

## User manual for the ZZS similarity tool

The ZZS structural similarity model is an online tool designed to predict whether a substance is structurally similar to a ZZS (*i.e. Dutch Substance of Very High Concern - in Dutch: Zeer Zorgwekkende Stof*). Structural similarity could be an indication of similar toxicity profiles. The model is based on the methodology as described and evaluated by Wassenaar et al. (2019<sup>1</sup>, 2021<sup>2</sup> and 2022<sup>3</sup>).

The online tool is available at: <https://rvszoekstelsysteem.rivm.nl/ZzsSimilarityTool> (see Figure 1). This user manual describes the technical procedure, the required input and the way the output is presented. It consists of the sections: Background, Input, Output and Technical remarks on the (implementation of the) methodology. To facilitate the implementation of the similarity models in this online tool, only some minor changes had to be made in the algorithm (see section “Technical remarks on the methodology”).



Figure 1. Link to the ZZS similarity tool.

### New in this version

A first version of the tool has been launched in January 2020. In January 2022, the tool has been updated. In this new version the following elements are added and updated:

- **CAS-search:** Added the possibility to search an input chemical based on CAS-number. In the previous version it was only possible to provide an input chemical based on SMILES.
- **Batch-search:** Added the possibility to calculate and retrieve the results for a list of input chemicals at once (with a maximum of 200 chemicals).
- **Update of the underlying dataset:** All ZZS chemicals identified till January 25, 2021 are added to the ZZS similarity tool. In the previous version only ZZS till March 1, 2018 were included.
- **Re-optimization of the models:** The ZZS data are separated into five categories to better reflect the chemical concerns (CM, R, PBT/vPvB, ED and Others), instead of three (CMR, PBT/vPvB and ED). As new ZZS are added and separated into new categories, the similarity models were re-optimized.
- **Improved outcome interpretation:** Addition of a quantitative confidence score, besides the qualitative conclusion (sufficiently similar: yes/no), to support better outcome interpretation.

<sup>1</sup> <https://doi.org/10.1016/j.comtox.2019.100110>

<sup>2</sup> <https://doi.org/10.1016/j.yrtph.2020.104834>

<sup>3</sup> <https://doi.org/10.1002/jcc.26859>

## Background

We have developed a structural similarity tool that allows a comparison between a non-classified substance and a ZS. The methodology as implemented in the ZS similarity tool is based on the similar property principle, which states that structural similar substances are likely to have similar toxic properties. Therefore, structural similarity between a substance and a known ZS might be an indication of comparable effects, and could be a trigger for further inspection and analysis.

ZS hazard properties are defined by REACH article 57:

- a-c: CMR (also known as: carcinogenic, mutagenic and reprotoxic).
- d-e: PBT/vPvB (also known as: Persistent, bioaccumulative and Toxic / very persistent and very bioaccumulative)
- f: equivalent level of concern. In the ZS similarity tool only ED-properties (also known as: endocrine disruptor) are modelled as endpoint of this group.

The ZS similarity tool is able to calculate whether a substance is structural similar to a ZS substance with CM, R, PBT/vPvB or ED properties – in this manual also referred to as “categories”. Substances on the ZS-list that do not belong to any of the above-mentioned categories were included in the ‘Other’-category, like substances on the European SVHC list with persistent, mobile and toxic (PMT) properties, specific target organ toxicity after repeated exposure (STOT-RE) or sensitizing properties. In addition, substances were also included in the ‘Other’-category when they were only included on the ZS-list based on other sources, like substances on the OSPAR list for priority action or priority hazardous substances according to the Water Framework Directive.

In this section we shortly describe the underlying methodology <sup>4</sup>. Structural similarity is expressed based on a similarity measure which consist of a binary fingerprint and a similarity coefficient. A fingerprint is a string of binary values that can be assigned to a structure and can be calculated based on the SMILES of a structure (i.e. a specific chemical identifier; see also section ‘Input’). These fingerprints have a specific length (e.g. 1024) and consist of 1’s and 0’s (see Figure 2 for an illustration). The fingerprints of a substance and a ZS substance can be compared by a similarity coefficient which calculates a similarity value between 0 and 1. In which 0 indicates that two substances are totally different and 1 indicates that two substances are identical (see Figure 3 for an illustration).

