Fortnightly meeting 22nd October 2019

Mark Nelms

22/10/2019

Contents

Questions from last time

After our last meeting we had some questions to answer:

- What is the overlap in chemicals between the Escher/Carthew datasets and the ToxVal Inhalation set?
- Read the Chebekoue (Occupational TTC) and Hoersch (Inhalation TTC using DNELs) papers
  - How did they split their chemicals?
  - Could we also use this method?
- Can we identify enriched chemotypes for chemicals where the local NOEC was driving the general NOEC in Escher et al?
- Start exploring the Acute Toxicity work Jeremy had started
  - What unsupervised learning approaches could we use to ID and create groups?
    * Group chemicals using ToxPrints - what groups form?
    * Remove ToxPrints with 0 and near-0 variance & ToxPrints with all 0 columns, create PCA

1 Inhalation TTC

Filtering criteria for inhalation studies:

- Test species: Rat, mouse, other rodents, or rabbit
- Route of exposure: Inhalation
- Study duration: subchronic, chronic, reproductive, developmental, and multigeneration
- POD type: NO(A)EL, NOEL, NO(A)EC, NOEC

1.1 What is the overlap in chemicals between the Escher/Carthew and ToxVal datasets?

One of the things I hadn't looked into was the overlap in the chemicals present in the Escher/Carthew datasets and what, if any, overlap there was with the chemicals in our ToxVal inhalation dataset. Let's do that quickly now. NB: All comparisons are going to be using the DTXSIDs I gathered from the CompTox Chemicals Dashboard.

There are:

- 37 chemicals present in both the Escher and Carthew datasets
  - 18.23% of chemicals in Escher

    - 40.22% of chemicals in Carthew
- 99 chemicals present in both the ToxVal inhalation and Escher datasets
    - 48.77% of chemicals in Escher
    - 20.71 of chemicals in ToxVal inhalation
- 52 chemicals present in both the ToxVal inhalation and Carthew datasets
    - 56.52% of chemicals in Carthew
    - 10.88% of chemicals in ToxVal inhalation

1.1.1 Compare the NO(A)ECs between Escher/Carthew and ToxVal inhalation datasets

Now we have a handle on the number of chemicals that overlap between the different datasets we can compare the toxicity values between the Escher/Carthew dataset and the ToxVal inhalation dataset.

Let's first compare the ToxVal inhalation toxicity values to the general NOEC values from Escher (Figure 1.). I chose to use the general NOECs as there is one of these for each chemical.

