Towards Quantifying the Uncertainty in *In Silico* Predictions using Bayesian Learning

Timothy E. H. Allen,[*1] *Alistair M. Middleton,*[2] *Jonathan M. Goodman,*[3] *Paul J. Russell,*[2] *Predrag Kukic,*[2] *and Steve Gutsell*[2]

[1]MRC Toxicology Unit, University of Cambridge, Gleeson Building, Tennis Court Road, Cambridge CB2 1QR, United Kingdom

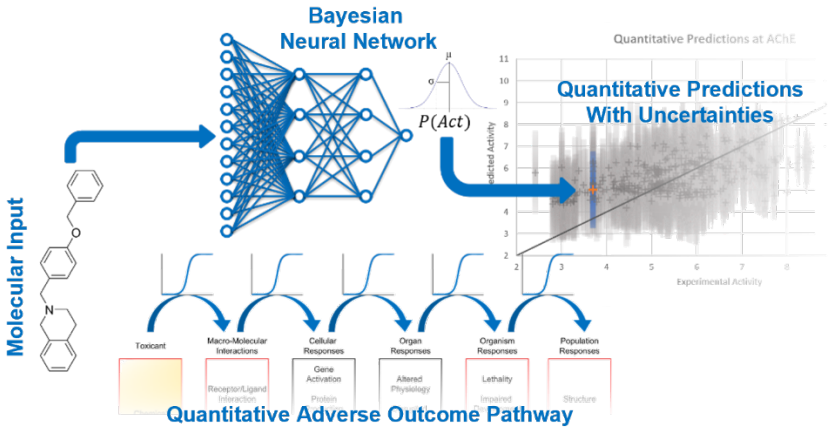
[2]Unilever Safety and Environmental Assurance Centre, Colworth Science Park, Sharnbrook, Bedfordshire, MK44 1LQ, United Kingdom

[3]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom

*E-mail: teha2@cam.ac.uk

KEYWORDS: Machine Learning, Bayesian Learning, Computational Toxicology, Molecular Initiating Event (MIE), Risk Assessment, Human Health.

Table of Contents Graphic



ABSTRACT

Next-generation risk assessment (NGRA) involves the combination of *in vitro* and *in silico* models for more human-relevant, ethical, and sustainable human chemical safety assessment. NGRA requires a quantitative mechanistic understanding of the effects of chemicals across human biology (be they molecular, cellular, organ-level or higher) coupled with a quantitative understanding of the uncertainty in any experimentally measured or predicted values. These values with their uncertainties can then be considered as a probability distribution, which can then be compared to exposure estimates to establish the presence or absence of a margin of safety. We have constructed Bayesian learning neural networks to provide such quantitative predictions and uncertainties for 20 pharmacologically important human molecular initiating events. These models produce high quality quantitative estimates (p(IC50), p(EC50), p(Ki), p(Kd)) of biochemical activity at a molecular initiating event (MIE) with average mean absolute errors (in Log units) of 0.625 ± 0.048 in test data and 0.941 ± 0.215 in external validation data. The key advantage of these models is their ability to also produce standard deviations and credible intervals (CIs) to quantify the uncertainty in these predictions, which we show to be able to distinguish between molecules close to the training data in chemical structure, those less similar to the training data, and decoy compounds drawn from the wider ChEMBL database. These uncertainty values mean that when a prediction is made a user can understand the certainty of the prediction, similar to a quantitative applicability domain, aiding prediction usefulness in NGRA. The ability for *in silico* methods to produce quantitative predictions with these kinds of probability distributions will be vital to their further use in NGRA, and here clear first steps have been taken.

INTRODUCTION

The safety evaluation of new chemicals is of the utmost importance for the protection of the health of consumers, workers, and the environment. To ensure these protections an understanding is required of the hazard associated with the test chemical and the exposure of the individual(s) to the test chemical. In order to use these quantities in risk assessment, it is vital associated uncertainties are appropriately quantified[1–6]. One approach is to adopt a probabilistic description of the uncertainty; by considering both the hazard and exposure as quantitative values with their own uncertainties which can be treated as probability distributions. Any overlap of these probability distributions is a potential concern to health.[7–9] A significant gap between these distributions represents a margin of safety that can be used to argue for regulatory acceptance of a chemical's safe use (Figure 1). As such, characterising these distributions quantitatively is of vital significance.

Toxicology is undergoing a paradigm shift, away from reliance on *in vivo* animal experiments and towards methods based on understanding the mechanisms behind toxicological effects.[10] One such framework which has gained traction over the last 10 years is the adverse outcome pathway (AOP).[11–13] AOPs link key events across the spectrum of biological organisation, from a molecular initiating event (MIE)[14,15] through cellular and tissue events to an adverse outcome at an organ, organism or population level. Most of the AOPs developed are currently qualitative,[16] but it is the ambition of safety scientists to move forward to a time when all of these relationships are quantified into a quantitative AOP (qAOP).[17–19] Many new approach methodologies (NAMs) are now being developed to measure the effects of chemicals throughout AOPs. *In vitro* assays and *in silico* calculations have a key role to play in this, and their importance in next-generation risk assessment (NGRA) is increasing. NAMs such as these are vital for the progress of NGRA,[3,4,20] and can have many advantages over *in vivo* animal experiments, including more rapid and

efficient chemical evaluation, more relevance to human health endpoints and fewer ethical concerns.[21–24] Another important driver for NAMs has been legislation, such as the 7th amendment to the European Cosmetics Directive, which completely outlawed animal testing on cosmetic ingredients in 2013,[25] the European Union Directive 2010/63/EU, which aims to further protections for animals used in scientific research,[26] and the commitment of the United States Environmental Protection Agency to phase out all animal tests on mammals by 2035.[27]

In silico toxicology methods are seeing more interest due to several factors, including an increase in the number of data points available for modelling, the amount of computer power available and new techniques such as machine learning. Many of the models developed focus on hazard prediction, making binary or categorical predictions of molecular activity, classifying molecules as "active" or "inactive", or categorizing them as "high", "moderate" or "low" risk. A non-exhaustive list of methods includes models for carcinogenicity,[28,29] mutagenicity,[30–32] hepatotoxicity,[33–35] cardiotoxicity,[36,37] developmental toxicity,[38] mitochondrial toxicity,[39,40] and MIEs such as receptor binding and enzyme inhibition.[41–45] A Bayesian neural network approach, with methodological similarity to the method presented here, has also been used to produce these classification outputs with estimations of uncertainty.[46] Classification predictions of toxicity are useful in certain circumstances, such as during compound development when a large number of molecules are screened to remove the ones most likely to lead to toxic endpoints or for the prediction of assays like the Ames test, which itself provides a binary output. Many circumstances though require a quantitative output, that can be used as an estimate for or to calculate, a point of departure (POD), the dose at which potentially adverse health conditions may be observed, which is required for use in risk assessment.[4]

Some computational models have been developed to deliver quantitative predictions, particularly in the field of quantitative structure-activity relationships (QSARs).[47] These tend to be associated with predictions for acute oral toxicity,[48] or sensitization.[49–51] As noted above, quantifying uncertainty in these computational predictions is of great importance for modern approaches to NAM development.[3,52–54] Some uncertainty estimation in computational modelling is qualitative rather than quantitative,[55,56] suggesting where confidence is "high" or "low", but lacking a quantitative measure of this. These classification predictions can be useful, but not are as beneficial in a risk assessment setting, or when developing and using qAOPs as quantitative modelling. Toxicity endpoints which are currently defined in a classification way, may not require a quantitative output, but could benefit from it. Quantitative uncertainties are available in a few cases,[57,58] including in acute fish toxicity estimation[59] and repeat-dose toxicity POD calculation[60] but to our knowledge not yet in the area of MIE predictions. Making such quantitative estimations of uncertainty is undoubtedly a challenge, but important to the use of *in silico* calculations in future risk assessments.[1]

In this work, we aim to investigate the feasibility of modelling quantitative molecular activity at human MIEs using neural networks. Modelling of these MIEs gives more mechanistic understanding of the effects a chemical is having when compared to the prediction of toxicological endpoints, as it allows exploration of a compounds toxicity through an AOP, and fits well into the goals of NGRA. Bayesian learning has been implemented, in which internal machine learning parameters (weights and biases), that are treated as fixed values in traditional machine learning, are replaced by probability distributions.[61,62] With these models we aim to provide quantification of overall uncertainty in the model predictions, which can either arise due to aleatoric uncertainty (which represents uncertainty caused by variability in experimental measurements) and epistemic uncertainty (which represents uncertainty caused by a lack of knowledge in how a particular

model describes the underlying processes and relationships of interest).[63,64] The resultant output prediction from the Bayesian neural network is itself a probability distribution (the so-called posterior predictive distribution) from which credible intervals (CIs), which are the Bayesian statistics equivalent to confidence intervals in non-Bayesian (i.e. frequentist) analysis, can be estimated. The networks were trained using variational inference (see materials and methods), resulting in the posterior predictive distributions taking the form of multivariate normal distributions. Thus, the outputs can be summarised in terms of a mean corresponding to the predicted activity and a standard deviation that provides measure of uncertainty in that prediction. From a risk assessment perspective, CIs are typically used to quantify the uncertainty in a particular outcome; typically the 95% credible interval is used as it presents a reasonably conservative range to base decisions on.[4,6–9] Such Bayesian learning algorithms also have the advantage of regularizing to prevent overfitting.[46,61,62] Bayesian neural networks have the potential to make quantitative predictions of molecular bioactivity suitable for NGRA. A workflow showing this procedure, from chemical structure to quantitative hazard assessment, is shown in Figure 2.

MATERIALS AND METHODS

Training Validation and Test Data

Bioactivity data for 21 pharmacologically important human biological targets were extracted from the publicly available database ChEMBL (Version 23, Table 1).[65] These targets are from the Bowes set, having been identified as important for *in vitro* pharmacological profiling at major pharmaceutical companies, have the potential to provide valuable toxicological information for risk assessment and fit within the definition of an MIE.[14,66] These MIEs have previously been modelled using structural alerts, random forests and neural networks for binary predictions of molecular activity and all had over 1000 quantitative datapoints (IC50, EC50, Ki, Kd) for

modelling.[43,44,67,68] Activity reports with a confidence score of less than eight were removed, leaving only reports from assays that measure changes to the function of a single protein target directly or through a homologous single protein target. This choice ensures experimental datapoints measure activity at specific and well-defined human targets. Quantitative activities (p(IC50), p(EC50), p(Ki), p(Kd)) were extracted for each compound and common salts and counterions were stripped with RDKit[69] Salt Stripper. Duplicate data points were removed using newly canonicalized SMILES and confirmed using InChIs. Where compounds did have duplicate datapoints, the highest activity was maintained for the most conservative estimate of molecular activity. The 21 targets used here all had datasets of over 1000 quantitative data points, which was found to be necessary for Bayesian learning neural networks in preliminary studies. In total 64133 unique compound-target quantitative relationships were obtained.

External Validation Data

Additional bioactivity data was downloaded from ChEMBL (Version 25). The updated version of the ChEMBL database contains additional datapoints for each target compared to ChEMBL 23. The ChEMBL 25 data was filtered and treated in the same way as previously with the ChEMBL 23 data. For each of the 21 biological targets, any chemicals with activity reported in the target's training, validation or test sets were removed to give only new data points. This gives a completely external data set totalling 5838 data points (Table 1). Despite being drawn from ChEMBL, as the training, validation and test sets are, the external validation set shows lower Tanimoto similarities to molecules in the training data when compared to the test set (Figure S2). This shows the external validation set is drawn from a different area of chemical space when compared to the training, validation, and test data, allowing us to explore how generalisable predictions of the Bayesian neural networks are.

Molecular Representation

Chemical fingerprints were generated using RDKit for Python.[70] A previous study concerning similar training data shows that the best neural network models are produced using ECFP4 fingerprints at length 10000[44] and hence these have been chosen for this study as well. This fingerprinting methods was used on each dataset for each target, meaning all models were trained on the same type of fingerprint.

