Central Lancashire Online Knowledge (CLoK)

Title	Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across: Recommendations for Optimal Use
Type	Article
URL	https://clok.uclan.ac.uk/24855/
DOI	https://doi.org/10.1016/j.yrtph.2018.11.002
Date	2019
Citation	Mellor, Claire, Marchese Robinson, Richard, R, Benigni, Ebbrell, D, Enoch, S.J., Firman, J.W., Madden, J.M., Pawar, G, Yang, C et al (2019) Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across: Recommendations for Optimal Use. Regulatory Toxicology and Pharmacology, 101. pp. 121-134. ISSN 0273-2300
Creators	Mellor, Claire, Marchese Robinson, Richard, R, Benigni, Ebbrell, D, Enoch, S.J., Firman, J.W., Madden, J.M., Pawar, G, Yang, C and Cronin, M.T.D.

1

Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across:

Recommendations for Optimal Use

C.L. Mellor[1], R.L. Marchese Robinson[1], R. Benigni[2], D. Ebbrell[1], S.J. Enoch[1], J.W. Firman[1], J.C. Madden[1], G. Pawar[1], C. Yang[3] and M.T.D. Cronin[1]*

[1]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England

[2]Alpha-Pretox, Via G. Pascoli 1, 00184, Rome, Italy

[3]Molecular Networks GmbH, Neumeyerstraße 28, 90411 Nürnberg, Germany

*Corresponding author (Mark Cronin): Tel. +44 151 231 2402; e-mail address: M.T.Cronin@ljmu.ac.uk

ABSTRACT

Computational approaches are increasingly used to predict toxicity, in part due to pressures to find alternatives to animal testing. Read-across is the "new paradigm" which aims to predict toxicity by identifying similar, data rich, source compounds. This assumes that similar molecules tend to exhibit similar activities, i.e. molecular similarity is integral to read-across. Various molecular fingerprints and similarity measures may be used to calculate molecular similarity. This study investigated the value and concordance of the Tanimoto similarity values calculated using six widely used fingerprints within six toxicological datasets. There was considerable variability in the similarity values calculated from the various molecular fingerprints for diverse compounds, although they were reasonably concordant for homologous series acting via a common mechanism. The results suggest generic fingerprint-derived similarities are likely to be optimally predictive for local datasets, i.e. following sub-categorisation. Thus, for read-across, generic fingerprint-derived similarities are likely to be most predictive after chemicals are placed into categories (or groups), then similarity is calculated within those categories, rather than for a whole chemically diverse dataset.

HIGHLIGHTS

- Molecular fingerprints to identify read-across analogues have been evaluated

- Identification of read-across analogues is dependent on the molecular fingerprint

- Commonly used molecular fingerprints may not address the mechanism of toxic action

- Commonly used molecular fingerprints are most likely to be predictive within a homologous series

- Similarity measures tailored to the endpoint are likely to be most useful

1. INTRODUCTION

The use of alternative approaches to assess chemical safety is growing due to legislation that requires greater knowledge of the harmful effects of chemicals, whilst also requiring a reduction in, or avoidance of, animal testing. Alternative methods, including *in vitro* assays, -omics and computational approaches ((Quantitative) Structure-Activity Relationships ((Q)SARs), read across etc.) have become integral to many hazard assessment strategies. Of these, computational or (Q)SAR (*in silico*) approaches aim to predict the toxicity of compounds from descriptors of chemical structure and thus reduce testing. In particular, read-across is at the forefront of the prediction of toxicity and has been seen as the "new paradigm" for hazard assessment (Cronin et al, 2013; Berggren et al., 2015; Schultz et al, 2015; Schultz and Cronin 2017; Patlewicz et al 2018). Read-across relies on the ability to identify similar molecules with the assumption that similar molecules will tend to exhibit similar activity or, at least, show similar trends in activity (OECD, 2014). Although the concept of similarity has growing acceptance for toxicity prediction, in reality there are still a number of barriers to acceptance of the predictions, especially for regulatory purposes (Bender and Glen, 2004; Spielmann et al., 2011; Teubner et al., 2015; Ball et al., 2016; Schultz and Cronin 2017; Chesnut et al 2018). Of the barriers identified by Ball et al (2016), some are more trivial to address than others, e.g. full documentation and ensuring the correct chemical structure is provided. The most difficult aspect of justifying a read-across argument is the assessment of "similarity" and being able to provide evidence for such, so to build scientific confidence (Patlewicz et al., 2015; Schultz et al 2018). For instance, there is a concern over effects such as activity cliffs, where structurally similar compounds have a significant difference in potency (Guha and van Drie, 2008; Stumpfe and Bajorath, 2011; Cruz-Monteagudo et al., 2014). In addition, there is the on-going problem of how to define similarity from a molecular level (Maggiora et al., 2014) as well as adequately for read-across (OECD, 2014; Shah et al., 2016; Patlewicz et al 2018; Schultz et al 2018). It is

important to note that the similarity between any two objects may be calculated in a variety of different ways and relies on a definable set of features (or descriptors), as well as a means of qualitatively or quantitatively defining similarity based upon those variables. Molecular similarity is no different and whilst two molecules may appear highly similar in one aspect, for instance they may have the same molecular weight, they can be dissimilar in other aspects, such as chemical structure. Thus, the means of defining similarity and providing a means to calculate it is essential. This study has focused on molecular fingerprints due to their increased use in read-across through techniques such as machine learning (Luechtefeld et al., 2018). However, in the context of the current work, the focus is upon read-across predictions made using pairwise comparison to one, or a few, suitably "similar" chemicals, as may well be the case for practical applications. Some of the insights presented herein, regarding the strengths and weaknesses of molecular fingerprint derived similarity measures, may also be applicable in the context of these machine learning studies. Still, detailed examination of the pros and cons of the use of molecular similarity in the context of supervised machine learning, where relationships may be found based on the similarity computed to multiple tested chemicals within a large database, is beyond the scope of the current paper. To assist the reader, definitions are stated in Table 1 that are pertinent to this investigation.

TABLE 1 HERE

The read-across approach may be broadly defined as one in which quantitative or qualitative predictions of an endpoint of interest are made for a target chemical using endpoint data for one or more sufficiently similar source chemicals (OECD, 2014). Usually, this approach is envisaged as only being suitable following grouping of related chemicals, e.g. to form a category (OECD, 2014). There are a number of means of identifying "similar" molecules for grouping and read-across which are deemed acceptable for regulatory purposes, including use

93   of common, mechanistically relevant, structural features and transformation to the same

94   metabolite or degradant (OECD, 2014). There is also the more general concept of "chemical

95   similarity", i.e. using measures of similarity based on common structural features,

96   physicochemical or biological properties and / or calculated variables related to molecular

97   structure (descriptors). This broader notion of "chemical similarity", in contrast to those which

98   are deemed acceptable for regulatory purposes, may be defined in terms of generic structural

99   features / properties / variables, which are not necessarily relevant to the endpoint of interest.

100   These approaches use chemometrics, the science of using mathematics and statistics to analyse

101   chemical data in order to obtain knowledge about chemical systems; elsewhere, the term

102   cheminformatics or chemoinformatics may be used.) Chemometric measures of similarity are

103   widely used as they are rapid and cost effective due to the availability of online tools, e.g.

104   ChemMine Tools (chemminetools.ucr.edu/) and MuDRA (Alves, 2018), and software that can

105   be freely downloaded, e.g. Toxmatch (Patlewicz, 2008; 2017). Whilst the use of analogues and

106   mechanistically relevant fragment based methods to identify similar molecules for read-across

107   is relatively well developed (Schultz et al., 2015), much less is known about the use of

108   "chemical similarity", as defined above, for read-across. This is an area that was founded in

109   the identification of new leads for drug development, thus the similarity measures were not

110   necessarily intended for the purpose for which they are currently applied. For grouping and

111   read-across, where there is no rational measure to find similar compounds, or where a large,

112   diverse inventory is being searched, chemometric methods may seem appealing. However,

113   there is no clear guidance on how they may be applied.

114   The generation of chemometric similarity requires the conversion of chemical structures into

115   machine readable representations which are then compared using one of the many available

116   similarity coefficients (Willett et al., 1998; Holliday et al., 2003). The calculated similarity can

117   vary depending on the type of representation chosen and which similarity coefficient is used.

118   Most similarity calculations rely on the use of (molecular) fingerprints in order to generate

119   machine readable bit representations from chemical structure. Fingerprints are based mostly on

120   2D representations of a molecule and are used due to their computational efficiency (Holliday

121   et al., 2003). The process of generating bits from chemical structure is illustrated by Figure 1,

122   for a scenario in which the corresponding structural features are molecular substructures A

123   fingerprint is typically a binary vector, with bits set to 1 or 0 depending on the presence or

124   absence of a structural feature (e.g. molecular substructure) within the molecule of interest. In

125   principle, there does not have to be a simple one-to-one correspondence between the presence

126   of a structural feature and the presence of a molecular substructure. For example, one of the

127   features employed in the RDKit implementation of the MACCS fingerprint corresponds to

128   "two   or   more   methyl   groups"   (https://github.com/rdkit/rdkit-

129   orig/blob/master/rdkit/Chem/MACCSkeys.py). Moreover, other fingerprints might encode the

130   occurrence count of structural features, rather than simply their presence or absence. However,

131   if the fingerprint only encodes the presence or absence of certain fragments and not their

132   quantity, this may be a limitation (Flower, 1998). For this scenario, a molecule can contain a

133   specific fragment 1 or 100 times and the resulting bit string will be set the same, thus giving

134   little information with regards to, for instance, molecule size and which fragments occur more

135   often within a molecule (Flower, 1988).

136   **FIGURE 1 HERE**

137   Many different types of molecular fingerprints are used to calculate the similarity between two

138   molecules. Two of the most widely used are the molecular access system (MACCS) fingerprint

139   and the extended connectivity fingerprint (ECFP). The MACCS fingerprint was one of the first

140   developed and is amongst the most commonly used for similarity calculations. MACCS is a

141   prototypic fingerprint, which typically contains 166 structural features, related to the presence

142 and occurrence count of substructures comprising a variety of non-hydrogen ("heavy") atoms

143 (Maggiora et al., 2014), albeit this may be implementation dependent

144 (http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html,

145 https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py). The ECFP

146 defines molecular features by assigning identifiers to each of the heavy atoms in the molecule,

147 based upon atomic properties and bonding arrangements, and then combining those identifiers

148 with those assigned to neighbouring heavy atoms up to a specified number of bonds away

149 (Rogers and Hahn, 2010). The most commonly used ECFP fingerprint is ECFP4, which has a

150 bond diameter of four. ECFP4 comprises features derived from the compounds in the analysed

151 dataset, which necessarily overlap, in contrast to the MACCS fingerprint, for which the features

152 are pre-defined (Maggiora et al., 2014). In simple terms, approaches such as ECFP are more

153 complex than MACCS, allowing for the generation of many different atom environments and

154 describe molecular structure more subtly. Finally, it should be noted that different variants of

155 both fingerprints may be computed by different software programs (Rosenbaum et al., 2011;

156 http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html,

157 https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py).

