# UVCB Project Description

## UVCB substances
## Characterisation, chemical representation and modelling


## Abstract

Substances of unknown or variable composition, complex reaction products or biological materials, collectively called UVCBs under REACH, pose the greatest challenge in a number of REACH processes, in particular in the context of dossier and substance evaluation but also in the priority setting activities of the Agency. The purpose of this contract is to develop a computational methodology for generating representative structures likely to be present in the UVCB and to perform a preliminary screening of their hazardous properties using non-testing methods. The outcome of the project will assist in assessing the homogeneity of UVCBs as far as the likely presence or absence of hazardous constituents is concerned, which can be subsequently applied in dossier and substance evaluation processes but also used by Industry in designing a comprehensive testing strategy of UVCB substances.


## Project summary

The successful registration and regulation of chemical substances depends crucially on their unambiguous structural description [1, 2]. Chemical structural information is essential not only for substance identification but also for establishing substance sameness and for the application of non-testing methods for hazard assessment. In this respect, the substances of unknown or variable composition, complex reaction products or biological materials (collectively called UVCBs) pose a significant challenge [3] for a number of REACH processes. This project is focused on the development of a methodology for structural identification of UVCB substances and their computational hazard assessment. The assessment process could be generalised by the scheme presented in Figure 1:


Figure 1. Illustration of the general UVCB hazard assessment process.

After an open procurement procedure, the project was awarded to the bid presented by the Laboratory of Mathematical Chemistry at the Prof. As. Zlatarov University of Bourgas, Bulgaria.

## Aims

The goal of the project is the development of a generic and flexible methodology for generating a computationally manageable number of representative structures that cover the chemical space defined by the UVCB identifiers and the computational assessment of these representative structures with respect to their physicochemical, fate and (eco)toxicological

properties. The major scientific challenge is to identify, and manage the level of structural variability of UVCB substances. The following specific objectives form the basis of the methodology development:

- introducing a generic computer description that is suitable for the wide range of industrial UVCB substances,
- generation of a computationally manageable number of UVCB representative structures, and
- hazard assessment of UVCBs based on their generated representative structures with respect to physicochemical, fate and (eco)toxicological properties.

The project is considered as a strategic development activity of the European Chemicals Agency aiming to propose a pragmatic and efficient methodology for the unique *in silico* representation and database storage of UVCB substances with the longer aim being to improve the hazard assessment, risk assessment and prioritisation of such substances. These developments are also expected to improve the derivation and assessment of read-across and category data, and also to allow the application of (Q)SARs for chemicals that lack explicit structural information.

**Background**

UVCB substances cannot be sufficiently identified by their chemical composition, because [3]:

- the number of constituents is relatively large, and/or
- the composition is, to a significant part, unknown, and/or
- the variability of composition is relatively large or poorly predictable.

As a result of compositional variability, the identification of UVCB substances is mostly based on their generic description [3]. The production source (biological or non-biological), process of formation (chemical reaction, extraction, fractioning, etc.), composition fingerprints (chromatographic or spectral information), physicochemical properties (optical activity, range of variation of boiling point, viscosity, molecular weight, etc.) are used to identify UVCB substances. Such a generic identification of UVCB substances causes significant difficulties for submitters and regulatory authorities to select representative chemical structures for assessing UVCB hazard properties.

It should be mentioned, that approaches for encoding UVCB substances have already been developed and used in pharmaceutical or patent applications (e.g. Markush code, SMARTS). However, these approaches have no general applicability for all UVCB cases; for example, where chemical processes are involved in the UVCB substance description.

**Significance and innovation**

The primary objective of the project is the development of a methodology for the structural representation of UVCB substances and their hazard assessment using computational methods and/or experimental data for specific representatives. Experience in the generic description of UVCB substances, the control of combinatorial problems, simulation of chemical reactions, categorisation, fragment search, databasing and information technologies

support the achievement of the project goals. The methodology developed in the project will assist registrants and regulators in assessing UVCB substances. They will have not only a convenient way to code, select representatives and perform hazard assessment, but they will also be able to store the results of the UVCB computational hazard assessment in archive files and to disseminate the results. In this respect, the treatment of UVCB substances will become transparent, allow easy crosschecking and facilitate the communication between industries and authorities. Despite the early stages of development in this project, if the methodology is successful and proves that it has added value, we can envisage the inclusion in the registration dossiers of UVCB structural information, using a format that would allow the generation of the constituents likely to be present in the UVCBs by other parties, including the regulatory authorities.

