Fortnightly meeting 5th November 2019

Mark Nelms

5/11/2019

Contents

Questions from last time 1

1 Acute toxicity **1**

1.1 Profiling additional chemicals 1

1.2 Converting experimental and TIMES LD50s from mg/kg to -log10(mol/kg) 2

1.3 How do the TEST/TIMES predictions compare to the experimental values? 2

1.4 Are there areas of chemical space where one prediction software works better? 10

Questions from last time

After our last meeting we had some questions to answer/actions to perform:

- Run the additional acute tox chemicals we now have QSAR-ready SMILES for through TEST and TIMES
- Convert the experimental and TIMES LD50s from mg/kg to -log10(mol/kg)
- How well are the TEST/TIMES predictions working? (Excluding those in the respective training sets)
  - What do the fits between the predicted and experimental LD50s look like?
  - What is the root mean square error (RMSE), r^2, and mean absolute error (MAE)
- Are there areas of chemical space where one prediction software works better?

1 Acute toxicity

1.1 Profiling additional chemicals

Before doing anything else I converted the QSAR-ready SMILES for the 1,541 additional chemicals into an SDF and then generated LD50 predictions for these chemicals using both TEST (v4.2.1) and TIMES (v2.28.1).

Below is a table (Table 1) that shows the number of chemicals that:

1) could be run through each piece of software, and
2) with an LD50 prediction (that wasn't in the training set)

Table 1: Counts of chemicals with QSAR-ready SMILES run through TEST and TIMES

		Number of chemicals	
Software		**Orginal SMILES**	**Updated SMILES**
TEST	Total chemicals	9332	10760
	With LD50 prediction (not in training set)	3194	3927
TIMES	Total chemicals	8893	10419
	With LD50 prediction (not in training set)	738	862

1.2 Converting experimental and TIMES LD50s from mg/kg to -log10(mol/kg)

As we talked about in our last meeting, we decided to convert the experimental and TIMES LD50 values from mass (i.e. mg/kg) to molar (i.e. -log10(mol/kg)) units. Using the molar units allows us to compare the toxicity of chemicals whilst taking the mass of the chemical into consideration, i.e. we are comparing the number of "functional units" of each chemical rather than their respective weights.

For example, if 2 chemicals have an LD50 of 10mg/kg but chemical A has a MW of 100 and chemical B has a MW of 50, there will be twice as many moles (i.e. "functional units") of chemical B present in the same mass compared to chemical A. Therefore, if we use molar units, chemical A will be shown to be more toxic than chemical B.

To do the conversion from mg/kg to -log10(mol/kg) I used the final part of this equation:

$$\begin{aligned} OralLD_{50}(\text{-log}_{10}[mol/kg]) &= \text{-log}_{10}\left(\frac{Concentration(g/kg)}{MM(g/mol)}\right) \\ &= \text{-log}_{10}\left(\frac{(Concentration(mg/kg)\,/\,1000)}{MM(g/mol)}\right) \\ &= \text{-log}_{10}\left(Concentration(mg/kg)\,/\,MM(g/mol)\,/\,1000\right) \end{aligned} \tag{1}$$

$$OralLD_{50}(-log_{10}[mol/kg]) = \text{-log}_{10}\left(Concentration(mg/kg)\,/\,MM(g/mol)\,/\,1000\right) \tag{2}$$

where Concentration = Oral LD_{50} of chemical (in mg/kg) and MM = molar mass of chemical (in g/mol).

The top equation above is expecting the concentration in g/kg; however, our values are in mg/kg, therefore we have to divide the concentration by 1000 to get the molarity in mol/kg[1].

1.3 How do the TEST/TIMES predictions compare to the experimental values?

1.3.1 What do the fits between the predicted and experimental LD50s look like?

Before I started visualising the fits, I calculated the residuals for each chemical by subtracting the predicted LD50 value from the experimental LD50 value (in -log10(mol/kg) space), i.e. $Residual_i = ExpLD50_i - PredLD50_i$. I did this separately for the TEST predictions and TIMES predictions.

Once I had both the LD50s (in -log10[mol/kg]) and the residuals for each piece of software, I generated scatter plots that compared the predicted values against the experimental values.

