Supplemental Material

Chemical Similarity to Identify Potential Substances of Very High Concern – an Effective Screening Method

Pim N.H. Wassenaar[1,2], Emiel Rorije[1], Nicole M.H. Janssen[1], Willie J.G.M. Peijnenburg[1,2], Martina G. Vijver[2]

[1] *National Institute for Public Health and the Environment (RIVM), Centre for Safety of Substances and Products, P.O. Box 1, 3720 BA Bilthoven, The Netherlands*

[2] *Institute of Environmental Sciences (CML), Leiden University, P. O. Box 9518, 2300 RA Leiden, The Netherlands*

Outline

S.1 Dutch national Substances of Very High Concern

Within the Netherlands, national policy is particularly focusing on Dutch national Substances of Very High Concern (nSVHC). These substances could seriously harm man and environment and are therefore of very high concern. Although the nSVHC substances cover a broader range of chemicals than the EU-SVHC substances under REACH, nSVHC substances are identified based on the same hazard criteria as the EU-SVHC substances (i.e. REACH article 57; 1907/2006):

a. Carcinogenic category 1A or 1B according to Regulation (EC) 1272/2008.
b. Mutagenic category 1A or 1B according to Regulation (EC) 1272/2008.
c. Toxic for reproduction category 1A or 1B according to Regulation (EC) 1272/2008.
d. Persistent, Bioaccumulative and Toxic in accordance with the criteria set out in REACH Annex XIII.
e. Very Persistent and Very Bioaccumulative in accordance with the criteria set out in REACH Annex XIII.
f. Substances for which there is scientific evidence of probable serious effects to human health or the environment which give rise to an equivalent level of concern to those of other substances listed above, like endocrine disruptors.

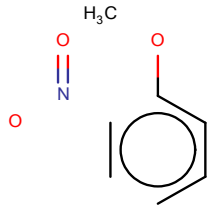
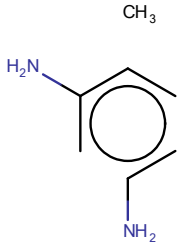
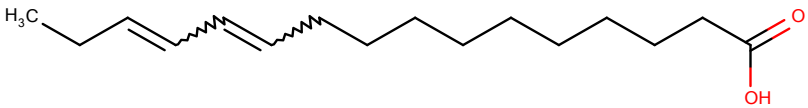
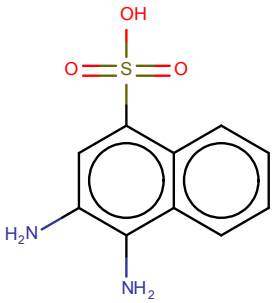
A substance is considered nSVHC when it is included on any of the following lists:

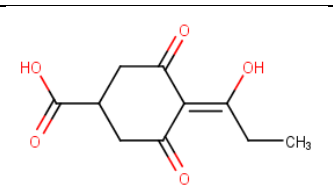
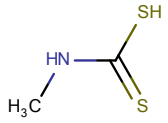
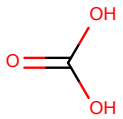
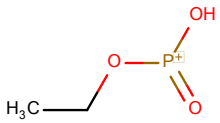
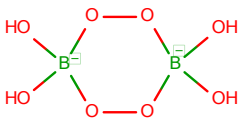
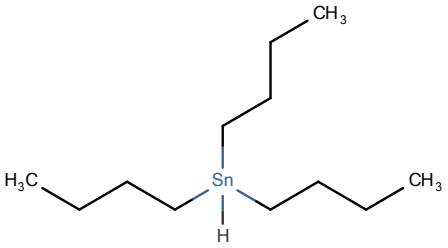
- Substances that are classified as C, M, or R category 1A or 1B according to Regulation (EC) 1272/2008.
- Substances on the candidate list for REACH Annex XIV.
- Substances that are identified as POP in the Stockholm Convention regulation (EC) 850/2004.
- Priority Hazardous substances according to the Water Framework Directive 2000/60/EC.
- Substances on the OSPAR list for priority action.

