Concawe
Environmental Science
for European Refining

Report

Report no. 4/23

Evaluation of aquatic toxicity monitoring techniques for refinery effluents

ISBN 978-2-87567-171-4

Evaluation of aquatic toxicity monitoring techniques for refinery effluents

E. Vaiopoulou (Concawe Science Executive)

This report was prepared by:

M. Griffiths, C. Hughes, E. Pemberton and E. Butcher (Ricardo)

Under the supervision of:

E. Vaiopoulou (Concawe Science Executive)
M. Hjort (Concawe Science Associate)

At the request of:

Concawe Special Task Force WQ/STF-32 and STF-34

Thanks for their contribution to:

Members of WQ/STF-32: M. Smit, C. Gelber, K. Colvin
Members of WQ/STF-34: A. Bohne-Kjersem, M-P. Campione

Members of Concawe Water Soil Waste Management Group (WSWMG): C. Dupre, E. Marin, J. Smith.

Brussels
April 2023

ABSTRACT

This report identifies and evaluates a selection of *in vivo* and *in vitro* aquatic toxicity testing methods, in order to assess which tests could present an interest if included in a test battery for toxicity monitoring of refinery effluents discharging to the receiving aquatic environment.

A total of 13 *in vivo* tests covering a range of trophic levels and endpoints, and 18 *in vitro* tests covering cytotoxicity, genotoxicity, oxidative stress, metabolism, and endocrine disruption have been evaluated against specific criteria based on a literature review and survey of commercial laboratories.

The suggested battery resulting from this evaluation includes the *umuC*, AhR activation, AREc32 activation, *Allivibrio fischeri* toxicity, algal growth inhibition assays, *Daphnia* immobilisation and (q)FET or bivalve embryo development assays. The latter two are considered interchangeable depending on the type of water sample tested. The battery covers *in vitro* modes of action (genotoxicity, metabolism and oxidative stress) as well as apical *in vivo* endpoints (cytotoxicity, developmental toxicity, immobilisation and growth inhibition). All tests are commercially available (some with more limited availability than others), commonly used for the assessment of environmental water samples, sufficiently validated, standardised to an ISO guideline (or one is currently in preparation) and are expected to be responsive to refinery effluent constituents and refinery effluents themselves. The test array, ultimately applied on a refinery site, could consist of one or more tests depending on the assessment objective, the protection goal of the monitoring campaign, the type of receiving water, and the activity undertaken (e.g. routine monitoring, full site risk assessment, detailed composition investigation etc.). We do not provide recommendations on the frequency of monitoring as this is out of scope of the current project.

KEYWORDS

Monitoring techniques, refinery effluents, toxicity tests, Effect-Based Methods (EBMs), *in vivo, in vitro,* Whole Effluent Toxicity (WET)

INTERNET

This report is available as an Adobe pdf file on the Concawe website (www.concawe.eu).

CONTENTS

SUMMARY

The Industrial Emissions Directive (IED; 2010/75/EU) prescribes that emissions from certain industrial processes should be treated using the Best Available Technology (BAT) and which parameters should be monitored to prove this. The parameters that refineries are obliged to monitor and control - and by which monitoring technique - are laid down in a Commission Implementing Decision BAT Conclusions (BATC) document for the mineral oil refining sector, which are published alongside the BAT Reference document (BREF) - the so-called REF BREF. BATC can contain BAT Associated Emission Limits (BAT-AELs), or other BAT-associated environmental performance levels (BAT-AEPLs), or may not be associated with BAT-AEPLs. This report identifies and evaluates a selection of *in vivo* and *in vitro* toxicity testing methods, in order to provide options for a battery of tests suitable for monitoring toxic effects of refinery effluents discharged to the aquatic environment. The evaluation could provide an important input to an upcoming review and revision of REF BREF and BATC scope.

A total of 13 *in vivo* tests covering a range of trophic levels and endpoints, and 18 *in vitro* tests covering cytotoxicity, genotoxicity, oxidative stress, metabolism, and endocrine disruption have been evaluated against specific criteria based on a literature review and survey of 14 commercial laboratories. These criteria were: commercial availability, application of test in a regulatory context, standardisation, validation maturity, sensitivity, potency of effects, use in environmental samples, use with refinery effluent constituents, use with refinery effluents/produced water, applicability to passive sampler extracts, time to run, cost per sample, Relevance, availability of trigger values, bioassay category, chemical analysis replacement, and confounding factors. In addition, toxicity testing using passive sample extracts, as well as complementary techniques such as biomimetic solid-phase microextraction and prediction methods have been discussed.

Based on this evaluation, a battery of test methods suitable for the monitoring of refinery effluents can be considered as a viable option. A battery should comprise tests that capture as many effects and/or active substances as possible that are relevant to the type of sample that battery is being designed for. Noting also that this must be carefully balanced with feasibility in terms of the costs and practical constraints associated with toxicity monitoring campaigns. The proposed battery includes the *umuC*, AhR activation, AREc32 activation, *Allivibrio fischeri* toxicity, algal growth inhibition assays, *Daphnia* immobilisation and (q)FET or bivalve embryo development assays. The latter two are considered interchangeable depending on the type of water sample tested. The battery covers *in vitro* modes of action (genotoxicity, metabolism and oxidative stress) as well as apical *in vivo* endpoints (cytotoxicity, developmental toxicity, immobilisation and growth inhibition). All tests are commercially available (some with more limited availability than others), commonly used for the assessment of environmental water samples, sufficiently validated, standardised to an ISO guideline (or one is currently in preparation) and are expected to be responsive to refinery effluent constituents and refinery effluents themselves. We do not provide recommendations on the frequency of monitoring as this is out of scope of the current project.

While all tests in the suggested battery are deemed suitable for toxicity monitoring of refinery effluents, they do not all necessarily need to be deployed as one fixed array. The exact choice of tests would be informed by the assessment objective, the protection goal of the monitoring campaign, the type of receiving water, and the activity undertaken (e.g. routine monitoring, full site risk assessment, detailed composition investigation etc.).

1. INTRODUCTION

1.1. BACKGROUND

1.1.1. EU industrial emissions policy

1.1.1.1. Industrial Emissions Directive

The Industrial Emissions Directive 2010/75/EU (IED) takes an integrated approach to regulating pollution to air, water, and land within the European Union (EU) to mitigate and reduce the environmental impacts from emissions from certain industrial processes. The directive seeks to prevent harmful emissions from industry, while promoting the use of energy and resource efficient techniques. All installations conducting activities listed in Annex I to the IED are required to operate according to a permit. The permit is issued by the relevant Member State authorities under national legislation that transposes the IED, and should reflect the principles and provisions stipulated by the IED.

All environmental aspects of the operating activities of the installation, including emissions, are covered by the permit. Larger industrial facilities undertaking specific activities are required to use Best Available Technology (BAT) to reduce their emissions. All permit conditions must be based on BAT conclusions (BATC) within four years of adoption of the BATC. The BATC are adopted by the European Commission following an exchange of information among technical experts culminating in BAT Reference Documents (BREFs). BREFs include BATC with associated emission limits; these must not be exceeded unless agreed by the appropriate competent authority. BATC can contain monitoring by using specific monitoring techniques at a given frequency, BAT Associated Emission Limits (BAT-AELs), or other BAT-associated environmental performance levels (BAT-AEPLs) such as consumption levels and abatement efficiency, or may not be associated with BAT-AEPLs (e.g. concerning monitoring, site remediation or environmental management systems) [1].

1.1.1.2. Mineral oil and gas refinery BREF and BAT Conclusions

The BREF document for the mineral oil and gas refinery sector (REF BREF) [2] specifies the requirements for wastewater. The document, via BATCs, imposes BAT-AELs and monitoring requirements, which – after its next review and revision – might include toxicity testing, depending on risk assessment after initial characterisation.

The reference document for Monitoring of Emissions to Air and Water from IED Installations (ROM REF) [3] was updated after REF BREF adoption and provides an overview of standard testing, including their costs, pros and cons or data treatment requirements. The document describes the use of toxicity testing to derive BAT and BAT-AEPLs and to monitor impacts from emissions, particularly where effluents are complex and not adequately assessed by the analysis of specific substances.

1.1.1.3. Preliminary determination of Key Environmental Issues (KEIs)

In 2018 a change was made to assess the identification of KEIs in the BREF process. The aim was to determine KEIs in advance of the information exchange of the BREF process as part of the approach of front-loading the efforts in the BREF review, and to take a more focussed approach. This approach was based on four criteria originally proposed by the European Commission in 2015. The criteria were refined following testing of the new approach to identifying KEIs in a study led by Ricardo, concluded in 2018 [4]. The four criteria that underpin the methodology are:

1. Are the environmental issues and associated parameters relevant for the activity or process concerned?
2. Is the industrial process and its pollution and consumption a significant part of industrial pollution and consumption in the EU, currently or trending?
3. The potential for identifying new or additional techniques that would further significantly reduce pollution.
4. The potential for defining BAT-AELs and BAT Associated Environmental Performance Levels (BAT-AEPLs) that would significantly improve the level of protection for the environment.

The core steps and milestones of the Seville process remain similar to the steps described in the guidance (CID 2012/119/EU). BREF Technical Working Groups (TWGs) remain the key pillar for each BREF review and TWG members approve BREF scope changes or data collection features. The latest BREF reviews, such as Common Waste Gas Management and Treatment Systems in the Chemical Sector BREF (WGC BREF), have incorporated a more systematic KEI selection to justify what is included in the data collection questionnaire.

1.1.2. Toxicity testing to assess water quality

1.1.2.1. Toxicity monitoring in oil & gas refineries

Within the petroleum water refining industry, water is a vital resource, and refineries produce aqueous effluent as part of their operations [5]. The efficient and responsible management of refinery wastewater has been a key focus of the oil industry for several decades, and considerable improvements have been made. Since 1969, Concawe has been gathering data on effluents from European oil refinery installations in the form of surveys, completed at three to five year intervals [6]. These surveys are continuously evolving to reflect scientific and legislative developments. In 2010, the Concawe survey was updated to request information on the toxicity testing of refinery effluents.

Toxicity testing is a means to assess hazard (level of potential adverse effects) posed by a substance or mixture and is frequently used for risk assessment. Whole Effluent Toxicity (WET) testing has been conducted with industrial effluents as part of broader effluent assessments. Toxicity tests used for regulatory purposes employ standardised test methods. The OECD test guidelines are typically used. EN ISO standards are generally recommended for effluent monitoring. The methods originating from these different standardisation bodies are based on common principles, and the EN ISO standards are recognised as acceptable alternatives to OECD tests for purposes of regulatory evaluation of chemical substances [7].

Effluent toxicity tests typically use aquatic organisms, such as the freshwater invertebrate *Daphnia magna*, luminescent bacteria or algal species. The use of *in vivo* fish toxicity tests has declined significantly in Europe following the introduction of EU Directive “on the protection of animals used for scientific purposes” (2010/63/EU) and has been replaced by tests on fish eggs, e.g. the fish embryo toxicity (FET) test using zebrafish [8, 9]. Direct toxicity testing of whole effluents is considered as one component of a broader effluent assessment toolkit that includes chemical monitoring of effluents, bioassays, mechanistic modelling, and testing and monitoring of the receiving environment.

Toxicity testing of refinery effluents can be driven by mandatory regulatory or national permit requirements, as well as voluntary company programmes. Refinery effluents originate from multiple sources and processes within the refinery and may undergo various treatment steps. The concentrations of contaminants in effluent varies according to factors such as the feedstock characteristics processed, the unit processes from which the wastewater originates and the wastewater treatment steps. Subsequent toxicity testing may cover some, all or none of a refinery's effluent streams, and testing may be carried out according to varying frequencies from daily to yearly [3, 7].

1.1.2.2. European policy on toxicity testing of industrial effluents

The REF BREF [2] contains no BAT for toxicity monitoring requirements. European refinery installations might also be covered by the CWW BREF, of which the BAT conclusion *can* impose toxicity testing, proposed by industrial site with justifications. The BATC for the CWW BREF motivated future work on effects-based methods and toxicity testing, listing ISO standardised tests with fish eggs, *Daphnia*, luminescent bacteria, duckweed and algae [10].

The way the latest BREFs manage the potential impacts of harmful chemicals on the environment has been changing. More recent BREF contain more BAT conclusions to minimise these impacts and also disaggregate measures based on chemicals characterisation. For example, the WGC BREF [9] sets different BAT AELs values on emission to air based on different CMR (Carcinogenic, Mutagenic or toxic for Reproduction) classification of chemical compounds.

Outside of BREF discussions, whole effluent toxicity testing is commonly applied as conditions of permitting across the EU (and US) [11]. Toxicity testing was frequently requested by national or regional regulations or guidance in several Member States such as Austria, Germany, Italy, Lithuania, or Ireland. Effects-based methods have recently been incorporated into the assessment of discharges from offshore installations [12]. There is also interest from the perspective of the implementation of the Water Framework Directive [13]. Finally, there has been considerable recent research activity in this space, one example being the EU SOLUTIONS programme, a 5-year collaboration between 39 institutes, concluding in 2018, with a heavy focus on effect-based tools and analysis.

1.1.2.3. Composition of refinery effluents

The selection of appropriate effects-based methods will primarily be informed by the assessment objective (e.g. full receiving water risk assessment, monitoring toxicity trends over time etc.), which should also consider the protection goal of the monitoring program (e.g. environmental health vs human health) and the relevance to the modes of action (MoA) of the refinery effluent constituents. The composition of refinery effluents depends on the complexity of the refining process, crude oil characteristics, plant configuration and process designs, and is thus subject to variability. However, the major chemical families are largely known and common to most refinery effluents. These include organics such as phenols, BTEX (benzene, toluene, ethylbenzene and xylenes), polycyclic aromatic hydrocarbons (PAHs) and others (e.g. esters, amides, aromatic ketones, alkanes etc.); inorganics such as phosphorus, sulphides, nitrogen (ammonium-N, nitrites-N, nitrates-N), chlorides, fluorides, cyanides and metallic elements (e.g. aluminium, magnesium, calcium, barium etc.); and metals such as mercury, cadmium, lead and nickel [2, 14]. The annual average concentrations of some of these, based on monitoring data for 2016 or 2019 provided to Concawe surveying 72 European refineries, are presented below in Table 1.

Table 1: Annual average concentrations of refinery effluent constituents, presented for 25th,50th and 75th percentile based on monitoring data for 2016 or 2019 surveying 72 European refineries.

Substance	Number of sources below Limit of Quantification*	Number of sources	Percentile 25th	Percentile 50th	Percentile 75th	Units
Benzene	17	28	0.2	0.7	5	µg/l
Total Nitrogen	0	55	3.5	6.9	10.6	mg/l
Lead	18	22	0.8	2	10	µg/l
Cadmium	20	26	0.1	0.5	2	µg/l
Nickel	8	38	4.4	6	10	µg/l
Mercury	15	28	0.08	0.3	1	µg/l

*Not included in percentile calculations

The different constituents may have different modes and mechanisms of action that will be captured by different bioassays with different organisms. These modes of action are physiological or behavioural changes at the cellular level, underpinned by specific 'mechanisms of toxicity' at the biochemical level, such as chemical-receptor interactions, chemical-enzyme interactions or chemical-endogenous molecule interactions [15]. It is important to understand which modes of action, where known, are linked to the constituents of refinery effluents, and affect which kind of organisms, as this will affect the relevance and thus appropriateness of any effects-based methods used to assess effluent toxicity. PAHs, for example, are known to be genotoxic and carcinogenic. They physically interact with DNA to change and/or damage its structure, such as through the formation of adducts [15]. This is particularly true when the PAHs are bioactivated to reactive diol epoxides during phase I of metabolism. The ensuing biochemical changes and cell damage lead to mutations, tumours, cancer and developmental malformations, for both humans and animals. Embryotoxic and reprotoxic effects of PAHs have also been reported in animals exposed to benz(a)anthracene, benzo(a)pyrene and naphthalene [16].

PAHs, dioxins and dioxin-like chemicals activate the aryl hydrocarbon receptor (AhR), a ligand-activated transcription factor that regulates the expression of genes involved in the oxidative stress response and inflammation, such as monooxygenases cytochrome P450 (CYP) 1A1, CYP1A2 and CYP1B1 [15, 17]. PAHs have also been reported to be immunotoxic, as they can bind to AhR in lymphocytes and accessory cells of the immune system and supress it [18]. For example, *in vivo*, phenanthrene and pyrene have altered the immune response in European clams [19]. Activation of the AhR by PAHs further promotes carcinogenicity as CYP can convert several of its ligands to DNA damaging reactive intermediates [15]. Antiestrogenic effects have also resulted from PAH activation of AhR responsive genes [16].

PAHs such as benzo[a]pyrene, benz[a]anthracene, fluoranthene and pyrene have also been linked to weak endocrine disruption via direct binding with estrogen and/or androgen receptors, eliciting antiestrogenic and/or antiandrogen effects, respectively [16]. Chrysene has also been found to be endocrine disruptive in the aquatic crustacean *Penaeus monodon* [20]. Endocrine disruption is a MoA relevant to aluminium and other metals similar to calcium ions, as these can interfere with the calcium-sensing receptors in the parathyroid gland involved in calcium homeostasis. Metals are also reported to be immunotoxic and linked to reproductive toxicity [15].

Metals usually elicit toxicity via either the generation of reactive oxygen species (ROS), the inactivation of enzymes, or the suppression of the antioxidant defence [21]. Lead causes cellular toxicity via oxidative stress; lead decreases the levels of antioxidant and increases the levels of ROS, to the point where ROS levels may be high enough to structurally damage cells, proteins, membranes etc., and where antioxidants levels are low enough to reduce protection against free radicals [22]. Lead toxicity also acts via an ionic mechanism, in which lead metal ions can replace other bivalent ions (e.g. iron (II), calcium and magnesium) and monovalent cations (e.g. sodium), potentially resulting in apoptosis or affecting enzyme regulation, ionic transportation, release of neurotransmitters etc. [22]. Mercury and cadmium elicit reactive toxicity; they can form complexes with thiol groups, damaging tertiary and quaternary protein structure which then hinders cellular structure, integrity and function [22]. If this protein damage affects an enzymatic site of a protein, then non-specific enzyme inhibition can also occur [15]. Mercury, as well as arsenic, lead, manganese and aluminium can also result in toxic neuronopathy, a form of neurotoxicity [15]. Aluminium toxicity is likely to result from ROS generation, inhibition of antioxidant enzyme activity, disruption of magnesium, calcium and iron homeostasis, apoptosis, inhibition of DNA repair enzymes, and cross-linking of DNA [23]. Finally, copper and arsenic are linked to oxidative stress-related DNA damage, mutations, carcinogenesis and cytotoxicity [24]. Benzene is also carcinogenic.

Other constituents, such as phenols and cyanide, can affect mitochondrial energy production. Phenols are respiratory uncouplers, affecting the protonphoric shuttle mechanism, while cyanide binds to quinone binding sites in the electron transfer chain and thus inhibit electron transport [15]. Both reduce the energy efficiency of mitochondria, with cyanide causing a complete cessation of aerobic cell metabolism.

The above are specific modes of action, caused by specific substance-receptor or substance-enzyme interactions. Baseline toxicity, on the other hand, operates via a non-specific mode of action (narcosis), which is how hydrocarbons can elicit hydrophobicity-dependent toxicity. Baseline toxicity is described as the minimum toxicity any substance elicits and results from the partitioning of substances into cell membranes, causing cells to lose their structural integrity [15]. Ion and proton gradients across the membrane thus cannot be maintained, leading to mitochondrial dysfunction and, ultimately, cytotoxicity. As such, cytotoxicity assays and any cell viability/proliferation assays can detect baseline toxicity. A summary of the modes of action of key refinery effluent constituents are provided in Table 2.

Table 2: Summary of modes of action of relevant refinery effluent constituents.

	Hydrocarbons	Metals	Inorganics
Narcosis and Cytotoxicity	Yes	Yes	Yes (e.g. sulphides)
Genotoxicity	Yes (e.g. PAHs, benzene)	Yes (e.g. nickel, copper, arsenic)	Yes (e.g. sulphides, nitrites)
Oxidative stress	Yes (e.g. PAHs)	Yes (e.g. lead, nickel, copper, arsenic)	
Metabolism	Yes (e.g. PAHs, biphenyls, phenols, polychlorinated biphenyls (PCBs), dioxins)		Yes (e.g. cyanides)
Endocrine disruption	Weak (e.g. PAHs, dioxins, phenols)	Yes (e.g. cadmium aluminium)	Yes (e.g. nitrates)
Immunotoxicity	Yes (e.g. PAHs, PCBs, benzene)	Yes (e.g. cadmium, mercury)	
Neurotoxicity		Yes (e.g. manganese, lead, mercury, arsenic, aluminium)	

It is also important to note that not all substances and their mode(s) of action will affect all organisms in the same way, or at all. For example metals can adversely affect invertebrates but algae can show tolerance [25]. Further, toxic effects may occur even when individual substances are present at concentrations below the toxicity threshold (e.g. no-observed effect concentration (NOEC)), owing to additive or synergistic effects of the substances. The former occurs when there are no interactions between the substances in the mixture, but the substances can act together via concentration addition if they have a similar mode of action, or by independent action if their mode of action is dissimilar. An example of concentration addition is baseline toxicity, and hence individual hydrocarbons can contribute to mixture toxicity even if they are not toxic at the individual level [26]. Synergistic effects, on the other hand, occur when the substances in the mixture interact, producing more of a toxic response than would be predicted by their individual toxicities. However, any interaction in chemical mixtures may also promote antagonistic effects, which lowers the mixture toxicity compared to the summed effect of each substance.

A key advantage of WET testing (i.e. toxicity testing of the whole mixture) is that it accounts for these additive, synergistic and/or antagonistic effects directly, without having to use predictive mathematical models as may be required for toxicity testing of individual substances. WET also incorporates testing of toxic components that may be unknown and provides direct information on toxic effects. However, this means that WET testing requires further investigation to isolate contributions to toxicity. Further, WET testing may not account for important fate processes that occur after the effluent has been discharged into the receiving environment, such as biodegradation, evaporation, volatilisation and partitioning, affecting bioavailability of the substances to aquatic organisms. Degradation and volatilisation of the sample can also present logistical challenges prior to and during WET testing, for example sample representativeness, transport conditions, holding

times and overall testing timelines will need to be considered. During testing the physico-chemical properties of the sample may also need to be considered, e.g. dissolved oxygen, pH and salinity.

1.2. AIMS AND OBJECTIVES OF THE PROJECT

This project was commissioned to evaluate *in vivo* and *in vitro* toxicity test methods and provide options on which tests could be included in a test battery for toxicity monitoring of refinery effluents discharged to the aquatic environment.

This project has three objectives:

1. Identification of potential toxicity methods.
2. Evaluation of identified methods against select criteria.
3. Selection for a battery of methods, including justifications for why some methods are excluded.

The evaluation considers the performance of the tests - sensitivity, repeatability, and specificity. Practical conduct of the tests such as the availability of standard guidelines and time and cost to run are also considered in the evaluation. The use of the test for environmental samples, and specifically refinery effluent samples in a regulatory context is also commented on. Of high importance to the evaluation is the relevance of the modes of action or endpoints of the tests to both the aquatic environment and the anticipated constituents of refinery effluents. Our commentary is based on review of available scientific literature and expert knowledge. A survey completed by commercial laboratories has been used to provide insights on costs, frequency of use, applicability and technical limitations of the tests. This report also includes discussion of further aspects related to toxicity testing, such as the use of passive sample extracts, biomimetic solid-phase microextraction and prediction methods that can complement the testing itself. Recommendations on the frequency of monitoring are not included.

1.3. LESSONS LEARNED FROM PREVIOUS WORK

The past few decades have observed a substantial increase in the evaluations of substances contaminating water sources including surface waters, drinking water, recycled water, and effluents. A high proportion of the research activity in this area, particularly in the EU, has been dedicated to the use of chemical analysis to quantify and identify single substances in order to assess the risk of impacts. In recent years there has been widespread demonstration of the use *in vivo* and *in vitro* effect-based methods to facilitate a more extensive interpretation of results. Laboratory-based experiments together with additional lines of evidence such as bioassays and other bioanalytical tools (e.g. WET testing) to test mixtures of contaminants have been able to take more unknown components into account.

Previous research by Concawe, and others, has significantly developed the knowledge base on measurement, interpretation and prediction of the toxicity of refinery effluents. Recent publications are described in **Appendix 1**, highlighting implications for this current project.

2. METHODOLOGY

2.1. LITERATURE SCREEN

To identify and evaluate potential methods for monitoring toxicity of refinery effluents, a limited literature review was undertaken. Title and abstract screening were conducted using the PubMed AbstractSifter and ScienceDirect, and the following search string:

Toxicity AND ("in vivo" OR "in vitro" OR bioassay or EBM or "effect-based method" OR WET) AND (effluent OR wastewater OR "produced water")

The date for the screen was limited to 2018 onwards to avoid repeating the work of literature reviews presented in previous publications [27-29], although some earlier papers were included from the grey literature and publications provided by Concawe. As seen in Table 3, this query generated 391 and 222 papers from AbstractSifter and ScienceDirect, respectively, of which 22 and 9 were identified as relevant to this work following abstract screening.

Table 3: Results of the literature screen.

Search engine used	Key words/search string used		Total papers found	Total relevant papers
AbstractSifter	Toxicity AND ("in vivo" OR "in vitro" OR bioassay or EBM or "effect-based method" OR WET) AND (effluent OR wastewater OR "produced water")	2018-2022	391	22
ScienceDirect	Toxicity AND ("in vivo" OR "in vitro" OR bioassay or EBM or "effect-based method" OR WET) AND (effluent OR wastewater OR "produced water")	2018-2022, in title, abstract or author-specified keywords	222	9

2.2. SURVEY

A survey was developed using the Alchemer[1] survey software in order to gain insights from commercial ecotoxicity laboratories about the running of the tests selected for evaluation. For every test that the respondent selects they conduct, the questions in Table 4 appeared.

[1] https://www.alchemer.com/

Table 4: Survey questions and their respective multiple-choice answer options.

How frequently do you conduct the test?	How much do you charge per sample for the test?	What kind of sample(s) do you use for the test? (E.g. whole water samples, solid phase extracts)	What type of water samples do you use for the test?	Are there any logistical or technical challenges that can affect the ease and promptness in which the test can be executed?
Very regularly (e.g. more than 50 per year)	Under 100 €		Freshwater	
Regularly (e.g. 10-50 per year)	100 - <300 €		Marine water	
Infrequently (less than 10 per year)	300 - <1,000 €		Brackish water	
Rarely (once or twice per year)	1,000 - <3,000 €		Wastewater	
Don't know	3,000 - <10,000 €		Drinking water	
	10,000 € or above		Recycled water	
	Don't know		Other	

2.3. EVALUATION

Based on the results of the literature review and survey, a total of 13 *in vivo* methods (see Section 4 and Table A2-1) and 19 *in vitro* methods (see Section 5 and Table A2-2) were taken forward for evaluation. These were selected to encompass some well utilised tests as well as some newer tests. Some tests, e.g. *Arcatia* tests, found as part of the literature review were excluded as they did not have enough information available in order to evaluate them. All selected tests were evaluated according to the same set criteria, with some less critical to the evaluation than others. For transparency and ease of interpretation, these are defined and discussed in more detail below.

General details:
The type of organism or cell line used, and the mode of action or endpoint measured by the test is recorded. For the *in vivo* tests, the species and trophic level were also captured, as well as whether the test could be applied to marine, freshwater, brackish water or all water types. Marine water is typically that of >30 parts per thousand (ppt), based on dissolved salts, while freshwater is typically <0.5 ppt. Brackish water describes water with salinity between that of marine water and freshwater (0.5-30 ppt), normally in transitional environments where these two waters mix, such as estuaries. The water type applicability was based on knowledge of the species used and the application of the test in the scientific literature. It was envisaged that a battery would not fit all water types, but that tests could be swapped in and out for different water types. These general details helped assess relevance of the assay.

Commercial availability:
The commercial availability, i.e. the ability of commercial laboratories to perform the test, was assessed as yes or no based on Concawe [30], Wernersson *et al.* [28] and Kokkali and van Delft [31]. The *Corophium volutator* test was not mentioned in any of the publications. We assumed it was available commercially because it is recommended for use under the Convention for the Protection of the Marine Environment of the North-East Atlantic (OSPAR). All the tests included were available commercially with the exception of the Zebrafish Toxarray.