158 A coefficient is used to assess the similarity of two, or more, molecules as defined by the

159 fingerprints. The similarity coefficient most frequently combined with the use of fingerprints

160 is the Tanimoto coefficient (Tc). (Elsewhere, this may be termed the Jaccard similarity (Willett

161 et al., 1998; Luechtefeld et al., 2018).) For molecules described in terms of bit-vector molecular

162 fingerprints, Tc is computed as per equation (1), albeit a more general definition exists for

163 continuous variables (Willett et al., 1998).

164

165 $$\text{Tc (A, B)} \quad = \frac{c}{a+b-c} \qquad (1)$$

166

In equation (1), the Tanimoto coefficient (Tc) for the similarity of two objects (molecules) A and B is a function of the number of features present within compounds A and B (a and b respectively), and the number of features shared by A and B (c). With regard to molecular fingerprints, a and b are the number of structural features, or bits set to 1, in each molecule, c is the number in common. Therefore, Tc quantifies the fraction of features common to A and B as a fraction of the total number of features of A or B, where the c term in the denominator corrects for double counting of the features (Willett et al., 1998; Maggiora et al 2014). It is obvious, therefore, that the Tc calculated is dependent on the type of fingerprint method applied. Thus, should Tc be used for grouping or read-across within a group, the type of fingerprint applied is vital. Also of relevance to read-across is the value of Tc that would constitute molecules being considered to be sufficiently similar for read-across predictions of a given endpoint to be made for a target compound based upon endpoint data for the similar source compounds (OECD, 2014). There is no definitive rule or guidance for use of Tc or specific fingerprints, in part due to the differences in calculated values. Within the drug design community, it is often considered that knowledge of the point at which the similarity of A and B reaches a 'threshold' point, where they exhibit similar biological activity, is required. For more than 15 years, a Tc value of 0.85 was widely considered this 'threshold' value for bioactivity (Maggiora et al 2014). However, studies have since shown that this value is not reliable, especially when different molecular representations are used (Eckert et al., 2007; Stumpfe et al. 2011; Martin et al., 2002). Despite these issues, Tc is widely used as a measure of molecular similarity as it is simple to calculate and is readily available in easy-to-use tools, some of which are online and some of which are freely available to download (Whittle et al., 2004; Salim et et al., 2006; Rogers and Hahn, 2010; Todeschini et al., 2012; Reisen et al., 2013; Willett, 2013; Bajusz et al., 2015, Cereto-Massague et al., 2015).

191 Whilst widely applied, a number of studies have shown that using Tc to calculate chemical

192 similarity has its limitations and weaknesses (Dixon and Koehler, 1999; Flower, 1998;

193 Holliday et al., 2002; Laiiness, 1997). Godden et al (2000) demonstrated that Tc has a tendency

194 to produce a similarity score of about 0.3 even for structurally distant molecules. It has also

195 been suggested that Tc calculations are biased towards smaller molecules when used for

196 selection according to diversity and that other coefficients may be more appropriate for some

197 data types (Dixon et al., 1999; Lajiness et al., 1997; Whittle et al., 2003). Moreover, as is

198 perhaps most relevant for the purposes of toxicity prediction, Tc is a generic measure of

199 molecular similarity which treats the shared presence of mechanistically irrelevant

200 substructures as equally important as the shared presence of mechanistically crucial

201 substructures, such as those corresponding to structural alerts (Alves et al., 2016). One way of

202 taking account of this is to compute a weighted Tanimoto index (Maunz et al., 2008).

203 Nonetheless, in spite of its known limitations, a Tanimoto similarity of 0.7 is elsewhere

204 considered as a cut-off for read-across (Enoch et al 2009; Hartung, 2016).

205 The aim of this study was to determine the value of different molecular fingerprints to assess

206 molecular similarity, in terms of the Tanimoto coefficient, in the context of read-across. In

207 particular, the focus of the study was to examine scenarios in which these similarity values

208 might be useful for read-across based upon pairwise comparison to one or a few chemicals,

209 with measured endpoint data, for the purpose of toxicological data gap filling. Specific

210 objectives were to assess the performance and reliability of different molecular fingerprints

211 used in similarity analysis, with a view to determine when similarity computed in this fashion

212 works well and does not work well, as well as to consider how molecular similarity can be

213 placed into a mechanistic framework to predict toxicity taking in account molecular initiating

214 events (MIEs) (Allen et al., 2016, Cronin et al., 2017; Cronin and Richarz, 2017). It should also

215 be made clear that the purpose of this study was not to conclusively establish an optimum

216   method for predicting toxicity. Rather, the purpose of this study was to gain a better

217   understanding of chemical similarity, calculated in terms of the widely used Tanimoto

218   coefficient and generic chemical fingerprints, its strengths, weaknesses and how best to make

219   use of it for read-across based upon pairwise comparisons to one, or a few, chemical(s).

220   To achieve the objectives of this study, six datasets were analysed and these are summarised

221   in Table 2. The datasets were small in size (from 7 to 211 compounds) compared to more

222   complex inventories, e.g. of REACH chemicals, or databases that may be investigated for drug

223   discovery. The selection of the datasets was influenced by a number of factors. Datasets were

224   chosen which had been the subject of previous read-across or QSAR analyses, or potentially

225   could be used as such. These were datasets that the authors were familiar with, hence allowing

226   for an understanding of the selection process for compounds as well as the quality of the

227   underlying biological data. They were also chosen to represent a range of mechanisms and

228   molecular initiating events which may influence the use of molecular similarity.

229

230   **2. METHODS**

231   **2.1 Data Sets Analysed**

232   In total six different datasets were chosen to calculate Tc in this study. These datasets were

233   chosen as they provided different read-across scenarios, thus allowing similarity calculations

234   based on different fingerprints to be assessed for reliability/ accuracy. The six data sets (Table

235   2) chosen were analysed and a Tanimoto score for each pair of chemicals within each data set

236   was calculated for the different fingerprints.

237   **TABLE 2 HERE**

11

238

2.2 Calculation of molecular fingerprints

Molecular fingerprints and Tanimoto similarities were calculated using the freely available KNIME software (version 3.3.0). A KNIME workflow (http://dx.doi.org/10.5281/zenodo.1401196) was developed that applied the CDK Fingerprints node to calculate 2D fingerprints and then to calculate different Tanimoto similarities, in terms of these fingerprints, between the molecules in a dataset provided as an SDF file. Tanimoto similarities (Tc) in terms of these bit-vector fingerprints were calculated as per equation (1). The CDK fingerprints calculated were the CDK Standard, CDK Extended, CDK PubChem, CDK FCFP6, CDK ECFP4 and the CDK MACCS fingerprints.

248

2.3 Analysis of Tanimoto coefficients.

The performance of the six different fingerprints to calculate Tc was analysed via the visualisation of the similarity matrices. This was performed by adding the following conditional formatting rules to cells within a Microsoft Excel spreadsheet: green (values between 0.75 and 1), yellow (values between 0.5 and 0.749), orange (values between 0.3 and 0.499) and red (values between 0 and 0.299). Whilst arbitrary, these conditions led to the colour green representing "highly similar" chemicals and red representing "highly dissimilar" chemicals. The ranges of Tc scores were subsequently calculated to determine if knowledge could be gained about which fingerprint works best for the different datasets.

258

3. RESULTS

260   The KNIME workflow produced a CSV file which contained calculated Tc values for the input

261   data sets. The Tc data matrices for the datasets are provided in the supplementary information.

262   Figures (2-6) show the visualisation of the calculated Tc similarity matrices for five different

263   datasets (perfluorinated acids, alkylphenols, saturated alcohols, unsaturated alcohols and the

264   non-polar narcotic datasets), full details of which can been found within the supplementary

265   information along with the matrices for the LLNA skin sensitisation dataset. (The size of the

266   LLNA dataset meant that it was not possible to produce an informative image of the similarity

267   matrices.) In each of these figures, the Tc scores for the same dataset using the six different

268   fingerprints are shown, where **A** was calculated using CDK Standard fingerprints, **B** was

269   calculated using CDK MACCS fingerprints, **C** was calculated using CDK Extended

270   fingerprints, **D** was calculated using CDK PubChem fingerprints, **E** was calculated using CDK

271   FCFP6 fingerprints and **F** was calculated using CDK ECFP4 fingerprints. Each figure shows

272   pairwise Tc values for all compounds in the dataset, with the similarity between compound i

273   and j being shown in the matrix element of row i and column j of the matrix, such that the Tc

274   values for the same compound compared to itself (Tc=1.0) lie along the diagonal elements.

275   N.B. (1) Each row (column) in these images is labelled by the name of the chemical for which

276   colour coded similarity values are reported within that row (column). (2) These images are

277   designed to illustrate the variation in pairwise similarity for the same pairs of compounds using

278   different fingerprints in terms of the corresponding colour patterns. The size of some datasets

279   necessarily makes it hard to read the individual pairwise similarity values from these images.

280   Hence, all pairwise similarity values are provided in an Excel workbook in the Supporting

281   Information. In addition, Tables 3 – 5 show the range of Tanimoto similarity values that can

282   be obtained for the same pairwise comparisons, between compounds in selected datasets, using

283   the different fingerprints.

284   **FIGURES 2-6 HERE**

285   **TABLES 3-5 HERE**

286

287   **4. DISCUSSION**

288   Chemical similarity is, in theory, a beguiling concept allowing for the identification of similar

289   molecules to those with existing information, whether it be biological activity (such as

290   pharmacological or toxicological effects), biokinetics, environmental fate or physico-chemical

291   properties. The science of molecular similarity is founded in drug discovery, where the aim

292   was to identify similar molecules to a known active compound. It mostly utilises easily

293   calculable parameters (descriptors), or fingerprint representations, of molecular structure. The

294   application of molecular similarity is typically based around the Tanimoto coefficient

295   computed from bit-vector fingerprints, as per the current work. As such, there has been a strong

296   interest in this approach in drug discovery for many years and there has been a recent growth

297   of interest in the field of toxicology to enable data gap filling. With regard to toxicity prediction,

298   the focus of the application of molecular similarity has shifted from being intended to identify

299   molecules highly similar to a known active (assuming a receptor mediated pharmacological

300   effect) to multiple uses ranging from searching for any "similar" molecules to a target query

301   with unknown activity, to serving as the input to grouping and/or read-across approaches (Gini

302   et al., 2014; Luechtefeld et al., 2016a-d; 2018). As use of these approaches grows, it is clear

303   that issues may arise with analogues being identified of little relevance, or important analogues

304   not being identified as the similarity measures are not appropriate. The purpose of this study,

305   therefore, was to assess the use of some commonly applied measures of similarity to investigate

306   their use and provide a means of making recommendations for their use for techniques such as

307   read-across, with a focus on read-across predictions made using pairwise similarity calculations

308   to one, or a few, chemical(s), rather than, say, supervised machine learning approaches using

14

309   large quantities of data. To this end, six datasets were analysed which have previously been

310   subject to some form of read-across or QSAR approaches. All have well defined endpoints

311   with varying levels of confidence in the mechanistic rationale.