The list of nSVHC substances is compiled and updated on https://rvszoeksysteem.rivm.nl/ZZSlijst/Index.

S.2 SMILES charge conversion

SMILES were adjusted to neutral versions where possible (see Table below).

Functional group or salts of the functional group	**Neutral or Charged representation**	**Final structure (examples)**
Nitro	Neutral	
Quaternary amine	Charged	
Quaternary amine with 1-3 hydrogen atoms	Neutral expressed as primary, secondary or tertiary amine	
Carboxylic acid	Neutral	
Sulfonic acid	Neutral	

Alcohol	Neutral	
Tertiary carbon	Charged	
Thiol	Neutral	
Carbonate	Neutral	
Phosphonic acid	Charged	
Boron(IV)	Charged	
Tin(III)	Neutral (as Tin(IV))	

48

49

S.3 Model application

1. Generate SMILES

For substances of interest, SMILES / .sdf files need to be generated. The applicability domain should be taken into account (section 4.3) and charged structures should be converted to their neutral versions where possible (see Supplemental Material S.2). There are multiple possibilities to generate a correct SMILES code (e.g. non-canonical or canonical), these should provide similar outcomes.

2. Generate Fingerprint

For the substances of interest, fingerprints need to be generated:

- Extended fingerprint for CMR model.
- MACCS fingerprint for PBT/vPvB model.
- FCFP4 for ED model.

The extended fingerprint and MACCS fingerprint can be generated using PaDEL-Descriptor [23] (http://www.yapcwsoft.com/dd/padeldescriptor/). The following settings were enabled: "remove salt", "detect aromaticity", "standardize all tautomers" and "standardize nitro groups".

The FCFP4 fingerprint can be generated by using RDkit in python [22]. Python version 2.7 and RDkit version 2017.09.3.0 were applied. The following script can be used to generate the FCFP4 fingerprint:

```
### Load packages
from __future__ import print_function
from rdkit import Chem
from rdkit.Chem import AllChem
import csv
import os

### Set working directory
os.chdir("C:/...)

### Import .sdf file
suppl = Chem.SDMolSupplier("C:/….sdf")

### Check SMILES
m = [x for x in suppl if x is not None]

### Calculate FCFP4 fingerprint
Fingerprint_FCFP4 = [AllChem.GetMorganFingerprintAsBitVect(x, 2,
useFeatures=True, nBits=1024) for x in m]

### Export fingerprint
with open('FCFP4_fp_TestCase.csv', 'w') as output:
    writer = csv.writer(output, lineterminator='\n')
    writer.writerows(Fingerprint_FCFP4)
```

3. Calculate similarity

In order to run the models, the generated fingerprints need to be order in separate .csv files with in the first three columns: "Name", "CAS or EC" and "SMILES" (Note: these columns could be left blank). In the other columns each fingerprint bit should be placed (n=166 for MACCS and n=1024 for the Extended Fingerprint and FCFP4).

The files need to be ordered in the following folder structure in order to run the R-script as shown below. Note that the working directory and files location need to be adjusted within this script.

- <u>Folder:</u> R_import_files:
    - CMR_ExtendedFingerprint (Sheet 3 from Supplemental Material Excel as .csv file)
    - PBT_MACCS (Sheet 4 from Supplemental Material Excel as .csv file)
    - ED_FCFP4 (Sheet 5 from Supplemental Material Excel as .csv file)
    - Subfolder: Test_data:
        - File_CMR (the ExtendedFingerprint file as generated for the substances of interest)
        - File_PBT (the MACCS file as generated for the substances of interest)

83 ▪ File_ED (the FCFP4 file as generated for the substances of interest)

