REACH*Across*™ Software

Artificial intelligence to combine read-across & (Q)SAR

UL Environment

The unprecedented rate of chemical classification and labelling performed as a result of research efforts and government regulation has uncovered new opportunities for toxicological modeling. REACH*Across* Software merges read-across approaches with machine learning. Hundreds of thousands of chemical labels and billions of chemical-chemical comparisons result in enormous chemical space networks which graph algorithms and machine learning can leverage to predict hazard.

Read-Across | QSAR | REACH | REACH*Across* Software 1.1.0

REACH*Across* software is a new kind of platform for predicting chemical hazards. UL Environment's integrated cheminformatics database merges research, government, and industry hazard data to provide a singular portal for modeling chemical hazards. Cluster computing in combination with UL's integrated database serves to create an enormous network of over 31 billion chemical similarities. REACH*Across* Software version 1.1.0 provides hazard estimation for eight human health hazards required by REACH regulation at all tonnage bands. Future releases will incorporate hazard estimations for all human health hazards and most ecological toxicity and environmental fate endpoints.

Concept. Chemical similarity claim that chemical structures that share many chemical features share biological activity. REACH employs chemical similarity in read-across submissions, wherein experts describe how similar chemicals should induce similar biological effects. These approaches only perform well when a large number of chemicals have been well described. It does no good to find similar chemicals about which nothing is known. Newfound success in these models is primarily due to impressive growth in the availability and size of chemical data with toxicological information. REACH*Across* Software breaks the traditional chemical similarity workflow into three main components:

1. **Fingerprinting** Chemical fingerprints are vectors describing features of a chemical. Is it a halogen? What is the molecular weight? How many rings does it contain? REACH*Across* Software 1.1.0 uses Pubchem2D a popular fingerprinter supported by Pubchem.

2. **Similarity** Chemical Similarity is a function of two chemical fingerprints. Similarity functions approximate a probability that two chemicals fingerprints will have the same hazard. REACH*Across* Software 1.1.0 uses tanimoto similarity which is simply the fraction of shared features over total number of features in both chemical fingerprints.

3. **Network Features** The UL integrated database contains 250,000 chemicals with hazard labelling data. Steps 1 and 2 are repeated for every pair of two chemicals. This results in a large network of 250,000 chemicals with 31 billion similarities.

4. **Machine Learning** Once the global similarity network is constructed in step 3. REACH*Across* Software 1.1.0 derives **network features** for each chemical. These features are chemical and endpoint specific. Figure 1 shows the local network for the target compound (1-DECENE). This local network shows two simple and powerful features, closest negative and closest positive. Chemicals tend to be hazardous when they are very similar to another hazardous compound, particularly when they are not close to any negative compounds. REACH*Across* Software 1.1.0 trains a statistical model (logistic regression) on a large training set of labeled chemicals with network features to predict probabilities of hazard.

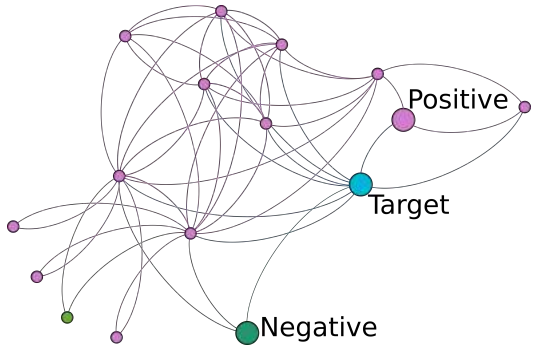


Fig. 1. Local similarity graph for 1-DECENE. Shows chemicals with similarity > 0.9 according to pubchem2d tanimoto. REACH*Across* Software uses similarity to the closest Positive (large pink node - 1,7-OCTADIENE) and closest Negative (large green node - MYRCENE) along with other features to characterize a local similarity space.

REACH*Across* Software is improving daily but already gives strong performance for eight REACH endpoints.

Significance Statement

REACH*Across* Software is a generalized cheminformatics platform that combines Read-Across with QSAR. It is built from UL's integrated cheminformatics database which combines some private and most publicly available academic, industrial, and government databases. It currently supports eight required REACH Endpoints required for all tonnage levels in REACH Annex VII.

- Skin sensitization
- Eye Irritation
- Acute Oral toxicity
- Mutagenicity
- Skin Irritation / Corrosion
- Acute Dermal toxicity
- Acute Aquatic
- Chronic Aquatic

These cover ANNEX VII all tonnage bands requirements.