All chemicals with LD50 predictions in TEST or TIMES

To begin with I removed chemicals that:

- Did not have a processed LD50 from the Acute Toxicity Working Group (ATWG), and
- Were part of the training set for that profiler

This reduced the number of chemicals down to:

- 1,619 chemicals with a predicted LD50 from TEST, and
- 503 chemicals with a predicted LD50 from TIMES

Then, to get a sense of how well each piece of software performed overall, I first generated a scatter plot for each software program looking at all of the chemicals with a predicted LD50 against the corresponding experimental LD50. Additionally, I used the residuals to calculate the number of chemicals with errors: 1) > 0.5 log units, 2) within ±0.5 log units, or 3) < -0.5 log units.

I arbitrarily chose 0.5 log units as a starting point based upon Katie Paul Friedman's paper that's currently in prep that we cited in our TTC work, which, admittedly, is in log(mg/kg-day). Once I have calculated the bootstrapped standard deviation in -log10(mol/kg) we can update this value and get a better feel of how good/bad the TEST and TIMES predictions actually are compared to the variability in the *in vivo* data.

[1] I think, technically speaking, we are calculating molality because our units are in mol/kg rather than mol/L (units for molarity); but, as 1L == 1kg for our purposes molarity and molality should be interchangeable.

Figure 1, shows all 1,619 chemicals with a predicted LD50 from TEST and an experimental LD50 from the ATWG. Here, and throughout, the solid line in the scatter plots is the zero variance line and the dashed lines denote ±0.5 log units. Table 2, provides a breakdown of the number and percentage of chemicals that fall within one of the three categories mentioned above.

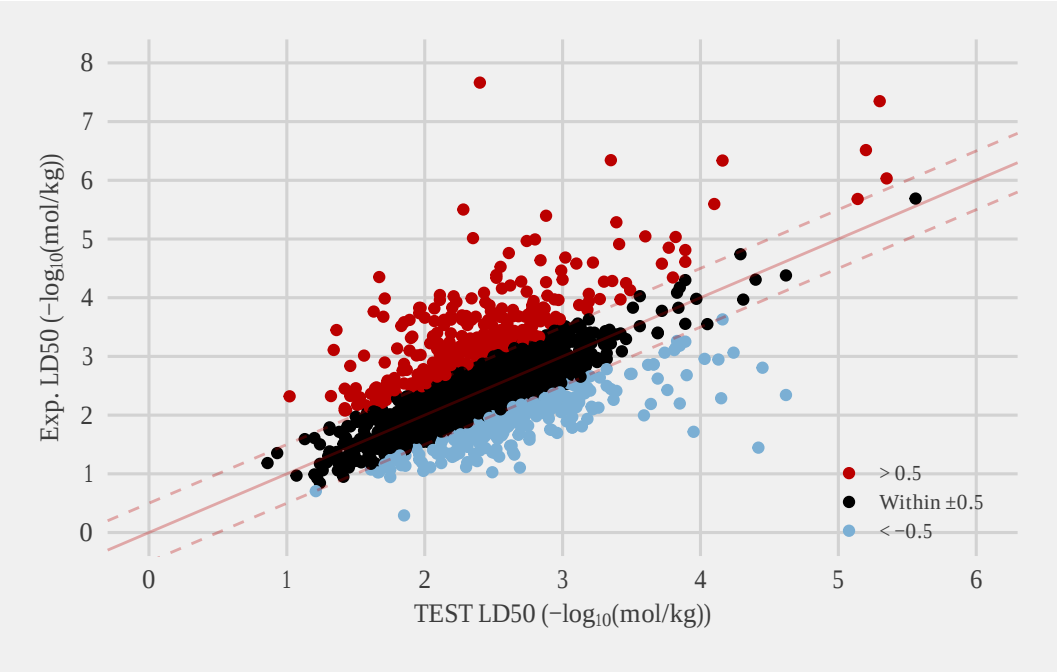

Figure 1: Comparing TEST predicted and experimental values -log10(mol/kg)

Figure 2, shows all 503 chemicals with a predicted LD50 from TIMES and an experimental LD50 from the ATWG. Again, Table 2 provides a breakdown of the number and percentage of chemicals that fall within one of the three categories.

