Transitioning to composite bacterial mutagenicity models in ICH M7 (Q)SAR analyses

Curran Landry\textsuperscript{a}, Marlene T. Kim\textsuperscript{a}, Naomi L. Kruhlak\textsuperscript{a}, Kevin P. Cross\textsuperscript{b}, Roustem Saiakhov\textsuperscript{c}, Suman Chakravarti\textsuperscript{c}, Lidiya Stavitskaya\textsuperscript{a,\textsuperscript{*}}

\textsuperscript{a}US Food and Drug Administration, Center for Drug Evaluation and Research, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA
\textsuperscript{b}Leadscope Inc., 1393 Dublin Road, Columbus, OH 43215, USA
\textsuperscript{c}Multicase Inc., 23811 Chagrin Boulevard, Suite 305, Beachwood, OH 44122, USA

Abstract

The International Council on Harmonisation (ICH) M7(R1) guideline describes the use of complementary (quantitative) structure-activity relationship ((Q)SAR) models to assess the mutagenic potential of drug impurities in new and generic drugs. Historically, the CASE Ultra and Leadscope software platforms used two different statistical-based models to predict mutations at G-C (guanine-cytosine) and A-T (adenine-thymine) sites, to comprehensively assess bacterial mutagenesis. In the present study, composite bacterial mutagenicity models covering multiple mutation types were developed. These new models contain more than double the number of chemicals (n = 9,254 and n = 13,514) than the corresponding non-composite models and show better toxicophore coverage. Additionally, the use of a single composite bacterial mutagenicity model simplifies impurity analysis in an ICH M7 (Q)SAR workflow by reducing the number of model outputs requiring review. An external validation set of 388 drug impurities representing proprietary pharmaceutical chemical space showed performance statistics ranging from 66 to 82\% in sensitivity, 91 to 95\% in negative predictivity and 96\% in coverage. This effort represents a major enhancement to these (Q)SAR models and their use under ICH M7(R1), leading to improved patient safety through greater predictive accuracy, applicability, and efficiency when assessing the bacterial mutagenic potential of drug impurities.

\textsuperscript{*}Corresponding Author. Lidiya.Stavitskaya@fda.hhs.gov.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher’s Disclaimer: FDA CDER Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA’s views or policies. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
1. Introduction

The bacterial reverse mutation assay is designed to detect and classify mutagens. Specifically, the test uses several auxotrophic strains of *Salmonella enterica* serovar Typhimurium and *Escherichia coli* to detect point and frame-shift mutations, which include substitution, addition, or deletion of one or more DNA base pairs (Ames et al., 1973; Green et al., 1976; Maron and Ames, 1983). The principle of the bacterial reverse mutation assay is to detect mutagens through the reversion of auxotrophic bacteria to wild type in the presence of the test substance. This assay can be conducted in *Salmonella enterica* Typhimurium strains TA98, TA100, TA1535, TA1537 (or TA97, or TA97a), and TA102 (or *E. coli* WP2 *uvrA* with or without pKM101) (ICH, 2011). The bacterial reverse mutation assay is one of the most widely used components of the International Council on Harmonisation (ICH) S2 genotoxicity test battery to assess the safety of pharmaceuticals prior to clinical exposure (ICH, 2011). The battery includes multiple assays to detect mutagenic, clastogenic and aneugenic effects *in vitro* and *in vivo*, where the most commonly used combination of tests comprises the bacterial reverse mutation assay, the mouse lymphoma assay, the *in vitro* chromosomal aberration assay, and the *in vivo* micronucleus assay (Gatehouse, 2012; Stavitskaya et al., 2015). The test battery is intended to identify genotoxic substances that exhibit a greater likelihood of subsequently causing carcinogenicity in humans.

A pivotal study conducted by Ashby and Tennant (1988) showed that although not all carcinogens are genotoxic, many genotoxic chemicals are carcinogenic in rodents. This was later confirmed by Kirkland et al. (2005), who examined the correlation between carcinogenicity and genotoxicity in at least one of the three assays (Ames + mouse lymphoma assay, *in vitro* micronucleus assay, and *in vitro* chromosomal aberration assay). The authors found that 93% of the examined carcinogens had positive results in one or more genotoxicity assays. Furthermore, the results showed that the Ames test had the best specificity, at 74%, for predicting the outcome of the rodent carcinogenicity 2-year bioassay when compared to the other genotoxicity assays, making it the most promising early-screening assay. Early screening is especially important in drug development where a positive mutagenic result is unfavorable for a pharmaceutical candidate unless the candidate’s benefit clearly outweighs the risk.

The use of (quantitative) structure-activity relationship ((Q)SAR) models in drug development has become increasingly important as it provides rapid, early screening of pharmaceutical candidates based upon their chemical structures (Stavitskaya, 2015). (Q)SAR models describe the correlation between chemical moieties and their biological activities under the general assumption that similar chemical structures exhibit similar biological activities (Benigni and Bossa, 2011; Enoch and Cronin, 2010; Kazius et al., 2005; Mortelmans and Zeiger, 2000). They have been used for a variety of endpoints related to...
drug safety, including several assays present in the ICH S2 battery such as the bacterial mutagenicity, mouse lymphoma assay, the in vitro chromosome aberration assay and the in vivo micronucleus assay (Hsu et al., 2018; Kruhlak et al., 2012; Matthews et al., 2006b). In a regulatory environment, (Q)SAR models can contribute to the weight of evidence for decision-making by predicting toxicological endpoints for chemicals with limited or no experimental data (Kruhlak et al., 2012; Rouse et al., 2018). While (Q)SAR models have historically been used for predicting the toxicological profiles of active pharmaceutical ingredients (APIs), models have more recently found mainstream use for predicting the mutagenic potential of drug impurities.

Under the ICH M7(R1) guideline, drug impurities and degradants can be assessed for bacterial mutagenicity using two complementary computational methodologies, statistical-based (QSAR) and expert rule-based (SAR) (ICH, 2017). Statistical-based models offer the benefit of being rapid to construct from extremely large and chemically diverse datasets, while expert-rule based models provide greater interpretability capturing human derived and often mechanistically defined structural alerts that contribute to biological activity. If, following expert review, an impurity is predicted to have mutagenic properties by either methodology and the carcinogenic potential is unknown, the ICH M7 guideline recommends that the compound be controlled to a level at or below the threshold of toxicological concern (TTC) (ICH M7, 2017; Muller et al., 2006). Under the ICH M7 guideline, (Q)SAR data may be submitted by pharmaceutical applicants in place of conventional in vitro bacterial mutagenicity assay data for drug impurities up to 1 mg (ICH, 2017).

Over the past several decades, numerous bacterial mutagenicity (Q)SAR models have been constructed using a variety of (Q)SAR modeling methodologies and data sets (Cariello et al., 2002; Chakravarti and Saiakhov, 2018; Chakravarti et al., 2012; Contrera et al., 2005; Hanser, Barber et al. 2014; Jolly et al., 2015; Marchant et al., 2008; Matthews et al., 2006b; Saiakhov et al., 2013; Stavitskaya et al., 2013a; Stavitskaya et al., 2013b; Valerio and Cross, 2012; Votano et al., 2004; Williams et al., 2016; Zeiger et al., 1996). In one of the earlier studies, Zeiger et al. (1996) described the development of Salmonella mutagenicity (Q)SAR models using CASE and TOPKAT, as well as an SAR model based on structural alerts extracted from the published literature (Ashby, 1985; Ashby and Tennant, 1988; Ashby and Tennant, 1991). Two versions of the CASE models were constructed: 1) CASE/n contained 820 NTP chemicals and 2) CASE/e contained 808 EPA GENE-TOX chemicals. The external validation performance statistics of these models ranged from 67%−78% in sensitivity and 66%−84% in specificity. Similarly, the TOPKAT model was also constructed using data primarily derived from the EPA GENE-TOX database. External validation was performed using a set of less than 100 chemicals (45% positive) and performance statistics for these models showed 71% in sensitivity and 76% in specificity (Zeiger et al., 1996).

A report by Votano et al. described the use of atom E-state descriptors and MDL (Q)SAR modeling software to predict Salmonella mutagenicity using both artificial neural networks and multiple linear regression-genetic algorithm modeling techniques. A model was constructed from 2693 compounds and 400 compounds were used for validation, where concordance ranged from 81%−91%, false positive rates ranged from 3%−11%, and false negative rates ranged from 6%−8% (Votano et al., 2004). In a subsequent report applying the
same approach, Contrera et al. constructed Salmonella mutagenicity models demonstrating sensitivity of 81% and specificity of 76% (Contrera et al., 2005). The authors also constructed E. coli and composite bacterial mutagenicity models; however, the E. coli training set was limited in the number of chemicals (n = 472) and the composite bacterial mutagenicity training set contained data from several strains (e.g., TA2638) and organisms (e.g., B. subtilis) that are inconsistent with current regulatory guidelines (Contrera et al., 2005; ICH, 2011). Furthermore, atom E-state indices do not provide sufficient transparency and interpretability for the qualification of pharmaceutical impurities under ICH M7 and therefore have limited utility in a regulatory environment.