There are many different fingerprints and similarity coefficients available and based on the analysis as described in <sup>3</sup>, the best combination was selected for chemicals of the different hazard categories: CM, R, PBT/vPvB, ED and Other. Next, we derived an optimum similarity threshold (value between 0-1) for the different categories, so that we optimized the accuracy (i.e. false/true positives/negatives). When a new chemical with unknown toxicity is compared to a ZS, and the similarity value is above this threshold, it is considered to be structurally comparable to the ZS. Consequently, this chemical could be of potential concern with respect to the ZS hazard endpoints in that category. These best performing models (with corresponding thresholds) have been implemented in the online ZS similarity tool (see also section ‘Technical remarks on the implementation of the methodology’). Note that there are some more details/specifications applicable to the incorporated models for CM, R, PBT/vPvB, ED and Other, which can be found in <sup>3</sup>. The full workflow of the ZS similarity tool is shown in Figure 4.

---

<sup>4</sup> More details can be found in <sup>3</sup>; in which we analyzed a large number of methodologies that are available in the open literature to asses structural similarity between substances.

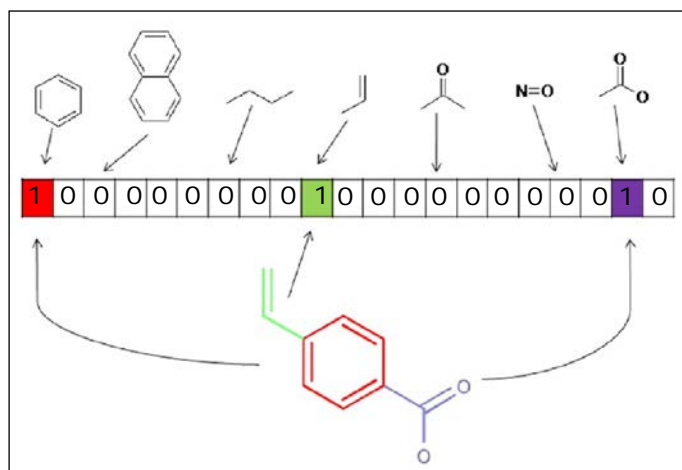
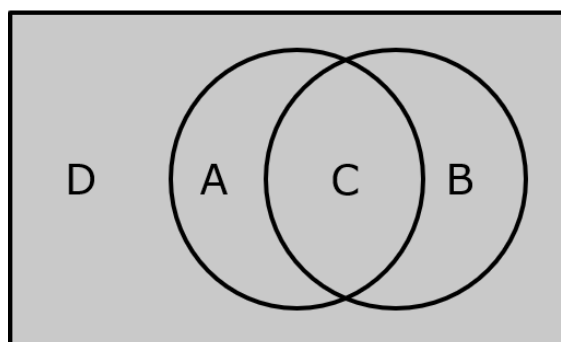


Figure 2: Illustration of a fingerprint, a numerical expression of a chemical structure (note that multiple types of fingerprints exist). Each bit in a bit-string is related to a specific fragment. If this fragment is present in a substance, the bit is set "On" and a score of 1 is provided, if a fragment is not present in a substance, the bit is set "Off" and a score of 0 is provided to the bit. A fingerprint of a structure thus consists of 1's and 0's. Fingerprints of two substances can be compared on structural similarity by using a similarity coefficient (see Figure 3). Figure adopted from Cao et al. [Analytica Chimica Acta 752 (2012): 1-10].




---

**Substance a:** 1 0 0 1 1 1 0 1 0 1  
**Substance b:** 1 0 1 0 0 1 1 1 0 0

---

C D B A A C B C D A

---

$$S_{ab} = \frac{N_C}{N_A + N_B + N_C}$$

$$S_{ab} = \frac{3}{3 + 2 + 3} = 0.375$$

Figure 3: Illustration of a similarity coefficient, a formula to quantitatively express the similarity between two fingerprints with a value between 0 and 1. A score of 0 indicates that two substances are totally different and 1 indicates that two substances are identical. On the top-right two fingerprints are shown. If a fragment is present in both substances, it is called a "C" fragment; if a fragment is not present in both substances, it is called a "D" fragment; if the fragment is only present in substance A, it is called a "A" fragment; and if the fragment is only present in substance B, it is called a "B" fragment (also see the left picture). The number of A, B, C and D fragments are counted and can be used in a similarity coefficient formula to express the similarity between substance A and B (see bottom-right for an example) (note that multiple different similarity coefficients exist).