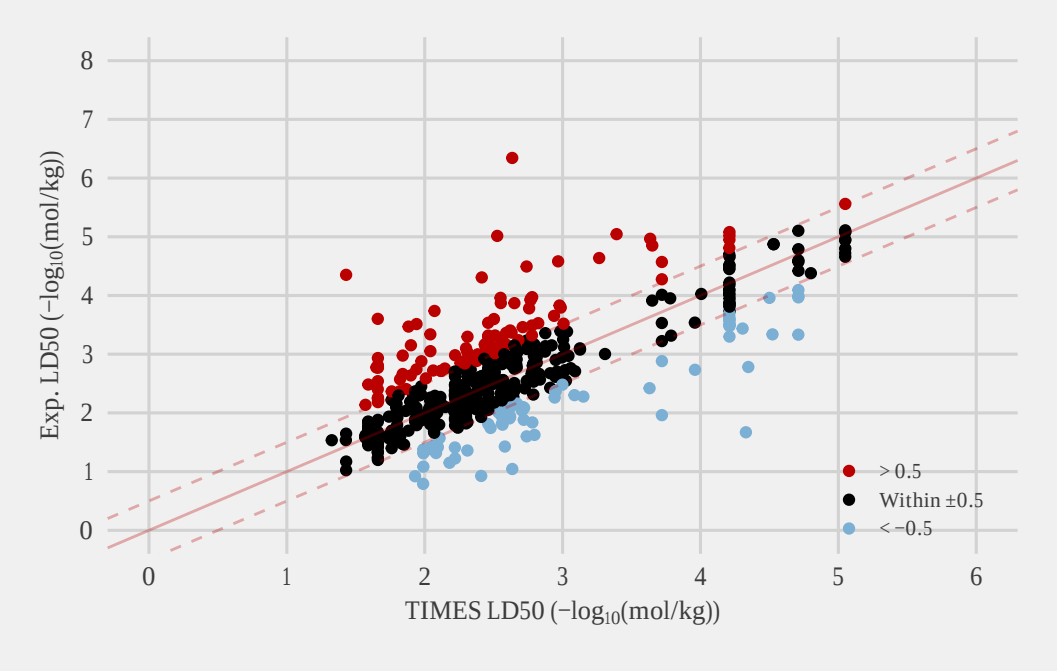

Figure 2: Comparing TIMES and experimental values -log10(mol/kg)

As you can see from Figures 1 and 2 and Table 2, the majority of chemicals have a prediction within ±0.5 log unit of the experimental

value, which is a good starting point. However, the next largest category for both TEST and TIMES is where the predicted LD50 is lower than the experimental value, with some having quite large difference (i.e. > 2 log units).

Table 2: Counts of chemicals with predictions in TEST or TIMES

Software	Above 0.5 log units		Within ±0.5 log units		Below -0.5 log units	
	Number of chemicals	%	Number of chemicals	%	Number of chemicals	%
TEST	321	19.83	1046	64.61	252	15.57
TIMES	98	19.48	339	67.40	66	13.12

As a sanity check that assuming the difference between the predicted and experimental values should be linear, which they theorectically should be, I plotted the residuals against the predicted values for both TEST and TIMES (Figure 3). The fact the residuals for both profilers form a random pattern of points is an indicator that our assumptions were correct.