In 2006, FDA/CDER developed Salmonella and E. coli mutagenicity models using the fragment-based MC4PC software. The models were validated by external cross-validation achieving specificities of 90% and 94% and sensitivities of 70% and 31% for Salmonella and E. coli mutagenicity, respectively (Matthews et al., 2006a; Matthews et al., 2006b). Although these models provided sufficient transparency and interpretability, they were tuned for specificity rather than sensitivity making them more suitable for early screening in drug development rather than regulatory decision-making.

To this end, in 2013, FDA/CDER enhanced the predictive performance profile of the Salmonella mutagenicity models using CASE Ultra and Leadscope by expanding the training sets with recently-marketed drugs, compounds containing previously out-of-domain (OOD) toxicophores, and previously unmodeled atoms such as boron, silicon, selenium, and tin (Stavitskaya et al., 2013b). Additionally, the models were refined to yield increased sensitivity (82%) and negative predictivity (up to 73%), which are performance statistics of greater importance for (Q)SAR models used for regulatory decision-making. The models also achieved greater coverage (up to 88%) during external validation using the Hansen data set (Hansen et al., 2009); however, lower specificity values ranging from 58–68% were reported. Subsequently, FDA/CDER constructed additional (Q)SAR models that predict A-T base pair mutations—based on a combination of E. coli and Salmonella TA 102 mutagenicity data—with improved coverage and performance over previous models. The models were enhanced over those previously described by Matthews et al. (2006b) in part by expanding the training set to include molecular features from more recently marketed drugs, as well as by targeting areas of chemical space where the previous models were known to have weaknesses (Stavitskaya et al., 2013a). Cross-validation performance statistics for these models ranged from 68% to 73% in sensitivity, 80% to 87% in specificity and 77% to 81% in negative predictivity.

The need for greater sensitivity in detecting potential mutagens resulted in a decrease in specificity (primarily due to additional false positives), which can be mitigated through the application of expert knowledge. The ICH M7(R1) guideline contains a provision for the application of expert knowledge to provide additional evidence to support the final conclusion, and several recent publications have described best practices for this process (Amberg et al., 2016; Barber et al., 2015; Bower et al., 2017; Kruhlak et al., 2012; Myatt et al., 2018; Powley, 2015; Sutter et al., 2013). Expert knowledge may be applied by: 1) examining the chemical environment of the training set compounds supporting an alert to ensure they are relevant to the query chemical; 2) reviewing additional analogs to identify
relationships between structures and their mutagenic activity and/or 3) reviewing publications to identify relevant mechanisms of genotoxicity to the query chemical (Hsu et al., 2018; Amberg et al., 2016; Bower et al., 2017; Myatt et al., 2018). In cases where additional evidence suggests that a prediction may be incorrect, or when the model outcome is ambiguous (e.g., equivocal or OOD) or conflicts with results from another model or alert set, expert knowledge may be used to support a revised conclusion.

In the present study, two statistical-based models for composite bacterial mutagenicity were constructed based upon Salmonella and E. coli mutagenicity data combined. The use of a composite bacterial mutagenicity model simplifies impurity analysis in an ICH M7 (Q)SAR workflow by reducing the number of model outputs requiring review. The new models contain more than double the number of chemicals than the earlier models to enhance the domain of applicability. Data gaps were identified and compounds were added to improve predictions in those areas of chemical space. In addition, discrepant and/or deficient studies for 1,140 chemicals were examined to resolve conflicting calls. The newly-constructed models were externally validated using a test set representing proprietary pharmaceutical chemical space. Furthermore, the external test set was used to examine the predictive performance of existing structural alerts for bacterial mutagenicity in a commercially available expert rule-based model. Finally, the use of multiple (Q)SAR models in various combinations was examined in accordance with ICH M7 guidelines. These models provide greater predictive accuracy, applicability, and efficiency when assessing the mutagenic potential of drug impurities under ICH M7(R1), consistent with FDA/CDER’s regulatory imperative to protect patient safety.

2. Methods

2.1 Data sources

All training set compounds used to construct (Q)SAR models were comprised of nonproprietary bacterial reverse mutation assay data harvested from US FDA approval packages and other regulatory authorities (e.g., the Japanese NIHs and the Japanese Ministry of Health), online repositories of genetic toxicology data (e.g., NTP, EPA GENE-TOX, and CCRIS), data sharing efforts, the published literature and MultiCASE and Leadscope internal databases. Data were harvested for the following strains Salmonella TA98, TA100, TA1535, TA1537 (or TA97, or TA97a), and/or TA102 (or E. coli WP2 uvrA, or WP2 uvrA (pKM101)) in the absence and presence of metabolic activation. All findings were captured using a binary scoring system for modeling purposes, where a “0” denotes a negative response and a “1” denotes a positive response as indicated by the author call. Chemicals with a positive response in the presence and/or absence of S9 were scored as overall positive. Chemicals with equivocal, ambiguous or conflicting study results from multiple sources were re-reviewed according to ICH S2(R1) guidance (ICH, 2011) and given a resolved binary call or removed from the database. All bacterial mutagenicity training sets were constructed by expanding previously published databases for Salmonella mutagenicity and E. coli/TA102 mutagenicity (n = 3,979 and n = 1,199, respectively) (Stavitskaya et al., 2013a; Stavitskaya et al., 2013b). References and activity scores are provided in Supplementary Table S1 for data harvested from the Japanese Ministry of Health, Labor and
2.2 Data review

Published study conclusions (i.e., mutagenic, non-mutagenic) were used for this investigation without being re-reviewed unless conflicting results were obtained from multiple sources. In those cases, studies were re-reviewed according to ICH S2(R1) guidance (ICH, 2011) and given a resolved binary call or removed from the database. Specifically, chemicals with a 2-fold dose-related increase in revertants in any strain in the presence or absence of S9 were scored as overall positive, while chemicals tested under standard conditions with negative study results were scored as negative. Chemicals tested in a single strain with an overall negative result, those tested with non-standard or modified tester strains, or those with equivocal or ambiguous study results were excluded from the database. Overall, the activity scores of 94 chemicals were updated, and 98 chemicals were removed due to unacceptable study design (e.g., chemicals tested as a mixture, use of non-standard strains). Detailed information about the chemicals with updated conclusions (i.e., mutagenic, non-mutagenic) is provided in Supplementary Table S2.

2.3 Chemical structure curation

Electronic representations of the chemical structures were created using MDL molfile format. Inorganic chemicals, mixtures, and high molecular weight compounds (e.g., peptides, polysaccharides, proteins, and polymers >1000 Daltons) were excluded from the training sets due to processing limitations within the (Q)SAR software used in this investigation. Furthermore, the neutralized free form of any simple salt was included.

2.4 (Q)SAR software

Two commercial (Q)SAR software platforms, CASE Ultra (CU) version 1.7.0.5 (MultiCASE Inc., USA) and Leadscope Enterprise (LS) version 3.6.3 (Leadscope Inc., USA) were used to construct two distinct composite bacterial mutagenicity models. Derek Nexus (DX) version 6.0.1 (Lhasa Limited, UK) was used concurrently with the new composite bacterial mutagenicity models as a complementary expert rule-based model when testing the predictive performance under ICH M7 guidelines. It is of note that each software platform contains both an expert rule-based as well as statistical-based models. Historically, FDA has used the DX expert rule-based system and CASE Ultra and Leadscope statistical-based models as a first pass for internal evaluations. All software were acquired and used under Research Collaboration Agreements between FDA/CDER and the software providers mentioned above.

2.4.1 CASE Ultra (CU)—CU includes a statistical-based (Q)SAR software platform that uses a machine-learning algorithm in combination with molecular descriptors generated by fragmentation of training set structures. Fragments that are identified as being statistically associated with active molecules in the training set are defined as structural alerts. Additional fragments are also identified as deactivating features that decrease the potency of
the alerts. During model application, the model generates a confidence score between 0 and 1 to indicate the likelihood of a test chemical being positive based on the presence of alerts and deactivating features. The model also verifies that all 3-non-hydrogen atom fragments present in the test compound are present among the training set structures. In cases where the model identifies no alerts or produces a non-positive confidence score and one or more “unknown fragment(s),” the model returns an OOD response.