312   A number of different molecular fingerprints were calculated to determine the advantages or

313   disadvantages of a single method. The similarity matrices in Figures 2-6 clearly demonstrate a

314   difference in Tc scores calculated for the same dataset when using different fingerprints. Closer

315   examination of the perfluorinated acids dataset (Figure 2, dataset 3 from Table 2) indicates a

316   concordance in the fingerprints with regard to in their Tc values as all data matrices are green

317   (values of between 0.75 and 1), showing chemicals are "highly similar". For this data set, the

318   Tc similarity matrices showed good concordance regardless of which fingerprint was chosen

319   i.e. the Tc based assessment of all chemicals as highly similar is in keeping with the assessment

320   which would be made by toxicological experts - since this dataset comprises a homologous

321   series, i.e. the same functional group with varying chain length, expected to act via a common

322   mechanism. As would be expected, variations in Tc scores were as a result of differences in

323   carbon chain length. Those chemicals with C6-C8 gave similarity scores of 1 when compared

324   with each other, those chemicals with C10-C12 gave similarity scores of 1 when compared

325   with each other and the chemical with C9 tended to only show a similarity score of 1 when

326   compared against itself (for CDK standard, CDK Extended fingerprints) or those with C10-

327   C12 (for the other fingerprints). Naturally, all fingerprints gave a Tc value of one for

328   comparisons of the same compound to itself. This trend was similar for all fingerprints applied

329   to this dataset. Thus, fingerprint similarity, in terms of Tc, is a reasonable measure when

330   applied to homologous, or highly similar, series of chemicals, regardless of the fingerprint

331   chosen With regard to read-across, this would indicate that it may be appropriate for "fine-

332   tuning" a read-across within such a preselected series of chemicals – the process sometimes

333   referred to as sub-categorisation.

334   Analysis of datasets with greater structural variability (cf. Figures 3 - 6) indicates a much higher

335   variability in the calculated Tc values depending on which fingerprint was chosen, with limited

336   concordance between them. For example, compare the Tc results for the alkylphenol dataset

337   calculated with CDK FCFP6 against those calculated using the CDK PubChem fingerprints.

338   For two chemicals, 3-methyl-6-n-butylphenol and 2,6-di-tert-butylphenol, CDK FCFP6

339   fingerprints gave a Tc score of 0.26, whereas CDK PubChem fingerprints gave a Tc score of

340   0.88. For both the alkylphenols (Figure 3) and saturated alcohols (Figure 4) datasets, the Tc

341   value computed from the CDK Standard, CDK MACCS, CDK Extended and, for Figure 4,

342   CDK PubChem fingerprints showed some concordance, with a similar pattern of colours

343   denoting the degree of similarity as indicated by the Tc values. However, for both these datasets

344   the calculated Tc values for CDK FCFP6 and the CDK ECFP4 fingerprints were significantly

345   different to the Tc values from the other four fingerprints, with the CDK ECFP4 giving many

346   values that would suggest "highly dissimilar" chemicals, which is not the case for these datasets

347   (based upon expert judgement). Similar discrepancies between fingerprints were seen for the

348   non-polar narcosis dataset (Figure 6). The reasons for such discrepancies undoubtedly reflect

349   the method of fingerprint calculation having an enormous impact on the identification of

350   analogues from large structurally heterogeneous datasets. It may even be an indicator for

351   consideration of composite Tc scores to capitalise on the different information contained.

352   However, that would not address the possibility that toxicologically irrelevant structural

353   variation is being reflected in these similarity values and that relevant structural variation may

354   not be being appropriately captured, even when the information from all fingerprints was

355   combined. Overall, care must be applied in using Tc values for structurally heterogeneous

356   datasets. To make optimal use of Tc values, the user should arguably decide carefully, and

357   rationally, on which fingerprint to use, requiring the user to first give some thought to the

358   fingerprints and mechanism of the endpoint to be read across.

359   For the unsaturated alcohols dataset (Figure 5), all the calculated Tc similarity matrices were

360   noticeably different for each of the six fingerprints used. This dataset consist of chemicals

361   which are, on the face of it, structurally similar but with subtle changes and differences not

362   only in chain length but also the position of the hydroxyl group, (primary or secondary alcohol),

363   branching, and position (internal or external) of the double bond. The positioning of the alcohol

364   group and double bond, as well as branching, will impact of toxicity (Schultz et al., 2017),

365   however none of the Tc values assisted in identifying rational, mechanistically similar

366   analogues across the group. Therefore, subtle, mechanistically relevant changes in molecular

367   structure, such as branching and positional effects may not be captured by any of the

368   fingerprints considered here. Moreover, these most relevant changes will be treated as equally

369   important to whether irrelevant molecular substructures are shared or not between two

370   molecules.

371   Using molecular similarity to assist in toxicity prediction is unlikely to be perfect. There are

372   many examples of highly similar chemicals, in terms of Tc value, having very different toxicity

373   profiles. For example, Table 5 lists four pairs of compounds, selected from the LLNA skin

374   sensitisation dataset, showing potential issues with activity cliffs, despite high Tc values from

375   some fingerprints. Comparison of 1,4-dihydroxyquinone, a strong skin sensitiser, with

376   resorcinol (1,3-dihydroxyquinone), a non-sensitiser, indicates both chemicals being highly

377   similar in structure with the only difference being the position of the hydroxyl groups on the

378   phenol ring (Table 5). The position of the hydroxyl groups in 1,4-dihydroxyquinone enables

379   this chemical to readily form benzoquinone, a reactive metabolite, whereas resorcinol does not

380   form this metabolite, leading to the difference in toxicity seen in regards to skin sensitisation

381   (Bajot et al., 2011, Enoch et al., 2011). However, the Tc scores for most fingerprints in Table 5

382   indicate high similarity, which could lead to false assumptions with regard to grouping and

383   read-across, unless the mechanism of action is known. The wide range of Tc scores calculated

384 also shows the variability of the Tc scores dependent upon the choice of fingerprint. This

385 emphasises the importance of choosing the most appropriate fingerprint, if any, for similarity

386 calculations. In the second comparison 3-phenylenediamine, a strong skin sensitiser, is

387 compared against aniline, a weak skin sensitiser. These chemicals are highly similar in structure,

388 with the main difference being the presence of an extra amine group (Table 5). It has been

389 demonstrated that the presence of the 2 amine groups in 3-phenylenediamine makes this

390 chemical more reactive and leads to its ability to induce strong skin sensitisation (Bajot et al.,

391 2011, Enoch et al., 2011). The Tc scores for this comparison again show variability dependent

392 upon fingerprint choice, with the majority of fingerprints giving a highly Tc score that could

393 be interpreted as indicating these chemicals should have highly similar sensitizing activity.

394 Clearly, this would be an incorrect conclusion.

395 The final two comparisons compare 3,4-dihydrocoumarin, a moderate skin sensitiser, against

396 coumarin and 6-methylcoumarin which are both non-sensitisers (Table 5). These chemicals are

397 all structurally similar with the main difference being the presence of a methyl group and the

398 presence of a double bond (Table 5). The presence of a double bond in the second ring of

399 coumarin causes it to be readily metabolised via Michael addition, into a non-sensitising

400 metabolite (Table 5). The absence of the double bond makes 3,4-dihydrocoumarin more

401 reactive, which accounts for its moderate skin sensitisation when compared to the other two

402 chemicals. The Tc scores calculated for these two comparisons again show variability

403 dependent on fingerprint choice (Table 5). Two of the six fingerprints (CDK MACCS and CDK

404 PubChem) resulted in high Tc scores; this would suggest these chemicals exhibit similar

405 endpoint values, which would be invalid with regards to skin sensitisation.

406 One means of addressing the problems with fingerprint based Tc values calculated for non-

407 homologous datasets, for which subtle changes in molecular structure may lead to significant

18

408   changes in toxicity for certain endpoints, would be to investigate similarity values calculated

409   using a limited number of mechanistically relevant descriptors chosen based on expert

410   judgement. For example, in the case of skin sensitization, the electrophilicity index could be

411   used (Enoch et al., 2008). Similarities might be computed based upon the more general

412   expression for the Tanimoto coefficient, for continuous variables (Willett et al., 1998),

413   following normalisation of different descriptors to the same scale. However, even under this

414   scenario, it is possible that grouping of the chemicals, to ensure that they acted via a common

415   MIE, would first be required before similarity coefficients could be computed for read-across

416   (Enoch et al., 2008).

417   The visualisation and practical handling of Tc values should be borne in mind. In this

418   investigation, due to the number of chemicals in the LLNA skin sensitisation (211 chemicals)

419   and the non-polar narcotic (87 chemicals) datasets (Figure 6 and supplementary data), both of

420   which are quite modest in size, visualisation was challenging which makes the analysis of

421   results difficult. This is an issue that needs to be addressed to ensure that Tc similarity matrices

422   can be used to their full potential. One approach could be to recognise the need to form

423   categories from larger datasets before Tc calculation, thus reducing the number of chemicals

424   within each matrix and making visualisation easier. One means of achieving this is that any

425   relevant knowledge of MIEs should be used to pre-categorise the datasets prior to calculating

426   Tc values. For example, Tc values might be computed for chemicals acting via a common MIE,

427   as indicated by a shared structural alert, and for which some other expert based rules reduced

428   mechanistically irrelevant structural variation that would reduce the information conveyed by

429   the Tc values. This is likely to be the case if the chemicals could be assigned to a homologous

430   series acting via a common mechanism, where the structural variation in chain length was

431   known to be biologically relevant.

In addition, in this study, arbitrary values were applied to visualise the data matrices. The range of 0.75 and 1 was chosen to highlight Tc scores green and show "highly similar" chemicals. It must be remembered that issue of which Tc score is the cut off point for "highly similar", assuming a simple approach based upon saying pairs of "highly similar" chemicals would tend to exhibit "highly similar" biological activity, is not well defined. It is clear from this study that it is very difficult to include a universal "cut-off" and a variable approach to similarity levels is preferable. This further assumes that such a simple approach to predicting similar toxicity, based upon any cut-off value using a fingerprint derived similarity calculation, is appropriate. If suitable cut-off values can be identified at all, the exact values will depend on the fingerprint method applied, endpoint analysed and types of chemical and dataset (Enoch et al., 2009, Nelms et al., 2015). Expert judgement is likely to also have a role to play when deciding whether any single pairwise similarity value is biologically significant, taking into account the observed differences in chemical structures, with reference to understanding of how this is likely to be mechanistically related to the toxicology.

Finally, recent work (Luechtefeld et al., 2016d) reported "read-across" predictions of skin sensitisation based upon the most similar chemicals, in terms of Tanimoto similarities computed from PubChem 2D molecular fingerprints, with available skin sensitisation data. Building upon that work, Luechtefeld et al. (2018) proposed approaches to "read-across" predictions of toxicity based upon supervised machine learning which incorporated Tanimoto similarity values, again calculated from PubChem 2D molecular fingerprints, to multiple compounds with experimental toxicity data. (Further work in that latter study also proposed a "data fusion" model, incorporating data for other endpoints, as well as similarity values.) In spite of the limitations of Tanimoto similarity values calculated from molecular fingerprints, which are highlighted above, they reported empirically good results.

456   It may be speculated that these empirically good results (Luechtefeld et al., 2016d, Luechtefeld

457   et al., 2018) could, in part, reflect the nature of the datasets investigated, e.g. those datasets

458   may comprise categories of structurally similar chemicals acting via a similar mechanism, with

459   structural differences within those categories being biologically relevant, for which Tanimoto

460   similarity values based on molecular fingerprints can be expected to work best. For example,

461   31% of the skin sensitisation dataset of Luechtefeld et al. (2016d) was composed of Michael

462   acceptors. However, further analysis is required to determine whether this is, indeed, the case.