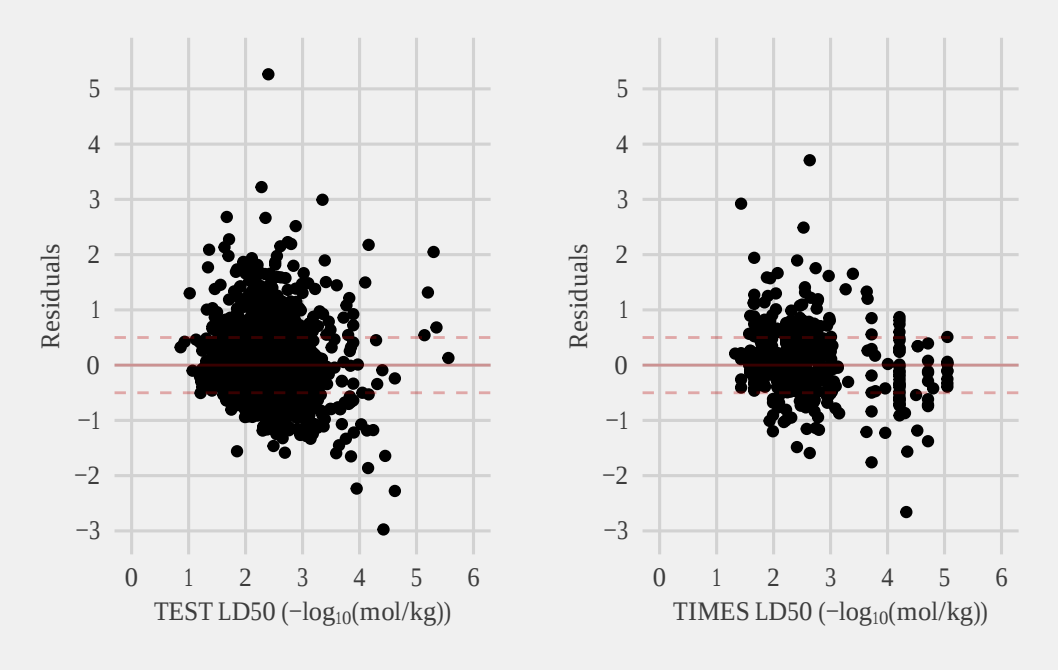

Figure 3: Residual plots for TEST and TIMES predictions

Figure 4 shows that the density plot of the residuals for both profilers have a large peak around 0 with tails that are longer than would be expected for a normal distribution. Both distributions are also slightly right-skewed (i.e. towards underestimating the experimental LD50).

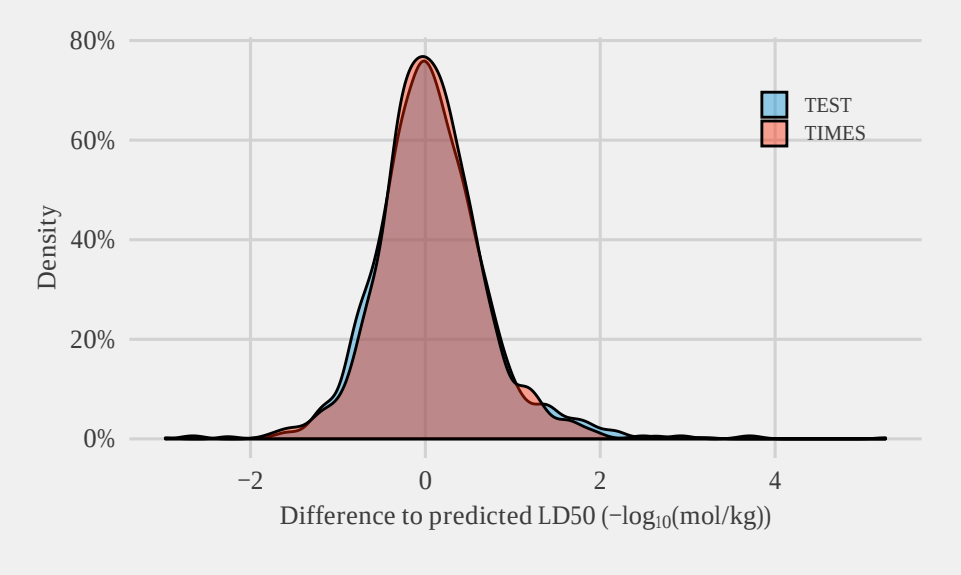

Figure 4: Density plot of residuals for chemicals with TEST or TIMES predictions

To further assess the performance of the TEST and TIMES predictions, I calculated 3 performance metrics for each software program: 1) the root mean square error (RMSE), which measures the overall accuracy of the model; 2) the traditional r^2, which measures the proportion of variation explained by the model; and 3) the mean absolute error (MAE), which measures the average magnitude of the errors.