The new composite bacterial mutagenicity model was constructed in CU using a training set of 13,514 chemicals. The model was cross-validated using a 10 by 10% leave-many-out (LMO) method. Briefly, the entire dataset was randomly divided into 10 equal subsets, with a single subset (10% of the total training set) set aside as a test set and the remaining 9 subsets (90% of the total training set) used to reconstruct a model. CU recalculated descriptor weights for each prediction cycle based solely on the remaining 9 subsets. This process was repeated 10 times, with a different training set for each iteration. The classification threshold was selected based on optimal balance between sensitivity and specificity on the receiver operating characteristic (ROC) curve. During model application, predictions were classified as equivocal when a predicted confidence was within ±0.1 of the classification threshold. Predicted values above the upper bound of this range were treated as positive, and those below this range were treated as negative. An OOD response was given to any chemicals that contained one or more unknown fragments not recognized by the model.

2.4.2 Leadscope Enterprise (LS)—LS is a data mining and visualization software package that includes a statistical-based (Q)SAR modeling functionality. To construct a bacterial mutagenicity (Q)SAR model, a training set (n = 9,254) was imported into LS and fingerprinted using a set of 27,142 pre-defined medicinal chemistry structural features as candidates for model building descriptors. A small predictive subset of these features was used to construct the model. Additionally, a set of unique scaffolds was automatically constructed from the pre-defined structural features that specifically defined structure-activity relationships in the training set. The unique set of scaffolds was generated using the following settings: 1) the minimum of compounds per scaffold (10); 2) the minimum number of atoms per scaffold (6); 3) the maximum number of rotatable bonds (unspecified); and 4) the minimum absolute Z-score (2.0). Additionally, inclusion of properties such as charge, hydrogen bonding, and lipid solubility were explored to improve predictive performance.

The highest predictive features were identified for retention while weakly predictive features were removed using Z-score and mean activity as constraints (Roberts et al., 2000). Additional pruning was manually performed to reduce the number of features while maintaining optimal predictive performance. Subsequently, additional structural features based on known mechanisms of chemical mutagenesis were manually identified and included in the model. Lastly, the total number of model features was reduced using a partial least-squared regression algorithm leaving only those that best fit the experimental activity scores in the training set (Cross et al., 2015).
The model was cross-validated 25 times using a 10 × 10% LMO method. This method sets aside 10% of the training set for testing and reconstructs a reduced model using the remaining 90% of the compounds recalculating the descriptor weights. This process was repeated 10 times with 10 different training sets ensuring that all of the training set compounds have been predicted. The entire process was then repeated 25 times and the average predicted values were used in calculating the Cooper statistics (Cooper et al., 1979).

A classification threshold was determined by varying the positive cutoff probability thresholds for equivocal results and analyzing the resulting Cooper statistics. The optimal probability range for indeterminate predictions was identified to be 0.4 to 0.6. Predictions that are above the 0.6 probability cutoff are classified as positive, while predictions below 0.4 are classified as negative. The domain of applicability is determined using two criteria. The first criterion is defined as the presence of at least one chemical descriptor (in addition to all property descriptors) for the test compound. The second criterion is defined as the presence of at least one structurally similar analog in the training set, defined as a structure within a Tanimoto distance of ≥ 0.3. If either criterion is not met by the test chemical, then an OOD response is generated.

2.4.3 Derek Nexus (DX)—DX is an expert rule-based system that identifies SAR alerts derived from mechanistic knowledge and relevant experimental data (Segall and Barber, 2014). The software uses a controlled vocabulary of confidence terms to express the likelihood that a prediction is correct based on the weight of evidence for and against it. Compounds that matched a structural alert for bacterial mutagenesis with a likelihood of “plausible” were treated as positive, “equivocal” predictions were treated as equivocal, “inactive” as negative, and compounds that were designated as “inactive with misclassified or unclassified features” were treated as negative, but were subjected to further investigation in an expert review workflow. DX v.6.0.1 contains a total of 135 structural alerts for bacterial mutagenesis.

2.5 Combining model outputs

The new bacterial mutagenicity models were compared to previous models for Salmonella mutagenicity, based on Salmonella TA97, TA97a, TA98, TA100, TA1535 and TA1537 (Stavitskaya et al., 2013a), and E. coli/TA102 mutagenicity, based on Salmonella TA102, E. coli WP2 uvrA, and/or WP2 uvrA (pKM101) (Stavitskaya et al., 2013b). To compare the new bacterial mutagenicity models to the previous Salmonella and E. coli/TA102 models, predictions from the Salmonella and E. coli/TA102 models were combined, where a positive prediction from either model within a given software platform justified an overall positive prediction for that software’s Salmonella/E. coli prediction. Note that this approach can mathematically improve the sensitivity when applying two models, but also increases the false positive rate. An equivocal prediction from any one model resulted in an overall equivocal prediction except in cases where the Salmonella model returned an OOD call and the E. coli/TA102 model returned an equivocal call, which resulted in an overall OOD call. Additionally, an overall OOD call was given when one of the two models generated a negative prediction and the other was OOD.
In a second exercise, CU and LS were each combined with DX giving one statistical-based model and one expert rule-based model combinations (CU/DX and LS/DX), consistent with the ICH M7 guideline. Additionally, all three models were combined into a single prediction for each query chemical (CU/LS/DX). When combining model outputs across different software platforms, a positive prediction from any one software platform was used to justify an overall positive prediction. Similarly, an equivocal prediction from any one software platform was used to justify an overall equivocal prediction, in the absence of a positive prediction. In cases where both statistical models returned an OOD call and DX returned a negative prediction, an overall OOD result was reported. However, if only one statistical model was OOD and the other statistical model generated a prediction, the OOD was disregarded and the remaining predictions were used to generate an overall call.

2.6 External validation

Predictive performance of the models was assessed using an external validation set comprised of 388 proprietary compounds (72 actives and 316 inactives). The compounds are pharmaceutical impurities that were harvested from New Drug Applications (NDAs) submitted to FDA/CDER. Chemicals that were already part of the CU and LS training sets were removed. The final set contains a range of chemical classes, including aromatic amines, aromatic nitro compounds, carbamates, and alkyl halides.

2.7 Performance statistics

Cooper statistics were used to evaluate the performance of individual and combined model outputs. Briefly, predictive performance was evaluated using a classic 2×2 contingency table containing counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Chemicals classified as OOD and equivocal were excluded from Cooper statistic calculations. Statistics such as sensitivity, specificity, positive predictivity, negative predictivity, and concordance were calculated as described by Cooper et al. (1979). Coverage was calculated as the percentage of all chemicals screened for which a prediction could be made (OOD results do not constitute a prediction).

To account for the bias present within the external validation set, the negative predictive value and positive predictive value were normalized using the following equations:

\[
\text{Normalized Positive Predictivity} = \frac{TP/\%\text{ Actives}}{TP/\%\text{ Actives} + FP/(1 - \%\text{ Actives})}
\]

\[
\text{Normalized Positive Predictivity} = \frac{TN/(1 - \%\text{ Actives})}{TN/(1 - \%\text{ Actives}) + FN/\%\text{ Actives}}
\]

3. Results

3.1 Overview of bacterial mutagenicity data sets

The bacterial mutagenicity training sets were expanded by each software provider, independently, from published databases (2013 Salmonella training set (n = 3,979) and 2013 Landry et al. Page 9

\text{Regul Toxicol Pharmacol. Author manuscript; available in PMC 2020 December 01.}
The final composite bacterial mutagenicity training sets were increased to 13,514 and 9,254 for CU and LS, respectively.

The active use of previously constructed models at FDA/CDER revealed some situations where previous training set classifications were inconsistent with newer study calls. In general, the reproducibility of the Ames test is about 85%, affected by a variety of test conditions (Benigni and Giuliani, 1988; Cooper et al., 1979; Piegorsch and Zeiger, 1991). In order to improve the predictive performance of the earlier (Q)SAR models, Ames data were re-evaluated for 1,140 chemicals (192 of which had conflicting data) in accordance with published methods (Gatehouse, 2012; ICH, 2011; Mortelmans and Zeiger, 2000) to enhance the overall quality of underlying training data. Of the 1,140 chemicals re-reviewed, 46 chemicals were changed to negative and 48 were changed to positive, while 95 were removed due to inadequate study data. Of those chemicals removed, 40 were because activity could not be assigned to a single structure. References for the updated chemicals are provided in Supplementary Table S2.

Among the re-reviewed chemicals was the acid halide class, specifically carboxylic and sulfonic acid halides. This class of chemicals is often used in the synthesis of organic chemicals and has been shown to produce a positive result in the bacterial reverse mutation assay when tested in DMSO. Furthermore, it has been recently reported that the acid halide class is capable of reacting with DMSO, via a Pummerer rearrangement, to form alkyl halides, a well-known class of mutagens (Amberg et al., 2015). However, when tested in solvents other than DMSO, the majority of these chemicals were experimentally negative. As a result, 5 chemicals were reclassified in the database as non-mutagens.