84 - Folder: R_export_files.

```
# --------------------------
# Load Packages
# --------------------------

### Load packages
library("caret")
library("ChemmineR")
library(caTools)
library(xlsx)
library(ROCR)
library(dplyr)

### Set working directory
setwd("C:..../R_export_files")

# --------------------------
# Load similarity measures
# --------------------------

### CMR
CMR_Substances <- read.csv("C:..../R_Import_files/CMR_ExtendedFingerprint.csv", sep=";")
CMR_Substances <- filter(CMR_Substances, CMR_Substances$CMR == 1)

### PBT/vPvB
PBT_Substances <- read.csv("C:..../R_Import_files/PBT_MACCS.csv", sep=";")
PBT_Substances <- filter(PBT_Substances, PBT_Substances$PBT.vPvB == 1)

### ED
ED_Substances <- read.csv("C:..../R_Import_files/ED_FCFP4.csv", sep=";")
ED_Substances <- filter(ED_Substances, ED_Substances$ED == 1)

### Similarity coefficients
SS3 <- function(a,b,c,d){ifelse(c==(a+b+c+d),1,ifelse(d==(a+b+c+d),1,ifelse(c==0 &
      d==0,0,ifelse(c==0 & a ==0,
      ((1/4)*(((c)/(c+b))+((d)/(a+d))+((d)/(b+d)))),((1/4)*(((c)/(c+a))+((c)/(c+b))+((d)/(a
      +d))+((d)/(b+d))))))))}
SM <- function(a,b,c,d){(c+d)/(c+a+b+d)}
CT4 <- function(a,b,c,d){(log(1+c))/(log(1+c+a+b))}

### Thresholds
CMR_Threshold_Below <- 0.85054337568321992
CMR_Threshold_Above <- 0.9443359375
PBT_Threshold <- 0.96987951807228912
ED_Threshold <- 0.86632190004714749
```

86

```
# --------------------------
# Compare similarity - Test data
# --------------------------

### CMR
CMR_test_data <- read.csv("C:..../R_Import_files/Test_data/File_CMR.csv", sep=";")
Top1_CMR_test_data <- apply(CMR_test_data[,c(4:1027)],MARGIN = 1, function(x) ifelse(sum(x)
      < 85,fpSim(x, y=data.matrix(CMR_Substances[,c(12:1035)]), method = CT4,
      top=1),fpSim(x, y=data.matrix(CMR_Substances[,c(12:1035)]), method = SM, top=1)))

CMR_Results <- CMR_test_data[,1:3]
names(CMR_Results) <- c("Identifier","CAS","SMILES")
CMR_Results$CMR_SimValue <- Top1_CMR_test_data
CMR_Results$CMR_Concern <- apply(CMR_test_data[,c(4:1027)],MARGIN = 1, function(x)
      ifelse(sum(x) < 85, ifelse(fpSim(x, y=data.matrix(CMR_Substances[,c(12:1035)]),
      method = CT4, top=1) >= CMR_Threshold_Below, "Yes", "No"),ifelse(fpSim(x,
      y=data.matrix(CMR_Substances[,c(12:1035)]), method = SM, top=1) >=
      CMR_Threshold_Above, "Yes", "No")))
CMR_Results$CMR_MostSimilar_Name <- c(NA)
CMR_Results$CMR_MostSimilar_SMILES <- c(NA)
MostSimilarID <- apply(CMR_test_data[,c(4:1027)],MARGIN = 1, function(x) which.max(fpSim(x,
      y=data.matrix(CMR_Substances[,c(12:1035)]), method = SM, sorted=FALSE)))
CMR_Results$CMR_MostSimilar_Name <- as.character(CMR_Substances[MostSimilarID,2])
CMR_Results$CMR_MostSimilar_SMILES <- as.character(CMR_Substances[MostSimilarID,3])
CMR_Results$CMR_NumberSimilar <- apply(CMR_test_data[,c(4:1027)],MARGIN = 1, function(x)
      ifelse(sum(x) < 85, sum(fpSim(x, y=data.matrix(CMR_Substances[,c(12:1035)]),method =
      CT4, sorted=FALSE)>= CMR_Threshold_Below), sum(fpSim(x,
      y=data.matrix(CMR_Substances[,c(12:1035)]),method = SM, sorted=FALSE)>=
      CMR_Threshold_Above)))

### PBT
PBT_test_data <- read.csv("C:..../R_Import_files/Test_data/File_PBT.csv", sep=";")
Top1_PBT_test_data <- apply(PBT_test_data[,4:169],MARGIN = 1, function(x) fpSim(x,
      y=data.matrix(PBT_Substances[,12:177]), method = SM, top=1))

PBT_Results <- PBT_test_data[,1:3]
names(PBT_Results) <- c("Identifier","CAS","SMILES")
PBT_Results$PBT_SimValue <- Top1_PBT_test_data
PBT_Results$PBT_Concern <- apply(PBT_test_data[,c(4:169)],MARGIN = 1, function(x)
      ifelse(fpSim(x, y=data.matrix(PBT_Substances[,12:177]), method = SM, top=1) >=
      PBT_Threshold, "Yes", "No"))
PBT_Results$PBT_MostSimilar_Name <- c(NA)
PBT_Results$PBT_MostSimilar_SMILES <- c(NA)
MostSimilarID <- apply(PBT_test_data[,c(4:169)],MARGIN = 1, function(x) which.max(fpSim(x,
      y=data.matrix(PBT_Substances[,c(12:177)]), method = SM, sorted=FALSE)))
PBT_Results$PBT_MostSimilar_Name <- as.character(PBT_Substances[MostSimilarID,2])
PBT_Results$PBT_MostSimilar_SMILES <- as.character(PBT_Substances[MostSimilarID,3])
PBT_Results$PBT_NumberSimilar <- apply(PBT_test_data[,c(4:169)],MARGIN = 1, function(x)
      sum(fpSim(x, y=data.matrix(PBT_Substances[,c(12:177)]),method = SM, sorted=FALSE)>=
      PBT_Threshold))
```

