

Article

Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across: Recommendations for Optimal Use

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.

Available at http://clok.uclan.ac.uk/24855/

Mellor, Claire ORCID: 0000-0002-7647-2085, 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

It is advisable to refer to the publisher's version if you intend to cite from the work. http://dx.doi.org/10.1016/j.yrtph.2018.11.002

For more information about UCLan's research in this area go to http://www.uclan.ac.uk/researchgroups/ and search for <name of research Group>.

For information about Research generally at UCLan please go to http://www.uclan.ac.uk/research/

All outputs in CLoK are protected by Intellectual Property Rights law, including Copyright law. Copyright, IPR and Moral Rights for the works on this site are retained by the individual authors and/or other copyright owners. Terms and conditions for use of this material are defined in the <u>http://clok.uclan.ac.uk/policies/</u>



2 3	Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-Across: Recommendations for Optimal Use
4	
5	C.L. Mellor ¹ , R.L. Marchese Robinson ¹ , R. Benigni ² , D. Ebbrell ¹ , S.J. Enoch ¹ , J.W.
6	Firman ¹ , J.C. Madden ¹ , G. Pawar ¹ , C. Yang ³ and M.T.D. Cronin ¹ *
7	
8	¹ School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom
9	Street, Liverpool L3 3AF, England
10	² Alpha-Pretox, Via G. Pascoli 1, 00184, Rome, Italy
11	³ Molecular Networks GmbH, Neumeyerstraße 28, 90411 Nürnberg, Germany
12	
13	
14	*Corresponding author (Mark Cronin): Tel. +44 151 231 2402; e-mail address:
15	M.T.Cronin@ljmu.ac.uk

17 ABSTRACT

Computational approaches are increasingly used to predict toxicity, in part due to pressures to 18 19 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 20 21 molecules tend to exhibit similar activities, i.e. molecular similarity is integral to read-across. 22 Various molecular fingerprints and similarity measures may be used to calculate molecular 23 similarity. This study investigated the value and concordance of the Tanimoto similarity values 24 calculated using six widely used fingerprints within six toxicological datasets. There was 25 considerable variability in the similarity values calculated from the various molecular fingerprints for diverse compounds, although they were reasonably concordant for homologous 26 series acting via a common mechanism. The results suggest generic fingerprint-derived 27 similarities are likely to be optimally predictive for local datasets, i.e. following sub-28 categorisation. Thus, for read-across, generic fingerprint-derived similarities are likely to be 29 30 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. 31

32

KEYWORDS: Read-across; toxicity; molecular fingerprint; regulatory acceptance;

34 molecular similarity; Tanimoto coefficient; *in silico*

35 HIGHLIGHTS

37	-	Molecular fingerprints to identify read-across analogues have been evaluated
38	-	Identification of read-across analogues is dependent on the molecular fingerprint
39	-	Commonly used molecular fingerprints may not address the mechanism of toxic action
40	-	Commonly used molecular fingerprints are most likely to be predictive within a
41		homologous series
42	-	Similarity measures tailored to the endpoint are likely to be most useful
43		

44 1. INTRODUCTION

The use of alternative approaches to assess chemical safety is growing due to legislation that 45 46 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, -47 omics and computational approaches ((Quantitative) Structure-Activity Relationships 48 ((Q)SARs), read across etc.) have become integral to many hazard assessment strategies. Of 49 these, computational or (Q)SAR (in silico) approaches aim to predict the toxicity of compounds 50 from descriptors of chemical structure and thus reduce testing. In particular, read-across is at 51 52 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 53 2017; Patlewicz et al 2018). Read-across relies on the ability to identify similar molecules with 54 the assumption that similar molecules will tend to exhibit similar activity or, at least, show 55 similar trends in activity (OECD, 2014). Although the concept of similarity has growing 56 57 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., 58 2011; Teubner et al., 2015; Ball et al., 2016; Schultz and Cronin 2017; Chesnut et al 2018). Of 59 the barriers identified by Ball et al (2016), some are more trivial to address than others, e.g. 60 full documentation and ensuring the correct chemical structure is provided. The most difficult 61 aspect of justifying a read-across argument is the assessment of "similarity" and being able to 62 provide evidence for such, so to build scientific confidence (Patlewicz et al., 2015; Schultz et 63 64 al 2018). For instance, there is a concern over effects such as activity cliffs, where structurally 65 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 66 67 of how to define similarity from a molecular level (Maggiora et al., 2014) as well as adequately 68 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 69 70 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 71 72 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, 73 such as chemical structure. Thus, the means of defining similarity and providing a means to 74 75 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). 76 77 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 78 case for practical applications. Some of the insights presented herein, regarding the strengths 79 80 and weaknesses of molecular fingerprint derived similarity measures, may also be applicable 81 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 82 83 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, 84 definitions are stated in Table 1 that are pertinent to this investigation. 85

86 **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

of common, mechanistically relevant, structural features and transformation to the same 93 metabolite or degradant (OECD, 2014). There is also the more general concept of "chemical 94 95 similarity", i.e. using measures of similarity based on common structural features, physicochemical or biological properties and / or calculated variables related to molecular 96 structure (descriptors). This broader notion of "chemical similarity", in contrast to those which 97 are deemed acceptable for regulatory purposes, may be defined in terms of generic structural 98 99 features / properties / variables, which are not necessarily relevant to the endpoint of interest. These approaches use chemometrics, the science of using mathematics and statistics to analyse 100 101 chemical data in order to obtain knowledge about chemical systems; elsewhere, the term cheminformatics or chemoinformatics may be used.) Chemometric measures of similarity are 102 widely used as they are rapid and cost effective due to the availability of online tools, e.g. 103 104 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 mechanistically relevant fragment based methods to identify similar molecules for read-across 106 is relatively well developed (Schultz et al., 2015), much less is known about the use of 107 "chemical similarity", as defined above, for read-across. This is an area that was founded in 108 the identification of new leads for drug development, thus the similarity measures were not 109 necessarily intended for the purpose for which they are currently applied. For grouping and 110 read-across, where there is no rational measure to find similar compounds, or where a large, 111 diverse inventory is being searched, chemometric methods may seem appealing. However, 112 there is no clear guidance on how they may be applied. 113

The generation of chemometric similarity requires the conversion of chemical structures into machine readable representations which are then compared using one of the many available similarity coefficients (Willett et al., 1998; Holliday et al., 2003). The calculated similarity can vary depending on the type of representation chosen and which similarity coefficient is used.

Most similarity calculations rely on the use of (molecular) fingerprints in order to generate 118 machine readable bit representations from chemical structure. Fingerprints are based mostly on 119 2D representations of a molecule and are used due to their computational efficiency (Holliday 120 et al., 2003). The process of generating bits from chemical structure is illustrated by Figure 1, 121 for a scenario in which the corresponding structural features are molecular substructures A 122 fingerprint is typically a binary vector, with bits set to 1 or 0 depending on the presence or 123 124 absence of a structural feature (e.g. molecular substructure) within the molecule of interest. In principle, there does not have to be a simple one-to-one correspondence between the presence 125 126 of a structural feature and the presence of a molecular substructure. For example, one of the features employed in the RDKit implementation of the MACCS fingerprint corresponds to 127 "two methyl groups" (https://github.com/rdkit/rdkit-128 or more orig/blob/master/rdkit/Chem/MACCSkeys.py). Moreover, other fingerprints might encode the 129 occurrence count of structural features, rather than simply their presence or absence. However, 130 if the fingerprint only encodes the presence or absence of certain fragments and not their 131 quantity, this may be a limitation (Flower, 1998). For this scenario, a molecule can contain a 132 specific fragment 1 or 100 times and the resulting bit string will be set the same, thus giving 133 little information with regards to, for instance, molecule size and which fragments occur more 134 often within a molecule (Flower, 1988). 135

136 FIGURE 1 HERE

Many different types of molecular fingerprints are used to calculate the similarity between two molecules. Two of the most widely used are the molecular access system (MACCS) fingerprint and the extended connectivity fingerprint (ECFP). The MACCS fingerprint was one of the first developed and is amongst the most commonly used for similarity calculations. MACCS is a prototypic fingerprint, which typically contains 166 structural features, related to the presence and occurrence count of substructures comprising a variety of non-hydrogen ("heavy") atoms
(Maggiora et al., 2014), albeit this may be implementation dependent
(http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html,

https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py). 145 The ECFP defines molecular features by assigning identifiers to each of the heavy atoms in the molecule, 146 based upon atomic properties and bonding arrangements, and then combining those identifiers 147 148 with those assigned to neighbouring heavy atoms up to a specified number of bonds away (Rogers and Hahn, 2010). The most commonly used ECFP fingerprint is ECFP4, which has a 149 150 bond diameter of four. ECFP4 comprises features derived from the compounds in the analysed dataset, which necessarily overlap, in contrast to the MACCS fingerprint, for which the features 151 are pre-defined (Maggiora et al., 2014). In simple terms, approaches such as ECFP are more 152 complex than MACCS, allowing for the generation of many different atom environments and 153 describe molecular structure more subtly. Finally, it should be noted that different variants of 154 both fingerprints may be computed by different software programs (Rosenbaum et al., 2011; 155 http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs key 44.html, 156

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

A coefficient is used to assess the similarity of two, or more, molecules as defined by the fingerprints. The similarity coefficient most frequently combined with the use of fingerprints is the Tanimoto coefficient (Tc). (Elsewhere, this may be termed the Jaccard similarity (Willett et al., 1998; Luechtefeld et al., 2018).) For molecules described in terms of bit-vector molecular fingerprints, Tc is computed as per equation (1), albeit a more general definition exists for continuous variables (Willett et al., 1998).

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

In equation (1), the Tanimoto coefficient (Tc) for the similarity of two objects (molecules) A 167 and B is a function of the number of features present within compounds A and B (a and b 168 respectively), and the number of features shared by A and B (c). With regard to molecular 169 fingerprints, a and b are the number of structural features, or bits set to 1, in each molecule, c 170 is the number in common. Therefore, Tc quantifies the fraction of features common to A and 171 B as a fraction of the total number of features of A or B, where the c term in the denominator 172 corrects for double counting of the features (Willett et al., 1998; Maggiora et al 2014). It is 173 174 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 175 applied is vital. Also of relevance to read-across is the value of Tc that would constitute 176 177 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 178 compounds (OECD, 2014). There is no definitive rule or guidance for use of Tc or specific 179 fingerprints, in part due to the differences in calculated values. Within the drug design 180 community, it is often considered that knowledge of the point at which the similarity of A and 181 182 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 183 bioactivity (Maggiora et al 2014). However, studies have since shown that this value is not 184 reliable, especially when different molecular representations are used (Eckert et al., 2007; 185 186 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, 187 188 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; 189 Willett, 2013; Bajusz et al., 2015, Cereto-Massague et al., 2015). 190

Whilst widely applied, a number of studies have shown that using Tc to calculate chemical 191 similarity has its limitations and weaknesses (Dixon and Koehler, 1999; Flower, 1998; 192 193 Holliday et al., 2002; Laiiness, 1997). Godden et al (2000) demonstrated that Tc has a tendency to produce a similarity score of about 0.3 even for structurally distant molecules. It has also 194 been suggested that Tc calculations are biased towards smaller molecules when used for 195 selection according to diversity and that other coefficients may be more appropriate for some 196 197 data types (Dixon et al., 1999; Lajiness et al., 1997; Whittle et al., 2003). Moreover, as is perhaps most relevant for the purposes of toxicity prediction, Tc is a generic measure of 198 199 molecular similarity which treats the shared presence of mechanistically irrelevant substructures as equally important as the shared presence of mechanistically crucial 200 substructures, such as those corresponding to structural alerts (Alves et al., 2016). One way of 201 202 taking account of this is to compute a weighted Tanimoto index (Maunz et al., 2008). Nonetheless, in spite of its known limitations, a Tanimoto similarity of 0.7 is elsewhere 203 considered as a cut-off for read-across (Enoch et al 2009; Hartung, 2016). 204

The aim of this study was to determine the value of different molecular fingerprints to assess 205 molecular similarity, in terms of the Tanimoto coefficient, in the context of read-across. In 206 particular, the focus of the study was to examine scenarios in which these similarity values 207 might be useful for read-across based upon pairwise comparison to one or a few chemicals, 208 with measured endpoint data, for the purpose of toxicological data gap filling. Specific 209 objectives were to assess the performance and reliability of different molecular fingerprints 210 used in similarity analysis, with a view to determine when similarity computed in this fashion 211 212 works well and does not work well, as well as to consider how molecular similarity can be placed into a mechanistic framework to predict toxicity taking in account molecular initiating 213 214 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 in Table 2. The datasets were small in size (from 7 to 211 compounds) compared to more 221 complex inventories, e.g. of REACH chemicals, or databases that may be investigated for drug 222 discovery. The selection of the datasets was influenced by a number of factors. Datasets were 223 224 chosen which had been the subject of previous read-across or QSAR analyses, or potentially could be used as such. These were datasets that the authors were familiar with, hence allowing 225 for an understanding of the selection process for compounds as well as the quality of the 226 underlying biological data. They were also chosen to represent a range of mechanisms and 227 molecular initiating events which may influence the use of molecular similarity. 228

229

230 **2. METHODS**

231 **2.1 Data Sets Analysed**

In total six different datasets were chosen to calculate Tc in this study. These datasets were chosen as they provided different read-across scenarios, thus allowing similarity calculations based on different fingerprints to be assessed for reliability/ accuracy. The six data sets (Table 2) chosen were analysed and a Tanimoto score for each pair of chemicals within each data set was calculated for the different fingerprints.