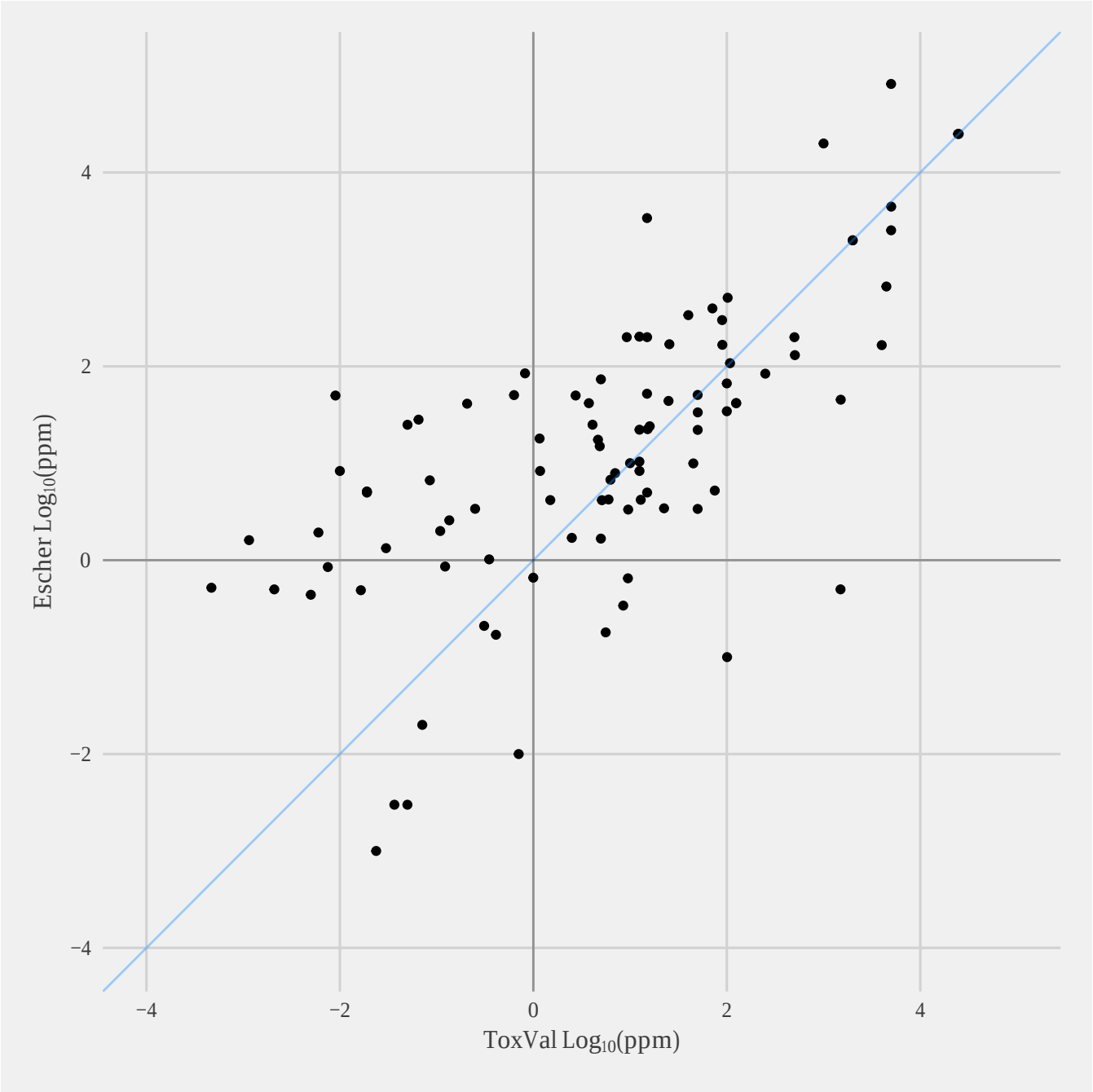

Figure 1: Comparing toxicity values for overlapping chemicals in ToxVal and Escher datasets.

In Figure 1, where a point is **above** the blue diagonal line this shows that the toxicity value from ToxVal is lower than that from Escher (i.e. the value from ToxVal is more potent than that from Escher). Meanwhile, if a point is **below** the blue diagonal line the toxicity value from Escher is lower than that from ToxVal (i.e. the value from Escher is more potent than that from ToxVal).

As you can see, there looks to be slightly more chemicals with a lower toxicity value in ToxVal than in the Escher dataset.

Now, let's have a look at the comparison between ToxVal and Carthew. So that we were comparing the most potent toxicity value from the Carthew dataset (whether that be the local or systemic value), I first had to make some adjustments to the Carthew dataset:

- Converted the systemic NO(A)EL in mg/kg-day to mg/m^3
- Compared this systemic NO(A)EC to the local NO(A)EC to identify the minimum toxicity value
- Converted the minimum NO(A)EC from mg/m^3 to ppm