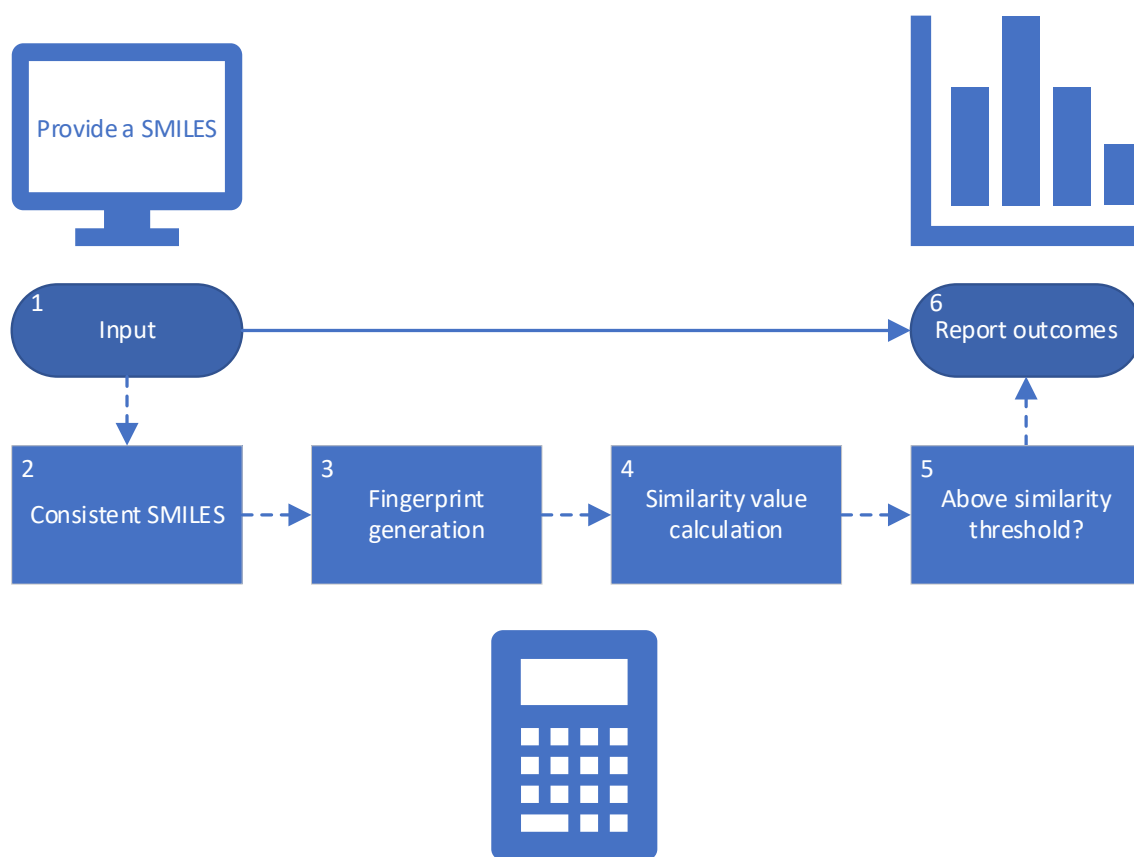


Figure 4. Workflow of the ZZS similarity model. Step 1 and 6 consider the input and output as shown by the ZZS similarity tool. Step 2-5 consider processes that take place automatically when pressing “calculate”. Step 1: Provide an input substance in the form of a CAS-number or SMILES (see section “Input”). Step 2: The computer standardizes the provided SMILES to ensure equal comparison to ZZS structures. Step 3: Two different fingerprints are generated for the consistent/standardized SMILES (see <sup>3</sup> for more details). Step 4: The fingerprints of the input substance are compared to the fingerprints of all the CM, R, PBT/vPvB, ED and Other ZZS, respectively. Different similarity coefficients are applied for the different categories (see <sup>3</sup> for more details). Step 5: The computer analyses whether the similarity values of the input substance to all ZZS are above or below the similarity threshold. Different similarity thresholds are applied for the different categories (see <sup>3</sup> for more details). Step 6: The results of the ZZS similarity tool are reported (see section “Output”).