Model Validation Strategy

Model statistical performance of these networks as quantitative biological activity predictors was evaluated primarily using mean absolute errors (MAEs), and mean squared errors (MSEs) and coefficients of determination for specific linear correlation between experimental and predicted activities (R^2) are also included for completeness.[71,72] The ChEMBL 23 dataset was split randomly into 75% training/validation data and 25% test data. The test data was not used during hyperparameter optimization to prevent hyperparameter overfitting. The training/validation data was then split randomly into 75% training and 25% validation data. After hyperparameter optimization, the test data and the ChEMBL 25 external validation data were evaluated using the best models.

Neural Network Architecture

Regression neural networks were constructed and trained using TensorFlow (TF) in Python 3. A grid search was used to obtain the highest performing architectures including one or two hidden layers, and 10, 50 and 100 neurons per hidden layer. Wider and deeper networks were constructed during preliminary studies, but the statistical performance was found to be poor and models tended not to train. The best performing networks were chosen based on having the

lowest MAE values. ReLU (rectified linear unit) activation functions were used to provide non-linearity based on our previous work.[44] The Adam optimizer and negative log-likelihood loss function were used in model training. Chemical features were input as ECFP4 fingerprints of length 10000 and quantitative estimation of biological activity at a target was provided as the output.

Bayesian Inference

To model the uncertainty in the neural network outputs given a chemical fingerprint input, the models were trained using Bayesian inference. This provides a natural self-contained framework for modelling epistemic and aleatoric uncertainty. For a brief overview of Bayesian inference in the context of neural network learning and additional details on the Bayesian neural network see the Supplementary Material. For a detailed introduction to Bayesian statistics, see Gelman *et al*.[73]

Bayesian neural network construction and training was implemented using TF Probability. The various neural network layers were implemented using Dense Variational Layers, while the final output layer contains the likelihood, which was implemented using the TF Lambda Distribution function.

Predictions made by the model for evaluation were calculated using a Monte Carlo simulation with 500 iterations. These samples were used to estimate the mean and standard deviation of each prediction output obtained from the posterior predictive distribution. The uncertainties produced by these models can be seen as a form of quantitative applicability domain (AD), assisting these models in meeting OECD guidelines for QSAR models.[74] While a classical AD defines input examples as within or outside the models predictive domain, suggesting cases within are likely to be well predicted and those outside are more likely to be incorrect or untrustworthy, these quantitative values can imitate this idea on a scale, where low values will be

given with confident predictions for cases similar to the training data and higher values correspond to less certain predictions.

7363,6461,6275,7673616174**Accounting for Uncertainty**

As previously mentioned, an important part of the implementation of Bayesian learning in this prediction task is the ability of the models to provide meaningful uncertainty estimates when predictions are made for use in NGRA. Model calibration was evaluated by considering various CIs and their fit to the experimental data,[77–79] whereby if a model is well-calibrated it is expected that an x% credible interval would contain the true outcome of the predicted activity x% of the time. Calibration plots, whereby this comparison is provided for a range of CIs, were generated for each model and data set (training, test, and external validation) as follows: for each target credible interval between 5% and 95%, in increments of 10%, the percentage of true outcomes (i.e. activities) contained within those intervals were calculated. In the calibration plots, target CIs are plotted against the true outcome percentages.

To evaluate the usefulness of the standard deviation values provided by the Bayesian learning model, predictions were made on test set and external validation set chemicals, and decoy compounds chosen randomly from among the ChEMBL data for other biological targets. For each model, a list of 1000 decoy compounds drawn from the ChEMBL data for other biological targets was generated for comparison as a sample of the total ChEMBL chemical space. These compounds are not part of, or drawn from, the training data for the Bayesian model and hence are expected to produce higher standard deviation values than the test set on average.

Applicability Domain

An AD score has been constructed using the Tanimoto similarity between the ECFP inputs used to train the Bayesian neural network models for comparison to the Bayesian neural networks provided standard deviation values.[80] Tanimoto similarity is a suitable choice for an AD for this model as it uses the same chemical information fed to the model (an ECFP). For each prediction, the Tanimoto similarity is calculated between the input molecule and all compounds in the training set. These values are then extracted to identify how similar the input compound is to the training set. The highest 10 values are then averaged to give a 10-nearest neighbour score (applicability domain 10, or AD10 value), to provide context of how well chemical space around the prediction is captured in the training data. These domains have been applied to the Bayesian neural networks models presented here to identify if a specific threshold can be used to define an applicability domain. As a comparison point to the standard deviation uncertainty values provided by the models, model calculated uncertainties (the standard deviation values) and the AD10 values, were plotted against absolute error in prediction and bin-based averages (BBAs) included across 10 equal folds in the dataset, as described by Sushko.[81]

Assessment of Best Models

The statistical performance of models was performed using MAE, where models with the lowest MAE values are considered the best. This evaluation, however, ignores the important issue of capturing quantitative uncertainty estimates within these Bayesian models. Because of this, two additional criteria were also considered:

1) Model predictions on the test set and decoy set should produce statistically significantly different distributions of standard deviations (evaluated using a 2-sample unequal variance T-Test due to the unequal variances of the approximately normal distributions of the predicted standard deviations and the large number of samples, at $\alpha = 0.01$).

2) The mean decoy set standard deviation should be higher than the mean test set standard deviation.

[43,44]The Python code for model construction and recall, the developed models, and the training, validation, test and external validation data are available through GitHub (https://github.com/teha2/chemical_toxicology).

RESULTS AND DISCUSSION

Data Distribution

The distribution of molecular activities across all datasets is shown in Figure 3 and appears close to a normal distribution. The distributions primarily fall between 4 and 9 Log units, peaking between 5 and 7. The distributions of the training, validation, test and external validation datapoints for each target dataset are available in the Supplementary Material (Figure S1). In several cases the distributions show high frequency at specific activities, and this is particularly pronounced for some of the external validation distributions. This is due to the reporting of the experimental data, such as reporting of experimental activities such as 10 μM or p(Activity) = 5 (in the case of the serotonin transporter (SLC6A4) and the serotonin 2a receptor (HTR2A) external validation charts), 30 μM or p(Activity) = 4.523 (tyrosine-protein kinase (LCK)) or 0.1 μM or p(Activity) = 7 (delta opioid receptor (OPRD1)). These compounds were checked manually to ensure no duplication of chemicals. Several series of similar chemicals were identified, suggesting that some of these results are coming single studies in the CHEMBL database – i.e. the data points refer to a chemical series. This potentially suggests a challenge for the models, as these chemicals are all represented by different chemical fingerprints but anticipate the same quantitative prediction output. This will be a particular challenge for the model if these chemicals are dissimilar to the training data.

Model Performance

Models were assessed based on the criteria outlined for model performance and uncertainty estimation above. Models meeting these criteria were produced for every target except KCNH2. In each case, the model passing the criteria with the lowest validation MAE is presented as the best, and the performance statistics of these best models for each biological target against the training, validation and test sets are shown in Table 2. All statistics for all trained models are included in the Supplementary Material (Table S1). The KCNH2 dataset has proven challenging to model in the past,[43,44] and perhaps this suggests why passing the criteria was challenging. As a result of this the KCNH2 model has been removed from subsequent analysis.

Overall, the best performing models for each MIE have produced MAEs well within one log unit, with an average MAEs of 0.486 (training set), 0.618 (validation) and 0.625 (test). Unsurprisingly the MAE increases going from training to validation to test data, but these increases are modest and do not indicate large levels of overfitting. Ultimately, all models are overfitted, the importance is to restrict overfitting to give the models generalizability, which is indicated by the modest increase in MAE going between each set (0.132 on average from training to validation and 0.007 from validation to test).

Example test set predictions are shown graphically in Figure 4 for four targets chosen at random, the acetylcholinesterase (AChE), HTR2A, the glucocorticoid receptor (NR3C1), and the dopamine transporter (SLC6A3). Graphs for all targets are shown in Figure S3 and predictions for all test set compounds combined in Figure S4 in the Supplementary Material. These graphs show a

general trend of the predictions aligning to the x = y diagonal. Molecules with low experimental activity values tend to be overpredicted and those with high activities tend to be underpredicted. This is due to the majority of the training data falling in the middle and pushing the predictions towards the centre. The gathering of additional datapoints at the ends of this distribution may assist this in the future, although the nature of such data will always make the ends of the distribution the most challenging to predict.

The test set statistical performance of these models can be compared to bespoke 3D QSAR models previously produced to predict MIE-level molecular activity using comparative molecular field analysis (CoMFA) for some of the same biological targets using similar data from ChEMBL.[82] CoMFA models were constructed using five structural categories for the μ-opioid receptor (average root MSE = 0.59), glucocorticoid receptor (0.52) and dopamine transporter (0.71). The Bayesian neural networks perform slightly less well for the μ-opioid receptor (average MSE = 0.753) and glucocorticoid receptor (0.798), and slightly better for the dopamine transporter (0.595). The Bayesian neural networks are also easier to implement, as all novel compounds can be assessed by a single neural network per target, while CoMFA requires molecular alignment amongst structurally similar training set compounds, and also provides uncertainty estimates with those predictions. The addition of uncertainty estimates in the predictions was a key driver in the choice of Bayesian neural networks for this study, and it is appreciated that other machine learning, classical or consensus approaches may be able to provide better predictive statistical performance (MAE, MSE or R^2) on the data considered in this work.

External Validation

The performance of the best models against the external validation data, alongside the figures for the test data, is shown in Table 2. A decrease in model performance on moving from the test set

to the external validation set is expected, and this is observed. On average MAE increases by approximately 50% to 0.942. The statistical performance on the external validation set is also potentially influenced by the size of these sets, as in six cases fewer than 100 external validation data points are evaluated. On average these predictions remain within one log unit of the experimental values.

Model performance appears less good when considering R^2, specifically when looking at the external validation data. Eight of the 20 cases have R^2 values greater than 0.4, suggesting similar model performance to the test set cases, five are between 0 and 0.4 showing a small amount of predictivity on these datasets and seven are negative. Meaning these models are not predictive on these datasets when compared to the mean activity of that dataset. This can be caused by small external validation datasets containing molecules with a similar experimental activity, such as the cases for ADRB1 and EDNRA. Otherwise, additional data may be able to assist with this in future work. Ultimately, when such predictions are not good the model recognises this and provides higher standard deviations with their predictions, alerting the user to this and urging more caution.

Some external validation set predictions are shown graphically in Figure 5 for the same targets as Figure 4. Graphs for all targets are shown in Figure S5 and a graph for all predictions combined in Figure S6 in the Supplementary Material. The trends of these predictions are more distant from x = y than for the test set as expected but are still generally close, with cases at high or low activities generally less well predicted. External validation dataset predictions are more challenging for the model as is to be expected. The range of predicted values is narrower than that observed for the test data (For AChE the range of predictions on the test set is around 6 log units, while for the external validation set it is around 4), however, the predictions are not all

clustered at the mean training set activity, which suggests that the model has learned about these molecular activities.

Averages of the standard deviation values provided by the Bayesian neural network are plotted against average model performance (MAE or R^2) for the test and external validation sets of each MIE in Figure 6. These plots do not show high correlation, as might be considered ideal, but do show that higher standard deviations generally correlate with poorer model performance in the cases of MAE on both the test and external validation sets, and R^2 on the external validation data.

Understanding Uncertainties

The average standard deviation values associated with each prediction were extracted and examined (Table 3). These values correspond well with what is observed regarding predictivity, with relatively high standard deviations produced for the external validation set predictions on the EDNRA, HRH1 and SLC6A4, cases where the model particularly struggled. These values provide a context for those predictions, allowing for caution when they are considered in NGRA, or other contexts.

Calibration plots were generated for each model to assess how well calibrated the generated models are. Figure 7 shows four examples and Figure S7 shows plots for all targets. Points on the plot are expected to approximately follow the x = y diagonal to indicate a well calibrated model, which is generally shown. Training set points are expected to lie above test set points which are in turn above external validation set points, as predictions on the test and particularly external validation sets are expected to be more challenging.