Application of test in a regulatory context:
If a test was known to be applied in a regulatory context, for example under the IED, then the test received a 'yes', and, where available, which regulation it related to. Of the *in vitro* assays, only the *umuC* and Yeast Estrogen/Androgen Screen (YES/YAS) assays were known to be applied in a regulatory context [30]. For the *in vivo* assays, it was further included whether or not they had been applied in European refinery effluents, based on the results of the Concawe surveys conducted in 2010, 2013, 2016 and 2019 [32].

Standardisation:
It was noted whether the assay had been standardised to an ISO guideline, and, if so, which one. In the cases where no ISO guideline existed for the assay, the availability of other guidelines (e.g. OECD or ASTM) were included where available, or whether an ISO guideline was currently under preparation, such as ISO/CD 24295 for AhR activation.

Validation maturity:
The maturity of each assay's validation was ranked according to the NORMAN Network validation guidance document [33]. That is, as 'high' if the assay had been validated at the routine level (i.e. inter-laboratory validation), as 'medium' if it had been validated at the expert level (i.e. basic external validation), and as 'low' if it had been validated at the research level (i.e. within-laboratory validation). The work of Wernersson *et al.* [28], Schriks *et al.* [34] and Concawe [27] helped inform this assessment.

Sensitivity:
Sensitivity in this context relates to how well the assay correctly identifies test substances that are active, i.e. substances that induce the biological effect the assay is designed to assess. A high sensitivity is therefore indicative of a higher chance of detecting an effect for particular substance(s). In cases where only specific compound types elicit the targeted mode of action, only the sensitivity of these compound types were considered. Since sensitivity is substance specific a global comparison cannot be made. Therefore, sensitivity was only discussed at high level in terms of being either 'high', 'variable', or 'low' as it can depend on not only the substance but also numerous other factors, such as cell line used and metabolic activation requirements of the substance. Where possible, the sensitivity of an assay was discussed in relation to that of assays that utilise the same mode of action (*in vitro*) or organism type (*in vivo*). The publications used to derive this information varied depending on the assay, but the work of Schriks *et al.* [34], Wieczerzak *et al.* [35], Brack *et al.* [29], Escher *et al.* [15] and Carere *et al.* [13] were useful sources.

Potency:
This criterion was used to assess whether the assay shows potency or magnitude of effect, or whether it generates a yes/no positive/negative result, such as is observed for many genotoxicity assays. An assay that shows quantitative potency of an effect/endpoint provides more useful toxicological information than a qualitative result.

Use in environmental samples:
The use in environmental samples was categorised as very common, common, occasional, or rare according to information provided by Schriks *et al.* [34] and Concawe [27], and examples given in Escher *et al.* [15]. These categories were expanded upon in the infocards to provide examples of the environmental samples that had been used with the assay, particularly in relation to water samples such as surface water, drinking water, recycled water and effluents. This category was less relevant in the evaluation than the two below.

Use with refinery effluent constituents:
Examples were given where the assay had been used for hydrocarbon samples, such as individual polyaromatic hydrocarbons (PAHs), oils and other mixtures of PAHs, as well as other constituents of refinery effluents such as metals. This was a key consideration for the evaluation.

Use with refinery effluents/produced water:
In addition to detailing whether the assay had successfully been applied to refinery constituents, it was distinguished if the assay had been used for whole effluent toxicity testing of refinery effluents or produced water. Produced water is a by-product of oil and gas extraction, whereby water trapped in underground formations is brought to the surface. As produced water contains substances similar to refinery effluent (e.g. hydrocarbons, phenols, inorganics), bioassays conducted on it are somewhat relevant in this context, although it must be highlighted that produced water also contains a wide range of chemicals not found in refinery effluents [36]. This evaluation criterion provides a much more relevant insight into the applicability of these tests for the REF BREF. It was noted where refinery effluents had been tested using the assay but where no response was observed, as this also provides useful insight and minimised positive bias, i.e. the reporting of only positive results.

Applicability to passive sampler extracts:
While not critical to the evaluation, the applicability of the test to passive sampler extracts was assigned ‘yes’ or ‘no’ in the evaluation tables, with references for studies that had successfully used passive sampler extracts for the test included in the infocards. Where it was not known if the test had previously been applied to passive sampler extracts, its applicability to these extracts was evaluated based on sample volumes [29] and high-throughput potential. This category was of low importance in the evaluation, as passive sampling is not yet considered ready for routine monitoring of refinery effluents [37].

Time to run:
Based on knowledge of the test guidelines, Schriks *et al.* [34], and use of the test in the academic literature, the time to run and receive results from the test was categorised into three categories: results within a day, results within a week, and results within two weeks. The exposure time was also included where available.

Cost per sample:
The costs were based on survey results, together with rankings provided by Wernersson *et al*. [28], who used price brackets of < 200 €, 200-500 €, and >500 € , and Schriks *et al*. [34], who used price brackets of <100 € per sample, 100-1,000 € per sample, and >1,000 € per sample. For Wernersson *et al*. [28], these cost estimates include the analyses involved and are primarily based on information from performing laboratories.

Relevance:
The relevance of the test endpoint or mode of action to ecotoxicological effects in the environment was evaluated and categorised as 'high', 'medium' or 'low'. This included consideration of a number of factors. The first being whether the effects result in a molecular initiating event or a key event that is linked to an adverse outcome, which can subsequently cause higher tier ecotoxicological impacts at e.g. the organism or population level. It was also considered whether the effect is likely to cause toxicity in the context of refinery effluents, i.e. a) in the aquatic environment, and b) by constituents found in refinery effluents.

Availability of trigger values:
Effect-based trigger values (EBT) that had been proposed (but not yet acknowledged) in the limited academic literature were collated. In particular, the extensive work of Van der Oost *et al*. [38] and Escher *et al*. [39] in this field were relied upon, amongst others. An EBT is a value below which a bioassay response is unlikely to produce adverse effects; it is comparable to a water quality guideline value but provides a specific threshold for each type and class of bioassay [15]. EBTs for the same assay may differ depending on the water type and its usage, for example EBTs for surface waters and effluents will be different to EBTs for drinking and recycled water, as the former are ecological health-relevant while the latter are human health-relevant. The derivation of the EBT will also differ depending on whether it is a category 1 or category 2 assay, see below. Owing to the uncertainty and lack of scientific consensus around EBTs, this category was only a minor consideration in the final evaluation.

"Bioassay category" [39]:
Escher *et al*. [39] has divided many of the evaluated assays into category 1 and category 2. Category 1 assays are highly specific. They respond to a limited set of substances, i.e. those that target a single highly specific molecular initiating event (MIE) such as estrogen, androgen or thyroid receptor binding. On the other hand, category 2 assays are applicable to a wide range of substances as they register more integrative effects, e.g. apical endpoints or activation of the oxidative stress response. This criterion therefore relates to an assay's suitability as a bioanalytical tool (ability to detect certain chemicals through effects).

Chemical analysis replacement:
Closely linked to the bioassay category is the possible replacement of chemical analysis. For each assay it was stated whether it could potentially replace chemical analysis. For assays that are specific to particular substances, i.e. category 1 bioassays [39], chemical analysis could possibly be replaced for the substances covered in the assay. However, it also needed to be considered whether the substances covered in the assay are likely to be present in refinery effluents.

Confounding factors:
The literature review and survey results informed the discussion on whether known confounding factors existed for the assays that would interfere with the interpretation and accuracy of results. For example, cytotoxicity can frequently mask effects in many of the assays evaluated here, such as in the oxidative stress and endocrine disruption assays.

The evaluation tables have been developed into the report in the form of infocards for each test method. This was considered the most systematic and comparable way to present the more detailed information that could not be included in the tables. Each infocard has the following sections:

- **Description.** This provides an overview of the test method, including the basic principle of the test, the organism/cell line used, mode of action (MoA) and endpoint measured.
- **Validation maturity.** This includes a ranking of the Validation maturity (from high to low), whether the test is available commercially, whether the test is standardised to an ISO/CEN guideline, whether it has been validated to water samples, and whether the test has been applied in a regulatory context.
- **Performance.** This describes how sensitive the test is and whether the effects show potency. Where available, comments on the predictability, specificity, repeatability (intra-laboratory variability), reproducibility (inter-laboratory variability) and high-throughput potential have been provided.
- **Use.** This provides information of the general frequency of use in environmental samples, whether the test has been used for refinery effluent constituents and/or whole samples of refinery effluents/produced water.
- **Conduct of test.** This describes the applicability to passive sample extracts, any sample processing steps required, and the time and cost to run the test.
- **Result interpretation.** This provides discussion of possible confounding factors that may interfere with the results, whether the results/endpoint used are relevant to the environment, the availability of trigger values, the bioassay category, and whether the test results provide details that could replace chemical analysis.
- **Possible application limitations.** This describes any practical or technical limitations to the application of the test, based on the available literature and results of the survey. Limitations could include timing constraints, intensive labour, specialist equipment, commercial licensing, sample handling/volumes, overall complexity etc.

3. EVALUATION OF *IN VIVO* TESTS

Out of approximately 20 tests initially identified, a total of 13 *in vivo* tests were selected for evaluation. Some tests did not have sufficient information available in the literature to warrant an infocard but have been discussed in the 'Other' and 'Novel developments' sections. It was not possible to cover every test or test organism, but it was ensured that the evaluated tests covered a range of trophic levels, from bacteria to fish, and included various endpoints, from acute assays measuring immobility and/or death, to chronic or sub-lethal endpoints such as growth and reproduction. Marine, brackish and freshwater test organisms have been considered so that tests may be swapped in or out of the proposed battery depending on the sample type and relevance for the receiving water body. Included in the full evaluations are decomposers (bacteria), primary producers (algae), primary consumers (daphnids and amphipods), and secondary consumers (fish, bivalves, echinoderms and nematodes). The effects on decomposers are important to elucidate as any cytotoxicity of such species will reduce the decomposer population, impacting decomposition and nutrient cycling within the aquatic ecosystem. Inhibition of primary producers reduces the amount of energy entering the food web, thus affecting higher trophic levels. Effects on primary and secondary consumers will similarly cause imbalance within the ecosystem. Some more novel assays have been included that may elucidate the genetic or biochemical mechanisms behind any observed toxicity in the other assays. However, genetic or biochemical changes may not always translate to whole organism effects. A summary of the evaluation can be found in the Appendix.

3.1. BACTERIA

3.1.1. Toxicity to *Allivibrio fischeri* e.g. MicroTox®

Description: This test uses the marine luminescent bacterium *Allivibrio fischeri*. The capacity of the bacteria to produce light determines the total toxicity of the sample. When the bacteria are exposed to a toxic sample, light production and luminescence decreases with time. The test is carried out using a temperature controlled photometric device and measures the luminescence of the bacteria over 30 minutes.
Validation maturity: The MicroTox® assay is available commercially and is standardised to ISO 11348-3. The test has been applied in regulatory context, under the IED [30]. It has been the subject of a number of inter-laboratory ring tests [27, 40], *Validation maturity*: High (routine lab) [27, 40].
Performance: *Sensitivity*: The test is highly sensitive to a wide range of compounds [41]. The commercially available kit is sensitive to over 2,700 "simple and complex" chemical compounds according to the manufacturer. It has been reported that the method is 1 to 3 times more sensitive than traditional animal tests for some substances (e.g. benzene) [42], and in refinery effluents, the Microtox® test was capable of detecting toxicity where acute *D. magna* tests did not [43]. *Potency*: The test shows potency of effect; effect concentrations are derived. The assay has low incidences of false positives and can be conducted in a 96-well plate for higher through-put capacity [43].

Use:

Use in environmental samples: The use of the assay in environmental samples is very common [27, 40]. The assay can be applied to both marine/brackish and freshwater samples. In surveys of EU refineries (2010, 2013, 2016, 2019), the test was applied to brackish/salt water at 11 sites, and to freshwater at 4 sites [32]. It has also been used for wastewater [41, 44-46] and various other water samples such as drinking water, river water and seawater [15, 44, 47]. Another relevant constituent of refinery effluents are metals, for which the test has also been applied to [14, 48].

Use with refinery effluent constituents: Allivibrio fischeri tests have produced responses with benz(a)pyrene (BaP) [47], benzo(b)fluoranthene [47], naphthenic acids [49], and class-fractionated oils [50]. In terms of metals, the MicroTox® assay has been reported to be very sensitive to metals (e.g. zinc, mercury, copper, cadmium, lead, nickel), with the exception of chromium [51, 52].

Use with refinery effluents/produced water: According to surveys conducted by Concawe in 2010, 2013, 2016 and 2019 [32], the assay has been commonly used to test European refinery effluents; it was the second most used test behind *Daphnia*. Academic literature supports these findings about the use of MicroTox® in toxicity testing of refinery effluents, e.g. Whale *et al*. [43], Comber *et al*. [53] and Whale *et al*. [5].

Conduct of test:

Applicability to passive sample extracts: The test utilises small sample volumes and is applicable to passive sample extracts [45, 54, 55].

Time to run: Results within a day; ≤30-minute exposure.

Sample processing: Samples can be stored; toxicity was not significantly affected by a three-week storage period under 4°C [5]. Test dilutions are prepared in glass cuvettes immediately prior to exposure, thus minimising loss via adsorption to test vessels. Turbid samples will need filtration or centrifugation [44]. Sample enrichment may be required.

Cost per sample: The assay is likely to be under 1,000 € per sample based on the results of the survey; the most commonly reported cost was 300 - <1,000 €, with 67% of laboratories that conduct the MicroTox® assay reporting this cost. One laboratory reporting a cost of under 100 € per sample.

Result interpretation:

Confounding factors: Turbidity and colour can interfere with the quantification of bioluminescence, however black microplates can be utilised for coloured samples [44]. Increased chemical oxygen demand (COD) is positively correlated to increased inhibition of bacterial bioluminescence [53]. Lipophilicity also has a strong correlation with response in the assay [56]. Further, dilution of saline effluent samples for the assay may mean the range of concentration that causes toxicity is not covered.

Relevance: High. Although it only provides information about non-specific effects, it measures lethality, which is of high (eco)toxicological relevance. Bacteria are key contributors to the functioning of an ecosystem; reduction of their population would incite changes in the food web as well as nutrient recycling. Further, MicroTox® is highly relevant to the constituents of refinery effluents, responsive to PAHs, oils, metals and refinery effluents.

Availability of trigger values: Trigger values of 0.05 toxic units (TU) [38], 1 toxic unit [41] and 1.2 mg baseline toxicant equivalent /L [39] have been proposed in the literature.

Bioassay category: Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
The assay has a solubility cut-off for baseline toxicants with high melting points and/or high hydrophobicity, so won't be applicable to all substances, and may neglect certain substances when testing whole effluents [57]. However the response seen in tests with refinery effluents suggest that hydrophobic constituents present in refinery effluents are not outside the applicability domain of MicroTox®. It uses a licensed strain and specialist equipment, which may make it difficult to apply to widescale assessment [40]. The results of the survey indicate that the salinity should be known/taken into account.

3.1.2. Multi-species microbial toxicity assay e.g. MARA/LumiMara

Description:
This is a 24-hour assay using multiple species, utilising 11 species of microorganism (10 bacteria and 1 fungus). Microorganism growth is measured by the reduction of the redox dye, tetrazolium red (TTC), which can be observed spectrophotometrically. Each different strain of microorganism shows different levels of sensitivity to a toxic sample and the growth inhibition values for the 11 species allows the identification of a 'toxic fingerprint' for the chemical tested.
MARA is the Microbial Assay for Risk Assessment. The test can also be conducted using 9 naturally occurring marine species of bioluminescent bacteria, where luminescence can be used as the endpoint to measure toxicity to the bacteria directly (the LumiMara).

Validation maturity:
The assay is available commercially but is not standardised to an ISO guideline. The test has had limited application in regulatory context [30] and an inter-laboratory ring-test has been performed [58].

Validation maturity: High (routine lab) [27, 40].

Performance:
Sensitivity: The assay is highly sensitive, especially towards *V. fischeri* [40].

Potency: The test shows potency of effect; effect concentrations are derived.

Use:
Use in environmental samples: The use of the assay in environmental samples is common [27, 40]. The assay can be applied to both marine and freshwater samples. It has been used for raw waters, industrial effluents, sewage sludge and soil leachates [40].

Use with refinery effluent constituents: The assay has been responsive to crude oil in wastewater treatment plant effluents [59]. LumiMara has been tested with PAHs and metals [60], both of which are constituents of refinery effluents.

Use with refinery effluents/produced water: No information could be found.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes and is applicable to passive sample extracts.

Time to run: Results within a day; 18- to 24-hour exposure.

Sample processing: No known sample processing considerations.

Cost per sample: The results of the survey (n=1) indicate a cost of 1,000 - <3,000 € per sample.

Result interpretation: *Confounding factors:* Turbidity and colour can interfere with the quantification of bioluminescence, however black microplates can be utilised for coloured samples [44]. Increased chemical oxygen demand (COD) is positively correlated to increased inhibition of bacterial bioluminescence [53]. *Relevance:* High. The assays expose many different species concurrently, introducing the concept of investigating community effects, rather than just effects in individuals. This is of very high ecological relevance. The assay is also highly relevant to the constituents of refinery effluents, responsive to PAHs, oils, and metals. *Availability of trigger values:* No trigger values were found in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The survey raised that sampling events are unpredictable, and that reagent supply from the supplier can be limitations affecting promptness/ease of conducting the assay.

3.2. ALGAE

3.2.1. Algal growth inhibition

Description: This assay determines the effects of a substance on the growth of algae, either freshwater or marine. Growing algae are exposed to varying concentrations of the test substance for 72 hours to determine a reduction in growth, which is deemed a chronic endpoint [61]. The growth of the test batches are compared to the average growth of the culture controls. Each culture is allowed unrestricted growth under nutrient sufficient conditions and continuous fluorescent illumination. Growth and growth inhibition can be quantified from measurements of algal biomass over time.
Validation maturity: The assay is available commercially and is standardised to ISO 8692 (freshwater) and 10253 (marine). The test has been applied in regulatory context, under the IED [28, 30]. A miniaturised version exists, but this introduces difficulties in reaching validity criteria of the ISO guideline. *Validation maturity:* High (routine lab) [27, 40].
Performance: *Sensitivity*: The test is highly sensitive [29], with greater sensitivity compared to the algal PAM test for many substances [62]. Further, bacteria, daphnids and algae have been reported to be more sensitive with respect to acute toxicity than fish [43]. *Skeletonema costatum* was found to be the most sensitive species among daphnids and fish tested in acute marine toxicity tests of a refinery effluent [43]. *Potency:* The test shows potency of effect; effect concentrations are derived.

Use:

Use in environmental samples: The use of the assay in environmental samples is very common [27, 40]. The assay can be applied to both marine/brackish and freshwater samples. In surveys of EU refineries (2010, 2013, 2016, 2019), the test was applied to brackish/salt water at 2 sites, and to freshwater at 1 site [32] .It has been used for wastewater effluent, recycled water, drinking water and surface water [28, 46, 63].

Use with refinery effluent constituents: The algal growth inhibition assay has been used for B(a)P and benzo(b)fluoranthene [47]. The test has produced responses with light catalytic cracked gas oil and light thermal cracked gas oil [64]. In a metal-contaminated river, the samples that inhibited algal growth (n=10), 8 of them also showed high concentrations of free ionic forms of lead or zinc [65]. Metals are thought to inhibit algal growth by affecting membrane integrity (cadmium, chromium and copper) and inhibiting esterase activity (cadmium, chromium, copper and zinc), as well as reducing chlorophyll a content, decreasing the maximum quantum yield of photosystem II and modifiying of mitochondrial membrane potential [66].

Use with refinery effluents/produced water: According to surveys conducted by Concawe in 2010, 2013, 2016 and 2019 [32], the assay has been commonly used to test European refinery effluents. The literature also reports its use for refinery effluents [5, 43, 53].

Conduct of test:

Applicability to passive sample extracts: The test utilises relatively small sample volumes (200–2,000µL typical final volume [29]) and is applicable to passive sample extracts [55].

Time to run: Results within a week; ≤72-hour exposure.

Sample processing: Sample preparation such as filtration to remove confounding factors (e.g. suspended particles or coloured samples) may be required.

Cost per sample: The assay is likely to be under 100 to under 10,000 € per sample based on the results of the survey; the most commonly reported cost was 1,000 - <3,000 €, with 50% of the laboratories that conduct the algal growth inhibition assay reporting this cost.

Result interpretation:

Confounding factors: High ammonium concentrations in samples may increase toxicity [43]. Increased COD is positively correlated to increased inhibition of algal growth [53]. The sensitivity of the species used may also affect results [43]. Other confounding factors include organic matter, alkalinity, water hardness and nutrient concentrations-particularly in cases where the native sample is used [29]. When testing heavy-metal contaminated river samples, most (20 out of 30) samples did not inhibit algal growth, thought to be inhibited by the presence of organic ligands, colloidal organic matter and antagonism effects among other metals, which reduce their bioavailability [65].

Relevance: High. Tests with algae have ecological relevance as algae are primary producers (an important link in the food web), are sensitive to a broad range of compounds and are good indicators of variations in the environment [43]. Further, algae are found in marine, brackish and freshwater environments where refineries release to. The assay is highly relevant to refinery effluents as it has been responsive to PAHs, metals, oils, and refinery effluents.

Availability of trigger values: Trigger values of 0.05 TU [38] and 116.5 ng Diuron equivalents /L [39] have been proposed in the literature.

Bioassay category: Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations: The results of the survey indicate that an adapted room may be required, specific analytical determination is needed during the experimental phase, and that it may be necessary to perform chlorophyll extraction to determine the cell density. Timing constraints may present issues, but generally the algae stocks are continuously maintained and easy and fast to set up for experiments when needed.

3.2.2. Combined algal assay

Description: This is a 96-well plate assay which uses the green algae *Raphidocelis subcapitata*. The test combines the inhibition of growth rate over 24 hours with the inhibition of photosynthesis (specifically photosystem II) detected after 2- and 24-hours, as measured by a pulse-amplitude modulated (PAM) fluorometer.
Validation maturity: The assay is available commercially and an ISO guideline is in preparation. The test has had limited application in regulatory context [30]. *Validation maturity:* Medium (expert lab) [27].
Performance: *Sensitivity*: The sensitivity is variable; it is generally less sensitive than the standard algal growth test, except for PSII-inhibitors such as herbicides, which are not relevant to refinery effluents [27]. *Potency:* The test shows potency of effect; effect concentrations are derived.
Use: *Use in environmental samples:* The use of the assay in environmental samples is occasional [27]. The assay can be applied to only freshwater samples. It has been used for wastewater effluent [67-70], surface water [68-70], drinking water [68] and recycled water [67, 68]. *Use with refinery effluent constituents:* The assay has been used to assess toxicity of oil spill-contaminated filtered seawater and sediment elutriates [71]. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes and is applicable to passive sample extracts [27]. *Time to run:* Results within a week; 24-hour exposure. *Sample processing:* Sample preparation such as filtration to remove confounding factors (e.g. suspended particles or coloured samples) may be required. *Cost per sample:* Unknown. While one laboratory in the survey conducted this assay, the cost was not known.

Result interpretation: *Confounding factors:* High ammonium concentrations in samples may increase toxicity [43]. Increased COD is positively correlated to increased inhibition of algal growth [53]. The sensitivity of the species used may also affect results [43]. Other confounding factors include organic matter, alkalinity, water hardness and nutrient concentrations-particularly in cases where the native sample is used [29]. *Relevance:* High. Tests with algae have ecological relevance as algae are primary producers (an important link in the food web), are sensitive to a broad range of compounds and are good indicators of variations in the environment [43]. Further, algae are found in marine, brackish and freshwater environments where refineries release to. Algae are expected to be responsive to oils, PAHs and whole refinery effluents. *Availability of trigger values:* Trigger values of 130 (growth) and 73.7 (PSII) ng Diuron equivalents /L ([39], and 0.05 TU [54] have been proposed in the literature. *Bioassay category:* Category 1 as it measures specific effects on an organism level (inhibition of photosynthesis) [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The application limitations of the algal growth inhibition assay are expected to be present for the combined algal assay too.

3.3. INVERTEBRATES

3.3.1. *Daphnia* immobilisation

Description: This is an acute toxicity assay to assess the effects of chemicals towards *Daphnia*, usually *Daphnia magna*. Daphnids aged < 24 hours are exposed to a range of concentrations (5+) of the test substance for duration of 48 hours. Immobilisation of the daphnids is recorded at 24 and 48 hours and is compared with the control batch. These results can then be used to calculate an EC50 at 24 and 48 hours.
Validation maturity: The assay is available commercially and is standardised to ISO 6341. The test has been applied in regulatory context, under the IED [30]. *Validation maturity:* High (routine lab) [27].
Performance: *Sensitivity*: The test is highly sensitive [29]. Further, bacteria, daphnids and algae have been reported to be more sensitive with respect to acute toxicity than fish [43]. *Potency:* The test shows potency of effect; effect concentrations are derived.
Use: *Use in environmental samples:* The use of the assay in environmental samples is very common [27]. *Daphnia magna* is a freshwater species, however, the assay has been used for other water samples too. In surveys of EU refineries (2010, 2013, 2016, 2019), the test was applied to brackish/salt water at 9 sites, and to freshwater at 12 sites [32]. It has been used for surface water quality assessment [54, 72, 73], as well as wastewaters and seawater [35].

Use with refinery effluent constituents: The test has produced responses with B(a)P, benzo(b)fluoranthene and other PAHs [47, 74]. The test has produced responses with light catalytic cracked gas oil and light thermal cracked gas oil [64]. In *D. magna* acute toxicity testing of 50 metals, the lowest EC50 values (<100 µg/L) were produced from copper, cadmium, mercury, silver, gold, beryllium and osmium [75]. This correlates well with previous research that found the order of metal toxicity to *D. magna* was mercury>silver>copper>zinc>cadmium>cobalt>chromium>lead>nickel>selenium [76].

Use with refinery effluents/produced water: According to surveys conducted by Concawe in 2010, 2013, 2016 and 2019, the assay has been commonly used to test European refinery effluents; in fact, it was the most used test every survey year [32]. Academic literature supports these findings about its use in toxicity testing of refinery effluents [43, 53].

Conduct of test:
Applicability to passive sample extracts: The test utilises relatively small sample volumes (300–4,000µL typical final volume [29]) and is applicable to passive sample extracts [54, 72, 73].

Time to run: Results within a week; 48-hour exposure.

Sample processing: No known sample processing considerations.

Cost per sample: The assay is likely to be above 100 to under 10,000 € per sample based on the results of the survey; the most commonly reported cost was 1,000 - <3,000 €, with 50% of the laboratories that conduct the *Daphnia* immobilisation assay reporting this cost.

Result interpretation:
Confounding factors: Confounding factors include the influence of constituents other than pollutants, and potential problems with pH or oxygen content (particularly in cases where the native sample is used) [29].

Relevance: High. Water fleas are very sensitive to toxic chemicals and occupy a central place in the food web. As aquatic organisms they are highly relevant to the testing of refinery effluents and are responsive to refinery effluent constituents such as oils, (heavy) metals and PAHs, as well as whole refinery effluents.

Availability of trigger values: Trigger values of 0.05 TU [38] and 15 ng Chlorpyrifos-equivalents /L [39] have been proposed in the literature.

Bioassay category: Category 1 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
May require maintenance of a breeding *Daphnia* population before the experiment, with daily sorting of mothers from neonates to keep track of neonate age for use in the test. The results of the survey indicate that the health of the culture can affect results or timings, pH adjustments may be required (which often require modification to culturing), specific analytical determination is needed, and that cultures bought on an ‘as and when’ basis may be contaminated with unwanted species.

3.3.2. *Daphnia* reproduction

Description:
This assay assesses the reproduction of *Daphnia magna* following exposure to a chemical in a semi-static test. Young female daphnids are exposed to the test medium at a range of concentrations (5+). The daphnids are exposed for 7-21 days and at the end of the test, the total number of living offspring from each parent animal and the number of living offspring per surviving parent animal are reported.

Validation maturity:
The assay is available commercially and although not ISO standardised, can be conducted according to OECD 211. The test has been applied in regulatory context [30]. A miniaturised version is available.

Validation maturity: High (routine lab) [27].

Performance:
Sensitivity: The assay's sensitivity is variable. Bacteria, daphnids and algae have been reported to be more sensitive with respect to acute toxicity than fish, however the test is not always responsive to refinery effluent samples where toxicity using the acute test has been demonstrated [43, 53].

Potency: The test shows potency of effect; effect concentrations are derived.

Use:
Use in environmental samples: The use of the assay in environmental samples is very common [27]. *D. magna* is a freshwater species, however, the assay has been used for other water samples too [32]. It has been used for effluents and surface waters [35, 43, 53, 77].