## Input

As input, a SMILES-code of a chemical structure or a CAS-number needs to be provided (see Figure 5).

Preference should be given to CAS-search. A list of 700,000 CAS numbers is connected to the database and have a related pre-programmed SMILES that will be used for the calculations. These SMILES are neutralized and standardized to best facilitate the similarity calculations.

If a CAS-number is not available, a SMILES (i.e. Simplified Molecular Input Line Entry System) can be used as input. A SMILES represents a chemical structure by a line notation. This line notation can be interpreted by computer systems. More information on SMILES expression can be found elsewhere (e.g. <sup>5,6,7</sup>).

For batch-searches, both SMILES and CAS-numbers could be used in combination as long as each identifier is divided by a space, a comma, a vertical line (|) or by a new line (see Figure 6).

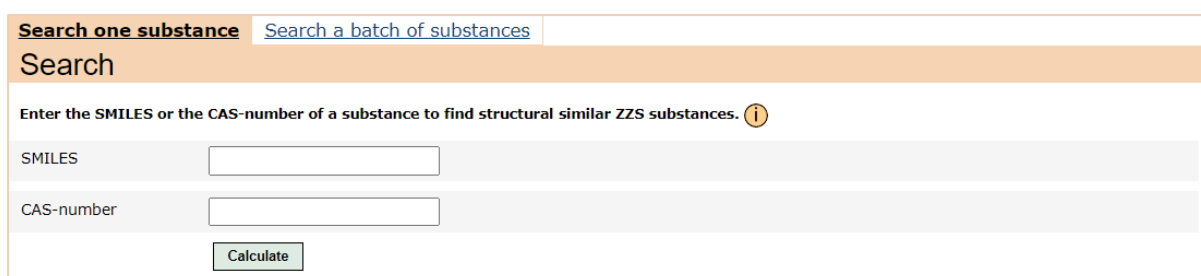


Figure 5. Input screen "Search one substance".

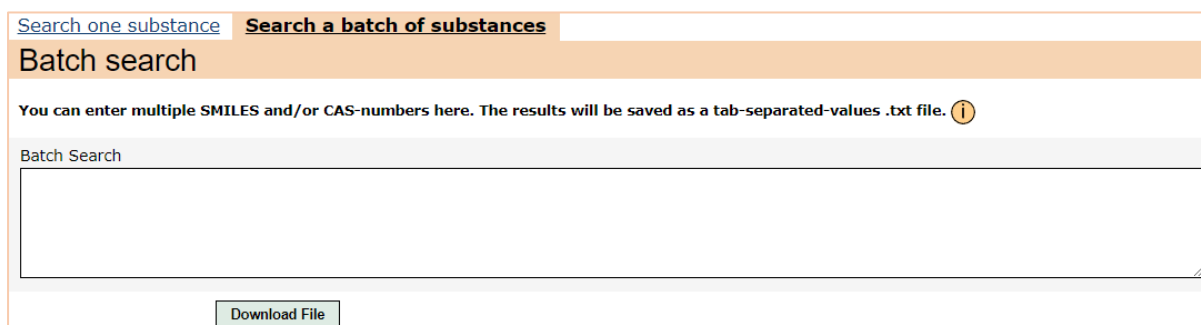


Figure 6. Input screen "Search a batch of substances".

If you do not know the SMILES of your substance you can search for the SMILES on several websites, like:

- EPA's chemistry dashboard/DSSTox (<https://comptox.epa.gov/dashboard>)  
Search compound by CAS or Name. SMILES can be found in section structural identifiers.
- Cactus (<https://cactus.nci.nih.gov/chemical/structure>)  
Search compound by CAS or Name and convert to SMILES.
- PubChem (<https://pubchem.ncbi.nlm.nih.gov/>)  
Search compound by CAS or Name. SMILES can be found in section 2.1.4.