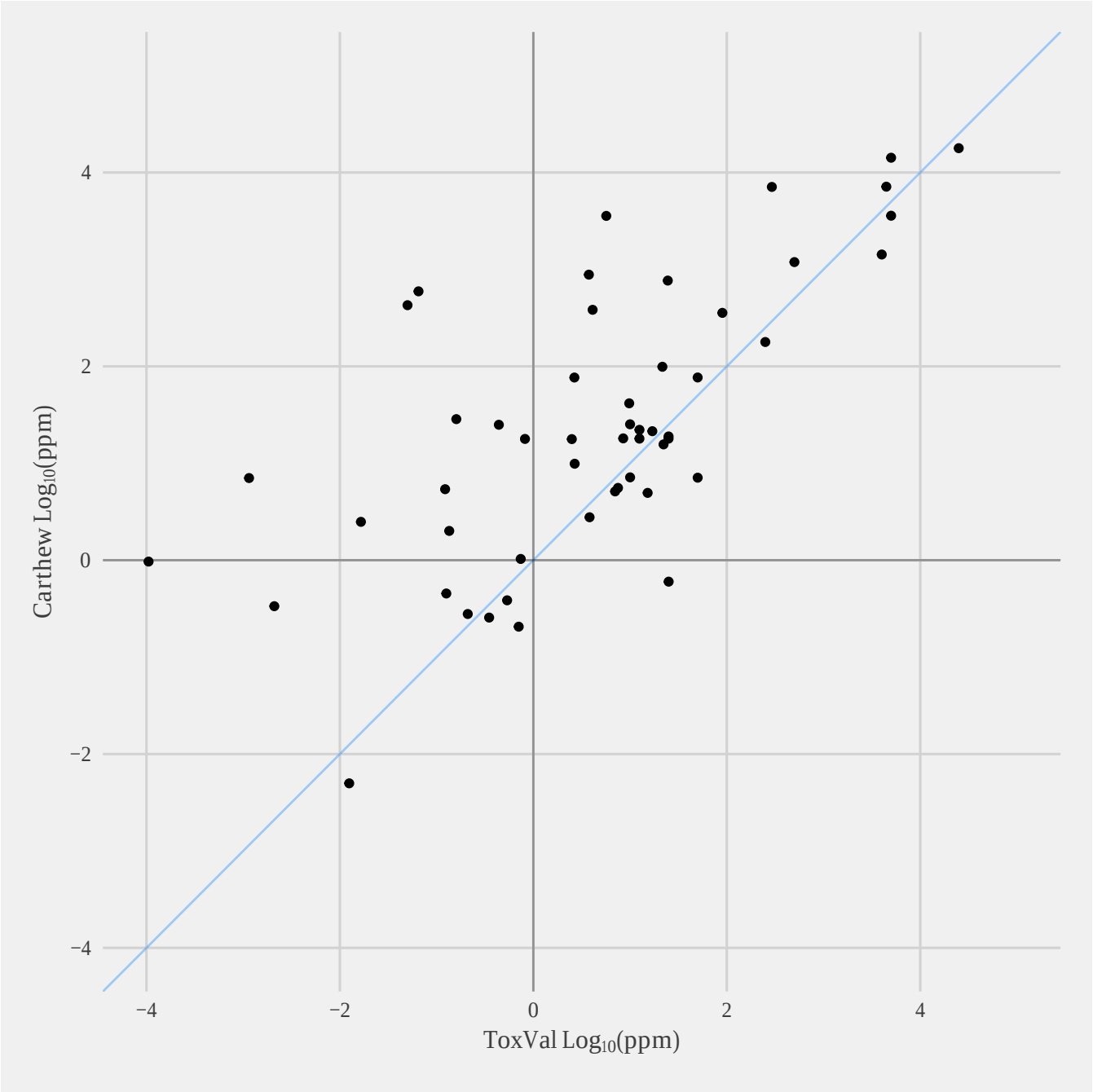

Figure 2: Comparing toxicity values for overlapping chemicals in ToxVal and Carthew datasets.

As with Figure 1, if a point is **above** the blue line in Figure 2 then the toxicity value is lower in ToxVal than in the Carthew dataset (i.e. the value from ToxVal is more potent than that from Carthew) and vice-versa when the point is below the diagonal line (i.e. the value from Carthew is more potent than that from ToxVal).

As you can see in Figure 2, the vast majority of chemicals have (slightly) more potent toxicity values in ToxVal than in the Carthew dataset.

Looking at both of these Figures, it appears that the toxicity values held in ToxVal are consistently lower (i.e. more potent) than those in both the Escher and Carthew datasets.