For each test and external validation set, the 95% CIs were calculated, using the approximation of two times the standard deviation, and the percentage of molecules whose experimental activity

values fall within this interval were calculated (Table 3). The average percentage obtained within two standard deviations was 96.63% for the test sets and 90.49% for the external validation sets. For seven of the models test set data, the percentage of predictions within the 95% CI values lie within +/- 1% of 95.45%. Seven more within a total of +/- 2%, five more within a total for +/- 4% and only one outside this (CHRM3). Interestingly, the more distant datasets are not particularly small, or poorly predicted based on their test set statistical performance values, showing the uncertainty element of this task is not entirely dependent on the dataset size of model statistical performance. This alludes to an important challenge in this probabilistic modelling – getting a good balance between quality of prediction and useful uncertainty estimates. The values for the external validation data are more distant from the expected value of 95.45%, showing that these distributions are not quite as well modelled. The small sizes of some of the external validation sets may be responsible for this.

Comparison histograms for four biological targets are shown in Figure 8. Test set data, external validation set data and decoy set data were evaluated as a point of comparison. Each set shows a slightly different distribution of standard deviation values. Kernel density estimates are also shown to highlight differences in the data distributions. The test set compounds tend to have the lowest standard deviation values, with external validation set compounds shifted slightly to the right, corresponding to overall higher standard deviation values, and decoy compounds shifted slightly further to the right. This is expected as the test set should be the most similar to the training set (and therefore should have activities predicted with the lowest uncertainty) with the eternal validation compounds being slightly more different, and decoy compounds the most different. The differences between these distributions for the test and decoy compounds were found to be statistically significant at $\alpha = 0.01$ in every case using a 2-sample unequal variance T-Test. Further plots like this are shown in the Supplementary Material (Figure S8) and show similar trends. Absolute values of the uncertainties for the decoy data depend on the neural network architecture,

with larger networks appearing to be able to better differentiate the decoy compounds from the test and external validation data (for example see in particular CHRM3 and EDNRA, but also AChE, CHRM1, CHRM2 and OPRD1 in Figure S8 – all modelled with a single layer of 100 neurons). Future work may investigate different network architectures and find the ones best able to model these uncertainties, and hence improve their use in NGRA.

Uncertainties and Applicability Domains

Graphs showing the relationship between model standard deviation values and AD10 values were plotted for each target against absolute error in prediction for both the test and external validation sets (Figures 9 and S9). [81]These plots show that neither the model standard deviation values or AD10 values directly correlation with absolute error, but in most cases average errors increase with lower AD10 value or higher uncertainty. Specifically average errors show a step change in BBA around a standard deviation value of 0.8-1.2 or an AD10 value of 0.2-0.3.

To assess the use of these values as absolute thresholds for an applicability domain, model statistical performance at each MIE was recalculated excluding compounds based on three criteria:

1) AD10 values less than 0.2
2) standard deviation values greater than 1.2
3) AD10 values less than 0.2 OR standard deviation values greater than 1.2 (both)

The results of this are shown in Table 4.. These results suggest using the standard deviation values as an absolute threshold for an applicability domain is unsuitable, as the MAE values are almost unchanged.The AD10 values do decrease MAE values, however the differences are modest while excluding a large number of predictions (almost 10% of the test set predictions and

nearly 40% of the external validation set predictions). Further investigations into these applicability domains may be able to provide a greater improvement in statistical performance – although the use of the in conjunction with these models is clearly a complex problem.

The 95% CIs calculated for each prediction can be added to the earlier prediction plots, as would be anticipated for the use of these models in NGRA. This is shown for acetylcholinesterase activity predictions on the training, test and external validation data in Figure 10. This graph shows that in the vast majority of cases the experimental activity lies within the 95% credible interval predicted by the model. These models can contribute to safety evaluation scenarios such as the case in Figures 1 and 2, where the hazard distribution produced by the model can be compared to the exposure distribution, perhaps calculated using physiologically based kinetic (PBK) modelling.[83–85] This is a significant step forward for *in silico* predictions of human MIEs and the provision of quantitative predictions and uncertainty estimates should be the ultimate goal of more modelling procedures moving forward.

CONCLUSIONS

Computational toxicology has key advantages when compared to *in vivo* or *in vitro* approaches, including reduced cost and faster evaluation, given that suitable modelling approaches are employed and appropriate data can be gathered.[1,3,21] *In silico* tools for the prediction of hazards associated with new chemicals are relatively common and have in general been well developed, but focus mainly on qualitative hazard prediction tasks.[28,32,39,41,44,86] This is especially true when compared to models developed to aid risk assessment through both i) the quantitative prediction of molecular activity and ii) the quantification of uncertainty in these predictions. Here we present a Bayesian learning-based approach to help answer these questions in the case of human MIE prediction. Publicly available data has been modelled for 20 human MIEs and produces models

with MAE values less than one log unit against external validation data. Models showing good statistical performance were generated for some MIEs and external validation sets, particularly the ADRB2, CHRM1, CHRM2, CHRM3, DD2R, HTR2A and OPRD1 providing external validation MAE values below 0.9 and R^2 above 0.4. These Bayesian neural networks also produce standard deviations associated with their predictions which can be used to quantify uncertainty. It has been shown that these values provide the uncertainties expected with the test and external validation data and are useful in distinguishing between molecules similar to the training data and decoy molecule input strings.

Quantification of uncertainty is considered one of the biggest challenges with NAMs and is required both to enable the wider adoption of NAMs and for NGRA to succeed.[3,5] The models presented here provide uncertainty values that can be fed into an NGRA procedure or a qAOP for the safety evaluation of a novel chemical. While the calibration curves presented in Figures 7 and S7 did show some deviation from 'perfect' calibration, the results are encouraging. Moreover, we obtained good agreement between the frequency of true outcomes at the 95% credible interval level, which is a common statistic for decision-making,[4,6–9] indicating the approach could suitable in risk assessment. This is a considerable step forward for computational toxicology and the use of Bayesian learning should be more widely used in model development moving forward. Coupling these methods with MIEs and AOPs gives an element of mechanistic understanding to the procedure, desirable in the protection of human health.

ASSOCIATED CONTENT

Supplementary Material.

Histograms showing the data distribution for compounds at each biological target, Tanimoto similarity distribution for the test and external validation sets to the training set, additional

information on the Bayesian neural network model, performance statistics for all models, graphs showing predicted vs experimental activity values for test set and external validation set compounds, calibration plots for every target and histograms showing the distributions of standard deviations for each model.

The Python code for ECFP fingerprint generation, model construction and recall, the developed models, and the training, validation, test and external validation data as SMILES strings are available through GitHub (https://github.com/teha2/chemical_toxicology) under Bayesian Learning.

AUTHOR INFORMATION

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Statement

The authors acknowledge the financial support of Unilever.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Statement

According to the University of Cambridge data management policy, all the data used in this paper is available either in the paper or in the Supplementary Material. A copy of the data is also available in the University of Cambridge repository at: https://www.repository.cam.ac.uk/

ABBREVIATIONS

AChE=acetylcholinesterase,

AD10=applicability domain 10,

ADORA2A=adenosine A2a receptor,

ADRB1=beta-1 adrenergic receptor,

ADRB2=beta-2 adrenergic receptor,

AOP=adverse outcome pathway,

AR=androgen receptor,

BBA=bin based average,

CHRM1=muscarinic acetylcholinesterase receptor M1,

CHRM2=muscarinic acetylcholinesterase receptor M2,

CHRM3=muscarinic acetylcholinesterase receptor M3,

CI=credible interval,

DD1R=dopamine D1 receptor,

DD2R=dopamine D2 receptor,

EDNRA=endothelin receptor ET-A,

HRH1=histamine H1 receptor,

HTR2A=serotonin 2a receptor,

KCNH2=human ether-a-go-go related gene channel,

LCK=tyrosine-protein kinase LCK,

MAE=mean absolute error,

MIE=molecular initiating event,

MSE=mean squared error

NAM=new approach methodology,

NGRA=next-generation risk assessment,

NR3C1=glucocorticoid receptor,

OPRD1=delta opioid receptor,

OPRM1=mu opioid receptor,

PBK=physiologically-based kinetic,

(Q)SAR=(quantitative) structure-activity relationship,

ReLU=rectified linear unit,

SLC6A2=norepinephrine transporter,

SLC6A=dopamine transporter,

SLC6A4=serotonin receptor,

TF=TensorFlow.

REFERENCES

(1) Thomas, R. S.; Bahadori, T.; Buckley, T. J.; Cowden, J.; Deisenroth, C.; Dionisio, K. L.; Frithsen, J. B.; Grulke, C. M.; Gwinn, M. R.; Harrill, J. A.; Higuchi, M.; Houck, K. A.; Hughes, M. F.; Sidney Hunter, E.; Isaacs, K. K.; Judson, R. S.; Knudsen, T. B.; Lambert, J. C.; Linnenbrink, M.; Martin, T. M.; Newton, S. R.; Padilla, S.; Patlewicz, G.; Paul-Friedman, K.; Phillips, K. A.; Richard, A. M.; Sams, R.; Shafer, T. J.; Woodrow Setzer, R.; Shah, I.; Simmons, J. E.; Simmons, S. O.; Singh, A.; Sobus, J. R.; Strynar, M.; Swank, A.; Tornero-Valez, R.; Ulrich, E. M.; Villeneuve, D. L.; Wambaugh, J. F.; Wetmore, B. A.; Williams, A. J. The next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. *Toxicological Sciences* **2019**, *169* (2), 317–332. https://doi.org/10.1093/toxsci/kfz058.

(2) Dong, Z.; Liu, Y.; Duan, L.; Bekele, D.; Naidu, R. Uncertainties in Human Health Risk Assessment of Environmental Contaminants: A Review and Perspective. *Environment International* **2015**, *85*, 120–132. https://doi.org/10.1016/j.envint.2015.09.008.

(3) Dent, M.; Teixeira, R.; Amores, P.; Silva, D.; Ansell, J.; Boisleve, F.; Hatao, M.; Hirose, A.; Kasai, Y.; Kern, P.; Kreiling, R.; Milstein, S.; Montemayor, B.; Oliveira, J.; Richarz, A.; Taalman, R.; Vaillancourt, E.; Verma, R.; Vieira, N.; Cabral, O. R.; Weiss, C.; Kojima, H. Principles Underpinning the Use of New Methodologies in the Risk Assessment of Cosmetic Ingredients. *Computational Toxicology* **2018**, *7* (June), 20–26. https://doi.org/10.1016/j.comtox.2018.06.001.

(4) Baltazar, M. T.; Cable, S.; Carmichael, P. L.; Cubberley, R.; Cull, T.; Delagrange, M.; Dent, M. P.; Hatherell, S.; Houghton, J.; Kukic, P.; Li, H.; Lee, M.; Malcomber, S.; Middleton, A. M.; Moxon, T. E.; Nathanail, A. v; Nicol, B.; Pendlington, R.;

Reynolds, G.; Reynolds, J.; White, A.; Westmoreland, C. A Next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products. *Toxicological Sciences* **2020**, *176* (1), 236–252. https://doi.org/10.1093/toxsci/kfaa048.

(5) Gosling, J. P. The Importance of Mathematical Modelling in Chemical Risk Assessment and the Associated Quantification of Uncertainty. *Computational Toxicology* **2019**, *10* (December 2018), 44–50. https://doi.org/10.1016/j.comtox.2018.12.004.

(6) Lazic, S. E.; Edmunds, N.; Pollard, C. E. Predicting Drug Safety and Communicating Risk: Benefits of a Bayesian Approach. *Toxicological Sciences* **2018**, *162* (1), 89–98. https://doi.org/10.1093/toxsci/kfx236.

(7) ECHA. Guidance on Information Requirements and Chemical Safety Assessment Chapter R . 19 : Uncertainty Analysis November 2012. **2012**, *2012* (November), 1–36.

(8) Bokkers, H. The Practicability of the Integrated Probabilistic Risk Assessment Approach for Substances in Food The Practicability of the Integrated Probabilistic Risk Assessment (IPRA) Approach for Substances in Food. *RIVM Report* **2009**, 16272–16277.

(9) SCCS. The SCCS Notes of Guidance for the Testing of Cosmetic Ingredients and Their Safety Evaluation 10th Revision. *Sccs* **2016**, *1564* (April), 151.

(10) Leist, M.; Hartung, T.; Nicotera, P. The Dawning of a New Age of Toxicology. *ALTEX* **2008**, *25* (2), 103–114.