237 TABLE 2 HERE

238

239 2.2 Calculation of molecular fingerprints

Molecular fingerprints and Tanimoto similarities were calculated using the freely available 240 **KNIME** software (version 3.3.0). **KNIME** workflow 241 А (http://dx.doi.org/10.5281/zenodo.1401196) was developed that applied the CDK Fingerprints 242 node to calculate 2D fingerprints and then to calculate different Tanimoto similarities, in terms 243 of these fingerprints, between the molecules in a dataset provided as an SDF file. Tanimoto 244 similarities (Tc) in terms of these bit-vector fingerprints were calculated as per equation (1). 245 The CDK fingerprints calculated were the CDK Standard, CDK Extended, CDK PubChem, 246 CDK FCFP6, CDK ECFP4 and the CDK MACCS fingerprints. 247

248

249 **2.3 Analysis of Tanimoto coefficients.**

The performance of the six different fingerprints to calculate Tc was analysed via the 250 visualisation of the similarity matrices. This was performed by adding the following 251 conditional formatting rules to cells within a Microsoft Excel spreadsheet: green (values 252 between 0.75 and 1), yellow (values between 0.5 and 0.749), orange (values between 0.3 and 253 254 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" 255 chemicals. The ranges of Tc scores were subsequently calculated to determine if knowledge 256 257 could be gained about which fingerprint works best for the different datasets.

258

259 **3. RESULTS**

The KNIME workflow produced a CSV file which contained calculated Tc values for the input 260 data sets. The Tc data matrices for the datasets are provided in the supplementary information. 261 262 Figures (2-6) show the visualisation of the calculated Tc similarity matrices for five different datasets (perfluorinated acids, alkylphenols, saturated alcohols, unsaturated alcohols and the 263 non-polar narcotic datasets), full details of which can been found within the supplementary 264 information along with the matrices for the LLNA skin sensitisation dataset. (The size of the 265 266 LLNA dataset meant that it was not possible to produce an informative image of the similarity matrices.) In each of these figures, the Tc scores for the same dataset using the six different 267 268 fingerprints are shown, where A was calculated using CDK Standard fingerprints, B was calculated using CDK MACCS fingerprints, C was calculated using CDK Extended 269 fingerprints, **D** was calculated using CDK PubChem fingerprints, **E** was calculated using CDK 270 271 FCFP6 fingerprints and **F** was calculated using CDK ECFP4 fingerprints. Each figure shows pairwise Tc values for all compounds in the dataset, with the similarity between compound *i* 272 and *j* being shown in the matrix element of row *i* and column *j* of the matrix, such that the Tc 273 values for the same compound compared to itself (Tc=1.0) lie along the diagonal elements. 274 N.B. (1) Each row (column) in these images is labelled by the name of the chemical for which 275 colour coded similarity values are reported within that row (column). (2) These images are 276 designed to illustrate the variation in pairwise similarity for the same pairs of compounds using 277 different fingerprints in terms of the corresponding colour patterns. The size of some datasets 278 279 necessarily makes it hard to read the individual pairwise similarity values from these images. Hence, all pairwise similarity values are provided in an Excel workbook in the Supporting 280 Information. In addition, Tables 3 - 5 show the range of Tanimoto similarity values that can 281 282 be obtained for the same pairwise comparisons, between compounds in selected datasets, using the different fingerprints. 283

284 FIGURES 2-6 HERE

285 **TABLES 3-5 HERE**

286

287 4. DISCUSSION

Chemical similarity is, in theory, a beguiling concept allowing for the identification of similar 288 289 molecules to those with existing information, whether it be biological activity (such as pharmacological or toxicological effects), biokinetics, environmental fate or physico-chemical 290 properties. The science of molecular similarity is founded in drug discovery, where the aim 291 was to identify similar molecules to a known active compound. It mostly utilises easily 292 calculable parameters (descriptors), or fingerprint representations, of molecular structure. The 293 294 application of molecular similarity is typically based around the Tanimoto coefficient computed from bit-vector fingerprints, as per the current work. As such, there has been a strong 295 interest in this approach in drug discovery for many years and there has been a recent growth 296 297 of interest in the field of toxicology to enable data gap filling. With regard to toxicity prediction, the focus of the application of molecular similarity has shifted from being intended to identify 298 molecules highly similar to a known active (assuming a receptor mediated pharmacological 299 effect) to multiple uses ranging from searching for any "similar" molecules to a target query 300 with unknown activity, to serving as the input to grouping and/or read-across approaches (Gini 301 302 et al., 2014; Luechtefeld et al., 2016a-d; 2018). As use of these approaches grows, it is clear that issues may arise with analogues being identified of little relevance, or important analogues 303 not being identified as the similarity measures are not appropriate. The purpose of this study, 304 305 therefore, was to assess the use of some commonly applied measures of similarity to investigate their use and provide a means of making recommendations for their use for techniques such as 306 read-across, with a focus on read-across predictions made using pairwise similarity calculations 307 308 to one, or a few, chemical(s), rather than, say, supervised machine learning approaches using

large quantities of data. To this end, six datasets were analysed which have previously been
subject to some form of read-across or QSAR approaches. All have well defined endpoints
with varying levels of confidence in the mechanistic rationale.

A number of different molecular fingerprints were calculated to determine the advantages or 312 disadvantages of a single method. The similarity matrices in Figures 2-6 clearly demonstrate a 313 difference in Tc scores calculated for the same dataset when using different fingerprints. Closer 314 examination of the perfluorinated acids dataset (Figure 2, dataset 3 from Table 2) indicates a 315 concordance in the fingerprints with regard to in their Tc values as all data matrices are green 316 317 (values of between 0.75 and 1), showing chemicals are "highly similar". For this data set, the Tc similarity matrices showed good concordance regardless of which fingerprint was chosen 318 i.e. the Tc based assessment of all chemicals as highly similar is in keeping with the assessment 319 which would be made by toxicological experts - since this dataset comprises a homologous 320 series, i.e. the same functional group with varying chain length, expected to act via a common 321 322 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 with each other, those chemicals with C10-C12 gave similarity scores of 1 when compared 324 with each other and the chemical with C9 tended to only show a similarity score of 1 when 325 compared against itself (for CDK standard, CDK Extended fingerprints) or those with C10-326 C12 (for the other fingerprints). Naturally, all fingerprints gave a Tc value of one for 327 comparisons of the same compound to itself. This trend was similar for all fingerprints applied 328 to this dataset. Thus, fingerprint similarity, in terms of Tc, is a reasonable measure when 329 330 applied to homologous, or highly similar, series of chemicals, regardless of the fingerprint chosen With regard to read-across, this would indicate that it may be appropriate for "fine-331 tuning" a read-across within such a preselected series of chemicals - the process sometimes 332 333 referred to as sub-categorisation.

Analysis of datasets with greater structural variability (cf. Figures 3 - 6) indicates a much higher 334 variability in the calculated Tc values depending on which fingerprint was chosen, with limited 335 336 concordance between them. For example, compare the Tc results for the alkylphenol dataset calculated with CDK FCFP6 against those calculated using the CDK PubChem fingerprints. 337 For two chemicals, 3-methyl-6-n-butylphenol and 2,6-di-tert-butylphenol, CDK FCFP6 338 fingerprints gave a Tc score of 0.26, whereas CDK PubChem fingerprints gave a Tc score of 339 340 0.88. For both the alkylphenols (Figure 3) and saturated alcohols (Figure 4) datasets, the Tc value computed from the CDK Standard, CDK MACCS, CDK Extended and, for Figure 4, 341 342 CDK PubChem fingerprints showed some concordance, with a similar pattern of colours denoting the degree of similarity as indicated by the Tc values. However, for both these datasets 343 the calculated Tc values for CDK FCFP6 and the CDK ECFP4 fingerprints were significantly 344 different to the Tc values from the other four fingerprints, with the CDK ECFP4 giving many 345 values that would suggest "highly dissimilar" chemicals, which is not the case for these datasets 346 (based upon expert judgement). Similar discrepancies between fingerprints were seen for the 347 non-polar narcosis dataset (Figure 6). The reasons for such discrepancies undoubtedly reflect 348 the method of fingerprint calculation having an enormous impact on the identification of 349 analogues from large structurally heterogeneous datasets. It may even be an indicator for 350 consideration of composite Tc scores to capitalise on the different information contained. 351 However, that would not address the possibility that toxicologically irrelevant structural 352 353 variation is being reflected in these similarity values and that relevant structural variation may not be being appropriately captured, even when the information from all fingerprints was 354 combined. Overall, care must be applied in using Tc values for structurally heterogeneous 355 356 datasets. To make optimal use of Tc values, the user should arguably decide carefully, and rationally, on which fingerprint to use, requiring the user to first give some thought to the 357 fingerprints and mechanism of the endpoint to be read across. 358

For the unsaturated alcohols dataset (Figure 5), all the calculated Tc similarity matrices were 359 noticeably different for each of the six fingerprints used. This dataset consist of chemicals 360 which are, on the face of it, structurally similar but with subtle changes and differences not 361 only in chain length but also the position of the hydroxyl group, (primary or secondary alcohol), 362 branching, and position (internal or external) of the double bond. The positioning of the alcohol 363 group and double bond, as well as branching, will impact of toxicity (Schultz et al., 2017), 364 365 however none of the Tc values assisted in identifying rational, mechanistically similar analogues across the group. Therefore, subtle, mechanistically relevant changes in molecular 366 367 structure, such as branching and positional effects may not be captured by any of the fingerprints considered here. Moreover, these most relevant changes will be treated as equally 368 important to whether irrelevant molecular substructures are shared or not between two 369 370 molecules.