In addition to the non-drug compounds containing under-represented functional groups, 150 drug substances approved between 2013 and 2018 were added to the database, including opioids, benzodiazepines, and nucleic acid analogs (Table 1).

### 3.2 Predictive performance using cross-validation and external validation

The predictive performance of the bacterial mutagenicity models based on 10 × 10% LMO cross-validation experiments are presented in Table 2. The CU composite bacterial mutagenicity model achieved a sensitivity of 90.6% and a negative predictivity of 88.9% in cross-validation, while the LS model achieved a sensitivity of 84.7% and a negative predictivity of 82.5%. It should be noted that the cross-validation statistics between the two software are not directly comparable because the total number of chemicals in cross-validation training sets are different between the two software platforms. Whereas the external validation statistics are directly comparable.

In a subsequent evaluation, the performance of the newly-constructed models was assessed using an external validation set of 388 proprietary compounds, where 19% were experimentally positive (Table 2). Using this test set, the CU composite bacterial mutagenicity model achieved a sensitivity of 65.6% and a negative predictivity of 90.8%, while the Salmonella and E. coli/TA102 in combination achieved a sensitivity of 84.7% and a negative predictivity of 89.4% Furthermore, the newly-constructed LS bacterial mutagenicity model achieved a sensitivity of 82.3% and a negative predictivity of 95.4% in
external validation, in contrast to the Salmonella and E. coli/TA102 models combined, which achieved a sensitivity of 69.4% and a negative predictivity of 92.8%. In both software platforms, transitioning to the combined bacterial mutagenicity models resulted in a decrease in false positive rates and an increase in positive predictivity as expected. Additionally, normalized positive and negative predictivity were determined to account for the small number of active chemicals in the external validation set (Table 2). Overall, the normalized positive predictive values substantially increased in both LS and CU while the normalized negative predictive values showed a decrease in CU but maintained good predictive performance.

The combined predictive performance of the CU and LS models and DX expert rule-based system was also assessed (Table 2). Both CU+DX and LS+DX showed an increase in sensitivity (19.9% and 8.6%, respectively), and normalized negative predictivity (12% and 5.9%, respectively) when compared to the bacterial mutagenicity statistical-based models alone. In contrast, a decrease in specificity was observed for both CU and LS when combined with DX (−12.5%, and −10.2%, respectively). Similarly, the normalized positive predictivity decreased by 4.6% in CU and 5.2% in LS. These results were expected given the practice of combining predictions across different platforms resulting in an increase of false positive predictions.

The combined use of all three (Q)SAR software resulted in similar overall predictive performance to the use of two, except that higher coverage was obtained. The number of chemicals that were classified as OOD are reported in Table 3. CU returned 21 OOD responses with the Salmonella/E. coli models and 16 OOD responses using the new bacterial model, whereas LS generated 57 OOD results with the previous Salmonella/E. coli models and 16 OOD results using the new bacterial model. When combined with DX predictions, the frequency of OODs remained the same since compounds that were predicted by DX as inactive with misclassified or unclassified features were treated as negative in the absence of expert review (see Section 2.4.3). However, when predictions from all three software platforms (CU, LS, DX) were combined, 4 chemicals generated an OOD in both statistical programs in the previous Salmonella/E. coli models whereas only one chemical was OOD in both statistical programs when the new bacterial mutagenicity models were applied. The three newly-predicted chemicals representing different chemical classes contained a true positive, a true negative, and a false positive.

4. Discussion

Computational methods that provide early screening of pharmaceutical candidates and impurities to predict the outcome of a bacterial mutagenicity assay have become increasingly important for industry as well as regulatory agencies. However, for regulatory purposes, it is desirable that models provide sufficient transparency and interpretability in their predictions to facilitate the use of expert knowledge for the qualification of pharmaceutical impurities under the ICH M7 guideline. Furthermore, the application of two complementary systems that use different methodologies to leverage different strengths has been generally shown to provide greater sensitivity in detecting mutagens (Amberg et al., 2016; Barber et al., 2015; Bower et al., 2017; Kruhlak et al., 2012; Myatt et al., 2018;
Powley, 2015; Sutter et al., 2013). In the current study, predictions from the newly-developed composite bacterial mutagenicity CU and LS models in combination with predictions from DX were examined. Additionally, selected toxicophores are presented to illustrate how the new models have improved for specific chemical classes. Lastly, a series of case studies are presented below for three newly approved drugs.

4.1 External Validation of (Q)SAR Models

In a regulatory environment, high sensitivity and negative predictivity are important characteristics of (Q)SAR models used to support drug safety decisions, thereby minimizing risk to patients. In the present study, negative predictivity was maintained by both CU and LS in external validation, while a 12.9% increase in sensitivity was observed in the LS bacterial mutagenicity model as compared to the Salmonella/E. coli models (Fig. 1). In contrast, CU showed a 19% decrease in sensitivity when compared to the previous models. However, it is noted that changes in sensitivity may be magnified in the current study due to the low number of active chemicals (n=72) in the external validation set. Also of note is the increase in positive predictivity, which shows that the models are more reliable in their positive predictions. This is due in part to the combined use of Salmonella and E. coli mutagenicity models in the previous version which mathematically resulted in an increase in false positive predictions. Furthermore, the LS model demonstrated a substantial increase in coverage (10.6%) as compared to the previous models, resulting in CU and LS now showing the same coverage of the external validation set (95.9%) (Table 2).

The combined use of two complementary (Q)SAR methodologies is recommended by ICH M7(R1) to take full advantage of the relative strengths of different model descriptors and algorithms, thereby providing a more robust assessment of mutagenic activity for impurities in the absence of empirical data. Performance of a single statistical methodology compared to the performance of two complementary (Q)SAR models is shown in Fig. 2. The combined use of CU with DX as well as LS with DX increased sensitivity by 19.9% and 8.6%, respectively, and the negative predictivity by 4.2% and 1.7%, respectively. This supports the regulatory imperative to protect patient safety by reducing the number of false negative predictions, which is particularly important as drug impurities provide no therapeutic benefit. As previously reported, the combined use of two methodologies also resulted in a decrease in specificity and positive predictivity, generally producing an increased number of false positive predictions; however, the false positive rate can be decreased through the application of expert knowledge.

In addition, the performance of all three (Q)SAR models was assessed with results showing a slight improvement in sensitivity and no further improvement in negative predictivity when predictions were combined from all three software platforms. However, using three software platforms instead of two substantially decreased the number of OOD calls (Fig. 3). Furthermore, the OOD results have decreased from 4 chemicals to 1 when comparing combined predictions from the previous CU and LS models and current DX model to the new composite bacterial mutagenicity CU and LS models and current DX model (Table 3). Of the three chemicals that were determined to be OOD by the previous Salmonella/E. coli models and DX, two were correctly predicted (1 TN and 1 TP) and one was incorrectly
predicted as positive (FP) by the composite bacterial models. While a false positive is not desirable, it can be mitigated through the application of expert knowledge, using structurally similar analogs to provide evidence to dismiss the positive prediction. Overall, this result demonstrates that negative predictivity and sensitivity can be increased by combining bacterial mutagenicity predictions from two software applications and better coverage can be obtained by using three (Q)SAR models.

### 4.3 Toxicophore analysis

A toxicophore analysis was performed to assess the predictive performance of known toxicophores in the new bacterial mutagenicity models. A toxicophore or a “structural alert” is a unique structural feature associated with toxicity. The new bacterial mutagenicity (Q)SAR models contain larger, more comprehensive training sets with more than double the number of chemicals and have a broader applicability domain which has not only introduced new structural alerts and deactivating fragments but also led to the refinement of previously identified alerts. Although CU and LS utilize different algorithms for identifying structural alerts, both software platforms generated a more comprehensive set of structural features to better represent mechanistic alerts and mitigating effects. As an example, the complexity of a primary aromatic amine alert and a set of associated features that were generated by CU and LS are shown in Fig. 4 and 5. Compounds containing primary aromatic amines have the potential to exhibit mutagenic activity, which is heavily dependent on the alert chemical environment, wherein minor structural modifications can either increase or decrease mutagenic activity (Ahlberg et al., 2016). As such, CU identified several related alerts and deactivating fragments that share a primary aromatic amine substructure (Fig. 4). In contrast, the deactivating features that were identified by LS (Fig. 5) may or may not be specific to aromatic amines. Each feature in LS is instead given a calculated mean that is based on all instances in the training set, rather than the instances related to the alerting substructure of concern (in this case, a primary aromatic amine). The LS deactivating features (Fig. 5) were selected to show that similar features may be present in CU and LS (Fig. 4).