(11) Ankley, G. T.; Bennett, R. S.; Erickson, R. J.; Hoff, D. J.; Hornung, M. W.; Johnson, R. D.; Mount, D. R.; Nichols, J. W.; Russom, C. L.; Schmieder, P. K.; Serrrano, J. A.; Tietge, J. E.; Villeneuve, D. L. Adverse Outcome Pathways: A Conceptual Framework to Support Ecotoxicology Research and Risk Assessment. *Environmental Toxicology and Chemistry* **2010**, *29* (3), 730–741. https://doi.org/10.1002/etc.34.

(12) OECD. *Proposal for a Template, and Guidance on Developing and Assessing the Completeness of Adverse Outcome Pathways, Appendix I, Collection of Working Definitions*.

(13) Leist, M.; Ghallab, A.; Graepel, R.; Marchan, R.; Hassan, R.; Hougaard, S.; Alice, B.; Mathieu, L.; Stefan, V.; Tanja, S.; Danen, E.; Ravenzwaay, B. Van; Kamp, H.; Gardner, I.; Godoy, P.; Bois, F. Y.; Leist, M. Adverse Outcome Pathways: Opportunities, Limitations and Open Questions. *Archives of Toxicology* **2017**, *204* (0123456789), 1–29. https://doi.org/10.1007/s00204-017-2045-3.

(14) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Defining Molecular Initiating Events in the Adverse Outcome Pathway Framework for Risk Assessment. *Chemical Research in Toxicology* **2014**, *27*, 2100–2112.

(15) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. A History of the Molecular Initiating Event. *Chemical Research in Toxicology* **2016**, *29* (12), 2060–2070. https://doi.org/10.1021/acs.chemrestox.6b00341.

(16) European Commission Institute for Health and Consumer Protection. AOP Wiki http://ihcp.jrc.ec.europa.eu/our_activities/alt-animal-testing-safety-assessment-chemicals/improved_safety_assessment_chemicals/adverse-outcome-pathways-aop.

(17) Perkins, E. J.; Ashauer, R.; Burgoon, L.; Conolly, R.; Landesmann, B.; Mackay, C.; Murphy, C. A.; Pollesch, N.; Wheeler, J. R.; Zupanic, A.; Scholz, S. Building and Applying Quantitative Adverse Outcome Pathway Models for Chemical Hazard and Risk Assessment. *Environmental Toxicology and Chemistry* **2019**, *38* (9), 1850–1865. https://doi.org/10.1002/etc.4505.

(18) Zgheib, E.; Gao, W.; Limonciel, A.; Aladjov, H.; Yang, H.; Tebby, C.; Gayraud, G.; Jennings, P.; Sachana, M.; Beltman, J. B.; Bois, F. Y. Application of Three Approaches for Quantitative AOP Development to Renal Toxicity. *Computational Toxicology* **2019**, *11* (February), 1–13. https://doi.org/10.1016/j.comtox.2019.02.001.

(19) Spinu, N.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Worth, A. P. Quantitative Adverse Outcome Pathway (QAOP) Models for Toxicity Prediction. *Archives of Toxicology* **2020**, *94* (5), 1497–1510. https://doi.org/10.1007/s00204-020-02774-7.

(20) Gilmour, N.; Reynolds, J.; Przybylak, K.; Aleksic, M.; Aptula, N.; Baltazar, M. T.; Cubberley, R.; Rajagopal, R.; Reynolds, G.; Spriggs, S.; Thorpe, C.; Windebank, S.; Maxwell, G. Next Generation Risk Assessment for Skin Allergy: Decision Making Using New Approach Methodologies. *Regulatory Toxicology and Pharmacology* **2022**, *131*, 105159. https://doi.org/10.1016/j.yrtph.2022.105159.

(21) Committee on Toxicity Testing and Assessment of Environmental Agents and National Research Council. *Toxicology Testing in the 21st Century: A Vision and a Strategy*; 2007.

(22) Hartung, T. Toxicology for the Twenty-First Century. *Nature* **2009**, *460*, 208–212.

(23) Gottmann, E.; Kramer, S.; Pfahringer, B.; Helma, C. Data Quality in Predictive Toxicology: Reproducibility of Rodent Carcinogenicity Experiments. *Environmental Health Perspectives* **2001**, *109* (5), 509–514.

(24) Smith, R. Animal Research: The Need for a Middle Ground. *British Medical Journal* **2001**, *322*, 248–249.

(25) The 7th Amendment to the Cosmetics Directive http://ec.europa.eu/consumers/sectors/cosmetics/files/doc/antest/(2)_executive_summary_en.pdf.

(26) European Union Directive 2010/63/EU https://ec.europa.eu/environment/chemicals/lab_animals/pdf/guidance/inspections/en.pdf.

(27) Grimm, D. U.S. EPA to Eliminate All Mammal Testing by 2035. *Science*. 2019. https://doi.org/10.1126/science.aaz4593.

(28) Li, X.; Du, Z.; Wang, J.; Wu, Z.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In Silico Estimation of Chemical Carcinogenicity with Binary and Ternary Classification Methods. *Molecular Informatics* **2015**, *34* (4), 228–235. https://doi.org/10.1002/minf.201400127.

(29) Zhang, H.; Cao, Z. X.; Li, M.; Li, Y. Z.; Peng, C. Novel Naïve Bayes Classification Models for Predicting the Carcinogenicity of Chemicals. *Food and Chemical Toxicology* **2016**, *97*, 141–149. https://doi.org/10.1016/j.fct.2016.09.005.

(30) Seal, A.; Passi, A.; Abdul Jaleel, U. C.; Wild, D. J.; Consortium, O. S. D. D. In-Silico Predictive Mutagenicity Model Generation Using Supervised Learning Approaches. *Journal of Cheminformatics* **2012**, *4* (5). https://doi.org/10.1186/1758-2946-4-10.

(31) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *Journal of Chemical Information and Modeling* **2018**, *58* (8), 1533–1543. https://doi.org/10.1021/acs.jcim.8b00338.

(32) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicological Sciences* **2018**, *165* (1), 198–212. https://doi.org/10.1093/toxsci/kfy152.

(33) Mellor, C. L.; Steinmetz, F. P.; Cronin, M. T. D. Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis. *Chemical Research in Toxicology* **2016**, *29* (2), 203–212. https://doi.org/10.1021/acs.chemrestox.5b00480.

(34) Ai, H.; Chen, W.; Zhang, L.; Huang, L.; Yin, Z.; Hu, H.; Zhao, Q.; Zhao, J.; Liu, H. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicological Sciences* **2018**, *165* (1), 100–107. https://doi.org/10.1093/toxsci/kfy121.

(35) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *Journal of Chemical Information and Modeling* **2015**, *55* (10), 2085–2093. https://doi.org/10.1021/acs.jcim.5b00238.

(36) Zhang, Y.; Zhao, J.; Wang, Y.; Fan, Y.; Zhu, L.; Yang, Y.; Chen, X.; Lu, T.; Chen, Y.; Liu, H. Prediction of HERG K+ Channel Blockage Using Deep Neural Networks. *Chemical Biology and Drug Design* **2019**, *94* (5), 1973–1985. https://doi.org/10.1111/cbdd.13600.

(37) Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *Journal of Chemical Information and Modeling* **2019**, *59* (3), 1073–1084. https://doi.org/10.1021/acs.jcim.8b00769.

(38) Zhang, H.; Ren, J. X.; Kang, Y. L.; Bo, P.; Liang, J. Y.; Ding, L.; Kong, W. B.; Zhang, J. Development of Novel in Silico Model for Developmental Toxicity Assessment by Using Naïve Bayes Classifier Method. *Reproductive Toxicology* **2017**, *71*, 8–15. https://doi.org/10.1016/j.reprotox.2017.04.005.

(39) Nelms, M. D.; Mellor, C. L.; Cronin, M. T. D.; Madden, J. C.; Enoch, S. J. Development of an in Silico Profiler for Mitochondrial Toxicity. *Chemical Research in Toxicology* **2015**, *28* (10), 1891–1902. https://doi.org/10.1021/acs.chemrestox.5b00275.

(40) Zhang, H.; Yu, P.; Ren, J. X.; Li, X. B.; Wang, H. L.; Ding, L.; Kong, W. B. Development of Novel Prediction Model for Drug-Induced Mitochondrial Toxicity by Using Naïve Bayes Classifier Method. *Food and Chemical Toxicology* **2017**, *110*, 122–129. https://doi.org/10.1016/j.fct.2017.10.021.

(41) Unterthiner, T.; Mayr, A.; Klambauer, G.; Hochreiter, S. Toxicity Prediction Using Deep Learning. **2015**.

(42) Steinmetz, F. P.; Mellor, C. L.; Meinl, T.; Cronin, M. T. D. Screening Chemicals for Receptor-Mediated Toxicological and Pharmacological Endpoints: Using Public Data to Build Screening Tools within a KNIME Workflow. *Molecular Informatics* **2015**, *34* (2–3), 171–178. https://doi.org/10.1002/minf.201400188.

(43) Wedlake, A. J.; Folia, M.; Piechota, S.; Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Structural Alerts and Random Forest Models in a Consensus Approach for Receptor Binding Molecular Initiating Events. *Chemical Research in Toxicology* **2020**. https://doi.org/10.1021/acs.chemrestox.9b00325.

(44) Allen, T. E. H.; Wedlake, A. J.; Goodman, J. M.; Russell, P. J. Neural Network Activation Similarity: A New Measure to Assist Decision Making in Chemical Toxicology. *Chemical Science* **2020**, *11*, 7335–7348. https://doi.org/10.1039/d0sc01637c.

(45) Allen, T. E. H.; Allen, T. E. H.; Nelms, M. D.; Nelms, M. D.; Edwards, S. W.; Goodman, J. M.; Gutsell, S.; Russell, P. J. In Silico Guidance for in Vitro Androgen and Glucocorticoid Receptor ToxCast Assays. *Environmental Science and Technology* **2020**, *54* (12), 7461–7470. https://doi.org/10.1021/acs.est.0c01105.

(46) Semenova, E.; Williams, D. P.; Afzal, A. M.; Lazic, S. E. A Bayesian Neural Network for Toxicity Prediction. *Computational Toxicology* **2020**, *16* (August), 100133. https://doi.org/10.1016/j.comtox.2020.100133.

(47) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min,

V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014**, *57* (12), 4977–5010. https://doi.org/10.1021/jm4004285.

(48) Nelms, M. D.; Karmaus, A. L.; Patlewicz, G. An Evaluation of the Performance of Selected (Q)SARs/Expert Systems for Predicting Acute Oral Toxicity. *Computational Toxicology* **2020**, *16* (September), 100135. https://doi.org/10.1016/j.comtox.2020.100135.

(49) Enoch, S. J.; Roberts, D. W.; Cronin, M. T. D. Electrophilic Reaction Chemistry of Low Molecular Weight Respiratory Sensitizers. *Chemical Research in Toxicology* **2009**, *22* (8), 1447–1453. https://doi.org/10.1021/tx9001463.

(50) Enoch, S. J.; Roberts, D. W. Predicting Skin Sensitization Potency for Michael Acceptors in the LLNA Using Quantum Mechanics Calculations. *Chemical Research in Toxicology* **2013**, *26* (5), 767–774. https://doi.org/10.1021/tx4000655.

(51) Ebbrell, D. J.; Madden, J. C.; Cronin, M. T. D.; Schultz, T. W.; Enoch, S. J. Development of a Fragment-Based in Silico Profiler for Michael Addition Thiol Reactivity. *Chemical Research in Toxicology* **2016**, *29* (6), 1073–1081. https://doi.org/10.1021/acs.chemrestox.6b00099.

(52) Rusyn, I.; Daston, G. P. Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. *Environmental Health Perspectives* **2010**, *118* (8), 1047–1050. https://doi.org/10.1289/ehp.1001925.

(53) Walker, J. D.; Carlsen, L.; Jaworska, J. Improving Opportunities for Regulatory Acceptance of QSARs: The Importance of Model Domain, Uncertainty, Validity and Predictability. *QSAR and Combinatorial Science* **2003**, *22* (3), 346–350. https://doi.org/10.1002/qsar.200390024.

(54) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discovery Today* **2020**, *00* (00). https://doi.org/10.1016/j.drudis.2020.11.027.