<sup>5</sup> <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

<sup>6</sup> [https://www.daylight.com/dayhtml\\_tutorials/languages/smiles/index.html](https://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html)

<sup>7</sup> [https://www.daylight.com/dayhtml\\_tutorials/languages/smiles/smiles\\_examples.html](https://www.daylight.com/dayhtml_tutorials/languages/smiles/smiles_examples.html)

Note that the ZZS similarity tool requires a SMILES notation that represents one specific structure. This means that mixtures or UVCBs<sup>8</sup> cannot be assessed as a whole. Therefore, dots – which separate multiple structures within one SMILES notation – are not allowed (i.e. “.”). For mixtures or UVCBs, it is therefore advised to provide all structures, or several representative structures to the ZZS similarity tool consecutively (or in batch-search). When the SMILES notation contains atoms that are expressed with a charge, the ZZS similarity tool will transform these atoms/SMILES to a neutral form, where possible (see also <sup>3</sup>).

It is advised to always evaluate whether the input structure resembles your chemical of interest. This can be done by using the visualization of the input structure in the output screen (see the next section “Output”).

The structure similarity models are not applicable to arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt-metal derivatives. For these chemicals, the metal atoms (or ions) are thought to be the cause of concern, irrespective of the (organic) groups present in the inorganic molecule. These metal-based complexes are by definition predicted to be ZZS (see Figure 7). However, the models can be used to generate a first prediction for non-dissociating metals (e.g. organotin substances).

Besides when entering a SMILES, several error messages could appear. When you enter a CAS-number that is not in the database, the error as shown in Figure 8 will appear. When you enter an invalid SMILES or CAS-number, or when the SMILES or CAS-number is not recognized as valid by the system, the errors as shown in Figure 9 will appear. When the entered SMILES is too long or could not be processed by the system, the errors as shown in Figure 10 will appear. Furthermore, you are only allowed to use one search field at a time (i.e. either using CAS-number or SMILES). If both fields are filled the error as shown in Figure 11 will appear.

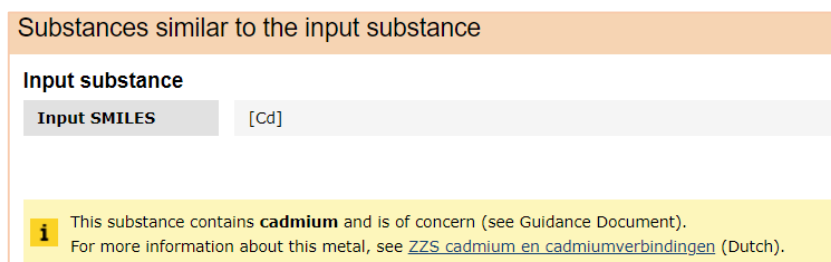


Figure 7. Example output when an arsenic, beryllium, cadmium, chromium, lead, mercury, nickel or cobalt-metal derivative is provided as input.

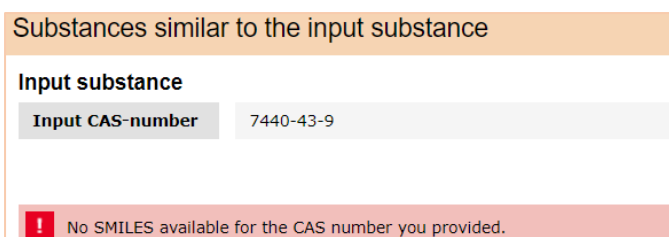


Figure 8. Example if a CAS number is no in the database

<sup>8</sup> Substances of unknown variable composition, complex reaction products or biological materials.

**Search**

Enter the SMILES or the CAS-number of a substance to find structural similar ZZS substances. ⓘ

SMILES  **The input is not a valid SMILES code.**

CAS-number  **The input is not a valid CAS-number.**

Figure 9. Errors when the input is not a valid SMILES code or CAS-number, or when the input is not recognized as a valid SMILES/CAS.

**Search**

Enter the SMILES or the CAS-number of a substance to find structural similar ZZS substances. ⓘ

SMILES  **The field SMILES must be a string or array type with a maximum length of '750'.**

CAS-number

**!** An error occurred while processing the SMILES you provided.

**!** The generation of fingerprints took longer than the maximum allowed duration of 30 seconds.

Figure 10. Errors when the input SMILES is too long (above 750 characters), or when the SMILES could not be processed.