463   Moreover, due to the inherent limitations of Tanimoto values of molecular similarities

464   computed from molecular fingerprints and the variation in similarity values which can be

465   obtained with different fingerprints, as highlighted in the current work, it is unlikely that read-

466   across predictions based upon these values using a single fingerprint would be optimal for all

467   relevant scenarios. Thus, for the examples that may be taken from the range of datasets

468   investigated in this study, different types of chemical similarity would be required for effective

469   and defensible analogue selection. Optimal read-across predictions are more likely to be

470   obtained if care is taken to use a similarity measure based upon consideration of the mechanism

471   of action. Indeed, providing a mechanistic rationale for the predictions, rather than just

472   statistical validation, is more likely to lead to acceptance in a regulatory context.

473   In terms of analogue selection, fingerprints may be developed that have a stronger focus on

474   mechanisms of action and thus are more applicable to address toxicological problems e.g.

475   toxicologically relevant structural features such as the ToxPrint chemotypes could be used as

476   a means of developing fingerprints (Richard et al., 2016). The assumption underpinning the

477   improvement that may be assumed in analogue selection and justification is that such

478   fingerprints, if used, would provide better focus on the MIE which is at the heart of mechanistic

479   similarity but which may not be captured by the commonly used methods investigated in this

480 study. It is further acknowledged that the use of a broad fingerprint method based around

481 known toxicologically relevant fragments could assist in situations where the precise MIE may

482 not be known. However, the development of new fingerprints to aid toxicological read-across

483 would most appropriately be carried out on an endpoint specific basis, rather than assuming a

484 single fingerprint could be developed for all endpoints.

485

486 **5. CONCLUSIONS**

487 In conclusion, molecular fingerprint similarity matrices can be used as a means of identifying

488 possible analogues in some contexts. However, on their own, it is difficult to use generic

489 similarity measures computed from generic, purely structurally based, fingerprints to support

490 a read-across hypothesis or justification. This is due to several known limitations of generic

491 similarity measures calculated from these fingerprints, which are highlighted in the current

492 work. They are liable to exhibit activity cliffs (where small changes to the overall molecular

493 structure, resulting in high similarity values, lead to significant changes in biological activity).

494 The fingerprints may not capture the relevant structural variation (depending upon the

495 fingerprint method) and treat mechanistically irrelevant structural variation equally to

496 mechanistically relevant structural variation. Similarity matrices, calculated from different

497 fingerprints, show greater concordance and are better suited to analogue identification for less

498 diverse datasets, especially homologous series. This suggests they could be most appropriate

499 for read-across within a homologous series, acting via a common mechanism, for which the

500 variation in chemical structure is known to be related to biological activity This could avoid

501 the pitfall of fingerprint based similarity measures reflecting biologically irrelevant structural

502 variation. Hence, for a read across setting, users of chemically diverse datasets could benefit

503 from first forming categories when using molecular fingerprint similarity values.

504 Whilst Tanimoto similarity values computed from generic molecular fingerprints have been

505 integrated into recent machine learning predictions of toxicity within diverse datasets with

506 empirically successful results, the limitations of these similarity values, highlighted in our work,

507 mean that other approaches to similarity assessment are preferable for read-across. Ideally,

508 similarity values which reflect biologically relevant information, informed by mechanistic

509 understanding, should be employed. This is especially the case in a regulatory context, where

510 a mechanistic justification is likely to be required. More preferable approaches to similarity

511 assessment could entail the previously outlined approach, i.e. first applying a mechanism based

512 categorisation of the dataset, such that the use of generic similarity values based on molecular

513 fingerprints would only be used to fine tune read-across within a homologous series.

514 More generally, when calculating similarity, the user needs to give careful consideration to the

515 selection of the most appropriate similarity measure to use and, where possible, link this to

516 rational consideration of the mechanism underpinning the endpoint, e.g. in terms of the

517 Molecular Initiating Event (MIE). Following the cautionary examples presented in this work,

518 the following recommendations are made concerning the use of generic similarity coefficients

519 based on molecular fingerprints for read-across predictions of toxicity.

520 - Fingerprint-derived measures of molecular similarity can be a useful means of identifying

521 close structural analogues and may have use in the application of read-across for data gap

522 filling. Such methods may provide a useful visual approach to molecular similarity.

523 - The similarity value is dependent on the type of fingerprint, or, if a more general similarity

524 value is computed, the descriptors and/or properties used for its calculation. The user

525 should acquaint themselves with the different fingerprint methods and their intended

526 purpose. A method tailored to the toxicity endpoint should ideally be applied.

527   -   Of the fingerprint methods considered in this study, there is evidence that Tanimoto

528   similarity values derived from CDK Standard, CDK MACCS, CDK Extended and CDK

529   PubChem fingerprints showed some concordance, for some scenarios, with similarity

530   values for CDK FCFP6 and the CDK ECFP4 providing different information. Further

531   work is required to understand the significance of these findings and at this time no single

532   fingerprint method from those investigated could be considered to be the most optimum.

533   These fingerprints may be appropriate to find "structural" analogues in terms of pure

534   chemistry, but these may not be appropriate for toxicological read-across without

535   interpretation and further mechanistic knowledge.

536   -   Where known, knowledge of the MIE will guide the successful application of molecular

537   similarities for toxicological read-across. Reference to the MIE will improve mechanistic

538   justification of the analogue selection and might be achieved with fingerprints that take

539   account of the structural basis of toxicity for specific endpoints. Fingerprints must be

540   chosen and interpreted such that they avoid pitfalls such as activity cliffs i.e. the selection

541   of close structural analogues, according to the fingerprint derived similarity measure,

542   which have different activity due to the effect of structural change on the MIE.

543   -   Whilst a justifiable means of identifying analogues, the use of the MIE is only appropriate

544   to relevant toxicological endpoints, i.e. where the MIE is known, and identifying the MIE

545   is only one step in the overall read-across process, which may involve the collation of

546   multiple lines of evidence.

547   -   Fingerprint-derived measures of similarity should be used to identify analogues for read-

548   across for large heterogeneous datasets with caution, unless the similarity measures can be

549   shown to clearly relate to biologically relevant structural variation and not to capture

550   biologically irrelevant variation. Where they are known, this justification should be made

551   with reference to relevant mechanism(s) of action, for instance relating to the MIE.

552   However, generic fingerprint similarity measures do not fulfil these criteria, so must be

553   used with caution for large, chemically diverse datasets.

554   -   Arguably, the most suitable use of generic fingerprint-derived similarity measures for read-

555   across within large, chemically diverse datasets is following sub-categorisation. (However,

556   further work is required to determine the extent to which this yields better predictive

557   performance than integrating these similarity measures within machine learning

558   approaches, which have recently been advocated. Moreover, sub-categorisation which

559   removes biologically irrelevant structural variation may result in the fingerprint-derived

560   similarity measures being optimally predictive, yet redundant if read-across is performed

561   by expert examination of the structures within the category.) Sub-categorisation should

562   preferably be performed using a mechanistically based method. If sub-categorisation

563   yields homologous series, acting via a common mechanism, for which all the structural

564   variation is expected to be biologically relevant, generic fingerprint-derived similarity

565   measures could be suitable for fine tuning and confirming analogue identification for read-

566   across.

567   -   However, even within categories of chemicals acting via a common mechanism, the use

568   of alternative similarity measures, based upon mechanistic understanding of the endpoint

569   of interest, should be considered for read-across purposes. For example, similarity

570   coefficients can be computed from mechanistically relevant fingerprints or descriptors.

571   Overall, fingerprint-derived measures of molecular similarity may be a useful method in the *in*

572   *silico* toolbox for data gap filling. However, they are likely to be optimally predictive within a

573   small, mechanistically derived category and, ideally, the specific similarity measure should be

574   appropriate to the chemistry and endpoint considered.

25

575

6. ACKNOWLEDGEMENTS

579

7. REFERENCES

581 Allen, T.E.H., Goodman, J.M., Gutsell, S., Russell, P.J., 2016. A history of the Molecular
582 Initiating Event. Chem. Res. Toxicol. 29, 2060–2070.

583 Alves, V.M., Muratov, E.N., Capuzzi, S.J., Politi, R., Low, Y., Braga, R.C., Zakharov, A.V.,
584 Sedykh, A., Mokshyna, E., Farag, S., Andrade, C.H., Kuz'min, V.E., Fourches, D., Tropsha
585 A., 2016. Alarms about structural alerts. Green Chem. 18, 4348-4360.

586 Alves, V.M., Golbraikh, A., Capuzzi, S.J., Liu, K., Lam, W.I., Korn, D.R., Pozefsky, D.,
587 Andrade, C.H., Muratov, E.N. Tropsha A., 2018. Multi-Descriptor Read Across (MuDRA):
588 A simple and transparent approach for developing accurate Quantitative Structure-Activity
589 Relationship models. J. Chem. Inf. Model. 58, 1214-1223.

590 Bajot, F., Cronin, M.T.D., Roberts, D.W., Schultz, T.W., 2011. Reactivity and aquatic
591 toxicity of aromatic compounds transformable to quinone-type Michael acceptors. SAR
592 QSAR Environ. Res. 22, 51–65.

593 Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for
594 fingerprint-based similarity calculations? J. Cheminf. 7, 20.

595 Ball, N., Cronin, M.T.D., Shen, J., Blackburn, K., Booth, E.D., Bouhifd, M., Donley, E.,

596 Egnash, L., Hastings, C., Juberg, D.R., Kleensang, A., Kleinstreuer, N., Kroese, E.D., Lee,

597 A.C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J.M., Palmer, J., Pamies, D., Penman,

598 M., Richarz, A.-N., Russo, D.,P.. Stuard, S.B., Patlewicz, G., van Ravenzwaay, B., Wu, S.,

599 Zhu, H., Hartung, T., 2016. Toward Good Read-Across Practice (GRAP) guidance. ALTEX,

600 33, 149-166.

601 Bender, A., Glen, R.C., 2004. Molecular similarity: a key technique in molecular informatics.

602 Org. Biomol. Chem. 2, 3204-3218.

603 Berggren, E., Amcoff, P., Benigni, R., Blackburn, K., Carney, E., Cronin, M., Deluyker, H.,

604 Gautier, F., Judson, R.S., Kass, G.E.N., Keller, D., Knight, D., Lilienblum, W., Mahony, C.,

605 Rusyn, I., Schultz, T., Schwarz, M., Schüürmann, G., White, A., Burton,J.,. Lostia, A.M.,

606 Munn, S., Worth, A., 2015. Chemical safety assessment using read-across: How can novel

607 testing methods strengthen evidence base for decision-making? Environ. Health. Perspect.

608 123, 1232-1240.

609 Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., Pujadasa, G.,

610 2015. Molecular fingerprint similarity search in virtual screening. Methods. 71, 58–63.

611 Chesnut, M., Yamada, T., Adams, T., Knight, D., Kleinstreuer, N., Kass, G., Luechtefeld, T.,

612 Hartung, T., Maertens, A., 2018. Regulatory acceptance of read-across. ALTEX. 35, 413-419.

613 Cronin, M.T.D., Madden, J.C., Enoch, S.J., Roberts, D.W., 2013. Chemical Toxicity

614 Prediction: Category Formation and Read-Across. Royal Society of Chemistry, Cambridge

615 UK.

616 Cronin, M.T.D., Richarz, A.-N., 2017. Relationship between Adverse Outcome Pathways and

617 chemistry-cased in silico models to predict toxicity. Appl. *in Vitro* Toxicol. 3, 286-297.