(55) Cronin, M. T. D.; Richarz, A. N.; Schultz, T. W. Identification and Description of the Uncertainty, Variability, Bias and Influence in Quantitative Structure-Activity Relationships (QSARs) for Toxicity Prediction. *Regulatory Toxicology and Pharmacology* **2019**, *106* (April), 90–104. https://doi.org/10.1016/j.yrtph.2019.04.007.

(56) Benfenati, E.; Pardoe, S.; Martin, T.; Diaza, R. G.; Lombardo, A.; Manganaro, A.; Gissi, A. Using Toxicological Evidence from QSAR Models in Practice. *Altex* **2013**, *30* (1), 19–40. https://doi.org/10.14573/altex.2013.1.019.

(57) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.

Environmental Health Perspectives **2003**, *111* (10), 1361–1375. https://doi.org/10.1289/ehp.5758.

(58) Baumann, D.; Baumann, K. Reliable Estimation of Prediction Errors for QSAR Models under Model Uncertainty Using Double Cross-Validation. *Journal of Cheminformatics* **2014**, *6* (1), 1–19. https://doi.org/10.1186/s13321-014-0047-1.

(59) Sahlin, U. Uncertainty in QSAR Predictions. *ATLA Alternatives to Laboratory Animals* **2013**, *41* (1), 111–125. https://doi.org/10.1177/026119291304100111.

(60) Pradeep, P.; Paul Friedman, K.; Judson, R. Structure-Based QSAR Models to Predict Repeat Dose Toxicity Points of Departure. *Computational Toxicology* **2020**, *16* (June), 100139. https://doi.org/10.1016/j.comtox.2020.100139.

(61) Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Networks. *ArXiv* **2015**.

(62) Overweg, H.; Popkes, A.-L.; Ercole, A.; Li, Y.; Hernández-Lobato, J. M.; Zaykov, Y.; Zhang, C. Interpretable Outcome Prediction with Sparse Bayesian Neural Networks in Intensive Care. *ArXiv* **2019**.

(63) Kiureghian, A. Der; Ditlevsen, O. Aleatory or Epistemic ? Does It Matter ? *Structural Safety* **2009**, *31* (2), 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.

(64) Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision ? *ArXiv* **2017**, No. Nips.

(65) ChEMBL database http://www.ebi.ac.uk/chembl/.

(66) Bowes, J.; Brown, A. J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing Safety-Related Drug Attrition: The Use of in Vitro Pharmacological Profiling. *Nat Rev Drug Discov* **2012**, *11* (12), 909–922. https://doi.org/10.1038/nrd3845.

(67) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Using 2D Structural Alerts to Define Chemical Categories for Molecular Initiating Events. *Toxicological Sciences* **2018**, *165* (1), 213–223.

(68) Allen, T. E. H.; Liggi, S.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Using Molecular Initiating Events to Generate 2D Structure-Activity Relationships for Toxicity Screening. *Chemical Research in Toxicology* **2016**, *29* (10), 1611–1627. https://doi.org/10.1021/acs.chemrestox.6b00101.

(69) Landrum, G. RDKit. 2018.

(70) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, 742–754.

(71) Golbraikh, A.; Tropsha, A. Beware of q 2 ! *Journal of Molecular Graphics and Modelling* **2002**, *20*, 2553.

(72) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling* **2015**, *55* (7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206.

(73) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. *Bayesian Data Analysis, Third Edition*, Third Edit.; CRC Press, 2013.

(74) OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. **2007**, No. February.

(75) Wenzel, F.; Roth, K.; Veeling, B. S.; Świątkowski, J.; Tran, L.; Mandt, S.; Snoek, J.; Salimans, T.; Jenatton, R.; Nowozin, S. How Good Is the Bayes Posterior in Deep Neural Networks Really? *37th International Conference on Machine Learning, ICML 2020* **2020**, *PartF16814* (1), 10179–10190.

(76) Izmailov, P.; Vikram, S.; Hoffman, M. D.; Wilson, A. G. What Are Bayesian Neural Network Posteriors Really Like? **2021**.

(77) Vaicenavicius, J.; Widmann, D.; Roll, J.; Andersson, C.; Schön, T. B. Evaluating Model Calibration in Classification. *ArXiv* **2019**, *89*.

(78) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On Calibration of Modern Neural Networks. *ArXiv* **1996**.

(79) Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. *ArXiv* **2017**.

(80) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science (1979)* **1960**, *132* (3434), 1115–1118.

(81) Sushko, I. Applicability Domain of QSAR Models, Technischen Universität München, 2010.

(82) Allen, T. E. H.; Goodman, J. M.; Gutsell, S.; Russell, P. J. Quantitative Predictions for Molecular Initiating Events Using Three- Dimensional Quantitative Structure – Activity Relationships. *Chemical Research in Toxicology* **2019**, *33* (2), 324–332. https://doi.org/10.1021/acs.chemrestox.9b00136.

(83) Louisse, J.; Beekmann, K.; Rietjens, I. M. C. M. Use of Physiologically Based Kinetic Modeling-Based Reverse Dosimetry to Predict in Vivo Toxicity from in Vitro Data. *Chemical Research in Toxicology* **2017**, *30*, 114–125. https://doi.org/10.1021/acs.chemrestox.6b00302.

(84) Paini, A.; Leonard, J. A.; Joossens, E.; Bessems, J. G. M.; Desalegn, A.; Dorne, J. L.; Gosling, J. P.; Heringa, M. B.; Klaric, M.; Kliment, T.; Kramer, N. I.; Loizou, G.; Louisse, J.; Lumen, A.; Madden, J. C.; Patterson, E. A.; Proença, S.; Punt, A.; Setzer, R. W.; Suciu, N.; Troutman, J.; Yoon, M.; Worth, A.; Tan, Y. M. Next Generation Physiologically Based Kinetic (NG-PBK) Models in Support of Regulatory

Decision Making. *Computational Toxicology* **2019**, *9* (June 2018), 61–72. https://doi.org/10.1016/j.comtox.2018.11.002.

(85) Punt, A.; Bouwmeester, H.; Blaauboer, B. J.; Coecke, S.; Hakkert, B.; Hendriks, D. F. G.; Jennings, P.; Kramer, N. I.; Neuhoff, S.; Masereeuw, R.; Paini, A.; Peijnenburg, A. A. C. M.; Rooseboom, M.; Shuler, M. L.; Sorrell, I.; Spee, B.; Strikwold, M.; van der Meer, A. D.; van der Zander, M.; Vinken, M.; Yang, H.; Bos, P. M. J.; Heringa, M. B. New Approach Methodologies (NAMs) for Human-Relevant Biokinetics Predictions : Meeting the Paradigm Shift in Toxicology Towards an Animal-Free Chemical Risk Assessment. *ALTEX* **2020**, *37* (4), 607–622. https://doi.org/10.14573/altex.2003242.

(86) Zhang, H.; Ding, L.; Zou, Y.; Hu, S. Q.; Huang, H. G.; Kong, W. B.; Zhang, J. Predicting Drug-Induced Liver Injury in Human with Naïve Bayes Classifier Approach. *Journal of Computer-Aided Molecular Design* **2016**, *30* (10), 889–898. https://doi.org/10.1007/s10822-016-9972-6.

FIGURES

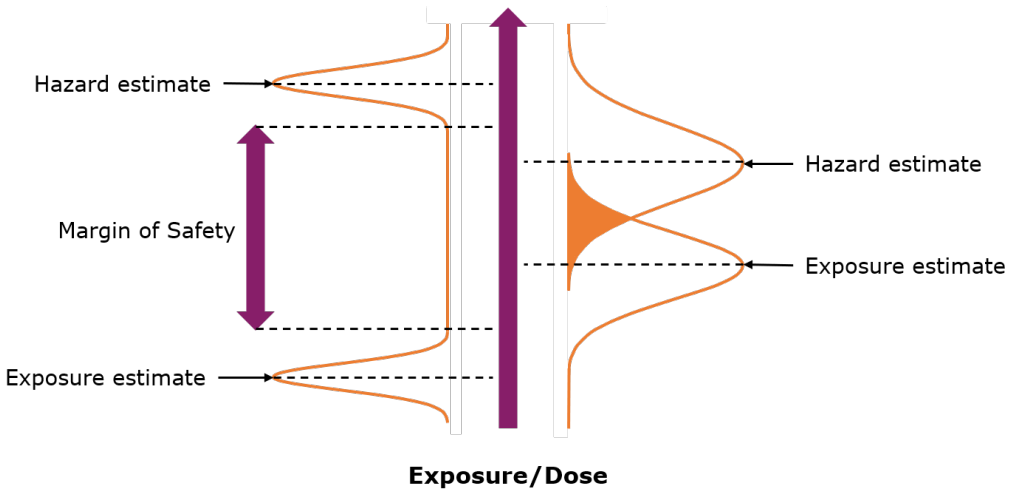


Figure 1. A graphical representation of scenarios where a safety evaluation would indicate a concern to health (right) and one with a large margin of safety (left).

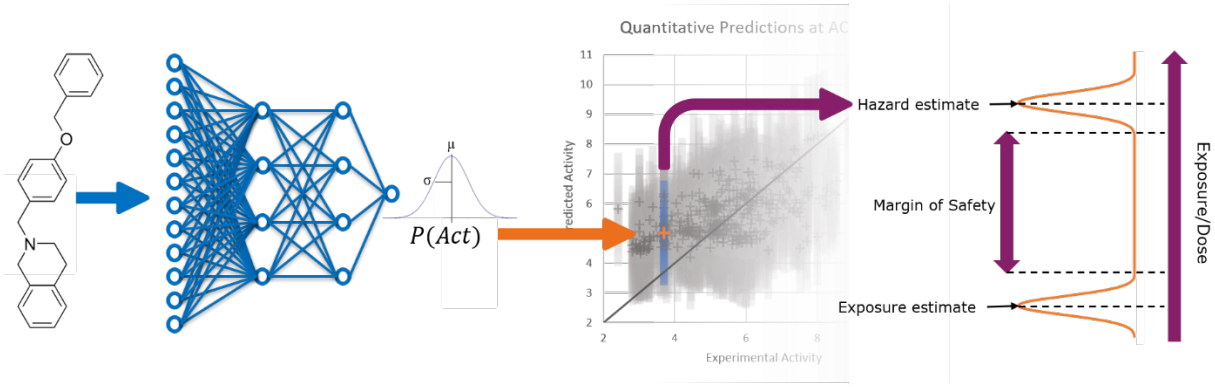


Figure 2. A workflow showing how Bayesian neural networks might be used in NGRA. Chemical structure is used by the neural network to make quantitative predictions of molecular activity (P(Act)) including uncertainties at biological targets which can feed into hazard estimates to be compared to exposure estimate for the identification of a margin of safety.

Figure 3. Histograms showing the data distribution for each data set.

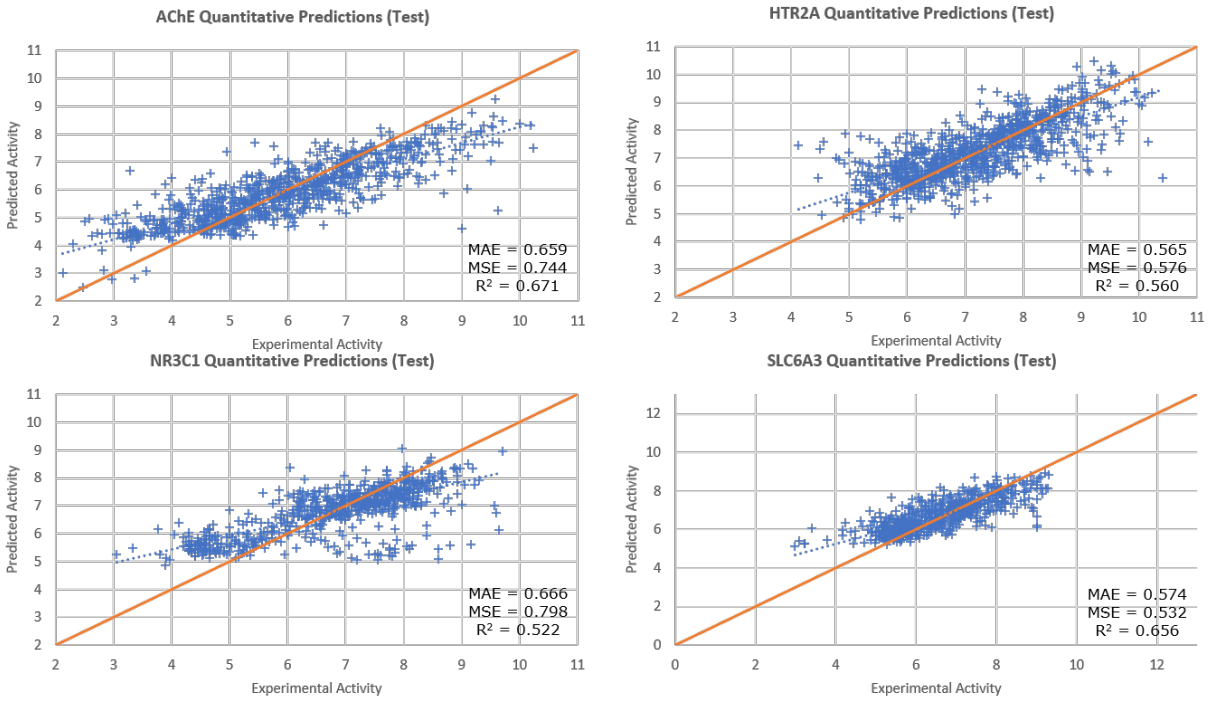


Figure 4. Graphs showing predicted vs experimental activity values for the test set chemicals. The ideal x = y diagonal is shown as an orange solid line and the trendline for the shown data as a blue dashed line.

