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

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

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

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

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

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

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

ABSTRACT

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

KEYWORDS: Read-across; toxicity; molecular fingerprint; regulatory acceptance; molecular similarity; Tanimoto coefficient; *in silico*

HIGHLIGHTS

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

1. INTRODUCTION

The use of alternative approaches to assess chemical safety is growing due to legislation that requires greater knowledge of the harmful effects of chemicals, whilst also requiring a reduction in, or avoidance of, animal testing. Alternative methods, including *in vitro* assays, -omics and computational approaches ((Quantitative) Structure-Activity Relationships ((Q)SARs), read across etc.) have become integral to many hazard assessment strategies. Of these, computational or (Q)SAR (*in silico*) approaches aim to predict the toxicity of compounds from descriptors of chemical structure and thus reduce testing. In particular, read-across is at the forefront of the prediction of toxicity and has been seen as the "new paradigm" for hazard assessment (Cronin et al, 2013; Berggren et al., 2015; Schultz et al, 2015; Schultz and Cronin 2017; Patlewicz et al 2018). Read-across relies on the ability to identify similar molecules with the assumption that similar molecules will tend to exhibit similar activity or, at least, show similar trends in activity (OECD, 2014). Although the concept of similarity has growing acceptance for toxicity prediction, in reality there are still a number of barriers to acceptance of the predictions, especially for regulatory purposes (Bender and Glen, 2004; Spielmann et al., 2011; Teubner et al., 2015; Ball et al., 2016; Schultz and Cronin 2017; Chesnut et al 2018). Of the barriers identified by Ball et al (2016), some are more trivial to address than others, e.g. full documentation and ensuring the correct chemical structure is provided. The most difficult aspect of justifying a read-across argument is the assessment of "similarity" and being able to provide evidence for such, so to build scientific confidence (Patlewicz et al., 2015; Schultz et al 2018). For instance, there is a concern over effects such as activity cliffs, where structurally similar compounds have a significant difference in potency (Guha and van Drie, 2008; Stumpfe and Bajorath, 2011; Cruz-Monteagudo et al., 2014). In addition, there is the on-going problem of how to define similarity from a molecular level (Maggiora et al., 2014) as well as adequately for read-across (OECD, 2014; Shah et al., 2016; Patlewicz et al 2018; Schultz et al 2018). It is important to note that the similarity between any two objects may be calculated in a variety of different ways and relies on a definable set of features (or descriptors), as well as a means of qualitatively or quantitatively defining similarity based upon those variables. Molecular similarity is no different and whilst two molecules may appear highly similar in one aspect, for instance they may have the same molecular weight, they can be dissimilar in other aspects, such as chemical structure. Thus, the means of defining similarity and providing a means to calculate it is essential. This study has focused on molecular fingerprints due to their increased use in read-across through techniques such as machine learning (Luechtefeld et al., 2018). However, in the context of the current work, the focus is upon read-across predictions made using pairwise comparison to one, or a few, suitably “similar” chemicals, as may well be the case for practical applications. Some of the insights presented herein, regarding the strengths and weaknesses of molecular fingerprint derived similarity measures, may also be applicable in the context of these machine learning studies. Still, detailed examination of the pros and cons of the use of molecular similarity in the context of supervised machine learning, where relationships may be found based on the similarity computed to multiple tested chemicals within a large database, is beyond the scope of the current paper. To assist the reader, definitions are stated in Table 1 that are pertinent to this investigation.

TABLE 1 HERE

The read-across approach may be broadly defined as one in which quantitative or qualitative predictions of an endpoint of interest are made for a target chemical using endpoint data for one or more sufficiently similar source chemicals (OECD, 2014). Usually, this approach is envisaged as only being suitable following grouping of related chemicals, e.g. to form a category (OECD, 2014). There are a number of means of identifying “similar” molecules for grouping and read-across which are deemed acceptable for regulatory purposes, including use of common, mechanistically relevant, structural features and transformation to the same metabolite or degradant (OECD, 2014). There is also the more general concept of “chemical similarity”, i.e. using measures of similarity based on common structural features, physicochemical or biological properties and / or calculated variables related to molecular structure (descriptors). This broader notion of “chemical similarity”, in contrast to those which are deemed acceptable for regulatory purposes, may be defined in terms of generic structural 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 chemical data in order to obtain knowledge about chemical systems; elsewhere, the term cheminformatics or chemoinformatics may be used.) Chemometric measures of similarity are widely used as they are rapid and cost effective due to the availability of online tools, e.g. ChemMine Tools (chemminetools.ucr.edu/) and MuDRA (Alves, 2018), and software that can 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 is relatively well developed (Schultz et al., 2015), much less is known about the use of “chemical similarity”, as defined above, for read-across. This is an area that was founded in the identification of new leads for drug development, thus the similarity measures were not necessarily intended for the purpose for which they are currently applied. For grouping and read-across, where there is no rational measure to find similar compounds, or where a large, diverse inventory is being searched, chemometric methods may seem appealing. However, there is no clear guidance on how they may be applied.

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 machine readable bit representations from chemical structure. Fingerprints are based mostly on 2D representations of a molecule and are used due to their computational efficiency (Holliday et al., 2003). The process of generating bits from chemical structure is illustrated by Figure 1, for a scenario in which the corresponding structural features are molecular substructures A fingerprint is typically a binary vector, with bits set to 1 or 0 depending on the presence or 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 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 "two or more methyl groups" (https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py). Moreover, other fingerprints might encode the occurrence count of structural features, rather than simply their presence or absence. However, if the fingerprint only encodes the presence or absence of certain fragments and not their quantity, this may be a limitation (Flower, 1998). For this scenario, a molecule can contain a specific fragment 1 or 100 times and the resulting bit string will be set the same, thus giving little information with regards to, for instance, molecule size and which fragments occur more often within a molecule (Flower, 1988).

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). The ECFP defines molecular features by assigning identifiers to each of the heavy atoms in the molecule, based upon atomic properties and bonding arrangements, and then combining those identifiers 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 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 are pre-defined (Maggiora et al., 2014). In simple terms, approaches such as ECFP are more complex than MACCS, allowing for the generation of many different atom environments and describe molecular structure more subtly. Finally, it should be noted that different variants of both fingerprints may be computed by different software programs (Rosenbaum et al., 2011; 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).

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).

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

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

Whilst widely applied, a number of studies have shown that using Tc to calculate chemical similarity has its limitations and weaknesses (Dixon and Koehler, 1999; Flower, 1998; 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 been suggested that Tc calculations are biased towards smaller molecules when used for selection according to diversity and that other coefficients may be more appropriate for some 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 molecular similarity which treats the shared presence of mechanistically irrelevant substructures as equally important as the shared presence of mechanistically crucial substructures, such as those corresponding to structural alerts (Alves et al., 2016). One way of 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 considered as a cut-off for read-across (Enoch et al 2009; Hartung, 2016).

