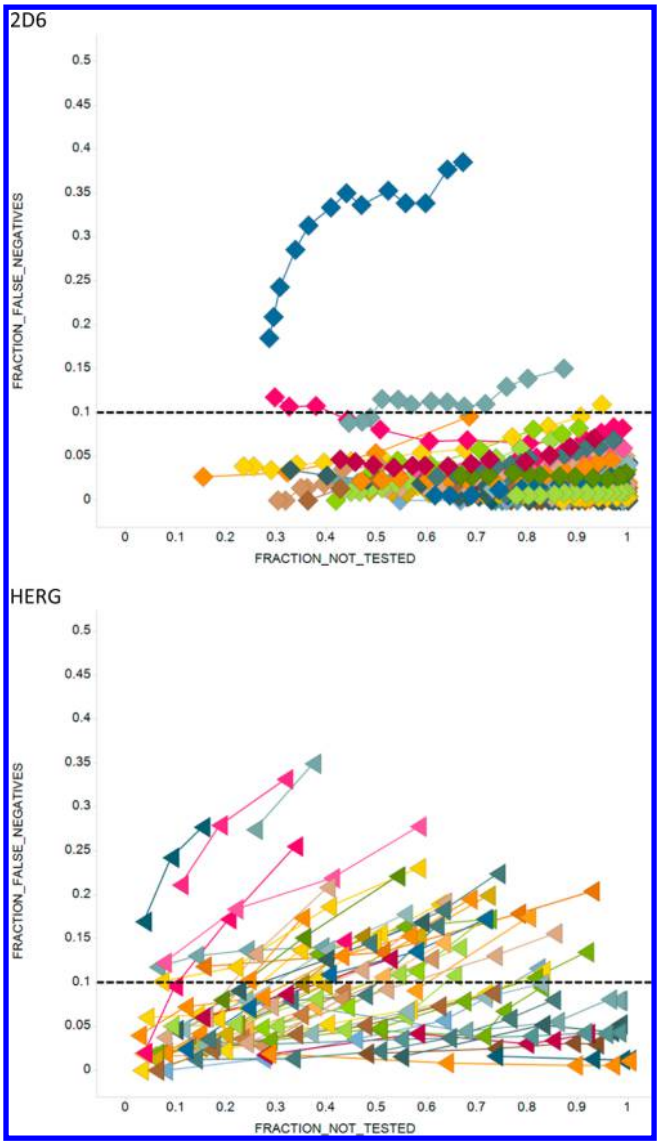


Figure 5. Example risk−benefit curves shown for all of the projects (distinguished by color) on two example counterscreen assays. In this and subsequent figures, the symbols on each curve mark, going right to left, the Z values 0.0, 0.25, 0.50, 0.75, etc.

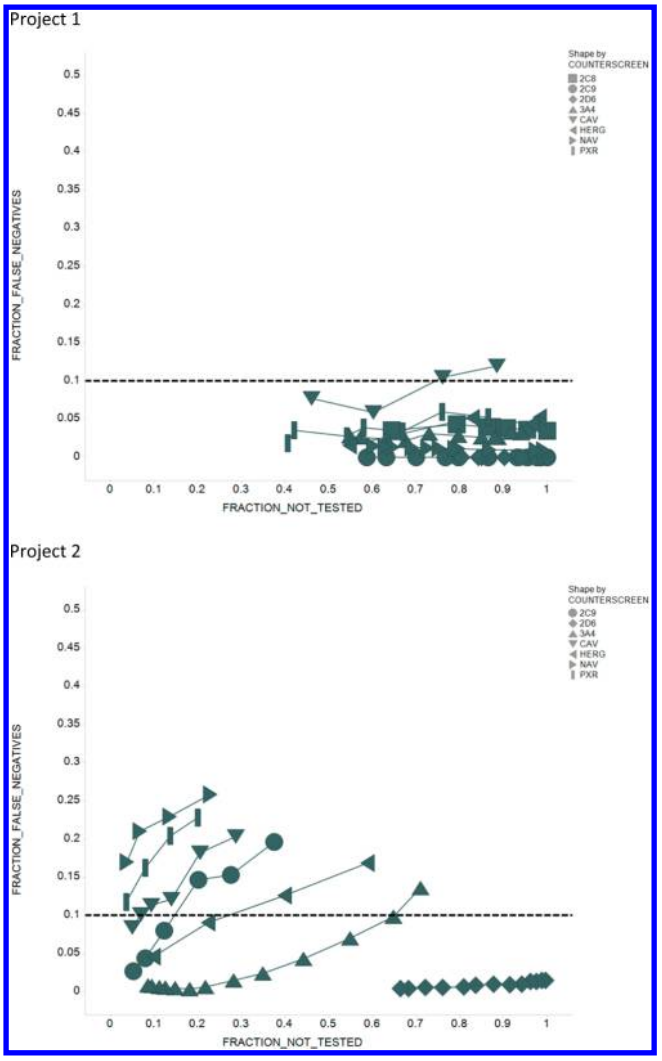


Figure 6. Example risk−benefit curves for individual projects showing all counterscreen assays (distinguished by symbol shape).