Use with refinery effluent constituents: The test has produced responses with light catalytic cracked gas oil [64]. In tests with water soluble fractions of diesel and biodiesel, the chronic test produced responses with diesel but not biodiesel, although the acute test detected toxicity to both diesel types [78].

Use with refinery effluents/produced water: It has also been responsive to flowback and produced water from hydraulic fractioning [79]. Toxicity (EC50 reproduction) was observed somewhat for two samples (mid-treatment and final effluent) in refinery effluent tested in Whale *et al. [5].*

Conduct of test:
Applicability to passive sample extracts: The test is not applicable to passive sample extracts, unless miniaturised [27].

Time to run: Results within a week; 7-day exposure (depending on guideline followed).

Sample processing: No known sample processing considerations.

Cost per sample: The assay is likely to be 1,000 to above 10,000 € per sample based on the results of the survey; the most commonly reported cost was >10,000 €, with 55% of the laboratories that conduct the *Daphnia* reproduction assay reporting this cost.

Result interpretation:
Confounding factors: Confounding factors include the influence of constituents other than pollutants (such as nitrite or ammonium), and potential problems with pH, conductivity levels or oxygen content (particularly in cases where the native sample is used) [5, 29].

Relevance: High. Water fleas are very sensitive to toxic chemicals and occupy a central place in the food web. Further, reproductive effects have the potential to cause effects at the population level and across generations. As aquatic organisms they are highly relevant to the testing of refinery effluents, and the assay has produced responses in oil and produced water. *Availability of trigger values:* A trigger value of 0.05 TU [38] has been proposed in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: May require maintenance of a *Daphnia* population before the experiment, with daily sorting of mothers to keep track of their age for use in the test. Further, a large sample volume is required due to the length of the test and the test medium renewals during the duration. This introduces questions regarding sample collection; a single large sample can be collected to ensure the test medium is homogeneous throughout, but storage and preservation of the sample may affect results, or smaller samples can be collected at various points to mitigate storage/preservation issues, but the composition of each sample may differ. The results of the survey indicate that the health of the culture can affect results or timings, pH adjustments may be required (which often require modification to culturing), specific analytical determination is needed, and that cultures bought on an 'as and when' basis may be contaminated with unwanted species.

3.3.3. *Daphnia magna* metabolic activity test e.g. IQ™ toxicity test

Description: This assay uses *Daphnia magna* exposed to a chemical at varying concentrations. After the organisms have been in contact with the water for 1 hour, a fluorometric biomarker (methylumbelliferyl galactoside (MUF)) is added, followed by illumination with a black light (longwave UV). Control organisms will emit a bright bluish-white light, indicating that they are healthy and have been able to feed normally, cleaving the marker from the substrate. Organisms in the sample water which glow less brightly are considered to be adversely affected.
Validation maturity: The assay is available commercially but is not ISO standardised. An inter-laboratory ring test has been performed [80]. *Validation maturity:* Medium (expert lab) [27].
Performance: *Sensitivity*: No information could be found. *Potency:* The test shows potency of effect; effect concentrations are derived.
Use: *Use in environmental samples:* The use of the assay in environmental samples is rare [27]. *Daphnia magna* is a freshwater species, however, the assay has been used for other water samples too [32]. It has thus far only been used intermittently to test single chemicals, synthetic mixtures and temperature effects, not for the testing of environmental samples. *Use with refinery effluent constituents:* No information could be found. *Use with refinery effluents/produced water:* No information could be found.

Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes and is applicable to passive sample extracts. *Time to run:* Results within a day; 1 hour 15-minute exposure. *Sample processing:* No known sample processing considerations. *Cost per sample:* Unknown. No survey respondents conducted this survey.
Result interpretation: *Confounding factors:* The test suffers from some interpretational difficulties [27]. *Relevance:* High. Water fleas are very sensitive to toxic chemicals and occupy a central place in the food web. As aquatic organisms they are highly relevant to the testing of refinery effluents. The relevance to the constituents of refinery effluents is unknown, however *Daphnia* are generally responsive to PAHs, oils, metals etc. *Availability of trigger values:* No trigger values were found in the literature. *Bioassay category*: Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: May require maintenance of a *Daphnia* population before the experiment, with associated limitations discussed in the infocards above.

3.3.4. Bivalve/echinoderm embryo development

Description: This assay utilises bivalves (e.g. mussels and oysters) or echinoderms (e.g. sea urchins), and their well-characterised development. Groups of bivalve or echinoderm embryos are exposed to the chemical sample at varying concentrations diluted with water for a period of 24-72 hours, to determine toxic effects on normal embryo development. Normal bivalve development involves transformation into "D-shaped" larvae. Abnormal development is considered to be embryos which die at an early stage or fail to reach the D-shaped stage. Embryonic development of echinoderms can similarly be monitored by looking at growth inhibition and skeletal malformations. This data can be used to determine ECx, NOEC and LOEC values. The suitability of different water types depends on the species used. Bivalve tests can be conducted with marine (e.g. oysters) or freshwater (e.g. freshwater mussel) species. Brackish water can also be tested using oysters, as these are euryhaline species that can tolerate a range of salinities. For the echinoderm test, marine water is generally used, as no freshwater echinoderm species exist.
Validation maturity: The assays are available commercially and standardised to ISO 17244 and ASTM E1563-21a. The test has been applied in regulatory context [30]. Measuring embryo-larval length instead of abnormalities has enabled considerable standardisation and highly consistent results [81]. *Validation maturity:* High (routine lab) for the bivalve assay and medium (expert lab) for echinoderms [27, 40].

Performance:
Sensitivity: Both assays utilise embryos, which are highly sensitive test organisms [82]. The sensitives of the bivalve and echinoderm assays are generally comparable, but will vary slightly depending on the substances tested [83].

Potency: The test shows potency of effect; effect concentrations are derived.

The assays are high-throughput [40].

Use:
Use in environmental samples: The use of the bivalve assay in environmental samples is very common, while the echinoderm version is used commonly [27, 40]. The assay with bivalves can be applied to marine, brackish and freshwater samples, while the assay with echinoderms are used for marine species. The tests have been used for wastewater effluents [84-86], seawater [71] and marine sediment elutriates [87, 88].

Use with refinery effluent constituents: Phenanthrene, naphthalene, pyrene and fluorene were toxic at the experimental concentrations to sea urchin and mussel embyros, while fluoranthene was toxic to sea urchin embryos but not mussel embryos [89]. Phenanthrene has been shown to affect oyster development [40]. Further, in toxicity tests with metals and bivalve embryos (*Ruditapes decussatus*) and (*Mytilus galloprovincialis*), the EC50 values ranged from 4 to 10 µg/l for mercury and copper, from 100 to 300 µg/l for zinc and lead, and from 400 to 2000 µg/l for cadmium [90].
The sea urchin assay has been used for various hydrocarbon-relevant samples. For example, Maya crude oil fractions (aliphatics, aromatics and polars), of which the aromatics were most toxic [81]. The joint toxicity of the mixtures was best explained by concentration addition [81]. The sea urchin assay has also been used to measure toxicity to oil spill-contaminated seawater [71].

Use with refinery effluents/produced water: The bivalve test has been used to test oysters with brackish and marine refinery effluent samples [91].

Conduct of test:
Applicability to passive sample extracts: The test is generally not applicable to passive sample extracts, however silicone passive sampler extracts of sediments have been used to test sea urchins in the test [88].

Time to run: Results within a week; 24-hour (bivalve) and 48-hour to 72- hour exposure (echinoderm).

Sample processing: The salinity of the sample may need to be adjusted depending on the species used [91].

Cost per sample: Both assays are likely to be 1,000 to under 10,000 € per sample based on the results of the survey; 50% of the respondents that conduct the embryo development assays indicated costs of 3,000 - <10,000 €, with the other 50% indicating costs of 1,000 - <3,000 €.

Result interpretation:
Confounding factors: Osmotic stress may confound toxicity if not adjusted for [91].

Relevance: High. Echinoderms and molluscs are ecologically important species, whose embryonic development is highly sensitive [81]. However, it should be noted that testing using embryos can be overprotective as they are often more sensitive than their adult counterparts. The test measures not only survival, but teratogenicity as well, which is of high ecotoxicological relevance. In terms of relevance to refinery effluents, this is also high as the assays are expected to be responsive to various constituents, e.g. PAHs, oils, metals, whole effluent samples etc.

Availability of trigger values: No trigger values were found in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: It has been reported that tests may have to wait until the reproductive season of bivalves or echinoderms (spring/summer in the UK), however, responses can still be seen after water samples are frozen and thawed [71]. The survey added that the spawning condition of the broodstock, and the availability of the adult broodstock, are potential limitations for both the bivalve and embryo test. It may also be necessary to order large batches of animals for the correct male/female ratio, and spawning during transport needs to be avoided.

3.3.5. Nematode growth and reproduction

Description: This assay assesses the toxicity of environmental samples (aqueous media, freshwater sediments and soils) on the growth, fertility and reproduction of *Caenorhabditis elegans* (nematoda). Growth (body length) and reproduction (number of juveniles generated from two adult hermaphrodite worms) can be determined after 72 or 96 hours of exposure. From this, EC50, NOEC and LOEC values can be calculated.
Validation maturity: The assay is available commercially and standardised to ISO 10872. The test has been applied in regulatory context [30]. The test has undergone an inter-laboratory ring test, demonstrating an acceptable reproducibility and repeatability for growth and reproduction [92]. *Validation maturity*: High (routine lab) [27].
Performance: *Sensitivity*: The sensitivity is variable. Its sensitivity is generally low to PAHs, and might be less sensitive to PAHs than the *D. magna* acute test [27]. *Potency:* The test shows potency of effect; effect concentrations are derived.
Use: *Use in environmental samples:* The use of the assay in environmental samples is occasional [27]. The assay can be applied to freshwater samples. It has been used for freshwater sediment [93], surface water [94] and wastewater effluent [95]. *Use with refinery effluent constituents:* PAHs demonstrate low acute toxicity to *C. elegans*, however prolonged exposure (72 hours) at high concentrations resulted in mortality; reproduction and growth of *C. elegans* were much more sensitive [96]. *C. elegans* was approximately 2-fold less sensitive to acenaphthene, 5-fold less sensitive to phenanthrene, and 20-fold less sensitive to fluoranthene than *D. magna* [96]. The fecundity and growth of *C. elegans* has been reported to be reduced by cadmium, excess copper and excess zinc [97]. *Use with refinery effluents/produced water: C. elegans* has been used to test refinery effluents [98]. According to surveys conducted by Concawe in 2010, 2013, 2016 and 2019, an assay with benthic invertebrates has been used to test European refinery effluents, although it is not known which test organism was used [32].

Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes and is applicable to passive sample extracts. *Time to run:* Results within a week; ≤96 hour-exposure. *Cost per sample:* The results of the survey (n=1) indicate a cost of 1,000 - <3,000 € per sample.
Result interpretation: *Confounding factors:* No information could be found. *Relevance:* Medium. Nematodes are often the dominant taxon in soil, can be present in aquatic sediment, occupy a key part of the benthic food web, and are sensitive to a wide range of pollutants [28]. Effects observed in a predominantly soil organism are, however, of lower relevance in the context of refinery effluents, although some response to PAHs and metals has been demonstrated. *Availability of trigger values:* No trigger values were found in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: No application limitations have been found in the literature or raised during the survey.

3.3.6. Amphipod (*Corophium volutator*) assay

Description: This assay is used to assess the acute toxicity of both contaminated freshwater and marine water sediments using *Corophium volutato*r, over an exposure time of 10 days. The test endpoint is the percentage of mortality observed in *Corophium volutato*r at the end of the study duration.
Validation maturity: The assay is available commercially and standardised to ISO 16712 and ASTM E1367-99. The test has been applied in a regulatory context within OSPAR, but for sediment only. *Validation maturity:* High (routine lab).
Performance: *Sensitivity*: The sensitivity is variable; it has been demonstrated to be much less sensitive that other invertebrates [99] *Potency:* The test shows potency of effect; effect concentrations are derived.
Use: *Use in environmental samples:* The use of the assay in environmental samples is common. The assay can be applied to both marine, estuarine and freshwater samples as it is a euryhaline species, i.e. able to tolerate a range of salinities [43]. *Use with refinery effluent constituents:* The assay has been used for crude oil-contaminated sand [100], oil-spiked marine sediment [101], hydrocarbon-contaminated sediments and historic drilling mud discharges in the North Sea [43, 53]. *Use with refinery effluents/produced water:* Refinery streams have been successfully tested using the assay [43, 53]. *C. volutator* is also sensitive to oil refinery effluents [102].

Conduct of test: *Applicability to passive sample extracts:* The test is not applicable to passive sample extracts, unless miniaturised. *Time to run:* Results within 2 weeks; 10-day exposure. *Cost per sample:* The assay is likely to be 100 to under 10,000 € per sample based on the results of the survey; the most commonly reported cost was 3,000 - <10,000 €, with 40% of the laboratories that conduct the *C. volutator* assay reporting this cost.
Result interpretation: *Confounding factors:* Low salinity was found to enhance the toxicity of ammonia for *C. volutator* [102]. Oxygen saturation and ammonium concentration can also influence the test result, although they can be controlled by a proper aeration and refreshing regime [103]. *Relevance:* Medium. Amphipods are sediment-dwelling organisms so are less relevant to the assessment of refinery effluents, however they are expected to be responsive to the constituents of refinery effluents. *Availability of trigger values:* No trigger values were found in the literature. *Bioassay category*: Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: Seasonality of the species is an application limitation; the survey raised that the reproductive season is typically April/May to August/September. It was further raised that it is very difficult to obtain the test organism (both in terms of season and sites); it is generally not possible during winter. Further, the test set-up is rather crude, requires a degree of handling, and takes up space.

3.4. FISH

3.4.1. Fish embryo toxicity test (FET)

Description: This assay utilises fish (e.g. *Danio rerio*) embryos. Newly fertilised fish eggs are exposed for a duration of 96 hours to a chemical at varying concentrations. Every 24 hours, four observations are made at each chemical concentration: 1. coagulation of fertilised eggs; 2. lack of somite formation; 3. lack of detachment of the tail-bud from the yolk sac, and; 4. lack of heartbeat. At the end of the 96 hours, acute toxicity can be determined based on a positive outcome in any of the four observations recorded, and the LC50 is calculated.
Validation maturity: The assay is available commercially and standardised to ISO 15088 and OECD 236. OECD 212 could also be used. The test has been applied in regulatory context, under the IED [30]. A round-robin study has been conducted, with intralaboratory variability of 13.6% and interlaboratory variability of 21.3% [104]. *Validation maturity*: High (routine lab) [27, 40].

Performance:
Sensitivity: The sensitivity is generally high [29], however, a number of studies (see below) suggest that the sensitivity of the FET to hydrocarbons and refinery effluents is typically lower than for tests with algae or *Daphnia*. Many chemicals act as baseline toxicants in the FET assay [47].

Potency: The test shows potency of effect; effect concentrations are derived.

It can be conducted in 96-well plates, which increases its high-throughput ability [47].

Use:
Use in environmental samples: The use of the assay in environmental samples is very common [40]. The assay is only applied to freshwater samples when using zebrafish, i.e. a freshwater species, but the protocol can be adapted to use other species, such as stickleback, to test a wider range of samples from the estuarine or marine environment [28]. The OECD 212 guideline would also present an alternative for using marine species, e.g. cod, tidewater silverside, herring or sheepshead minnow. In surveys of EU refineries (2010, 2013, 2016, 2019), the test was applied to brackish/salt water at no sites, and to freshwater at 7 sites [32]. It has been used for surface waters [47, 63, 105] and wastewater effluents [68].

Use with refinery effluent constituents: The test has produced responses with light catalytic cracked gas oil [64], but not with B(a)P or benzo(b)fluoranthene [47].

Use with refinery effluents/produced water: According to surveys conducted by Concawe [32], ISO 15088 has been used to test European refinery effluents. However, academic literature reports low or no response to refinery effluents, when responses were seen in other test organisms [5, 53]

Conduct of test:
Applicability to passive sample extracts: The test utilises relatively small sample volumes (300–4,000µL typical final volume [29]) and is applicable to passive sample extracts [72]. In a 48 hour FET assay, styrene divinylbenzene 'Speedisk' samplers were more responsive than silicone rubber samplers [72].

Time to run: Results within a week; 96-hour exposure.

Cost per sample: The literature reports a cost of <200 € [28]. However, results of the survey indicate the assay is likely to be 1,000 to above 10,000 € per sample; the most commonly reported cost was 3,000 - <10,000 €, with 38% of the laboratories that conduct the FET assay reporting this cost. 25% of the laboratories indicated costs of 1,000 - <3,000 €, and a further 25% indicated costs of >10,000 €.

Result interpretation:
Confounding factors: Reproducibility of results decreased for substances tested close to their water solubility limit, and for very toxic or volatile substances [104]. Other confounding factors include the influence of constituents other than pollutants, and potential problems with pH or oxygen content (particularly in cases where the native sample is used) [29]. The test set-up applied can also affect results, with EC50 values lower in tests conducted in glass vials than those run in 96-well polystyrene plates, likely due to chemical sorption to the plastic [47].

Relevance: High. No evidence of MoA, however tests with fish are highly relevant for the assessment of water samples. While testing using embryos can be overprotective for some organisms, fish embryo toxicity and acute toxicity of adult fish are closely related (r^2=0.95) [106].

Availability of trigger values: A trigger value of 0.183 mg Bisphenol A equivalent /L [39] has been proposed in the literature. *Bioassay category:* Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The use of stored samples is not recommended [28] and experience is required [29]. The survey results agreed that experience was required, and further mentioned that the developmental times vary depending on the species used- 3 to 4 days for zebrafish but up to 30 days for lumpfish.

3.4.2. Zebrafish qFET

Description: This assay is based on the Fish Embryo Acute Toxicity (FET); however it has been adapted to be used with low volume samples, and uses a 24-well microplate. Embryos < 4 hours post-implantation are used, with each well containing 12-13 embryos and 2ml of test solution. The assay duration is up to 120 hours and each 24 hours the embryos are scored for visible malformations. The LOEC, NOEC and ECx values can then be calculated.
Validation maturity: The assay is available commercially and standardised to ISO 15088 and OECD 236. The test has been applied in regulatory context, under the IED [30]. *Validation maturity:* High (routine lab) [27].
Performance: *Sensitivity*: The sensitivity is generally high [29], however, a number of studies (see below) suggest that the sensitivity of the FET, and thus qFET, to hydrocarbons and refinery effluents is typically lower than for tests with algae or Daphnia. Many chemicals act as baseline toxicants in the FET assay [47]. *Potency:* The test shows potency of effect; effect concentrations are derived. It can be conducted in 96-well plates, which increases its high-throughput ability.
Use: *Use in environmental samples:* The use of the assay in environmental samples is very common [27]. The assay is only applied to freshwater samples when using zebrafish, i.e. a freshwater species, but the protocol can be adapted to use other species, such as stickleback, to test a wider range of samples from the estuarine or marine environment [28]. It has been used for surface waters [72]. *Use with refinery effluent constituents:* The test was not responsive to B(a)P or benzo(b)fluoranthene [47]. *Use with refinery effluents/produced water:* According to surveys conducted by Concawe [32], ISO 15088 has been used to test European refinery effluents, although this is more likely to be the FET assay.

Conduct of test: *Applicability to passive sample extracts:* The test utilises relatively small sample volumes (300–4,000µL typical final volume [29]) and is applicable to passive sample extracts [72]. In a 48 hour FET assay, styrene divinylbenzene 'Speedisk' samplers were more responsive than silicone rubber samplers [72]. *Time to run:* Results within a week; 96-hour exposure. *Cost per sample:* Unknown. No survey respondents conducted this survey.
Result interpretation: *Confounding factors:* Reproducibility of results decreased for substances tested close to their water solubility limit, and for very toxic or volatile substances [104]. Other confounding factors include the influence of constituents other than pollutants, and potential problems with pH or oxygen content (particularly in cases where the native sample is used) [29]. The test set-up applied can also affect results, with EC50 values lower in tests conducted in glass vials than those run in 96-well polystyrene plates, likely due to chemical sorption to the plastic [47]. *Relevance:* High. No evidence of MoA, however tests with fish are highly relevant for the assessment of water samples. While testing using embryos can be overprotective for some organisms, fish embryo toxicity and acute toxicity of adult fish are closely related (r^2=0.95) [106]. *Availability of trigger values:* A trigger value of 0.183 mg Bisphenol A equivalent /L [39] has been proposed in the literature for zebrafish toxicity. *Bioassay category:* Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The same limitations as discussed for the FET are anticipated.

3.4.3. Zebrafish Toxarray

Description: This assay combines a zebrafish toxicity assay and a quantitative PCR test, with an exposure time of 5 days. The Toxarray allows the screening of up to 42 target genes which covers multiple toxicity pathways. MoA covered general stress, apoptosis, oxidative stress, DNA damage, metabolism, metabolism phase II/III, endocrine disruption, adipogenesis, insulin signalling and neurotoxicity. The target genes selected can be adapted in order to select the relevant mode(s) of action of interest.
Validation maturity: The assay is not available commercially and not standardised to any ISO, OECD or other guidelines. The test has not been applied in regulatory context. It has undergone initial validation and optimisation using model substances [107]. *Validation maturity:* Low (research lab) [27].
Performance: *Sensitivity*: The sensitivity is generally high [107], but the same lack of sensitivity to PAHs and refinery effluents as seen with the other zebrafish assays may still be present. *Potency:* The test shows potency of effect; effect concentrations are derived.

Use: *Use in environmental samples:* The use of the assay in environmental samples is occasional [27]. The assay is only applied to freshwater samples when using zebrafish, i.e. a freshwater species. It has been used for drinking water, surface water and wastewater samples [107]. *Use with refinery effluent constituents:* No information could be found. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises relatively small sample volumes is applicable to passive sample extracts [27]. *Time to run:* Results within a week; ≤120-hour exposure. *Cost per sample:* Unknown. No survey respondents conducted this survey.
Result interpretation: *Confounding factors:* No information could be found. *Relevance:* High. Tests with fish are highly relevant for the assessment of water samples, and while this test does not provide information on an adverse effect, it provides supporting interpretive information for MoA which the other *in vivo* assays do not [29]. Sometimes where no visual malformation can be seen, strong gene expression changes occur, giving new toxic insights [107]. Relevance to the constituents of refinery effluents is not known, although PAHs are known to be genotoxic and PAHs, dioxins, dioxin-like chemicals and metals cause oxidative stress which may be picked up in the Toxarray depending on the target genes chosen [15] *Availability of trigger values:* A trigger value of 0.183 mg Bisphenol A equivalent /L [39] has been proposed in the literature for zebrafish toxicity. *Bioassay category*: Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: No application limitations have been found in the literature or were raised during the survey.

3.5. OTHER

Tests with other taxonomic groups were also considered but not evaluated in as much depth, owing to their reduced applicability to the testing of refinery effluents or the existence of more developed counterparts. These included assays with marine invertebrates, amphibians and aquatic plants.

Freshwater, sediment and soil invertebrates have been evaluated above in the form of tests with daphnids, amphipods and nematodes, respectively. Here we consider the use of marine invertebrates for those refineries that release to seawater. One of the most commonly used marine invertebrates in toxicity testing is the mysid shrimp (*Americamysis bahia*), owing to its short life cycle, ease of maintenance/cultivation and small size [108]. Mysid shrimps have been used to assess the toxicity of crude oil-contaminated water samples, proving more sensitive to oil than inland silversides (*Menidia beryllina*) [40]. However, the responsiveness of mysid shrimps to oil constituents is generally minimal [40], reducing their

relevance to the monitoring of refinery effluents. Another marine alternative for tests with aquatic invertebrates is *Acartia*, marine calanoid copepods. There are acute (24-48 hour exposures) as well as semi-chronic (7 day exposures, static renewal) test set-ups described in the literature that monitor immobilisation of the nauplius stage of *A. tonsa* [109]. *A. tonsa* has been used for acute marine toxicity assessments of refinery effluent, but it was less sensitive than algae [53]. Rotifers are another alternative; the acute (24–48 hour) toxicity test with the marine rotifer *Branchionus plictalis* is commercially available, standardised to ISO 19820, well validated and used in routine monitoring. Further, *B. plictalis* is a euryhaline species, tolerating salinities of 5 to 35 parts per thousand [40]. The use of rotifers to test refinery effluents has not yet been demonstrated in the scientific literature. These marine invertebrate assays may present attractive marine alternatives for the *Daphnia* tests.

A further freshwater alternative to the *Daphnia* tests is the use of *Ceriodaphnia dubia*, another member of the *Daphniidae* family. Traditionally *D. magna* and *D, pulex* were considered the standard test organisms for acute and chronic freshwater invertebrate tests under OECD, ISO and ATSM guidelines. However *C. dubia* has since been mentioned in the OECD 211 *D. magna* reproduction test guideline and included in a separate ISO 20665:2008 guideline for determination of chronic toxicity to *C. dubia*. The required test volumes for *C. dubia* are reduced owing to their smaller body size compared to *D. magna*. Their smaller size also means a shorter time to reach reproductive maturity, and hence a shorter experiment duration of 7 days can be used for chronic reproductive tests. However, there have previously been concerns regarding the suitability of this shorter exposure duration for hydrophobic substances [110]. Nonetheless, chronic *C. dubia* assays have been successfully used to assess toxicity of oil refinery effluents [111] and PAH-containing refinery wastewater [112]. Further, both the acute and chronic sensitivities of *C. dubia* are comparable to *D. magna* [110], so should be viewed as viable and cost-effective alternatives to the *D. magna* assays.

Amphibians, while not fully aquatic organisms, may also be exposed to refinery effluents intermittently and hence were considered as part of this work. Amphibians such as the African clawed frog (*Xenopus laevis)* are often tested at their tadpole life stage to elucidate changes in gene and/or enzyme activity [47]. An example is the *X. laevis* metamorphosis assay (XEMA). *X. laevis* tadpoles are sensitive [29] and have been used for the testing of weathered bunker and unweathered refinery crude oil water accommodated fractions (WAFs) [113] and crude oil and its fractions [114]. However, as the use of vertebrates is highly discouraged for effluent assessment in the EU, and vertebrate tests are already covered in the evaluation by well-validated fish embryo tests, amphibians were not considered further.

Aquatic plants are also a relevant group when considering the discharge of refinery effluents in freshwater. A routinely used test organism in toxicity tests of aquatic plants is common duckweed (*Lemna minor*) or gibbous duckweed (*Lemna gibba*), which float on the surface of relatively stagnant freshwater such as slow-moving streams, ponds and lakes [115]. The test is standardised to ISO 20079 and has been applied in a regulatory context under the IED. ISO 20079 is listed as part of the BATC for the CWW BREF. Duckweed tests typically last 7 days and measure fronds or biomass. The test is well validated and available commercially [27]. Tests with *L. gibba* have also successfully shown growth inhibition of oil refinery effluents [98]. However, the test requires a relatively large amount of time, needs sample volumes of at least 100 mL per test vessel, and, while not a critical factor, is not applicable to passive sampler extracts, [116]. Primary producers are already covered extensively by the algal tests. An alternative approach that may be more attractive for routine monitoring use has recently been developed, discussed below.

3.6. NOVEL DEVELOPMENTS

The drive to reduce vertebrate animal testing and increase the high-throughput potential of *in vivo* toxicity testing under monitoring campaigns has seen the recent emergence of several novel techniques.

One of these include the use of a low-cost bacterial-decorated bioluminescent nanopaper (BLN). It is a rapid (15-minute) growth inhibition assay measuring bioluminescent *A. fischeri* immobilised on the nanopaper, a nanocellulose-based membrane device [117]. The assay is simple, fast, non-invasive, and does not employ vertebrate test species. Further, the BLN maintains its function during freezing, increasing its applicability for use in laboratories. The sensitivity has been reported as comparable to that of the MicroTox® test, albeit slightly lower. The main hinderance preventing its use in routine monitoring is inter-assay variability, thought to be caused by differences in assay conditions and/or *A. fischeri* batches [117].

As mentioned in the above section, tests with duckweed are commonly applied in the assessment of surface- and wastewaters. However, these tests typically last 7 days and require volumes of at least 100 mL per test vessel. To increase the usability of duckweed tests, a *L. minor* root-regrowth test has recently undergone interlaboratory validation [115]. The test works by removing existing roots and measuring new roots that subsequently develop over the course of 3 days. In comparison to standard duckweed tests, this novel assay can be conducted within 72 hours in 24-well cell plates and requires a smaller volume of test water samples (3 mL). Repeatability was 21.3% and reproducibility was 18.6% for wastewater, well within the generally accepted levels of <30% to 40%. The results of the interlaboratory validation demonstrate its validity and reliability [115].