The aim of this study was to determine the value of different molecular fingerprints to assess molecular similarity, in terms of the Tanimoto coefficient, in the context of read-across. In particular, the focus of the study was to examine scenarios in which these similarity values might be useful for read-across based upon pairwise comparison to one or a few chemicals, with measured endpoint data, for the purpose of toxicological data gap filling. Specific objectives were to assess the performance and reliability of different molecular fingerprints used in similarity analysis, with a view to determine when similarity computed in this fashion 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 events (MIEs) (Allen et al., 2016, Cronin et al., 2017; Cronin and Richarz, 2017). It should also be made clear that the purpose of this study was not to conclusively establish an optimum method for predicting toxicity. Rather, the purpose of this study was to gain a better understanding of chemical similarity, calculated in terms of the widely used Tanimoto coefficient and generic chemical fingerprints, its strengths, weaknesses and how best to make use of it for read-across based upon pairwise comparisons to one, or a few, chemical(s).

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 complex inventories, e.g. of REACH chemicals, or databases that may be investigated for drug discovery. The selection of the datasets was influenced by a number of factors. Datasets were 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 for an understanding of the selection process for compounds as well as the quality of the underlying biological data. They were also chosen to represent a range of mechanisms and molecular initiating events which may influence the use of molecular similarity.

2. METHODS

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.

TABLE 2 HERE

2.2 Calculation of molecular fingerprints

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

2.3 Analysis of Tanimoto coefficients.

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

3. RESULTS

The KNIME workflow produced a CSV file which contained calculated Tc values for the input data sets. The Tc data matrices for the datasets are provided in the supplementary information. 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 non-polar narcotic datasets), full details of which can been found within the supplementary information along with the matrices for the LLNA skin sensitisation dataset. (The size of the 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 fingerprints are shown, where **A** was calculated using CDK Standard fingerprints, **B** was calculated using CDK MACCS fingerprints, **C** was calculated using CDK Extended fingerprints, **D** was calculated using CDK PubChem fingerprints, **E** was calculated using CDK 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* and *j* being shown in the matrix element of row *i* and column *j* of the matrix, such that the Tc values for the same compound compared to itself (Tc=1.0) lie along the diagonal elements. N.B. (1) Each row (column) in these images is labelled by the name of the chemical for which colour coded similarity values are reported within that row (column). (2) These images are designed to illustrate the variation in pairwise similarity for the same pairs of compounds using different fingerprints in terms of the corresponding colour patterns. The size of some datasets 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 Information. In addition, Tables 3 – 5 show the range of Tanimoto similarity values that can be obtained for the same pairwise comparisons, between compounds in selected datasets, using the different fingerprints.

FIGURES 2-6 HERE

TABLES 3-5 HERE

4. DISCUSSION

Chemical similarity is, in theory, a beguiling concept allowing for the identification of similar molecules to those with existing information, whether it be biological activity (such as pharmacological or toxicological effects), biokinetics, environmental fate or physico-chemical properties. The science of molecular similarity is founded in drug discovery, where the aim was to identify similar molecules to a known active compound. It mostly utilises easily calculable parameters (descriptors), or fingerprint representations, of molecular structure. The 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 interest in this approach in drug discovery for many years and there has been a recent growth 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 molecules highly similar to a known active (assuming a receptor mediated pharmacological effect) to multiple uses ranging from searching for any "similar" molecules to a target query with unknown activity, to serving as the input to grouping and/or read-across approaches (Gini 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 not being identified as the similarity measures are not appropriate. The purpose of this study, 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 read-across, with a focus on read-across predictions made using pairwise similarity calculations 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 disadvantages of a single method. The similarity matrices in Figures 2-6 clearly demonstrate a difference in Tc scores calculated for the same dataset when using different fingerprints. Closer examination of the perfluorinated acids dataset (Figure 2, dataset 3 from Table 2) indicates a concordance in the fingerprints with regard to in their Tc values as all data matrices are green (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 i.e. the Tc based assessment of all chemicals as highly similar is in keeping with the assessment which would be made by toxicological experts - since this dataset comprises a homologous series, i.e. the same functional group with varying chain length, expected to act via a common mechanism. As would be expected, variations in Tc scores were as a result of differences in 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 with each other and the chemical with C9 tended to only show a similarity score of 1 when compared against itself (for CDK standard, CDK Extended fingerprints) or those with C10-C12 (for the other fingerprints). Naturally, all fingerprints gave a Tc value of one for comparisons of the same compound to itself. This trend was similar for all fingerprints applied to this dataset. Thus, fingerprint similarity, in terms of Tc, is a reasonable measure when 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-tuning” a read-across within such a preselected series of chemicals – the process sometimes referred to as sub-categorisation.

Analysis of datasets with greater structural variability (cf. Figures 3 - 6) indicates a much higher variability in the calculated Tc values depending on which fingerprint was chosen, with limited 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. For two chemicals, 3-methyl-6-n-butylphenol and 2,6-di-tert-butylphenol, CDK FCFP6 fingerprints gave a Tc score of 0.26, whereas CDK PubChem fingerprints gave a Tc score of 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, 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 the calculated Tc values for CDK FCFP6 and the CDK ECFP4 fingerprints were significantly different to the Tc values from the other four fingerprints, with the CDK ECFP4 giving many values that would suggest “highly dissimilar” chemicals, which is not the case for these datasets (based upon expert judgement). Similar discrepancies between fingerprints were seen for the non-polar narcosis dataset (Figure 6). The reasons for such discrepancies undoubtedly reflect the method of fingerprint calculation having an enormous impact on the identification of analogues from large structurally heterogeneous datasets. It may even be an indicator for consideration of composite Tc scores to capitalise on the different information contained. However, that would not address the possibility that toxicologically irrelevant structural 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 combined. Overall, care must be applied in using Tc values for structurally heterogeneous 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 fingerprints and mechanism of the endpoint to be read across.

For the unsaturated alcohols dataset (Figure 5), all the calculated Tc similarity matrices were noticeably different for each of the six fingerprints used. This dataset consist of chemicals which are, on the face of it, structurally similar but with subtle changes and differences not only in chain length but also the position of the hydroxyl group, (primary or secondary alcohol), branching, and position (internal or external) of the double bond. The positioning of the alcohol group and double bond, as well as branching, will impact of toxicity (Schultz et al., 2017), however none of the Tc values assisted in identifying rational, mechanistically similar analogues across the group. Therefore, subtle, mechanistically relevant changes in molecular 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 important to whether irrelevant molecular substructures are shared or not between two molecules.

Using molecular similarity to assist in toxicity prediction is unlikely to be perfect. There are 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 sensitisation dataset, showing potential issues with activity cliffs, despite high Tc values from some fingerprints. Comparison of 1,4-dihydroxyquinone, a strong skin sensitiser, with resorcinol (1,3-dihydroxyquinone), a non-sensitiser, indicates both chemicals being highly similar in structure with the only difference being the position of the hydroxyl groups on the phenol ring (Table 5). The position of the hydroxyl groups in 1,4-dihydroxyquinone enables this chemical to readily form benzoquinone, a reactive metabolite, whereas resorcinol does not 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 indicate high similarity, which could lead to false assumptions with regard to grouping and read-across, unless the mechanism of action is known. The wide range of Tc scores calculated also shows the variability of the Tc scores dependent upon the choice of fingerprint. This emphasises the importance of choosing the most appropriate fingerprint, if any, for similarity calculations. In the second comparison 3-phenylenediamine, a strong skin sensitiser, is compared against aniline, a weak skin sensitiser. These chemicals are highly similar in structure, with the main difference being the presence of an extra amine group (Table 5). It has been demonstrated that the presence of the 2 amine groups in 3-phenylenediamine makes this 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 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. Clearly, this would be an incorrect conclusion.