Using molecular similarity to assist in toxicity prediction is unlikely to be perfect. There are 371 372 many examples of highly similar chemicals, in terms of Tc value, having very different toxicity profiles. For example, Table 5 lists four pairs of compounds, selected from the LLNA skin 373 sensitisation dataset, showing potential issues with activity cliffs, despite high Tc values from 374 some fingerprints. Comparison of 1,4-dihydroxyquinone, a strong skin sensitiser, with 375 resorcinol (1,3-dihydroxyquinone), a non-sensitiser, indicates both chemicals being highly 376 similar in structure with the only difference being the position of the hydroxyl groups on the 377 phenol ring (Table 5). The position of the hydroxyl groups in 1,4-dihydroxyquinone enables 378 379 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 (Bajot et al., 2011, Enoch et al., 2011). However, the Tc scores for most fingerprints in Table 5 381 indicate high similarity, which could lead to false assumptions with regard to grouping and 382 read-across, unless the mechanism of action is known. The wide range of Tc scores calculated 383

also shows the variability of the Tc scores dependent upon the choice of fingerprint. This 384 emphasises the importance of choosing the most appropriate fingerprint, if any, for similarity 385 calculations. In the second comparison 3-phenylenediamine, a strong skin sensitiser, is 386 compared against aniline, a weak skin sensitiser. These chemicals are highly similar in structure, 387 with the main difference being the presence of an extra amine group (Table 5). It has been 388 demonstrated that the presence of the 2 amine groups in 3-phenylenediamine makes this 389 390 chemical more reactive and leads to its ability to induce strong skin sensitisation (Bajot et al., 2011, Enoch et al., 2011). The Tc scores for this comparison again show variability dependent 391 392 upon fingerprint choice, with the majority of fingerprints giving a highly Tc score that could be interpreted as indicating these chemicals should have highly similar sensitizing activity. 393 Clearly, this would be an incorrect conclusion. 394

The final two comparisons compare 3,4-dihydrocoumarin, a moderate skin sensitiser, against 395 coumarin and 6-methylcoumarin which are both non-sensitisers (Table 5). These chemicals are 396 397 all structurally similar with the main difference being the presence of a methyl group and the presence of a double bond (Table 5). The presence of a double bond in the second ring of 398 coumarin causes it to be readily metabolised via Michael addition, into a non-sensitising 399 metabolite (Table 5). The absence of the double bond makes 3,4-dihydrocoumarin more 400 reactive, which accounts for its moderate skin sensitisation when compared to the other two 401 chemicals. The Tc scores calculated for these two comparisons again show variability 402 dependent on fingerprint choice (Table 5). Two of the six fingerprints (CDK MACCS and CDK 403 PubChem) resulted in high Tc scores; this would suggest these chemicals exhibit similar 404 405 endpoint values, which would be invalid with regards to skin sensitisation.

One means of addressing the problems with fingerprint based Tc values calculated for non-homologous datasets, for which subtle changes in molecular structure may lead to significant

changes in toxicity for certain endpoints, would be to investigate similarity values calculated 408 using a limited number of mechanistically relevant descriptors chosen based on expert 409 judgement. For example, in the case of skin sensitization, the electrophilicity index could be 410 used (Enoch et al., 2008). Similarities might be computed based upon the more general 411 expression for the Tanimoto coefficient, for continuous variables (Willett et al., 1998), 412 following normalisation of different descriptors to the same scale. However, even under this 413 414 scenario, it is possible that grouping of the chemicals, to ensure that they acted via a common MIE, would first be required before similarity coefficients could be computed for read-across 415 416 (Enoch et al., 2008).

The visualisation and practical handling of Tc values should be borne in mind. In this 417 investigation, due to the number of chemicals in the LLNA skin sensitisation (211 chemicals) 418 and the non-polar narcotic (87 chemicals) datasets (Figure 6 and supplementary data), both of 419 which are quite modest in size, visualisation was challenging which makes the analysis of 420 421 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 categories from larger datasets before Tc calculation, thus reducing the number of chemicals 423 within each matrix and making visualisation easier. One means of achieving this is that any 424 relevant knowledge of MIEs should be used to pre-categorise the datasets prior to calculating 425 Tc values. For example, Tc values might be computed for chemicals acting via a common MIE, 426 as indicated by a shared structural alert, and for which some other expert based rules reduced 427 mechanistically irrelevant structural variation that would reduce the information conveyed by 428 429 the Tc values. This is likely to be the case if the chemicals could be assigned to a homologous series acting via a common mechanism, where the structural variation in chain length was 430 431 known to be biologically relevant.

In addition, in this study, arbitrary values were applied to visualise the data matrices. The range 432 of 0.75 and 1 was chosen to highlight Tc scores green and show "highly similar" chemicals. It 433 must be remembered that issue of which Tc score is the cut off point for "highly similar", 434 assuming a simple approach based upon saying pairs of "highly similar" chemicals would tend 435 to exhibit "highly similar" biological activity, is not well defined. It is clear from this study that 436 it is very difficult to include a universal "cut-off" and a variable approach to similarity levels 437 438 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. 439 440 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, 441 Nelms et al., 2015). Expert judgement is likely to also have a role to play when deciding 442 whether any single pairwise similarity value is biologically significant, taking into account the 443 observed differences in chemical structures, with reference to understanding of how this is 444 likely to be mechanistically related to the toxicology. 445

Finally, recent work (Luechtefeld et al., 2016d) reported "read-across" predictions of skin 446 sensitisation based upon the most similar chemicals, in terms of Tanimoto similarities 447 computed from PubChem 2D molecular fingerprints, with available skin sensitisation data. 448 Building upon that work, Luechtefeld et al. (2018) proposed approaches to "read-across" 449 predictions of toxicity based upon supervised machine learning which incorporated Tanimoto 450 similarity values, again calculated from PubChem 2D molecular fingerprints, to multiple 451 compounds with experimental toxicity data. (Further work in that latter study also proposed a 452 453 "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, 454 455 which are highlighted above, they reported empirically good results.

It may be speculated that these empirically good results (Luechtefeld et al., 2016d, Luechtefeld et al., 2018) could, in part, reflect the nature of the datasets investigated, e.g. those datasets may comprise categories of structurally similar chemicals acting via a similar mechanism, with structural differences within those categories being biologically relevant, for which Tanimoto similarity values based on molecular fingerprints can be expected to work best. For example, 31% of the skin sensitisation dataset of Luechtefeld et al. (2016d) was composed of Michael acceptors. However, further analysis is required to determine whether this is, indeed, the case.

Moreover, due to the inherent limitations of Tanimoto values of molecular similarities 463 computed from molecular fingerprints and the variation in similarity values which can be 464 obtained with different fingerprints, as highlighted in the current work, it is unlikely that read-465 across predictions based upon these values using a single fingerprint would be optimal for all 466 relevant scenarios. Thus, for the examples that may be taken from the range of datasets 467 investigated in this study, different types of chemical similarity would be required for effective 468 469 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 of action. Indeed, providing a mechanistic rationale for the predictions, rather than just 471 statistical validation, is more likely to lead to acceptance in a regulatory context. 472

In terms of analogue selection, fingerprints may be developed that have a stronger focus on mechanisms of action and thus are more applicable to address toxicological problems e.g. toxicologically relevant structural features such as the ToxPrint chemotypes could be used as a means of developing fingerprints (Richard et al., 2016). The assumption underpinning the improvement that may be assumed in analogue selection and justification is that such fingerprints, if used, would provide better focus on the MIE which is at the heart of mechanistic 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

In conclusion, molecular fingerprint similarity matrices can be used as a means of identifying 487 possible analogues in some contexts. However, on their own, it is difficult to use generic 488 489 similarity measures computed from generic, purely structurally based, fingerprints to support a read-across hypothesis or justification. This is due to several known limitations of generic 490 similarity measures calculated from these fingerprints, which are highlighted in the current 491 492 work. They are liable to exhibit activity cliffs (where small changes to the overall molecular structure, resulting in high similarity values, lead to significant changes in biological activity). 493 The fingerprints may not capture the relevant structural variation (depending upon the 494 fingerprint method) and treat mechanistically irrelevant structural variation equally to 495 mechanistically relevant structural variation. Similarity matrices, calculated from different 496 497 fingerprints, show greater concordance and are better suited to analogue identification for less diverse datasets, especially homologous series. This suggests they could be most appropriate 498 for read-across within a homologous series, acting via a common mechanism, for which the 499 500 variation in chemical structure is known to be related to biological activity This could avoid the pitfall of fingerprint based similarity measures reflecting biologically irrelevant structural 501 variation. Hence, for a read across setting, users of chemically diverse datasets could benefit 502 503 from first forming categories when using molecular fingerprint similarity values.

Whilst Tanimoto similarity values computed from generic molecular fingerprints have been 504 integrated into recent machine learning predictions of toxicity within diverse datasets with 505 empirically successful results, the limitations of these similarity values, highlighted in our work, 506 mean that other approaches to similarity assessment are preferable for read-across. Ideally, 507 similarity values which reflect biologically relevant information, informed by mechanistic 508 understanding, should be employed. This is especially the case in a regulatory context, where 509 510 a mechanistic justification is likely to be required. More preferable approaches to similarity assessment could entail the previously outlined approach, i.e. first applying a mechanism based 511 512 categorisation of the dataset, such that the use of generic similarity values based on molecular fingerprints would only be used to fine tune read-across within a homologous series. 513

More generally, when calculating similarity, the user needs to give careful consideration to the selection of the most appropriate similarity measure to use and, where possible, link this to rational consideration of the mechanism underpinning the endpoint, e.g. in terms of the Molecular Initiating Event (MIE). Following the cautionary examples presented in this work, the following recommendations are made concerning the use of generic similarity coefficients based on molecular fingerprints for read-across predictions of toxicity.

Fingerprint-derived measures of molecular similarity can be a useful means of identifying
 close structural analogues and may have use in the application of read-across for data gap
 filling. Such methods may provide a useful visual approach to molecular similarity.

The similarity value is dependent on the type of fingerprint, or, if a more general similarity
 value is computed, the descriptors and/or properties used for its calculation. The user
 should acquaint themselves with the different fingerprint methods and their intended
 purpose. A method tailored to the toxicity endpoint should ideally be applied.

Of the fingerprint methods considered in this study, there is evidence that Tanimoto 527 similarity values derived from CDK Standard, CDK MACCS, CDK Extended and CDK 528 PubChem fingerprints showed some concordance, for some scenarios, with similarity 529 values for CDK FCFP6 and the CDK ECFP4 providing different information. Further 530 work is required to understand the significance of these findings and at this time no single 531 fingerprint method from those investigated could be considered to be the most optimum. 532 533 These fingerprints may be appropriate to find "structural" analogues in terms of pure chemistry, but these may not be appropriate for toxicological read-across without 534 535 interpretation and further mechanistic knowledge.

Where known, knowledge of the MIE will guide the successful application of molecular
similarities for toxicological read-across. Reference to the MIE will improve mechanistic
justification of the analogue selection and might be achieved with fingerprints that take
account of the structural basis of toxicity for specific endpoints. Fingerprints must be
chosen and interpreted such that they avoid pitfalls such as activity cliffs i.e. the selection
of close structural analogues, according to the fingerprint derived similarity measure,
which have different activity due to the effect of structural change on the MIE.

Whilst a justifiable means of identifying analogues, the use of the MIE is only appropriate
to relevant toxicological endpoints, i.e. where the MIE is known, and identifying the MIE
is only one step in the overall read-across process, which may involve the collation of
multiple lines of evidence.

Fingerprint-derived measures of similarity should be used to identify analogues for readacross for large heterogeneous datasets with caution, unless the similarity measures can be
shown to clearly relate to biologically relevant structural variation and not to capture
biologically irrelevant variation. Where they are known, this justification should be made

with reference to relevant mechanism(s) of action, for instance relating to the MIE.
However, generic fingerprint similarity measures do not fulfil these criteria, so must be
used with caution for large, chemically diverse datasets.

Arguably, the most suitable use of generic fingerprint-derived similarity measures for read-554 across within large, chemically diverse datasets is following sub-categorisation. (However, 555 further work is required to determine the extent to which this yields better predictive 556 performance than integrating these similarity measures within machine learning 557 approaches, which have recently been advocated. Moreover, sub-categorisation which 558 removes biologically irrelevant structural variation may result in the fingerprint-derived 559 similarity measures being optimally predictive, yet redundant if read-across is performed 560 by expert examination of the structures within the category.) Sub-categorisation should 561 preferably be performed using a mechanistically based method. If sub-categorisation 562 yields homologous series, acting via a common mechanism, for which all the structural 563 variation is expected to be biologically relevant, generic fingerprint-derived similarity 564 measures could be suitable for fine tuning and confirming analogue identification for read-565 566 across.