as the target, so it would not be surprising that 2D6 might be inhibited also. In contrast, HERG is an example of a counterscreen that is problematic for many (although not all) projects, and many risk−benefit curves are above the risk threshold.

Figure 6 shows the risk−benefit curves for all of the counterscreens (distinguished by symbol shape) for two specific projects. It should be noted that in the case of specific projects, not all of the curves are present. Most often, 2C8 and/or TDI are missing. This generally occurs because, as shown in Table 1, there are the fewest compounds tested in those counterscreens overall. TDI has an additional issue in that the activity cutoff is set so that the majority of M are predicted to be active. In either case, the number of X/P combinations "not tested" is often smaller than the 50 molecules required to calculate a point on the risk−benefit curve. A missing risk−benefit curve for X will always have a rule $R(\mathrm{X}, \mathrm{P}) = 99$, i.e., all M will be tested because there is not enough X/P data to set a more specific rule.

In the case of Project 1, which is an anti-infective target, very few tests need to be made while still keeping the risk low. There are two exceptions. Since TDI is missing, all M would have to be tested there. For CAV, only some molecules would be tested depending on their $z(\mathrm{M}, \mathrm{X})$ values. The full set of rules at a risk threshold of 0.1 is

$$R(\mathrm{2C8,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{2C9,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{2D6,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{3A4,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{CAV,\ PROJECT\ 1}) = 0.5$$

$$R(\mathrm{HERG,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{NAV,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{PXR,\ PROJECT\ 1}) = 0.0$$

$$R(\mathrm{TDI,\ PROJECT\ 1}) = 99.0$$

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231−238

Table 2. Results for Compounds Tested on Counterscreens for Compounds Registered between January and July 2014 Where a Recommendation Can Be Compared to an Observed Counterscreen Result

assay	total verifiable recommendations	unique projects	fraction predicted ACTIVE	fraction NOT TEST at 0.2	fraction NOT TEST at 0.1	fraction NOT TEST at 0.05	fraction NOT TEST at 0.025	ratio of false negatives at 0.2	ratio of false negatives at 0.1	ratio of false negatives at 0.05	ratio of false negatives at 0.025
2C8	798	17	0.24	0.55	0.49	0.32	0.19	0.07	0.05	0.02	0.01
2C9	4203	37	0.17	0.83	0.69	0.55	0.42	0.11	0.08	0.05	0.04
2D6	4203	37	0.02	0.93	0.93	0.86	0.66	0.05	0.05	0.03	0.02
3A4	6622	40	0.04	0.92	0.90	0.77	0.58	0.05	0.04	0.03	0.02
HERG	8346	43	0.37	0.58	0.37	0.19	0.06	0.10	0.05	0.02	0.01
CAV	4377	33	0.29	0.58	0.37	0.22	0.09	0.10	0.07	0.05	0.03
NAV	4389	34	0.16	0.69	0.46	0.33	0.27	0.09	0.05	0.03	0.02
TDI	924	32	0.58	0.14	0.08	0.07	0.07	0.15	0.06	0.04	0.03
PXR	5722	40	0.54	0.40	0.29	0.15	0.10	0.12	0.09	0.05	0.05
total	39584			0.67	0.54	0.41	0.28				

Project 2, which involves a nuclear hormone receptor target, is an example where counterscreen activity is more problematic. 2C8 and TDI are missing, so M would always be tested on those counterscreens. All M would always be tested on NAV and PXR since their curves are always above the threshold of 0.1. 2D6 is always under the threshold, and compounds predicted to be inactive would not be tested. Compounds would be tested on other counterscreens depending on their z(M, X) values. The full set of rules is