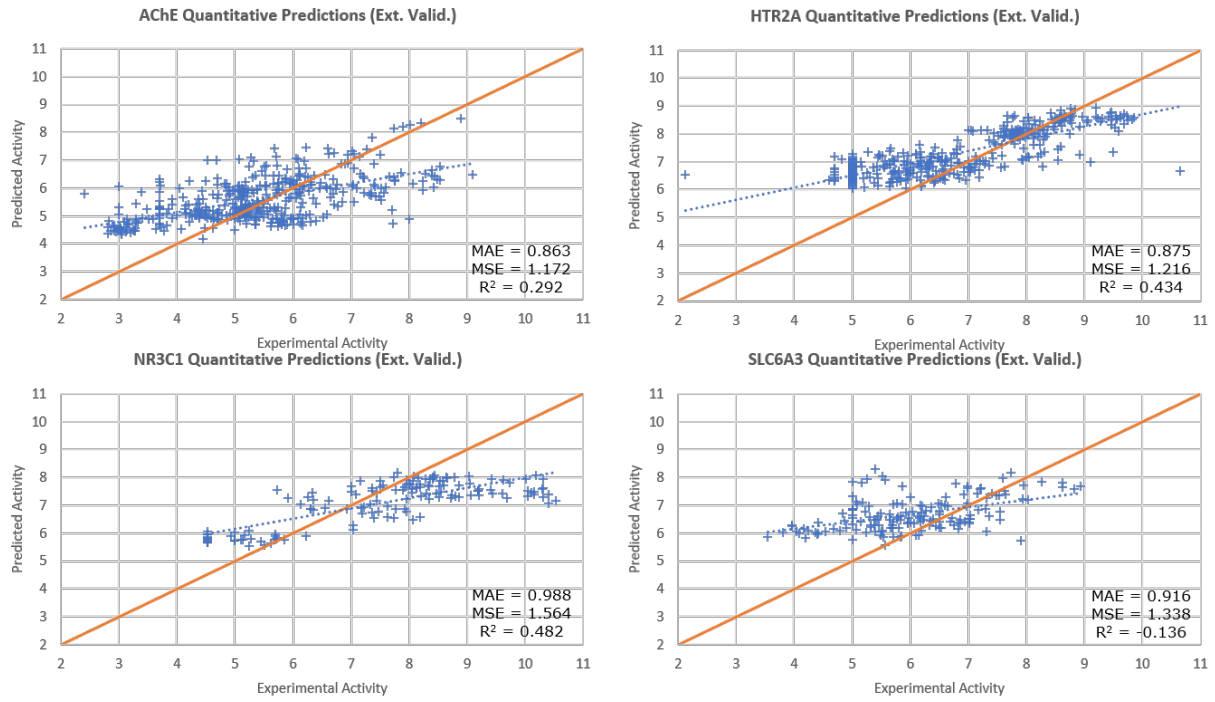


Figure 5. Graphs showing predicted vs experimental activity values for the external validation set chemicals. The ideal x = y diagonal is shown as an orange solid line and the trendline for the shown data as a blue dashed line.

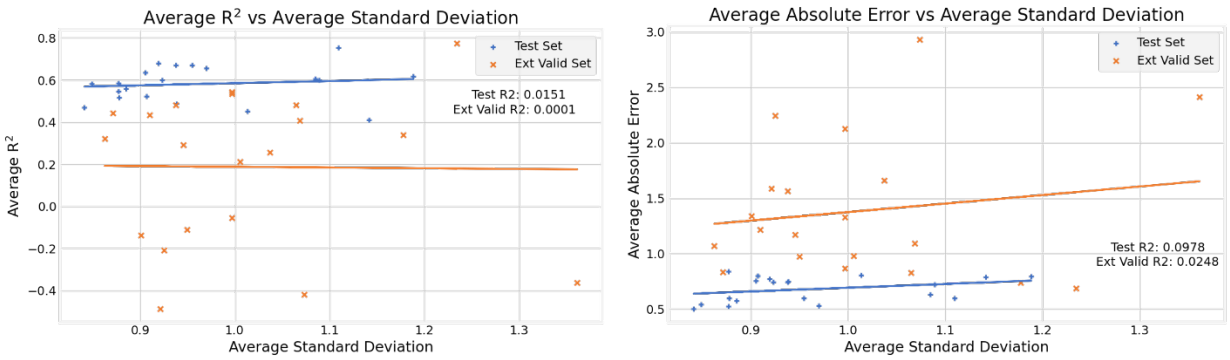


Figure 6. Graphs showing average standard deviation values against MAE or R^2 values for each test and external validation set for each MIE, including trendlines.

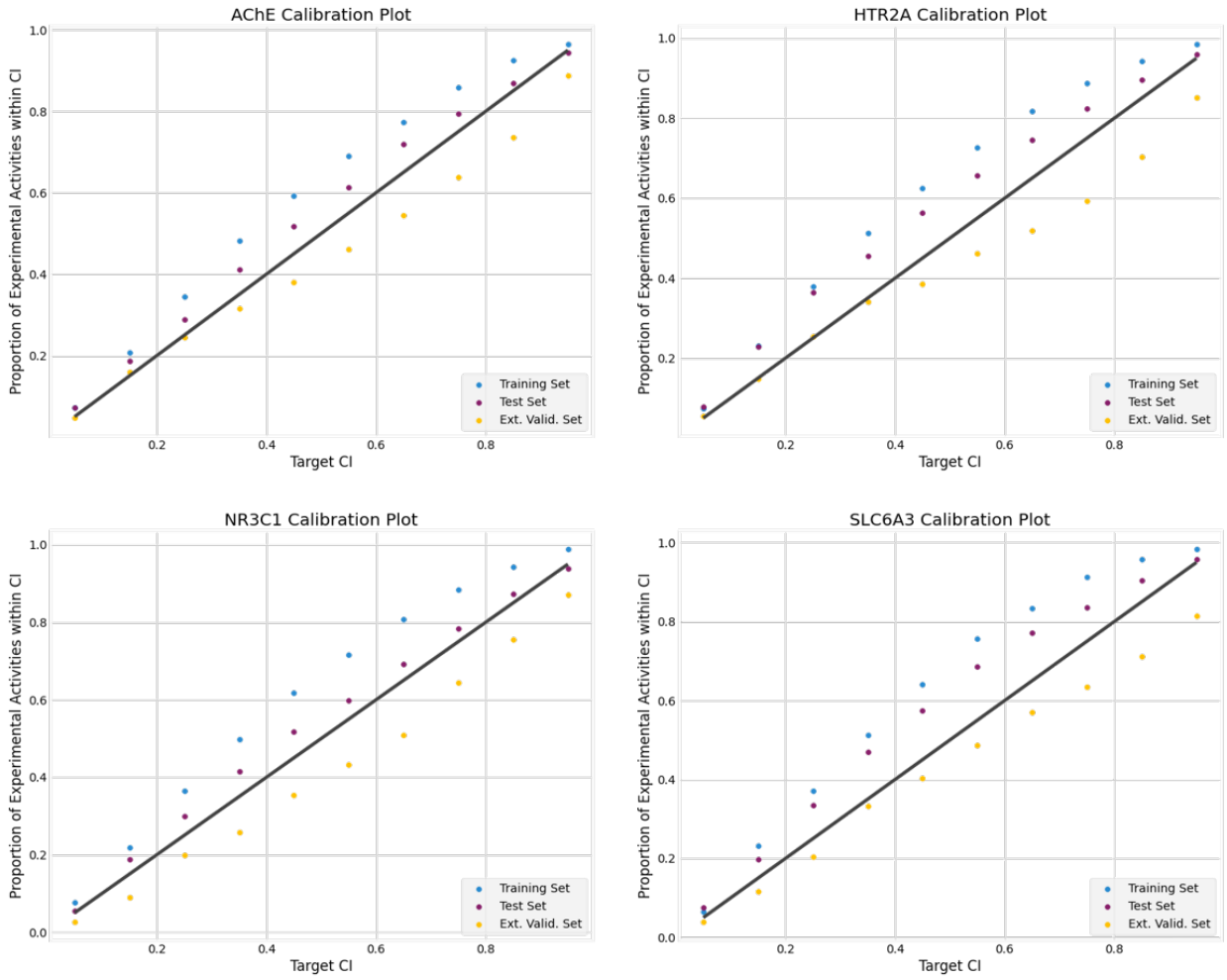


Figure 7. Calibration plots showing the proportion of experimental activities within credible intervals (CI) between 5% and 95% at 10% intervals for the training, test and external validation datasets. The ideal x = y diagonal is shown as a black solid line.

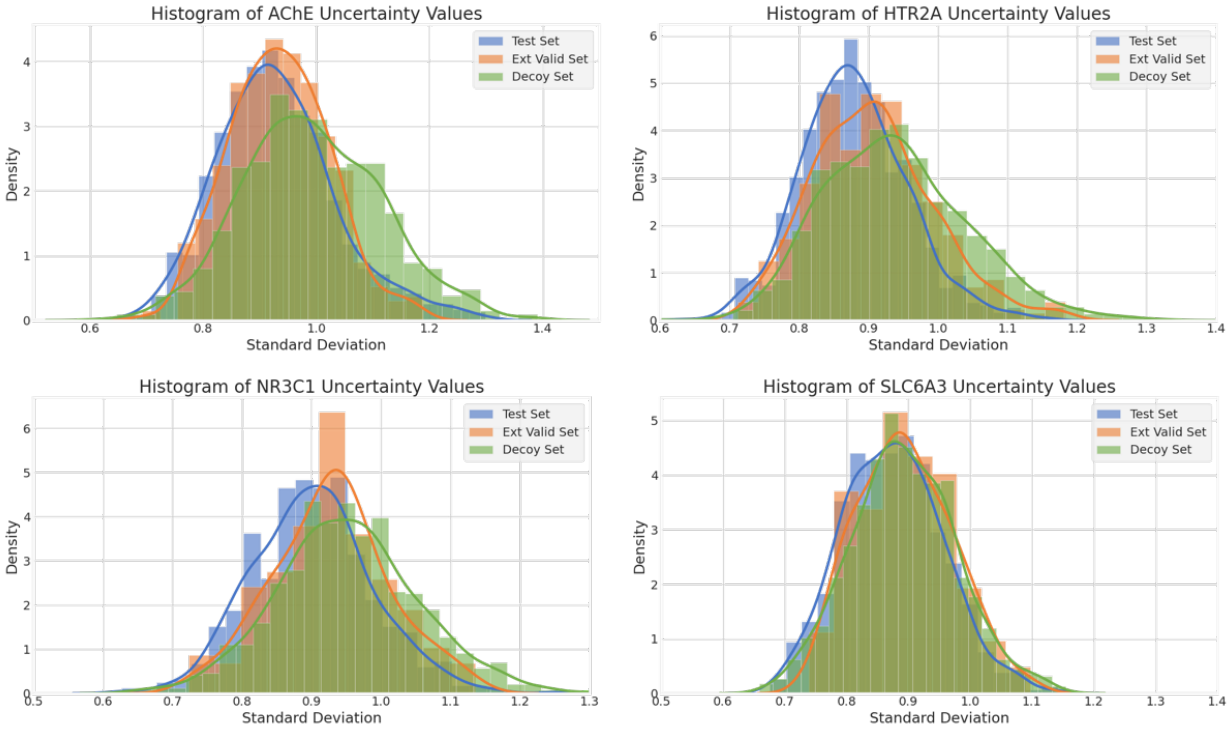


Figure 8. Histograms showing differences in standard deviation distributions between test, external validation, and decoy data sets overlaid with kernel density estimates.