However, even within categories of chemicals acting via a common mechanism, the use
 of alternative similarity measures, based upon mechanistic understanding of the endpoint
 of interest, should be considered for read-across purposes. For example, similarity
 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.

575

576 6. ACKNOWLEDGEMENTS

577 The authors would like to thank Dr David Roberts, Liverpool John Moores University, for578 sharing his valuable knowledge of skin sensitisation reaction chemistry.

579

580 7. REFERENCES

- 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.
- 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
- fingerprint-based similarity calculations? J. Cheminf. 7, 20.

- 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.,
- 599Zhu, H., Hartung, T., 2016. Toward Good Read-Across Practice (GRAP) guidance. ALTEX,
- 600 <u>33</u>, 149-166.
- Bender, A., Glen, R.C., 2004. Molecular similarity: a key technique in molecular informatics.
 Org. Biomol. Chem. 2, 3204-3218.
- Berggren, E., Amcoff, P., Benigni, R., Blackburn, K., Carney, E., Cronin, M., Deluyker, H.,
- Gautier, F., Judson, R.S., Kass, G.E.N., Keller, D., Knight, D., Lilienblum, W., Mahony, C.,
- Rusyn, I., Schultz, T., Schwarz, M., Schüürmann, G., White, A., Burton, J., Lostia, A.M.,
- Munn, S., Worth, A., 2015. Chemical safety assessment using read-across: How can novel
- 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.,
- 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, Cambridge615 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.

- Cronin, M.T.D., Enoch, S.J., Mellor, C.L., Przybylak, K.R., Richarz, A.-N., Madden, J.C.,
 2017. *In silico* prediction of organ level toxicity: linking chemistry to adverse effects.
 Toxicological Res. 33,173-182.
- 621 Cruz-Monteagudo, M., Medina-Franco, J.L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro,
- 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
- fragment descriptors: side effects on sampling in bioactive libraries. J. Med. Chem. 42, 2887–
 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.
- 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.
- 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.
- Enoch, S.J, Cronin, M.T.D., Madden, J.C., Hewitt, M., 2009. Formation of structural
- 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
- reaction chemistry involved in covalent protein binding relevant to toxicity. Crit. Rev.
- 639 Toxicol. 41, 783-802.

- 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
- of skin sensitization alternative methods. Dermatitis. 16, 157-202.
- Gini, G., Franchi, A.M., Manganaro, A., Golbamaki, A., Benfenati, E., 2014. ToxRead: A tool
 to assist in read across and its use to assess mutagenicity of chemicals. SAR QSAR Environ.
 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.
- Guha, R., van Drie, J.H., 2008. Structure-activity landscape index: Identifying and
 quantifying activity cliffs. J. Chem. Inf. Model. 48, 646–658.
- Hartung, T., 2016. Making big sense from big data in toxicology by read-across. ALTEX. 33,83-93.
- 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.
- Holliday, J.D., Hu, C.-Y., Willett, P., 2003. Analysis and display of the size dependence of
 chemical similarity coefficients. J. Chem. Inf. Comput. Sci. 43, 819–828.
- Lajiness, M.S., 1997. Dissimilarity-based compound selection techniques. Perspect. Drug
 Discov. Des. 7/8, 65–84.

- Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016a. Global
 analysis of publicly available safety data for 9,801 substances registered under REACH from
 2008-2014. ALTEX. 33, 95–109.
- Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016b. Analysis
 of public oral toxicity data from REACH registrations 2008-2014. ALTEX. 33, 111–122.
- Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016c. Analysis
 of Draize eye irritation testing and its prediction by mining publicly available 2008-2014
 REACH data. ALTEX. 33, 123–134.
- Luechtefeld, T., Maertens, A., Russo, D.P., Rovida, C., Zhu, H., Hartung, T., 2016d. Analysis
 of publically available skin sensitization data from REACH registrations 2008-2014. ALTEX.
 33, 135–148.
- Luechtefeld, T., March, D., Rowlands, C., Hartung, T., 2018. Machine learning of toxicological
 big data enables read-across structure activity relationships (RASAR) outperforming animal
 test reproducibility. Toxicol Sci. 165, 198-212.
- Maggiora, M. Vogt, M., Stumpfe, D., Bajorath, J., 2014. Molecular similarity in medicinal
 chemistry. J. Med. Chem. 57, 3186–3204.
- 678 Martin, Y.C., Kofron, J.L., Traphagen, L.M., 2002. Do structurally similar compounds have
- similar biological activity? J. Med. Chem. 45, 4350–4358.
- 680 Maunz, A., Helma, C., 2008. Prediction of chemical toxicity with local support vector
- regression and activity-specific kernels. SAR QSAR Environ. Res. 19, 413-431.

- Mellor, C.L., Schultz, T.W., Przybylak, K.R., Richarz, A.-N., Bradbury, S.P., Cronin, M.T.Da.,
 2017. Read-across for rat oral gavage repeated-dose toxicity for short-chain monoalkylphenols: A case study. Comput. Toxicol. 2, 1-11.
- Nelms, M.D., Ates, G., Madden, J.C., Vinken, M., Cronin, M.T.D., Rogiers, V., Enoch, S.J.,
- 2015. Proposal of an *in silico* profiler for categorisation of repeat dose toxicity data of hair
 dyes. Arch. Toxicol. 89, 733-741.

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

- Grouping of Chemicals, 2nd Edition, Series on Testing and Assessment No. 194. OECD,
 Paris, France.
- Patlewicz, G., Jeliazkova, N., Gallegos-Saliner A., Worth, A.P., 2008. Toxmatch a new
 software tool to aid in the development and evaluation of chemically similar groups. SAR
 QSAR Environ. Res. 19, 397-412.
- Patlewicz, G., Ball, N., Boogaard, P.J., Becker, R.A., Hubesch, B., 2015. Building scientific
 confidence in the development and evaluation of read-across. Reg. Toxicol. Pharmacol. 72,
 117-133.
- Patlewicz, G., Helman, G., Pradeep, P., Shah, I., 2017. Navigating through the minefield of
 read-across tools: A review of in silico tools for grouping. Comput. Toxicol. 3, 1-18
- Patlewicz, G., Cronin, M.T.D., Helman, G., Lambert, J.C., Lizarraga, L.E., Shah I., 2018.
 Navigating through the minefield of read-across frameworks: A commentary perspective.
 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.,
- 2017. Read-across of 90-day rat oral repeated-dose toxicity: A case study for selected β-olefinic
 alcohols. Comput. Toxicol. 1, 22-32.
- Reisen, F., Zhang, X., Gabriel, D., Selzer, P., 2013. Benchmarking of multivariate similarity
- measures for high-content screening fingerprints in phenotypic drug discovery. J. Biomol.Screen. 18, 1284–1297.
- 708 Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I.,
- 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.
- Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50,
 742–754.
- Rosenbaum, L., Hinselmann, G., Jahn, A., Zell A., 2011. Interpreting linear support vector
- machine models with heat map molecule coloring. J. Cheminform. 3, 11.
- Salim, N., Holliday, J., Willett, P., 2003. Combination of fingerprint-based similarity
 coefficients using data fusion. J. Chem. Inf. Comput. Sci. 43, 435–442.
- Schultz, T.W., Cronin M.T.D., 2017. Lessons learned from read-across case studies for
 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.,
- Schwarz, M., White, A., Cronin, M.T.D., 2015. A strategy for structuring and reporting a
- 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
- 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:
- knowledge gained from case studies. Comput. Toxicol. submitted
- 729 Shah, I., Liu, J., Judson, R.S., Thomas, R.S., Patlewicz, G., 2016. Systematically evaluating
- read-across prediction and performance using a local validity approach characterized by
- 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
- reports on compliance with the REACH and CLP regulations and on the use of alternatives to
- testing on animals for compliance with the REACH regulation. ATLA Altern. Lab. Anim,
 39, 481-493.
- Stumpfe, D., Bajorath, J., 2011. Similarity searching. Wiley Interdiscip. Rev.: Comput. Mol.
 Sci. 1, 260–282.
- Teubner, W., Landsiedel, R., 2015. Read-across for hazard assessment: the ugly duckling is
 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
- 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
- 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
- of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients J.
- 748 Chem. Inf. Comput. Sci. 44, 1840–1848.
- Willett, P., 2013. Combination of similarity rankings using data fusion. J. Chem. Inf. Model.
 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.

Table 1. Definitions of terms using in this investigation.

Term	Definitions used for this study
Analogue (for read-	A similar compound, with measured endpoint data, to that for
across)	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 between two molecules calculated from
molecular similarity	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
Wolceular similarity	The similarity, of degree of overlap, between two of 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.

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 <i>Tetrahymena pyriformis</i> . 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			metabolism or the reactivity of the metabolite is	
	acrolein			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.	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
		1.00.1	0.02.1	0.01.1	0.01.1	0.01.1	0.01.1
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
DELLA						1 00 1	1.00.1
PFUA						1.00-1	1.00-1
PFDoA							1.00-1
							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 (PFDA).

	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-IsopropyI-5-methylphenol (thymol)	2-Methyl-5-isopropylphenol (carvacrol)	3-Methyl-6-n-butylphenol	2-Ethyl-5-methylphenol	2-IsopropyIphenol	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- 0.91	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

Table 4: Shows the range of the Tc scores calculated when utilising the different fingerprints for the alkylphenols dataset (dataset 6).

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.

Chemica	als Compared	Shows	Tc Scores a	nd the finge	rprint used to	o calculate	e Tc.	Range of Tc
(LLNA score, sensitiser clas	sification (Gerberick et al., 2005))	CDK Standard	CDK MACCS	CDK Extended	CDK PubChem	CDK FCFP6	CDK ECFP4	across fingerprints
1,4- dihydroxyquinone (0.1, strong sensitiser)	Resorcinol (5.0, non-sensitiser) 아							
OH OH OH	HO	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

H ₂ N	NH ₂							
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.

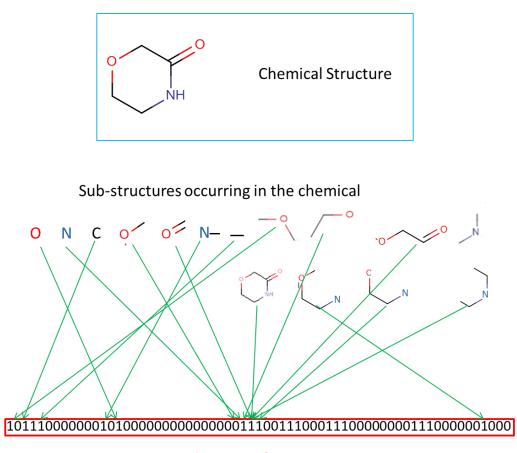
Figure 3: Shows overview of the Tc similarity matrices for the alkylphenols dataset (dataset 6), in terms of each of the computed fingerprints: (A) CDK Standard fingerprints; (B) CDK MACCS fingerprints; (C) CDK Extended fingerprints; (D) CDK PubChem fingerprints; (E) CDK FCFP6 fingerprints; (F) CDK ECFP4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





Bit value set in fingerprint

	C DK 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 N	ACCS			
	DELLA					DELLA	
	PFHX/	РЕН РА	PFUA	PFINA	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									
	PFHx	PFHp	DEOA		PFDA	PFUA	PFDo		
	Α	Α	FFUA	FFINA			Α		
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								
	PFHx	PFHp	DEOA			D.C.I.A	PFDo	
	Α	A		PFOA PFNA	PFUA	PFUA	Α	
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)

		C	DK FC	FP6			
	PFHx	PFHp	DFO 4			DELLA	PFDo
	Α	Α	PFUA	PFNA	PFDA	PFUA	Α
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								
	PFHx	PFHp	PFHp PFOA			DELLA	PFDo	
	Α	Α	PTUA	PTINA	PTUA	PFUA	Α	
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