The performance of selected toxicophores was assessed by comparing the frequencies and the mean activities of previously defined Salmonella mutagenicity model features and the new composite bacterial mutagenicity model features for both CU and LS (Fig. 6). Specifically, the features that were examined in this study were considered closely related between model generations. The results of this investigation showed that the sulfonate toxicophore is now present as a general and a more specific fragment in the new CU model. Both of the new fragments contain a methylene carbon atom in the alpha position relative to the ester oxygen atom, which is essential for the nucleophilic substitution reaction to take place (Benigni and Bossa, 2011).

Similarly, there are now two vinyl halide features in the LS bacterial mutagenicity model: a more specific vinyl halide feature, and a vinyl geminal dihalide feature (both excluding fluorine). Simple vinyl halides can be metabolized by cytochrome P450 to halo oxiranes, which are directly able to alkylate DNA (Guengerich, 1994), while more highly halogenated vinyl halides can take several active forms, including halo oxiranes and acyl halides (Benigni and Bossa, 2011). Furthermore, the new vinyl halide feature has been restricted to
bromides, chlorides and iodides due to the chemicals’ potential to alkylate DNA. Fluorides have been excluded because fluorine atoms are believed to be biologically inert (Benigni and Bossa, 2011) and many of the training set chemicals containing a vinyl fluoride were found to be negative.

Additionally, the new CU model contains a refined feature for the aromatic diazo toxicophore. Aromatic diazos are metabolized by azo reductase to form aromatic amines, which are then further metabolized to reactive electrophiles capable of directly reacting with DNA (Benigni and Bossa, 2011). The definition of the previous diazo feature in the Salmonella CU model was not specific and based on aromatic imines. In contrast, the new fragments in the composite bacterial mutagenicity model are specific to aromatic diazos and show increased mean activities.

Another toxicophore that has been refined in the LS model is the aromatic hydrazine. Hydrazines are often metabolized to azo compounds, which can generate reactive carbocations and free radical species capable of interacting with DNA (Benigni and Bossa, 2011). By limiting the substitution patterns present on both nitrogens the LS feature becomes more specific and the mean activity increased. Moreover, another new terminal aromatic hydrazine feature showed even greater positive predictivity and mean activity.

In addition to the refined features, the new models contain new toxicophores as well as deactivating features. An example of a toxicophore that has been introduced into both software platforms is the aromatic amide. Aromatic amides can be metabolized to nitrenium ions, which can then directly react with DNA (Benigni and Bossa, 2011).

The enhanced training set and improved model features facilitate expert analysis of (Q)SAR predictions. A more refined feature set simplifies the identification of toxicophores and evaluation of the training set chemicals supporting predictions. Furthermore, an increase in the number of training set chemicals has resulted in more structurally relevant analogs to support or modulate predictions.

### 4.4 Case Studies

To further explore the practical application of the newly-developed models, the (Q)SAR methods described in the ICH M7 guideline were applied to a set of three recently-approved drugs (Fig 7–9). The guideline recommends the use of two complementary (Q)SAR methodologies and states that the results may be further examined using “expert knowledge” to provide “additional supportive evidence on the relevance of any positive, negative, conflicting or inconclusive prediction and to provide a rationale to support the final conclusion” (ICH, 2017). Furthermore, the use of expert knowledge has been shown to improve the overall performance of models used under ICH M7 (Barber et al., 2016; Sutter et al., 2013).

The Ames negative drug, solriamfetol, was predicted to be negative by the Salmonella/E. coli CU models and DX, while predicted to be positive by the Salmonella/E. coli LS models (Fig. 7). In contrast, the new CU bacterial mutagenicity model generated an equivocal prediction based on the presence of a carbamate moiety (highlighted below), while LS’s
bacterial mutagenicity model gave a negative prediction. A review of the training set compounds supporting the carbamate alert revealed that carbamates that contain simple alkyl chains on the oxygen substituent (e.g., urethane) are more likely to exhibit a positive response. However, a review of additional training set analogs (e.g., felbamate) indicated that larger molecules containing an aromatic ring have limited mutagenic activity. A supplemental substructure search of an external database identified a relevant analog, phenylalanine, which was negative in TA98, TA100, TA1537 and TA1535. Although phenylalanine does not contain a carbamate it does provide evidence that the backbone is likely to be negative. Based on the entire weight-of-evidence, solriamfetol was predicted to be overall negative for bacterial mutagenicity.

Amifampridine, a newly-approved, non-mutagenic drug, was predicted to be equivocal by the CU and LS Salmonella/E. coli models and positive by DX based on the presence of an aromatic amine (Fig. 8). However, the new bacterial mutagenicity models both predicted the drug to be negative. The new models identified the diamine moiety as a structural alert and a heterocyclic nitrogen in the para position to the amine as a deactivating fragment, which is consistent with the observed trends for anilines (Ahlberg et al., 2016). A review of structurally similar analogs from additional databases revealed that the presence an amine group in the ortho position to the heterocyclic nitrogen can result in mutagenicity. However, the strong deactivating effect of a heterocyclic nitrogen in the para position is supported by the aminopyralid analog when compared to the effects of chloramben. Based on the weight-of-evidence, amifampridine was predicted to be non-mutagenic.

Triclabendazole, a recently approved drug, tested negative in the bacterial reverse mutation assay; however, no prediction could be made by either of the LS and CU Salmonella and E. coli models (Fig. 9). Additionally, triclabendazole was outside the applicability domain of the LS bacterial mutagenicity model. In contrast, triclabendazole was within the applicability domain and predicted to be negative by the new LS bacterial mutagenicity model. DX also generated a negative prediction but identified a misclassified feature, based on an experimentally positive structural analog, carmethizole. However, this analog lacks a fused ring system and contains two carbamate moieties that are not present in the triclabendazole, suggesting a lack of relevance. Furthermore, two structurally similar analogs in the bacterial mutagenicity model training sets, 2-benzimidazolethiol and benzothiazole, support a negative prediction for the fused ring system. Supplemental searching of additional databases identified 5-methoxy-2-aminobenzimidazole, which shows that the methoxy group is unlikely to function as an activating group. Lastly, a review of training set chemicals containing the dechlorinated moiety (e.g., dichlorophenol) revealed that it is unlikely to be mutagenic. Considering the negative (Q)SAR predictions in combination with the negative structural analogs, triclabendazole was predicted to be overall negative for bacterial mutagenicity.

### 4.5 Impact of new models on equivocal and out-of-domain results for drug impurities

In 2017, a retrospective analysis was conducted by FDA/CDER to assess the impact of applying expert review to (Q)SAR predictions for 519 drug impurities evaluated in new and generic drug applications (Amberg et al., 2019). The ICH M7 guideline includes a provision...
for the application of expert knowledge to increase prediction confidence and resolve conflicting calls. Expert knowledge, which includes structural analog searching and mechanistic interpretation, has been particularly valuable in situations where models return an indeterminate (equivocal) result or are unable to generate a prediction due to a lack of relevant training set analogs (OOD). The use of expert knowledge was previously found to change the overall predictions 13% of the time and to resolve 72% of equivocal predictions (Amberg et al., 2019) and 95% (103/108) of OOD results (unpublished data).

This assessment was repeated using the new models described in this report. The percentage of the 519 chemicals that gave at least one OOD result decreased substantially, from 21% to only 6%. Similarly, the number of chemicals that gave an equivocal prediction dropped from 31% to 25%, indicating improved prediction resolution. In contrast, the number of chemicals whose overall predictions changed through the application of expert knowledge increased slightly from 13% to 18%, suggesting that expert review of predictions still plays an important role in resolving conflicting predictions in an ICH M7 (Q)SAR analysis workflow.

5. Conclusions

Computational toxicology continues to evolve as an increasingly valuable and important tool for both drug development as well as regulatory review. Determining the risk associated with pharmaceutical candidates for a variety of endpoints, including bacterial mutagenicity, is invaluable for preliminary high-throughput screening during development, as well as for the evaluation of pharmaceutical impurities. From a regulatory perspective, the ability to predict the toxicological profile of impurities greatly enhances the review process.

In the present study, two statistical software platforms were utilized to construct two composite bacterial mutagenicity (Q)SAR models. This was achieved by enhancing the previous model training sets and improving upon them through data collection and curation efforts. The newly-constructed bacterial mutagenicity models maintain good sensitivity and negative predictivity while showing greater coverage of proprietary pharmaceutical chemical space. Sensitivity and negative predictivity were further improved by applying two different (Q)SAR software in accordance with ICH M7 recommendations, and the use of three software platforms was found to increase overall coverage and the ability to obtain at least two valid, complementary predictions. Additionally, when two or more software platforms were in consensus, greater confidence could be inferred. In cases where the results are inconclusive or conflicting, the case studies demonstrated that expert review remains a critical step in providing additional evidence to support a final conclusion in an ICH M7 workflow.