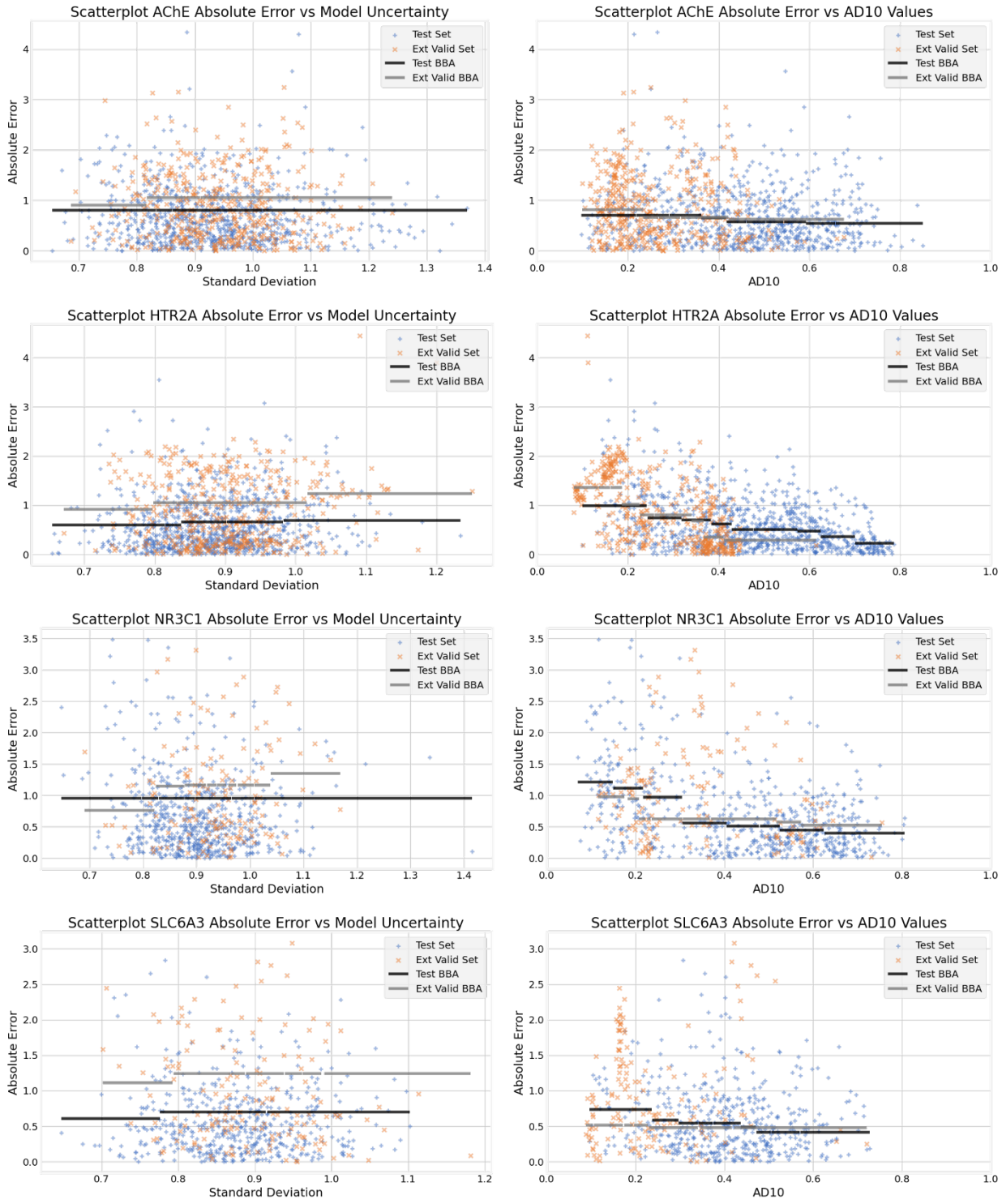
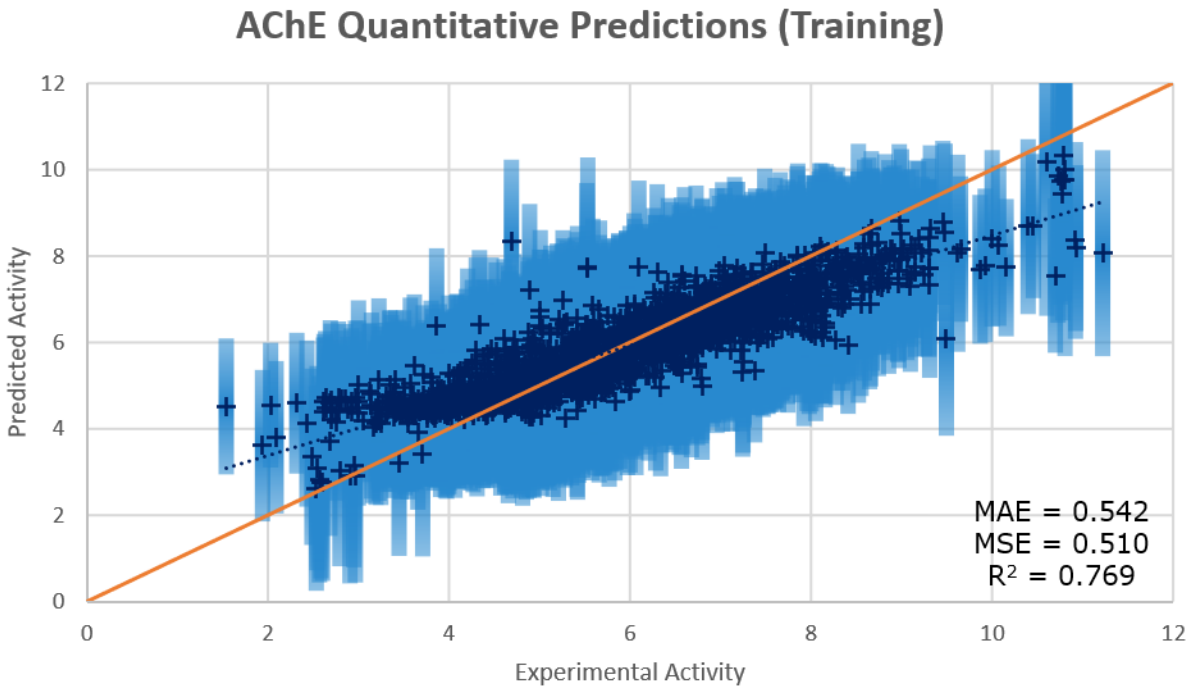
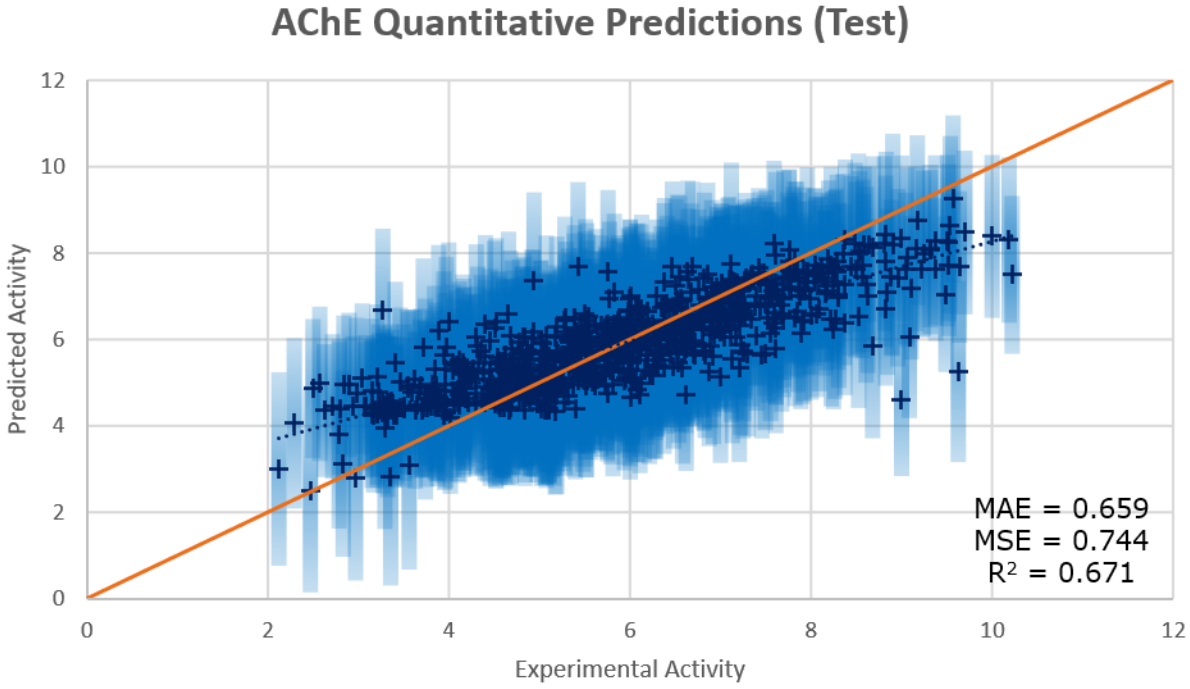


Figure 9. Scatterplots showing the relationship between prediction absolute error and model standard deviation or applicability domain 10 (AD10) values for the test and external validation set predictions. Bin-based averages (BBAs) are shown for 10 equal folds of the dataset.


AChE Quantitative Predictions (Training)
Predicted Activity
Experimental Activity
0
2
4
6
8
10
12
MAE = 0.542
MSE = 0.510
R² = 0.769


AChE Quantitative Predictions (Test)
Predicted Activity
Experimental Activity
0
2
4
6
8
10
12
MAE = 0.659
MSE = 0.744
R² = 0.671

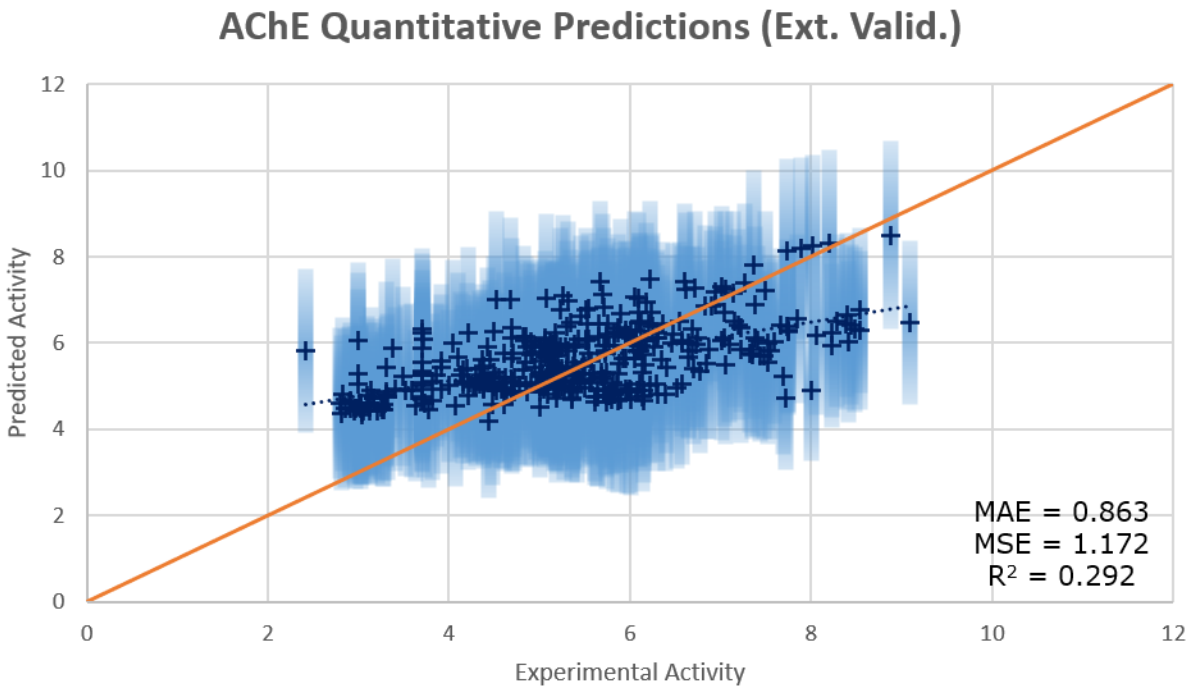


Figure 10. Graphs showing predicted vs experimental activity values and 95% credible intervals for the acetylcholinesterase training, test and external validation sets. The ideal x = y diagonal is shown as an orange solid line and the trendline for the shown data as a blue dashed line.