	C0135mded		77 W/73	COX Extended
(A)	4 cont: hony 2 analyt parent 3 cont: hony 2 analyt parent 3 cont: hony 4 analyt 3 control proper physics 3 control physics 3		A cent, huny 2 wasted pie ref a cent, huny 2 wasted pie ref 3 di Borris and a cent 3 di Borris and a central a 3 di Borris and a central a central 4 di Borris and a central a central 3 di Borris and a central a central 4 di Borris and 4 di Borri and 4 di Borris and 4 di Borris and 4 di Borris and 4 di	4 cort Lang 2 and by 2 and 2 and 3 and
2-tert.Subyl 8-methylphenel 2-tert-Subyl-4-methylphenel	120 0.77 0.77 0.87 0.82 0.85 0.71 0.84 0.39 0.85 0.85 0.73 0.82 0.70 0.85 0.24 0.71 0.85 0.71 0.77 1.00 0.21 0.77 0.87 0.82 0.73 0.86 0.81 0.74 0.85 0.82 0.73 0.88 0.90 0.89 0.81 0.87 0.89 0.89	361.54/340/#end 10 10 10 10 10		2-ler6su0/-3-mech/phenal 1.00 0.75 0.79 0.85 0.87 0.59 0.72 0.85 0.59 0.85 0.71 0.94 0.82 0.85 0.71 0.54 0.70 0.71 0.71 2-ler6su0/-4-mech/phenal 0.75 1.00 0.79 0.78 0.85 0.85 0.72 0.85 0.25 0.72 0.87 0.35 0.72 0.81 0.35 0.55 0.57 0.82 0.89 0.28
2-tot-54yletenel	0.77 0.51 1.00 0.55 0.79 0.55 0.75 0.55 0.50 0.54 0.55 0.57 0.75 0.55 0.57 0.54 0.54 0.51 0.51 0.52 0.75	3ct 42(dec) 10 10 10 10 1		2-ter-Sub/phenal 0.79 0.79 0.79 0.80 0.85 0.85 0.75 0.65 0.75 0.65 0.75 0.85 0.57 0.75 0.85 0.85 0.85 0.85 0.85 0.87 0.88
2,8-di-tert-Butylpherel	0.57 0.77 0.55 1.00 0.59 0.59 0.54 0.57 0.55 0.54 0.59 0.50 0.55 0.75 0.75 0.58 0.54 0.59 0.70 0.59	155251400 10 10 10 10 1		2,5-d-ler-Sulyiphenel 0.65 0.75 0.85 1.00 0.70 0.85 0.64 0.56 0.52 0.65 0.59 0.50 0.65 0.75 0.72 0.47 0.54 0.70 0.87 0.87
2-tot-Amylphenol	0.55 0.57 0.79 0.55 1.00 0.55 0.52 0.52 0.75 0.56 0.50 0.70 0.55 0.78 0.54 <mark>0.47 0.45</mark> 0.55 0.55 0.55	3614mjilovi 18 18 18 18	a ia ia ia <mark>ia ia</mark> ta ia <mark>ia ia ia ia ia ia ia ia</mark> ia ia ia ia	I.GerC.Am/sphenel 0.59 0.55 0.50 0.70 1.00 0.51 0.91 0.79 0.75 0.51 0.59 0.55 0.55 0.52 0.45 0.45 0.45 0.55 0.55
2,4-á a href="https://www.statics.com">a href="https://www.statics.com"/>a href="https://www.statics.com"/>a href="https://www.statics.com"/>a href="https://www.statics.com"/>a	036 0.65 0.65 0.69 0.53 1.00 0.75 0.71 0.66 0.55 0.52 0.61 0.56 0.65 0.69 0.40 0.40 0.55 0.60 0.62	245254mjded 1010101	a in in in <mark>th the la by the the la the la by the laboration and the laboration of the laboration of</mark>	2,4-d-3cr6-Amplphenel 0.59 0.65 0.65 0.65 0.81 100 0.74 0.65 0.82 0.55 0.55 0.57 0.60 0.70 0.40 0.59 0.55 0.52 0.55
2-sco-Sul) (pitenol	0.71 0.75 0.75 0.84 0.92 0.78 1.00 0.89 0.82 0.75 0.87 0.77 0.72 0.88 0.71 0.54 0.85 0.81 0.81 0.89	Sectoriation DA DA DA DA DA		2-sec-8ut/phenol 0.72 0.72 0.73 0.84 0.91 0.74 1.00 0.87 0.50 0.74 0.65 0.75 0.74 0.85 0.70 0.55 0.55 0.65 0.65
2-o-Sut)(phere) 2-o-Pentylpheno I	054 0.55 0.57 0.52 0.71 0.59 1.00 0.92 0.56 0.51 0.58 0.55 0.74 0.55 0.46 0.49 0.55 0.40 0.45 0.55 0.40 0.55 0.55 0.55	Defadyleten Det		2>du/(phena) 0.55 0.55 0.56 0.57 0.56 0.57 1.00 092 0.55 0.50 0.25 0.55 0.55 0.55 0.45 0.47 0.54 0.55 0.50 2.50 2.57 1.00 092 0.55 0.55 0.55 0.55 0.45 0.55 0.55 0.55
2-isop ropyl-5-meth y phen ol	055 0.1 050 0.5 0.5 0.5 0.5 0.5 0.5 1.0 0.1 100 0.1 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5	Proping in the second s		2-isoprop/-5-methylphend 0.55 0.55 0.55 0.55 0.55 0.55 0.55 0.5
2-Meth vi-5-lso pro pvibh enol	045 0.4 0.4 0.4 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5	Web (Starts idea)		2-Mathyf-S-Isopropy(phenol ass ast ass ass ass ass ass ass ass ass
3-Meth yl-6-n-butyl phen ol	0.73 0.55 0.57 0.50 0.70 0.61 0.77 0.55 0.50 0.77 0.71 1.00 0.76 0.55 0.55 0.55 0.45 0.55 0.51 0.55	Histyler-tutylpand tes ces ces		3-Methy+6-n-bub/phenol 0.71 0.55 0.57 0.50 0.59 0.50 0.75 0.55 0.79 0.75 0.69 1.00 0.74 0.54 0.55 0.55 0.55 0.51 0.52 0.57
2-Eth yl-5-meth ylph enol	035 0.75 0.75 0.85 0.85 0.75 0.85 0.72 0.85 0.80 0.89 0.91 0.76 1.00 0.84 0.71 0.72 0.85 0.85 0.85 0.71	18 hiff-mathighend ans are are ar		2-8th/t-5-meth/tphenol 0.54 0.72 0.75 0.65 0.65 0.57 0.74 0.65 0.60 0.59 0.29 0.29 0.24 0.70 0.75 0.56 0.65 0.67 0.72
2-isop ropyl phen ol	0.55 0.55 0.55 0.75 0.55 0.55 0.55 0.55	Peoplagistan di ser ser ser ser	8 13 13 17 <mark>19 19</mark> 18 18 <mark>10 1</mark> 3 20 18 19 19 18 18 18	2-issprapy(pheno) 0.52 0.51 0.55 0.76 0.74 0.60 0.55 0.75 0.57 0.55 0.76 0.64 0.54 1.00 0.79 0.64 0.65 0.72 0.76 0.75
2,4-Disopropylpherol	0.70 0.80 0.75 0.78 0.84 0.89 0.71 0.88 0.81 0.72 0.88 0.82 0.71 0.88 1.00 0.85 0.80 0.81 0.88 0.89			2,4-Ciargerapy(phane) 0.55 0.55 0.55 0.57 0.57 0.57 0.57 0.57
1,5-Directhylphonol	0.59 0.55 0.54 0.45 0.47 0.40 0.54 0.49 0.45 0.71 0.71 0.55 0.72 0.55 0.55 1.00 0.73 0.55 0.51 0.59	13 Dechglorel 104 104 104 104		2.5-0imethylphenel 0.71 0.55 0.55 0.47 0.45 0.40 0.55 0.47 0.45 0.44 0.75 0.55 0.75 0.54 0.52 1.00 0.74 0.55 0.55 0.57 0.55 0.55 0.55 0.55 0.55
1,8-Directhylphanel 3-tat-butylphanel	034 0.51 0.54 0.54 0.54 0.45 0.40 0.55 0.40 0.55 0.45 0.55 0.45 0.55 0.5	1954chiptori 19 19 19 19 19 19 19		2,5-Cimethylghemel 0.54 0.57 0.55 0.54 0.44 0.53 0.55 0.47 0.45 0.55 0.54 0.45 0.55 0.55 0.57 0.74 100 0.47 0.47 0.52 3-CerC-bullghemel 0.70 0.65 0.51 0.70 0.67 0.55 0.61 0.54 0.50 0.67 0.75 0.51 0.66 0.72 0.60 0.55 0.47 1.00 0.55
4-tot-5dyletenel	011 050 051 050 050 050 051 051 050 050	4cr42/gbrd 10 10 10 10		4-ter-Sulyiphenel 0.71 0.65 0.17 0.65 0.51 0.55 0.51 0.55 0.55 0.55 0.55 0.5
4-tot-6dy-2-methylphonel	0.71 0.89 0.75 0.89 0.51 0.52 0.59 0.55 0.55 0.71 0.72 0.55 0.71 0.52 0.59 0.59 0.55 0.55 0.74 1.00			4-tert-8-uty-1-methylehenel 0.71 0.88 0.89 0.87 0.39 083 0.88 0.88 0.88 0.88 0.38 0.37 0.72 0.77 0.72 0.78 0.88 0.89 0.82 0.85 0.75 1.00
	CDC Pablices		COKYCYPS	COX 80774
(D)	2 form: barry 2 method barrel 2 form: barry 2 method barrel 2 form: barry 2 method barrel 2 form: barry 4 menil	(E)	4 cor, huy 2 worky 2	(F)
5-tart 54) i S-methylphemi	d con hay 2 antisfaeni ay 4 antisfaeni ay 5 an	2-tert Subj-Smethylphenel	4 corr bay's methyl bend 3 corr bay's methyl bend 3 corr bay's methyl bend 3 corr bay's methyl bend 3 corr bay 4 methy	Re + Courty + Ferral Re + Cour
2-larl. Sul-1-5-mcHypherel 2-larl-Sul-1-4-mcHypherel	4 cont. Key 2 webdy kenni 4 cont. Key 3 kenni 3 cont. Key 3 kenni 3,5 control y kenni 3,5 control y kenni 3,4 thug 5 mer hyl ble od 3,4 thug 6 mer hyl ble od 4,4 thug 6 mer hyl ble od 4,4 thug 6 mer hyl ble od 5 mer hyl ble od 4 mer hyl ble od 5 mer hyl ble od 5 mer hyl ble od 5 mer hyl ble od 4 mer hyl ble od 5 mer hyl	2-lor: 50/j-5-roth/jelona 2-lor: 50/j-6-roth/jelona	4 core knyty ke mit 3 ke - knyty ke mit 3, de kar y ky ky ke mit 3, de kar y ky ke mit 4, de ky ky ke mit 4, de ky ky ky ke mit 4, de ky	1 1