---

2 Occupational TTC and Inhalation TTC using DNELs papers

2.1 Hoersch et al (2018)

Retrieved DNEL values for 4,799 chemicals from the GESTIS DNEL list

- The TTC value they calculate is for workers not the general public
- These DNELs are for workers so (probably) have been calculated using the workers adjustment factor of 5 rather than the adjustment factor of 10 for the general public

They removed chemicals that are:

- Polymers
- Radioactive
- Non-isolated intermediates
- Food/feeding stuffs
- Medicinal products
- Substances in Annex IV (e.g. natural substances)
- Substances in Annex V (e.g. where registration is deemed inappropriate/unnecessary)
- Biocides/pharmaceutical actives - as not covered by REACH
- Suspected carcinogens, mutagens, or respiratory sensitisers (according to Annex VI of the Regulation on Classification, Labelling, and Packaging)

After employing these filtering criteria they had 1,876 chemicals remaining. Chemicals were **NOT** split into separate Cramer classes - all data were used to calculate one overarching TTC value, similar to the original Threshold of Regulation by FDA.

The **distribution-free** Excel function `QUANTIL` was used to estimate percentiles. They calculated the 0.1%, 1%, and 5% percentiles (described as the 99.9%, 99%, and 95% percentiles, respectively, in the paper). No additional adjustment factors were required as the DNEL values already contain assessment factors described by the ECHA guidance R.8 document.

The 1% percentile DNEL was chosen for their TTC value. This was chosen because of the uncertainties associated with individual DNELs. For example:

1) DNEL values are often based on data from related chemicals and not the test chemical itself (i.e. read-across)
    - according to ECHA this was used in about 75% of analysed dossiers for at least one endpoint.
2) Quality of DNEL depends on expertise of the submitter - this can lead to fluctuations
3) Transcription errors may occur - due to the large volume of data

The **1**% percentile TTC calculated was **50**μg/m^3. This can be converted to a value of 500μg/person/day (at breathing volume under light activity of 10m^3), or 7μg/kg-day (with a body weight of 70kg and breathing volume of 10m^3). At the 1% percentile the systemic inhalation TTC was also protective of reproductive toxicity effects - this wasn't the case when using the 5% percentile.

2.2 Chebekoue and Krishnan (2017)

Retrieved 8hr TLV-TWA (Threshold Limit Value - Time Weighted Average) data for 508 chemicals from the ACGIH booklet. Again the TTC values they derive are for workers not the general public. TLVs are health-based values that are given with the indication of target organs, this information was used to identify those chemicals that were based on systemic effects.

They removed chemicals that had:

- A CAS not referring to a single substance/not recognised by programs used to complete the work
- A TLV soley based on irritation, sensitisation, or dental erosion

These filters removed 139 chemicals. The remaining chemicals were then processed through the Kroes workflow. A total of 280 chemicals were retained and assigned to one of the three Cramer classes.

Physicochemical properties were calculated using EPA's EPI Suite (v.4.11). These properties were used to derive the Pulmonary Retention Factor (PRF) and workers daily dose (i.e. the amount absorbed by a worker whilst performing light activity during an 8hr workday).

For each chemical these three steps were followed:

1) QPPR from Buist et al. (2012) was used to calculate log blood:air coefficient ($LogP_{ba}$) (Eq. (1))

2) $LogP_{ba}$ was then used to calculate PRF

i) Eq. (2) used for substances with water solubility > 10mg/l
ii) Eq. (3) used for substances with water solubility < 10mg/l

3) PRF then used to calculate workers daily dose, after TLV had been converted to mg/m^3, if necessary. (Eq. (4))

They had experimental pulmonary retention factor data for a subset of the data (27 chemicals). These 27 chemicals were used to test how well the predicted values did against the experimental PRF values. They showed that the predicted PRF values were reasonably close to the predicted PRF values so the predicted PRF was used for the entire set.

As with the Hoersch et al. paper, no uncertainty factors were used as these were already taken into account when calculating the TLVs.

The results and discussion were kind of confusing and a little contradictory. Apparently, normality tests showed the data distributions for each Cramer class were *not* normal - either for the daily dose or for its logarithm. Therefore, they used nonparametric methods to calculate the 5th percentiles.