88

```
### ED
ED_test_data <- read.csv("C:..../R_Import_files/Test_data/File_ED.csv", sep=";")
Top1_ED_test_data <- apply(ED_test_data[,4:1027],MARGIN = 1, function(x)      fpSim(x,
      y=data.matrix(ED_Substances[,12:1035]), method = SS3, top=1))

ED_Results <- ED_test_data[,1:3]
names(ED_Results) <- c("Identifier","CAS","SMILES")
ED_Results$ED_SimValue <- Top1_ED_test_data
ED_Results$ED_Concern <- apply(ED_test_data[,c(4:1027)],MARGIN = 1, function(x)
      ifelse(fpSim(x, y=data.matrix(ED_Substances[,12:1035]), method = SS3, top=1) >=
      ED_Threshold, "Yes", "No"))
ED_Results$ED_MostSimilar_Name <- c(NA)
ED_Results$ED_MostSimilar_SMILES <- c(NA)
MostSimilarID <- apply(ED_test_data[,c(4:1027)],MARGIN = 1, function(x) which.max(fpSim(x,
      y=data.matrix(ED_Substances[,c(12:1035)]), method = SS3, sorted=FALSE)))
ED_Results$ED_MostSimilar_Name <- as.character(ED_Substances[MostSimilarID,2])
ED_Results$ED_MostSimilar_SMILES <- as.character(ED_Substances[MostSimilarID,3])
ED_Results$ED_NumberSimilar <- apply(ED_test_data[,c(4:1027)],MARGIN = 1, function(x)
      sum(fpSim(x, y=data.matrix(ED_Substances[,c(12:1035)]),method = SS3, sorted=FALSE)>=
      ED_Threshold))


# --------------------------
# Export data
# --------------------------

TestData_Results <- cbind(CMR_Results, PBT_Results[,c(4:8)], ED_Results[,c(4:8)])
write.xlsx(TestData_Results, "TestData_Results.xlsx", col.names = TRUE, row.names = TRUE)
```