$$R(\text{2C8, PROJECT 2}) = 99.0$$

$$R(\text{2C9, PROJECT 2}) = 0.75$$

$$R(\text{2D6, PROJECT 2}) = 0.0$$

$$R(\text{3A4, PROJECT 2}) = 0.25$$

$$R(\text{CAV, PROJECT 2}) = 1.25$$

$$R(\text{HERG, PROJECT 2}) = 0.5$$

$$R(\text{NAV, PROJECT 2}) = 99.0$$

$$R(\text{PXR, PROJECT 2}) = 99.0$$

$$R(\text{TDI, PROJECT 2}) = 99.0$$

Overall Benefits and Risks in Simulated Prospective Prediction. Sets of R(X, P) were derived from the risk−benefit curves at risk thresholds of 0.2, 0.1, 0.05, and 0.025 for all projects using data from compounds registered from July 2011 to December 2013. This follows sections 1−4 in Figure 1. Recommendations based on the X/P rules were made for the 39 585 M/X combinations for compounds registered between January and July 2014. None of these compounds were in the training set of the QSAR models and none were used in the construction of the risk−benefit plots. That is, these are the "new molecules" in section 5 of Figure 1. This gives us an opportunity to measure the savings and actual risk we would have undertaken for compounds with a NOT TEST recommendation (section 6 in Figure 1). We emphasize that we are measuring the potential savings in M/X counterscreening that was actually done in 2014, not all possible M/X combinations during that time. Details of the analysis (section 7 of Figure 1) are shown in Table 2. Following the recommendations for risk thresholds of 0.2, 0.1, 0.05, and 0.025, we find that 0.67, 0.54, 0.41, and 0.28, respectively, of the testing could be saved overall. The fractions of testing saved as functions of the risk threshold are shown in Figure 7. Gratifyingly, one can ask for a fairly low risk (say 0.05) and still have substantial savings in counterscreen testing.

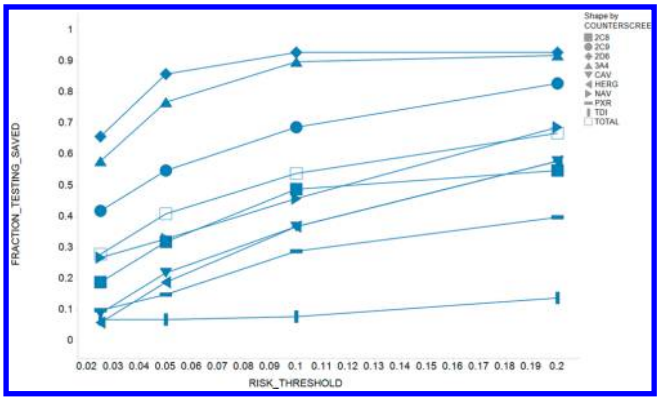


Figure 7. Fractions of testing saved vs the user-set risk threshold for January−July 2014 for all counterscreen assays (distinguished by symbol shape). The total values over all counterscreens are shown as open squares.

What fraction of those recommended NOT TEST would have turned out to be active in their respective assays? We see in Table 2 that at risk thresholds of 0.2, 0.1, and 0.05, the actual fraction of false negatives never exceeds the expected risk threshold. At a risk threshold of 0.025, the fraction of actives exceeds the expected risk threshold for four assays. This is perhaps not surprising since, as the risk threshold decreases, the rules are supported by fewer "not tested" compounds and may become less accurate.

Figure 8 shows some example histograms of observed activity for those recommended TEST versus NOT TEST. Individual histograms are shown for risk thresholds of 0.2, 0.1, 0.05, and 0.025. Generally, the lines for both TEST (blue) and NOT TEST (red) for 0.025 are lowest on the page, and the lines for 0.2 are highest on the page. The area of the NOT TEST histogram to the right of the activity cutoff (dashed line) is reflected as the "ratio false negative" column in Table 2. As expected, the compounds recommended NOT TEST show lower observed activities than the compounds recommended TEST.

DISCUSSION

There are many possible applications of QSAR models in a pharmaceutical environment, which have been discussed at length in recent reviews (e.g., refs 16 and 17). One application in recent discussion is how to prioritize the synthesis or