However, when they fit different distributions to the data the lognormal distribution was the best fit. Additionally, the chi-square and a one-sample K-S test indicated the daily doses were lognormally distributed. Maybe the data are lognormally distributed after all?? 🤷‍♂️

They calculate the TTC values at the 5th, 10th, and 25th percentiles by fitting a lognormal distribution to the CDF. At the level of the 5th percentile the TTC values they come up with are: Cramer class I - 0.07mmol/day, Cramer class II - 0.004mmol/day, and Cramer class III - 0.003mmol/day.

2.3 Equations

Blood:air coefficient

$$LogP_{ba} = 6.96 - 1.04 \times Log(VP) - 0.533 \times LogP_{OW} - 0.00495 \times MW \tag{1}$$

where $LogP_{ba}$ (unitless), is blood:air partition coefficient; VP (Pa at 25°C), is vapour pressue; $LogP_{OW}$ (unitless), is octanol:water partition coefficient; and MW (g/mol), is the molecular weight.

Pulmonary Retention Factor

$$PRF_{w+} = (36.608 + 9.799 \times LogP_{ba})/100 \tag{2}$$

$$PRF_{w-} = (26.810 + 21.022 \times LogP_{ba})/100 \tag{3}$$

Worker's daily dose

$$AD = \frac{TLV \times T \times V \times PRF}{MW} \tag{4}$$

where AD (mmol/day), is the absorbed dose; TLV (mg/m^3), is the 8hr TLV-TWA value; T (hr), is duration of exposure (8-hr workday used); V (m^3/hr), is the lung ventilation rate (10m^3/8hr/day for light activity); PRF, is the fraction of dose absorbed by inhalation route; and MW (g/mol), is the molecular weight

---

3 Exploring the Acute Toxicity work Jeremy had started

3.1 Data

3.1.1 LD50 values and chemical structures

The acute data came from the acute tox work group and consists of 21,200 rat oral LD50 values for 15,688 unique chemicals, which were collected from a variety of sources, including OECD's eChemPortal, ChemIDplus, and JRC's Acutoxbase. The CompTox dashboard and other public sources were utilised to identify the structures for these chemicals (represented as SMILES strings).

There were a total of 11,992 unique chemicals (16,173 LD50 values) for which SMILES strings could be found. For those chemicals with ≥ 3 LD50 values, Agnes calculated the median of the lower quartile and this value was retained as the representative LD50.

To ensure consistency when making predictions with the TEST and TIMES softwares the QSAR-ready SMILES were retrieved from the CompTox dashboard and used when profiling the chemicals. This had the added benefit of removing chemical mixtures; however, it did reduce the number of chemicals profiled to 9,345[1]. Additionally, the molecular weight for each chemical was also retrieved from the Dashboard.

3.1.2 TEST and TIMES

As the 9,345 chemicals for which we had QSAR-ready SMILES were profiled in batches of either 250 (TEST) or 1000 (TIMES) chemicals, there are mutltiple results files for each software: 38 for TEST and 10 for TIMES. Jeremy (or someone) had written a Python script to concatenate these results into .tsv files - one for the TEST results and one for the TIMES results.

3.2 Replicating Jeremy's current results

When I started trying to replicate Jeremy's work I used the .tsv files present in the Acute toxicity software evaluation folder on the O: drive. However, a stumbling block I came across was when reading in the "TESTresults.tsv" file:

- Jeremy used `read.csv()` separated using a tab and had 14,670 observations
- However, I used `read_tsv()` and only had 9,332 observations

After investigating, some of the rows in the "TESTresults.tsv" file contained an extra column with no column header. When Jeremy used `read.csv()` the values in the extra column were being silently wrapped underneath creating extra rows with only 1 value (Table 1).

Table 1: How read.csv handles rows with extra column - silently wraps data onto new line

X.	ID	Exp	Pred_Hierarchical.clustering	Pred_Nearest.neighbor	Pred_Consensus
1	30544-47-9	3.10	3.10	1.98	1.96
2.34					

Going back to the original set of 38 .txt files that had been concatenated to create the "TESTresults.tsv" file I saw that the first 16 files in the TESTresults folder contained only 3 predictions columns (i.e. hierarchical clustering, nearest neighbour, and consensus predictions). Whilst the remaining files contained 4 predictions columns: FDA predictions, as well as the hierarchical clustering, nearest neighbour, and consensus predictions (Table 2).