2-lor: Sub 1-5-mcHujelenel 2-lor: Sub 1-4-mcHujelenel 2-lor: Sub Jelenel	d con hay 2 antisfaeni ay 4 antisfaeni ay 5 an	2-tert Subj-Smethylphenel	4 core havy3 mentyl head 3 core individual 3 core inditindividua	Local And Analysis Local Analysis
2-12-50-1-5-mc1hiptonel 2-12-50-1-1-4-mc1hiptonel 2-12-50-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-	d con hay 2 antisfaeni ay 4 antisfaeni ay 5 an	2.655 80/1-8=551/piteral 2.655 80/1-8=551/piteral 2.655 80/10001 2.655 4075 81/piteral 2.655 4075 91/piteral	4 ctor Link <	Local Anglemal Local Anglemal <thlocal anglemal<="" th=""> Local An</thlocal>
2-101-544) 1-5-mc5hipherel 2-101-544) 1-4-mc5hipherel 2-101-544) 1-61-mai 2-5-1-544) Iphenel 2-101-544, Iphenel 2-4-4-1-11-1440 Jacob	d con hay 2 antisfaeni ay 4 antisfaeni ay 5 an	2-5ct 5:0/-5mcth/phres 2-5ct 5:0/-6mcth/phres 2-5ct 5:0/-6mcth/phres 2-5ct 5:0/-6mcth/ 2-5ct 5:0/-6mcth/ 2-5ct 5:0/-6mcth/ 2-6ct 5:0/-6mcth/ 2-6ct 5:0/-6mcth/ 2-6ct 5:0/-6mcth/phres	4 crist bary2 analysis and 3 c to max bary3 and 3 c to max bary3 and 3 c to max bary3 and 3 c to bary 4 c to max bary 1 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 bened 3 c to bary 4 c to bary 4 c to bary 4 bened 3 c to bary 4 c to	Lottauj-15-methylemel 100 0.00<
Dat Sal-Sechipten Jar-Sal-Sechipten Dat-Sal-Sechipten Za-Sac-Sal-Steel Za-Sac-Sal-Steel Za-Sac-Sal-Steel Za-Sac-Sal-Steel	d con hay 2 antisfaeni ay 4 antisfaeni ay 5 an	2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord	4 ctor 4 ctor<	Local and Analysis Local Analy
20ar. Saki Swethelsen 20ar. Saki Swethelsen		2-5ct 8.0/-3-mcH/phrei 2-5ct 8.0/-4-mcH/phrei 2-5ct 8.0/4-mcH/phrei 2-5ct-8.0/4-phrei 2-5ct-8-1/4-phrei 2-8-50/4-phrei 2-8-50/4-md	4 corr. brightendi 3 corr. brightendi 3 corr. brightendi 3 corr. brightendi 3 corr. brightendi	Local_0[46-mc0]/gitterel Local_0[46-mc0]/gitterel <thlocal_0[46-mc0] gitterel<="" th=""> <thlocal_0[4< td=""></thlocal_0[4<></thlocal_0[46-mc0]>
Side Sdui-Seechighenel Side Sdui-Hendhighenel Side Sduishenel Sde Sduighenel Sde Sduighenel Saedunghenel Saedunghenel Saedunghenel	4 con bay 2 antivitation 20 </td <td>2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord</td> <td>4 ctor 4 ctor<</td> <td>Local And Composition Local <thlocal< th=""> Local Local</thlocal<></td>	2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Smc1/ Jahnel 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gort 20/3-Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord 2 Gord	4 ctor 4 ctor<	Local And Composition Local Local <thlocal< th=""> Local Local</thlocal<>
24st 504 Seechighteel 24st 504 Seechighteel 24st 504 Seechigteel 25st 504 Seechigteel 25st 504 Seechigteel 24st 504 Seechigteel 25st 504 Seechigteel 25st 504 Seechigteel 25st 504 Seechigteel 25st 504 Seechigteel 25st 504 Seechigteel	4 con bay 2 antivitation 20 </td <td>2 der 50/3-meth/piterel 2 der 50/1-meth/piterel 2 der 60/14 meth/piterel 2 der 64 der 14 mig berei 2 der 64 mig berei 2 aus 60/14 met 2 aus 60/14 met</td> <td>4 ctor k-y-2 - an-xy/yke mil 3 ctor k-y-2 - an-xy/yke mil 3 ctor 3 ctor<</td> <td>2 2</td>	2 der 50/3-meth/piterel 2 der 50/1-meth/piterel 2 der 60/14 meth/piterel 2 der 64 der 14 mig berei 2 der 64 mig berei 2 aus 60/14 met 2 aus 60/14 met	4 ctor k-y-2 - an-xy/yke mil 3 ctor k-y-2 - an-xy/yke mil 3 ctor 3 ctor<	2 2
Side Sdui-Seechighenel Side Sdui-Hendhighenel Side Sduishenel Sde Sduighenel Sde Sduighenel Saedunghenel Saedunghenel Saedunghenel		24ct 8/(-3+cth/phrm) 24ct 8/(-4+cth/phrm) 24ct 8/(-4+cth/phrm) 24ct 8/(-4+cth/phrm) 24ct 8/(-4+cth/phrm) 24ct 8/(-4+cth/phrm) 24ct 8/(-4+cth/phrm) 2-4-cth/phrm) 2-4-cth/phrm)	4 - et al. bary 2 - arenytyberedi 3 - be - barytyberedi 3 - be - barytyberedi<	Location (-5-methylene) Los Location (-5-methylene) Location (-5-methylene) <thlocation (-5-methylene)<="" th=""> <thlocation (-5-<="" td=""></thlocation></thlocation>
2-fact. 54/-i-Sect:helicerel 2-fact. 54/-i-Sect:helicerel 2-fact. 54/-i-Sectional 2-fact. 54/-i-Sectio	4 con b con kay 2 a constant wang 8 <t< td=""><td>2-test 50/-5-meth/planel 2-test 50/-5-meth/planel 2-test 50/-14-meth/planel 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth/planel 2-test 50/-3-meth/planel</td><td>$\begin{array}{ c c c c c c c c c c c c c c c c c c c$</td><td>Local Applied Company Company</td></t<>	2-test 50/-5-meth/planel 2-test 50/-5-meth/planel 2-test 50/-14-meth/planel 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth 2-test 50/-14/meth/planel 2-test 50/-3-meth/planel	$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Local Applied Company
2-tot: 50-15-exc2highted 2-tot-50-15-exc2highted 2-tot-50-16-bight		5 Sach Sul - Samah Jahon I 2 Sach Sul - Ameth Jahon I 2 Sach Jahon Jahon I 2 Sach Jahon Jahon I 2 Sach Jahon Jahon Jahon I 2 Sach Jahon Jahon Jahon J 2 Sach Jahon J 2 Sach Jahon J 2 Sach Jahon J 2 Sach Jahon Jahon J 2 Sach Jahon Jahon J 2 Sach Jahon Jahon J 2 Sach Jahon J 2 Sach Jahon J 2 Sach Jahon Jahon J 2 Sach Jahon Jahon J 2 Sach Jahon Jah	4 - et - kury 2 - mery (he ord) 3 - e - kury 2 - mery	Local of A-mathylebroil Source A-mathylebroil
2-5art 3-6-1-5-eschleiten 2-5art 3-6-1-4-eschleiten 2-5art 3-6-2-5-1-4-eschleiten 2-5art 3-6-2-5-1-4-1-1-2-5-5-1-2-5-2-5		2-5ct 5:0/-5mcH/piters 2-5ct 5:0/-4mcH/piters 2-5ct 5:0/4bcd 2-5ct 5:0/4bcd 2-5ct 5:0/4bcd 2-5ct 4:0/4bcd 2-5ct 4:0/4bcd	$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Locata Add Add<
2-tot: 50-1-5-exc2highted 2-tot-50-1-6-exc2highted 2-tot-50-2-bighted		5 Sach Sul - Samah Jahon I 2 Sach Sul - Amerik Jahon I 2 Sach Sul - Amerik Jahon I 2 Sach Sul - Sach Jahon I 2 Sach Sul - Jahon I 2 Sach Jahon Jahon I 2 Sach Jahon	4 et e. hury he with 3 et e.	Local dividend State
2.5ar.5.84.15-erc.fujel.cerj 2.5ar.5.84.14-er.dh.jeteral 2.5ar.5.84.14-er.dh.jeteral 2.5ar.5.84.14-er.dh.jeteral 2.5ar.5.84.14-14-14-14-14-14-14-14-14-14-14-14-14-1		2-5ct 5:0/-5mcH/plenel 2-5ct 5:0/-4mcH/plenel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/4bcel 2-5ct 5:0/2bcel 2-5ct 5:0/2bcel	$ \begin{array}{ $	Locate Applies Applies <th< td=""></th<>
2-tot: 50-1-5-exc2highted 2-tot-50-1-6-exc2highted 2-tot-50-2-bighted		5 Sach Sul - Samah Jahon I 2 Sach Sul - Amerik Jahon I 2 Sach Sul - Amerik Jahon I 2 Sach Sul - Sach Jahon I 2 Sach Sul - Jahon I 2 Sach Jahon Jahon I 2 Sach Jahon	4 et e. hury he with 3 et e.	Local of A-mathyle Control of
2-tot: 50-15-excMpB cell 2-tot: 50-15-excMpB cell 2-tot-50-16-bit Merel 2-tot-50-16-bit Merel		2455.00/30mcH/phonel 2455.00/30mcH/phonel 2456.00/30mcH	4 et e. hury he ent 3 et e. hury he ent	Local_u 4>-mcl/ylemi Loc Loc <thloc< th=""> Loc <thloc< th=""></thloc<></thloc<>