TABLES

Target Gene	Target Name	Training Set	Validation Set	Test Set	Ext. Valid. Set
AChE	**Acetylcholinesterase**	2024	675	842	423
ADORA2A	**Adenosine A2a receptor**	2271	758	1017	270
ADRB1	**Beta-1 adrenergic receptor**	747	250	309	26
ADRB2	**Beta-2 adrenergic receptor**	1641	548	749	78
AR	**Androgen receptor**	2754	919	1208	163
CHRM1	**Muscarinic acetylcholinesterase receptor M1**	1201	401	553	96
CHRM2	**Muscarinic acetylcholinesterase receptor M2**	946	316	423	69
CHRM3	**Muscarinic acetylcholinesterase receptor M3**	926	309	358	129
DD1R	**Dopamine D1 receptor**	801	268	326	109
DD2R	**Dopamine D2 receptor**	3255	1086	1458	514
EDNRA	**Endothelin receptor ET-A**	754	252	336	16
HRH1	**Histamine H1 receptor**	732	245	327	87
HTR2A	**Serotonin 2a receptor**	2097	700	964	433
KCNH2	**Human ether-a-go-go related gene channel**	3983	1328	1871	1032
LCK	**Tyrosine-protein kinase LCK**	1091	364	497	214
NR3C1	**Glucocorticoid receptor**	1492	498	662	147
OPRD1	**Delta opioid receptor**	1784	595	812	612
OPRM1	**Mu opioid receptor**	2182	728	1014	915
SLC6A2	**Norepinephrine transporter**	1652	551	681	121
SLC6A3	**Dopamine transporter**	1376	459	614	156
SLC6A4	**Serotonin receptor**	2356	786	1011	227

Table 1. Numbers of compounds extracted for each biological target within each set.

		Training Set			Validation Set			Test Set			Ext. Validation Set		
Target	**Arch**	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
AChE	[100]	0.542	0.510	0.769	0.682	0.814	0.660	0.659	0.744	0.671	0.863	1.172	0.292
ADORA2A	[50]	0.495	0.406	0.712	0.605	0.643	0.555	0.580	0.540	0.582	0.819	1.070	0.321
ADRB1	[10]	0.432	0.335	0.743	0.650	0.729	0.479	0.641	0.751	0.487	0.904	0.972	-0.111
ADRB2	[100]	0.526	0.493	0.795	0.596	0.668	0.680	0.660	0.773	0.679	0.845	1.329	0.534
AR	[100]	0.477	0.432	0.743	0.630	0.862	0.492	0.628	0.837	0.546	1.077	1.589	-0.487
CHRM1	[100]	0.537	0.507	0.682	0.625	0.700	0.569	0.693	0.806	0.452	0.751	0.830	0.482
CHRM2	[100]	0.486	0.367	0.791	0.654	0.665	0.632	0.665	0.719	0.601	0.871	1.094	0.408
CHRM3	[100]	0.461	0.353	0.864	0.556	0.727	0.719	0.616	0.597	0.755	0.683	0.689	0.776
DD1R	[100]	0.494	0.410	0.708	0.657	0.741	0.421	0.663	0.788	0.412	0.715	0.739	0.341
DD2R	[50]	0.475	0.388	0.599	0.560	0.549	0.480	0.533	0.502	0.471	0.718	0.836	0.443
EDNRA	[100]	0.472	0.374	0.816	0.663	0.733	0.654	0.678	0.792	0.619	1.267	2.415	-0.362
HRH1	[50]	0.421	0.313	0.808	0.546	0.503	0.691	0.569	0.630	0.607	1.466	2.932	-0.417
HTR2A	[100]	0.472	0.377	0.708	0.554	0.538	0.548	0.565	0.576	0.560	0.875	1.216	0.434
LCK	[50,50]	0.554	0.507	0.756	0.726	0.844	0.633	0.678	0.745	0.601	1.319	2.130	-0.053
NR3C1	[50]	0.495	0.413	0.746	0.680	0.830	0.503	0.666	0.798	0.522	0.988	1.564	0.482
OPRD1	[100]	0.474	0.359	0.799	0.613	0.595	0.666	0.608	0.599	0.672	0.759	0.865	0.545
OPRM1	[50]	0.479	0.396	0.815	0.612	0.647	0.687	0.665	0.753	0.636	1.187	2.248	-0.207
SLC6A2	[50,50]	0.504	0.410	0.682	0.605	0.599	0.542	0.610	0.595	0.516	0.771	0.978	0.212
SLC6A3	[10]	0.457	0.364	0.724	0.562	0.549	0.581	0.550	0.524	0.585	0.916	1.338	-0.136
SLC6A4	[100]	0.458	0.348	0.780	0.571	0.559	0.642	0.574	0.532	0.656	1.032	1.663	0.258

Table 2. Statistical performance of the best models against the training, validation, test and external validation data sets. The best model architecture for each target is also shown. AChE=acetylcholinesterase, ADORA2A=adenosine A2a receptor, ADRB1=beta-1 adrenergic receptor, ADRB2=beta-2 adrenergic receptor, AR=androgen receptor, CHRM1=muscarinic acetylcholinesterase receptor M1, CHRM2=muscarinic acetylcholinesterase receptor M2, CHRM3=muscarinic acetylcholinesterase receptor M3, DD1R=dopamine

D1 receptor, DD2R=dopamine D2 receptor, EDNRA=endothelin receptor ET-A, HRH1=histamine H1 receptor, HTR2A=serotonin 2a receptor, LCK=tyrosine-protein kinase LCK, MAE=mean absolute error, MSE=mean squared error, NR3C1=glucocorticoid receptor, OPRD1=delta opioid receptor, OPRM1=mu opioid receptor, SLC6A2=norepinephrine transporter, SLC6A=dopamine transporter, SLC6A4=serotonin receptor.

Target		Average SD			% within 95% CI	
	Arch	Test Set	Ext. Val. Set	Random Set	Test Set	Ext. Val. Set
AChE	[100]	0.9376	0.9457	0.9933	95.96	93.16
ADORA2A	[50]	0.8491	0.8625	0.8849	97.54	90.51
ADRB1	[10]	0.9385	0.9498	0.9538	94.82	96.15
ADRB2	[100]	0.9194	0.9966	0.9597	94.53	90.00
AR	[100]	0.8771	0.9214	0.8959	95.03	83.44
CHRM1	[100]	1.0134	1.0647	1.0910	96.02	98.96
CHRM2	[100]	1.0888	1.0683	1.1573	98.58	100.00
CHRM3	[100]	1.1096	1.2344	1.2161	99.72	100.00
DD1R	[100]	1.1412	1.1774	1.1882	96.63	100.00
DD2R	[50]	0.8413	0.8714	0.8868	97.19	93.39
EDNRA	[100]	1.1884	1.3614	1.4345	97.92	93.75
HRH1	[50]	1.0848	1.0734	1.1208	96.33	73.56
HTR2A	[100]	0.8854	0.9099	1.1208	97.10	89.38
LCK	[50,50]	0.9229	0.9966	1.0429	94.77	78.60
NR3C1	[50]	0.9069	0.9376	0.9527	94.11	87.76
OPRD1	[100]	0.9546	0.9967	1.0399	97.91	95.92
OPRM1	[50]	0.9053	0.9253	0.9148	95.86	81.31
SLC6A2	[50,50]	0.8778	1.0057	0.9976	96.77	89.34
SLC6A3	[10]	0.8773	0.9007	0.8927	96.74	85.90
SLC6A4	[100]	0.9697	1.0371	1.0824	99.11	88.65

Table 3. Average standard deviation values produced by each model against the test and external validation sets, and a set of synthetic randomized input strings alongside the percentage of the test and external validation set activity predictions falling within the 95% credible interval (CI) at each biological target.

		AD10 > 0.2				StDev < 1.2				AD10 > 0.2 OR StDev < 1.2			
		Test Set		Ext. Validation Set		Test Set		Ext. Validation Set		Test Set		Ext. Validation Set	
Target	**Arch**	**MAE**	**attrition**	**MAE**	**attrition**	**MAE**	**attrition**	**MAE**	**attrition**	**MAE**	**attrition**	**MAE**	**attrition**
AChE	[100]	0.663	7.4	0.812	39.2	0.669	3.4	0.870	0.5	0.667	10.7	0.813	39.4
ADORA2A	[50]	0.557	3.6	0.778	18.2	0.580	0.0	0.819	0.0	0.557	3.6	0.778	18.2
ADRB1	[10]	0.624	2.7	1.130	53.8	0.628	0.3	0.904	0.0	0.625	3.1	1.130	53.8
ADRB2	[100]	0.692	34.1	0.919	33.8	0.641	0.6	0.852	16.3	0.690	34.7	0.887	50.0
AR	[100]	0.607	26.4	0.973	36.8	0.622	0.8	1.068	1.2	0.604	26.9	0.978	38.0
CHRM1	[100]	0.665	7.2	0.663	74.5	0.678	5.7	0.792	15.3	0.670	11.9	0.739	79.6
CHRM2	[100]	0.656	7.6	0.865	77.1	0.690	18.1	0.859	17.1	0.668	24.6	0.897	80.0
CHRM3	[100]	0.608	8.3	0.641	49.2	0.630	26.6	0.752	45.4	0.624	32.1	0.713	76.9
DD1R	[100]	0.597	26.0	0.723	19.3	0.708	29.0	0.679	40.4	0.622	48.0	0.606	53.2
DD2R	[50]	0.523	2.2	0.697	10.2	0.535	0.1	0.718	0.0	0.522	2.2	0.697	10.2
EDNRA	[100]	0.672	2.8	1.267	0.0	0.667	41.9	1.895	87.5	0.665	42.2	1.895	87.5
HRH1	[50]	0.498	11.7	1.077	59.8	0.580	13.7	1.462	8.0	0.511	24.8	1.103	60.9
HTR2A	[100]	0.554	4.1	0.593	35.6	0.578	0.1	0.871	0.2	0.555	4.2	0.593	35.6
LCK	[50,50]	0.670	15.2	1.325	73.5	0.676	3.2	1.331	7.0	0.665	16.4	1.332	74.9
NR3C1	[50]	0.572	17.2	0.977	18.2	0.670	0.5	0.976	0.0	0.572	17.2	0.977	18.2
OPRD1	[100]	0.601	5.2	0.844	28.8	0.615	1.2	0.760	1.8	0.599	6.0	0.840	30.0
OPRM1	[50]	0.654	4.2	1.197	12.3	0.670	0.0	1.187	0.0	0.654	4.2	1.197	12.3
SLC6A2	[50,50]	0.603	4.8	0.597	39.7	0.604	6.4	0.779	5.0	0.598	10.5	0.579	42.1
SLC6A3	[10]	0.552	6.1	0.789	51.9	0.563	0.0	0.916	0.0	0.552	6.1	0.789	51.9
SLC6A4	[100]	0.570	3.0	0.871	26.4	0.582	3.5	0.958	14.1	0.574	6.0	0.875	29.1

Table 4. Statistical performance of the best models against the test and external validation data sets, considering three applicability domain thresholds. Attrition rates show the percentage of compounds in case that have been excluded as outside the threshold. The best model architecture for each target is also shown.