The final two comparisons compare 3,4-dihydrocoumarin, a moderate skin sensitiser, against coumarin and 6-methylcoumarin which are both non-sensitisers (Table 5). These chemicals are 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 coumarin causes it to be readily metabolised via Michael addition, into a non-sensitising metabolite (Table 5). The absence of the double bond makes 3,4-dihydrocoumarin more reactive, which accounts for its moderate skin sensitisation when compared to the other two chemicals. The Tc scores calculated for these two comparisons again show variability dependent on fingerprint choice (Table 5). Two of the six fingerprints (CDK MACCS and CDK PubChem) resulted in high Tc scores; this would suggest these chemicals exhibit similar 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 using a limited number of mechanistically relevant descriptors chosen based on expert judgement. For example, in the case of skin sensitization, the electrophilicity index could be used (Enoch et al., 2008). Similarities might be computed based upon the more general expression for the Tanimoto coefficient, for continuous variables (Willett et al., 1998), following normalisation of different descriptors to the same scale. However, even under this 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 (Enoch et al., 2008).

The visualisation and practical handling of Tc values should be borne in mind. In this investigation, due to the number of chemicals in the LLNA skin sensitisation (211 chemicals) and the non-polar narcotic (87 chemicals) datasets (Figure 6 and supplementary data), both of which are quite modest in size, visualisation was challenging which makes the analysis of results difficult. This is an issue that needs to be addressed to ensure that Tc similarity matrices 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 within each matrix and making visualisation easier. One means of achieving this is that any relevant knowledge of MIEs should be used to pre-categorise the datasets prior to calculating Tc values. For example, Tc values might be computed for chemicals acting via a common MIE, as indicated by a shared structural alert, and for which some other expert based rules reduced mechanistically irrelevant structural variation that would reduce the information conveyed by 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 known to be biologically relevant.

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

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

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 computed from molecular fingerprints and the variation in similarity values which can be obtained with different fingerprints, as highlighted in the current work, it is unlikely that read-across predictions based upon these values using a single fingerprint would be optimal for all relevant scenarios. Thus, for the examples that may be taken from the range of datasets investigated in this study, different types of chemical similarity would be required for effective and defensible analogue selection. Optimal read-across predictions are more likely to be 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 statistical validation, is more likely to lead to acceptance in a regulatory context.

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 study. It is further acknowledged that the use of a broad fingerprint method based around known toxicologically relevant fragments could assist in situations where the precise MIE may not be known. However, the development of new fingerprints to aid toxicological read-across would most appropriately be carried out on an endpoint specific basis, rather than assuming a single fingerprint could be developed for all endpoints.

5. CONCLUSIONS

In conclusion, molecular fingerprint similarity matrices can be used as a means of identifying possible analogues in some contexts. However, on their own, it is difficult to use generic 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 similarity measures calculated from these fingerprints, which are highlighted in the current 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). The fingerprints may not capture the relevant structural variation (depending upon the fingerprint method) and treat mechanistically irrelevant structural variation equally to mechanistically relevant structural variation. Similarity matrices, calculated from different 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 for read-across within a homologous series, acting via a common mechanism, for which the 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 variation. Hence, for a read across setting, users of chemically diverse datasets could benefit from first forming categories when using molecular fingerprint similarity values.

Whilst Tanimoto similarity values computed from generic molecular fingerprints have been integrated into recent machine learning predictions of toxicity within diverse datasets with empirically successful results, the limitations of these similarity values, highlighted in our work, mean that other approaches to similarity assessment are preferable for read-across. Ideally, similarity values which reflect biologically relevant information, informed by mechanistic understanding, should be employed. This is especially the case in a regulatory context, where 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 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.

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 similarity values derived from CDK Standard, CDK MACCS, CDK Extended and CDK PubChem fingerprints showed some concordance, for some scenarios, with similarity values for CDK FCFP6 and the CDK ECFP4 providing different information. Further work is required to understand the significance of these findings and at this time no single fingerprint method from those investigated could be considered to be the most optimum. 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 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 read-across 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-across within large, chemically diverse datasets is following sub-categorisation. (However, further work is required to determine the extent to which this yields better predictive performance than integrating these similarity measures within machine learning approaches, which have recently been advocated. Moreover, sub-categorisation which removes biologically irrelevant structural variation may result in the fingerprint-derived similarity measures being optimally predictive, yet redundant if read-across is performed by expert examination of the structures within the category.) Sub-categorisation should preferably be performed using a mechanistically based method. If sub-categorisation yields homologous series, acting via a common mechanism, for which all the structural variation is expected to be biologically relevant, generic fingerprint-derived similarity measures could be suitable for fine tuning and confirming analogue identification for read-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.

Overall, fingerprint-derived measures of molecular similarity may be a useful method in the *in silico* toolbox for data gap filling. However, they are likely to be optimally predictive within a small, mechanistically derived category and, ideally, the specific similarity measure should be appropriate to the chemistry and endpoint considered.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

Allen, T.E.H., Goodman, J.M., Gutsell, S., Russell, P.J., 2016. A history of the Molecular 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., Sedykh, A., Mokshyna, E., Farag, S., Andrade, C.H., Kuz'min, V.E., Fourches, D., Tropsha A., 2016. Alarms about structural alerts. Green Chem. 18, 4348-4360.

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

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

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., Egnash, L., Hastings, C., Juberg, D.R., Kleensang, A., Kleinstreuer, N., Kroese, E.D., Lee, A.C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J.M., Palmer, J., Pamies, D., Penman, M., Richarz, A.-N., Russo, D.,P.. Stuard, S.B., Patlewicz, G., van Ravenzwaay, B., Wu, S., Zhu, H., Hartung, T., 2016. Toward Good Read-Across Practice (GRAP) guidance. ALTEX, 33, 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. 123, 1232-1240.

Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., Pujadasa, G., 2015. Molecular fingerprint similarity search in virtual screening. Methods. 71, 58–63.

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.

Cronin, M.T.D., Madden, J.C., Enoch, S.J., Roberts, D.W., 2013. Chemical Toxicity Prediction: Category Formation and Read-Across. Royal Society of Chemistry, Cambridge UK.

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

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 Discov. Today. 19, 1069-1080.

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.

Eckert, H., Bajorath, J. 2007. Molecular similarity analysis in virtual screening: foundations, 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 structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. SAR QSAR Environ. Res. 19, 751-783.

Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via 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.

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. Toxicol. 41, 783-802.

Flower, D.R., 1998. On the properties of Bit string-based measures of chemical similarity. J. Chem. Inf. Comput. Sci. 38, 379–386.

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

Godden, J.W., Xue, L., Bajorath, J., 2000. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J. Chem. 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 similarity and dissimilarity using 2D fragment Bit-strings. Comb. Chem. High Throughput 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.

Martin, Y.C., Kofron, J.L., Traphagen, L.M.., 2002. Do structurally similar compounds have similar biological activity? J. Med. Chem. 45, 4350−4358.

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 mono-alkylphenols: 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.

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.

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., Mansouri, K., Patlewicz, G., Williams, A.J., Little, S.B., Crofton, K.M., Thomas, R.S., 2016. ToxCast chemical landscape: paving the road to 21st Century Toxicology. Chem. Res. 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.

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.