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231−238

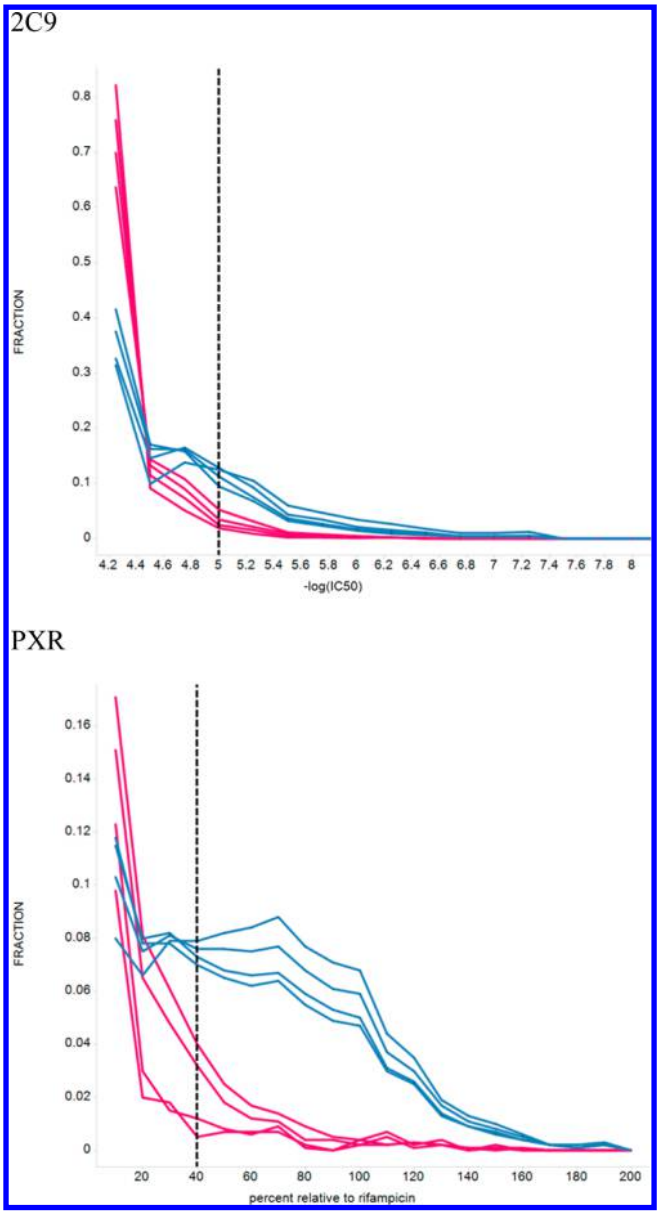


Figure 8. Histograms of observed activities for compounds registered between January and July 2014, recommended TEST (blue) vs NOT TEST (red). Results for two example counterscreen assays are shown. Separate lines are drawn for risk thresholds of 0.2, 0.1, 0.05, and 0.025. The vertical dashed line in each panel is the activity cutoff for that assay.

development of compounds on the basis of multiple predicted activities, both on-target and off-target (ADMET), simultaneously.[18,19] The application of QSAR discussed in this paper is specifically concerned with prioritizing the testing of already registered compounds against a specific set of essential off-target assays with the goal of reducing the cost of that particular type of testing.

The expense of testing every compound on every counterscreen (or at least every compound on some arbitrary selection of counterscreens) can be addressed in a number of ways. Any of them assumes some risk that an important undesired activity will be undetected until later in development. One approach, which is much simpler because it does not use any predictions but only the observed historical data, is to not test any molecule in project P for counterscreen X where X/P has a frequency of observed activity less than the risk threshold, and to test otherwise. The simpler approach affords much lower fractions of testing saved in our example: 0.42, 0.25, 0.12, and 0.08 for risk thresholds of 0.2, 0.1, 0.05, and 0.025, respectively, compared with 0.67, 0.54, 0.41, and 0.28 using eCounterscreening. eCounterscreening allows more compounds to be not tested for any given X/P combination because we can make a determination for an individual compound. We emphasize that for eCounterscreening to work most effectively, the screening protocol must allow any individual compound to be tested on an individual assay. Thus, situations where whole plates of compounds need to be tested together or a given compound has to be tested on, for instance, all of the CYPs would realize less savings. The downside of eCounterscreening is the extra complexity: we require QSAR predictions for M and some idea of the reliability of an "inactive" prediction for M, which in turn requires QSAR models and some way of calculating a prediction uncertainty.

The amount of counterscreening one can potentially save versus the risk one must assume depends on the assay and the project, and it is up to project chemists to assign the amount of risk they are willing to assume. The optimum balance is not entirely straightforward; while one can easily assign a dollar amount to the cost of testing a single compound on a specific assay, the monetary cost of having an undetected undesired activity is harder to estimate. We have been able to demonstrate using in-house data, with the assumption of what seem to be reasonable activity cutoffs and reasonable risk thresholds, a substantial amount of counterscreen testing can be saved, given QSAR models with only moderate predictivity.

We will next discuss simplifying assumptions in our paradigm that might be modified. We have assumed that the activity cutoff for a given X would be independent of the project. That seems reasonable but might not always be true. Also, for this exercise we have assumed a single risk threshold for X for every P. That also might vary from project to project and with time within a project.