Recent years have also seen the advancement of methods used to elucidate the modes of action behind the apical effects typically observed in *in vivo* testing, such as the *Daphnia* IQ™ toxicity test and zebrafish Toxarray. Such a method has recently been proposed for mussels. The 'STressResponse Microarray' (STREM) is a targeted low-density oligonucleotide microarray comprises probes covering 465 stress response target genes [118]. It combines *in vivo* and *in vitro*; mussels are exposed *in vivo* before the STREM is used on target tissues, such as gills. The microarray has been tested with B(a)P, providing mechanistic clues about other biological processes, e.g. immune response and mitochondrial activity.

While not sufficiently validated for use in a regulatory context thus far, these recent developments present promising future additions to the battery of *in vivo* tests that can be used for routine monitoring.

4. EVALUATION OF *IN VITRO* TESTS

Eighteen *in vitro* tests have been evaluated below and summarised in an overview table (Appendix), out of approximately 21 tests initially identified. The tests cover a range of cell lines and test organisms, which may affect their relevance to the environment (e.g. fish cell lines) versus human health (e.g. human or rodent cell lines). The selected tests also cover cytotoxicity, genotoxicity, oxidative stress, metabolism and endocrine disruption as modes of actions, as these were considered the most relevant to the constituents of refinery effluents.

Neurotoxicity (e.g. via an acetylcholinesterase inhibition assay) was excluded from the evaluation as only the metals were identified as potential neurotoxicants, however metal toxicity also acts via oxidative stress, cytotoxicity and genotoxicity which are covered in a number of the evaluated assays. Immunotoxicity (e.g. via a KappaB assay) was similarly excluded as the only constituents of refinery effluents expected to be immunotoxic would be the PAHs, however their immunotoxicity stems from activation of the aryl hydrocarbon receptor (AhR), which is already included in the evaluation below. Metabolism was covered in the evaluation, but with a focus on the aryl hydrocarbon receptor (AhR) pathway as opposed to the pregnane X receptor (PXR) pathway. The PXR is activated by vitamins, steroids, antibiotics, and other pharmaceutical compounds, rather than typical refinery effluent constituents [34]. Further, the PXR pathway is less linked to adverse effects than the AhR pathway which can contribute to developmental, immunotoxic, reprotoxic, and carcinogenic effects [34]. The peroxisome proliferator-activated receptor (PPAR) assays, e.g. PPARγ CALUX, were similarly not taken forward in the evaluation of metabolism bioassays. For endocrine disruption modes of action, assays investigating estrogen-, androgen- and thyroid-related endpoints were selected over assays looking at the glucocorticoid receptor (GR), e.g. GR CALUX, and the progesterone receptor (PR), e.g. anti-PR CALUX. Refinery effluents would not be expected to produce responses in GR or PR assays, as it is predominantly natural and synthetic pharmaceutical hormones (e.g. prednisolone, prednisone, dexamethasone and triamcinolone acetonide) that act as glucocorticoids or progesterones [15, 34].

4.1. CYTOTOXICITY

4.1.1. Cytotoxicity in a Rainbow Trout cell line (RT Gill-W1)

Description: This is a fish cell line assay whereby rainbow trout gill cells (RT Gill-W1) are exposed to an aqueous sample in 24-well tissue culture plates for 24 hours. Cytotoxicity is assessed in 3 ways using different fluorescent indicator dyes. Cell plasma membrane integrity is measured using the esterase substrate 5-carboxyfluorescein diacetate acetoxymethyl ester (CFDA-AM), which is converted to a fluorescent product by esterases in intact membranes. Cell metabolism is measured using reduction of the substrate Alamar Blue™ to the red fluorescent substance resorufin. Lysosomal activity is measured by staining with neutral red; the less dye retained, the more permeable and damaged the membrane is.
Validation maturity: RT gill-W1 is available commercially and is standardised to ISO 21115. It has recently undergone a round-robin study (n= 6 test chemicals, n= 6 laboratories) [119]. *Validation maturity:* Medium (expert lab) [27].

Performance:
Sensitivity: The sensitivity is variable, as it depends on the substance and target site. The assay is not sensitive to some PAHs due to lack of certain receptors and metabolic capabilities [120].

Potency: The test shows potency of effect; an EC50 value is derived.

A round-robin study demonstrated the robustness of the assay; it was easy to establish, and the repeatability (intralaboratory variability) and reproducibility (interlaboratory variability) were well within the range of those previously reported for other fish bioassay-focused round-robin analyses, such as the FET [119]. Further, close correlation between cytotoxicity in this test and acute fish toxicity has been demonstrated [120]. The predictability of the test has been validated by using an independent set of previously untested 38 fragrance chemicals; a strong correlation between *in vivo* LC50 values (fish mortality) and *in vitro* EC50 values (cell viability) was observed [121].

Use:
Use in environmental samples: The use of the assay in environmental samples is occasional [15]. It has been used for effluent testing of a nitrogen-producing facility [122], paper mill [123] and mining facility [124], as well as testing of oil sands process-affected water [125, 126], a lab-scale wastewater treatment plant [127], Norwegian coastal sediments [128] and lagoon sediments [129].

Use with refinery effluent constituents: The assay has been demonstrated to be responsive to 2- and 3- ring PAHs, e.g. naphthalene [130, 131]. Creosote has also been tested with this assay, with all the aromatic hydrocarbons in creosote contributing to the observed cytotoxicity [132]. As described under 'result interpretation', other PAHs can be responsive in this assay after metabolic activation, although the enzymes required for this may not always be present [120]. The RTgill-W1 assay been used to assess cytotoxicity of metals (copper, zinc, cadmium, iron and nickel); the most toxic metal was copper, followed by zinc, cadmium, iron and nickel, with the same rank order when using RTL-W1 liver cell, suggesting no particular metal sensitivity between lines [133]. The toxicity of copper and zinc was strongly dependent on the exposure medium used [133].

Use with refinery effluents/produced water: It has not been reported whether the assay is responsive in tests with refinery effluents or produced water, although Schirmer *et al.* [132] reported the use of fish cell lines for the assessment of petroleum refinery effluent.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes and is applicable to passive sample extracts.

Time to run: Results within a week; 24-hour exposure.

Sample processing: Whole water samples can be evaluated without extraction or concentration steps.

Cost per sample: Unknown. No survey respondents conducted this survey.

Result interpretation:
Confounding factors: The water solubility and lipophilicity of the substances present in the effluent can affect their results in the assay, as the assay centres around lipid-containing cell membranes. Further, metabolic activation may be required for some substances that are indirectly cytotoxic. Tanneberger *et al.* [120], for example, found good correlation between the *in vitro* RT gill-W1 assay and the *in vivo* OECD 203 test effect concentrations when testing 35 organic chemicals, with the notable exception allyl alcohol, which needs to be metabolically activated via an oxidation step. Tanneberger *et al.* [120] also found that RT gill-W1 can be exposed in serum-free conditions, thus eliminating interference of serum proteins with toxicants.

Dayeh et al. [122] showed that ammonia is toxic to RT Gill-W1 cells and rainbow trout via increased vacuolisation, so this may also confound results. Further, the use of cells from a fish species means that osmotic stress may affect the results, if not adjusted for (see below). However, increasing salt concentrations may reduce the sensitivity of the assay, as the bioavailability of the contaminants may be altered; this is of particular importance for samples containing metals, which may complex with anions added to the exposure medium [134]. Bioavailability may also be affected by organics when complexing at varying water hardness [135].

Relevance: Medium. The effects measured in the test are highly relevant to the environment. Fish are present in marine, fresh and brackish waters that refinery effluents and produced water are released to. Fish gill epithelial cells are the first organ exposed to toxicants and are of high physiological importance to fish [135]. Any cell damage and subsequent impairment/loss of gill function can lead to fish death. However, the assay is deemed less relevant in the context of refinery effluents. It is a non-specific form of toxicity with no specific MoA, that requires metabolic activation to be responsive to PAHs, although it is responsive to metals.

Availability of trigger values: A trigger value of 0.05 TU has been proposed in the literature [38].

Bioassay category: Not stated but assume Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
The osmotic tolerance of the RT Gill-W1 cells has been reported as 150 to 450 mOsm/kg [135]. Thus, in order to prevent osmotic stress to the cells, it may be necessary to add salts to the samples to adjust the osmolarity of the exposure medium [136].

4.2. GENOTOXICITY

4.2.1. *umuC* assay

Description:
This is a bacterial assay using *Salmonella typhimurium*. It is used as an indirect measure of DNA damage (or genotoxicity) by the inducible SOS response. The assay indirectly measures the activation and the SOS repair response to DNA damage, or interruption of DNA replication, via the induction of the enzyme ß-galactosidase. The test is conducted with and without metabolic activation, using liver S9. It is a rapid test, which responds to chemicals with various mutational mechanisms, including cross linking events.

Validation maturity:
The *umuC* assay is available commercially and is standardised to ISO 13829. The test has been applied in regulatory context, for wastewater from the chemical industry [30] and is validated to water samples [34]. It is the most mature of the genotoxicity assays.

Validation maturity: High (routine lab) [27].

Performance:
Sensitivity: The sensitivity is generally very high, and higher than the Ames assay, but sensitivity can be variable, as it depends on both the substance (including metabolic activation) and strain of *Salmonella* used.

Potency: The test does not show potency of effect; the assay gives a positive or negative result based on induction ratio.

The assay is highly reproducible [137] and highly specific [29].

Use:
Use in environmental samples: The use of the assay in environmental samples is very common, both with and without metabolic activation [34]. It has been used for:

- Wastewater [41, 67-69, 138-143];
- Drinking water [68, 139, 144, 145];
- Tap water [145];
- Surface water [68, 139, 141, 144, 146-151];
- Recycled water [67, 68, 139, 140, 142, 144].

Use with refinery effluent constituents: The assay has been demonstrated to be responsive to metabolically activated B(a)P [147] and other metabolically activated PAHs [150], as well as to heavy fuel oil residues in fresh and estuarine surface waters [152]. Further, metal salts were positive both without ($K_2Cr_2O_7$, $RhCl_3$, $IrCl_4$, and $MgCl_2$) and with ($CuCl_2$, VCl_3, CuCl, $RhCl_3$, $K_2Cr_2O_7$, and $IrCl_4$) metabolic activation [153], therefore metal ions in refinery effluents may contribute to genotoxicity in this assay.

Use with refinery effluents/produced water: It has not been reported whether the assay is responsive in tests with refinery effluents or produced water.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (200µL typical final volume, [29]) and is applicable to passive sample extracts [45, 150, 152]. A miniaturised version is available.

Time to run: Results within a day; 2-hour exposure.

Sample processing: Minimal advance preparation required. Sample enrichment may be required. Metabolic activation may be required.

Cost per sample: <100 € [34].

Result interpretation:
Confounding factors: The species used can affect the sensitivity of results for different types of water; *Salmonella* sp. is halo-sensitive and hence survives poorly in marine waters, thus water samples from refineries discharging into marine water may not perform well in this assay [40].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic. *Availability of trigger values:* Trigger values of 0.005 genoTU [38] and 1 genoTU [41] have been proposed in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The assay is based on genetically modified organisms, therefore permission is required and a safety level 1 laboratory is also needed [29, 34]. A cytotoxicity control may be required.

4.2.2. Ames assay

Description: This is a bacterial reverse mutation assay using *Salmonella typhimurium* strains, including TA98, TA100, TA1535, TA97 and 98NR. The strains are exposed to toxicants in a microplate system for 48 hours with and without metabolic activation, using liver S9. The assay measures the ability of toxicants to mutate a histidine-dependent strain of the *Salmonella* to grow on a histidine-deficient substrate. Each mutation can be reverted only by a specific transition or transversion, meaning that all possible base pair changes can be detected and identified.
Validation maturity: The Ames assay is available commercially and is standardised to ISO 16240. The assay is validated to water samples [34]. *Validation maturity:* High (routine lab) [27, 34].
Performance: *Sensitivity*: The sensitivity is variable, as it depends on the *Salmonella* strain used. It is less sensitive than the *umuC* assay which can detect mutagenic substances which do not illicit responses in the Ames assay [35]. *Potency:* The assay does not show potency of effect; it gives a positive or negative result. A positive result is when there is doubling of the number of revertant colonies at any concentration of the test sample compared to the negative control. The assay is specific to different types of gene mutations, and thus different strains can be used to provide information on the exact mutagenic mode of action. For example, *S. Typhimurium* strain TA100, TA102 and TA1535 can detect base-exchange mutations, while strains TA98 and TA1537 can detect frame shift mutagens [155].
Use: *Use in environmental samples:* The use of the assay in environmental samples is very common, both with and without metabolic activation [34]. Dating back to the 1970s, it has been used for: • Wastewater [59, 67, 68]; • Drinking water [68]; • Surface water [35, 68, 150, 156]; • Recycled water [157]; • Marine water [158].

Use with refinery effluent constituents: The assay has been demonstrated to be responsive to metabolically activated B(a)P [47, 147, 159] and to suspended particulate matter collected downstream of petrochemical industries [160]. It has also been responsive to class-fractionated oils extracted from creosote- and petroleum-contaminated soils [50].

Use with refinery effluents/produced water: The assay has been used to assess toxicity of petroleum refinery effluents [161] and the suspended particulate matter within these [162, 163], as well as river samples close to petrochemical industries [164, 165] and wastewater from petroleum oil plants [59, 166].

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [150, 167].

Time to run: Results within a week; 48-hour exposure.

Sample processing: Metabolic activation may be required. Sample enrichment may be required. The extraction method is important for this test [13].

Cost per sample: The literature reports a cost of 100-1,000 € [34]. The results of the survey indicate the assay is likely to be 3,000 to under 10,000 € per sample, with 100% of the laboratories that conduct the Ames assay (n=2) reporting this cost.

Result interpretation:
Confounding factors: Microbial contamination of the sample may give a positive result, and inhibitory effects of the sample may mask mutagenicity to give a negative effect [29, 168]. It has also been noted that unexpected high reversion rates can occur in negative controls [29]. The species used can affect the sensitivity of results for different types of water; *Salmonella* sp. is halo-sensitive and hence survives poorly in marine waters, thus refineries discharging to marine water may not perform well in this assay [40]. The strain of Salmonella used can also influence results, for example, strain TA98 gave a positive result in environmental water samples while strains TA100 and YG1041 were negative [150]. The different strains can detect different mutagenic modes of action; for wastewater (petroleum refinery wastewater mixed with domestic sewage), it is predominantly frameshift and missense mutations that are detected, and hence the strain used will have an important bearing on the results [166]. The extraction method can also influence results [13].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic.

Availability of trigger values: A trigger value of 0.005 genoTU has been proposed in the literature [38].

Bioassay category: Not stated but assume Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
It may be required to use at least five different strains of bacteria, with and without metabolic activation, in order to avoid false negative results, which can increase the workload [169]. The assay is already quite laborious and does not lend itself to high-throughput [34]. Compared to the Ames II/fluctuation assay, it uses more test chemical, more labour time, more S9 and more plasticware [170]. A cytotoxicity control may be required. Further, the assay is based on genetically modified organisms, therefore permission is required [34].

4.2.3. Ames II/Ames fluctuation assay

Description:
This is a bacterial reverse mutation assay using *Salmonella typhimurium* strains, including TA7001-TA7006. The strains are exposed to toxicants in a microplate system for 48 hours with and without metabolic activation, using liver S9. The assay measures the ability of toxicants to mutate a histidine-dependent strain of the *Salmonella* to grow on a histidine-deficient substrate. Each mutation can be reverted only by a specific transition or transversion, meaning that all possible base pair changes can be detected and identified.

Validation maturity:
The Ames II/fluctuation assay is available commercially and is standardised to ISO 11350. It has undergone a round-robin study [171] and is validated to water samples [34].

Validation maturity: High (routine lab) [27, 34].

Performance:
Sensitivity: The sensitivity is variable, as it depends on the *Salmonella* strain used. It is less sensitive than the *umuC* assay which can detect mutagenic substances which do not illicit responses in the Ames assay [35].

Potency: The assay does not show potency of effect; it gives a positive or negative result. A positive result is when there is doubling of the number of revertant colonies at any concentration of the test sample compared to the negative control.

It is highly specific [171] and much more capable of high-throughput samples than the classic Ames test as it can be conducted in 96-well plates [34]. The assay is specific to different types of gene mutations, and thus different strains can be used to provide information on the exact mutagenic mode of action. For example, *S. Typhimurium* strain TA100, TA102 and TA1535 can detect base-exchange mutations, while strains TA98 and TA1537 can detect frame shift mutagens [155]. The Ames II/fluctuation assay and classic Ames assay perform similarly; 84% agreement identifying mutagens and non-mutagens (equivalent to the intra- and interlaboratory reproducibility of 87% for the Ames assay) and similarly predict rodent carcinogenicity [170].

Use:
Use in environmental samples: The use of the assay in environmental samples is very common, both with and without metabolic activation [34]. It has been used for groundwater [172], surface waters [173], and effluent [173, 174].

Use with refinery effluent constituents: Total petroleum hydrocarbon, benzene, toluene, phenol index, PAHs, cadmium, mercury, nickel, lead, and vanadium contents analysed in refinery effluent samples were believed to be responsible for the observed toxicity in an Ames fluctuation assay [175]. Metabolically activated B(a)P has also produced a positive result in this assay [47].

Use with refinery effluents/produced water: The Ames fluctuation test was responsive to petroleum refinery effluent [175].

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes and is applicable to passive sample extracts [156].

Time to run: Results within a week; 48-hour exposure.

Sample processing: Metabolic activation may be required, although it requires less S9 than the classic Ames assay [170]. Sample enrichment is often required. The extraction method is important for this test [13].

Cost per sample: 100-1,000 € [34] or <200 € [28].

Result interpretation:
Confounding factors: As with the Ames and *umuC* assays, the strain of bacteria used can significantly influence whether a positive or negative result will be observed. For example, in a study with refinery wastewater and river water, the TA100 strain was more responsive than the TA98 strain [175]. The different strains can detect different mutagenic modes of action; for wastewater (petroleum refinery wastewater mixed with domestic sewage), it is predominantly frameshift and missense mutations that are detected, and hence the strain used will have an important bearing on the results [166].
The extraction method can also influence results [13].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic.

Availability of trigger values: A trigger value of 0.005 genoTU has been proposed in the literature [38].

Bioassay category: Not stated but assume Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
The assay is based on genetically modified organisms, therefore permission is required [34]. Its application is much easier than the classic Ames assay, e.g. less labour/time, less test chemical, less S9, less plasticware, and can be automated [170]. A cytotoxicity control may be required.

4.2.4. Micronucleus assay

Description:
The micronucleus assay is used to determine if a compound is genotoxic by evaluating the presence of micronuclei. The assay detects the frequency of formation of the micronuclei and other nuclear abnormalities within cells. Micronuclei are caused by DNA damage by the action of a genotoxic agent, and so the presence of an increased number indicates the presence of a genotoxic compound. The assay often uses the Chinese hamster lung fibroblast cell line V79, but multiple cell types can be used to make it more relevant to the aquatic environment.

Validation maturity:
The micronucleus assay is available commercially and is standardised to ISO 21427. It is validated to water samples [34]. It has undergone a round-robin study [176].

Validation maturity: High (routine lab) [27, 34].

Performance:
Sensitivity: The assay is highly sensitive [34].

Potency: The test does not show potency of effect; although it is possible to get the % micronucleated erythrocytes, the assay ultimately gives a yes/no result.

The assay is considered more robust than the Comet assay and has high specificity [29, 34]. It also has a positive correlation with mammalian tests [30].

Use:
Use in environmental samples: The use in environmental samples is common [27]. It has been used for surface water [148, 151, 177, 178], wastewater [176] and refinery effluents (see below).

Use with refinery effluent constituents: It has been demonstrated that petroleum hydrocarbons in refinery waste and petroleum-polluted water can cause micronuclei formation [179, 180]. Total petroleum hydrocarbon, benzene, toluene, phenol index, PAHs, cadmium, mercury, nickel, lead, and vanadium contents analysed in refinery effluent samples were believed to be responsible for the observed toxicity in a micronucleus assay [175]. Further, B(a)P and benzo(b)fluoranthene have both been responsive in the assay [28, 47].

Use with refinery effluents/produced water: The assay has been responsive to petroleum refinery effluents [175, 181, 182]. In a study by Hara and Marin-Morales [183], a response in petroleum refinery effluent was only detected after the first physico-chemical treatment- not in the source water, post-biological treatment effluent, final effluent or downstream of the discharge site.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts.

Time to run: Results within a week; the exact duration depends on sample incubation and purpose of the study.

Sample processing: Metabolic activation may be required.

Cost per sample: The literature reports a cost of 100-1,000 € [34] or <200 € [28]. The results of the survey indicate the assay is likely to be above 10,000 € per sample, with 100% of the laboratories that conduct the Micronucleus assay (n=2) reporting this cost.

Result interpretation:
Confounding factors: The subjective evaluation of results (frequency of stained micronucleated cells) may produce variability in results. Further, the cell line used can generate different responses; in one study, dibenzofuran and other PAHs tested positive in a fish cell line (RTL-W1), but negative in mammalian systems [184].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. Brack et al. [29] assigned the micronucleus assay high toxicological relevance, while Carere *et al.* [13] ranked it as medium-high relevance, as it responds to clastogenic substances that in the long term can give rise to negative effects. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic.

Availability of trigger values: A trigger value of 0.005 genoTU has been proposed in the literature [38].

Bioassay category: Not stated but assume Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
The assay is not very high-throughput, is labour intensive and time-consuming, due to the visual assessment of results. The assay is based on genetically modified organisms, therefore permission is required [34]. A cytotoxicity control may be required. Further, it has been reported that samples should be analysed within 3 days [28].

4.2.5. Comet assay

Description:
The comet assay (single-cell gel electrophoresis) is a simple way to measure single or double DNA strand breaks within eukaryotic cells, including human, other mammalian and fish cells in order to determine if a compound is genotoxic. This is transient genetic damage that does not permanently change or alter the DNA. Cells are embedded in agarose on a microscope slide and are lysed with detergents to form nucleoids which consist of supercoiled loops of DNA. These loops of DNA are linked to the nuclear matrix. Electrophoresis at a high pH is followed by DNA staining, which results in structures resembling comets, which can be observed by fluorescence microscopy. The colour intensity of the comet tail versus the head signifies the number of DNA strand breaks.

Validation maturity:
The comet assay is available commercially. It is not ISO standardised, but can be conducted according to OECD 489. It is in the process of being validated to water samples [34], but has been used to assess surface water and sediments in the Netherlands [13].

Validation maturity: High (routine lab) [34].

Performance:
Sensitivity: The assay is highly sensitive [13, 29].

Potency: The test does not show potency of effect; it gives a positive or negative result, based on the number of damaged cells (damage frequency).

The assay has high specificity [34]. The comet assay has been shown to be more sensitive than the Ames test for genotoxicity detection in drinking water samples [185].

Use:
Use in environmental samples: The use in environmental samples is common [27]. It has been used in surface waters [147, 178, 186, 187], groundwater [172], drinking water [185] and effluent [188]. The assay has been successfully introduced to assess sediments and surface water in The Netherlands [13].

Use with refinery effluent constituents: The assay has been responsive to suspended particulate matter collected downstream of petrochemical industries [160] and oil-containing drill cuttings [189].

Use with refinery effluents/produced water: No information could be found, other than Vincent-Hubert *et al.* (2012).

Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts. *Time to run:* Results within a week; the exact duration depends on sample incubation and purpose of the study, but is often a 24 hour exposure. *Sample processing:* Metabolic activation may be required. *Cost per sample:* <100 € [34].
Result interpretation: *Confounding factors:* The cell line used can affect the results; a mammalian cell line (monkey kidney) and fish cell line (RT gill-W1) gave different results in a comet assay assessing coal mining effluent [188]. Brack et al. [29] further remarks that cytotoxicity can lead to strand break formation, and that a distinction between apoptotic and necrotic is recommended. *Relevance:* Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. Brack et al. [29] assigned the Comet assay low toxicological relevance, while Carere *et al.* [13] ranked it as medium-high relevance, owing to the impairment of DNA, and thus cellular function and effects at the individual level. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic. *Availability of trigger values:* A trigger value of 0.005 genoTU has been proposed in the literature [38]. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The assay is not very high-throughput [34]. However, it is not based on genetically modified organisms, unlike many of the other genotoxicity assays, so may be more available to more laboratories [34].

4.2.6. hGADD45 activation assay

Description: This assay is based on the expression of the human GADD45a (hGADD45) gene, which has a role in the genotoxicity stress response. It is a Green Fluorescent Protein (GFP) assay. Different versions of the assay are based on yeast cultures of *Saccharomyces cerevisiae* or human cell lines. The test is conducted with and without metabolic activation using liver S9.
Validation maturity: The hGADD45 activation assay is available commercially. It is not ISO standardised; however the human cell line version of the assay is currently undergoing standardisation. The assay is not well validated for water samples [34]. *Validation maturity:* Medium (expert lab) [27, 34].

Performance: *Sensitivity*: The assay is highly sensitivity [34, 190]. *Potency:* The test does not show potency of effect; it gives a positive or negative result based on expression of the *hGADD45a* gene [190]. It is high-throughput and highly specific [34, 190].
Use: *Use in environmental samples:* The use of the human cell line version in environmental samples is occasional [34]. The yeast version has been used for industrial effluents [191]. *Use with refinery effluent constituents:* It has been used with metabolically activated B(a)P [192, 193]. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes so should be applicable to passive sample extracts. *Time to run:* Results within a week. *Sample processing:* Metabolic activation may be required. *Cost per sample:* Unknown. No survey respondents conducted this survey.
Result interpretation: *Confounding factors:* Bacterial contamination, particulate matter and cytotoxic effects may confound results [191]. *Relevance:* Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk. The *GADD45a* gene is an important contributor to DNA repair, apoptosis and cell-cycle regulation; it is one of the most robustly induced genes by genotoxic substances [193]. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic. *Availability of trigger values:* A trigger value of 0.005 genoTU has been proposed in the literature [38]. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The assay is based on genetically modified organisms, therefore permission is required [34].

4.2.7. p53-pathway activation assay (e.g. p53 CALUX®)

Description: This is a human cell line (U2OS) derived pathway selective reporter gene assay using an ELISA format (96-well plate), in which a firefly luciferase gene has been coupled to p53 Responsive Elements. p53 is detected by a primary antibody that recognises p53 only when the protein is activated and bound to its target DNA. A secondary antibody provides the colorimetric readout at OD 450 nm. The test is conducted with and without metabolic activation using liver S9.
Validation maturity: The p53-pathway activation assay is available commercially. It is not ISO standardised (although it is undergoing standardisation) and has not undergone any formal validation to water samples [34]. *Validation maturity:* Medium (expert lab) [27, 34].
Performance: *Sensitivity*: The assay is highly sensitive [29, 34, 194]. However, using S9 can increase the number of false positives, decreasing sensitivity slightly (e.g. from 95% to 82%) [194]. A false positive has been reported for phenanthrene (without metabolic activation) [194]. *Potency:* The test does not show potency of effect; it gives a positive or negative result, based on % induction with or without S9. It has high specificity and is high-throughput [29, 34, 194].
Use: *Use in environmental samples:* The use of the assay in environmental samples is occasional [34]. p53 assays have been used for surface water [68, 70, 105, 195, 196], drinking water [68, 195, 197], recycled water [68], and wastewater [68, 70, 198]. *Use with refinery effluent constituents:* It has been reported that metabolically activated B(a)P is responsive to this assay [194]. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes (40-200µL typical final volume [29]) so should be applicable to passive sample extracts. *Time to run:* Results within a week. *Sample processing:* Metabolic activation may be required. *Cost per sample:* 100-1,000 € [34].
Result interpretation: *Confounding factors:* Chemicals that activate p53 are often cytotoxic, which can mask the effects, particularly when applied to surface water samples [15, 70]. p53 activation was masked by cytotoxicity in most disinfected water samples [199]. Cross over luminescence measurement has been reported [29]. Further, metabolic activation may give false positive results [194].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, and non-carcinogens have also been known to induce positive results in genotoxicity assays [154]. However, a positive result is generally highly indicative of carcinogenesis risk; the p53 gene is the "tumour suppressor gene" which regulates cell cycle arrest and apoptosis. The use of a human cell line makes it less relevant to the aquatic environment. The assay is relevant to the constituents of refinery effluents, as PAHs and some metals are genotoxic/mutagenic. *Availability of trigger values:* Trigger values of 0.005 genoTU [38] and 0.005 TU [54] have been proposed in the literature. *Bioassay category:* Not stated but assume Category 2 [39]. *Chemical analysis:* The test cannot replace chemical analysis.
Possible application limitations: The assay is based on genetically modified organisms, therefore permission and a commercial licence is required [29, 34]. A cytotoxicity control may be required.