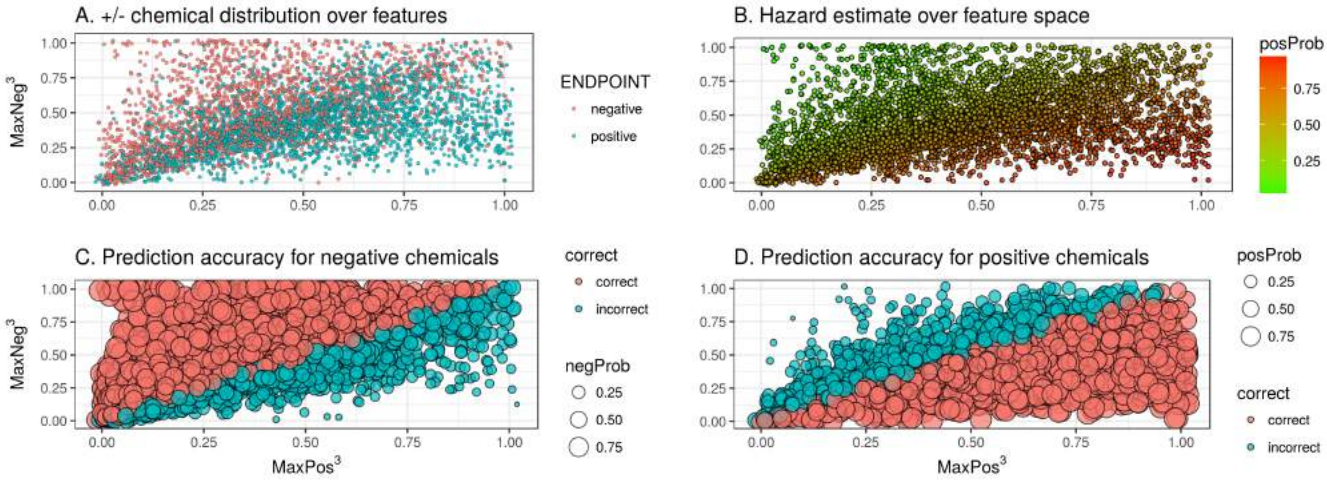


Fig. 2. These graphs show how skin sensitizers and non sensitizers distribute over features describing the closest negative and positive chemicals.

The Case for Similarity. REACH*Across* Software 1.1.0 operates by building network features for chemicals. These network features describe chemicals by their local similarity network. Figure 2 describes two network features $\mathbf{maxPos^3}$ and $\mathbf{maxNeg^3}$. These features describe a chemicals similarity to the closest positive chemical and closest negative chemical (in this case the closest sensitizer and non sensitizer).

Figure 2 A. shows that there are many more negatives in the upper left quadrant portion and many more positives in the lower right portion with a dividing line on the diagonal. This fits our intution. Chemicals that are very similar to a negative, and simultaneously not similar to any positives will tend to be negative (and vice versa). Chemicals that have a high degree of similarity to both positives and negatives do not have a clearly defined difference.

Figure 2 B. shows the REACH*Across* Software 1.1.0 hazard probability estimate for each chemical in the skin sensitization dataset. A quick visual inspection sees high hazard probabilities in the lower right and low hazard probabilities in the upper left. This fits well with the data shown in 2 A.

Figure 2 C. and D. Show how negatives and positives (respectively) distribute across this feature space. As expected the majority of negatives are in the upper left and have high negative probabilities. There are relatively few negative chemicals receiving a high (small diameter) hazard probability.

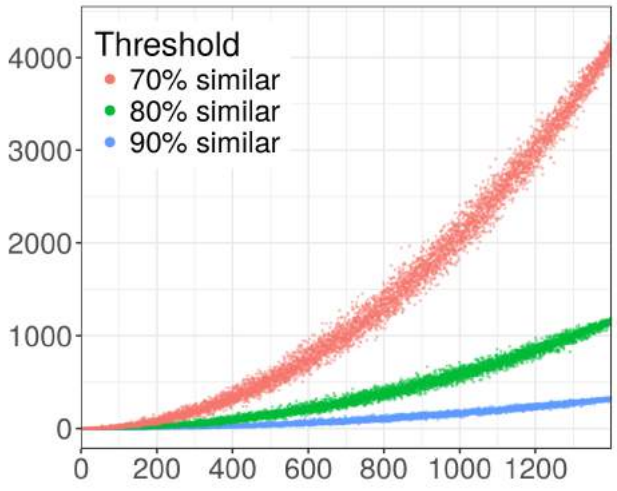
REACH*Across* Software uses these network features in logistic regression to make probabilistic estimates for chemical hazards. Each network feature is evaluated for its ability to contribute to accurate probability estimates. These figures demonstrate that for skin sensitization REACH*Across* Software 1.1.0 is able to make some high confidence estimates (upper left and lower right chemicals) and able to identify those chemicals for which it cannot make high confidence estimates (near diagonal). In this sense REACH*Across* Software knows what it knows.