**!** > You cannot search for a SMILES and a CAS-number at the same time.

Enter the SMILES or the CAS-number of a substance to find structural similar ZZS substances. ⓘ

SMILES

CAS-number

Figure 11. Error message if both input fields are used simultaneously.

## Output

The output screen consists of information on the input substance and information on the results.

### Input substance

Within the output screen, also information on the input structure is provided (see Figure 12). This information can be used to check whether the input structure resembles your chemical of interest (specifically the molecular structure):

- **Input SMILES/CAS-number:** Shows the SMILES notation or CAS-number you provided.
- **Consistent SMILES:** Shows the SMILES notation that is used by the model to estimate chemical similarity. This step is used to ensure uniformity in SMILES notation for the input structure and ZZS SMILES. Standardization for instance includes the neutral expression of atoms, where possible (see also <sup>3</sup> for more details).
- **Molecular structure:** Shows the molecular structure of the input substance.

Input substance	
Input SMILES	<chem>c1ccccc1CC</chem>
Consistent SMILES	<chem>c1ccc(cc1)CC</chem>
Molecular structure	

Figure 12. Example of the output screen, showing information on the input substance.

### Similarity to ZZS substances

The results are prioritized (i.e. most similar ZZS on top) and categorized into similarity with respect to ZZS hazard categories CM<sup>9</sup>, R<sup>10</sup>, PBT/vPvB<sup>11</sup>, ED<sup>12</sup> and/or Others<sup>13</sup>.

As illustrated in Figure 13, each category (i.e. CM, R, PBT/vPvB, ED and Others) has its own outcome table, and shows the most similar ZZS to the input structure (on top). When at least one ZZS within a category is considered to be structurally similar to the input structure, the top three most similar ZZS are shown in that category. Structural similar substances are highlighted in orange (instead of grey; see Figure 13). In case that more than three ZZS substance are considered to be structurally similar to the input structure, their information can be obtained by pressing the button “Show all xxx similar substances” (see Figure 14). When no structural similar substances are identified in a category this will be shown (see Figure 15). The top three most similar ZZS of this category can still be shown by pressing the button “Show best matches” (see Figure 15).

<sup>9</sup> CM: Carcinogenic, Mutagenic

<sup>10</sup> R: Reprotoxic

<sup>11</sup> PBT/vPvB: Persistent, Bioaccumulative and Toxic / very Persistent and very Bioaccumulative

<sup>12</sup> ED: Endocrine disruptor

<sup>13</sup> Other: Substances that do not belong to any of the other categories, like European recognized ZZS with persistent, mobile and toxic (PMT) properties, specific target organ toxicity after repeated exposure (STOT-RE) or sensitizing properties; and ZZS that are derived from other sources than REACH, CLP or POP-regulations.



As described above, structural similarity could be an indication of similar toxicity profiles. It should be noted that the absence of structural similarity to a ZZS does not per definition means no concerns. Vice versa, similarity does not per definition means that a substance exerts a specific effect, it only provides a trigger for further inspection and analysis.

Similarity to CM substances								
Substance	CAS number(s)	EC number(s)	SMILES	Similar	Confidence in structural similarity	Details	Molecular structure	Possible toxicity
styrene oxide	96-09-3	202-476-7	O1CC1c2ccccc2	Yes	64 %		<a href="#">Compare</a>	☞
chlorotoluene	100-44-7	202-853-6	c1ccc(cc1)CCl	Yes	51 %		<a href="#">Compare</a>	☞
5-nitroacenaphthene	602-87-9	210-025-0	O=[N+] ([O-])c3ccc2c1c3(cccc1CC2)	No	39 %		<a href="#">Compare</a>	☞

Figure 13. Example of the output screen, showing two substances that are considered to be structurally similar to the input substance (in orange) and one substance that is not structurally similar (in grey).

Similarity to ED substances								
Substance	CAS number(s)	EC number(s)	SMILES	Similar	Confidence in structural similarity	Details	Molecular structure	
Representing 4-nonylphenol, branched	84852-15-3	284-325-5	Oc1ccc(cc1)CCCCC(C)C(C)C	Yes	100 %		<a href="#">Compare</a>	
Representing 4-heptylphenol, branched and linear			Oc1ccc(cc1)CCCCC	Yes	100 %	View	<a href="#">Compare</a>	
Representing 4-nonylphenol, branched	84852-15-3	284-325-5	Oc1ccc(cc1)CCCCC(C)C	Yes	99 %		<a href="#">Compare</a>	

[Show all 22 similar substances](#)

Figure 14. Example of the output screen when more than 3 ZZS substances of a category are considered to be structurally similar to the input structure.