92 *S.4 Symmetric coefficient bias*

93 For the CMR dataset specifically, we adjusted the best performing model by using a symmetric-
94 asymmetric coefficient combination as all small substances were classified as positive. Although the
95 PBT/vPvB and ED models are also based on a symmetric similarity coefficient, they do not require a
96 symmetric-asymmetric combination, as the models have slightly different characteristics compared to the
97 CMR subgroup. The PBT/vPvB model is based on the MACCS fingerprint, which consists of only 166
98 bits. With a similarity threshold of 0.970, substances with five or less different bit-pairs will always be
99 considered as similar. As the lowest number of fragments in any of the PBT/vPvB substances is already
100 six, small substances in the reference datasets are not automatically identified as structurally similar to
101 PBT/vPvB SVHCs (as was the case for the CMR SVHC subgroup). The ED subgroup, where the FCFP4
102 fingerprint gave the best predictive performance, has a much better balance in ED and non-SVHC
103 fragment distribution (Figure S.2). Additionally, no ED substances with a low fragment count are
104 included and the fragments are more specific. Furthermore, the optimal ED model uses the SS3
105 coefficient, which takes *c* and *d* bit-pairs equally into account, but does not consider them as exactly
106 similar, as the SM coefficient does (Table 2).The PBT/vPvB and ED models therefore do not require a
107 combination of asymmetric and symmetric coefficients.

108

S.5 CMR model extension with ToxTree and DART Structural Alerts (Addition of extra fingerprint)

The best observed accuracy for the subset of CMR substances was 0.819, and is lowest for all subsets (i.e. CMR, PBT/vPvB and ED). A test was conducted in order to analyze whether the accuracy could be improved by adding a CMR specific fingerprint – containing (larger/specific) structural alerts that are related to CMR properties. Potentially, such CMR-specific fragments could improve the performance and fill the information gap of the plain similarity measures.

We developed a CMR-specific dictionary-based fingerprint, based on structural alerts as included in ToxTree (for C and M) [7] and DART classification scheme (for R) [34]. The CMR-fingerprint contained a total of 115 bits (35 CM related from ToxTree; 80 R related from DART). This fingerprint was combined with the seven selected similarity coefficients (Table 2), resulting in seven different "fingerprint-coefficient" combinations. Subsequently, these seven "fingerprints-coefficient" combinations were combined with the CMR model (i.e. "Extended fingerprint – SM coefficient" combination) using different weights, by using the following equation:

$$S = S_{CMR-FP} * W_{CMR-FP} + S_{Overall\ CMR} * W_{Overall\ CMR}$$

Where, S represents the final similarity value per substance. This similarity value is subsequently used to determine the final model performance similar as described in section 2.4 (i.e. determination of optimal threshold and calculation of balanced accuracy) $S_{CMR\text{-}FP}$ are the highest similarity values for a substance to a CMR-SVHC substance, as obtained by using the CMR-specific fingerprint and one of the seven similarity coefficients. $S_{Overall\ CMR}$ are the highest similarity values for a substance to a CMR-SVHC substance, as obtained by using the "Extended-fingerprint - SM coefficient" combination. $W_{CMR\text{-}FP}$ and $W_{Overall\ CMR}$, represent the weights given to the different similarity values. The applied weight combinations are shown in the Table below. By using this scheme the performance of 71 models was obtained (i.e. 10 weight combination * 7 coefficients + 1 weight combination [$W_{CMR\text{-}FP} = 0$, $W_{Overall\ CMR} = 1$]).