4.3. OXIDATIVE STRESS

4.3.1. AREc32 activation assay

Description: This assay uses the Nrf2-Antioxidant Response (ARE) pathway, which is responsive to chemicals which induce oxidative stress. This oxidative stress response pathway is measured in a human breast cancer cell line, MCF7. Nrf2 induction is proportionate to the amount of luciferase produced by the cells and can be identified by a bioluminescence fluorimeter. Tert-Butylhydroquinone (tBHQ) is used as a positive control for the test.
Validation maturity: The AREc32 activation assay is available commercially. It is not ISO standardised, although it is currently undergoing standardisation. It is validated to water samples and is more mature than other oxidative stress assays [34]. *Validation maturity*: Medium (expert lab) [27, 34].
Performance: *Sensitivity*: The assay is highly sensitive [29, 34]. *Potency:* The test shows potency of effect; effect concentrations are derived. The test is high-throughput, has a high responsiveness and has high specificity [29, 34]. It is relevant to water quality assessment [200].
Use: *Use in environmental samples:* The use of the assay in environmental samples is common [27]. It has been used for: • Surface water [144, 146, 195, 200-206]; • Recycled water [67, 142, 200, 205]; • Wastewater effluent [46, 67, 142, 200, 203, 205, 207]; • Drinking water [144, 145, 195, 197, 200, 205, 208].

Use with refinery effluent constituents: B(a)P and benzo(b)fluoranthene have been responsive to the assay [47]. Further, PAHs and metal/loid(s) activated the Nrf2 antioxidant pathway in ARE reporter-HepG2 cells, with cadmium being the most potent inducer followed by B(a)P and arsenic [209]. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes (40-200µL typical final volume [29]) so should be applicable to passive sample extracts. *Time to run:* Results within a week; up to 24-hour exposure. *Sample processing:* No sample processing considerations identified. *Cost per sample:* <100 € [34].
Result interpretation: *Confounding factors:* Cytotoxic effects [47, 200] and cross over luminescence measurement have been reported [29]. *Relevance:* Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, however, oxidative stress response markers give an early indication of higher tier effects such as impaired immune response, reproduction, cellular repair mechanisms etc. [154]. The use of a human cell line makes it less relevant to the aquatic environment. It is expected to be relevant to refinery effluent constituents such as PAHs and metals. *Availability of trigger values:* Trigger values of 156µg Dichlorvos equivalents/L [39, 46], 1.4 mg Dichlorvos equivalents/L [210], and an effective concentration causing an induction ratio of 1.5 = 6 Relative Enrichment Factor [200] have been proposed in the literature. *Bioassay category:* Category 2 [39]. *Chemical analysis:* The test cannot *replace* chemical analysis.
Possible application limitations: The assay is based on genetically modified organisms, therefore permission is required [34]. A cytotoxicity control may be required.

4.3.2. Nrf2-pathway activation assay e.g. Nrf2-CALUX®

Description: This assay measures Nrf2 (nuclear factor erythroid 2–related factor 2) transcriptional activity, which indicates that a chemical causes oxidative stress in the U2OS human cell line. This cell line contains the firefly luciferase gene which is under the control of four Electrophile Responsive Elements (EpREs). The luciferase gene acts as the reporter gene for the activation on the Nrf2 pathway. The test uses an ELISA format (96-well plate) and a colorimetric readout can be observed at OD 450 nm.
Validation maturity: The Nrf2-pathway activation assay is available commercially. It is not yet ISO standardised, and is still in the process of being validated for water samples [34]. It is less mature than the AREc32 assay. *Validation maturity:* Medium (expert lab) [27, 34].

Performance:
Sensitivity: The sensitivity of the assay is variable, with many substances responsive in the AREc32 assay not responsive in the Nrf2-pathway activation assay [15]. For example, B(a)P is very responsive in the AREc32 assay [200] but not in the Nrf2-pathway activation assay [194]. The reasons for this have not yet been confirmed, but the cell line used may be the reason (see *Confounding factors* below).

Potency: The test shows potency of effect; effect concentrations are derived.

The assay is high-throughput [34, 194] and is relevant to water quality assessment [47, 200, 211].

Use:
Use in environmental samples: The use of the assay in environmental samples is occasional [27]. It has been used for surface water [105, 212, 213], recycled water [68], drinking water [200, 213], effluent [68, 214].

Use with refinery effluent constituents: Neither phenanthrene, pyrene, B(a)P, and dibenzo[a,h]pyrene had induced the transcription factor Nrf2 [215]. However, the Nrf2-CALUX® assay was responsive to light naphthenic North Sea crude oil using chemically enhanced Water-Accommodated Fractions and refined fuel oil (IFO 180) using low energy Water-Accommodated Fractions [216].

Use with refinery effluents/produced water: No information could be found.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes so should be applicable to passive sample extracts.

Time to run: Results within a week.

Sample processing: Metabolic activation may be required to see results with refinery effluents.

Cost per sample: The literature reports a cost of <100 € [34]. The results of the survey (n=1) indicate the assay is likely to be 3,000 - <10,000 € per sample.

Result interpretation:
Confounding factors: The use of U2OS cells may confound results. It has been considered that they have a limited capability for metabolism, hence why PAHs needing to be metabolically activated, such as B(a)P, showed response in the AREc32 assay but no response in the Nrf2 assay [215]. It has been suggested that U2OS cell line tests with petrochemical substances should use S9 fraction obtained from rat livers [216].

Relevance: Medium. Sub-cellular effects are sometimes deemed as less ecologically relevant than those measuring higher tier effects, however, oxidative stress response markers give an early indication of higher tier effects such as impaired immune response, reproduction, cellular repair mechanisms etc. [154]. The use of a human cell line makes it less relevant to the aquatic environment.

Availability of trigger values: Trigger values of 10 µg curcumin equivalents /L [38] and 26 µg Dichlorvos equivalents /L [39] have been proposed in the literature.

Bioassay category: Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations: The assay is based on genetically modified organisms, therefore permission is required [34]. A cytotoxicity control may be required.

4.4. METABOLISM

4.4.1. AhR activation assay e.g. DR CALUX®/DR Luc®

Description: The AhR activation assay is dioxin responsive and typically uses a H4IIE rat hepatoma cell line. Binding activity between the cell line and the aryl hydrocarbon receptor (AhR) is measured using a luminometer and results are expressed as a percentage of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) binding activity.
Validation maturity: The AhR activation assay is available commercially. It is not currently ISO standardised, but an ISO guideline is in prep (ISO/CD 24295) and is in the process of being validated for water samples [34]. A TIMES protocol is available: ICES TIMES No. 55. Protocol for measuring dioxin-like activity in environmental samples using in vitro reporter gene DR-LUC assays. *Validation maturity:* High (routine lab) [27, 34].
Performance: *Sensitivity*: The test is highly sensitive [29, 34, 35]. However, different DR LUC cell lines exist, which may differ in sensitivity [27], for example, the H4L1.1c4 (rat) cell line is more sensitive than a human liver cell line [39]. *Potency:* The test shows potency of effect; effect concentrations can be derived. The AhR activation assay is highly specific for AhR receptor agonists, such as dioxins and dioxin-like substances [29, 34]. It is also highly reproducible between laboratories and has high repeatability [217]. It is high-throughput [29, 34]. No false positive or false negative samples occurred in the EC project HORIZONTAL [28], although fresh growth medium can also induce the AhR and give false positive results [218].
Use: *Use in environmental samples:* The use of the assay in environmental samples is common [27, 34]. AhR CALUX assays have been used for surface water [47, 105, 202-204, 219], drinking water [13], wastewater effluent [46, 70, 196, 203, 207], and produced water, see below. *Use with refinery effluent constituents:* AhR assays have been responsive to B(a)P [47, 218], benzo[k]fluoranthene [220], other PAHs [220-222], distillate marine grade A (DMA) oil [223], and crude oil and refined petroleum products (including bunker oils) [218]. Jonker *et al*. [223] found that the observed toxicity of the oil was likely due to the specific oil constituents rather than the analysed PAHs. Bekki *et al*. [224] found that, in a DR-CALUX® assay, 11 PAH derivatives had AhR agonist activity, and 6 had AhR antagonist activity. *Use with refinery effluents/produced water:* Produced water from oil and gas installations on the UK Continental Shelf were responsive to the DR-CALUX® assay [36]. All of the 22 produced water effluents elicited a response before clean-up, whereas only 13 elicited a response following the use of an acid silica column to remove PAHs [36].

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [54, 72, 88, 225].

Time to run: Results within a week; 24-hour exposure (although this can vary depending on the exact assay and cell line used).

Sample processing: Samples need to be cleaned up by a sulphuric acid pre-treatment, which removes some PAHs and other less stable compounds [28, 36, 226]. Afterwards, an additional step to separate dioxin-like PCBs from polychlorinated dibenzo-p-dioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) may be required [28].

Cost per sample: The literature reports a cost of 100-1,000 € [34] or <200 € [28]. The results of the survey (n=1) indicate the assay is likely to be 300 - <1,000 € per sample.

Result interpretation:
Confounding factors: AhR antagonists can confound results, underestimating the effects [218]. The exposure time can also influence the results generated, for example, Machala *et al*. [220] found that the 6-hour exposure to PAHs led to significantly higher activity than the 24-hour exposure, likely due to the high rate of PAH metabolism. Different cell types and tissues can result in different transcriptome profiles induced by the same AhR activator substance, and not all of these substances initiate similar molecular and cellular responses; AhR modulation occurs in a ligand-specific manner, not just in a cell- and/or species-specific manner [17]. Many small molecular weight compounds can bind to AhRs, including tryptophan metabolites, which may be present in fresh growth medium, generating false positive results [17, 218].
Cytotoxic effects may also occur [47], however, owing to the sulphuric acid clean-up step, cytotoxicity is rare [28]. Cytotoxicity was not observed in Vrabie *et al*. [218] or [227], although induction was reduced at the highest test concentrations of oils. Other factors considered for this decrease include specific additives, inhibition of cellular processes, and/or inhibition of the AhR induction pathway by petroleum compounds/oil additives, by increased interference of the light detection due to the darker colour of the medium (high oil concentrations) [218]. This latter issue will be less relevant for testing whole effluents.

Relevance: Medium. Activation of the AhR is arguably of higher ecological relevance than sub-cellular effects, but it is just receptor mediation rather than upregulation [27]. The use of rodent cell lines is of limited relevance to aquatic assessment.

Availability of trigger values: Trigger values of 50 pg 2,3,7,8- Tetrachlorodibenzo-pdioxin (TCDD) equivalents /L [38, 54], 6.36 ng BaP equivalents/L [39] and 16.2 pg TCDD-equivalents /L [72] have been proposed in the literature.

Bioassay category: Category 2 [39].

Chemical analysis: The test cannot replace chemical analysis.

Possible application limitations:
The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required and Brack *et al*. [29] recommends also conducting experiments with antagonists for specificity evaluation. An increased exposure time may also be necessary to distinguish between chemicals of different degradabilities [218]. Vrabie *et al*. [218] also utilised pre-conditioned medium to avoid a high background signal induced by tryptophan products present in the fresh growth medium. Further, the assay is based on genetically modified organisms, therefore permission is required [34]. CALUX® assays require a commercial license [29].

4.4.2. PAH-specific AhR activation assay e.g. PAH CALUX®

Description: The PAH-specific AhR activation assay is based on an AhR activation assay. It is optimised to respond specifically to AhR activity from readily biodegradable PAHs by using a cell line which is dioxin responsive and uses a H4IIE rat hepatoma cell line which has lower metabolic activity than a typical AhR assay. The test also has a shorter duration than the AhR assay. Binding activity between the cell line and the aryl hydrocarbon receptor (AhR) is measured using a luminometer and results are expressed as a percentage of benzo(a)pyrene binding activity.
Validation maturity: The assay is available commercially but not standardised to ISO guidelines. *Validation maturity:* Medium (expert lab) [27].
Performance: *Sensitivity*: The PAH CALUX® has high sensitivity [226]. *Potency:* The test shows potency of effect; effect concentrations can be derived. The PAH CALUX® is optimised for use with more biodegradable PAHs, but is specific for any AhR receptor agonists, including polychlorinated biphenyls, dioxins, and pharmaceuticals [38]. It has high repeatability and reproducibility, and is also highly predictable, especially for carcinogenic PAHs [226]. It is high-throughput and has high specificity [34].
Use: *Use in environmental samples:* The use of the assay in environmental samples is occasional [27]. It has been applied to surface water [39, 73] and wastewater effluent [228]. *Use with refinery effluent constituents*: The assay responds to B(a)P-like compounds, especially the higher aromatic PAHs [28]. It has been response to crude oil-contaminated water [229]. *Use with refinery effluents/produced water:* No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [27]. *Time to run:* Results within a day; 6-hour exposure time. *Sample processing:* Samples can be frozen and still elicit a response [229]. *Cost per sample:* <200 € [28].
Result interpretation: *Confounding factors:* Different cell types and tissues can result in different transcriptome profiles induced by the same AhR activator substance, and not all of these substances initiate similar molecular and cellular responses; AhR modulation occurs in a ligand-specific manner, not just in a cell- and/or species-specific manner [17]. Many small molecular weight compounds can bind to AhRs, including tryptophan metabolites, which may be present in fresh growth medium, generating false positive results [17, 218]. Cytotoxic effects may also occur [47], and a darker coloured sample may interfere with the light detection of results [218]. AhR antagonists can also inhibit activity [27].

Relevance: Medium. Activation of the AhR is arguably of higher ecological relevance than sub-cellular effects, but it is just receptor mediation rather than upregulation [27]. The use of rodent cell lines is of limited relevance to aquatic assessment, and it has been reported that dioxins and dioxin-like substances do not significantly contribute to risks in surface waters [54]. *Availability of trigger values:* Trigger values of 150 ng B(a)P equivalents /L (overall PAH activity) [38] 6.21 ng B(a)P equivalents /L [39] and 62.1 ng B(a)P equivalents /L [54] have been proposed in the literature. *Bioassay category:* Category 2 [39]. *Chemical analysis:* A response in this assay is indicative that there are PAH substances, and other substances such as polychlorinated biphenyls and dioxins, present in the sample, therefore it could replace chemical analysis of PAHs.
Possible application limitations: CALUX® assays require a commercial license [29]. A cytotoxicity control may be required. Further, the assay is based on genetically modified organisms, therefore permission is required [34].

4.4.3. EROD assay

Description: This assay is used to measure the ability of a substance to induct a P450 monooxygenase 1A (CYP1A). This induction is caused by the activity of 7-ethoxyresorufin-O-deethylase (EROD) in a rainbow trout (*Onchorhynchus mykiss*) liver cell line (RTL-W1). Induction of CYP1A can lead to the development of cancer or toxicity, through direct or indirect mechanisms. EROD activity is measured using fluorescence.
Validation maturity: The EROD assay is available commercially. It is ISO standardised *in vivo* to ISO/TS 23893-2, but not *in vitro*. *Validation maturity*: Medium (expert lab) [27].
Performance: *Sensitivity*: The test is highly sensitive [29, 40], but less sensitive than e.g. the DR CALUX or DR LUC assays [29]. *Potency:* The test shows potency of effect; effect concentrations can be derived. The assay is also moderately specific; it does not differentiate between dioxin substances and PAHs [28, 29]. However, it has been demonstrated that varying the exposure duration can increase the specificity of the assay, as PAHs show higher potency after a 4-hour exposure, and dioxins after 4- or 24-hour exposure [230].
Use: *Use in environmental samples:* The use of the assay in environmental samples is common [27]. The EROD assay has been used to assess surface water [167, 231], lake water receiving paper mill effluents [232] and sediments [128, 230, 233].

Use with refinery effluent constituents: The EROD assay is responsive to dioxins, planar PCBs and PAHs, however branched PAHs often do not induce the same response [28]. It has been responsive to B(a)P, 3-methylcholanthrene, benzo(b)fluoranthene, chrysene and benzo(a)anthracene [234], as well as other PAHs [167].

Use with refinery effluents/produced water: The assay has been used to assess produced water from a North Sea offshore oil production platform [235] and fractionated petroleum refinery effluent [132]. Whilst not testing refinery effluent itself, Suares-Rocha *et al*. [233] evaluated EROD activity of sediments receiving oil refinery effluent.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [167].

Time to run: Results within a day; 3-hour exposure.

Sample processing: No known sample processing considerations.

Cost per sample: The literature reports a cost of <200 € [28]. The results of the survey (n=1) indicate the assay is likely to be 300 - <1,000 € per sample.

Result interpretation:
Confounding factors: Protein concentration can influence results, therefore protein normalisation is usually required [236]. Results can also be affected by certain substances that cause substrate inhibition, although this can be mitigated by using reporter gene cell lines [217]. Further, the induction can be inhibited by too high concentrations of the Ah-receptor agonists, as well as other chemicals like metals and xenoestrogens [28]. In the environment, EROD activity can be affected by water temperature, age, and reproductive cycle [128].

Relevance: High. Activation of the CYP1A is arguably of higher ecological relevance than sub-cellular effects, and this assay measures protein upregulation rather than receptor mediation as with the AhR activation assays [27]. Further, 7-ethoxyresorufin-Odeethylase is an enzyme related to PAHs, so is of high relevance of refinery effluents. As cells originate from an organism with biotransformation capabilities, it has high environmental relevance [29].

Availability of trigger values: No trigger values were found in the literature.

Bioassay category: Not stated but assume Category 2 [39].

Chemical analysis: The test could potentially replace chemical analysis as it responds to PAHs and dioxin-like compounds [28].

Possible application limitations:
The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required and Brack *et al*. [29] recommends also conducting experiments with antagonists for specificity evaluation.
Protein concentration needs to be measured, as results should be normalised to protein concentration. Unlike the other metabolism assays evaluated here, it does not use genetically modified organisms, making it more applicable to more laboratories.

4.5. ENDOCRINE DISRUPTION

4.5.1. Yeast Estrogen Screen (YES) and Arxula-yeast estrogen screen (A-YES)

Description:
The YES assay is a 96-well assay which uses a recombinant yeast strain to allow the identification of substances which can interact with the human oestrogen receptor alpha (hERa). The concentration required to elicit a 50% response (EC50) in the exposed organisms for the positive control (17ß-oestradiol (E2)) and the percentage of the sample required to give the equivalent response in the test sample allows the calculation of oestradiol equivalent factors (EEQ). Cytotoxicity is determined by determining the growth of the yeast cells at 620 nm and comparing this to the enzyme activity measured at 540 nm.

A-YES is a 96-well assay and uses the recombinant yeast *Arxula adeninivorans* as an oestrogen-responsive biosensor to allow the identification of substances which can interact with the human oestrogen receptor alpha (hERa). It determines the calculation of oestradiol equivalent factors (EEQ), using 17ß-oestradiol (E2) as a control. Detection is performed photometrically.

Validation maturity:
The YES and A-YES assays are available commercially, and standardised to ISO 19040-1 and ISO 19040-2, respectively. The YES assay has been applied in regulatory context, for wastewater [30], and is in the process of being validated to water samples [34].

Validation maturity: High (routine lab) [27, 34, 40].

Performance:
Sensitivity: The assays are less sensitive than mammalian cell-based assays [29, 40, 237]. The YES assay was an order of magnitude less sensitive than ER-CALUX, MELN, T47D-KBluc and E-SCREEN assays, producing a lot of non-detects [238].

Potency: The test shows potency of effect; effect concentrations are derived.

The assay is high-throughput, selective, and highly reproducible between laboratories [27, 29, 34].

Use:
Use in environmental samples: The use of the YES assays in environmental samples is common [27]. It has been used for:

- Surface water [167, 239];
- Wastewater effluent [41, 141];
- Drinking water [68].

Use with refinery effluent constituents: PAHs in lake water demonstrated estrogenic potential in a YES assay [167]. While not a YES assay, Vrabie et al. [240] found that oils were estrogenic in a yeast assay using the ß estrogen receptor. More recently, native and chemically dispersed crude oil WAFs activated the ERa in an A-YES assay; resulting estradiol equivalents of the WAFs were above the established effect-based trigger values [241].

Use with refinery effluents/produced water: Produced water effluents from Norwegian North Sea oil production platforms found no estrogenic activity in the YES assay with filtered oil droplets, but did in the dissolved phase [242]. Further, all five samples of produced water effluents from British and Norwegian North Sea oil production platforms were responsive in YES assays [36]. The 17 beta-estradiol equivalent concentration was 0.01-91 ng/L, thus some were above the EBT values of 0.2–0.4 ng 17 beta-estradiol equivalents /L.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [167].

Time to run: Results within a week; ≤72-hour exposure for YES and 48-hour exposure A-YES.

Sample processing: Often sample enrichment prior to testing environmental samples is required, as concentrations in native samples are very low.

Cost per sample: The literature reports a cost of 100-1,000 € [34] for YES and A-YES, or <200 € [28] for YES. The results of the survey (n=1) indicate the assay is likely to be 1,000 - <3,000 € per sample.

Result interpretation:
Confounding factors: Anti-estrogenic substances, receptor subtype, and chlorination can also confound results. For example, the responsiveness of PAHs depends on whether the α or ß estrogen receptor is used in the assay; in the yeast assay using the α receptor, none of the oils were responsive, whereas almost all of them were responsive in the yeast assay using the ß receptor [240]. As the α estrogen receptor is the most commonly applied, estrogenic substances acting via the ß estrogen receptor may not be captured, and hence estrogenicity will be underestimated. Further, estrogenic compounds may be chlorinated during wastewater treatment, altering the potency of their effects [36].

Relevance: Medium. There is strong evidence that ER activation as a molecular initiating event is linked to adverse outcomes at higher biological levels [47]. However, the reported potencies of refinery effluent-relevant samples are low and unlikely to be responsive to whole refinery effluent samples.

Availability of trigger values: A trigger value of 0.2–0.4 ng 17 beta-estradiol equivalents /L [243] has been proposed for the YES assay, while values of 0.56 ng 17 beta-estradiol equivalents /L [39] and 0.4 ng 17 beta-estradiol equivalents /L [13] have been proposed for A-YES.

Bioassay category: Category 1 [39].

Chemical analysis: The test is Category 1 and could possibly replace chemical analysis for e.g. phenols found in refinery effluents [244].

Possible application limitations:
The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required [40] and Brack *et al.* [29] recommends also conducting experiments with antagonists for specificity evaluation. The assay is based on genetically modified organisms, therefore permission is required [34]. However, unlike the CALUX assays, it is license free.

4.5.2. ER activation assay e.g. ER CALUX®/ER Luc®

Description:
This assay uses a human bone marrow cell line (U2OS) to determine oestrogen receptor binding activity. It utilises the firefly luciferase gene coupled with oestrogen responsive elements (EREs) as the reporter gene, which allows the identification of the presence of oestrogens and oestrogen-like compounds. Binding is based on 17ß-estradiol (E2) equivalents or E2 equivalence factors (EEFs). The assay can also be used to identify anti-oestrogenic activity.

Validation maturity:
The assay is available commercially and can be conducted according to ISO 19040-3. The test is in the process of being validated to water samples [34].

Validation maturity: High (routine lab) [27, 34].

Performance:
Sensitivity: The sensitivity is high, and more sensitive than yeast cell-based assays [29, 40, 237].

Potency: The test shows potency of effect; effect concentrations are derived.

The assay is high-throughput and moderately specific [29, 34]. The ERα-CALUX® was the most repeatable and precise of five estrogenic bioassays tested (ERα-CALUX®, YES, T47D-KBluc, MELN and GeneBLAzer-ERα) [245]. The ER CALUX® was found to be a good predictor of *in vivo* estrogenic activity in a range of compounds [246].

Use:
Use in environmental samples: The use of the assay in environmental samples is common [27]. It has been used for:

- Surface water [247-250];
- Drinking water [68, 248, 251, 252];
- Recycled water [67, 68, 253];
- Wastewater effluent [68, 198, 250, 254-257].

Use with refinery effluent constituents: B(a)P, benzo(b)fluoranthene and crude and refined oils have all been responsive to ER assays using human cell lines [47, 258]. When testing crude oils, there were no major differences in response whether using the a or ß estrogen receptor, contrasting results seen with yeast-based assays [258]. More recently, native and chemically dispersed crude oil WAFs activated the ERa in an ER CALUX® assay; resulting estradiol equivalents of the WAFs were above the established effect-based trigger values [241].

Use with refinery effluents/produced water: No information could be found.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [54, 72, 88].

Time to run: Results within a week; 24-hour exposure.

Sample processing: Often sample enrichment prior to testing environmental samples is required, as concentrations in native samples are very low.

Cost per sample: The literature reports a cost of 100-1,000 € [34] or <200 € [28]. The results of the survey (n=1) indicate the assay is likely to be 1,000 - <3,000 € per sample.

Result interpretation:
Confounding factors: Cytotoxicity and matrix effects are higher than for YES, which can confound results [257]. Brack *et al.* [29] further noted the possibility of cross-talk interaction between different nuclear receptor pathways within the same cell line, such as between estrogen or androgen receptors and aryl hydrocarbon receptors. Anti-estrogenic substances, receptor subtype, and chlorination can also confound results (REF). For example, the responsiveness of PAHs depends on whether the α or ß estrogen receptor is used in the assay [258]. As the α estrogen receptor is the most commonly applied, estrogenic substances acting via the ß estrogen receptor may not be captured, and hence estrogenicity will be underestimated. The cell line can also affect results; B(a)P and benzo(b)fluoranthene were

active in a human cell line but not zebrafish cell lines in estrogenic assays [47]. Further, estrogenic compounds may be chlorinated during wastewater treatment, altering the potency of their effects [36].

Relevance: Medium. While human cell lines might not be the most reliable proxy for effects in the environment, owing to inter-species differences in estrogen receptor subtypes between fish and humans, there is strong evidence that ER activation as a molecular initiating event is linked to adverse outcomes at higher biological levels [47]. The correlation between effects seen in the ER CALUX® and *in vivo* data using the Allen-Doisy assay was high (r^2 =0.87) [246]. However, the reported potencies of refinery effluent-relevant samples are low and unlikely to be responsive to whole refinery effluent samples.

Availability of trigger values: Trigger values have been proposed in the literature:

- 0.2-0.5 ng 17 beta-estradiol equivalents /L [72, 259];
- 0.28 ng 17 beta-estradiol equivalents /L[260] ;
- 0.2-0.4 ng 17 beta-estradiol equivalents /L [243];
- 0.1 ng 17 beta-estradiol equivalents /L [39];
- 0.5 ng 17 beta-estradiol equivalents /L [38];
- 0.283 ng E2 equivalents /L [13].

Bioassay category: Category 1 [39].

Chemical analysis: The test is Category 1 and could possibly replace chemical analysis for e.g. phenols found in refinery effluents [244].

Possible application limitations:
The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required [34, 40] and Brack *et al.* [29] recommends also conducting experiments with antagonists for specificity evaluation. The assay is based on genetically modified organisms, therefore permission is required [34]. CALUX® assays require a commercial license [29].

4.5.3. Yeast Androgen Screen (YAS)

Description:
This is a 96-well assay which uses a recombinant yeast strain to allow the identification of substances which can interact with the human androgen receptor alpha (hAR). The concentration required to elicit a 50% response (EC50) in the exposed organisms for the DHT positive control and the percentage of the sample required to give the equivalent response in the test sample allows the calculation of dihydrotestosterone (DHT) equivalent factors (EEQ). Cytotoxicity is determined by determining the growth of the yeast cells at 620 nm and comparing this to the enzyme activity measured at 540 nm.

Validation maturity:
The YAS assay is available commercially and an ISO guideline is in preparation. The test has been applied in regulatory context, for wastewater [30], and is in the process of being validated to water samples [34].

Validation maturity: High (routine lab) [27, 34, 40].

Performance:
Sensitivity: The assay is less sensitive than mammalian cell-based assays [29, 40, 237].

Potency: The test shows potency of effect; effect concentrations are derived.

The assay is high-throughput, selective, specific, and is highly reproducible between laboratories [29, 34].

Use:
Use in environmental samples: The use of the YAS assay in environmental samples is common [40]. It has been used for:

- Surface water [68, 239, 261, 262];
- Recycled water [68];
- Drinking water [68];
- Wastewater effluent [68, 141, 239, 263].

Use with refinery effluent constituents: Oil SARA ('saturates', 'aromatics', 'resins', and 'asphaltenes') fractions show synergistic effect related to the androgenic system [264]. Vrabie *et al.* [240] showed that crude oils and some refined products (including bunker oils) produced responses in androgen receptor-mediated yeast assays.