The reason for calculating both the RMSE and MAE is because, the MAE is a simple measure that can tell us a lot about the average error between the predicted and experimental LD50s; however, it has the potential for understating large, but infrequent, errors. RMSE, meanwhile, allows us to identify if any large, infrequent, errors are present by giving them more weight - this is because all the errors are squared before calculating the mean and then square rooting it.

Table 3 shows that the RMSE for all the TEST predictions is 0.642 -log10(mol/kg) and the MAE is 0.469 -log10(mol/kg). Both the RMSE (0.62 -log10[mol/kg]) and the MAE (0.447 -log10[mol/kg]) for all the TIMES predictions are comparable to those for TEST. However, TIMES has a larger r^2 value (0.54) than TEST (0.296), meaning that the TIMES predictions fit the experimental data better than the TEST predictions.

Table 3: Performance metrics for all chemicals with TEST or TIMES prediction

Software	RMSE	R squared	MAE
TEST	0.642	0.296	0.469
TIMES	0.620	0.540	0.447

One reason for the difference between the RMSE and MAE values for both the TEST and TIMES predictions is likely due to the few chemicals with large discrepancies between their predicted and actual LD50s, which we can see in Figures 1 and 2.

The figure below shows how well correlated the TEST/TIMES predictions are to the experimental values and to each other (Figure 5). I used the `corrr::correlate` function to calculate the Pearson correlation coefficient for all observations that had pairwise complete observations (i.e. if both values weren't present that row was ignored).

As you can see, the TIMES predictions are more highly correlated to the experimental values than the TEST predictions. However, the TEST and TIMES predictions are also moderately correlated with one another, which is not wholly unsurprising.

Figure 5: Correlation of pairwise complete observations of TEST/TIMES predictions and experimental values

Chemicals with LD50 predictions in both TEST and TIMES

After looking at the performance of the two profilers across all chemicals for which they generated an LD50 prediction, I filtered the chemicals down to only those that had an LD50 prediction in both TEST and TIMES. There were a total of **274** chemicals that met the following criteria:

- Had a processed LD50 from the Acute Toxicity Working Group (ATWG)
- Were not part of the TEST or TIMES training sets, and
- Had an LD50 prediction from both TEST and TIMES

Figure 6, displays the predicted LD50 from TEST against the associated experimental LD50 from the ATWG for each of the 274 overlapping chemicals. As in the previous section, Table 4, provides a breakdown of the number and percentage of the overlapping chemicals that have a residual that falls within one of the three categories (i.e. above 0.5 log units, within ±0.5 log units, and below -0.5 log units).