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

Schultz, T.W., Richarz, A.-N., Cronin, M.T.D., 2018. Assessing uncertainty in read-across: knowledge gained from case studies. Comput. Toxicol. submitted

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.

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.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P.., 2012. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J. Chem. Inf. Model. 52, 2884–2901.

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, 449-457.

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

Willett, P., 2013. Combination of similarity rankings using data fusion. J. Chem. Inf. Model. 53, 1–10.

Willett, P., Barnard, J.M., Downs, G.M., 1998. Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38, 983-996.

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

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

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

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

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

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

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

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

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

Abbreviations relate to the following : Perfluorohexanoic acid (PFHxA), Perfluoroheptanoic acid (PFHpA), Perfluorooctanoic acid (PFOA), Perfluorononanoic acid (PFNA), Perfluorodecanoic acid (PFDA), Perfluoroundecanoic acid (PFUA) and Perfluorododecanic acid (PFDoA).

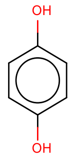
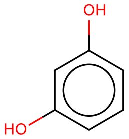
Table 4: Shows the range of the Tc scores calculated when utilising the different fingerprints for the alkylphenols dataset (dataset 6).

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

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

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

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

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

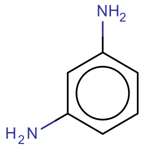
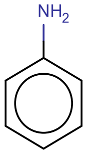
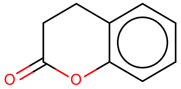
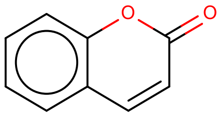
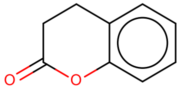
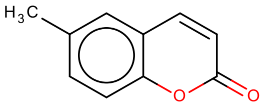
								
3,4-dihydrocoumarin (2.5, moderate sensitiser) 	Coumarin (5.0, non-sensitiser) 	0.43	0.73	0.48	0.86	0.40	0.35	**0.35-0.86**
3,4-dihydrocoumarin (2.5, moderate sensitiser) 	6-methylcoumarin (5.0, non-sensitiser) 	0.40	0.74	0.43	0.83	0.27	0.21	**0.21-0.83**

Figure Captions:

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

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

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

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

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

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

Figure 1

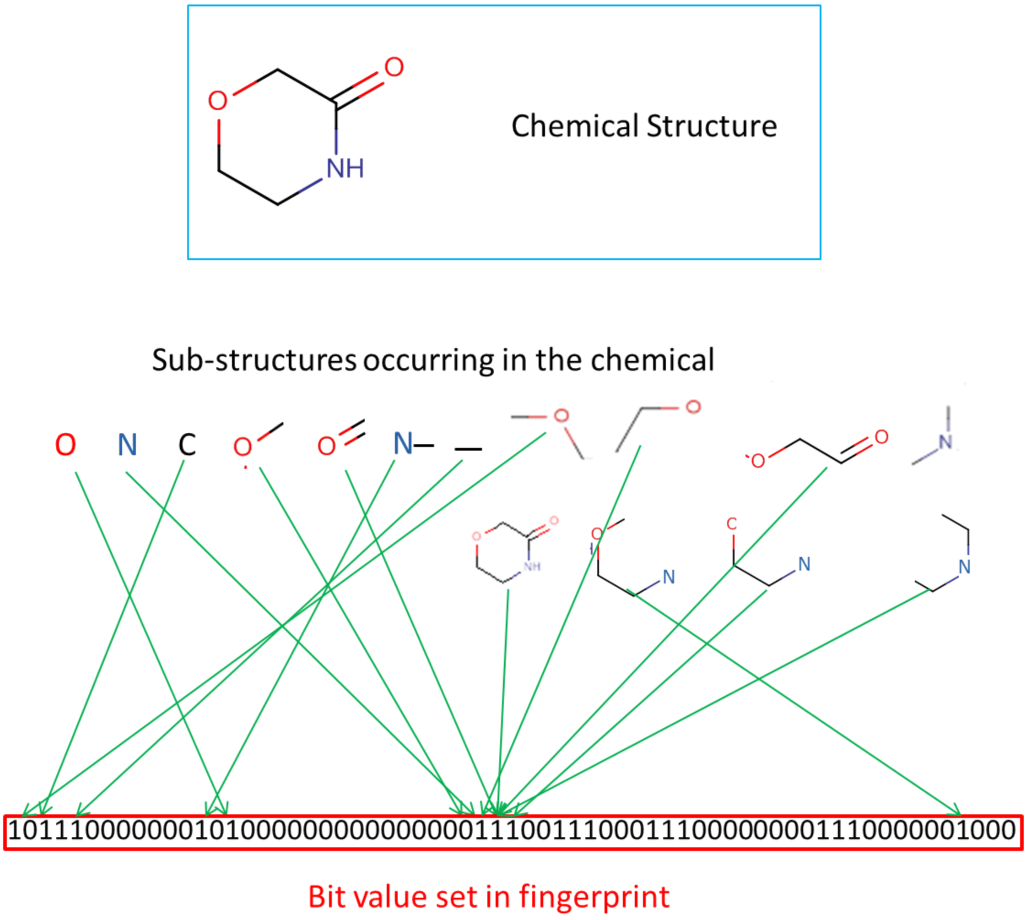
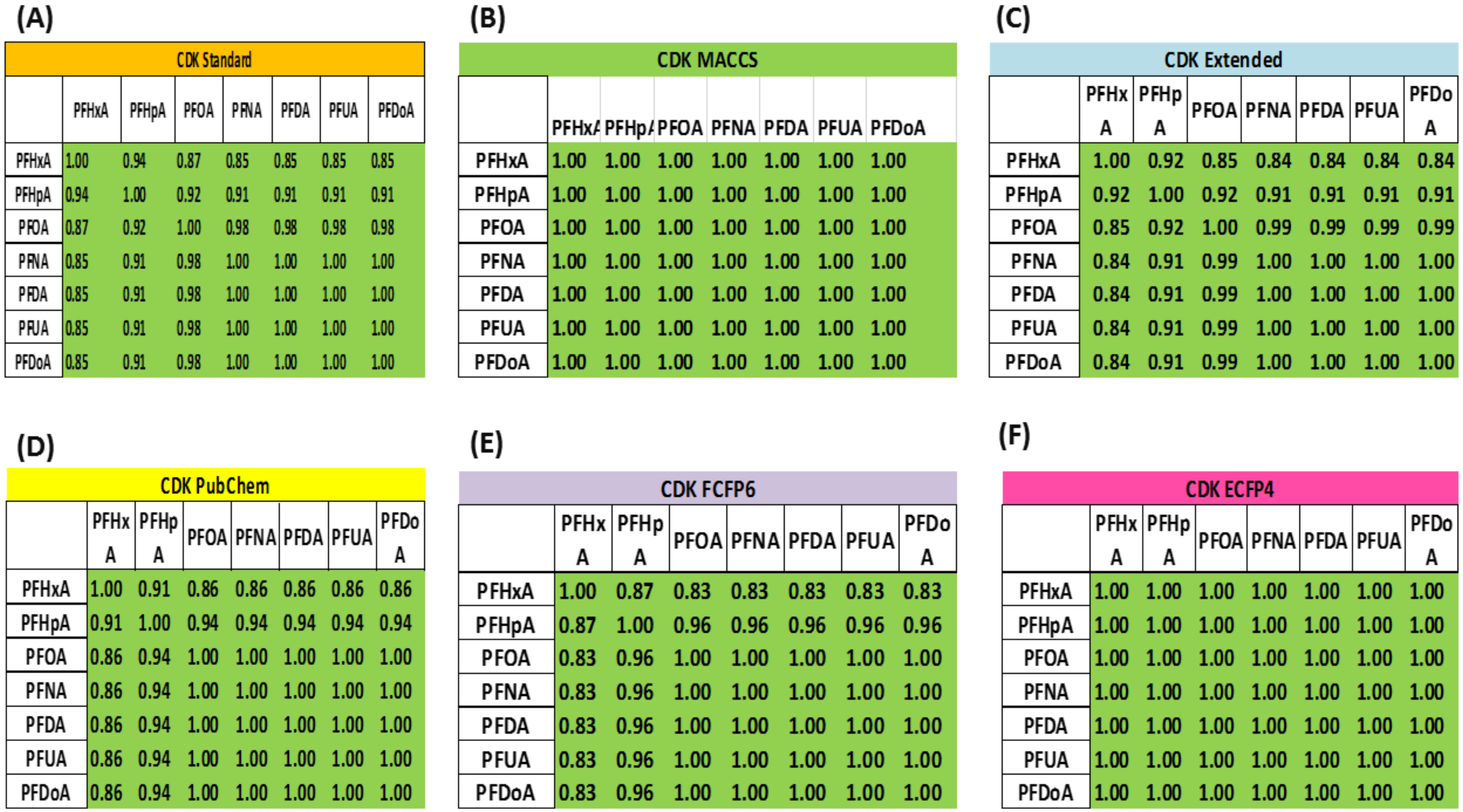