27

618   Cronin, M.T.D., Enoch, S.J., Mellor, C.L., Przybylak, K.R., Richarz, A.-N., Madden, J.C.,

619   2017. *In silico* prediction of organ level toxicity: linking chemistry to adverse effects.

620   Toxicological Res. 33,173-182.

621   Cruz-Monteagudo, M., Medina-Franco, J.L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro,

622   M.N.D.S., Borges, F., 2014. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug

623   Discov. Today. 19, 1069-1080.

624   Dixon, S.L., Koehler, R.T., 1999. The hidden component of size in two-dimensional

625   fragment descriptors: side effects on sampling in bioactive libraries. J. Med. Chem. 42, 2887–

626   2900.

627   Eckert, H., Bajorath, J. 2007. Molecular similarity analysis in virtual screening: foundations,

628   limitations and novel approaches. Drug Discov. Today. 12, 225−233.

629   Ellison, C.M., Cronin, M.T.D., Madden, J.C., Schultz, T.W., 2008. Definition of the

630   structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. SAR

631   QSAR Environ. Res. 19, 751-783.

632   Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. Quantitative and

633   mechanistic read across for predicting the skin sensitization potential of alkenes acting via

634   Michael addition. Chem. Res. Toxicol. 21, 513–520.

635   Enoch, S.J, Cronin, M.T.D., Madden, J.C., Hewitt, M., 2009. Formation of structural

636   categories to allow for read-across for teratogenicity. QSAR Combin. Sci. 28, 696-708.

637   Enoch, S.J, Ellison, C.M., Schultz, T.W., Cronin, M.T.D., 2011. A review of the electrophilic

638   reaction chemistry involved in covalent protein binding relevant to toxicity. Crit. Rev.

639   Toxicol. 41, 783-802.

640   Flower, D.R., 1998. On the properties of Bit string-based measures of chemical similarity. J.

641   Chem. Inf. Comput. Sci. 38, 379–386.

642   Gerberick, G.F., Ryan, C.A., Kern, P.S., Schlatter, H., Dearman, R.J., Kimber, I., Patlewicz,

643   G.Y., Basketter, D.A., 2005. Compilation of historical Local Lymph Node data for evaluation

644   of skin sensitization alternative methods. Dermatitis. 16, 157-202.

645   Gini, G., Franchi, A.M., Manganaro, A., Golbamaki, A., Benfenati, E., 2014. ToxRead: A tool

646   to assist in read across and its use to assess mutagenicity of chemicals. SAR QSAR Environ.

647   Res. 25, 999-1011.

648   Godden, J.W., Xue, L., Bajorath, J., 2000. Combinatorial preferences affect molecular

649   similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J. Chem.

650   Inf. Comput. Sci. 40, 163–166.

651   Guha, R., van Drie, J.H., 2008. Structure-activity landscape index: Identifying and

652   quantifying activity cliffs. J. Chem. Inf. Model. 48, 646−658.

653   Hartung, T., 2016. Making big sense from big data in toxicology by read-across. ALTEX. 33,

654   83-93.

655   Holliday, J.D., et al., 2002. Grouping of coefficients for the calculation of inter-molecular

656   similarity and dissimilarity using 2D fragment Bit-strings. Comb. Chem. High Throughput

657   Screen. 5, 155–166.

658   Holliday, J.D., Hu, C.-Y., Willett, P., 2003. Analysis and display of the size dependence of

659   chemical similarity coefficients. J. Chem. Inf. Comput. Sci. 43, 819–828.

660   Lajiness, M.S., 1997. Dissimilarity-based compound selection techniques. Perspect. Drug

661   Discov. Des. 7/8, 65–84.

662   Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016a. Global

663   analysis of publicly available safety data for 9,801 substances registered under REACH from

664   2008-2014. ALTEX. 33, 95–109.

665   Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016b. Analysis

666   of public oral toxicity data from REACH registrations 2008-2014. ALTEX. 33, 111–122.

667   Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016c. Analysis

668   of Draize eye irritation testing and its prediction by mining publicly available 2008-2014

669   REACH data. ALTEX. 33, 123–134.

670   Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016d. Analysis

671   of publically available skin sensitization data from REACH registrations 2008-2014. ALTEX.

672   33, 135–148.

673   Luechtefeld, T., March, D., Rowlands, C., Hartung, T., 2018. Machine learning of toxicological

674   big data enables read-across structure activity relationships (RASAR) outperforming animal

675   test reproducibility. Toxicol Sci. 165, 198-212.

676   Maggiora, M. Vogt, M., Stumpfe, D., Bajorath, J., 2014. Molecular similarity in medicinal

677   chemistry. J. Med. Chem. 57, 3186–3204.

678   Martin, Y.C., Kofron, J.L., Traphagen, L.M.., 2002. Do structurally similar compounds have

679   similar biological activity? J. Med. Chem. 45, 4350−4358.

680   Maunz, A., Helma, C., 2008. Prediction of chemical toxicity with local support vector

681   regression and activity-specific kernels. SAR QSAR Environ. Res. 19, 413-431.

682   Mellor, C.L., Schultz, T.W., Przybylak, K.R., Richarz, A.-N., Bradbury, S.P., Cronin, M.T.Da.,

683   2017. Read-across for rat oral gavage repeated-dose toxicity for short-chain mono-

684   alkylphenols: A case study. Comput. Toxicol. 2, 1-11.

685   Nelms, M.D, Ates, G., Madden, J.C., Vinken, M., Cronin, M.T.D., Rogiers, V., Enoch, S.J..,

686   2015. Proposal of an *in silico* profiler for categorisation of repeat dose toxicity data of hair

687   dyes. Arch. Toxicol. 89, 733-741.

688   OECD (Organisation for Economic Cooperation and Development), 2014. Guidance on

689   Grouping of Chemicals, 2nd Edition, Series on Testing and Assessment No. 194. OECD,

690   Paris, France.

691   Patlewicz, G., Jeliazkova, N., Gallegos-Saliner A., Worth, A.P., 2008. Toxmatch - a new

692   software tool to aid in the development and evaluation of chemically similar groups. SAR

693   QSAR Environ. Res. 19, 397-412.

694   Patlewicz, G., Ball, N., Boogaard, P.J., Becker, R.A., Hubesch, B., 2015. Building scientific

695   confidence in the development and evaluation of read-across. Reg. Toxicol. Pharmacol. 72,

696   117-133.

697   Patlewicz, G., Helman, G., Pradeep, P., Shah, I., 2017. Navigating through the minefield of

698   read-across tools: A review of in silico tools for grouping. Comput. Toxicol. 3, 1-18

699   Patlewicz, G., Cronin, M.T.D., Helman, G., Lambert, J.C., Lizarraga, L.E., Shah I., 2018.

700   Navigating through the minefield of read-across frameworks: A commentary perspective.

701   Comput. Toxicol. 6, 39-54.

702 Przybylak, K.R., Schultz, T.W., Richarz, A.-N., Mellor, C.L., Escher, S.E., Cronin, M.T.Da.,

703 2017. Read-across of 90-day rat oral repeated-dose toxicity: A case study for selected β-olefinic

704 alcohols. Comput. Toxicol. 1, 22-32.

705 Reisen, F., Zhang, X., Gabriel, D., Selzer, P., 2013. Benchmarking of multivariate similarity

706 measures for high-content screening fingerprints in phenotypic drug discovery. J. Biomol.

707 Screen. 18, 1284–1297.

708 Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I.,

709 Yang, C., Rathman, J., Martin, M.T., Wambaugh, J.F., Knudsen, T.B., Kancherla, J.,

710 Mansouri, K., Patlewicz, G., Williams, A.J., Little, S.B., Crofton, K.M., Thomas, R.S., 2016.

711 ToxCast chemical landscape: paving the road to 21st Century Toxicology. Chem. Res.

712 Toxicol. 29, 1225-1251.

713 Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50,

714 742–754.

715 Rosenbaum, L., Hinselmann, G., Jahn, A., Zell A., 2011. Interpreting linear support vector

716 machine models with heat map molecule coloring. J. Cheminform. 3, 11.

717 Salim, N., Holliday, J., Willett, P., 2003. Combination of fingerprint-based similarity

718 coefficients using data fusion. J. Chem. Inf. Comput. Sci. 43, 435–442.

719 Schultz, T.W., Cronin M.T.D., 2017. Lessons learned from read-across case studies for

720 repeated-dose toxicity. Reg. Toxicol. Pharmacol. 88, 185-191.

721 Schultz, T.W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D.J., Mahony, C.,

722 Schwarz, M., White, A., Cronin, M.T.D., 2015. A strategy for structuring and reporting a

723 read-across prediction of toxicity, Reg. Toxicol. Pharmacol. 72, 586-601.

724 Schultz, T.W., Przybylak, K.R., Richarz, A.-N., Mellor, C.L., Escher, S.E., Bradbury, S.P.,

725 Cronin, M.T.D., 2017. Read-across of 90-day rat oral repeated-dose toxicity: A case study for

726 selected n-alkanols. Comput. Toxicol. 2, 12-19.

727 Schultz, T.W., Richarz, A.-N., Cronin, M.T.D., 2018. Assessing uncertainty in read-across:

728 knowledge gained from case studies. Comput. Toxicol. submitted

729 Shah, I., Liu, J., Judson, R.S., Thomas, R.S., Patlewicz, G., 2016. Systematically evaluating

730 read-across prediction and performance using a local validity approach characterized by

731 chemical structure and bioactivity information. Reg. Toxicol. Pharmacol. 79, 12-24.

732 Spielmann, H., Sauer, U.G., Mekenyan, O., 2016. A critical evaluation of the 2011 ECHA

733 reports on compliance with the REACH and CLP regulations and on the use of alternatives to

734 testing on animals for compliance with the REACH regulation. ATLA – Altern. Lab. Anim,

735 39, 481-493.

736 Stumpfe, D., Bajorath, J., 2011. Similarity searching. Wiley Interdiscip. Rev.: Comput. Mol.

737 Sci. 1, 260−282.

738 Teubner, W., Landsiedel, R., 2015. Read-across for hazard assessment: the ugly duckling is

739 growing up. ATLA – Altern. Lab. Anim, 43, 67-71.

740 Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P.., 2012.

741 Similarity coefficients for binary chemoinformatics data: overview and extended comparison

742 using simulated and real data sets. J. Chem. Inf. Model. 52, 2884–2901.

743 Whittle, M., Willett, P., Klaffke, W., van Noort, P., 2003. Evaluation of similarity measures

744 for searching the Dictionary of Natural Products Database. J. Chem. Inf. Comput. Sci. 43,

745 449-457.

746 Whittle, M., Gillet, V.J., Willett, P., Alex, A., Loesel, J., 2004. Enhancing the effectiveness

747 of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients J.

748 Chem. Inf. Comput. Sci. 44, 1840–1848.

749 Willett, P., 2013. Combination of similarity rankings using data fusion. J. Chem. Inf. Model.

750 53, 1–10.