Similarity to PBT/vPvB substances	
<b>i</b>	No similar PBT/vPvB substances found.
<a href="#">Show best matches</a>	

Figure 15. Example of the output screen when no structurally similar substances are identified.

The following information is included in the results tables:

- **Substance:** Name of the ZZS substance to which the input structure is compared.
- **CAS number(s):** CAS number(s) of the ZZS substance to which the input structure is compared. The CAS number(s) is/are hyperlink(s) to the specific ZZS-substance page.
- **EC number(s):** EC number(s) of the ZZS substance to which the input structure is compared. The EC number(s) is/are hyperlink(s) to the specific ZZS-substance page.
- **SMILES:** SMILES of the ZZS substance to which the input structure is compared.
- **Similar:** Yes or No. When the input structure is considered to be structurally similar by the computer model to the ZZS the outcome is 'Yes', otherwise the outcome is 'No'. Besides this statement, the color of the row is influenced by the similarity (i.e. Yes = orange; No = grey).
- **Confidence in structural similarity:** A score is provided that indicates the model's confidence in the structural similarity, with a confidence score equal to or above 50% when the model considers the structures to be structurally similar. The provided confidence scores are related to the calculated similarity scores between two chemicals, and the underlying distribution is model specific (See <sup>3</sup> for more details and Figure 16 for an example).

- **Details:** When no CAS or EC number is available for the ZZS substance or group of ZZS substances, a link is provided to the ZZS-substance page (see Figure 14).
- **Molecular structure:** By pressing the “Compare” button, the chemical structures of the input substance and ZZS are shown (see Figure 17).
- **Possible toxicity:** ZZS substances as included in the CM, PBT/vPvB and Other dataset are included based on a specific concern. For CM, this can either be carcinogenic or mutagenic properties. For PBT/vPvB, this can either be Persistent, Bioaccumulative and Toxic properties, or very Persistent and very Bioaccumulative properties. For Other, this can be related to C-, M-, R-, PBT-, PBT/vPvB- or ED-related properties or can be based on PMT and other types of properties, like sensitizing or STOT-RE properties. Within this column, the possible toxicity of the ZZS substance is listed (multiple properties could apply). This could provide additional information on potential concerns for an input structure.

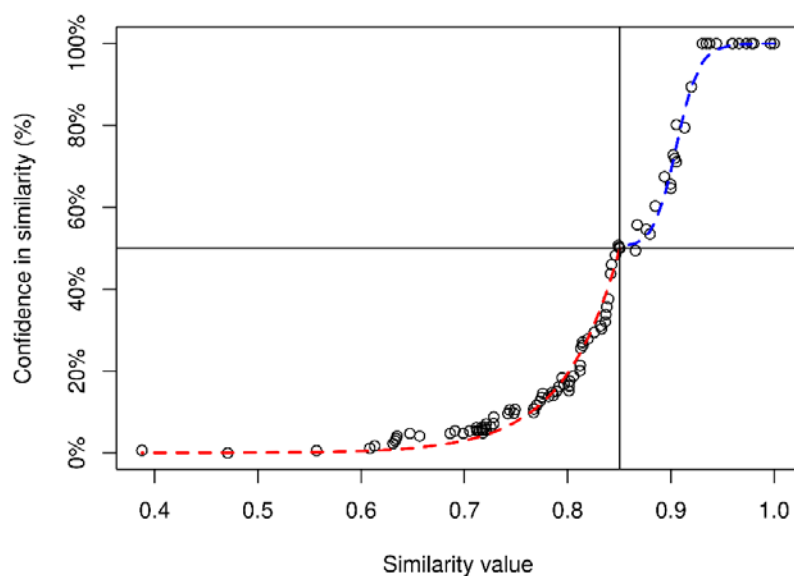


Figure 16. Relation between the structural similarity value and the confidence in the predicted structural similarity between a chemical and a ZZS. Each similarity model has its own optimized distribution.