Chemical Structure
Sub-structures occurring in the chemical
101110000000101000000000000000111001110001110000000011100000001000
Bit value set in fingerprint

Figure 2



(A)

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

(B)

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

(C)

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

(D)

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

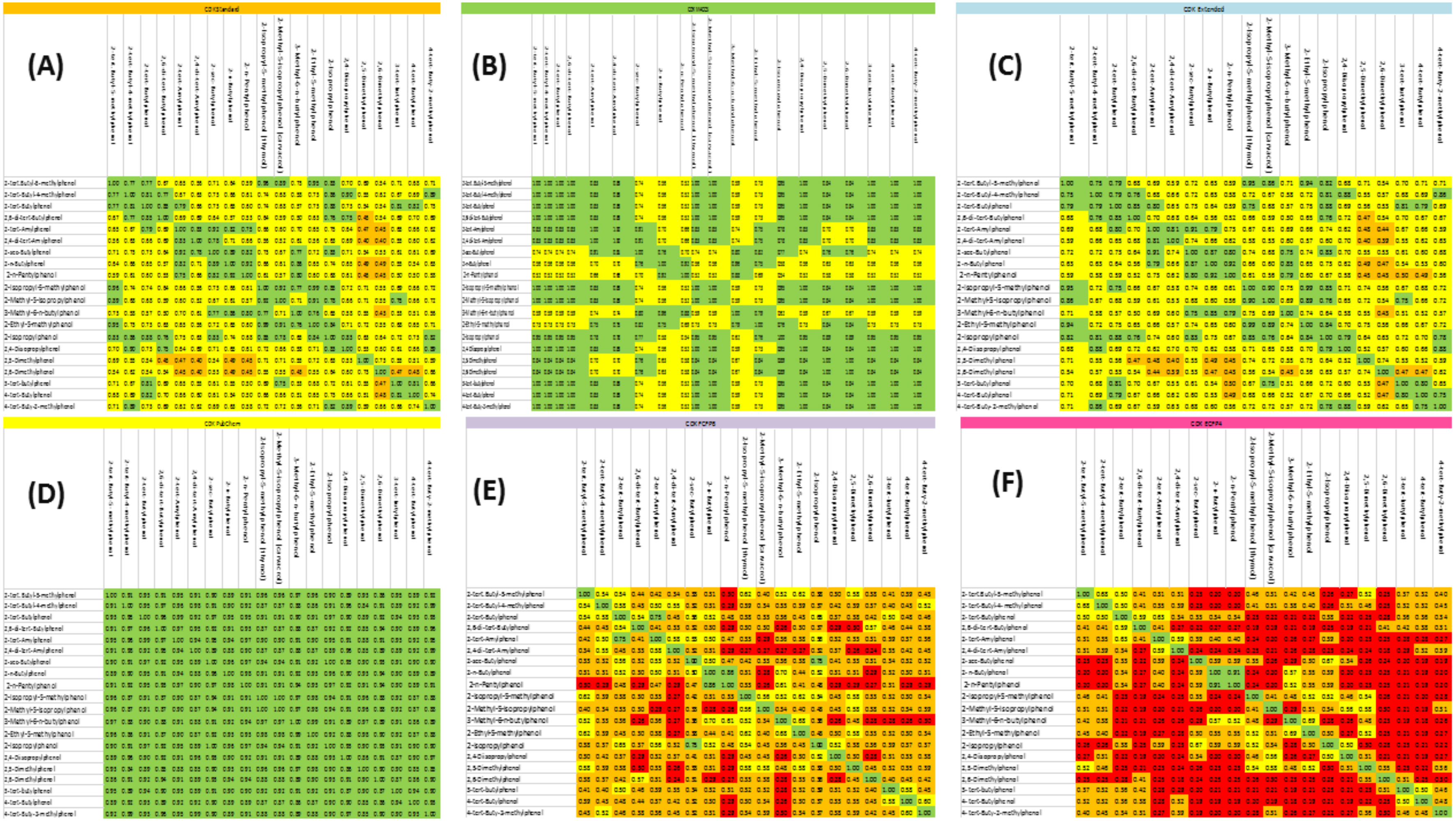
(E)

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

(F)