Use with refinery effluents/produced water: AR antagonists were detected in both the dissolved and oil associated phase of produced water from North Sea oil production platforms [242]. no AR agonist activity could be detected in another study with produced water from North Sea oil production platform [265].

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [263].

Time to run: Results within a week; ≤72-hour exposure.

Sample processing: Sample enrichment would likely be required to detect an effect in surface water [15].

Cost per sample: The literature reports a cost of 100-1,000 € [34] or <200 € [28]. The results of the survey (n=1) indicate the assay is likely to be 1,000 - <3,000 € per sample.

Result interpretation:
Confounding factors: Cytotoxicity can confound results. For example, fatty acids have identified as one source of cytotoxicity in the YAS assay [266]. Anti-androgenic substances can also mask effects [29, 240].

Relevance: Low. Inhibition or activation of the androgen receptor can interfere with reproduction, metabolism and the immune system, all of which have high relevance to effects in the environment. However, androgenic or anti-androgen activity is arguably less relevant than its estrogen equivalent, as generally androgen-related activity is not often detectable in environmental water samples [15, 41].

Availability of trigger values: No trigger values have been proposed in the literature.

Bioassay category: Category 1 [39].

Chemical analysis: The test is Category 1 and could therefore possibly replace chemical analysis for some constituents of refinery effluents, such as PAHs which are anti-androgenic [267].

Possible application limitations: The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required [40] and Brack *et al.* [29] recommends also conducting experiments with antagonists for specificity evaluation. The assay is based on genetically modified organisms, therefore permission is required [34]. However, unlike the CALUX assays, it is license free.

4.5.4. AR activation assay e.g. AR CALUX®/AR Luc®

Description: This assay can use both hamster and human cell lines. In the hamster cell line (CHO-K1 hamster ovarian cancer cells), androgen receptor-mediated and gene expression can be determined by evaluating binding on 5-alpha dihydrotestosterone (DHT) equivalents. The assay can also be used to detect anti-androgenic activity. In the human cell line (human bone marrow (U2OS) cells), the firefly luciferase gene was coupled with Androgenic Receptor Elements (AREs) which acted as a reporter gene for the presence of androgens and androgen-like compounds.
Validation maturity: The assay is available commercially and can be conducted according to OECD 458 for the hamster line. No standard exists yet for the human bone marrow cell line. The test is in the process of being validated to water samples [34]. *Validation maturity:* Medium (expert lab) [27].
Performance: *Sensitivity*: The sensitivity is high, and more sensitive than yeast cell-based assays [29, 40, 237]. *Potency:* The test does shows potency of effect; effect concentrations are derived. The assay is high-throughput and moderately specific [29, 34].
Use: *Use in environmental samples:* The use of the assay in environmental samples is common with the human bone marrow cell line and occasional with the hamster cell line [27]. It has been used for: • Surface water [41, 68, 250-252, 255, 259, 268]; • Drinking water [68, 252]; • Recycled water [41, 67, 68, 253]; • Wastewater effluent [41, 67, 68, 198, 250, 251, 253, 254, 257]; • Sediments [88]. *Use with refinery effluent constituents*: B(a)P possessed anti-androgenic potency in an AR assay [72]. Use with refinery effluents/produced water: No information could be found.
Conduct of test: *Applicability to passive sample extracts:* The test utilises small sample volumes (40-200µL typical final volume [29]) and is applicable to passive sample extracts [88]. *Time to run:* Results within a week; 24-hour exposure. *Sample processing:* Sample enrichment would likely be required to detect an effect in surface water [15]. *Cost per sample:* 100-1,000 € [34]for both cell lines, and <200 € for the human bone marrow cell line [28].

Result interpretation: *Confounding factors:* Cytotoxicity and matrix effects are higher than for YAS, which can confound results [257]. Antagonists and partial agonists of the androgen receptor can mask effects [72, 196, 269]. Weiss *et al.* [269] found that the androgenic potential of the fractions could only be elucidated after sequential fractionation, thus testing of whole-effluent samples would not give a good indication of the full androgenic potential of components within it. Brack *et al.* (2016) further noted the possibility of cross-talk interaction between different nuclear receptor pathways within the same cell line, such as between estrogen or androgen receptors and aryl hydrocarbon receptors. *Relevance:* Low. Inhibition or activation of the androgen receptor can interfere with reproduction, metabolism and the immune system, all of which have high relevance to effects in the environment. However, androgenic or anti-androgen activity is arguably less relevant that its estrogen equivalent, as generally androgen-related activity is not often detectable in environmental water samples [15, 41]. Further, the use of a human or hamster cell line makes it less relevant to the aquatic environment. *Availability of trigger values:* Trigger values of 25 mg flutamide equivalents /L [38] and 14.4 µg flutamide equivalents /L [39] (antagonistic activity on the androgen receptor) have been proposed in the literature [38] . *Bioassay category:* Category 1 [39]. *Chemical analysis:* The test is Category 1 and could therefore possibly replace chemical analysis for some constituents of refinery effluents, such as PAHs which are anti-androgenic [267].
Possible application limitations: The assay may become more laborious with the addition of more steps; a cytotoxicity control may be required [34, 40] and Brack *et al.* [29] recommends also conducting experiments with antagonists for specificity evaluation. The assay is based on genetically modified organisms, therefore permission is required [34]. CALUX® assays require a commercial license [29].

4.5.5. TTR binding assay

Description: This assay involves the displacement of the thyroid hormone precursor thyroxine (T4) from its plasma transport protein, transthyretin (TTR). The sample competitively binds to human TTR and can be identified by ^{125}I-labelled T4 and unlabelled T4. Once equilibrium is reached, the binding of ^{125}I-labelled T4 can be determined using a gamma counter. The result is expressed as a percentage of the control and allows the identification of a T4-equivalent concentration.
Validation maturity: The TTR binding assay is available commercially. It is not ISO standardised. *Validation maturity:* Low (research lab) [27].
Performance: *Sensitivity*: The sensitivity of the assay is variable; it's less sensitive than ER- and AR-assays [270]. *Potency:* The test shows potency of effect; IC50 and EC50 values can be derived. It is high-throughput, available in a 96-well plate.

Use:
Use in environmental samples: The use of the TTR binding assay in environmental samples is rare. It has only recently been used to assess surface water [270], marine sediments [88], estuarine sediments [271], treated wastewater [270], wastewater treatment plant sewage sludge [272] and wastewater effluents [273].

Use with refinery effluent constituents: It is responsive to PAH quinones (e.g. benzo[c]phenanthrene-[1,4]-quinone) and hydroxides [224]. It has been reported that compounds with no functional group and that are non-polar have lower affinity for binding to TTR; this includes PAHs, PCBs and PBDEs [274]. However, metabolic activation was demonstrated to increase the TTR binding potency of non-polar fractions of POP-polluted sediments up to 100 times [275].

Use with refinery effluents/produced water: No information could be found.

Conduct of test:
Applicability to passive sample extracts: The test utilises small sample volumes and is applicable to passive sample extracts [72, 88].

Time to run: Results within a day.

Sample processing: An additional T4 clean-up step is sometimes required [271, 276]. Careful handling is required to prevent TTR denaturation, such as avoiding mixing with a vortex during of the procedure [273]. Metabolic activation may be required for non-polar samples [275]. Further, as with all endocrine disruption assays, sample enrichment may be required.

Cost per sample: Unknown. No survey respondents conducted this survey.

Result interpretation:
Confounding factors: TTR displacement assays cannot differentiate agonists or antagonists [270].

Relevance: Medium. Interference of hormone receptors is a molecular initiating event of high toxicological relevance. However, TTR binding is not very relevant to PAHs and other non-polar compounds, as would be found in refinery effluents, as they have low binding affinity for TTR.

Availability of trigger values: Trigger values of 58 ng T4 equivalents /L [13, 39] and 49 ng Thyroxine equivalents /L [13, 39] have been proposed in the literature.

Bioassay category: Category 1 [39].

Chemical analysis: The test is Category 1 and could therefore possibly replace chemical analysis for some constituents of refinery effluents.

Possible application limitations:
No application limitations have been found in the literature or were raised during the survey.

4.6. NOVEL DEVELOPMENTS

Some of the evaluated assays are relatively novel, such as the PAH-CALUX®, however, new methods are being developed and validated all the time. Research exploring new methods for the assessment of effluent toxicity has focused largely on *in vitro* methods than can be used as *in vivo* alternatives. A few assays have been developed using the rat cardiomyoblast H9c2(2-1) cell line. One assay utilises a sulforhodamine B (SRB) cell mass colorimetric method to measure growth inhibition of the cell line exposed to water samples. Its sensitivity was comparable to that of acute toxicity tests with fish when testing metal, oil, municipal and paper effluents [277]. The rat cardiomyoblast H9c2(2-1)-based sulforhodamine B assay has been demonstrated for use with monitoring biological effluent toxicity, with the best results attained by a lyophilising pre-treatment [278].

5. SURVEY RESULTS

A total of 14 laboratories participated in the survey, comprising contract research organisations (n=8), research institutes (n=4) and industry laboratories (n=2) from Europe, North America and Asia. Out of the 14 laboratories, 13 conducted *in vivo* tests and 4 conducted in *vitro* tests. The number of laboratories surveyed that conducted each *in vivo* and *in vitro* test is shown in Figure 1 and 2, respectively.

Figure 1: The number of laboratories conducting each in vivo test.

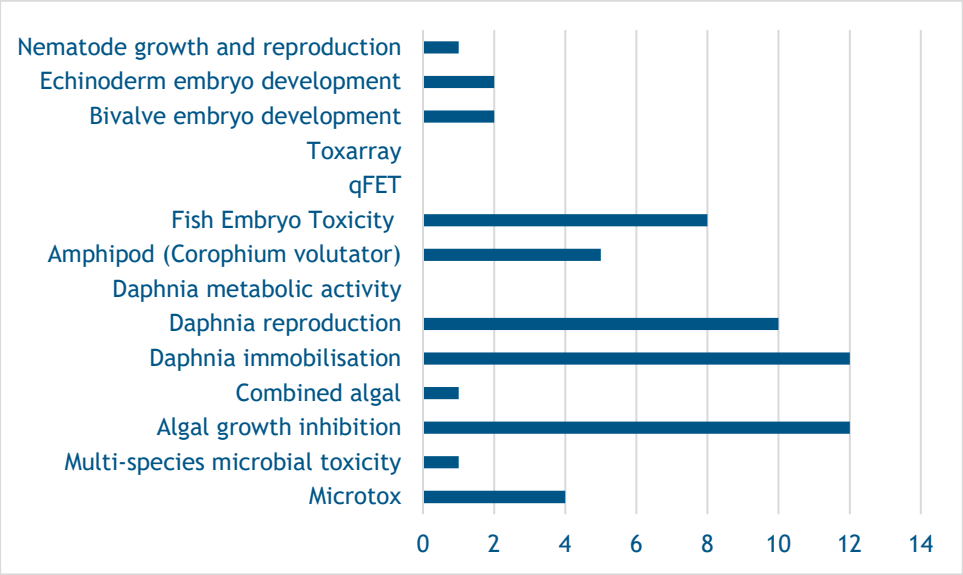
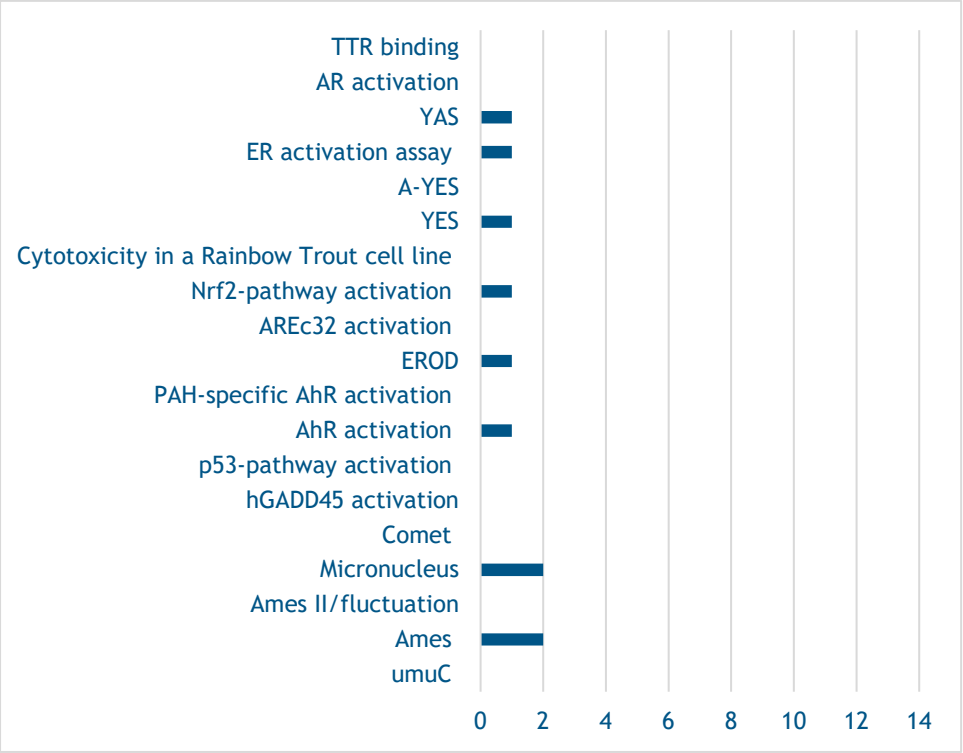


Figure 2: The number of laboratories conducting each in vitro test.



The *Daphnia* reproduction and algal growth inhibition assays were the most common *in vivo* tests, with 12 out of the 13 *in vivo*-conducting laboratories running them. None of the laboratories surveyed conducted the Toxarray, qFET or *Daphnia* metabolic activity assays. The micronucleus and Ames assays were the most commonly performed *in vitro* tests, with 2 out of the 4 *in vitro*-conducting laboratories running them.

The Ames, micronucleus and AhR activation assays were all conducted very regularly (more than 50 per year), the Nrf2-pathway activation assay was conducted regularly (10-50 per year), and the YES, YAS and ER activation assays were conducted infrequently (less than 10 per year). The MicroTox® test was conducted infrequently (50% of answers) or rarely (50%), the multi-species microbial toxicity and bivalve embryo development assays were conducted rarely, the nematode growth and reproduction test was conducted regularly, and the frequency of conducting the other tests varied greatly depending on the laboratory.

In terms of costs per sample, the MicroTox® test was the cheapest *in vivo* test (under 100 € to under 1,000 €), while the reported costs of the fish embryo toxicity (FET) and *Daphnia* reproduction tests encompassed the range above 1,000 to above 10,000 €. The AhR activation and EROD assays appeared to be the cheapest of the 8 *in vitro* assays the surveys laboratories conducted (above 300 to under 1,000 €), while the two respondents for the micronucleus assay both reported cost per sample of above 10,000 €.

Most of the tests were used for both marine and freshwater samples, with some also being used for brackish water, wastewater, drinking water, recycled water and storm water. It should be noted that it is not known whether these are effluent samples or samples from receiving water bodies. The one laboratory conducting the nematode growth and reproduction test only used freshwater. For the amphipod test, 80% of the respondents conducted the test with marine and/or brackish water, while 20% used freshwater as well as marine water.

Some of the challenges given included some general concerns relating to all tests, e.g. availability of testing supplies (other scheduled testing), the presence of volatile or adsorptive components, and timing constraints, as well as some test-specific challenges that are mentioned in the individual evaluations, e.g. seasonality of the test species.

6. CONSIDERATIONS FOR USING WHOLE EFFLUENT AND PASSIVE SAMPLE EXTRACTS

An effective water/effluent sampling strategy is required to ensure meaningful bioassay results are obtained. This may include the use of whole effluent samples or extracts from passive samplers.

Whole effluent toxicity testing (WET) uses effluent water that undergoes serial dilution to produce a range of effluent concentrations for use as the test media. This approach measures the aggregate effect of all constituents present in the effluent sample. However, samples may possess bulk qualities unfavourable for the performance of some bioassays, such as poor nutrient content or high salinity. For *in vivo* assays, the careful selection of test organism can negate many of these issues. For other assays, some sample preparation may be required in the form of pH adjustment, filtration, or nutrient media addition [15]. As noted in section 4, some *in vitro* assays (e.g. genotoxicity and endocrine disruption assays) may already require additional sample preparation steps, such as sample enrichment, when whole effluent samples are used owing to the dosing factor employed in these assays. To avoid interference from matrix effects and ensure any observed toxicity is caused by the constituents of the effluent the assay's robustness to any potential matrix interference should be investigated in these cases [15].

In addition to WET tests, several effects-based monitoring studies describe the use of passive sampling followed by toxicity profiling of the extract, e.g. Hamers *et al.* [72] and De Baat *et al.* [73], as an effective approach for collection of a representative water or effluent sample. A review of the scientific literature on the use of effect-based tools in combination with passive samplers has also been performed [27]. An analysis of the strengths and weakness of the approach and a case for it being a more cost-effective monitoring strategy are presented by Hamers *et al.* [72] and evidenced through the TIPTOP (Time-Integrated Passive sampling combined with Toxicity Profiling) study. Finally, a recent Concawe report [37] has critically assessed the use of effects-based methods in combination with passive sampling and chemical analysis, using effluent streams from three refineries.

Passive samplers are sampling devices deployed *in situ* for several days to weeks, providing a time-integrated aqueous sample. They accumulate freely dissolved contaminants over time through the movement of the substance from the sampled aqueous media – water or effluent, to the sampler substrate. The nature of the passive sampler substrate and the hydrophobicity of the contaminants will determine the composition of the sample according to the hydrophobicity of contaminants. For example, partition-based samplers such as silicone rubber sheets and semi-permeable membrane devices (SPMDs) accumulate hydrophobic compounds (typically log Kow > 3), whilst adsorption based passive samplers such as Speedisks or the polar organic chemical integrative sample (POCIS) are more suitable for hydrophilic organic compounds such as many pesticides and pharmaceuticals. More than one type of passive sampler may need to be deployed at a site to capture a fully representative sample.

Passive sampler extracts offer several advantages over samples collected through traditional spot sampling techniques. A large volume of water or effluent is sampled over several days to weeks, providing a time-integrated sample and maximising the opportunity to detect episodic events and contaminants which may be missed with spot sampling. Any peak concentrations are, however, averaged out over a longer sampling period.

Many contaminants are present at trace (ng/L) levels in surface waters and potentially below the limits of (bio)analytical detection. This is particularly true of surface and groundwater samples from areas with relatively few anthropogenic inputs that typically have lower contaminant concentrations, but may be less relevant for refinery effluents as these are anticipated to have higher contaminant loads before dilution. The accumulation of contaminants over time in a passive sampler extract provides an enriched sample, increasing the sensitivity and likelihood of detection of effects in any subsequent bioassay or chemical analysis.

Despite these advantages there are technical limitations which need to be considered when using passive sampling extracts. Although enriched samples produced from passive sampling increase sensitivity, they can also lead to results that are difficult to interpret. Testing with concentrated, and often highly hydrophobic, oily extracts can also present technical challenges. For example it is challenging to conduct tests with oily extracts and care must be taken to dose appropriately to ensure relevant exposure and avoid oiling of the test organisms during the bioassay. The stability of the extracts in the culture media should also be confirmed.

The transformation of bioassay response from a passive sampler extract to a corresponding water phase concentration is also challenging since it is not possible to readily determine an accurate ambient water concentration from a passive sampler extract. Indeed, Hamers *et al.* [72] describe it as a 'theoretical impossibility' to translate exactly the observed toxicity in the passive sampler to a corresponding toxicity in the water. In the Concawe work [37], the concentrations from the passive samplers were recalculated to original waters using average concentration factors for the bioassays, as the compounds causing the toxicity were not known. However, this introduces large uncertainties into the recalculated bioassay results, and it is questionable whether the results can be used as a quantitative prediction of toxicity in the original samples [37].

Further, quantitative analysis of passive sampler extracts requires knowledge of the identity of all substances present in an extract and the sampling rates, which are specific to the type of sampler and substance. A performance reference compound (PRC) is spiked onto partition-based sampling devices when quantitative chemical analysis is required. This cannot be done on samplers where the extract is subsequently used for toxicity profiling, and so additional samplers must be deployed for chemical analysis alongside those deployed for bioassay sample collection. However, the lack of quantitative chemical information from passive samplers is not necessarily a major limitation. The use of passive samplers in combination with toxicity screening can be used effectively as a screening tool to identify sites where effect-based trigger values have been exceeded and to prioritise them for further investigation, although the partitioning of substances/phases in passive sampling does contradict the WET approach [73]. This is particularly true when the preparation of an extract forms essentially a 'new' substance that exists in an oily state that was not present in the effluent. It is a different composition than was in the effluent, and the extract will not include other constituents present in the effluent such as salts, ammonia and organic acids.

7. BATTERY RESULTING FROM EVALUATION

7.1. TEST BATTERY DESIGN

Any given water sample, including industrial and refinery effluents, will contain complex mixtures of substances. The composition of these mixtures may vary considerably over time in substances and in concentrations. It is therefore impossible to know the exact composition (chemical substances, organic carbon content, ions etc.) of an environmental or industrial sample when designing a battery of tests, especially one that can be applied routinely across different sites and at different times. A battery should therefore either comprise tests that capture as many effects and/or active substances as possible, or the main drivers, that are relevant to the type of sample that battery is being designed for, depending on the goal of the monitoring. Noting also that this must be carefully balanced with feasibility in terms of the costs and practical constraints associated with toxicity monitoring campaigns.

Design of test batteries can generally be distinguished into two main approaches [15]. The first is a "chemical-group motivated" approach that is driven by groups of 'priority' substances of concern in the sample that have a common toxic MoA. Bioassays that are specific to a MoA typically indicate that the substances producing a response are structurally similar, hence enabling identification of certain substance groups present in the sample. Utilising this approach can lead to a better understanding of the true toxic potential of the sample than chemical analysis can in isolation. Chemical analysis can only pick up specific structures tested for, but there could be many similarly acting structures not detected by analysis which all contribute additively to the effects observed in the environment. Further, some MoAs can occur via direct and indirect pathways, e.g. estrogenicity can be caused directly by binding of the estrogen receptor, or indirectly by activation of the aryl hydrocarbon receptor. Thus, including multiple bioassays covering the same MoA can capture substances that may not otherwise be detected [15]. The second approach to designing a test battery is motivated by the protection goal. For example, in surface water quality assessments, a protection goal could be protection of drinking water sources for the safety of human health, rather than safety of the environment as would be expected for monitoring of refinery effluents. Another example could be focusing on bacterial toxicity tests to predict toxicity to microorganisms in wastewater treatment plants. This approach necessitates the careful consideration of endpoints relevant for the protection goal, the exposure route, and biological target/tissue [15]. Overall, the test battery should cover different MoAs to detect the effect of as many active substances in a sample as possible, in tests deemed to be of ecological and/or toxicological relevance in the context. It should be emphasised that a test battery, or choice of tests selected for use within a given battery, will vary depending on the assessment objective of the biomonitoring program and how the generated results will be utilized, e.g. risk assessment (baseline or routine monitoring), impact prevention, toxicity screening, toxicity source identification, random spot identification of toxicity, or routine monitoring of the main driver to check deviation and adjust operation settings. Further, the tests do not all necessarily need to be deployed as one fixed array-one or a few may be sufficient in some cases, such as in the latter example.

A combination of the two approaches ("chemical-group motivated" and "protection goal motivated") was used. To be a robust bioassay battery for the assessment of toxicity in refinery effluents, the battery must 1) be relevant to the health of the receiving aquatic environment, across the trophic chain, i.e. "protection goal

motivated", and 2) be relevant to compounds typically found in such refinery effluents, i.e. "chemical-group motivated". Regarding the former, this immediately ruled out the nematode growth and reproduction test, as *C. elegans* is a predominantly soil organism, and considerably lowered the relevance of sediment-dwelling organisms and amphibians. These tests were further discounted from the final battery due to their cumbersome nature, variable sensitivity, and difficulties in interpretation. Further, the vertebrate nature of amphibians reduced their suitability for regulatory monitoring, as vertebrate use is discouraged for effluent assessment in the EU, and zebrafish assays fulfil this trophic level while being more widely used and validated.

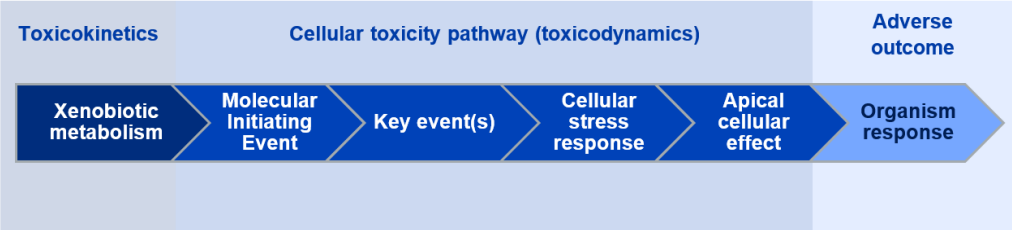
Relevance to the aquatic environment also encompassed consideration of whether marine, brackish and/or freshwater species could be used in each *in vivo* assay. For a lot of the assays, different halo-specific species could be selected depending on the sample type, for example the algae *Skeletonema costatum* can be used for brackish and coastal waters, *S. tropicum* for more oceanic waters, and *Raphidocelis subcapitata* for freshwater samples. However, for some, e.g. the *Daphnia* and zebrafish tests, only freshwater samples would be scientifically valid, reducing their applicability to samples from refineries discharging into the marine or brackish environment. Consideration of the sample salinity extended to the *in vitro* tests using fish cell lines too, such as for cytotoxicity in a rainbow trout cell line (RT Gill-W1). Osmotic stress can affect the results, however adjusting salt concentrations could reduce the bioavailability of the compounds and reduce the performance of the assay, hence it may be unsuitable for marine samples [135].

Knowing that an assay, or battery of assays, is relevant to the selected water sample of interest is necessary, but it also needs to be known if the assay(s) can detect effects that are likely to translate into adverse impacts in the selected environment. It is therefore highly beneficial to include MoAs of different stages of cellular toxicity pathways (Figure 3) for the *in vitro* assays, and whole organisms of different trophic levels for the *in vivo* assays, for the battery to have the highest likelihood of capturing ecotoxicologically relevant effects [47]. *In vivo* assays cover multiple toxicity pathways and can relate to the ultimate step in the adverse outcome pathway (AOP), the adverse outcome, i.e. an apical endpoint at the organism level. On the other hand, *in vitro* assays cover the toxicokinetic and toxicodynamic steps of the AOP, as shown in Figure 3. In general, this means that *in vitro* effects can translate into effects seen at the *in vivo* level, e.g. estrogenic activity in *in vitro* assays has been correlated with *in vivo* estrogenic activity [260], although for some genotoxicity assays measuring a reversible stress reaction this link is not seen. It is also acknowledged that most assays cannot be used to directly inform an AOP, only provide associated inferences.

The first step of the AOP, the toxicokinetics, relate to uptake and metabolism, whether this be activation or detoxification. Examples of assays covering this stage include the AhR assays. In the next stage, i.e. the toxicodynamics, the molecular initiating event (MIE) starts the cellular toxicity pathway. This comprises inhibition of enzymes, binding to hormone receptors (e.g. as in the ER and AR activation assays), and covalent interaction with DNA or proteins [47]. This then leads to a key event, or intermediate effect, such as disruption of cellular homeostasis, alternation of hormone signalling pathways, and DNA damage or mutagenicity, such as is measured in the Ames assays. It should be noted that in some cases this is an adaptation reaction and only a reversible response, e.g. for genotoxicity assays measuring primary DNA damage that can be repaired, rather than irreversible damage, e.g. for genotoxicity assays measuring gene mutations and chromosome aberrations. The intermediate effects incite a cellular stress response in the form of inflammation, oxidative stress response (e.g. AREc32 or Nrf2-pathway activation)

or p53-mediated DNA repair. Apical effects follow, such as cytotoxicity (e.g. MicroTox®), which leads to the adverse outcome ultimately observed in whole organisms. A test battery comprising tests along this pathway thus increases the ecotoxicological relevance of the endpoints measured.

Figure 3: Schematic showing different stages of the Adverse Outcome Pathway (AOP). Adapted from Neale et al. [47].