**Approach and training**

The development of a methodology for the generation of representative structures and hazard assessment of UVCB substances is based on the following tasks addressed in the project:

- coding of chemical variability of UVCB substances into the GSMILES format; this is a succinct notation that captures the structural variability of UVCB substances and can technically be included in registration dossiers, databases or dissemination portals,
- generation of a computationally manageable number of representatives,
- simulation of reactions with UVCB reactants, by introducing RGRAPHs,
- UVCB hazard assessment, and
- reporting of results.

**Coding of chemical variability of UVCB substances.** Computer representation of UVCB substances is based on an extension of the SMILES code, named Generic SMILES (GSMILES). The new features implemented in this encoding of structurally-related sets of chemical structures reflect the generic description of UVCB substances. The variable parts of the UVCB constituents are coded in generic fragments, which to some extent could be considered as “superatoms”. The generic fragments are determined by their name, structural units, coordination bonds and boundaries. Currently, enumeration, positional, chain, polycyclic and bond migration fragments are determined. The list of fragments is open and could be expanded when necessary. SMILES coding is used to determine the connectivity of GSMILES elements (atoms and generic fragments). A 2D molecular editor is used to facilitate the editing of GSMILES. The basic functionalities of the editor allow the automatic generation of GSMILES based on the user drawn UVCBs, adding, editing or deleting atoms and bonds, depictions of the UVCB structure, as well as importing and exporting GSMILES into a text format that can easily be stored in databases and used for generating representative structures when the need arises.

**Generation of representatives.** Generation of UVCB constituents is based on GSMILES. Three situations could arise:

1. the number of all constituents can be generated in a reasonable time and the computational hazard assessment is also computationally feasible. In this case, the UVCB can be represented by **all** generated structures that correspond to the structural boundaries and parametric ranges specified in the generic UVCB description.

2. it is still possible to generate all constituents but their number is so large that it is impossible to perform computational hazard assessments for all the generated structures within an acceptable timeframe. In this respect, constituents of greater concern have to be selected. Selection is foreseen to be endpoint and/or statistically driven (evenly, normally or empirically distributed representatives by using a characteristic property such as log Kow or degrees of branching as a means to calculate representative structure distributions to allow sampling).

3. combinatorial explosion occurs and the generation of all possible constituents is not feasible with reasonable computing resources. Information about the generic fragments contributing to the combinatorial problem will be used to adjust their boundaries in order to reduce the structural variability and make the generation feasible (i.e. reducing the third case into the second one by limiting the structural variability of the individual structural features that lead to combinatorial explosion; for instance, by allowing discrete degrees of branching of a particular hydrocarbon chain instead of generating all possible branching alternatives).

**Simulation of reactions with UVCB reactants.** The identification of UVCB substances includes a description of physical and/or chemical processes used for their production. Chemical processes are usually reactions where some or all reactants are themselves UVCB substances. In the proposed approach, the computer coding of the generic description of UVCB substances that are the product of reactions is based on the GSMILES of reactants combined with a list of molecular transformations (named Generic Reaction Graph – GR Graph). The generation of constituents starts with the generation of representative structures for the reactants followed by the computational application of the chemical transformations involved in the manufacture of UVCBs. The resulting products could be subject of another set of chemical reactions, etc. Theoretical chemistry and knowledge for the simulated chemical processes are used to reduce the combinatorial problem.

**UVCB hazard assessment.** The UVCB hazard assessment is based on all generated constituents or a set of representative structures. Their SMILES files are used by other programs (e.g. CATALOGIC [4], Derek [5], EPI Suite [6], QSAR Toolbox [7], TIMES [8], etc.) to perform a search for experimental data, to apply category approaches or to predict the targeted endpoints by (Q)SAR models.

**Reporting of results.** The methodology includes that generation of a report including the description of the definition of GSMILES, generation of constituents (including the undertaken reduction of the combinatorial problem if needed), simulation of reaction processes, computational hazard assessment of constituents (representatives) and the overall hazard assessment of the UVCB substance.

The principal stages of the methodology for the computational hazard assessment of UVCB substances, which are not defined reaction products, are illustrated in Figure 2. Depending on the complexity of the UVCB some of these stages could be unnecessary. As can be seen from the figure, the assessment could be an iterative procedure where some of the stages could be repeated until an unambiguous result is obtained.