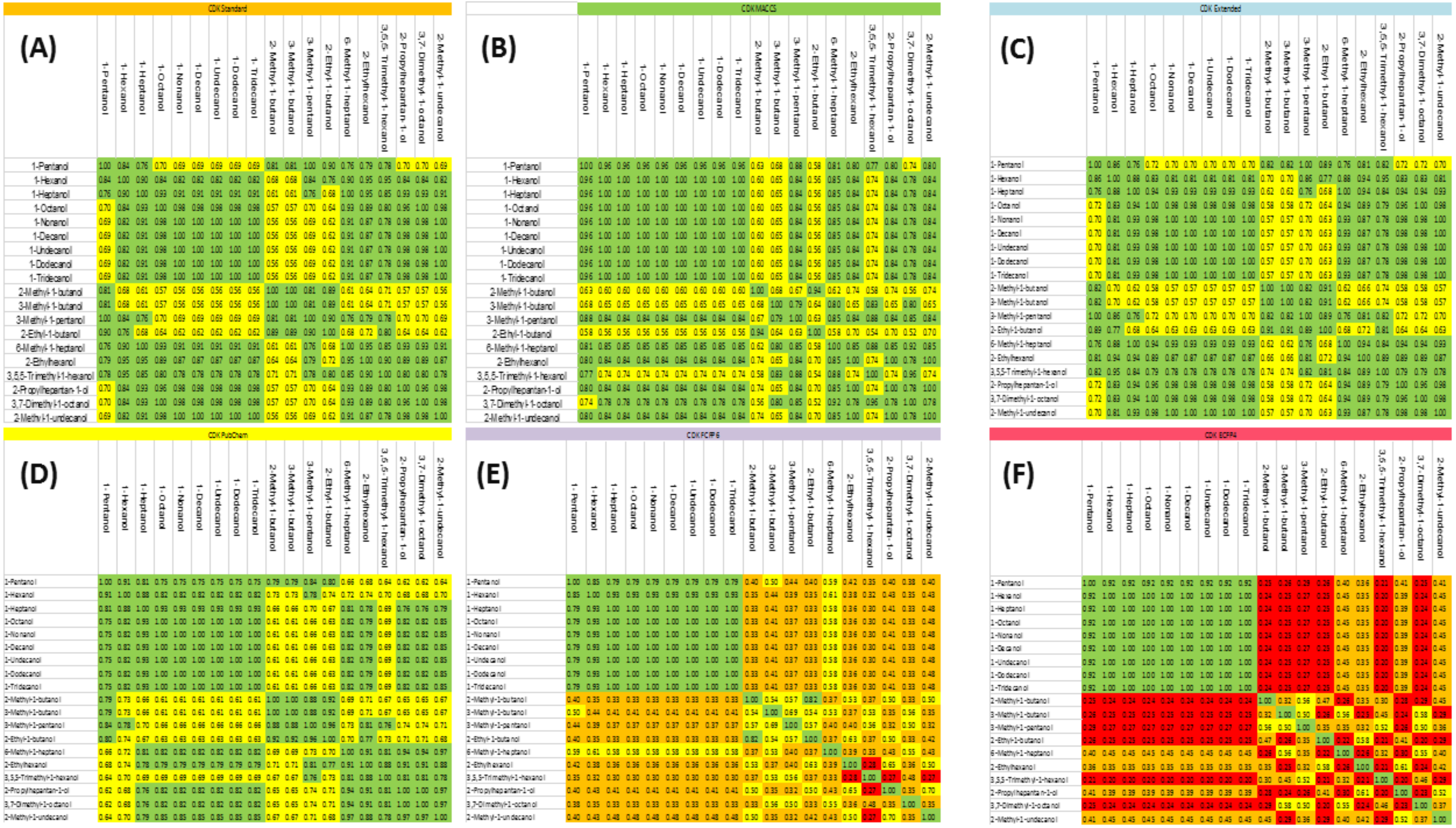
CDK ECFP4							
	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUA	PFDoA
PFHxA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFHpA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFOA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFNA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFUA	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PFDoA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

Figure 3



CDK Standard
(A)
CDK MACCS
(B)
CDK Extended
(C)
CDK PubChem
(D)
CDK FCFP6
(E)
CDK ECFP4
(F)

Figure 4



CDK Standard

(A)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	1.00	0.84	0.76	0.70	0.69	0.69	0.69	0.69	0.69	0.81	0.81	1.00	0.90	0.76	0.79	0.78	0.70	0.70	0.69
1-Hexanol	0.84	1.00	0.90	0.84	0.82	0.82	0.82	0.82	0.82	0.68	0.68	0.84	0.76	0.90	0.95	0.95	0.84	0.84	0.82
1-Heptanol	0.76	0.90	1.00	0.93	0.91	0.91	0.91	0.91	0.91	0.61	0.61	0.76	0.68	1.00	0.95	0.85	0.93	0.93	0.91
1-Octanol	0.70	0.84	0.93	1.00	0.98	0.98	0.98	0.98	0.98	0.57	0.57	0.70	0.64	0.93	0.89	0.80	0.96	1.00	0.98
1-Nonanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00
1-Decanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00
1-Undecanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00
1-Dodecanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00
1-Tridecanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00
2-Methyl-1-butanol	0.81	0.68	0.61	0.57	0.56	0.56	0.56	0.56	0.56	1.00	1.00	0.81	0.89	0.61	0.64	0.71	0.57	0.57	0.56
3-Methyl-1-butanol	0.81	0.68	0.61	0.57	0.56	0.56	0.56	0.56	0.56	1.00	1.00	0.81	0.89	0.61	0.64	0.71	0.57	0.57	0.56
3-Methyl-1-pentanol	1.00	0.84	0.76	0.70	0.69	0.69	0.69	0.69	0.69	0.81	0.81	1.00	0.90	0.76	0.79	0.78	0.70	0.70	0.69
2-Ethyl-1-butanol	0.90	0.76	0.68	0.64	0.62	0.62	0.62	0.62	0.62	0.89	0.89	0.90	1.00	0.68	0.72	0.80	0.64	0.64	0.62
6-Methyl-1-heptanol	0.76	0.90	1.00	0.93	0.91	0.91	0.91	0.91	0.91	0.61	0.61	0.76	0.68	1.00	0.95	0.85	0.93	0.93	0.91
2-Ethylhexanol	0.79	0.95	0.95	0.89	0.87	0.87	0.87	0.87	0.87	0.64	0.64	0.79	0.72	0.95	1.00	0.90	0.89	0.89	0.87
3,5,5-Trimethyl-1-hexanol	0.78	0.95	0.85	0.80	0.78	0.78	0.78	0.78	0.78	0.71	0.71	0.78	0.80	0.85	0.90	1.00	0.80	0.80	0.78
2-Propylhepantan-1-ol	0.70	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
3,7-Dimethyl-1-octanol	0.70	0.84	0.93	1.00	0.98	0.98	0.98	0.98	0.98	0.57	0.57	0.70	0.64	0.93	0.89	0.80	0.96	1.00	0.98
2-Methyl-1-undecanol	0.69	0.82	0.91	0.98	1.00	1.00	1.00	1.00	1.00	0.56	0.56	0.69	0.62	0.91	0.87	0.78	0.98	0.98	1.00

CDK MACCS

(B)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.63	0.68	0.88	0.58	0.81	0.80	0.77	0.80	0.74	0.80
1-Hexanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Heptanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Octanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Nonanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Decanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Undecanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Dodecanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
1-Tridecanol	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60	0.65	0.84	0.56	0.85	0.84	0.74	0.84	0.78	0.84
2-Methyl-1-butanol	0.63	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	1.00	0.68	0.67	0.94	0.62	0.74	0.58	0.74	0.56	0.74
3-Methyl-1-butanol	0.68	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.68	1.00	0.79	0.64	0.80	0.65	0.83	0.65	0.80	0.65
3-Methyl-1-pentanol	0.88	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.67	0.79	1.00	0.63	0.85	0.84	0.88	0.84	0.85	0.84
2-Ethyl-1-butanol	0.58	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.94	0.64	0.63	1.00	0.58	0.70	0.54	0.70	0.52	0.70
6-Methyl-1-heptanol	0.81	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.62	0.80	0.85	0.58	1.00	0.85	0.88	0.85	0.92	0.85
2-Ethylhexanol	0.80	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.74	0.65	0.84	0.70	0.85	1.00	0.74	1.00	0.78	1.00
3,5,5-Trimethyl-1-hexanol	0.77	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.58	0.83	0.88	0.54	0.88	0.74	1.00	0.74	0.96	0.74
2-Propylhepantan-1-ol	0.80	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.74	0.65	0.84	0.70	0.85	1.00	0.74	1.00	0.78	1.00
3,7-Dimethyl-1-octanol	0.74	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.56	0.80	0.85	0.52	0.92	0.78	0.96	0.78	1.00	0.78
2-Methyl-1-undecanol	0.80	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.74	0.65	0.84	0.70	0.85	1.00	0.74	1.00	0.78	1.00