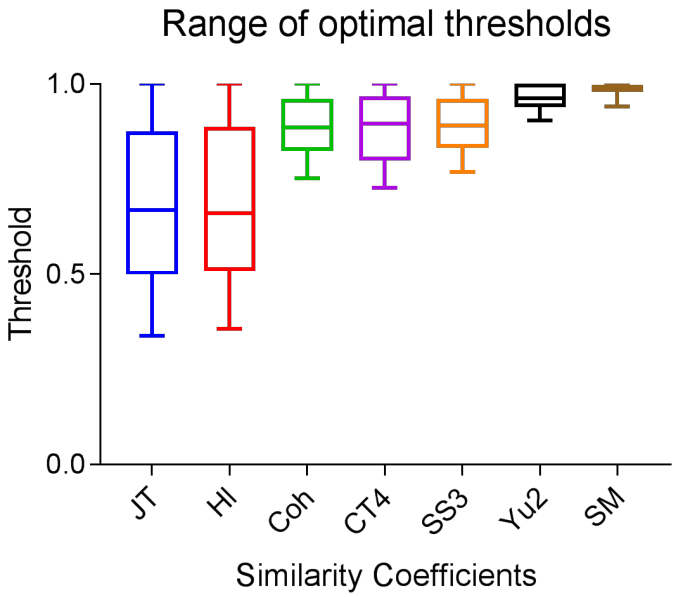
$W_{CMR\text{-}FP}$	$W_{Overall\ CMR}$
1	0
0.9	0.1
0.8	0.2
0.7	0.3
0.6	0.4
0.5	0.5
0.4	0.6
0.3	0.7
0.2	0.8
0.1	0.9
0	1

Of all models, the $W_{CMR\text{-}FP}$ = 0 resulted in highest balanced accuracy (0.819). This model is exactly similar to the best overall model ("Extended-fingerprint - SM coefficient" combination; thus without inclusion of the CMR-specific fingerprint). In addition, all models based on the Yu2-coefficient (except Yu2 with $W_{CMR\text{-}FP}$ = 1) and the SM-coefficient $W_{CMR\text{-}FP}$ = 0.1 had a similar accuracy to the best model, indicating that these models do not influence the model performance. All other models resulted in a lower balanced accuracy, with a lowest balanced accuracy for all $W_{CMR\text{-}FP}$ = 1 models. This indicates that the CMR-FP do not provide additional information for an improved distinction between CMR and non-CMR substances (see Table below). All weighing values in between resulted in balanced accuracies between the extreme values. It is observed that the asymmetric coefficient (i.e. JT and CT4) perform much better than the symmetric coefficient. This can be explained by the fact that only a few alerts are present per substance, and thus many zero fingerprint bit values are included.

$W_{CMR\text{-}FP}$ = 1	Balanced Accuracy
CMR-FP_JT	0.651
CMR-FP_CT4	0.651
CMR-FP_Hl	0.501
CMR-FP_SS3	0.501
CMR-FP_Coh	0.501
CMR-FP_SM	0.500
CMR-FP_Yu2	0.500

146 *Figure S.1. Optimal threshold values for the analyzed similarity coefficients in combination with the*

147 *sixteen investigated fingerprints.*



148

149

Figure S.2. Distribution of fragments (i.e. "1-bits") across TP, FP, TN and FN substances. 1) for PBT/vPvB using the MACCS fingerprint, 2) for ED using the FCFP4 fingerprint, and 3) for CMR using the extended fingerprint and CT4-SM combination.