In conclusion, the new composite bacterial mutagenicity models represent a major improvement over previous models for AT and GC mutations, providing better predictions and more efficient evaluation of drug impurities under ICH M7.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.
Acknowledgements

CASE Ultra, Leadscope Enterprise, and Derek Nexus software were used by CDER under Research Collaboration Agreements with MultiCASE Inc., Leadscope Inc., and Lhasa Limited, respectively. This project was supported by FDA’s Safety Research Interest Group and an appointment to the Research Participation Programs at the Oak Ridge Institute for Science and Education through an interagency agreement between the Department of Energy and FDA.

This project was supported by FDA’s Safety Research Interest Group and an appointment to the Research Participation Programs at the Oak Ridge Institute for Science and Education through an interagency agreement between the Department of Energy and FDA. Additionally, research reported in this article was partially supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number 2R44ES026909.

References

1. Ahlberg E; Amberg A; Beilke LD; Bower D; Cross KP; Custer L; Ford KA; Van Gompel J; Harvey J; Honma M; Jolly R; Joossens E; Kemper RA; Kenyon M; Kruhlak N; Kuhnke L; Leavitt P; Naven R; Neilan C; Quigley DP; Sluey D; Spirkl HP; Stavitskaya L; Teasdale A; White A; Wichard J; Zwickl C; Myatt GJ, Extending (Q)SARs to incorporate proprietary knowledge for regulatory purposes: A case study using aromatic amine mutagenicity. Regul Toxicol Pharmacol 2016, 77, 1–12. https://doi.org/10.1016/j.yrtph.2016.02.003 [PubMed: 26879463]

2. Amberg A; Andaya RV; Anger LT; Barber C; Beilke L; Bercu J; Bower D; Brigo A; Cammerer Z; Cross KP; Custer L; Dobo K; Gerets H; Gervais V; Glowienke S; Gomez S; Van Gompel J; Harvey J; Hasselgren C; Honna M; Johnson C; Jolly R; Kemper R; Kenyon M; Kruhlak N; Leavitt P; Miller S; Muster W; Naven R; Nicolette J; Parenty A; Powley M; Quigley DP; Reddy MV; Sasaki JC; Stavitskaya L; Teasdale A; Trejo-Martin A; Weiner S; Welch DS; White A; Wichard J; Woolley D; Myatt GJ, Principles and procedures for handling out-of-domain and indeterminate results as part of ICH M7 recommended (Q)SAR analyses. Regul Toxicol Pharmacol 2019, 102, 53–64. https://doi.org/10.1016/j.yrtph.2018.12.007 [PubMed: 30562600]

3. Amberg A; Beilke L; Bercu J; Bower D; Brigo A; Cross KP; Custer L; Dobo K; Dowdy E; Ford KA; Glowienke S; Van Gompel J; Harvey J; Hasselgren C; Honna M; Jolly R; Kemper R; Kenyon M; Kruhlak N; Leavitt P; Miller S; Muster W; Naven R; Nicolette J; Parenty A; Powley M; Quigley DP; Reddy MV; Stavitskaya L; Teasdale A; Weiner S; Welch DS; White A; Wichard J; Myatt GJ, Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses. Regul Toxicol Pharmacol 2016, 77, 13–24. 10.1016/j.yrtph.2016.02.004 [PubMed: 26877192]

4. Amberg A; Harvey JS; Czich A; Spirkl H-P; Robinson S; White A; Elder DP, Do Carboxylic/Sulfonic Acid Halides Really Present a Mutagenic and Carcinogenic Risk as Impurities in Final Drug Products? Organic Process Research & Development 2015, 19 (11), 1495–1506. 10.1021/acs.oprd.5b00106

5. Ames BN; Lee FD; Durston WE, An improved bacterial test system for the detection and classification of mutagens and carcinogens. Proceedings of the National Academy of Sciences of the United States of America 1973, 70 (3), 782–6. [PubMed: 4577135]

6. Araya S; Lovsin-Barle E; Glowienke S, Mutagenicity assessment strategy for pharmaceutical intermediates to aid limit setting for occupational exposure. Regul Toxicol Pharmacol 2015, 73 (2), 515–20. 10.1016/j.yrtph.2015.10.001 [PubMed: 26454093]

7. Ashby J, Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. Environmental mutagenesis 1985, 7 (6), 919–21. [PubMed: 4065064]

8. Ashby J; Tennant RW, Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. Mutation research 1988, 204 (1), 17–115. [PubMed: 3277047]

9. Ashby J; Tennant RW, Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. Mutation research 1991, 257 (3), 229–306. [PubMed: 1707500]

10. Barber C; Amberg A; Custer L; Dobo KL; Glowienke S; Van Gompel J; Gutsell S; Harvey J; Honma M; Kenyon MO; Kruhlak N; Muster W; Stavitskaya L; Teasdale A; Vessey J; Wichard J,
Establishing best practice in the application of expert review of mutagenicity under ICH M7. Regul Toxicol Pharmacol 2015, 73 (1), 367–77. 10.1016/j.yrtph.2015.07.018 [PubMed: 26248005]

11. Barber C; Cayley A; Hanser T; Harding A; Heghes C; Vessey JD; Werner S; Weiner SK; Wichard J; Giddings A; Glowienke S; Parenty A; Brigo A; Spirkl HP; Amberg A; Kemper R; Greene N, Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained. Regul Toxicol Pharmacol 2016, 76, 7–20. https://doi.org/10.1016/j.yrtph.2015.12.006 [PubMed: 26708083]

12. Benigni R; Bossa C, Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. Chemical reviews 2011, 111 (4), 2507–36. 10.1021/cr100222q [PubMed: 21265518]

13. Benigni R; Giuliani A, Computer-assisted analysis of interlaboratory Ames test variability. Journal of toxicology and environmental health 1988, 25 (1), 135–48. 10.1080/15287398809531194 [PubMed: 3418743]

14. Bower D; Cross KP; Escher S; Myatt GJ; Quigley DP, In silico Toxicology: An Overview of Toxicity Databases, Prediction Methodologies, and Expert Review In Johnson D; Richardson R, Eds. Royal Society of Chemistry: London, UK, 2017; pp 209–242.

15. Cariello NF; Wilson JD; Britt BH; Wedd DJ; Burlinson B; Gombar V, Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. Mutagenesis 2002, 17 (4), 321–329. 10.1093/mutage/17.4.321 [PubMed: 12110629]

16. Chakravarti SK; Saiakhov RD, Computing similarity between structural environments of mutagenicity alerts. Mutagenesis 2018, gey032–gey032. 10.1093/mutage/gey032

17. Chakravarti SK; Saiakhov RD; Klopman G, Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. Journal of chemical information and modeling 2012, 52 (10), 2609–18. 10.1021/ci300111r [PubMed: 22947043]

18. Contrera JF; Matthews EJ; Kruhlak NL; Benz RD, In silico screening of chemicals for bacterial mutagenicity using electrotopological E-state indices and MDL QSAR software. Regul Toxicol Pharmacol 2005, 43 (3), 313–23. 10.1016/j.yrtph.2005.09.001 [PubMed: 16242226]

19. Cooper JA 2nd, Saracci R; Cole P, Describing the validity of carcinogen screening tests. British journal of cancer 1979, 39 (1), 87–9. [PubMed: 758931]

20. Cross KP; Kruhlak N; Myatt GJ; Stavitskaya L; White A, Ensuring Regulatory Acceptable (Q)SAR Models and Expert Alerts for ICH M7 Reflect Proprietary Chemical Space. In 35th Annual Meeting of the Americina College of Toxicology, 2015; Vol. 34, pp 83–84.