Network Effect. The REACH*Across* Software platform improves as more labeled chemicals enter into the system. The algorithm uses local similarity networks to make predictions. As more connections are added to the network REACH*Across* Software improves. Metcalfe's law suggests that the number of connections in a network is proportional to the square of the number of entities. In this context the law states that the number of chemical-chemical similarities increases quadratically everytime a new labeled compound enters the UL integrated database.

This "network effect" is visualized in Figure 3 which uses the European Chemical Agencies short list of labeled compounds in REACH Annex VI table 3.1. Multiple random samples (10) are taken of increasing sample sizes from table 3.1. The pink, green, and blue points show the number of chemical-chemical similarities (or "connections") which have similarity $\geq$ 70%,80%, and 90%. The figure demonstrates that chemical similarity networks appear to obey Metcalfe's law.

Currently the REACH*Across* Software platform integrates 250,000 compounds with over 31 billion connections, although the number of compounds and type of connections vary depending on the endpoint being queried.

Fig. 3. Random samples of **sample size** are taken from a set of 1300 chemicals in REACH Annex VI table 3.1 of chemicals publicly classifed by ECHA. **Connections** counts the number of connections of $\geq$ **Threshold**. Connections are seen to increase as the square power of the sample size.



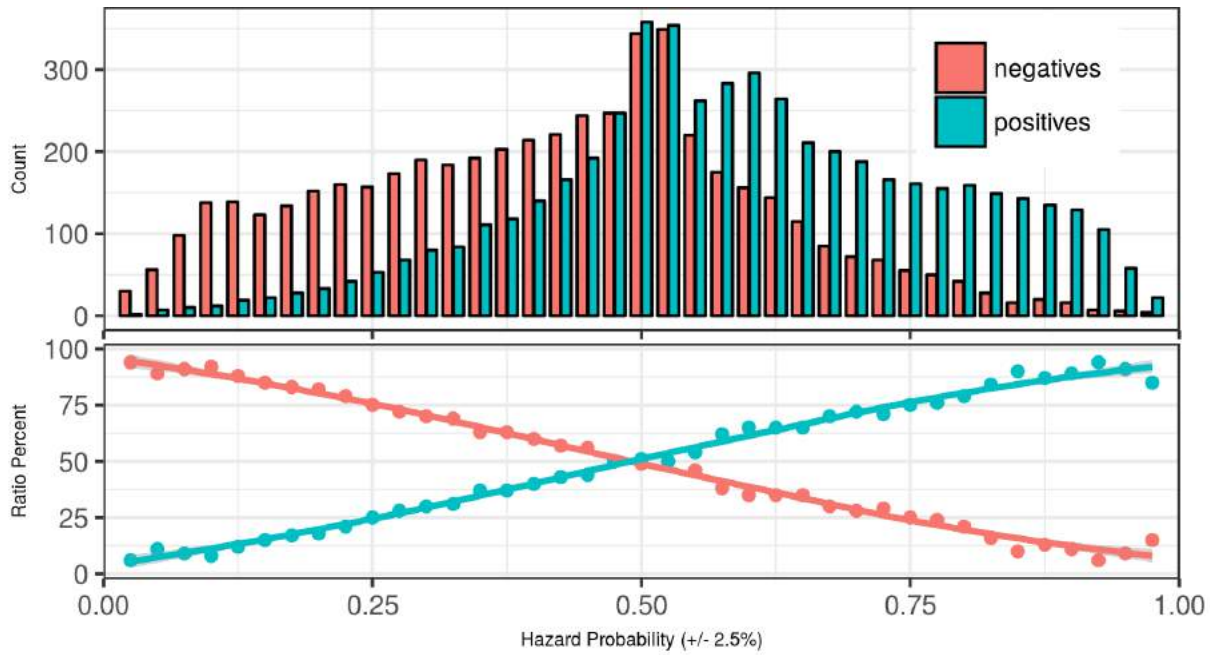


Fig. 4. These graphs show how sensitizers/nonsensitizers distribute over REACH*Across* Software hazard estimates. The left figure is a histogram counting the number of +/- chemicals receiving different probabilistic estimates (+/-2.5%). The right figure shows the percentage of +/- chemicals at each hazard probability estimate.