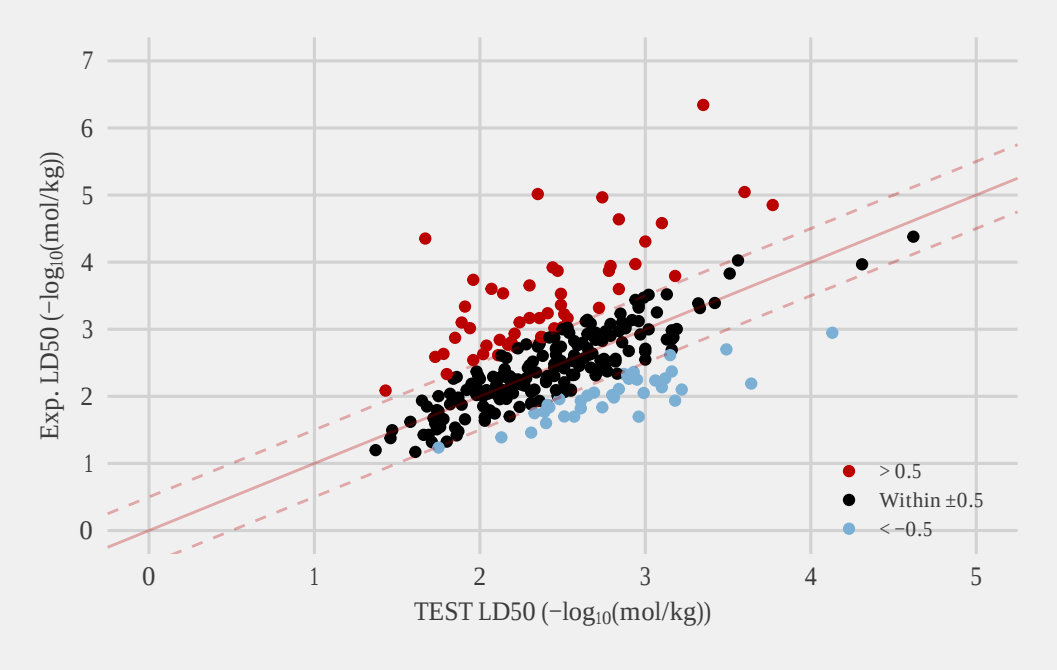

Figure 6: Comparing TEST and experimental values -log10(mol/kg) - filtered for overlapping chemicals

Figure 7, displays the predicted LD50 from TIMES against the associated experimental LD50 from the ATWG for each of the overlapping chemicals. Table 4, provides a breakdown of the number and percentage of the overlapping chemicals that have a residual that falls within one of the three categories.

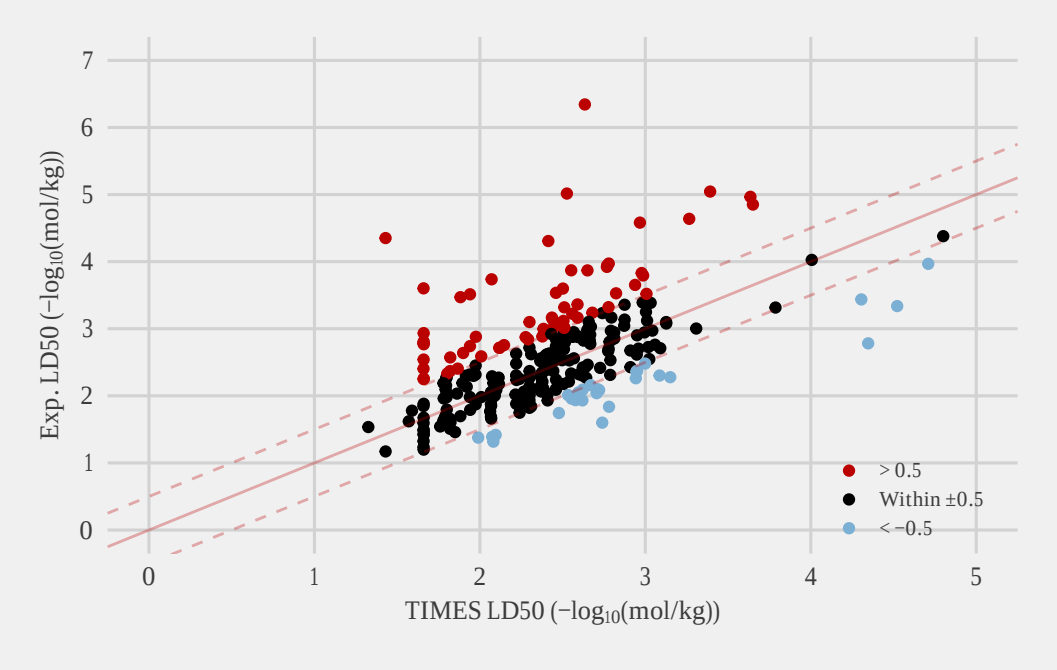

Figure 7: Comparing TIMES and experimental values -log10(mol/kg) - filtered for overlapping chemicals

The percentage of chemicals that fall into the "Within ±0.5 log units" category are approximately the same between the 2 profilers (68.61% and 67.52% for TEST and TIMES, respectively). There is also a slight improvement in the percentage of chemicals with TEST predictions for the overlapping set (68.61%) compared to all of the chemicals with a TEST prediction (64.61%).

The TIMES predictions look to have a slight increase in the percentage of chemicals that underestimate the experimental values (i.e. the "Above 0.5 log units" category), with a corresponding decrease in the percentage of chemicals present in the "Below -0.5 log units" category (i.e. overestimation of experimental values).