751 Willett, P., Barnard, J.M., Downs, G.M., 1998. Chemical similarity searching. J. Chem. Inf.

752 Comput. Sci. 38, 983-996.

753

754 Table 1. Definitions of terms using in this investigation.

Term	Definitions used for this study
Analogue (for read-across)	A similar compound, with measured endpoint data, to that for which read-across predictions are required for the endpoint in question. So-called "data rich" analogues are often most useful, as relevant physicochemical and biological data, in addition to endpoint data, may complement calculated measures of structural similarity.
Fingerprint-derived molecular similarity	Molecular similarity between two molecules calculated from molecular fingerprints. In this study, all similarity values were calculated in terms of the widely used Tanimoto coefficient (defined below).
Grouping	The process of assigning chemicals to a category of related compounds. This is usually based upon the hypothesis that the chemicals assigned to the category exhibit common properties with regard to the endpoint of interest, or exhibit simple trends in the endpoint related to structural variation. Similarity calculations within that category may then be used to make read-across predictions.
Molecular fingerprint	Typically, a binary vector with bits (0 or 1) calculated from the presence (1) or absence (0) of structural features. Six different types of fingerprints were investigated in this study.

Molecular similarity	The similarity, or degree of overlap, between two or more molecules. Similarity is defined in terms of a set of features, properties or calculated descriptors. In this investigation, molecular similarity was quantified by the Tanimoto coefficients calculated from the molecular fingerprints.
Tanimoto coefficient	A value calculated to represent the similarity between two objects represented as two vectors. For the purposes of this study, the objects were molecules and the vectors were the binary vectors corresponding to one out of many possible molecular fingerprints. An equation for calculating this coefficient, for binary vectors, is provided below.
Read-across	The process of interpolating or extrapolating a value of some endpoint of interest between similar compounds. This investigation focussed on read-across for various toxicological endpoints. In the context of the current work, the focus is upon read-across predictions made using pairwise comparison to one, or a few, suitably "similar" chemicals.

755

Table 2: The datasets investigated in this study with a description of the toxicological effect and mechanistic hypothesis for the factors which would need to be captured by a similarity approach employed for read-across.

Data Set No.	Effect / Toxicity / MIE if known	Number of Chemicals	Types of Chemicals	Mechanistic hypothesis for similarity for read-across	Reference
1	40 hour inhibition of growth to the ciliated protozoan *Tetrahymena pyriformis*. All chemicals are assumed to act by non-polar narcosis, although the exact MIE is unknown is is assumed to induce perturbation of cellular membranes.	87	Unreactive e.g. saturated alcohols and ketones	Toxicity is assumed to be a function of distribution to the active site (e.g. accumulation within membranes). Therefore, compounds fitting the non-polar narcosis domain should exhibit similar toxicity, if they have similar properties relating to distribution.	Ellison et al., 2008
2	Local LLNA skin sensitisation dataset of chemicals that have both chemical and biological diversity. The MIE is the (electrophilic) interaction of the toxicant with the immunoprotein	211	In terms of chemical diversity, the database contains aldehydes, ketones, aromatic amines, quinones, and acrylates, as well as compounds that have different reactivity mechanisms.	Compounds are required to be protein reactive, or be metabolised to a reactive form, to elicit skin sensitisation. Hence, molecules should be similar in a manner which reflects these requirements in order to cause similar skin sensitisation.	Gerberick et al., 2005

3	A category of perfluorinated acids on which read-across has been performed for repeat dose toxicity data. The MIE following repeated dose exposure is assumed to be binding to the peroxisome proliferator–activated receptor and other nuclear receptors.	7	A congeneric series of perfluorinated acids with a carbon chain length of between C6 – C12	PFAAs are chemically unreactive and assumed to be active by a similar mechanism (binding to nuclear receptor(s)). Hence, molecules should be similar in a manner which is related the degree of nuclear receptor binding, in order to exhibit similar toxicity.	Berggren et al., 2015
4	Alkanols (saturated aliphatic alcohols). This chemical category represents analogues with low general or no toxicity (i.e., toxicants which are non-reactive and exhibit no specific mode of action). There is no specific MIE other than that associated with perturbation of cellular membranes in the same manner as non-polar narcosis.	19	n-Alkanols within the range C5-C12	Alkanols form a homologous series of compounds associated with low toxicity..	Berggren et al., 2015; Schultz et al 2017
5	Unsaturated aliphatic alcohols, exhibiting hepatotoxicity (toxicity to the liver). The MIE assumes metabolic transformationin the liver, to reactive electrophilic toxicants which react with biological macromolecules	26	Small (C3 to C6) primary and secondary β-olefinic alcohols.	Compounds are assumed to be metabolised to a common reactive metabolite which is responsible for their toxicity to the liver. Hence, similarity in terms of structural factors which affect the degree of	Berggren et al., 2015; Przybylak et al 2017

	in a mechanistically similar manner to acrolein			metabolism or the reactivity of the metabolite is required for toxicological similarity.	
6	Alkyl phenols read-across case study for repeated dose toxicity. A precise MIE is unknown, however they are associated with perturbation of cellular membranes in the same manner as polar narcosis.	20	Alkyl-substituted phenols	These compounds are non-reactive and exhibit an unspecific, reversible polar narcosis mode of toxic action. Toxicity is reliant on their distribution to the site of action. Hence, similarity with respect to factors which affect distribution will be required for biological similarity.	Berggren et al., 2015; Mellor et al 2017

Table 3: Shows the range of the Tc scores calculated when utilising the different fingerprints for the perfluorinated acids dataset (dataset 3).

	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00-1	0.87-1	0.83-1	0.83-1	0.83-1	0.83-1	0.83-1
PFHpA		1.00-1	0.92-1	0.91-1	0.91-1	0.91-1	0.91-1
PFOA			1.00-1	0.98-1	0.98-1	0.98-1	0.98-1
PFNA				1.00-1	1.00-1	1.00-1	1.00-1
PFDA					1.00-1	1.00-1	1.00-1
PFUA						1.00-1	1.00-1
PFDoA							1.00-1

Abbreviations relate to the following : Perfluorohexanoic acid (PFHxA), Perfluoroheptanoic acid (PFHpA), Perfluorooctanoic acid (PFOA), Perfluorononanoic acid (PFNA), Perfluorodecanoic acid (PFDA), Perfluoroundecanoic acid (PFUA) and Perfluorododecanic acid (PFDoA).
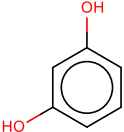
Table 4: Shows the range of the Tc scores calculated when utilising the different fingerprints for the alkylphenols dataset (dataset 6).

	2-tert.Butyl-5-methylphenol	2-tert-Butyl-4-methylphenol	2-tert-Butylphenol	2,6-di-tert-Butylphenol	2-tert-Amylphenol	2,4-di-tert-Amylphenol	2-sec-Butylphenol	2-n-Butylphenol	2-n-Pentylphenol	2-Isopropyl-5-methylphenol (thymol)	2-Methyl-5-isopropylphenol (carvacrol)	3-Methyl-6-n-butylphenol	2-Ethyl-5-methylphenol	2-Isopropylphenol	2,4-Diisopropylphenol	2,5-Dimethylphenol	2,6-Dimethylphenol	3-tert-butylphenol	4-tert-Butylphenol	4-tert-Buty-2-methylphenol
2-tert.Butyl-5-methylphenol	1.00-1	0.54-1	0.50-1	0.41-1	0.31-0.95	0.31-0.91	0.23-0.9	0.20-0.89	0.20-0.91	0.46-1	0.31-1	0.42-0.97	0.45-0.96	0.26-0.95	0.27-1	0.52-0.93	0.25-0.86	0.37-1	0.32-1	0.40-1
2-tert-Butyl-4-methylphenol	0.54-1	1.00-1	0.50-1	0.41-1	0.35-0.96	0.39-0.98	0.23-0.91	0.20-0.9	0.20-0.92	0.39-1	0.31-1	0.33-0.88	0.39-0.86	0.26-0.95	0.31-1	0.39-0.84	0.25-0.91	0.32-1	0.32-1	0.45-1
2-tert-Butylphenol	0.50-1	0.50-1	1.00-1	0.54-1	0.63-0.99	0.34-0.92	0.33-0.97	0.34-0.95	0.34-0.95	0.23-1	0.22-1	0.21-0.9	0.22-0.91	0.38-0.97	0.22-1	0.25-0.89	0.28-0.92	0.36-1	0.36-1	0.34-1
2,6-di-tert-Butylphenol	0.41-1	0.41-1	0.54-1	1.00-1	0.41-0.97	0.27-0.95	0.22-0.92	0.27-0.91	0.27-0.93	0.19-1	0.19-1	0.21-0.88	0.19-0.87	0.25-0.95	0.19-1	0.21-0.85	0.41-0.94	0.42-1	0.38-1	0.31-1
2-tert-Amylphenol	0.31-0.95	0.35-0.96	0.63-0.99	0.41-0.97	1.00-1	0.58-1	0.39-0.95	0.40-0.94	0.40-0.97	0.24-0.9	0.20-0.9	0.26-0.91	0.27-0.9	0.39-0.95	0.20-0.91	0.23-0.88	0.25-0.91	0.28-0.93	0.28-0.92	0.27-0.95

2,4-di-tert-Amylphenol	0.31-0.91	0.39-0.98	0.34-0.92	0.27-0.95	0.58-1	1.00-1	0.24-0.91	0.24-0.88	0.24-0.9	0.25-0.87	0.21-0.87	0.26-0.88	0.27-0.87	0.23-0.89	0.24-0.96	0.24-0.85	0.18-0.89	0.29-0.89	0.32-0.92	0.39-0.99
2-sec-Butylphenol	0.23-0.9	0.23-0.91	0.33-0.97	0.22-0.92	0.39-0.95	0.24-0.91	1.00-1	0.39-0.96	0.39-0.97	0.35-0.94	0.26-0.94	0.29-0.91	0.30-0.92	0.67-1	0.34-0.93	0.26-0.9	0.24-0.93	0.20-0.91	0.19-0.9	0.19-0.9
2-n-Butylphenol	0.20-0.89	0.20-0.9	0.34-0.95	0.27-0.91	0.40-0.94	0.24-0.88	0.39-0.96	1.00-1	0.86-0.98	0.24-0.91	0.20-0.91	0.57-0.96	0.35-0.93	0.39-0.96	0.20-0.9	0.23-0.93	0.25-0.94	0.21-0.9	0.19-0.89	0.20-0.89
2-n-Pentylphenol	0.20-0.91	0.20-0.92	0.34-0.95	0.27-0.93	0.40-0.97	0.24-0.9	0.39-0.97	0.86-0.98	1.00-1	0.24-0.91	0.20-0.91	0.52-0.94	0.35-0.93	0.39-0.97	0.20-0.92	0.23-0.91	0.25-0.94	0.21-0.9	0.19-0.89	0.20-0.91
2-Isopropyl-5-methylphenol (thymol)	0.46-1	0.39-1	0.23-1	0.19-1	0.24-0.9	0.25-0.87	0.35-0.94	0.24-0.91	0.24-0.91	1.00-1	0.41-1	0.48-0.97	0.52-0.99	0.52-0.95	0.43-1	0.54-0.96	0.26-0.88	0.21-1	0.20-1	0.28-1
2-Methyl-5-isopropylphenol (carvacrol)	0.31-1	0.31-1	0.22-1	0.19-1	0.20-0.9	0.21-0.87	0.26-0.94	0.20-0.91	0.20-0.91	0.41-1	1.00-1	0.29-0.97	0.31-0.98	0.34-0.95	0.43-1	0.58-0.96	0.30-0.88	0.21-1	0.19-1	0.31-1
3-Methyl-6-n-butylphenol	0.42-0.97	0.33-0.88	0.21-0.9	0.21-0.88	0.26-0.91	0.26-0.88	0.29-0.91	0.57-0.96	0.52-0.94	0.48-0.97	0.29-0.97	1.00-1	0.68-0.99	0.28-0.91	0.26-0.89	0.48-0.97	0.23-0.89	0.19-0.91	0.18-0.86	0.26-0.89
2-Ethyl-5-methylphenol	0.45-0.96	0.39-0.86	0.22-0.91	0.19-0.87	0.27-0.9	0.27-0.87	0.30-0.92	0.35-0.93	0.35-0.93	0.52-0.99	0.31-0.98	0.68-0.99	1.00-1	0.30-0.92	0.27-0.88	0.52-0.98	0.25-0.9	0.21-0.92	0.19-0.87	0.27-0.88
2-Isopropylphenol	0.26-0.95	0.26-0.95	0.38-0.97	0.25-0.95	0.39-0.95	0.23-0.89	0.67-1	0.39-0.96	0.39-0.97	0.52-0.95	0.34-0.95	0.28-0.91	0.30-0.92	1.00-1	0.50-0.95	0.30-0.9	0.28-0.93	0.23-0.95	0.21-0.95	0.22-0.95
2,4-Diisopropylphenol	0.27-1	0.31-1	0.22-1	0.19-1	0.20-0.91	0.24-0.96	0.34-0.93	0.20-0.9	0.20-0.92	0.43-1	0.43-1	0.26-0.89	0.27-0.88	0.50-0.95	1.00-1	0.30-0.86	0.21-0.91	0.21-1	0.19-1	0.27-1