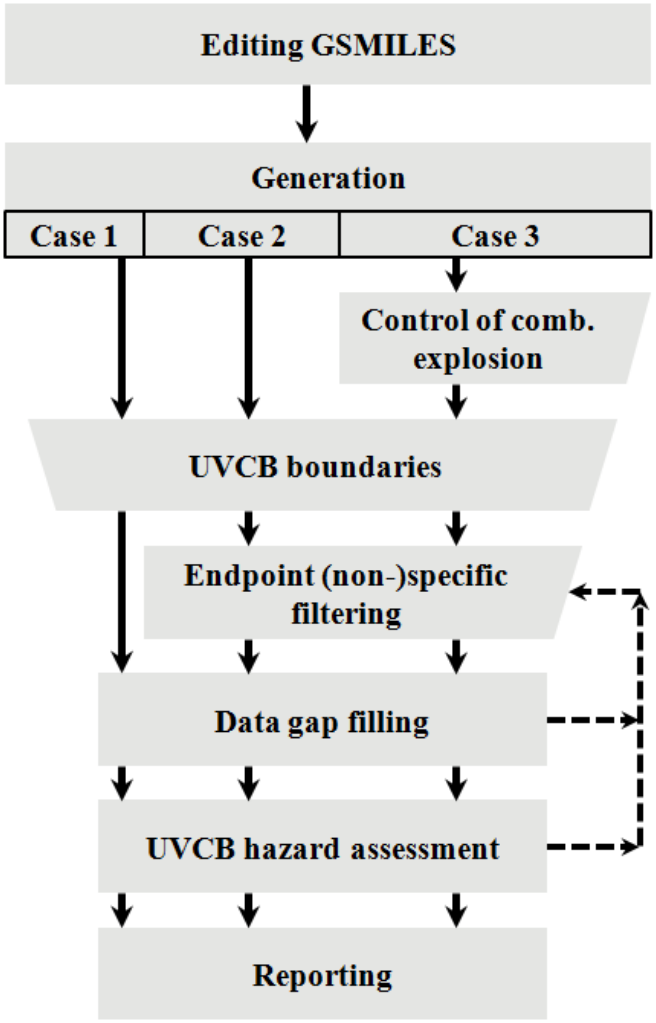


Figure 2. Stages for generation of representatives and hazard assessment of UVCB substances that are not encoded as a chemical reaction product. The assumption behind this figure is that in cases 2 and 3 (complex UVCBs), it is not computationally feasible to generate and computationally characterise the hazards of all representative structures and that different structure generation and filtering options are tested so that the optimal choice for sufficiently identifying hazard properties can be identified.

Figure 3 illustrates the reaction graph for a particular case of formation of UVCBs as a result of two consecutive reaction processes. Dashed lines illustrate the stages for generation of reactant representatives and their hazard assessment.