CDK Extended

(C)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	1.00	0.86	0.76	0.72	0.70	0.70	0.70	0.70	0.70	0.82	0.82	1.00	0.89	0.76	0.81	0.82	0.72	0.72	0.70
1-Hexanol	0.86	1.00	0.88	0.83	0.81	0.81	0.81	0.81	0.81	0.70	0.70	0.86	0.77	0.88	0.94	0.95	0.83	0.83	0.81
1-Heptanol	0.76	0.88	1.00	0.94	0.93	0.93	0.93	0.93	0.93	0.62	0.62	0.76	0.68	1.00	0.94	0.84	0.94	0.94	0.93
1-Octanol	0.72	0.83	0.94	1.00	0.98	0.98	0.98	0.98	0.98	0.58	0.58	0.72	0.64	0.94	0.89	0.79	0.96	1.00	0.98
1-Nonanol	0.70	0.81	0.93	0.98	1.00	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
1-Decanol	0.70	0.81	0.93	0.98	1.00	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
1-Undecanol	0.70	0.81	0.93	0.98	1.00	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
1-Dodecanol	0.70	0.81	0.93	0.98	1.00	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
1-Tridecanol	0.70	0.81	0.93	0.98	1.00	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
2-Methyl-1-butanol	0.82	0.70	0.62	0.58	0.57	0.57	0.57	0.57	0.57	1.00	1.00	0.82	0.91	0.62	0.66	0.74	0.58	0.58	0.57
3-Methyl-1-butanol	0.82	0.70	0.62	0.58	0.57	0.57	0.57	0.57	0.57	1.00	1.00	0.82	0.91	0.62	0.66	0.74	0.58	0.58	0.57
3-Methyl-1-pentanol	1.00	0.86	0.76	0.72	0.70	0.70	0.70	0.70	0.70	0.82	0.82	1.00	0.89	0.76	0.81	0.82	0.72	0.72	0.70
2-Ethyl-1-butanol	0.89	0.77	0.68	0.64	0.63	0.63	0.63	0.63	0.63	0.91	0.91	0.89	1.00	0.68	0.72	0.81	0.64	0.64	0.63
6-Methyl-1-heptanol	0.76	0.88	1.00	0.94	0.93	0.93	0.93	0.93	0.93	0.62	0.62	0.76	0.68	1.00	0.94	0.84	0.94	0.94	0.93
2-Ethylhexanol	0.81	0.94	0.94	0.89	0.87	0.87	0.87	0.87	0.87	0.66	0.66	0.81	0.72	0.94	1.00	0.89	0.89	0.89	0.87
3,5,5-Trimethyl-1-hexanol	0.82	0.95	0.84	0.79	0.78	0.78	0.78	0.78	0.78	0.74	0.74	0.82	0.81	0.84	0.89	1.00	0.79	0.79	0.78
2-Propylhepantan-1-ol	0.72	0.83	0.94	0.96	0.98	0.98	0.98	0.98	0.98	0.58	0.58	0.72	0.64	0.94	0.89	0.79	1.00	0.96	0.98
3,7-Dimethyl-1-octanol	0.72	0.83	0.94	1.00	0.98	0.98	0.98	0.98	0.98	0.58	0.58	0.72	0.64	0.94	0.89	0.79	0.96	1.00	0.98
2-Methyl-1-undecanol	0.70	0.81	0.93	0.98	1.00	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

CDK PubChem

(D)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	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
1-Hexanol	0.91	1.00	0.88	0.82	0.82	0.82	0.82	0.82	0.82	0.73	0.73	0.78	0.74	0.72	0.74	0.70	0.68	0.68	0.70
1-Heptanol	0.81	0.88	1.00	0.93	0.93	0.93	0.93	0.93	0.93	0.66	0.66	0.70	0.67	0.81	0.78	0.69	0.76	0.76	0.79
1-Octanol	0.75	0.82	0.93	1.00	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
1-Nonanol	0.75	0.82	0.93	1.00	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
1-Decanol	0.75	0.82	0.93	1.00	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
1-Undecanol	0.75	0.82	0.93	1.00	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
1-Dodecanol	0.75	0.82	0.93	1.00	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
1-Tridecanol	0.75	0.82	0.93	1.00	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
2-Methyl-1-butanol	0.79	0.73	0.66	0.61	0.61	0.61	0.61	0.61	0.61	1.00	1.00	0.88	0.92	0.69	0.71	0.67	0.65	0.65	0.67
3-Methyl-1-butanol	0.79	0.73	0.66	0.61	0.61	0.61	0.61	0.61	0.61	1.00	1.00	0.88	0.92	0.69	0.71	0.67	0.65	0.65	0.67
3-Methyl-1-pentanol	0.84	0.78	0.70	0.66	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
2-Ethyl-1-butanol	0.80	0.74	0.67	0.63	0.63	0.63	0.63	0.63	0.63	0.92	0.92	0.96	1.00	0.70	0.77	0.73	0.71	0.71	0.68
6-Methyl-1-heptanol	0.66	0.72	0.81	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
2-Ethylhexanol	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
3,5,5-Trimethyl-1-hexanol	0.64	0.70	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.67	0.67	0.76	0.73	0.81	0.88	1.00	0.81	0.81	0.78
2-Propylhepantan-1-ol	0.62	0.68	0.76	0.82	0.82	0.82	0.82	0.82	0.82	0.65	0.65	0.74	0.71	0.94	0.91	0.81	1.00	1.00	0.97
3,7-Dimethyl-1-octanol	0.62	0.68	0.76	0.82	0.82	0.82	0.82	0.82	0.82	0.65	0.65	0.74	0.71	0.94	0.91	0.81	1.00	1.00	0.97
2-Methyl-1-undecanol	0.64	0.70	0.79	0.85	0.85	0.85	0.85	0.85	0.85	0.67	0.67	0.71	0.68	0.97	0.88	0.78	0.97	0.97	1.00

CDK FCFP6

(E)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	1.00	0.85	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.40	0.50	0.44	0.40	0.59	0.42	0.35	0.40	0.38	0.40
1-Hexanol	0.85	1.00	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.35	0.44	0.39	0.35	0.61	0.38	0.32	0.43	0.35	0.43
1-Heptanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Octanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Nonanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Decanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Undecanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Dodecanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
1-Tridecanol	0.79	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.33	0.41	0.37	0.33	0.58	0.36	0.30	0.41	0.33	0.48
2-Methyl-1-butanol	0.40	0.35	0.33	0.33	0.33	0.33	0.33	0.33	0.33	1.00	0.54	0.57	0.82	0.37	0.53	0.37	0.50	0.33	0.50
3-Methyl-1-butanol	0.50	0.44	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.54	1.00	0.69	0.54	0.53	0.37	0.53	0.35	0.56	0.35
3-Methyl-1-pentanol	0.44	0.39	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.57	0.69	1.00	0.57	0.40	0.40	0.56	0.32	0.50	0.32
2-Ethyl-1-butanol	0.40	0.35	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.82	0.54	0.57	1.00	0.37	0.63	0.37	0.50	0.33	0.42
6-Methyl-1-heptanol	0.59	0.61	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.37	0.53	0.40	0.37	1.00	0.39	0.33	0.43	0.55	0.43
2-Ethylhexanol	0.42	0.38	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.53	0.37	0.40	0.63	0.39	1.00	0.28	0.65	0.36	0.50
3,5,5-Trimethyl-1-hexanol	0.35	0.32	0.30	0.30	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
2-Propylhepantan-1-ol	0.40	0.43	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.50	0.35	0.32	0.50	0.43	0.65	0.27	1.00	0.35	0.70
3,7-Dimethyl-1-octanol	0.38	0.35	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.56	0.50	0.33	0.55	0.36	0.48	0.35	1.00	0.35
2-Methyl-1-undecanol	0.40	0.43	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.50	0.35	0.32	0.42	0.43	0.50	0.27	0.70	0.35	1.00