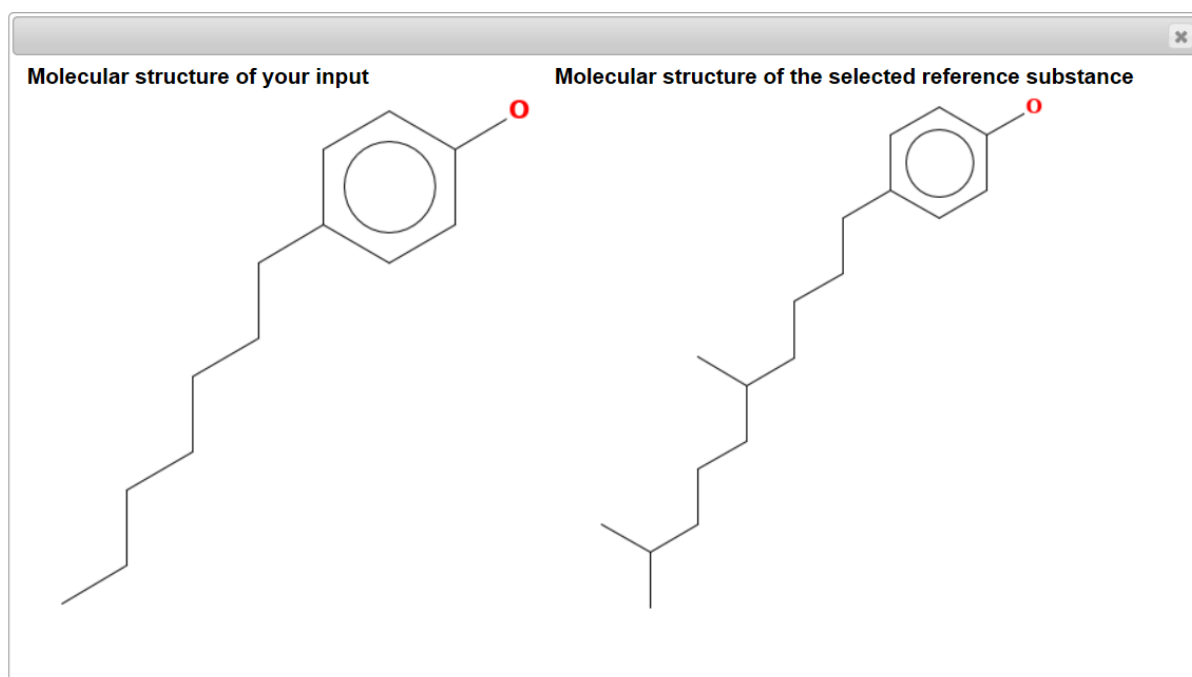


Figure 17. Example of the output screen when pressing the “Compare” button in the Molecular structure column.

### Technical remarks on the (implementation of the) methodology

The model is based on the methodology as analyzed and evaluated by Wassenaar et al. (2022)<sup>3</sup> and makes use of the PaDEL functionalities (based on the chemistry development kit libraries)<sup>14</sup>. Accordingly, all fingerprints within the ZZS similarity tool are all based on one programming language.

During implementation of the CDK fingerprints and corresponding PaDEL libraries, the fingerprints of 19 ZZS substances could not be reproduced as applied in Wassenaar et al. (2022) (probably due to small differences in the incorporation of the underlying PaDEL libraries for SMILES standardization). Therefore, for these 19 ZZS substances, the Extended fingerprints (n=19) and PubChem fingerprints (n=17) have been adjusted in the underlying dataset of the ZZS similarity tool.

Furthermore, it should be noted that not all ZZS substances are (yet) incorporated in the ZZS similarity model. All ZZS that were identified prior to 25-01-2021 are currently included. New ZZS substances will be added during an update of the ZZS similarity tool.

It should be noted – with respect to the interpretation of the results – that the absence of structural similarity to a ZZS does not per definition means no concerns. Vice versa, similarity does not per definition means that a substance exerts a specific effect, it is a trigger for further inspection and analysis. The ZZS similarity tool is meant to be applied as a screening model.

---

<sup>14</sup> PaDEL descriptor: <http://www.yapcwsoft.com/dd/padeldescriptor/>