Table 4: Counts of chemicals with predictions in TEST and TIMES

Software	Chemical Set	Above 0.5 log units		Within ±0.5 log units		Below -0.5 log units	
		Number of chemicals	%	Number of chemicals	%	Number of chemicals	%
TEST	Full	321	19.83	1046	64.61	252	15.57
	Overlap	50	18.25	188	68.61	36	13.14
TIMES	Full	98	19.48	339	67.40	66	13.12
	Overlap	62	22.63	185	67.52	27	9.85

When comparing the RMSEs and MAEs, the results for both profilers are fairly consistent between the overlapping LD50 predictions (Table 5) and the full set of predictions (Table 3), with only marginal changes occurring.

However, the r^2 for the TIMES predictions of the overlapping chemicals (0.255) is half that of the full set of chemicals with TIMES predictions (0.54). I don't know if that is statistically significant but it does seem like a rather large drop-off in how much of the variance the TIMES predictions can account for.

Table 5: Performance metrics for overlapping chemicals with TEST and TIMES predictions

Software	Chemical set	RMSE	R squared	MAE
TEST	Full	0.642	0.296	0.469
	Overlap	0.643	0.270	0.457
TIMES	Full	0.620	0.540	0.447
	Overlap	0.650	0.255	0.459

As you can see in Figure 8, the density plot of the residuals for both profilers has a large peak around 0 with a longer right tail than would be expected if the values were normally distributed. This time though, the residuals have a much larger right-skew (i.e. towards underestimation of experimental LD50) than before when we looked across the full set of chemicals with a prediction.

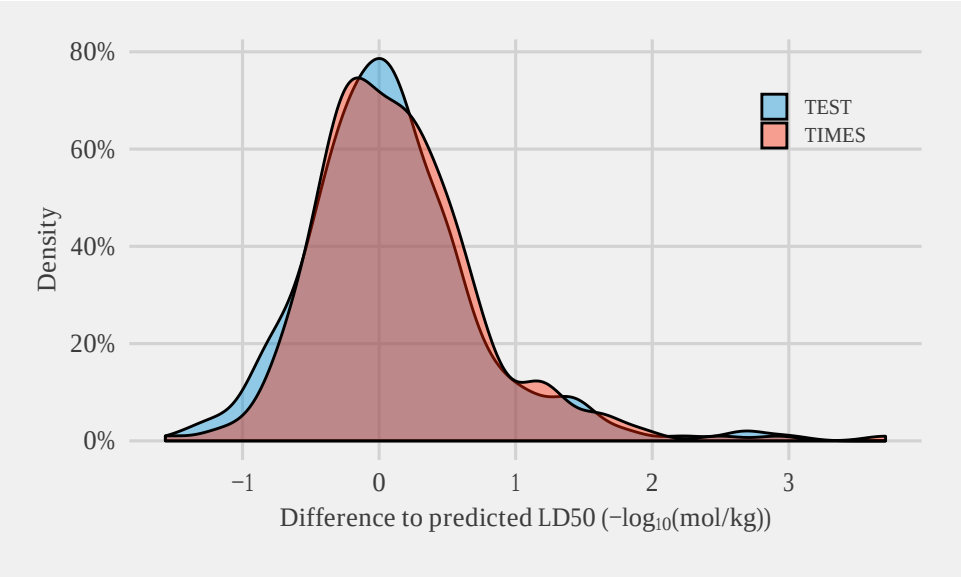

Figure 8: Density plot of residuals of overlapping chemicals with TEST and TIMES predictions

The figure below shows how well correlated the TEST/TIMES predictions for the overlapping chemicals are to the experimental values and to each other (Figure 9).

The correlation coefficient has decreased by quite a bit for the TIMES predictions of the overlapping chemicals compared to the full set of chemicals with TIMES predictions. Meanwhile, the correlation coefficient for the TEST predictions has only marginally dropped. The correlation between the TEST and TIMES predictions, however, has stayed the same.