21. Ellis P; Kenyon M; Dobo KL; Kenyon MO; Cheung J; Munzner J; Sutter A; Wichard J, A practical application of two in silico systems for identification of potentially mutagenic impurities. Regul Toxicol Pharmacol 2015, 72 (2), 335–49. https://doi.org/10.1016/j.yrtph.2015.05.008 [PubMed: 25980641]

22. Guengerich FP, Mechanisms of Formation of DNA Adducts from Ethylene Dihaudes, Vinyl Halides, and Arylamines. Drug Metabolism Reviews 1994, 26 (1–2), 47–66. 10.3109/03602539409029784 [PubMed: 8082581]

23. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N and Muller KR (2009). “Benchmark data set for in silico prediction of Ames mutagenicity.” J Chem Inf Model 49 (9): 2077–2081. [PubMed: 19702240]
28. Hanser T; Barber C; Rosser E; Vessey JD; Webb SJ; Werner S, Self-organising hypothesis networks: a new approach for representing and structuring SAR knowledge. J Cheminform 2014, 6, 21. [PubMed: 24959206]
29. Haworth S; Lawlor T; Mortelmans K; Speck W; Zeiger E, Salmonella mutagenicity test results for 250 chemicals. Environmental mutagenesis 1983, 5 Suppl 1, 1–142.
30. Honma M; Kitazawa A; Cayley A; Williams RV; Barber C; Hanser T; Saiakhov R; Chakravarti S; Myatt GJ; Cross KP; Benfenati E; Raitano G; Mekenyan O; Petkov P; Bossa C; Benigni R; Battistelli CL; Giuliani A; Tcheremenskaia O; DeMeo C; Norinder U; Koga H; Jeliazkova N; Kochev N; Paskaleva V; Yang C; Daga PR; Clark RD; Rathman J, Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis 2019, 34 (1), 3–16. 10.1093/mutage/gey031 [PubMed: 30357358]
31. Hsu CW; Hewes KP; Stavitskaya L; Kruhlak NL, Construction and application of (Q)SAR models to predict chemical-induced in vitro chromosome aberrations. Regul Toxicol Pharmacol 2018, 99, 274–288. 10.1016/j.yrtph.2018.09.026 [PubMed: 30278198]
32. ICH M7, Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk In International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), Group IEW, Ed. 2017; pp 1–27.
33. ICH S2(R1), Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use S2(R1) In International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals, Group IEW, Ed. 2011; pp 1–25.
34. Jolly R; Ahmed KB; Zwickl C; Watson I; Gombar V, An evaluation of in-house and off-the-shelf in silico models: implications on guidance for mutagenicity assessment. Regul Toxicol Pharmacol 2015, 71 (3), 388–97. 10.1016/j.yrtph.2015.01.010 [PubMed: 25656493]
35. Kazius J; McGuire R; Bursi R, Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry 2005, 48 (1), 312–20. 10.1021/jm040835a [PubMed: 15634026]
36. Kirkland D; Aardema M; Henderson L; Muller L, Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. Mutation research 2005, 584 (1–2), 1–256. https://doi.org/10.1016Z.mrgentox.2005.02.004 [PubMed: 15979392]
37. Kruhlak NL; Benz RD; Zhou H; Colatsky TJ, (Q)SAR Modeling and Safety Assessment in Regulatory Review. Clinical Pharmacology & Therapeutics 2012, 91 (3), 529–534. 10.1038/clpt.2011.300 [PubMed: 22258468]
38. Marchant CA; Briggs KA; Long A, In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. Toxicol Mech Methods. 2008, 18(2–3), 177–87. [PubMed: 20020913]
39. Maron DM, Ames BN, 1983 Revised methods for the Salmonella mutagenicity test. Mutat Res. 113, 173–215. [PubMed: 6341825]
40. Matthews EJ; Kruhlak NL; Cimino MC; Benz RD; Contrera JF, An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints. Regul Toxicol Pharmacol 2006a, 44 (2), 83–96. 10.1016/j.yrtph.2005.11.003 [PubMed: 16386343]
41. Matthews EJ; Kruhlak NL; Cimino MC; Benz RD; Contrera JF, An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. Identification of genotoxicants, reprotoxicants, and carcinogens using in silico methods. Regul Toxicol Pharmacol 2006b, 44 (2), 97–110. 10.1016/j.yrtph.2005.10.004 [PubMed: 16352383]
42. Mortelmans K; Haworth S; Lawlor T; Speck W; Tainer B; Zeiger E, Salmonella mutagenicity tests: II. Results from the testing of 270 chemicals. Environmental mutagenesis 1986, 8 Suppl 7, 1–119.
43. Mortelmans K; Zeiger E, The Ames Salmonella/microsome mutagenicity assay. Mutation research 2000, 455 (1–2), 29–60. [PubMed: 11113466]
44. Muller L; Mauthe RF; Riley CM; Andino MM; Antonis DD; Beels C; DeGeorge J; De Knaep AG; Ellison D; Fagerland JA; Frank R; Frischel B; Galloway S; Harpur E; Humphrey CD; Jacks AS; Jagota N; Mackinnon J; Mohan G; Ness DK; O’Donovan MR; Smith MD; Vudathala G; Yotti L, A

Regul Toxicol Pharmacol. Author manuscript; available in PMC 2020 December 01.
rationale for determining, testing, and controlling specific impurities in pharmaceuticals that possess potential for genotoxicity. Regul Toxicol Pharmacol 2006, 44 (3), 198–211. https://doi.org/10.1016/j.yrtph.2005.12.001 [PubMed: 16412543]

45. Myatt GJ; Ahlberg E; Akahori Y; Allen D; Amberg A; Anger LT; Aptula A; Auerbach S; Beilke L; Bellion P; Benigni R; Bercu JP; Booth ED; Bower D; Brigo A; Burden N; Cammerer Z; Cronin MTD; Cross KP; Custer L; Dettwiler M; Dobo K; Ford KA; Fortin MC; Gad-McDonald SE; Gellatly N; Gervais V; Glover KP; Glowienke S; Van Gompel J; Gutsell S; Hardy B; Harvey JS; Hillegass J; Honna M; Hsieh JH; Hsu CW; Hughes K; Johnson C; Jolly R; Jones D; Kemper R; Kenyon MO; Kim MT; Kruhlak NL; Kulkarni SA; Kummerer K; Leavitt P; Major B; Masson S; Miller S; Moser J; Mustaz M; Muster W; Neilson L; Oprea TI; Patlewicz G; Paulino A; Lo Piparo E; Powley M; Quigley DP; Reddy MV; Richarz AN; Ruiz P; Schilter B; Serafimova R; Simon W; Stavitskaya L; Stidl R; Suarez-Rodriguez D; Szabo DT; Teasdale A; Trejo-Martin A; Valentin JP; Vuorinen A; Wall BA; Watts P; White AT; Wichard J; Witt KL; Woolley A; Woolley D; Zwickl C; Hasselgren C. In silico toxicology protocols. Regul Toxicol Pharmacol 2018, 96, 1–17. 10.1016/j.yrtph.2018.04.014 [PubMed: 29678766]

46. NTP https://tools.niehs.nih.gov/cebs3/ntpViews/?activeTab=detail&studyNumber=A37789 (accessed 05/10/2019). 80.

47. Piegorsch WW; Zeiger E, Measuring intra-assay agreement for the Salmonella assay. Lecture Notes in Medical Informatics 1991, 43 (35).

48. Powley MW, (Q)SAR assessments of potentially mutagenic impurities: a regulatory perspective on the utility of expert knowledge and data submission. Regul Toxicol Pharmacol 2015, 71 (2), 295–300. https://doi.org/10.1016/j.yrtph.2014.12.012 [PubMed: 25545315]

49. Roberts G; Myatt GJ; Johnson WP; Cross KP; Blower PE Jr., LeadScope: software for exploring large sets of screening data. Journal of chemical information and computer sciences 2000, 40 (6), 1302–14. [PubMed: 11128088]

50. Rouse R; Kruhlak N; Weaver J; Burkhart K; Patel V; Strauss DG, Translating New Science Into the Drug Review Process: The US FDA’s Division of Applied Regulatory Science. Therapeutic innovation & regulatory science 2018, 52 (2), 244–255. 10.1177/2168479017720249 [PubMed: 29568713]

51. Saiakhov R; Chakravarti S; Klopman G, Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs. Molecular informatics 2013, 32 (1), 87–97. 10.1002/minf.201200081 [PubMed: 27481026]

52. SCoCP Opinion on 2-Amino-3-hydroxypyridine. https://ec.europa.eu/health/scientific_committees/consumer_safety_en

53. Scott H; Walmsley RM, Ames positive boronic acids are not all eukaryotic genotoxins. Mutation research. Genetic toxicology and environmental mutagenesis 2015, 777, 68–72. 10.1016/j.mrgentox.2014.12.002 [PubMed: 25726177]

54. Segall MD; Barber C, Addressing toxicity risk when designing and selecting compounds in early drug discovery. Drug Discovery Today 2014, 19 (5), 688–693. [PubMed: 24451294]

55. Seifried HE; Seifried RM; Clarke JJ; Junghans TB; San RH, A compilation of two decades of mutagenicity test results with the Ames Salmonella Typhimurium and L5178Y mouse lymphoma cell mutation assays. Chemical research in toxicology 2006, 19 (5), 627–44. 10.1021/tx0503552 [PubMed: 16696565]

56. Stavitskaya L, Aubrecht J, Kruhlak, Naomi L, Chemical Structure-Based and Toxicogenomic Models In Genotoxicity and Carcinogenicity Testing of Pharmaceuticals, Graziano MJ, Jacobson-Kram, David, Ed. Springer: 2015; pp 13–34. 10.1007/978-3-319-22084-0

57. Stavitskaya L; Minnier BL; Benz RD; Kruhlak N, Development of Improved QSAR Models for Predicting A-T Base Pair Mutations In 2013 Genetic Toxicology Association Meeting, Newark, DE, 2013a.