2,5-Dimethylphenol	0.52-0.93	0.39-0.84	0.25-0.89	0.21-0.85	0.23-0.88	0.24-0.85	0.26-0.9	0.23-0.93	0.23-0.91	0.54-0.96	0.58-0.96	0.48-0.97	0.52-0.98	0.30-0.9	0.30-0.86	1.00-1	0.35-1	0.23-0.9	0.22-0.85	0.36-0.85
2,6-Dimethylphenol	0.25-0.86	0.25-0.91	0.28-0.92	0.41-0.94	0.25-0.91	0.18-0.89	0.24-0.93	0.25-0.94	0.25-0.94	0.26-0.88	0.30-0.88	0.23-0.89	0.25-0.9	0.28-0.93	0.21-0.91	0.35-1	1.00-1	0.31-0.87	0.25-0.86	0.30-0.9
3-tert-butylphenol	0.37-1	0.32-1	0.36-1	0.42-1	0.28-0.93	0.29-0.89	0.20-0.91	0.21-0.9	0.21-0.9	0.21-1	0.21-1	0.19-0.91	0.21-0.92	0.23-0.95	0.21-1	0.23-0.9	0.31-0.87	1.00-1	0.50-1	0.45-1
4-tert-Butylphenol	0.32-1	0.32-1	0.36-1	0.38-1	0.28-0.92	0.32-0.92	0.19-0.9	0.19-0.89	0.19-0.89	0.20-1	0.19-1	0.18-0.86	0.19-0.87	0.21-0.95	0.19-1	0.22-0.85	0.25-0.86	0.50-1	1.00-1	0.48-1
4-tert-Buty-2-methylphenol	0.40-1	0.45-1	0.34-1	0.31-1	0.27-0.95	0.39-0.99	0.19-0.9	0.20-0.89	0.20-0.91	0.28-1	0.31-1	0.26-0.89	0.27-0.88	0.22-0.95	0.27-1	0.36-0.85	0.30-0.9	0.45-1	0.48-1	1.00-1

Table 5: Shows chemicals compared from the LLNA skin sensitisation dataset (dataset 2) and the range of Tc scores calculated with different fingerprints.

Chemicals Compared		Shows Tc Scores and the fingerprint used to calculate Tc.						Range of Tc across fingerprints
		CDK Standard	CDK MACCS	CDK Extended	CDK PubChem	CDK FCFP6	CDK ECFP4	
(LLNA score, sensitiser classification (Gerberick et al., 2005))								
1,4- dihydroxyquinone (0.1, strong sensitiser)	Resorcinol (5.0, non-sensitiser)	0.79	0.88	0.79	0.87	0.54	0.43	**0.43-0.88**
3-phenylenediamine (2.5, strong sensitiser)	Aniline (5.0, weak sensitiser)	0.89	0.78	0.88	0.92	0.75	0.53	**0.53-0.92**
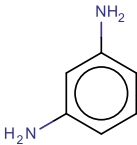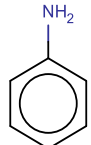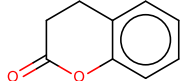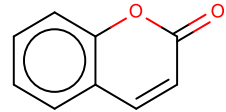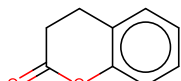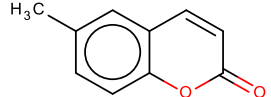
3,4-dihydrocoumarin (2.5, moderate sensitiser)	Coumarin (5.0, non-sensitiser)	0.43	0.73	0.48	0.86	0.40	0.35	**0.35-0.86**
3,4-dihydrocoumarin (2.5, moderate sensitiser)	6-methylcoumarin (5.0, non-sensitiser)	0.40	0.74	0.43	0.83	0.27	0.21	**0.21-0.83**

Figure Captions:

Figure 1. Diagrammatic illustration of how a chemical structure may be converted into a bit string.

Figure 2: Shows overview of the Tc similarity matrices for the perfluorinated acids dataset (dataset 3), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4 fingerprints.
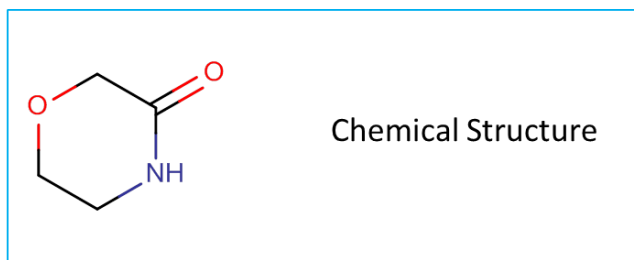
Figure 3: Shows overview of the Tc similarity matrices for the alkylphenols dataset (dataset 6), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4 fingerprints.
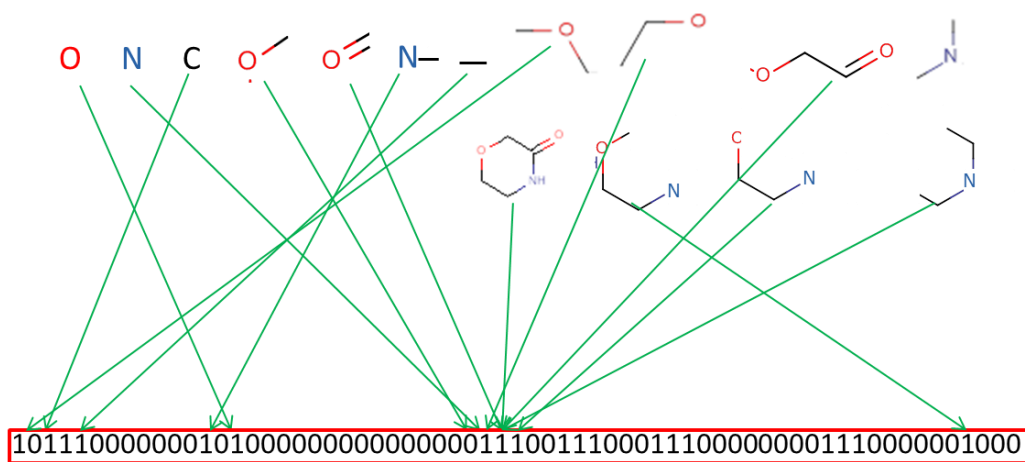
Figure 4: Shows overview of the Tc similarity matrices for the saturated alcohols dataset (dataset 4), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4 fingerprints.

Figure 5: Shows overview of the Tc similarity matrices for the unsaturated alcohols dataset (dataset 5), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4 fingerprints.

Figure 6: Shows overview of the Tc similarity matrices for the non-polar narcotic dataset (dataset 1), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4 fingerprints.

Figure 1

Chemical Structure

Sub-structures occurring in the chemical

10111000000010100000000000000001110011100011100000000111000001000

Bit value set in fingerprint

Figure 2

(A)

CDK Standard							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	0.94	0.87	0.85	0.85	0.85	0.85
PFHpA	0.94	1.00	0.92	0.91	0.91	0.91	0.91
PFOA	0.87	0.92	1.00	0.98	0.98	0.98	0.98
PFNA	0.85	0.91	0.98	1.00	1.00	1.00	1.00
PFDA	0.85	0.91	0.98	1.00	1.00	1.00	1.00
PFUA	0.85	0.91	0.98	1.00	1.00	1.00	1.00
PFDoA	0.85	0.91	0.98	1.00	1.00	1.00	1.00

(B)

CDK MACCS							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFHpA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFOA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFNA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFUA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDoA	1.00	1.00	1.00	1.00	1.00	1.00	1.00

(C)

CDK Extended							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	0.92	0.85	0.84	0.84	0.84	0.84
PFHpA	0.92	1.00	0.92	0.91	0.91	0.91	0.91
PFOA	0.85	0.92	1.00	0.99	0.99	0.99	0.99
PFNA	0.84	0.91	0.99	1.00	1.00	1.00	1.00
PFDA	0.84	0.91	0.99	1.00	1.00	1.00	1.00
PFUA	0.84	0.91	0.99	1.00	1.00	1.00	1.00
PFDoA	0.84	0.91	0.99	1.00	1.00	1.00	1.00

(D)

CDK PubChem							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	0.91	0.86	0.86	0.86	0.86	0.86
PFHpA	0.91	1.00	0.94	0.94	0.94	0.94	0.94
PFOA	0.86	0.94	1.00	1.00	1.00	1.00	1.00
PFNA	0.86	0.94	1.00	1.00	1.00	1.00	1.00
PFDA	0.86	0.94	1.00	1.00	1.00	1.00	1.00
PFUA	0.86	0.94	1.00	1.00	1.00	1.00	1.00
PFDoA	0.86	0.94	1.00	1.00	1.00	1.00	1.00

(E)

CDK FCFP6							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	0.87	0.83	0.83	0.83	0.83	0.83
PFHpA	0.87	1.00	0.96	0.96	0.96	0.96	0.96
PFOA	0.83	0.96	1.00	1.00	1.00	1.00	1.00
PFNA	0.83	0.96	1.00	1.00	1.00	1.00	1.00
PFDA	0.83	0.96	1.00	1.00	1.00	1.00	1.00
PFUA	0.83	0.96	1.00	1.00	1.00	1.00	1.00
PFDoA	0.83	0.96	1.00	1.00	1.00	1.00	1.00

(F)