Handling Uncertainty. Chemical similarity approaches to hazard models suffer from poor probabilistic estimation. Most similarity methods make determinstic class predictions. The final stage of the REACH*Across* Software pipeline uses a probabilistic model (logistic regression) on network features. Figure 4 shows how sensitizers distribute over hazard probabilities.

$$\frac{1}{1+e^{-(B_1x_1+B_2x_2+\ldots)}} \quad [1]$$

Note on Coverage. REACH*Across* Software does not define a mechanistic domain of applicability. Instead, a probabilistic domain of applicability is defined via negative and positive probability thresholds. The REACH*Across* Software domain is defined by those chemicals resting below the negative threshold and above the positive threshold. Equation 1, describes the general logistic equation which makes up the REACH*Across* Software probabilistic model. The B_i values are optimized coefficients and the x_i variables are network features. This domain of applicability determines the number of chemicals for which REACH*Across* Software is valuable. Appendix figure 5 shows the impressive level of coverage achieved with a relatively small number of labeled compounds.

Results

REACH*Across* Software 1.1.0 has been tested on the current European Chemical Agency Classification and labelling data for eight endpoints. Table 1 shows the results of this analysis along with the counts of positives and negatives for each endpoint. The balanced accuracies for every endpoint are greater than 70% and increase for stricter domains of applicability. Strong sensitivities and specificities exist for every endpoint, these can also be manipulated by changing the desired negative or positive threshold. At the chosen thresholds a large percentage of the chemicals in the ECHA C&L are covered. Additionally, a large percentage of chemicals in the European Inventory of existing commercial chemical substances (EINECS) are covered. EINECS is representative of available commercial chemical substances and the REACH*Across* Software applicability to new REACH submissions.

Links

1. **ulreachacross.com**: Purchase chemical reports.
2. **ulreachacross.com/documents**: View more analysis.

Table 1. Leave one out cross validation results. Se = Sensitivity, Sp = Specificity, Cover = ECHA C&L chemicals in REACH*Across* Software domain of applicability. eCover = percent of EINECS in REACH*Across* Software . ePos/eNeg = eCoverage predicted to be +/-.

Endpoint	Positive	Negative	Total	negT	posT	Se	Sp	TestCover
skin sensitization	2886	1897	4783	43%	50%	80%	50%	85%
eye damage	14794	966	15760	47%	55%	81%	51%	88%
acute oral	10225	1932	12157	40%	50%	80%	64%	87%
mutagenicity	600	2795	3395	42%	50%	80%	55%	83%
skin corr. / irrit.	13846	1377	15223	37%	52%	80%	51%	75%
acute dermal	4430	1997	6427	40%	60%	80%	69%	73%
acute aquatic	1129	926	2055	40%	50%	80%	52%	82%
chronic aquatic	2582	262	2844	40%	54%	80%	50%	80%

Appendix Figure - Coverage of EINECS

A graphical representation of the rapid coverage of the chemical universe by the REACH*Across* Software tool is shown below. In each figure labeled compounds from REACH Annex VI table 3.1 are added to 33,000 compounds selected from the European INventory of Existing Commercial chemical Substances (EINECS). Connections are shown between unlabeled EINECS compounds (blue) and highly similar Annex compounds (red). The top figure helps to visualize how EINECS and Annex compounds cluster with no minimum distance between chemicals. The bottom figure shows the coverage of the chemical space. We can see some ANNEX chemicals cover a very large number of EINECs chemicals. By using only 1000 labeled chemicals we are able to cover 33,000 unknowns. The UL integrated database can cover a much larger chemicals space (which is unfortunately difficult to visualize).

Fig. 5. Visualizations of EINECS coverage. 33,000 EINECS compounds are represented in blue and labeled ANNEX VI table 3.1 compounds are in red. Edges represent similarities between EINECS compounds and Annex compounds. Top graph shows clustering (no minimum node distance). Bottom graph shows high resolution nodes (forced distance between chemicals). Larger red nodes cover more EINECS chemicals.