Table 2: FDA predictions column (in red) is additional column present in about half of the TEST rsults files

ID	Exp	Pred_Hierarchical clustering	Pred_FDA	Pred_Nearest neighbor	Pred_Consensus
3393-59-7	4.66	4.92	5.45	5.23	5.20
3393-60-0	4.52	4.77	4.87	4.82	4.82
35822-46-9	4.83	5.61	6.06	5.85	5.84

[1]When I re-ran the CAS numbers through the Dashboard to get the MW I retrieved QSAR-ready SMILES for 10,886 chemicals (10,212 unique QSAR-ready SMILES), do we want to make predictions for these additional 1,541 chemicals?

Because the "Pred_FDA" column (when present) occurred between the hierarchical clustering and nearest neighbour prediction columns (Table 2), the nearest neighbour and consensus values were not correct for the rows with the extra column in the TESTresults.tsv file.

> i.e. the FDA predictions were under the nearest neighbour heading, the nearest neighbour predictions were under the consensus heading, and the consensus predictions were being removed as they were the values that made up the extra rows.

To overcome this, I went back and read in the .txt files directly using `read_delim()` as this allows you to specify which columns to read in based upon the column name; because of this I decided to go back to the original files for the TIMES data as well.

3.2.1 Comparing predicted and experimental LD50s

The first thing Jeremy did was merge the predicted LD50s with the experimental LD50s. The consensus prediction was the LD50 taken from the TEST dataset.

Next, he converted the experimental LD50s and the TEST/TIMES predictions into $\log_{10}$(mg/kg) space. This was straightforward for the experimental and TIMES values as these LD50s were already in mg/kg. The TEST predictions, however, were in -$\log_{10}$(mol/kg) space, to convert these values into $\log_{10}$(mg/kg) he used this equation (Eq. (5)):

$$log10(mg/kg) = \left(\frac{-log10(mol/kg)^{10} \times MW}{1000}\right) \tag{5}$$

This is another stumbling block as I don't think this is right: the result of the top portion of the equation doesn't equal the prediction in mg/kg from TEST in the Dashboard. For example, CAS 100-21-0, -$\log_{10}$(mol/kg) = 1.63, LD50(mg/kg) = 3913.28:

$$\begin{aligned} log10(mg/kg) &= \left(\frac{-1.67^{10} \times 166.1308}{1000}\right) \\ &= \left(\frac{21,995}{1000}\right) \\ &= log_{10}(21.995) \\ &= 1.34 \end{aligned}$$

As you can see none of the values in the top portion of the equation equal, or are even close, to the LD50(mg/kg) value of 3913.28 from the TEST predictions in the Dashboard.

I think the equation should be:

$$log10(mg/kg) = 10^{-log10(mol/kg)} \times MW \times 1000 \tag{6}$$

If we use the same chemical as above, i.e. CAS 100-21-0, and plug the values into this new equation we can see that the value that we $\log_{10}$ transform is close to that for the mg/kg TEST prediction from the Dashboard. The difference between the mg/kg value calculated using equation (6) and that in the Dashboard can be explained by a rounding error of the -$\log_{10}$(mol/kg) value. If we plug the mg/kg value from the Dashboard (3913.28) into the inverse formula we end up with a -$\log_{10}$(mol/kg) value of 1.6279, which gets rounded to 1.63.

$$\begin{aligned} log10(mg/kg) &= 10^{-1.63} \times 166.1308 \times 1000 \\ &= log_{10}(3894.406) \\ &= 3.59 \end{aligned}$$

Once converting the values into $\log_{10}$(mg/kg) space, Jeremy had calculated the difference between the TEST/TIMES predictions and the experimental values, and plotted this as a dot plot histogram.

After our talk yesterday, I have started running the chemicals for which we now have QSAR-ready SMILES through the acute toxicity profiler in TIMES and I'll bring them into the fold.

Also, rather than converting the TEST predictions from $-\log_{10}$(mol/kg), I'm going to instead convert the experimental and TIMES predictions from $\log_{10}$(mg/kg) to $-\log_{10}$(mol/kg).