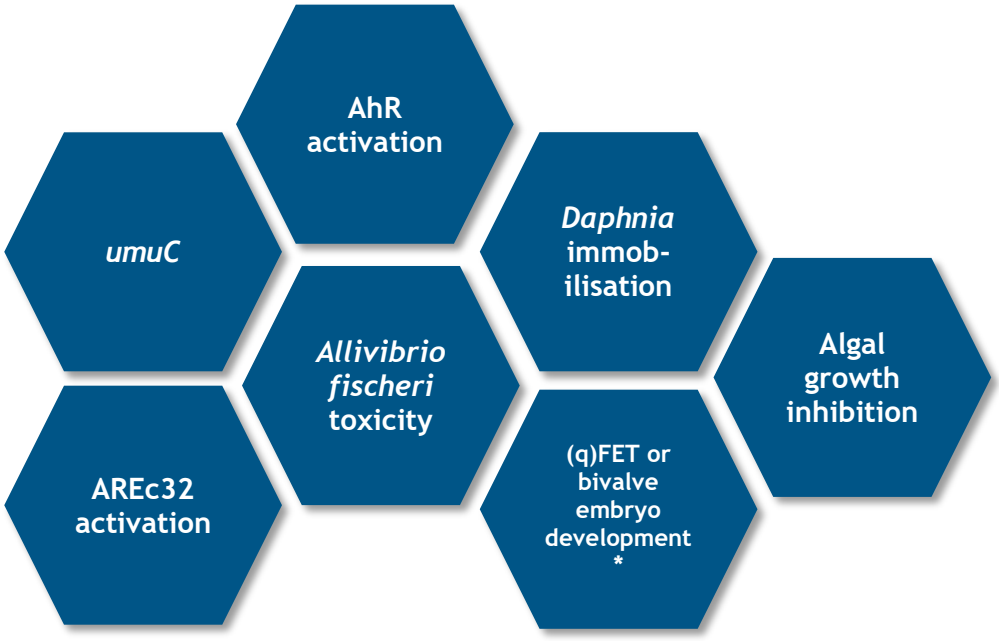


Regarding relevance to the constituents typically found in effluents related to the petrochemical refinery sector, this is much harder to define owing to the complex and often variable composition of different effluents [14]. Some assays are highly specific to certain substance types, notably the PAH CALUX® and DR CALUX®. A battery consisting of all specific assays, however, neglects non-specific assays, e.g. those measuring genotoxicity, likely allowing non-assessed compound classes to go undetected [15]. This is particularly true of refinery effluents, which although comprise various PAHs that can elicit responses in specific assays, normally consist of additional compounds such as other organic substances (e.g. phenols, esters, amides, alkanes etc.), metals, total nitrogen, sulphides, cyanides and other inorganics, to name a few [14].

7.2. TEST BATTERY SELECTION

There are several factors to consider when designing a test battery, with no single universal solution, even for specific industries with highly similar processes. The design of the battery, and the utilisation of one or more of the tests within it, will change according to the nature of samples, the assessment objective and protection goal of the (regulatory) monitoring campaign. Economic and ethical practicalities may further constrain the number and type of tests that can be conducted. Taking these factors into consideration we present a battery of suitable bioassays, shown in Figure 4 below. All tests included are available commercially. Where a test is suited to only one type of water sample or receiving water body (e.g. only freshwater), an alternative has been recommended.

Figure 4: Battery of *in vitro* and *in vivo* bioassays suggested. * the assay used will be selected based on the type of water in the sample; the FET for freshwater samples or the bivalve embryo.



The suggested battery includes:

- ***umuC***. This is an *in vitro* genotoxicity assay that relates to the key event/intermediate effect step of the AOP. The *umuC* assay has already been applied to various environmental samples, including in a regulatory context and for refinery effluents [30]. While it does not show potency of effect, neither do any other genotoxicity assays, yet it does measure a molecular initiating event that has the potential to cause an adverse outcome at the organism level. Exclusion based on lack of potency would discount all genotoxicity assays. The *umuC* assay is ISO standardised, highly validated and the most sensitive and mature of the genotoxicity assays. Unlike some other genotoxicity assays, it has few confounding factors.
- **AhR activation.** This is an *in vitro* metabolism assay that relates to the metabolic/toxicokinetics stage of the AOP. It is highly relevant and sensitive to refinery effluent constituents, for which its use has been documented [36]. A number of factors have the possibility of confounding results, but these can all be reduced by e.g. an acid clean-up step, changing the exposure time etc. The PAH-specific assay is not considered sufficiently validated enough currently to be included.
- **AREc32 activation.** This is an *in vitro* oxidative stress assay that relates to the cellular stress response part of cellular toxicity pathway. The AREc32 activation assay is one of two assays considered for oxidative stress. It is more sensitive than the Nrf2-pathway activation assay, especially for PAHs. It has been widely applied to environmental water samples, including wastewater effluents, but no information on its use with refinery effluents could be found.

- ***Allivibrio fischeri* toxicity (MicroTox®).** This is an *in vivo* bacterial assay that relates to the apical effect stage of the cellular toxicity pathway in the AOP. The MicroTox® assay is used routinely, including in a regulatory context under the IED, and to test refinery effluents. It is highly sensitive, highly validated, ISO standardised and produces results rapidly. Further, bacterial toxicity covers a stage of the AOP that none of the other evaluated *in vivo* assays do.
- **Algal growth inhibition.** This is an *in vivo* algal assay that covers the whole organism response, of the AOP. The assay is used routinely, including in a regulatory context under the IED, and to test refinery effluents. It is ISO standardised, highly validated, and its confounding factors can be mitigated with adequate sample preparation. It is applicable to marine, brackish, and freshwater samples. Further, the algal test has a short duration (72 hours), but the EC10 generated is considered a chronic endpoint [61]. While other chronic tests were available, the associated length and cost of these meant that the algal test was preferred for the purpose of routine monitoring.
- ***Daphnia magna* immobilisation.** This is an *in vivo* invertebrate assay that covers the AO of the AOP. The *D. magna* immobilisation assay is used routinely, including in a regulatory context under the IED, and to test refinery effluents. It was the most commonly used assay in a survey of European refinery effluents [32]. It is highly sensitive, highly validated, ISO standardised and has trigger values proposed. *D. magna* is a freshwater species, however, in the survey of oil refineries, the *Daphnia* test was applied to marine/brackish samples too, so its applicability may not be as limited as previously thought. The immobilisation test was selected over the *Daphnia* reproduction assay as it can be applied to passive sample extracts, generates results quickly, and is cheaper. Acute tests with *C. dubia* may also be used here instead of with *D. magna*, although generally it is within chronic tests that *C. dubia* is seen as advantageous over *D. magna,* due to a much shorter test duration (7 vs 21 days).
- **Fish Embryo Toxicity (FET/qFET).** This is an *in vivo* fish embryo assay that covers the AO of the AOP. It measures a different endpoint to that of the other AO included in the battery (growth inhibition in algae), in the form of developmental toxicity. The assay is ISO standardised, used routinely, has been used under the IED, has been applied to refinery effluents, and covers a highly relevant and sensitive endpoint. The qFET would be preferred for the purpose of toxicity monitoring owing to the small sample volumes required.
- **Bivalve embryo development.** This is an *in vivo* invertebrate assay that covers the AO of the AOP. The bivalve test was preferred to the echinoderm version owing to its higher validation maturity, availability of an ISO standard, higher frequency of use, including with refinery effluents, and its applicability to both marine and freshwater organisms. It was included to be used interchangeably with the (q)FET based on the salinity of the water sample used. The (q)FET can only be used with freshwater, whereas the bivalve assay can be used with freshwater, brackish water or marine water depending on the species utilised.

Further justification for inclusion of the suggested bioassays, and for exclusion of those not included, are provided in the sections below.

7.2.1. Maturity and performance of battery

Maturity and performance of a test can be evaluated by looking for tests with test guidelines standardised to ISO guidelines. This means the performance of the assays have been validated in an inter-laboratory ring test evaluation and reproducibility and variability of the test are within acceptable limits. All assays proposed have an

ISO standardised test guideline or it is currently in preparation, such as the AREc32 and AhR activation assays. The algal growth inhibition test has two ISO test guidelines for freshwater (ISO 8692) and marine water (ISO 8692). When debating the inclusion of the bivalve versus the echinoderm embryo development test, the lack of ISO guideline for the echinoderm embryo development test was a contributing factor, although it is standardised to ASTM E1563-21a. Its lower validation maturity score ("medium") also favoured the use of the bivalve version of the test, which is more mature.

The *in vivo* tests were all scored "high" against validation maturity. This means they had been validated at the routine level (inter-laboratory validation) [33]. The only suggested *in vitro* assay to not score "high" was the AREc32 activation assay at "medium", however it is more mature than the other oxidative stress assay evaluated (Nrf2-pathway activation assay). Similarly, the PAH-specific AhR activation was not chosen over the standard AhR activation assay as it is currently not well validated enough for use in routine monitoring.

Considering sensitivity, *in vitro* battery assays are very sensitive, and, crucially, higher than their alternatives within the same MoA. For example, the *umuC* assay is more sensitive than the Ames assay [29, 40, 237]. As expected, the sensitivity of some tests can vary depending on the strain or cell line used, so this should be considered when designing the tests [39]. MicroTox® and the *Daphnia* immobilisation assays are highly sensitive to a wide range of compounds [41], and the algal growth inhibition is considered to have greater sensitivity compared with algal PAM for many substances [62]. The sensitivity of the FET assays is generally high, however their sensitivity to hydrocarbons and refinery effluents is typically lower than for tests with algae or *Daphnia*. Less is known about the sensitivity of the bivalve embryo development test, but embryos are known to provide sensitive indicators of early life stage effects [40].

7.2.2. Use of battery

The suitability and applicability of tests for use within monitoring of refinery effluents and produced water can be gleaned from their history of use in a) a regulatory context, b) in environmental samples, c) with refinery effluent constituents including oils, and d) with whole refinery effluent samples.

All proposed tests are routinely used for the assessment of environmental water samples. The *in vivo* tests are all used routinely and in a regulatory context, such as under the IED. Of the *in vitro* assays recommended, only *umuC* has been applied in a regulatory context thus far. This reflects a general lack of use of *in vitro* assays in a regulatory context, rather than the unsuitability of the other assays for this purpose. The only other evaluated *in vitro* assays to have been used for regulation are the YES/YAS assays.

The evaluation criteria applied in this project (use in regulation, use in environmental samples, use with refinery effluent constituents, use with refinery effluents) provide an insight into the use of the assays for regulatory monitoring of aquatic samples. To further understand their applicability to the refinery sector, their responsiveness to refinery effluent constituents and refinery effluents was also evaluated. It is important to note that use with refinery effluent constituents and refinery effluents were evaluated distinctly, as not all of the many constituents of refinery effluents could be evaluated separately [14].

All suggested assays have all been used to test hydrocarbons; further, they have all elicited responses in these tests. For example, the *umuC* assay requires metabolic

activation but is then responsive to B(a)P [147] and other metabolically activated PAHs [150]. All other suggested assays have also produced responses in tests with B(a)P and other PAHs. Hydrocarbons are not the only constituent of refinery effluents, and so other constituents were considered, literature permitting. For example, the MicroTox® assay is very sensitive to (heavy) metals (e.g. zinc, mercury, copper, cadmium, lead, nickel) [51, 52].

Literature describing the use of the tests with refinery effluents is notably scarcer than for hydrocarbons, metals, surface waters, industrial effluents and other water sources. No information could be found for the *umuC* and AREc32 assays. However, they are responsive to crude oil, refined petroleum products and/or heavy fuel oil residues in surface waters [152, 218, 223, 258]. The remainder of the recommended test battery have been used for testing refinery effluents. The AhR assay was used to test 22 produced water samples from oil and gas installations on the UK Continental Shelf, of which all elicited a response before clean-up, and 13 after, although refinery effluent composition will differ from that of produced water [36]. Surveys conducted by Concawe demonstrate that the MicroTox®, *Daphnia* immobilisation and the algal growth inhibition assays have commonly been used to test effluents of European refineries [32]. This is further supported by the academic literature, such as Whale *et al*. [43], Comber *et al*. [53], Whale *et al*. [5]. The bivalve embryo development test has also been proven for use with refinery effluents in a marine context [91].

An assay that was excluded from the considered test battery based on its use was the ER activation assay. Of the two estrogen assays, it was preferred above the YES assay owing to its higher sensitivity and wider use, particularly with oils, PAHs, produced water and refinery effluents. However, while responses were reported in the literature, they were considered low potency and/or unlikely to produce a response when using whole refinery effluent samples.

For example, solid phase extracts of produced water from an oil production facility and refinery effluent from a land-based oil refinery contained chemicals able to induce estrogenic effects in a rainbow trout assay [279]. The potency of such induction was low, at approximately 40 ng estrogen equivalents (ng 17 ß-estradiol /L), whereas the control was approximately 25 ng EE and untreated refinery water was near 150 ng EE. 4-tert-butylphenol was suggested to be contributing to the activity, however it has a much lower displacement potency than 17 ß-estradiol, by 4-5 orders of magnitude [279]. Also, solid-phase extracts were tested rather than the actual effluent, so the relevance of the results is further questioned. In another study, ER activity was reported when using the dissolved phase (attributed to methyl- to nonyl-substituted alkylphenol isomers) of produced water, but no activity when using filtered oil droplets [242]. Investigations using four crude oils and seven refined oils found activity to the ERß receptor in yeast screens in all oils (no activity in the ERa), and activity of three crude oils and five refined oils to both receptors in ER-LUC assays [240, 258]. The calculated estrogenic potencies of the oils were six to nine orders of magnitude lower than the potency of 17 ß-estradiol [258]. Other publications also found weak estrogenic activity of produced water extracts [36], and of B(a)P and benzo(b)fluoranthene in a human cell line ER activation assay (MELN) with no activity in zebrafish cell line assays (ZELH-zfERalpha and ZELHzfERbeta2) [47]. A recent study found that naturally and chemically dispersed crude oil WAFs activated the ERα in A-YES and ERα-CALUX® assays [241]. Results obtained from the ERα-CALUX® assay indicate that both the WAF free from droplets and the chemically dispersed WAF interact with the ERα in both U2-OS and T47D cells. For both bioassays the estradiol equivalents were above the EBT values, but still much lower potency than 17 ß-estradiol. Further, these WAFs were high concentration so the relevance may be questioned.

Overall, while some refinery-relevant samples show estrogenic activity, potency is generally low [210]. It is therefore considered unlikely that concentrations in refinery effluents would be high enough or potent enough to elicit a response in an estrogen receptor activation assay.

7.2.3. Conduct of battery

Test duration can have a significant bearing on the ease of use and application of a test in routine monitoring. The *in vitro* assays recommended here all have exposure times of equal to or under 24 hours. All five *in vivo* assays are also relatively quick, with exposures of equal to or under 96 hours. While longer than the *in vitro* assays, this is rapid when compared to other *in vivo* assays, such as the amphipod (*Corophium volutator*) test which has a test duration of 10 days and the *Daphnia* reproduction test (7 days). MicroTox® is known for its rapid exposure time of up to 30 minutes. It is also important to consider the duration of sample shipping and the duration to produce and distribute the results, although this will be site- and laboratory-specific.

Several recent studies describe the use of effect-based methods in conjunction with passive sampling techniques, to facilitate time-integrated water quality monitoring. With one exception, all the assays recommended here are applicable to passive sample extracts. The use of passive sample extracts with this the bivalve embryo development test is anticipated to be difficult. If a bioassay suitable for use with passive sampler extracts is required, the (q)FET could be used instead as this has been proven fit for purpose [72]. Compatibility with passive sampling was not one of the highest priorities when evaluating the tests.

While it is useful to note that the recommended battery is applicable to passive sampler extracts, direct testing of refinery effluents may be preferred in some instances. As discussed in Section 7, passive sampling extracts provides a time weighted average sample. This offers the advantage of increasing the likelihood of detection of transient pollution events and providing a more representative sample over time. In addition, use of passive sampling extracts may increase the sensitivity of both chemical analysis and bioassays since contaminants are typically preconcentrated within the sampler. Whilst this offers a benefit for cleaner environmental samples such as groundwater or drinking water with much lower concentrations of contaminants it may not be necessary for refinery effluents.

The composition of a passive sampling extract will not be the same as a spot sample and this must be considered when interpreting results. Sample composition will vary according to the type of sampler, the physico-chemical properties of the substances present and the time the sampler is deployed. Commonly, more than one type of sampler is deployed at the same location simultaneously to capture both hydrophobic and hydrophilic contaminants. Fractionation of the sample in this manner may provide an additional useful line of evidence in the interpretation of any effects observed and elucidation of which components effects may be attributed to. However, partitioning and the subsequent potential to not capture certain components on a passive sampler is contradictory to the whole effluent approach.

Conduct of the battery also comprises the cost per sample. As few survey respondents conducted *in vitro* assays, cost estimates for most had to be based on those found in the literature. The *umuC* and AREc32 assays are both likely to be <100 € [34], whereas the AhR assay is slightly more expensive at 100-1,000 € [34] or <200 € [28]. However, according to the survey results, the AhR assay is likely to be 300 - <1,000 € per sample. Evidently, the literature values may underestimate the

costs observed for laboratories. Less cost information was available in the literature for the in vivo assays. The literature reports a cost of <200 € for the FET assay [28]. However, results of the survey indicate the assay is likely to be 1,000 to above 10,000 € per sample; the most commonly reported cost was 3,000 - <10,000 €. According to the survey, the MicroTox® assay is likely to be under 1,000 € per sample, the algal growth inhibition assay is likely to be under 100 to under 10,000 €, the Daphnia immobilisation assay is likely to be above 100 to under 10,000 €, and the bivalve embryo development assay is likely to be above 1,000 to under 10,000 €. One of the reasons the Daphnia immobilisation assay was recommended over the reproduction assay was cost; the latter was the most expensive in vivo assay in the survey, most frequently above 10,000 € (55% of responses). The overall cost of the battery will be informed by the frequency of the monitoring, the tests utilised, and the existing facilities/capacity of the laboratories.

7.2.4. Results interpretation of battery

7.2.4.1. Confounding factors

The performance of many bioassays is dictated by the concentration of the responsive substances. If the concentration is too high cytotoxic effects may occur and negate any response that may otherwise be seen. If the concentration is too low, there will be insufficient substrate to induce a measurable response. The breadth of this 'goldilocks' window can determine how useful an assay is. No matter how wide this window is, results can be confounded by other entities present in the sample and decrease, or in some cases completely diminish, the potency of response observed. As seen in Sections 4 and 5, responses can be confounded by other entities present in the sample, for example COD, antagonists, particulate matter etc., and other non-sample factors such as the cell line used, duration of test, absence of metabolic activation etc. Recognising that these confounding factors can affect data quality is critical when assembling a test battery. Tests deemed unreliable can be excluded from the battery entirely, and tests with known weakness can be complemented with further tests and appropriate sample preparation. Understanding the influence of confounding factors is critical for interpretation of test results.

The *umuC* assay has few confounding factors compared to other well-validated genotoxicity assays such as the Ames, micronucleus, and comet assays. The only confounding factor identified was the species used in relation to the type of water in the sample. *Salmonella* sp. is a halo-sensitive freshwater species that would be unlikely to survive in samples of produced water from offshore installations [40]. This is also a confounding factor for the Ames/Ames II tests, so was not considered to be sufficient reason for excluding the *umuC* assay from the battery. Further, the Ames/Ames II and remaining genotoxicity assays had a higher number of confounding factors influencing results observed when compared with the *umuC* assay. These included the strain used in the Ames tests, microbial contamination, masking effects, subjective evaluations, cell line used, particulate matter and cytotoxic effects in some of the other tests.

The AhR activation assay has a comparable number of confounding factors to the two other metabolism assays. For example, Machala *et al.* [220] found that a shorter exposure time to PAHs led to significantly higher activity than the longer exposure. However, these differences can be avoided if a standard exposure time is employed. In addition, tryptophan metabolites may be present in fresh growth medium. These metabolites can also activate the AhR and result in an overestimation of effects of the sample itself [17, 218]. This can be mitigated by use of a different medium. Also, while the risk of cytotoxicity in the AhR activation assay is not zero, use of a

sulphuric acid clean-up step during sample preparation minimises the risk, as reflected in a lack of cytotoxicity in Vrabie *et al*. [218] and Ziccardi *et al*. [227]. The PAH-AhR assay has similar challenges. In contrast, the EROD assay is associated with confounding factors that are difficult to mitigate or account for, such as EROD activity being affected by metals, xenoestrogens, water temperature etc. [28, 128].

One of the reasons that the AREc32 assay was included in the suggested battery over the Nrf2-pathway activation assay was the lack of confounding factors. Cytotoxic effects may occur, but this is seen in a lot of assays [47, 200]. In contrast, the results of the Nrf2-pathway activation assay are confounded for substances relevant to refinery effluents, for example, B(a)P is very responsive in the AREc32 assay [200] but not in the Nrf2-pathway activation assay, despite the tests utilising the same pathway [194]. It is putatively suggested that the use of U2OS cells in the Nrf2-pathway activation may be the root of these differences [215, 216].

The recommended *in vivo* tests have fewer known confounding factors. The responses of the MicroTox® and algal growth inhibition assays were both noted to be inhibited by COD, although this could be avoided by sample preparation [53]. The *Daphnia* and (q)FET assays could also potentially be confounded by constituents other than pollutants and by pH or oxygen content, particularly in cases where the native sample is used [29]. Neale et al. [47] reported lower EC50 values in the FET in tests conducted in glass vials compared with tests run in 96-well polystyrene plates. A trade-off between high throughput capability and test set-up may be required in some instances.

Common to all tests will be any confounding factors related to the choice of sampling method used. These are discussed in Section 6. Briefly, when passive sample extracts are used the relevance or representativeness of the sample compared to the effluent can be questioned, owing to the enriched, oily nature of these extracts and lack of various effluent constituents (e.g. salts). Meanwhile, whole effluent samples will include all constituents present at the time of sampling, but these can be unfavourable in assays at the undiluted level (e.g. high salinity, low nutrient concentration). Whichever sampling method is used, it is recommended to test and account for any confounding sources as far as is possible.

7.2.4.2. Availability of Effect-Based Trigger values

If a sample produces a response in an assay, it is critical to know what this means. A response may not always infer unacceptable water quality. Effect based trigger values (EBTs) have been developed, albeit by few studies, to aid interpretation of test data and determine if a quantitative bioassay response is of concern [15]. It should be highlighted that EBTs are not formally acknowledged but there is a need to define them and to make a link with *in vivo*, organism-level adverse effect. Generally, EBTs are more readily available for the *in vitro* assays than the *in vivo* tests evaluated in this work. All the assays bar one has EBTs proposed in the literature. The bivalve embryo development assay does not currently have an EBT, although methods used to calculate other *in vivo* EBTs could be applied.

7.2.4.3. Relevance of results to refinery effluents

EBTs are numerical indicators of what a response in an assay may mean in terms of effects in the environment. The relevance of effects to the aquatic environment and to refinery effluents was also considered for the battery of assays. An assay measuring genotoxicity was included. It is a sub-cellular effect - sometimes deemed less ecologically relevant that those further along the AOP. However, a positive result is generally highly indicative of carcinogenesis risk, of high ecological and toxicological importance if it is indeed observed in the environment downstream

the effluent discharge. The AhR activation assays are highly relevant to AhR receptor agonists, such as dioxins and dioxin-like substances [29, 34]. The assays are just receptor mediated, as opposed to upregulation as seen in the EROD assay, but the EROD assay has been deemed less suitable for inclusion in the assay for other reasons [27]. The AREc32 assay measures oxidative stress, which gives an early indication of higher tier effects such as impaired immune response, reproduction, cellular repair mechanisms etc. [154], thus was deemed ecotoxicologically relevant.

In vivo tests are automatically of high relevance due to their apical nature. All five *in vivo* tests selected are highly sensitive and have produced responses in refinery effluents. They were deemed relevant to the toxicity testing of refinery effluents. Moreover, they represent different endpoints. MicroTox® measures apical cellular effect, a different stage of the AOP to the other *in vivo*, and *in vitro*, tests included in the battery. The algal assay and *Daphnia* and (q)FET/bivalve assays all measure an AO in a whole organism - growth inhibition, immobilisation, and developmental toxicity/teratogenicity, respectively.

Teratogenicity is a highly relevant endpoint as it has the capacity to induce population level effects over multiple generations, for both humans and wildlife [40]. Another reason for inclusion of the algal test in terms of result interpretation was that it is considered chronic under REACH and CLP [61]. Longer term testing was excluded on economic and practical grounds, but the battery still covers a chronic endpoint (algal growth) and captures sublethal effects (developmental toxicity).

One notable exception from the AOP is an assay that covers the MIE of the cellular toxicity pathway. The inclusion of an endocrine disruption assay would have covered this. However, assays measuring (anti-)androgenic activity were excluded as only a relatively low number of substances cause androgenic responses, and anti-androgenic activity is typically only detected at high concentrations (Scott et al., 2014). Further, assays measuring (anti-)estrogenic activity were excluded as although ISO standardised, highly validated, and commonly used, they were deemed unlikely to cause effects in refinery effluents, and hence not relevant for the monitoring of such water.

Table 5: Overview evaluation of the toxicity tests in the considered test battery.

	umuC	**AhR activation**	**AREc32 activation**
Mode of action	Genotoxicity	Metabolism	Oxidative stress
Standardisation	ISO 13829	ISO in preparation (ISO/CD 24295)	ISO in preparation
Validation maturity	High	High	Medium
Use in a regulatory context	Yes	No	No
Use in environmental water samples	Very common	Common	Common
Responsive to refinery effluent constituents	Yes (PAHs, metals)	Yes (PAHs, oils)	Yes (PAHs, metals)
Responsive to refinery effluents	Yes	Yes	Unknown
Trigger value proposed	Yes	Yes	Yes
Possible chemical analysis replacement	No	Yes	No
Relevance	Medium	Medium	Medium

	A. fischeri toxicity	Algal growth inhibition	*Daphnia* immobilisation	q(FET)	Bivalve embryo development
Endpoint	Cytotoxicity	Growth inhibition	Immobilisation	Developmental toxicity	Developmental toxicity
Type of organism	Bacteria	Algae	Invertebrate	Fish	Invertebrate
Acute or chronic	Acute	Chronic	Acute	Acute	Chronic
Type of water	Marine, freshwater, brackish	Marine, freshwater, brackish depending on species	Freshwater	Marine, freshwater, brackish depending on species	Marine, freshwater, brackish depending on species
Standardisation	ISO 11348-3	ISO 8692 (freshwater) 10253 (marine)	ISO 6341	ISO 15088	ISO 17244
Validation maturity	High	High	High	High	High
Use in a regulatory context	Yes	Yes	Yes	Yes	Yes
Use in environmental water samples	Very common	Very common	Very common	Very common	Very common
Responsive to refinery effluent constituents	Yes (PAHs, oils, metals)	Yes (PAHs, oils, metals)	Yes (PAHs, oils, metals)	Yes (oils)	Yes (PAHs, oils, metals)
Responsive to refinery effluents	Yes	Yes	Yes	Yes	Yes
Trigger value proposed	Yes	Yes	Yes	Yes	No
Possible chemical analysis replacement	No	No	No	No	No
Relevance	High	High	High	High	High

7.3. OTHER CONSIDERATIONS

As observed with passive samplers, hydrophobic compounds can bind to equipment. This should be avoided in the context of toxicity testing as it reduces the bioavailability of the substances to the test organism and can hence reduce toxicity, as seen in the FET using polystyrene plates. It is therefore recommended to use glass test apparatus, e.g., glass vials, where possible, as many constituents of refinery effluents are anticipated to be hydrophobic. However, it is acknowledged that the use of glass apparatus may reduce the high throughput potential of some of the assays where plastic-based well plates may be used.

Further, enriching the samples by using passive sampling, or by extracting the samples using e.g., XAD® resin to remove metals and other confounding factors, is considered not appropriate under a WET approach.

8. COMPLEMENTARY METHODS

Toxicity testing of refinery effluents provides a snapshot of mixture toxicity, but does not provide information on chemical composition of the sample or evidence that can be used to investigate which components of a mixture may be responsible for observed effects, or quantification of test results in terms of dose. However, results can be strengthened further though the use of complementary methods that provide additional lines of evidence on sample composition and provide a wider context for interpretation of results. Complementary methods include biomimetic solid phase microextraction for chemical analysis and PETROTOX for bioavailability and toxicity prediction. These are discussed further below.

8.1. BIOMIMETIC SOLID PHASE MICRO EXTRACTION (BE-SPME)

Biomimetic solid phase micro extraction (BE-SPME) is an organism-free method used to quantify and extract the bioavailable fraction of organic contaminants in effluent. This includes hydrocarbons, non-polar organics and to a lesser extent, polar organics, and ionic organics. Cationic constituents and metals are out of scope. BE-SPME can be used in combination with laboratory studies, including those assessing biodegradability, to understand the fate of the constituents of the effluent [53].