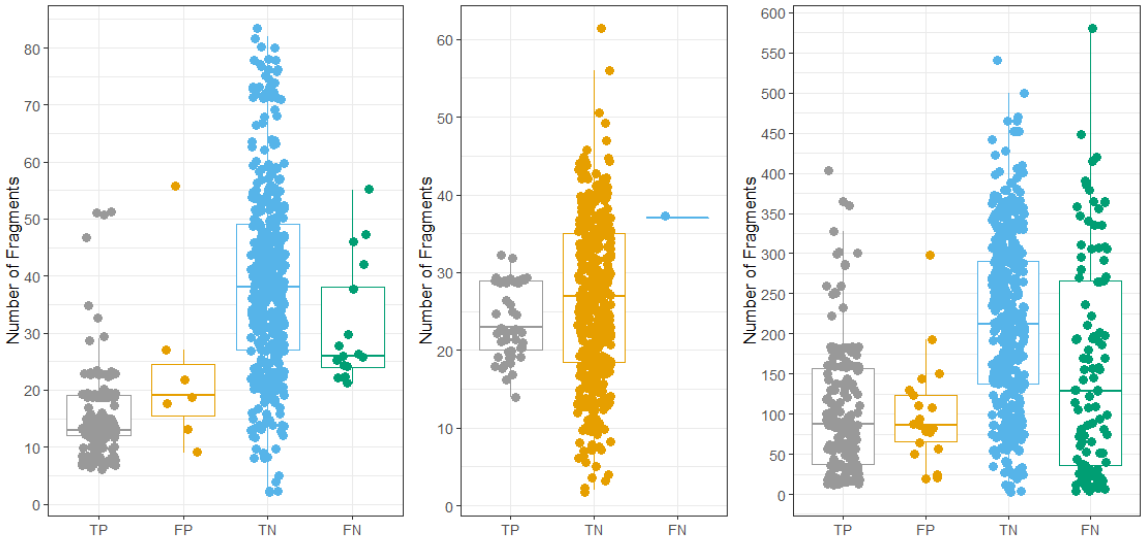
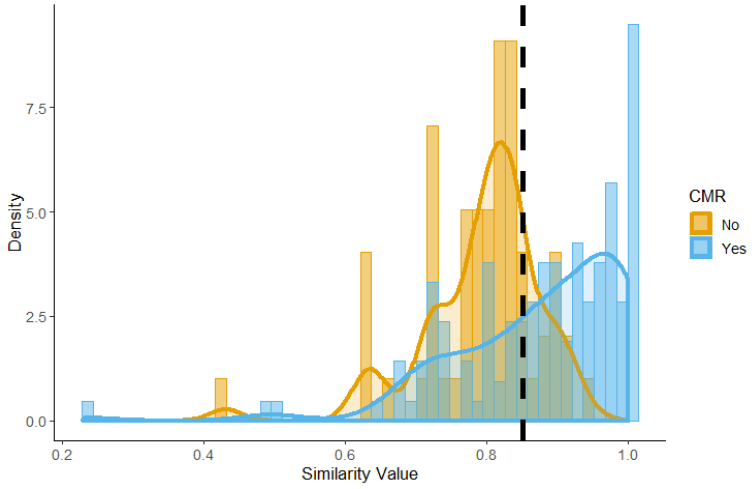
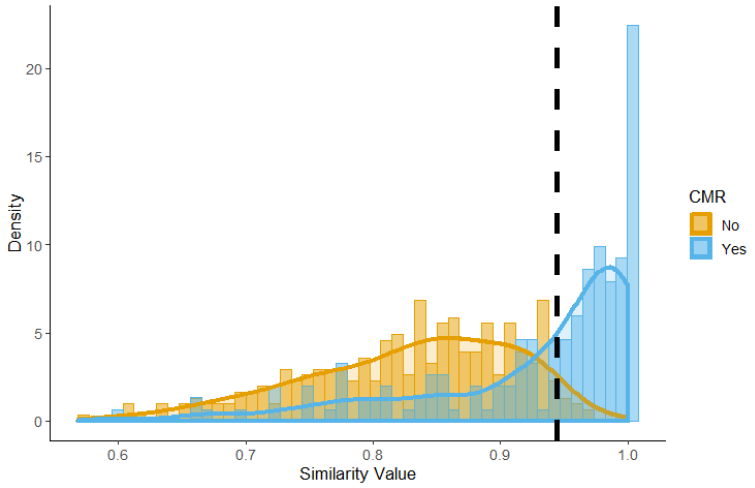
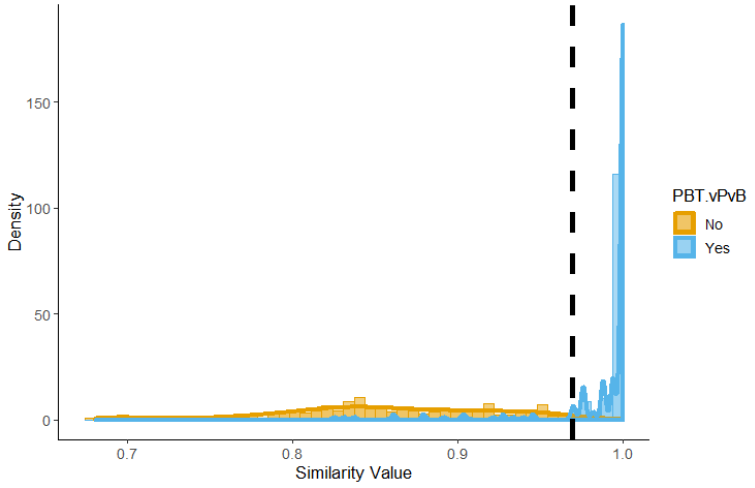
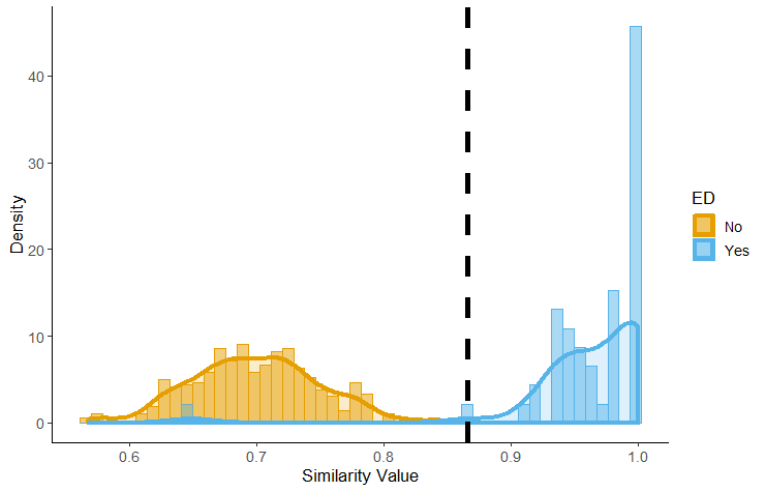