58. Stavitskaya L; Minnier BL; Benz RD; Kruhlak N, Development of Improved Salmonella Mutagenicity QSAR Models Using Structural Fingerprints of Known Toxicophores In 52nd Annual Society of Toxicology Annual Meeting and ToxExpo, San Antonio, TX, 2013b.

59. Sutter A; Amberg A; Boyer S; Brigo A; Contrera JF; Custer LL; Dobo KL; Gervais V; Glowienke S; van Gompel J; Greene N; Muster W; Nicolette J; Reddy MV; Thybaud V; Vock E; White AT; Muller L, Use of in silico systems and expert knowledge for structure-based assessment of
potentially mutagenic impurities. Regul Toxicol Pharmacol 2013, 67 (1), 39–52. https://doi.org/10.1016/j.yrtph.2013.05.001 [PubMed: 23669331]

60. Valerio LG Jr.; Cross KP, Characterization and validation of an in silico toxicology model to predict the mutagenic potential of drug impurities. Toxicology and applied pharmacology 2012, 260 (3), 209–21. 10.1016/j.taap.2012.03.001 [PubMed: 22426359]

61. Votano JR; Parham M; Hall LH; Kier LB, New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. Molecular diversity 2004, 8 (4), 379–91. [PubMed: 15612642]

62. Williams RV; Amberg A; Brigo A; Coquin L; Giddings A; Glowienke S; Greene N; Jolly R; Kemper R; O’Leary-Steele C; Parenty A; Spirkl HP; Stalford SA; Weiner SK; Wichard J, It’s difficult, but important, to make negative predictions. Regul Toxicol Pharmacol 2016, 76, 79–86. [PubMed: 26785392]

63. Zeiger E; Anderson B; Haworth S; Lawlor T; Mortelmans K, Salmonella mutagenicity tests: IV. Results from the testing of 300 chemicals. Environmental and molecular mutagenesis 1988, 11 Suppl 12, 1–157.

64. Zeiger E; Anderson B; Haworth S; Lawlor T; Mortelmans K; Speck W, Salmonella mutagenicity tests: III. Results from the testing of 255 chemicals. Environmental mutagenesis 1987, 9 Suppl 9, 1–109.

65. Zeiger E; Ashby J; Bakale G; Enslein K; Klopmann G; Rosenkranz HS, Prediction of Salmonella mutagenicity. Mutagenesis 1996, 11 (5), 471–84. [PubMed: 8921509]

66. Zhu Q; Li T; Wei X; Li J; Wang W, In silico and in vitro genotoxicity evaluation of descarboxyl levofloxacin, an impurity in levofloxacin. Drug and chemical toxicology 2014, 37 (3), 311–5. 10.3109/01480545.2013.851691 [PubMed: 24224725]
Highlights

• (Q)SAR models for predicting bacterial mutagenicity of drug impurities under ICH M7
• Published study calls with discrepant and/or deficient experimental studies were re-reviewed for 1,140 chemicals
• Expanded bacterial mutagenicity training sets (n = 9,254 and n = 13,514)
• Case studies illustrating model application
• Impact of new models on frequency of equivocal and out-of-domain results for drug impurities
Figure 1.
Changes (Δ) in sensitivity, specificity, positive predictivity, negative predictivity, and coverage between *Salmonella/E. coli* cumulative predictions and bacterial mutagenicity models in external validation.
Figure 2.
Changes (Δ) in model statistics when a single statistical model is compared against the model combined with Derek Nexus.
Figure 3.
Percentage of out-of-domain (OOD) calls for different model combinations in external validation.
Figure 4.
Selected primary aromatic amine fragments present in the CU composite bacterial mutagenicity model. Mean activity values are shown adjacent to or below each feature. Values above 0.5 are considered positive or activating features, while features below 0.5 are considered negative or deactivating features. Asterisks represent non-hydrogen atoms.
Figure 5.
Selected primary aromatic amine features present in the LS composite bacterial mutagenicity model. Mean values are presented above activating features and below deactivating features. Mean values above 0.5 are considered positive or activating features, while features below 0.5 are consider negative or deactivating features. (Ak = Alkyl carbon, Ar = Aromatic carbon)
Figure 6.
Model fragments and features representing toxicophores across CASE Ultra and Leadscope, and their mean activity values when transitioning from the Salmonella mutagenicity model to composite bacterial mutagenicity models. Asterisks represent non-hydrogen atoms, X represents any halogen, Ar represents an aromatic ring, and Q represents any atom other than carbon or hydrogen.
Figure 7:
Case Study 1: Software predictions and supporting information for the assessment of solriamfetol
Figure 8:
Case study 2: Software predictions and supporting information for the assessment of amifapridine
Figure 9:
Case Study 3: Software predictions and supporting information for the assessment of triclabendazole
Table 1.
Selected examples of newly-approved drugs and non-drugs added to the bacterial mutagenicity training sets. Previously under-represented functional groups are highlighted in red for non-drug compounds.

| New Drug Additions | Non-drug Additions |
|--------------------|--------------------|
| Structure          | Drug Name          | Drug Class | Structure          | Chemical Name                  | Functional group |
| ![Naloxegol](image1) | Naloxegol          | Opioids    | ![4-Methoxybenzenediazonium chloride](image2) | 4-Methoxybenzenediazonium chloride | Diazonium       |
| ![Clobazam](image3) | Clobazam           | Benzodiazepines | ![S-(1,2,2-trichlorovinyl)-L-cysteine](image4) | S-(1,2,2-trichlorovinyl)-L-cysteine | Vinyl Dichloride |
| ![Uridine Triacetate](image5) | Uridine Triacetate | Nucleic Acid Analog | ![N-(4-(2-chloropropanoyl) phenyl) acetamide](image6) | N-(4-(2-chloropropanoyl) phenyl) acetamide | Aromatic Amide  |
Table 2.
Validation statistics for single and combined models. Columns 2 and 3 show cross-validation performance statistics and columns 4–9 show external validation performance statistics.

| Software Platforms | Cross-Validation | External Validation (n = 388) |
|--------------------|------------------|-------------------------------|
|                    | CU               | LS | CU | LS | CU/DX | LS/DX | CU/LS/DX |
| **Model Type**     | **Bacterial**    |     | **Salm/E. coli** | **Bacterial** | **Salm/E. coli** | **Bacterial** | **Bacterial** | **Bacterial** |
| Sensitivity        | 90.6%            | 84.7% | 84.7% | 65.6% | 69.4% | 82.3% | 85.5% | 90.9% | 92.6% |
| Specificity        | 86.6%            | 86.1% | 72.0% | 84.7% | 80.3% | 83.2% | 72.2% | 73.0% | 68.4% |
| Positive Predictivity | 88.6%        | 87.9% | 42.0% | 51.9% | 42.0% | 52.6% | 44.7% | 44.4% | 42.3% |
| Normalized Pos Pred | NA              | NA | 75.5% | 82.1% | 75.5% | 82.5% | 77.5% | 77.3% | 75.7% |
| Negative Predictivity | 88.9%       | 82.5% | 89.4% | 90.8% | 92.8% | 95.4% | 95.0% | 97.1% | 97.4% |
| Normalized Neg Pred | NA              | NA | 82.2% | 69.7% | 75.0% | 82.9% | 81.7% | 88.8% | 89.7% |
| Concordance        | 77.5%            | 85.3% | 74.4% | 80.9% | 78.5% | 83.0% | 75.0% | 76.5% | 73.2% |
| Coverage           | 85.5%            | 92.8% | 94.6% | 95.9% | 85.3% | 95.9% | 95.9% | 95.9% | 99.7% |

Regul Toxicol Pharmacol. Author manuscript; available in PMC 2020 December 01.
Table 3.
Total number of out-of-domain (OOD) results in the external validation set.

| Model               | Frequency of OOD |
|---------------------|------------------|
| CU Salm/E.coli      | 21               |
| CU Bacterial        | 16               |
| LS Salm/E.coli      | 57               |
| LS Bacterial        | 16               |
| LS/CU/DX Salm/E.coli| 4                |
| LS/CU/DX Bacterial  | 1                |