CDK ECFP4

(F)

	1-Pentanol	1-Hexanol	1-Heptanol	1-Octanol	1-Nonanol	1-Decanol	1-Undecanol	1-Dodecanol	1-Tridecanol	2-Methyl-1-butanol	3-Methyl-1-butanol	3-Methyl-1-pentanol	2-Ethyl-1-butanol	6-Methyl-1-heptanol	2-Ethylhexanol	3,5,5-Trimethyl-1-hexanol	2-Propylhepantan-1-ol	3,7-Dimethyl-1-octanol	2-Methyl-1-undecanol
1-Pentanol	1.00	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.25	0.26	0.29	0.26	0.40	0.36	0.21	0.41	0.25	0.41
1-Hexanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Heptanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Octanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Nonanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Decanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Undecanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Dodecanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
1-Tridecanol	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.24	0.25	0.27	0.25	0.45	0.35	0.20	0.39	0.24	0.45
2-Methyl-1-butanol	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	1.00	0.32	0.56	0.47	0.26	0.35	0.30	0.28	0.29	0.45
3-Methyl-1-butanol	0.26	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.32	1.00	0.50	0.26	0.56	0.25	0.45	0.24	0.58	0.29
3-Methyl-1-pentanol	0.29	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.56	0.50	1.00	0.35	0.35	0.32	0.52	0.26	0.50	0.36
2-Ethyl-1-butanol	0.26	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.47	0.26	0.35	1.00	0.22	0.58	0.21	0.41	0.20	0.29
6-Methyl-1-heptanol	0.40	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.26	0.56	0.35	0.22	1.00	0.26	0.32	0.30	0.55	0.40
2-Ethylhexanol	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.25	0.32	0.58	0.26	1.00	0.21	0.61	0.24	0.42
3,5,5-Trimethyl-1-hexanol	0.21	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.30	0.45	0.52	0.21	0.32	0.21	1.00	0.20	0.46	0.29
2-Propylhepantan-1-ol	0.41	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.28	0.24	0.26	0.41	0.30	0.61	0.20	1.00	0.23	0.52
3,7-Dimethyl-1-octanol	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.29	0.58	0.50	0.20	0.55	0.24	0.46	0.23	1.00	0.37
2-Methyl-1-undecanol	0.41	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.29	0.36	0.29	0.40	0.42	0.29	0.52	0.37	1.00

Figure 5

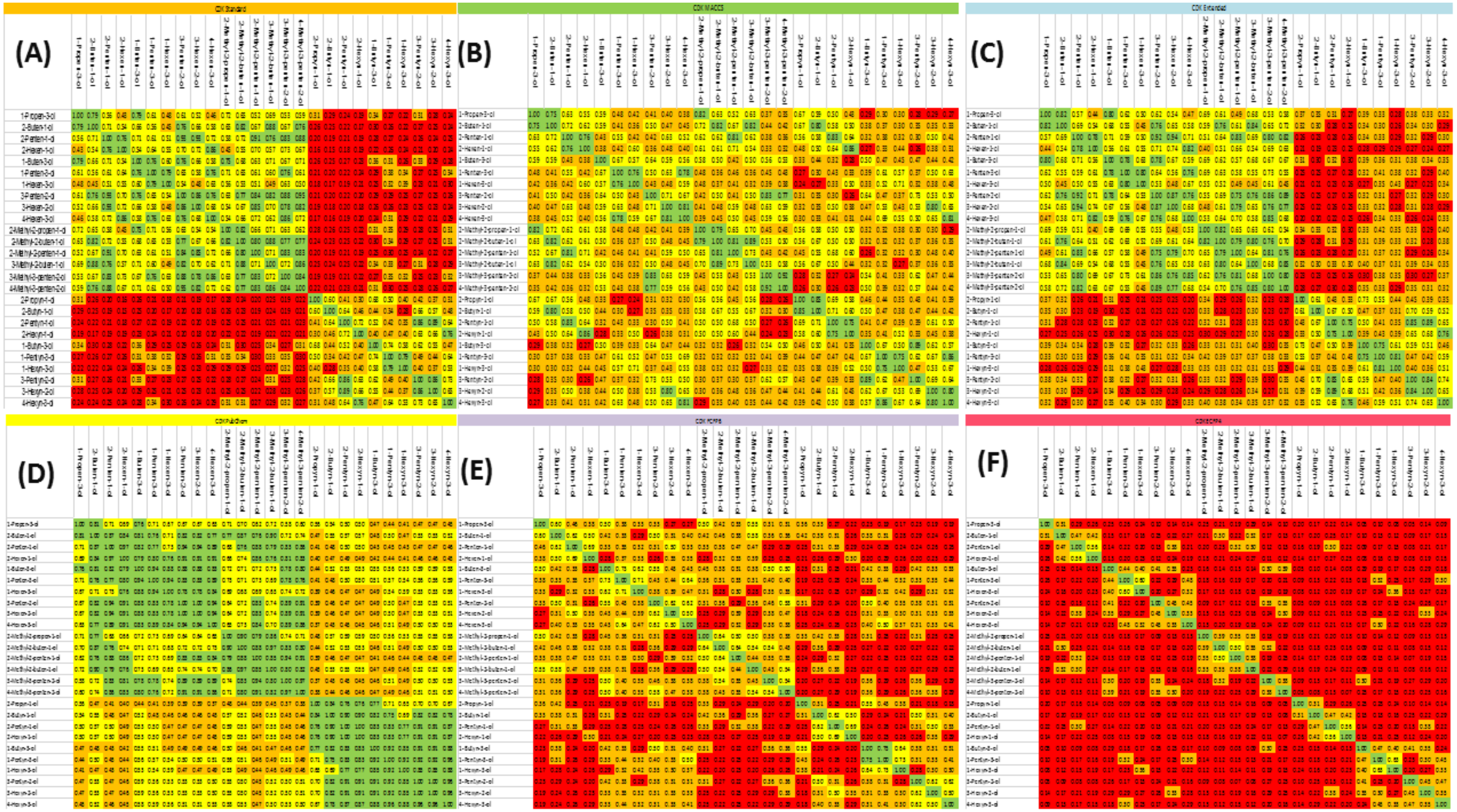

(A)
CDK Standard
(B)
CDK MACCS
(C)
CDK Extended
(D)
CDK PubChem
(E)
CDK FCFP6
(F)
CDK ECFP4

Figure 6