Figure 4

	CDK Standard		COKMACCS	CDX Extended
(A)	2. Methyl + undecannol 2. Propylnepartan + 1 2. Propylnepartan + 1 35. Timethyl + necanol 2. Bethyl + nectanol 2. Methyl + nectanol 3. Methyl + butanol 3. Methyl + butanol 3. Methyl + butanol 3. Methyl + butanol 4. Prodecanol 4. Prodecanol 4. Undecanol 4. Undecanol 4. Necanol 4. Necanol 4. Hectanol 4. Hectanol 4. Hectanol 4. Hectanol 4. Hectanol	(B)	2 Methyl-1 unde canol 3.7 Dimethyl 1 o danol 2.P ropylhe pantan 1 of 3.6,6 Trimethyl 1 hexanol 2.Ethyl 1 hexanol 3.Methyl 1 hertanol 3.Methyl 1 hertanol 3.Methyl 1 hertanol 4. Tridecanol 1.Doecanol 4. Undecanol 1.Decanol 4. Undecanol 1. Hexanol 1. Hexanol 1. Hexanol 1. Hexanol	2 Methyl 1-undecanol 3,7 Dimethyl 1-oda nol 2 Propythepartan 1-ol 3,62 Timethyl 1-beanol 6 Methyl 1-beanol 3 Methyl 1-butanol 3 Methyl 1-butanol 1 Tidecanol 1 Tidecanol 1 Didecanol 1 Didecanol 1 Didecanol 1 Didecanol 1 Didecanol 1 Didecanol 1 Hecanol 1 Hecanol 1 Hecanol 1 Hecanol 1 Hecanol 1 Heptanol 1 Heptanol
1-Pertand 1-Heard 1-Heard 1-Otarol 1-Otarol 1-Doleand 1-Troleand 3-Toroleand 3-Heart-1-board 3-Heart-1-board 3-Heart-1-board 3-Heart-1-Heard 2-Bryheard 2-Bryhearta-1-d 2-Bryhearta-1-Heard 2-Bryhearta-1-Heard	070 0.84 0.93 0.96 0.98 0.98 0.98 0.98 0.98 0.57 0.57 0.70 0.64 0.93 0.89 0.80 1.00 0.96 0.98	1-Hestanol 1-Octanol 1-Decanol 1-Decanol 1-Decleanol 1-Decleanol 2-Methyl-1-Octanol 3-Methyl-1-Octanol 3-Methyl-1-Octanol 2-Bhyl-1-Octanol 2-Bhyl-1-Boctanol 3-Bethyl-1-Bectanol 3-Bethyl-1-Bectanol 3-Bethyl-1-Bectanol 3-Bethyl-B	956 100 <th>Personal 100 0.66 0.76 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.82 0.82 0.70 0.81 0.82 0.71 0.70 0.81</th>	Personal 100 0.66 0.76 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.82 0.82 0.70 0.81 0.82 0.71 0.70 0.81
3,7-Dimethy H1-oct and	070 084 093 100 096 098 098 098 098 098 097 057 070 0.64 0.93 0.89 0.80 0.96 100 0.98 069 082 091 098 100 100 100 100 100 056 055 069 0.62 0.91 0.67 0.78 0.98 0.98 100	3,7-Dimethy I-1-octanol		7-Dimethyl-schand 0.72 0.83 0.94 1.00 0.98 0.98 0.98 0.98 0.98 0.58 0.58 0.58 0.72 0.64 0.94 0.89 0.79 0.96 1.00 0.98 Methyl-undeand 0.70 0.81 0.93 0.98 1.00 1.00 1.00 1.00 0.57 0.57 0.70 0.63 0.93 0.87 0.78 0.98 0.98 1.00
21/eth/1-urdecand	2. Avethyst - Lundscanol 3.7. Clinidityst - Lectanol 2. Propythepanitar. Foil 3.6.5. "Immethyst - Inegranol 3.4.6." Nethyst - Inegranol 3.4.6."thyst - Inegranol 3.4.6."thyst - Indicennol 1. Tridecennol 1. Decemnol 1. Decemnol 1. Decemnol 1 Decemno	(E)	2-Methyl 1-undecanol 3,7-Cinnethyl 1-chand 2-Propylhepantar. 1-ol 3,6,5-Titmethyl 1-beanol 5-Methyl 1-beanol 3-Methyl 1-beanol 3-Methyl 1-beanol 1-Tridecanol 1-Dedecanol 1-Dedecanol 1-Decanol 1-beanol 1-beanol 1-beanol	2 - Methyl - Lundecarrol 3,7: Crmethyl - Loctarol 2: Popylineganitari - Loi 3,6: 5: Trimethyl - Loctarol 3,6: 5: Trimethyl - Loctarol 3,6: 5: Trimethyl - Loctarol 3: Bhyle seniol 1: Tridecarrol 1: Nonarol 1: Octarol 1: Heplanol 1: Heplanol 1: Heplanol 1: Heplanol
1-Pentano I	1.00 0.91 0.81 0.75 0.75 0.75 0.75 0.75 0.75 0.79 0.79 0.84 0.80 0.66 0.68 0.64 0.62 0.62 0.64			Pertanol 1.00 0.92 0.92 0.92 0.92 0.92 0.92 0.92 0
1-Hexanol	0.91 100 0.88 0.82 0.82 0.82 0.82 0.82 0.82 0.			Herenal 0.92 100 100 100 100 100 100 100 100 100 24 025 027 025 0.45 035 020 0.39 024 0.45
1-Heptanol 1-Octanol	0.81 0.88 100 0.93 0.93 0.93 0.93 0.93 0.93 0.93 0.			Hagtanol 0.52 100 100 100 100 100 100 100 100 100 024 025 027 025 0.45 035 020 0.35 024 0.45 Octanol 0.52 100 100 100 100 100 100 100 100 024 025 027 025 0.45 035 020 0.35 024 0.45
1-Nonanol	0.75 0.82 0.93 1.00 1.00 1.00 1.00 1.00 0.01 0.01 0.0	1-Nonanol 0.7	9 0.93 100 100 100 100 100 100 100 <mark>0.33 0.41 0.37 0.33 0.58</mark> 0.36 0.30 0.41 0.33 0.48	Moranol 0.92 100 100 100 100 100 100 100 100 024 025 027 025 0.45 035 0.20 0.39 0.24 0.45
1-Decanol	0.75 0.82 0.93 1.00 1.00 1.00 1.00 1.00 0.61 0.61 0.66 0.63 0.82 0.79 0.69 0.82 0.82 0.85			Gerand 0.52 100 100 100 100 100 100 100 100 024 025 027 025 0.45 035 020 0.39 024 0.45
1-Undecanol	0.75 0.82 0.93 100 100 100 100 100 100 0.61 0.61 0.63 0.82 0.79 0.69 0.82 0.82 0.82			Undecanol 0.92 100 100 100 100 100 100 100 100 0.24 0.25 0.27 0.15 0.45 0.35 0.20 0.39 0.24 0.45
1-Dodecanol 1-Tridecanol	0.75 0.82 0.93 1.00 1.00 1.00 1.00 1.00 1.00 0.61 0.61	1-Bode canol 0.7		Codecanol 0.92 100 100 100 100 100 100 100 100 24 025 027 025 0.45 035 020 0.39 024 0.45
2-Methy+1-butanol	0.75 0.82 0.95 100 100 100 100 100 100 100 0.61 0.61 0			Tride can cl 0.52 100 100 100 100 100 100 100 100 100 10
3-Methy+1-butano i	0.79 0.73 0.66 0.61 0.61 0.61 0.61 0.61 0.61 100 100 0.88 0.92 0.69 0.71 0.67 0.65 0.67			Web/y-1-dutahol 012 014 014 014 014 014 014 014 014 014 014
3-Methyl-1-pentanol	0.84 0.78 0.70 0.66 0.66 0.66 0.66 0.66 0.88 0.88 1.00 0.96 0.73 0.81 0.76 0.74 0.74 0.71			Machyl-Courol 100 010 010 010 010 010 010 010 010 01
2-Ethy(+1-butano I	0.80 <mark>0.74</mark> 0.67 0.63 0.63 0.63 0.63 0.63 0.63 0.63 0.92 0.92 0.96 1.00 <mark>0.70</mark> 0.77 0.73 0.71 0.71 0.68	2-Ethy+1-buta nol 0.4	0 035 033 033 033 033 033 033 033 <mark>082</mark> 054 057 <mark>100 037 063 037 050</mark> 038 042 2.	Ethyl+1-butanol 0 25 0 25 0 25 0 25 0 25 0 25 0 25 0 2
6-Methy+1-heptano I	0.66 0.72 0.81 0.82 0.82 0.82 0.82 0.82 0.82 0.82 0.69 0.69 0.73 0.70 1.00 0.91 0.81 0.94 0.94 0.97		19 0.61 0.58 0.58 0.58 0.58 0.58 0.58 0.58 0.37 0.53 0.40 0.37 <mark>1.00</mark> 0.39 <u>0.33 0.43 0.55</u> 0.43	Methy/-1-heptanol 0.40 0.45 0.45 0.45 0.45 0.45 0.45 0.45
2-Ethylhexanol			12 0.38 0.36 0.36 0.36 0.36 0.36 0.36 0.33 0.37 0.40 0.63 0.39 1.00 0.28 0.65 0.36 0.30 2.	
	0.68 0.74 0.78 0.79 0.79 0.79 0.79 0.79 0.79 0.71 0.71 0.81 0.77 0.91 1.00 0.88 0.91 0.91 0.88			Ethyhexanol 0.36 0.35 0.35 0.35 0.35 0.35 0.35 0.35 0.35
3, 5, 5-Trimethyl-1-hexanol	0.64 0.70 0.69 0.69 0.69 0.69 0.69 0.69 0.69 0.6	3,5,5-Trimethyi-1-hexanol 0.3	IS 0.32 0.30 0.30 0.30 0.30 0.30 0.37 0.53 0.56 0.37 0.33 0.28 1.00 0.27 0.48 0.27	5,5-Trimeth y-1-hevanol 021 020 020 020 020 020 020 020 020 020
3, 5,5-Trimethy+1-hexanol 2-Pro py/hepantan-1-ol	064 070 069 069 069 069 069 069 069 069 067 067 076 073 081 088 100 081 081 078 062 068 076 082 082 082 082 082 082 065 065 074 071 094 091 081 100 100 097	3,5,5-Trimethyi-1-hexano I 0.3 2-Propythepantan-1-oI 0.4	s 0.22 0.80 0.80 0.80 0.80 0.80 0.80 0.80	55-Trimeth y-1-henanol 021 020 020 020 020 020 020 020 020 020
3, 5, 5-Trimethyl-1-hexanol	0.64 0.70 0.69 0.69 0.69 0.69 0.69 0.69 0.69 0.6	3,5,5-Trimethyl-1-hexanol 0.3 2-Propyhepantan-1-ol 0.4 3,7-Olmethyl-1-octanol 0.3	5 02 02 02 02 02 02 02 02 02 02 02 02 02	5,5-Trimeth y-1-hevanol 021 020 020 020 020 020 020 020 020 020