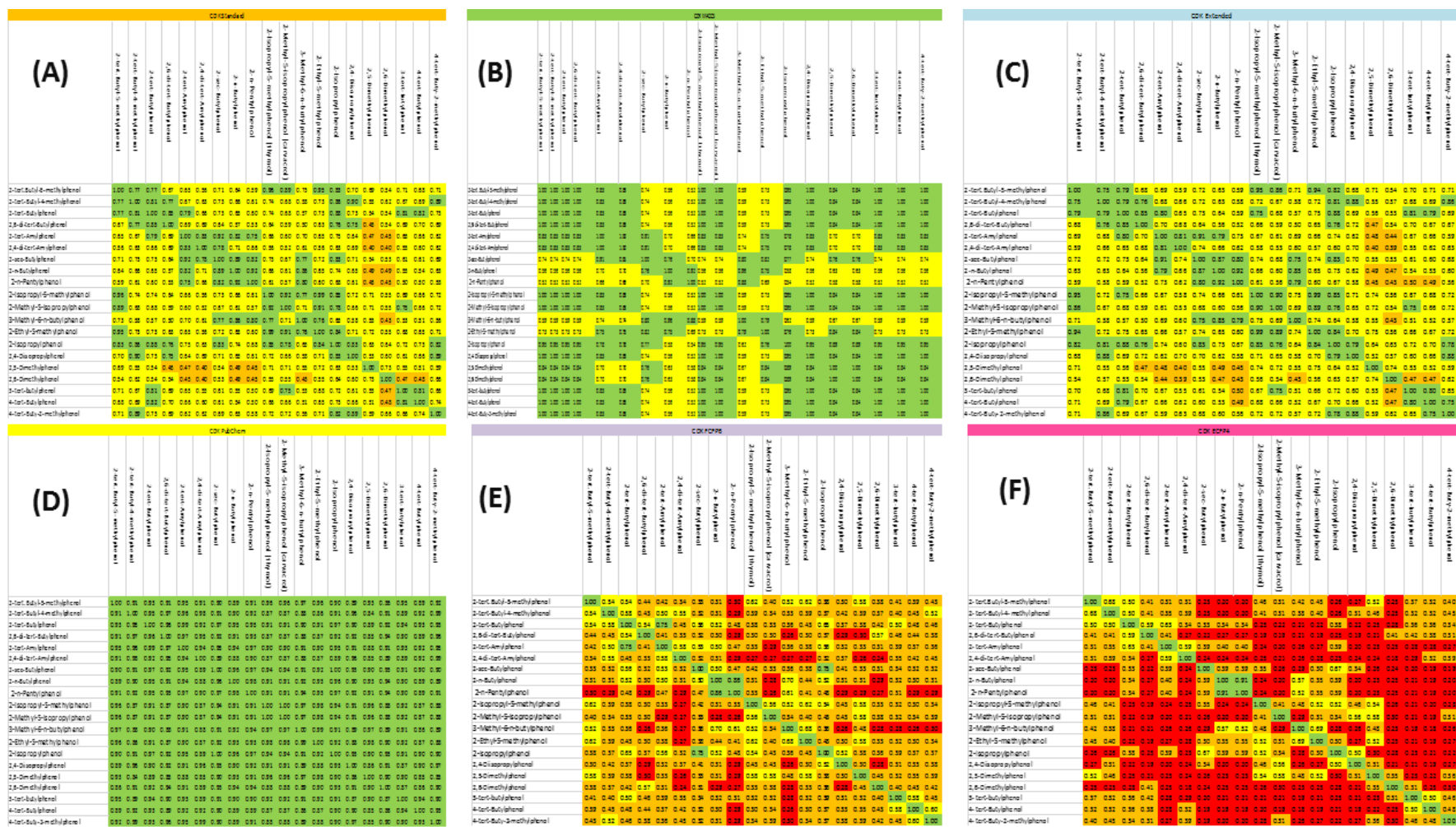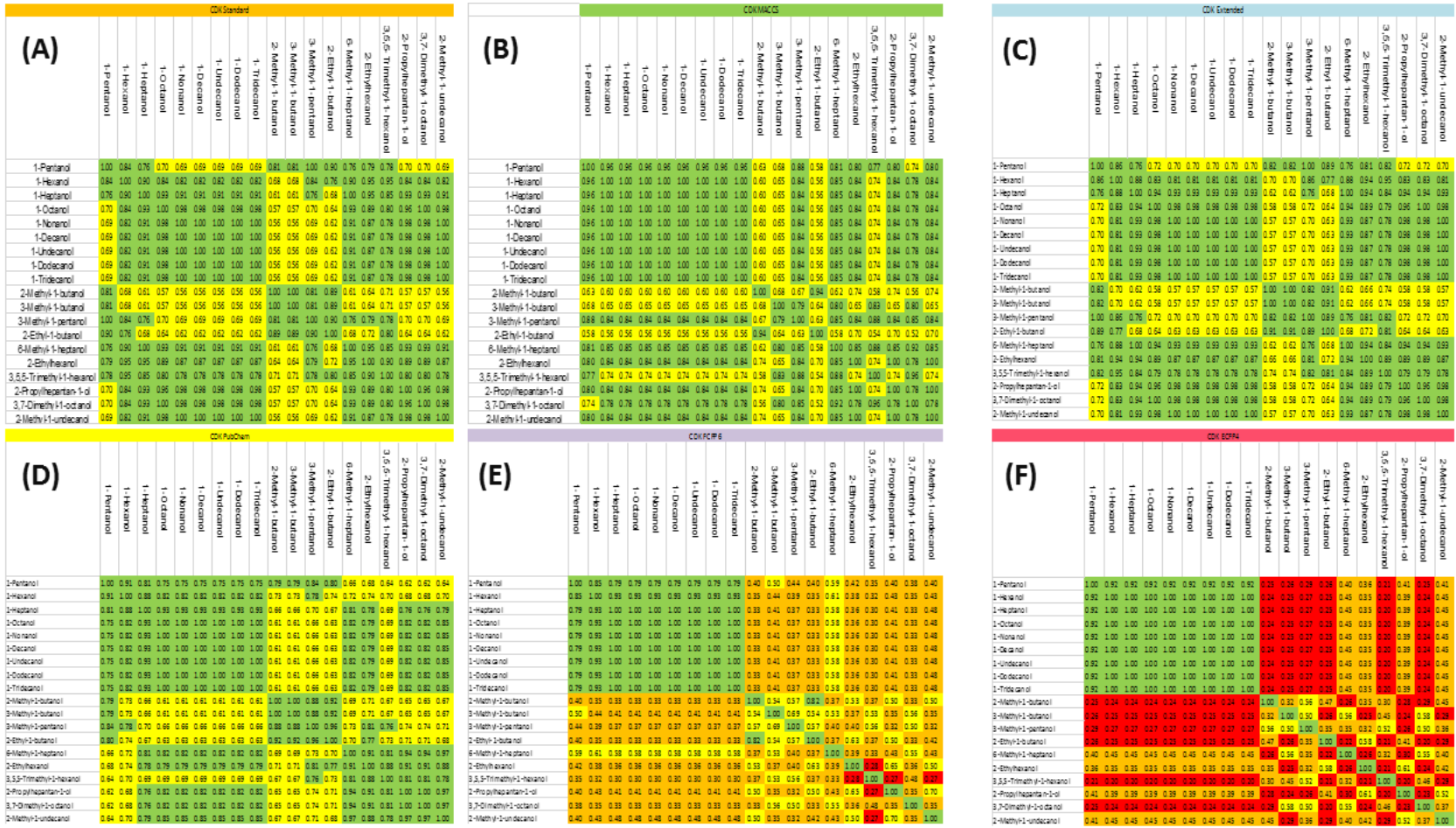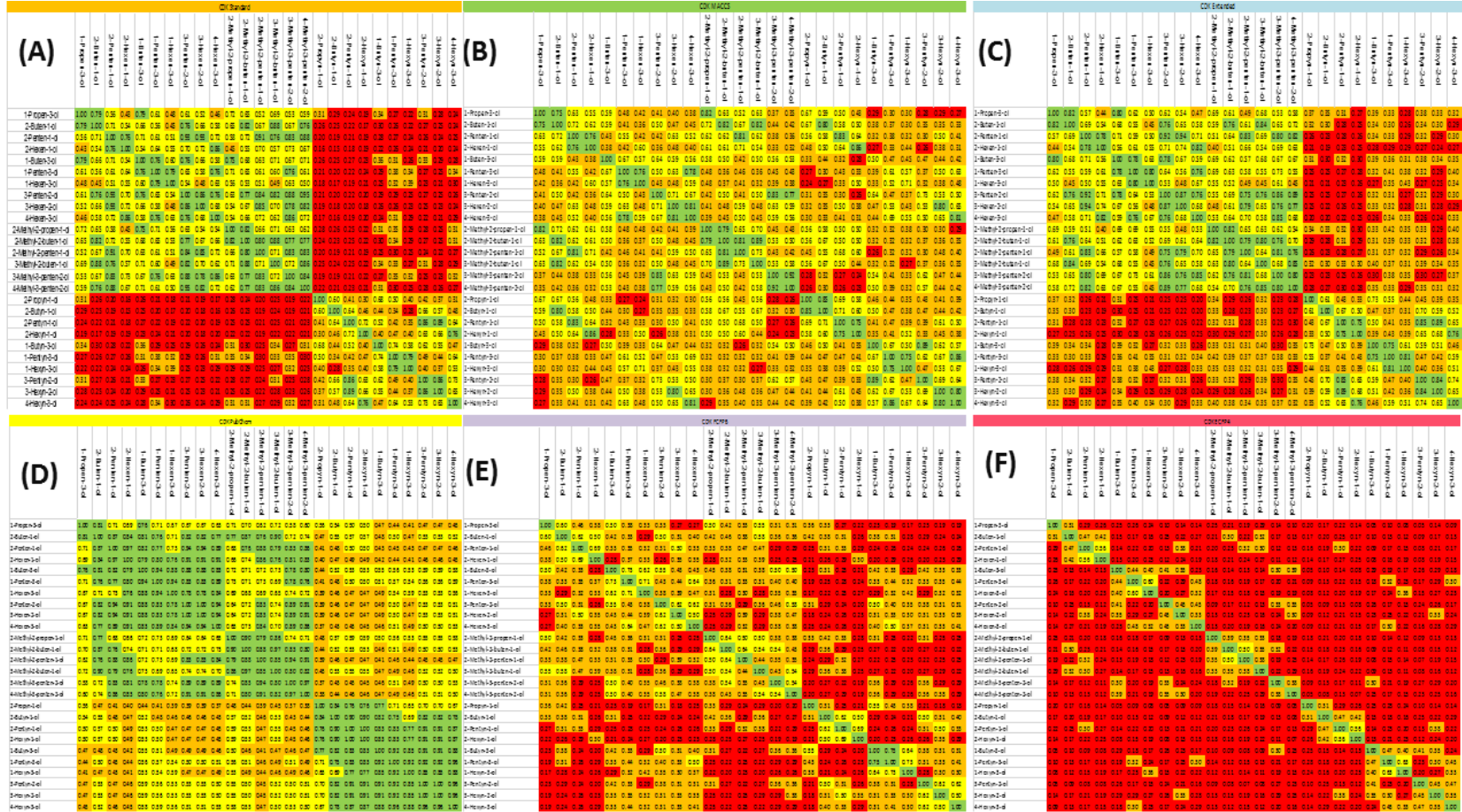| CDK ECFP4							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFHpA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFOA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFNA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFUA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDoA	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Figure 3

Figure 4

Figure 5

Figure 6