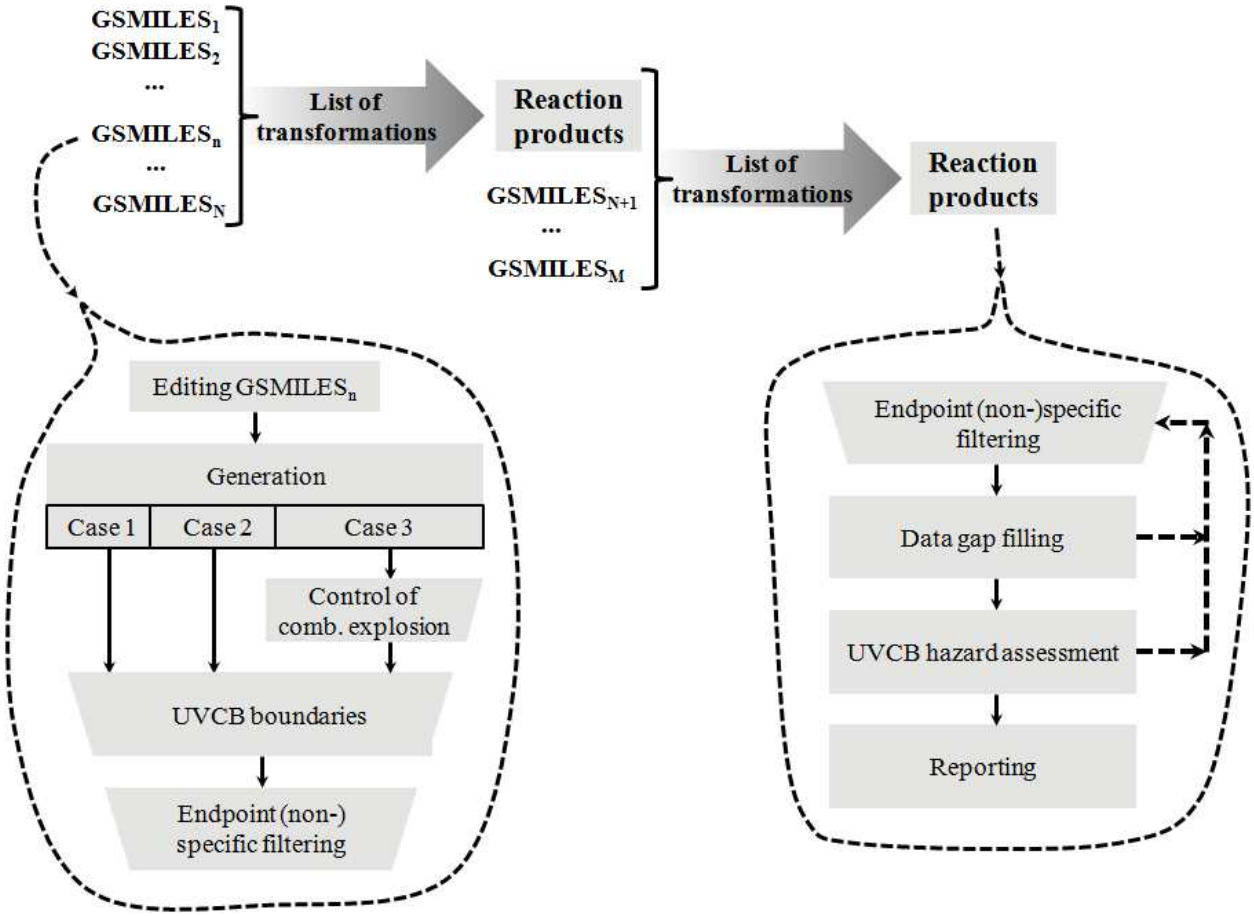


Figure 3. Stages for generation of representatives and hazard assessment of UVCB substances that are encoded as a chemical reaction product. In the figure, it is assumed that the final UVCB is obtained after two distinct reaction steps. The possible reactant structures are generated using the approach in Figure 1 (shown with the dashed enclosure at the lower left part of the figure). After the chemical transformation of the reactants, it may be necessary to only select a reasonable number of reaction products for computational hazard assessment (shown with the dashed enclosure at the lower right part of the figure)

The methodology developed in this project will be demonstrated by its application on eight UVCB cases specified by the Agency. These cases include substances formed by physical and chemical processes and are considered as representative examples for different types of UVCBs.

**Expected outcomes**

The expected outcome from the project is the development of a methodology for the generation of UVCB constituents (or representatives) and their use for assessing the fate and (eco)toxicological properties of UVCB substances. The methodology will affect the *in silico* representation, databasing, and hazard assessment of UVCB substances and assist the generation and assessment of data using read-across or category approaches. It is envisaged that such a methodology, if successful, may assist both the registrants fulfilling the information requirements with regard to substance identity and the regulatory authorities

assessing this information. Communication of results is foreseen by a presentation in a regulatory conference and scientific paper.

## References


1. Official Journal of the European Union, No 1907/2006, 2006, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=oj:l:2006:396:0001:0849:en:pdf.
2. Official Journal of the European Union, No 1272/2008, 2008, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:353:0001:1355:en:PDF.
3. ECHA, Guidance for identification and naming of substances under REACH and CLP, ECHA-11-G-10-EN, 2011.
4. S. Dimitrov, T. Pavlov, N. Dimitrova, D. Georgieva, D. Nedelcheva, A. Kesova, R. Vasilev, O. Mekenyan, Simulation of chemical metabolism for fate and hazard assessment. II. CATALOGIC simulation of abiotic and microbial degradation, SAR and QSAR in Environ. Res. (2011) 22(7-8): 719-755.
5. Lhasa Limited; https://www.lhasalimited.org/
6. Estimation Program Interface (EPI) Suitehttp://www.epa.gov/opptintr/exposure/pubs/episuite.htm
7. QSAR Toolbox, http://www.qsartoolbox.org/
8. O. Mekenyan, S. Dimitrov, T. Pavlov, G. Dimitrova, M. Todorov, P. Petkov, S. Kotov, Simulation of chemical metabolism for fate and hazard assessment. V. Mammalian Hazard Assessment, SAR and QSAR in Environ. Res. (2012) 23(5–6): 553–606.