Figure 5

	D(Red)		CDK M ACCS	CDX Extended
(A)	THINK 201 SPERV 201 I-PREV 201 I-PREV 201 I-PREV 201 T-BREV 201 SPERV 101 SPERV 201 I-BREV 201 SPERV 201 I-BREV 201	(B)	4.18094.3.01 3.18044.3.01 1.18094.3.01 1.18044.3.01 1.18044.3.01 2.280494.1.01 2.280494.1.01 2.280494.1.01 2.380494.2.01 2.380494.2.01 2.380494.2.01 2.380494.2.01 3.38044.2.0104.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.01044.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.0104.2.01	4 Inny 3 of 3 Ferry 3 of 1 Ferry 3 of 1 Ferry 3 of 1 Ferry 3 of 2 Ferry 1 of 3 F
1-Pitopen-3-ol	100 079 036 048 079 041 048 041 052 046 072 045 052 059 053 059 031 029 024 089 034 027 022 031 028 03		100 075 053 055 059 048 042 041 040 038 052 053 053 053 053 057 059 050 048 038 0.30 030 038 039 039	1-Reguné-ci 100 082 057 04 080 052 050 052 054 047 059 051 049 058 053 038 037 038 031 037 039 038 038 038 033 032
2-Buten-1-ol	075 100 071 024 026 025 026 075 026 028 025 020 057 028 057 075 025 022 010 020 026 020 020 020 020 020 020 020 02		0.75 100 0.72 0.62 0.55 0.41 0.56 0.50 0.47 0.45 0.72 0.82 0.57 0.82 0.44 0.42 0.57 0.80 0.58 0.50 0.38 0.37 0.30 0.35 0.35 0.35 0.35 0.35 0.35 0.35	Dråtter-Bel 082 100 055 054 055 054 055 055 055 055 055 0
2Perten-1-d 2Hexen-1-d	043 054 075 120 056 071 051 051 051 055 077 056 045 055 070 057 073 057 015 015 015 015 015 015 015 015 015 015		CSS 052 076 100 038 042 050 036 048 040 051 051 051 051 052 058 050 054 056 057 0.5 044 055 044 055 054	2-Marge-bel 0.44 0.54 0.78 100 0.55 0.61 0.55 0.71 0.74 0.82 0.40 0.51 0.55 0.54 0.55 0.61 0.12 0.15 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.2
1-Buten-3-ol	0.75 0.66 0.71 0.54 1.00 0.76 0.60 0.76 0.66 0.38 0.75 0.68 0.63 0.71 0.67 0.71 0.26 0.25 0.27 0.28 0.35 0.31 0.38 0.38 0.39 0.		059 059 043 038 100 067 057 064 059 056 058 050 042 050 056 059 033 044 032 038 050 0.47 0.45 0.47 0.44 0.42	1-&tw-Bel 050 055 071 055 100 076 052 057 059 059 059 057 057 057 057 051 050 052 030 039 036 031 038 034 035
1-Perten-3-di	0.61 0.56 0.61 0.54 0.76 1.00 0.79 0.65 0.76 0.71 0.65 0.61 0.60 0.76 0.61 0.22 0.20 0.29 0.38 0.38 0.34 0.27 0.25 0.3	1-Rentan-Biol	0.48 0.41 0.55 0.42 0.67 1.00 0.76 0.50 0.53 0.78 0.48 0.36 0.46 0.36 0.45 0.48 0.37 0.80 0.43 0.88 0.39 0.41 0.57 0.87 0.30 0.42	1-%rfan9-cl 052 055 059 061 078 100 080 054 056 075 069 063 058 055 073 055 025 025 025 025 041 038 032 029 040
1Hexen-3ol	045 045 051 055 050 055 050 054 045 055 055 055 051 045 055 050 018 017 018 010 055 052 056 056 056 056 056 056		042 036 042 060 057 076 100 043 048 059 048 037 041 032 039 038 034 037 033 050 033 0.52 0.71 0.52 038 0.48	1-Were-Bol 050 045 050 055 08 080 100 055 046 057 055 052 045 045 051 045 010 010 015 035 027 035 034
3Perten-2-d 3Hexen-2-d		S 3-Renter-2-ol 3 3-Horan-7-ol	041 050 042 056 054 050 045 100 071 067 042 050 041 050 055 077 031 055 050 <mark>075</mark> 054 0.47 037 0.75 053 0.30 040 047 053 048 059 063 048 071 100 051 041 045 055 048 053 059 032 055 030 058 047 059 043 059 050 056	5-Reference 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
4Herer-30		8 4-Heilen-B-ol	038 045 032 040 056 078 059 067 081 100 039 045 050 045 059 056 030 085 041 081 0.44 0.8 055 0.80 055 0.81	4-Herer-Bel 04T 058 071 022 029 078 057 076 058 100 053 054 070 058 085 082 020 020 020 020 020 020 020 020 020
24Methyl-2-propen-1-d			082 072 053 051 058 048 048 042 041 039 100 079 055 070 045 048 056 058 050 050 033 0.32 038 0.30 030 039	2-Nath/2grogan/2-ol 0.69 0.59 0.51 0.40 0.89 0.55 0.55 0.48 0.55 0.60 0.65 0.65 0.65 0.54 0.54 0.55 0.57 0.58 0.42 0.55 0.58 0.40
2Methyl-2-buten 1-ol			053 082 052 061 050 036 037 050 048 045 079 100 081 089 053 050 056 067 050 050 032 0.32 0.37 036 0.35	2-Waty#74cam-1-cl 051 076 054 051 052 058 052 059 051 054 022 100 079 080 075 070 028 038 031 038 033 035 033 038
2-MEhyl-2perten-1-d 3-Mehyl-2buten-1-d			CSI 057 081 071 042 045 041 041 059 050 085 081 100 073 043 044 045 085 086 080 <mark>088</mark> 081 032 030 048 040 CSI 082 052 054 050 086 082 050 048 045 070 088 073 100 053 088 055 087 050 044 032 082 027 087 086 083	2-Naty-Perturbelei 048 051 053 066 057 052 048 075 079 070 055 075 100 054 051 075 078 028 028 028 037 032 029 034 035 034 055 034 055 056 055 055 055 055 056 056 056 055 055
3-Methyl-200er-10			037 044 038 033 054 055 045 039 088 053 055 045 053 043 053 100 053 028 035 037 034 034 044 033 0.52 047 0.4	5-Mathfridestrivel 053 052 059 059 059 059 059 059 059 059 059 059
44Viethyl-3-penten-2-ol			035 042 036 032 035 048 038 077 039 056 043 050 042 058 092 100 026 030 026 03 050 0.8 032 0.57 044 0.42	4-Natuf-Beartan2el 058 072 082 08 087 055 045 089 077 068 054 070 076 085 080 100 028 077 080 012 085 035 035 035 033
2Picpyn-1-d	031 026 020 016 026 011 018 011 019 017 028 034 020 025 019 022 100 050 041 030 058 030 040 042 037 03		CET CET CEE C48 C33 027 024 C31 C32 C3C C56 C45 C45 C56 028 C38 100 C85 C59 C38 C46 C44 C35 C48 C41 C39	2-Repreiel 037 032 028 021 021 022 021 025 020 024 029 028 028 022 023 028 100 041 042 038 075 055 044 045 039 035
2-8 <i>ut</i> /n-1-ol	213 023 025 025 025 025 025 025 025 025 025 025		CSE 080 058 050 044 030 <mark>028</mark> 035 035 039 058 057 055 067 032 030 085 100 071 080 050 047 038 047 044 042 050 058 083 064 032 048 033 030 050 041 050 050 058 050 <mark>027 035</mark> 059 071 100 075 041 047 039 039 051 030	2*#tm/-tcl 035 030 029 019 019 019 019 019 019 029 029 039 028 029 030 028 017 011 100 017 030 047 037 031 070 029 032 D*#tm/-tcl 031 028 029 029 029 029 039 029 039 029 039 029 039 029 039 029 040 047 100 078 050 041 055 029 029 055
2Pettyn-1-d 2Hexyn-1-d	019 011 011 015 019 019 010 010 010 010 010 010 010 010		CAS CES CES CES CES CES CES CES CES CES CE	Creatives Cal
1-8ut/n-3-ol	034 050 032 032 035 035 035 035 035 035 034 031 031 030 035 034 037 031 058 044 053 040 100 034 052 042 055 04		015 038 032 037 050 039 033 064 047 044 032 032 038 032 034 030 046 030 041 035 100 0.67 050 0.8 052 0.57	1-&t/#5-cl 035 034 034 038 035 037 032 035 028 035 031 031 040 050 035 075 047 050 0.28 100 075 041 059 051 045
1-Pertyn-8-d	027 028 037 038 031 038 032 038 033 038 038 031 038 034 030 033 033 035 030 034 042 047 074 100 079 049 044 04		030 037 038 033 047 061 052 047 053 069 032 032 032 032 041 039 044 047 041 041 057 <mark>1.00 075</mark> 0.52 0.57 0.85	1-Reference 1 033 030 033 035 036 041 035 031 032 034 042 039 037 037 038 039 035 031 041 0.46 075 100 081 047 042 039
1-Hexpn-3-ol	022 022 034 034 035 034 035 035 035 035 035 035 035 035 035 027 032 035 040 028 035 040 032 075 100 040 037 03		030 030 032 044 045 057 071 037 043 055 038 032 032 <mark>030</mark> 033 021 035 038 039 051 050 0.75 100 0.47 053 0.57 018 035 030 032 041 037 032 073 033 050 030 037 030 037 052 057 043 047 039 038 0.59 0.5 047 130 055 0.54	1-Harme-ci 228 028 029 029 029 021 028 021 028 021 028 033 035 033 032 031 035 229 044 021 025 026 031 021 100 040 035 031 5-Refer-3ci 038 034 033 027 038 021 027 032 031 028 033 029 038 030 025 045 070 028 0.5 047 040 100 024 074
3-Pertyn-2-d 3-Hexyn-2-d	031 027 026 021 025 027 025 027 025 027 025 027 026 027 024 031 025 028 042 055 025 025 025 046 040 100 026 03 028 025 024 020 025 025 025 021 025 025 021 025 025 027 028 023 025 025 025 025 025 026 025 046 027 025 100 02		018 035 030 038 044 050 038 053 050 055 030 036 048 038 047 044 041 046 051 045 052 057 053 0.8 100 080 048	
4Heyn-3-d	024 024 025 034 035 034 035 034 035 034 035 031 031 031 037 029 032 037 031 048 054 035 047 054 035 075 055 10		017 033 041 031 042 055 048 050 055 081 019 035 040 035 044 042 039 042 050 038 057 0.86 057 0.84 080 1.00	
(D)	Teneryo T	(E)	4 insyrnal 3 insyrnal 3 insyrna 1 insyrna 2 in	C () 2 () 1000 () 2 () 1000 () 2 () 1000 () 2 () 1000 (
	2 a a a a a a a a a a a a a a a a a a a	1	3.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1	3.3.4 3.3.4 1.1.4 1.1.4 1.1.4 1.1.4 1.1.4 1.1.4 1.1.4 1.1.4 3.3.4 1.1.4
14mpr×3d	100 CH 01 CH 040 CH 01 CH		2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	
1840-1el		2 2-84(cn-1-ol		
	100 CH 01 CH 040 CH 01 CH	5 3-54cm-1-ol 6 3-7cm/cm-3-ol	2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	
14um-tel 14mm-tel 14mm-tel 14um-tel		2 1-84(cn-1-e) 6 1-8:n(cn-3-e) 8 1-9:n(cn-3-e) 2 1-84(cn-3-e)		
1400-14 1400-14 1400-54 1400-54 1400-54		2-54cm-1-e 2-7cmcm-1-e 2-7cmcm-1-e 3-7cmcm-1-e 1-54cm-3-e 1-54cm-3-e 1-7cmcm-3-e		Ja Ja<
14um-tel 14mm-tel 14mm-tel 14um-tel		1 3-540-1-0 2 3-50100-3-0 3 3-50100-3-0 3 3-50100-3-0 3 3-50100-3-0 3 3-50100-3-0 1 3-50100-3-0		
1845-10 1705-10 1845-10 1845-10 1705-10 1705-10 1705-10 1705-10 1705-10		1-542cn-1-0 2-7612x-3-0		
1845-14 1955-14 1845-14 1845-44 1855-44 1955-44 1955-44 1955-14 1955-14 1955-14 1955-14		1-54cm-1-el 1-54cm-1-el 1-5cms1-el		Ja Ja<
240m344 240m344 240m344 240m344 240m344 240m344 240m344 240m344 240m344 240m344 240m344 240m344		1-542cn-1-0 2-7612x-3-0		Ja Ja <thja< th=""> Ja Ja Ja<!--</td--></thja<>
1845-14 1955-14 1845-14 1845-44 1855-44 1955-44 1955-44 1955-14 1955-14 1955-14 1955-14		2 2-datanal-el 8 2-datanal-el 2 2-datanal-el		Ja Ja <thja< th=""> Ja Ja Ja<!--</td--></thja<>
144m4 el 144m4 el 144m4 el 144m4 el 144m4 el 144m5 el 146m5 el		2 2-datanal-el 8 2-datanal-el 2 2-datanal-el		Ja Ja <thja< th=""> Ja Ja Ja<!--</td--></thja<>
244m-4 el 244m-5		2 2-datanal-el 8 2-datanal-el 2 2-datanal-el		$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$
144m4 el 144m4 el 144m4 el 144m4 el 144m4 el 144m5 el 146m5 el		2 3-2400-04 2 3-8010-34 2 3-8010-34 2 3-8010-34 2 3-8010-34 3 3-8010-34		Ja Ja <thja< th=""> Ja Ja Ja<!--</td--></thja<>
244m44 24mm54		2 Patters-lei 2 Patters-leite-		Ja Ja<
24mm kel 24mm kel 24m		2 - Datas-Jai 2-Astas-Jai		$ \begin{array}{c c c c c c c c c c c c c c c c c c c $
244m44 24mm54		2 Patters-lei 2 Patters-leite-		
Partin - I de Partin		2 - Datas-Jai 2-Astas-Jai		
Adamski Pratovsk		2 - Datas-Jai 2-Astas-Jai		
Partin - I de Partin		2 - Datas-Jai 2-Astas-Jai		

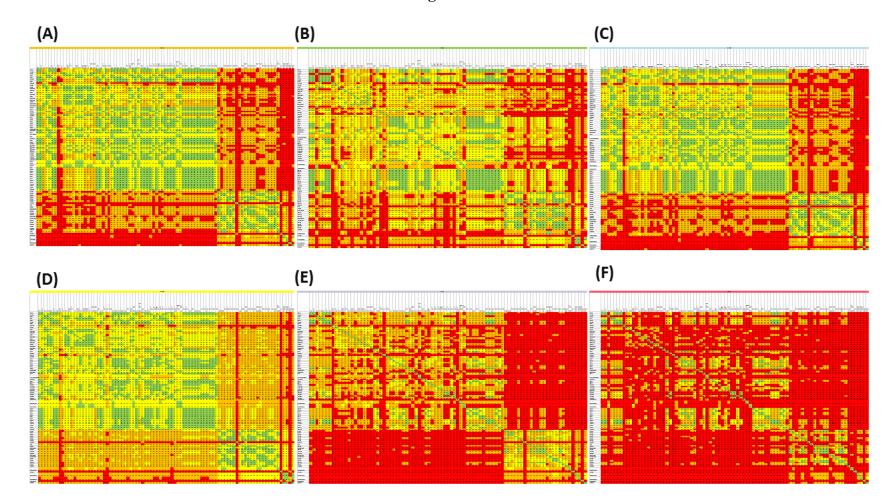


Figure 6