We have also assumed that one would use all of the available historical prospective prediction data to calculate risk–benefit plots. It is also possible to restrict the historical data to a specific time period, say the past year. Of course, when one introduces such restrictions, fewer data are available to calculate the risk–benefit curves, and fewer rules can be assigned with certainty.

Finally, our rules are based on X/P combinations. An alternative to "project" would be to use chemical series (S), so the risk–benefit curve would be for each X/S combination instead of each X/P combination. Some preliminary evidence indicates that more discrimination is available among series than among projects. Two issues are introduced, however. One must find a way of unambiguously defining series that will persist over time, perhaps as a set of substructure queries, and be able to use the queries to unambiguously assign a new molecule to a series. Also, since the number of compounds in a series is generally smaller than the number of compounds in a project, the statistics for the risk–benefit of X/S will be worse than for X/P.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sheridan@merck.com.

Present Address

[‖]J.H.V.: Structural Chemistry, Gilead Sciences Inc., Foster City, CA 94404.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231–238

■ ACKNOWLEDGMENTS

The authors thank Joseph Shpungin for parallelizing the random forest model so that it can handle very large data sets. The QSAR infrastructure used in this work depends on the MIX modeling infrastructure, and authors are grateful to other members of the MIX team. A large number of Merck biologists, over many years, generated the data for the examples used in this paper. The authors thank Michael Altman, Eric Gifford, Prabha Karnachi, Andy Liaw, and Brian Mattioni for helpful discussions.

■ REFERENCES

(1) De Ponti, F.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. Safety of non-antiarrhythmic drugs that prolong the QT interval or induce Torsade de Pointes. *Drug Saf.* **2002**, *25*, 263−286.

(2) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **2006**, *440*, 463−469.

(3) Elkins, R. C.; Davies, M. R.; Brough, S. J.; Gavaghan, D. J.; Cui, Y.; Abi-Gerges, N.; Mirams, G. R. Variability in high-throughput ion-channel screening data and consequences for cardiac safety assessment. *J. Pharmacol. Toxicol. Methods* **2013**, *68*, 112−122.

(4) Bertz, R. J.; Granneman, G. R. Use of in vitro and in vivo data to estimate the likelihood of metabolic pharmacokinetic interactions. *Clin. Pharmocokinet.* **1997**, *32*, 210−258.

(5) Zhou, S.-F.; Xue, C. C.; Yu, X.-Q.; Li, C.; Wang, G. Clinically important drug interactions potentially involving mechanism-based inhibition of cytochrome P450 3A4 and the role of therapeutic drug monitoring. *Ther. Drug Monit.* **2007**, *29*, 687−710.

(6) Zhou, S.; Chan, S. Y.; Goh, B. C.; Chan, E.; Duan, W.; Huang, M.; McLeod, H. L. Mechanism-based inhibition of cytochrome P450 3A4 by therapeutic drugs. *Clin. Pharmacokinet.* **2005**, *44*, 279−304.

(7) Waxman, D. J. P450 gene induction by structurally diverse xenochemicals: Central role of nuclear receptors CAR, PXR, and PPAR. *Arch. Biochem. Biophys.* **1999**, *369*, 11−23.

(8) Guengerich, F. P. Cytochrome P-450 3A4: Regulation and role in drug metabolism. *Annu. Rev. Pharmacol. Toxicol.* **1999**, *39*, 1−17.

(9) Handschin, C.; Meyer, U. A. Induction of drug metabolism: The role of nuclear receptors. *Pharm. Rev.* **2003**, *55*, 649−673.

(10) Sheridan, R. P. Global quantitative structure−activity relationship models vs selected local models as predictors of off-target activities for project compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1083−1092.

(11) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(12) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(13) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(14) *Molecular Operating Environment (MOE)*, version 2008; Chemical Computing Group: Montreal, Canada, 2009; www.chemcomp.com.

(15) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.

(16) Michielan, L.; Moro, S. Pharmaceutical perspectives of nonlinear QSAR strategies. *J. Chem. Inf. Model.* **2010**, *50*, 961−978.

(17) Sprous, D. G.; Palmer, R. K.; Swanson, J. T.; Lawless, M. QSAR in the pharmaceutical research settings: QSAR models for broad, large problems. *Curr. Top. Med. Chem.* **2010**, *10*, 619−637.

(18) Ekins, S.; Honeycutt, J. D.; Metz, J. T. Evolving molecules using multi-objective optimization: Applying to ADME/Tox. *Drug Discovery Today* **2010**, *15*, 451−460.

(19) Segall, M. D. Multiparameter optimization: Identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* **2012**, *18*, 1292−1310.

DOI: 10.1021/ci500666m
J. Chem. Inf. Model. 2015, 55, 231−238