Figure S.3. Highest similarity values as calculated for 1) CMR CT4, 2) CMR SM, 3) PBT/vPvB, and 4) ED substances and non-SVHC substances (based on the best performing models). The vertical dashed line represents the optimal threshold.








Density
40
30
20
10
0
0.6
0.7
0.8
0.9
1.0
Similarity Value
ED
No
Yes

164 *Table S.1. Best performing fingerprint-coefficient combination for the CMR subgroups based on one*
165 *similarity coefficient; and the improved CMR model by combining a symmetric and asymmetric*
166 *coefficient in order to prevent symmetric coefficient bias. In total, 411 non-SVHC substances were*
167 *included. '-' means that it is not possible to calculate a single AUC or threshold value for a combination of*
168 *two models. AUC is the area under the curve of ROC-plot.*

Subset	Model		Threshold	Sensitivity	Specificity	Precision	AUC (ROC)	Balanced accuracy
	Fingerprint	Coefficient						
CMR (n=306)	Extended	SM (<85)	0.944	0.978	0.222	0.728	0.832	0.600
		SM (≥85)	0.944	0.634	0.968	0.908	0.826	0.801
		Total	0.944	0.784	0.854	0.800	0.859	0.819
CMR improved (n=306)	Extended	CT4 (<85)	0.851	0.672	0.841	0.900	0.748	0.756
		SM (≥85)	0.944	0.634	0.968	0.908	0.826	0.801
		Total	-	0.650	0.949	0.905	-	0.800

169

170 *Table S.2. Physicochemical applicability domain for the similarity models based on the 95th percentiles of*
171 *the dataset substances.*

Properties	CMR	PBT/vPvB	ED
Molecular weight	59 – 632	100 – 717	70 – 556
Log K_{ow}	2.19 – 9.40	-1.62 – 10.20	-2.42 – 7.7
Number of atoms	7 – 84	12 – 70	11 – 84
Number of rings	0 – 5	0 – 6	0 – 4
Number of aromatic rings	0 – 5	0 – 4	0 – 3

172