Figure 9: Correlation of pairwise complete observations of overlapping TEST/TIMES predictions and experimental values

The reduction in the correlation coefficient and r^2 (and the marginal increase in RMSE) of the TIMES predictions, seems to suggest that there may be something in the overlapping chemicals that makes them more difficult to accurately predict the LD50.

1.4 Are there areas of chemical space where one prediction software works better?

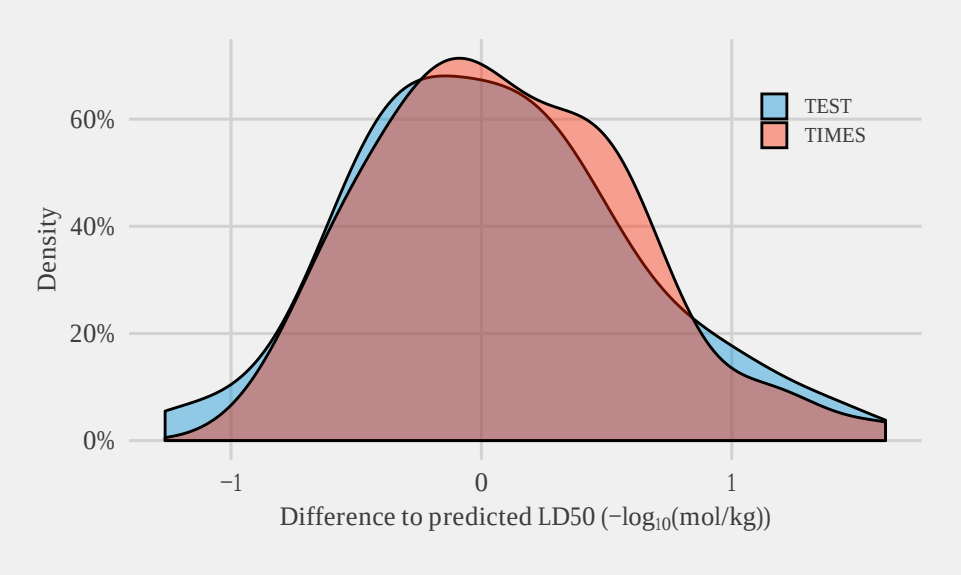

Figure 10: Distribution of residuals for chemicals containing bond 'COC_ether'

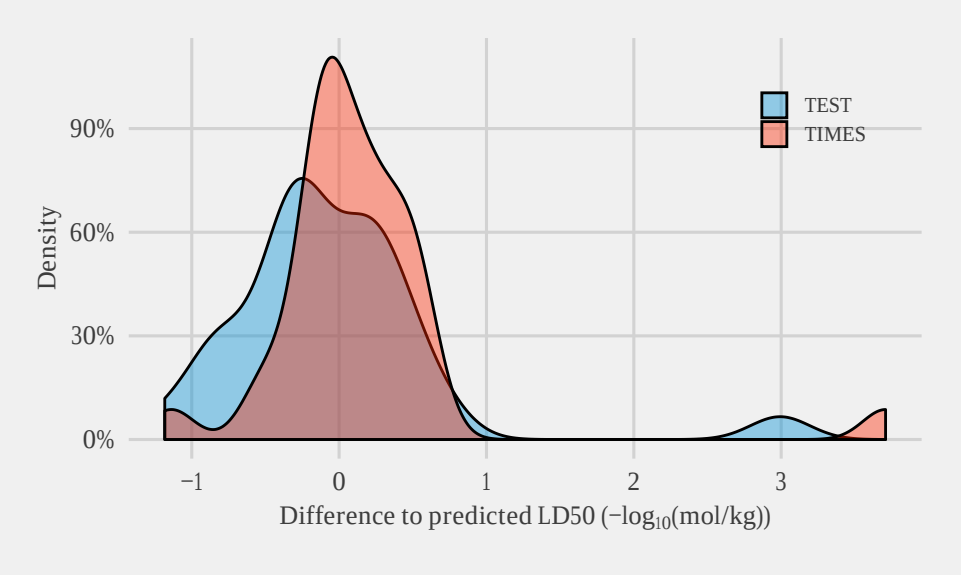

Figure 11: Distribution of residuals for chemicals containing bond 'CS_sulfide'