BE-SPME utilises the partitioning of dissolved phase chemicals onto a polydimethylsiloxane (PDMS)-coated SPME fibre to mimic the uptake of contaminants by an organism. The method relies on the relationship between the log Kow of a chemical and its potential to bioaccumulate and its baseline toxicity (non-specific narcosis). The molar concentration adsorbed to the SPME fibre (C_{fibre}) is then analysed by gas chromatography-flame ionization detector (GC-FID) and quantified using an external standard (2,3-dimethylnaphthalene). This allows the chemical constituents of the effluent to be identified, as well as an estimation of baseline toxicity [53]. Further analysis to determine concentrations of constituents which may bioaccumulate is also possible [43, 53].

BE-SPME allows samples to be assessed without undergoing filtration and is of value when petroleum products are the main source of contaminants, as would be anticipated at refineries. In addition, the method is particularly suitable for use in whole effluent assessments as a proxy for bioaccumulation, similar to providing a quantification of the bioavailable hydrocarbon fraction, therefore providing analytical coverage of chemical exposure [5].

The selectivity of BE-SPME is comparable to extractable gas-chromatographic organic matter liquid-liquid extraction (EGOM LLE) [280]. The performance of the method may be less reliable for very water-soluble chemicals and charged substances with multiple charges that may have additional modes of uptake and toxicity, although their partitioning will still governed by their log Kow and concentration [15]. Anionic constituents will also absorb to the fibre, if the samples are acidified.

8.2. PETROTOX

PETROTOX is a model used to predict the effect of petroleum hydrocarbon substances to aquatic organisms, based on substance composition [53, 281]. The model predicts the dissolution and toxicity of petroleum substances by modelling equilibrium partitioning assuming a closed water accommodated fraction (WAF) system at steady-state [282]. The hydrocarbon block method, used for conducting

petroleum substance risk assessments [283], is used to characterise substance composition through definition of discrete hydrocarbon blocks with similar physiochemical properties. These compositional data are used together with total load (sum of individual blocks) to calculate dissolved concentration profiles for the entire substance [284].

PETROTOX uses the Target Lipid Model (TLM) [284, 285] to convert the dissolved profiles to toxic units. The TLM uses the inverse relationship between toxicity endpoint (LC50) and Kow to account for the varying toxicity of individual hydrocarbons. Non-specific narcosis (baseline toxicity) is the assumed mode of action for acute toxicity, while empirical acute-to-chronic ratios (ACRs) are used for prediction of chronic toxicity. Toxic units are the sum of the ratio of the predicted dissolved concentration for each HC block to the predicted toxicity endpoint for each block [284]. Further details are provided by Redman *et al*. [281, 283].

9. CONCLUSIONS

This project has identified and evaluated a selection of toxicity testing methods for their applicability and suitability for toxicity monitoring of refinery effluents discharging to the aquatic environment, such as may be required in future EU legislation.

A total of 13 *in vivo* tests covering a range of trophic levels and endpoints, and 18 *in vitro* tests covering cytotoxicity, genotoxicity, oxidative stress, metabolism, and endocrine disruption have been evaluated against specific criteria. These criteria were: commercial availability, application of test in a regulatory context, standardisation, validation maturity, sensitivity, potency of effects, use in environmental samples, use with refinery effluent constituents, use with refinery effluents/produced water, applicability to passive sampler extracts, time to run, cost per sample, relevance, availability of trigger values, bioassay category, chemical analysis replacement, and confounding factors. In addition, toxicity testing using passive sample extracts, as well as complementary techniques such as biomimetic solid-phase microextraction and prediction methods have been discussed.

The last decade has seen the development of various bioassay test batteries for effect-based monitoring of environmental water quality and associated guidance. There is no universally applicable test battery, even within a specific industry; they should be sample-specific to be successfully applied. A fit-for-purpose battery should include tests that capture as many effects and/or active substances as possible that are relevant to the nature of the sample.

Refinery effluents are complex and often variable in nature, comprising aliphatic and aromatic hydrocarbons, other organics, nitrogen, phosphorus, sulphides, metals to name a few. A suitable test battery must be applicable to measurement of effects in the aquatic environment including marine, brackish and freshwater environments. A test battery should comprise multiple tests to increase the likelihood of capturing a wide range of effects and substances in the sample. We have proposed a battery of suitable tests, although this does not necessarily mean that all tests must be deployed together. We do not provide recommendations on the frequency of monitoring as this is out of scope of the current project. Further work may also include a comparison of assay response sensitivities to known contaminant concentrations in refinery effluents.

The suggested test battery resulting from the evaluation is summarised in Figure 4. These are the *umuC*, AhR activation, AREc32 activation, *Allivibrio fischeri* toxicity, algal growth inhibition assays, *Daphnia* immobilisation and (q)FET or bivalve embryo development assays. The latter two are considered interchangeable depending on the type of water sample tested and the relevance of the receiving water body; both tests provide assessment of sublethal developmental effects in secondary consumers. The battery relates to 3 *in vitro* MoAs (genotoxicity, metabolism and oxidative stress) as well as apical *in vivo* endpoints (cytotoxicity, developmental toxicity, immobilisation and growth inhibition), including chronic endpoints in the form of the algal test and bivalve embryo development test. They cover all but one stage (the MIE) of the AOP. The choice of chronic or acute tests within an applied battery will be closely related to the specific goal of the routine biomonitoring.

The suggested battery includes at least three trophic levels (algae, crustaceans, and fish/bivalve embryos) to address the main exposure routes in aquatic ecosystems. Furthermore, and of great importance, all tests are commercially available (some with more limited availability than others), commonly used for the assessment of environmental water samples, sufficiently validated, standardised to an ISO guideline (or one is currently in preparation) and demonstrated for use in refinery-relevant samples and/or refinery effluents themselves. The applied test battery can consist of one or more assays depending on the assessment objective (e.g. full receiving water risk assessment, monitoring toxicity trends over time etc.), the protection goal of the monitoring campaign, the type of sample and receiving water, and the activity undertaken (e.g. routine monitoring, full site risk assessment, detailed composition investigation etc.).

10. GLOSSARY

AhR	Aryl hydrocarbon Receptor
AOP	Adverse Outcome Pathway
AR	Androgen Receptor
BAT	Best Available Techniques
BAT-AELs	BAT-Associated Emissions Levels
BAT-AEPLs	BAT-Associated Environmental Performance Levels
BE-SPME	Biomimetic Solid Phase Micro Extraction
BOD	Biochemical Oxygen Demand
BREF	Best Available Techniques (BAT) Reference document (BREF)
BTEX	Benzene, Toluene, Ethylbenzene and Xylenes
CIS	Common Implementation Strategy
COD	Chemical Oxygen Demand
CWW	Common Waste Water and Waste Gas Treatment/Management Systems in the Chemical Sector
EBM	Effect-Based Method
EBT	Effect-Based Trigger value
ER	Estrogen Receptor
EROD	7-ethoxyresorufin-O-deethylase
FET	Fish Embryo Toxicity
IED	Industrial Emissions Directive
KEI(s)	Key Environmental Issue(s)
MIE	Molecular Initiating Event
MoA	Mode of Action
Nrf2	Nuclear factor erythroid 2-related factor 2
PAH	Polycyclic Aromatic Hydrocarbons
REF (BREF)	Refining of Mineral Oil and Gas (BREF)

SPME	Solid Phase Micro Extraction
TOC	Total Organic Carbon
TPH	Total Petroleum Hydrocarbons
TSS	Total Suspended Solids
TTR	Transthyretin
TU	Toxic Unit
WAF	Water Accommodated Fraction
WET	Whole Effluent Toxicity
WWTP	Waste Water Treatment Plant

11. ACKNOWLEDGEMENTS

The authors would like to acknowledge Concawe's Water, Soil and Waste Management Group and the members of STF-32 and STF-34 groups for the useful discussions and feedback.

12. REFERENCES

1. JRC, COMMISSION IMPLEMENTING DECISION of 9 October 2014 establishing best available techniques (BAT) conclusions, under Directive 2010/75/EU of the European Parliament and of the Council on industrial emissions, for the refining of mineral oil and gas. 2014.

2. JRC, Joint Research Centre. Best available techniques (BAT) reference document for the refining of mineral oil and gas industrial emissions. Industrial Emissions Directive 2010/75/EU (integrated pollution prevention and control). 2015.

3. JRC, Joint Research Centre Reference Report on Monitoring of Emissions to Air and Water from IED Installations. 2018.

4. Ricardo, Preliminary determination of Key Environmental Issues (KEI) for industrial sectors in BREF reviews under the IED. Report for the European Commission. 2018.

5. Whale, G.F., et al., Assessment of oil refinery wastewater and effluent integrating bioassays, mechanistic modelling and bioavailability evaluation. Chemosphere, 2022. **287**: p. 132146.

6. Concawe, 2016 Survey of Effluent Quality and Water Use at European Refineries. Report 10/20. 2020: Brussels.

7. ECHA, Guidance on Information Requirements and Chemical Safety Assessment - Chapter R.7b: Endpoint specific guidance. 2017.

8. Norberg-King, T.J., et al., An International Perspective on the Tools and Concepts for Effluent Toxicity Assessments in the Context of Animal Alternatives: Reduction in Vertebrate Use. Environmental Toxicology and Chemistry, 2018. **37**(11): p. 2745-2757.

9. JRC, Joint Research Centre. Best Available Techniques (BAT) Reference Document for Common Waste Gas Management and Treatment Systems in the Chemical Sector. Industrial Emissions Directive 2010/75/EU (Integrated Pollution Prevention and Control). 2019.

10. JRC, Best Available Techniques (BAT) Reference Document for Common Waste Water and Waste Gas Treatment/Management Systems in the Chemical Sector. 2016.

11. Radelyuk, I., et al., Oil refinery and water pollution in the context of sustainable development: Developing and developed countries. Journal of Cleaner Production, 2021. **302**: p. 126987.

12. OSPAR, Risk-Based Approach to the Management of Produced Water Discharges from Offshore Installations. 2013.

13. Carere, M., et al., Technical Proposal for Effect-Based Monitoring and Assessment under the Water Framework Directive. Report to the Common Implementation Strategy (CIS) Working Group Chemicals on the outcome of the work performed in the subgroup on Effect-Based Methods (EBM). 2021.

14. Concawe, Assessment of chemical water emission parameters for refinery discharges. in prep.

15. Escher, B., P. Neale, and F. Leusch, Bioanalytical Tools in Water Quality Assessment. 2021.

16. Bolden, A.L., et al., Polycyclic aromatic hydrocarbons and female reproductive health: A scoping review. Reproductive Toxicology, 2017. **73**: p. 61-74.

17. Vogel, C.F.A., et al., The aryl hydrocarbon receptor as a target of environmental stressors - Implications for pollution mediated stress and inflammatory responses. Redox Biol, 2020. **34**: p. 101530.

18. Patel, A.B., et al., Polycyclic Aromatic Hydrocarbons: Sources, Toxicity, and Remediation Approaches. Frontiers in Microbiology, 2020. **11**.

19. Sekiguchi, T., et al., Effect of Polycyclic Aromatic Hydrocarbons on Development of the Ascidian Ciona intestinalis Type A. Int J Environ Res Public Health, 2020. **17**(4).

20. Vrinda, S., et al., Neuroendocrine and immunotoxicity of polyaromatic hydrocarbon, chrysene in crustacean post larvae. Ecotoxicology, 2019. **28**(8): p. 964-972.

21. Balali-Mood, M., et al., Toxic Mechanisms of Five Heavy Metals: Mercury, Lead, Chromium, Cadmium, and Arsenic. Front Pharmacol, 2021. **12**: p. 643972.

22. Jaishankar, M., et al., Toxicity, mechanism and health effects of some heavy metals. Interdiscip Toxicol, 2014. **7**(2): p. 60-72.

23. Ungureanu, E. and G. Mustatea, Toxicity of Heavy Metals. 2022.

24. Akhtar, M.F., et al., Textile industrial effluent induces mutagenicity and oxidative DNA damage and exploits oxidative stress biomarkers in rats. Environmental Toxicology and Pharmacology, 2016. **41**: p. 180-186.

25. Xiao, X., et al., Responses and tolerance mechanisms of microalgae to heavy metal stress: A review. Marine Environmental Research, 2023. **183**: p. 105805.

26. European, C., et al., Modes of action of the current priority substances list under the Water Framework Directive and other substances of interest. 2018: Publications Office.

27. Concawe, Review of options for refinery effluent assessment using effect-based tools in combination with passive samplers. Report 18/14. 2018: Brussels.

28. Wernersson, A.-S., et al., Technical report on aquatic effect-based monitoring tools. Office for Official Publications of the European Communities, 2014.

29. Brack, W., et al., Effect-directed analysis supporting monitoring of aquatic environments — An in-depth overview. Science of The Total Environment, 2016. **544**: p. 1073-1118.

30. Concawe, Review of effect-based methods in the context of refinery effluents. in prep.

31. Kokkali, V. and W. van Delft, Overview of commercially available bioassays for assessing chemical toxicity in aqueous samples. TrAC Trends in Analytical Chemistry, 2014. **61**: p. 133-155.

32. Concawe, A review of toxicity testing conducted on European refinery effluents in 2010, 2013, 2016, and 2019. in prep.

33. NORMAN, Network of reference laboratories and related organisations for monitoring and bio-monitoring of emerging environmental pollutants. 2009.

34. Schriks, M., et al., Selection criteria to select in vitro bioassays for implementation and use. DEMEAU report. 2015.

35. Wieczerzak, M., J. Namieśnik, and B. Kudłak, Bioassays as one of the Green Chemistry tools for assessing environmental quality: A review. Environment International, 2016. **94**: p. 341-361.

36. Balaam, J.L., et al., Identification of nonregulated pollutants in north sea-produced water discharges. Environmental Toxicology and Chemistry, 2009. **28**(6): p. 1159-1167.

37. Concawe, Effects based methods (EBMs) in combination with passive sampling of refinery streams. Report 11/22. 2022.

38. van der Oost, R., et al., SIMONI (smart integrated monitoring) as a novel bioanalytical strategy for water quality assessment: Part i-model design and effect-based trigger values. Environ Toxicol Chem, 2017. **36**(9): p. 2385-2399.

39. Escher, B.I., et al., Effect-based trigger values for in vitro and in vivo bioassays performed on surface water extracts supporting the environmental quality standards (EQS) of the European Water Framework Directive. Sci Total Environ, 2018. **628-629**: p. 748-765.

40. Colvin, K.A., C. Lewis, and T.S. Galloway, Current issues confounding the rapid toxicological assessment of oil spills. Chemosphere, 2020. **245**: p. 125585.

41. Leusch, F.D.L., et al., Assessment of wastewater and recycled water quality: A comparison of lines of evidence from in vitro, in vivo and chemical analyses. Water Research, 2014. **50**: p. 420-431.

42. Komori, K., et al., A rapid and simple evaluation system for gas toxicity using luminous bacteria entrapped by a polyion complex membrane. Chemosphere, 2009. **77**(8): p. 1106-12.

43. Whale, G., et al., Assessment of refinery effluents and receiving waters using biologically-based effect methods. CONCAWE Reports, 2012.

44. Abbas, M., et al., Vibrio fischeri bioluminescence inhibition assay for ecotoxicity assessment: A review. Science of The Total Environment, 2018. **626**: p. 1295-1309.

45. Muller, R., et al., Combining passive sampling and toxicity testing for evaluation of mixtures of polar organic chemicals in sewage treatment plant effluent. J Environ Monit, 2007. **9**(1): p. 104-9.

46. Völker, J., et al., Systematic Review of Toxicity Removal by Advanced Wastewater Treatment Technologies via Ozonation and Activated Carbon. Environmental Science & Technology, 2019. **53**(13): p. 7215-7233.

47. Neale, P.A., et al., Development of a bioanalytical test battery for water quality monitoring: Fingerprinting identified micropollutants and their contribution to effects in surface water. Water Res, 2017. **123**: p. 734-750.

48. Fulladosa, E., et al., Patterns of metals and arsenic poisoning in Vibrio fischeri bacteria. Chemosphere, 2005. **60**(1): p. 43-8.

49. Pinzón-Espinosa, A. and R. Kanda, Naphthenic acids are key contributors to toxicity of heavy oil refining effluents. Science of The Total Environment, 2020. **729**: p. 138119.

50. Zemanek, M.G., et al., Toxicity and mutagenicity of component classes of oils isolated from soils at petroleum- and creosote-contaminated sites. J Air Waste Manag Assoc, 1997. **47**(12): p. 1250-8.

51. Codina, J.C., et al., A comparison of microbial bioassays for the detection of metal toxicity. Archives of Environmental Contamination and Toxicology, 1993. **25**(2): p. 250-254.

52. Ishaque, A.B., et al., Assessment of individual and combined toxicities of four non-essential metals (As, Cd, Hg and Pb) in the microtox assay. Int J Environ Res Public Health, 2006. **3**(1): p. 118-20.

53. Comber, M.H., et al., Oil refinery experience with the assessment of refinery effluents and receiving waters using biologically based methods. Integr Environ Assess Manag, 2015. **11**(4): p. 653-65.

54. De Baat, M.L., et al., Advancements in effect-based surface water quality assessment. Water Research, 2020. **183**: p. 116017.

55. Vermeirssen, E.L., et al., Linking toxicity in algal and bacterial assays with chemical analysis in passive samplers deployed in 21 treated sewage effluents. Environ Toxicol Chem, 2010. **29**(11): p. 2575-82.

56. Tang, J.Y.M., et al., Mixture effects of organic micropollutants present in water: Towards the development of effect-based water quality trigger values for baseline toxicity. Water Research, 2013. **47**(10): p. 3300-3314.

57. Escher, B.I., et al., General baseline toxicity QSAR for nonpolar, polar and ionisable chemicals and their mixtures in the bioluminescence inhibition assay with Aliivibrio fischeri. Environmental Science: Processes & Impacts, 2017. **19**(3): p. 414-428.

58. Wadhia, K., T. Dando, and K.C. Thompson, Intra-laboratory evaluation of Microbial Assay for Risk Assessment (MARA) for potential application in the implementation of the Water Framework Directive (WFD). Journal of environmental monitoring : JEM, 2007. **9**: p. 953-8.

59. Steliga, T., P. Jakubowicz, and P. Kapusta, Changes in toxicity during treatment of wastewater from oil plant contaminated with petroleum hydrocarbons. Journal of Chemical Technology & Biotechnology, 2015. **90**(8): p. 1408-1418.

60. Hurtado-Gallego, J., et al., Luminescent Microbial Bioassays and Microalgal Biosensors as Tools for Environmental Toxicity Evaluation. 2019. p. 1-58.

61. ECHA, Guidance on the Application of the CLP Criteria. Guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. Version 5.0. 2017.

62. Di Paolo, C., et al., Bioassay battery interlaboratory investigation of emerging contaminants in spiked water extracts - Towards the implementation of bioanalytical monitoring tools in water quality assessment and monitoring. Water Res, 2016. **104**: p. 473-484.

63. Tousova, Z., et al., European demonstration program on the effect-based and chemical identification and monitoring of organic pollutants in European surface waters. Sci Total Environ, 2017. **601-602**: p. 1849-1868.

64. Comber, M., et al., Aquatic ecotoxicity and biodegradability of cracked gas oils summary of relevant test data. CONCAWE Reports, 2013.

65. Magdaleno, A., et al., Assessment of heavy metal contamination and water quality in an urban river from Argentina. Brazilian Journal of Aquatic Science and Technology, 2014. **18**: p. 113.

66. Machado, M.D., A.R. Lopes, and E.V. Soares, Responses of the alga Pseudokirchneriella subcapitata to long-term exposure to metal stress. Journal of Hazardous Materials, 2015. **296**: p. 82-92.

67. Jia, A., et al., In vitro bioassays to evaluate complex chemical mixtures in recycled water. Water Res, 2015. **80**: p. 1-11.

68. Escher, B.I., et al., Benchmarking Organic Micropollutants in Wastewater, Recycled Water and Drinking Water with In Vitro Bioassays. Environmental Science & Technology, 2014. **48**(3): p. 1940-1956.

69. Escher, B.I., et al., Monitoring of the ecotoxicological hazard potential by polar organic micropollutants in sewage treatment plants and surface waters using a mode-of-action based test battery. Journal of Environmental Monitoring, 2008. **10**(5): p. 622-631.

70. Neale, P.A., et al., Integrating chemical analysis and bioanalysis to evaluate the contribution of wastewater effluent on the micropollutant burden in small streams. Science of The Total Environment, 2017. **576**: p. 785-795.

71. Beiras, R. and L. Saco-álvarez, Toxicity of seawater and sand affected by the Prestige fuel-oil spill using bivalve and sea urchin embryogenesis bioassays. Water, Air, and Soil Pollution, 2006. **177**(1): p. 457-466.

72. Hamers, T., et al., Time-Integrative Passive sampling combined with TOxicity Profiling (TIPTOP): an effect-based strategy for cost-effective chemical water quality assessment. Environmental Toxicology and Pharmacology, 2018. **64**: p. 48-59.

73. De Baat, M.L., et al., Effect-based nationwide surface water quality assessment to identify ecotoxicological risks. Water Research, 2019. **159**: p. 434-443.

74. Lampi, M.A., et al., Photoinduced toxicity of polycyclic aromatic hydrocarbons to Daphnia magna: ultraviolet-mediated effects and the toxicity of polycyclic aromatic hydrocarbon photoproducts. Environ Toxicol Chem, 2006. **25**(4): p. 1079-87.

75. Okamoto, A., M. Yamamuro, and N. Tatarazako, Acute toxicity of 50 metals to Daphnia magna. Journal of Applied Toxicology, 2015. **35**(7): p. 824-830.

76. Khangarot, B.S., P.K. Ray, and H. Chandr, Daphnia magna as a Model to Assess Heavy Metal Toxicity: Comparative Assessment with Mouse System. Acta hydrochimica et hydrobiologica, 1987. **15**(4): p. 427-432.

77. Chamorro, S., et al., Sublethal Effects of Chlorine-Free Kraft Mill Effluents on Daphnia magna. Bulletin of Environmental Contamination and Toxicology, 2016. **97**(6): p. 843-847.

78. Müller, J.B., et al., Comparative assessment of acute and chronic ecotoxicity of water soluble fractions of diesel and biodiesel on Daphnia magna and Aliivibrio fischeri. Chemosphere, 2019. **221**: p. 640-646.

79. Blewett, T.A., et al., Sublethal and Reproductive Effects of Acute and Chronic Exposure to Flowback and Produced Water from Hydraulic Fracturing on the Water Flea Daphnia magna. Environmental Science & Technology, 2017. **51**(5): p. 3032-3039.

80. Hayes, K.R., W.S. Douglas, and J. Fischer, Inter- and intra-laboratory testing of the Daphnia magna IQ toxicity test. Bull Environ Contam Toxicol, 1996. **57**(4): p. 660-6.

81. Rial, D., et al., Toxicity of binary mixtures of oil fractions to sea urchin embryos. J Hazard Mater, 2013. **263 Pt 2**: p. 431-40.

82. Bellas, J., et al., Toxicity of Organic Compounds to Marine Invertebrate Embryos and Larvae: A Comparison Between the Sea Urchin Embryogenesis Bioassay and Alternative Test Species. Ecotoxicology, 2005. **14**(3): p. 337.

83. His, E., et al., A comparison between oyster (Crassostrea gigas) and sea urchin (Paracentrotus lividus) larval bioassays for toxicological studies. Water Research, 1999. **33**(7): p. 1706-1718.

84. De Nicola, E., et al., Wastewater toxicity of tannin- versus chromium-based leather tanneries in Marrakesh, Morocco. Arch Environ Contam Toxicol, 2007. **53**(3): p. 321-8.

85. Meriç, S., et al., Toxicity of leather tanning wastewater effluents in sea urchin early development and in marine microalgae. Chemosphere, 2005. **61**(2): p. 208-217.

86. Mijangos, L., et al., Application of the Sea Urchin Embryo Test in Toxicity Evaluation and Effect-Directed Analysis of Wastewater Treatment Plant Effluents. Environ Sci Technol, 2020. **54**(14): p. 8890-8899.

87. Beiras, R., et al., Biological effects of contaminants: Paracentrotus lividus sea urchin embryo test with marine sediment elutriates. ICES Tech Mar Environ Sci, 2012. **51**.

88. Vethaak, A.D., et al., Toxicity profiling of marine surface sediments: A case study using rapid screening bioassays of exhaustive total extracts, elutriates and passive sampler extracts. Marine Environmental Research, 2017. **124**: p. 81-91.

89. Bellas, J., et al., Ecotoxicological evaluation of polycyclic aromatic hydrocarbons using marine invertebrate embryo–larval bioassays. Marine Pollution Bulletin, 2008. **57**(6): p. 493-502.

90. Beiras, R. and M. Albentosa, Inhibition of embryo development of the commercial bivalves Ruditapes decussatus and Mytilus galloprovincialis by trace metals; implications for the implementation of seawater quality criteria. Aquaculture, 2004. **230**(1): p. 205-213.

91. Leonards, P.E.G., et al., Impact of biodegradation on the potential bioaccumulation and toxicity of refinery effluents. Environmental Toxicology and Chemistry, 2011. **30**(10): p. 2175-2183.

92. Höss, S., et al., Interlaboratory comparison of a standardized toxicity test using the nematode Caenorhabditis elegans (ISO 10872). Environ Toxicol Chem, 2012. **31**(7): p. 1525-35.

93. Tejeda-Benitez, L., et al., Pollution by metals and toxicity assessment using Caenorhabditis elegans in sediments from the Magdalena River, Colombia. Environmental Pollution, 2016. **212**: p. 238-250.

94. Clavijo, A., et al., The nematode Caenorhabditis elegans as an integrated toxicological tool to assess water quality and pollution. Science of The Total Environment, 2016. **569-570**: p. 252-261.

95. Abbas, A., et al., Ecotoxicological impacts of surface water and wastewater from conventional and advanced treatment technologies on brood size, larval length, and cytochrome P450 (35A3) expression in Caenorhabditis elegans. Environ Sci Pollut Res Int, 2018. **25**(14): p. 13868-13880.

96. Sese, B.T., A. Grant, and B.J. Reid, Toxicity of polycyclic aromatic hydrocarbons to the nematode Caenorhabditis elegans. J Toxicol Environ Health A, 2009. **72**(19): p. 1168-80.

97. Moyson, S., et al., The effect of metal mixture composition on toxicity to C. elegans at individual and population levels. PLoS One, 2019. **14**(6): p. e0218929.

98. Sherry, J., B. Scott, and B. Dutka, Use of various acute, sublethal and early life-stage tests to evaluate the toxicity of refinery effluents. Environmental Toxicology and Chemistry, 1997. **16**(11): p. 2249-2257.

99. Cesar, A., et al., Amphipod and sea urchin tests to assess the toxicity of Mediterranean sediments: The case of Portmán Bay. Scientia Marina, 2004. **68**: p. 205-213.

100. Roddie, B., et al., The toxicity to Corophium volutator (Pallas) of beach sand contaminated by a spillage of crude oil. Chemosphere, 1994. **29**(4): p. 719-727.

101. Brils, J.M., et al., Oil effect in freshly spiked marine sediment on Vibrio fischeri, Corophium volutator, and Echinocardium cordatum. Environmental Toxicology and Chemistry, 2002. **21**(10): p. 2242-2251.

102. Wake, H., Oil refineries: a review of their ecological impacts on the aquatic environment. Estuarine, Coastal and Shelf Science, 2005. **62**(1): p. 131-140.

103. van den Heuvel-Greve, M., et al., A chronic bioassay with the estuarine amphipod Corophium volutator: Test method description and confounding factors. Chemosphere, 2007. **66**(7): p. 1301-1309.

104. Busquet, F., et al., OECD validation study to assess intra- and inter-laboratory reproducibility of the zebrafish embryo toxicity test for acute aquatic toxicity testing. Regul Toxicol Pharmacol, 2014. **69**(3): p. 496-511.

105. Neale, P.A., et al., Linking in Vitro Effects and Detected Organic Micropollutants in Surface Water Using Mixture-Toxicity Modeling. Environmental Science & Technology, 2015. **49**(24): p. 14614-14624.

106. Belanger, S.E., J.M. Rawlings, and G.J. Carr, Use of fish embryo toxicity tests for the prediction of acute fish toxicity to chemicals. Environ Toxicol Chem, 2013. **32**(8): p. 1768-83.

107. Legradi, J., et al. A novel ZF-Tox-Array to identify toxicity of environmental contaminants and complex mixtures. SETAC presentation. in SETAC Europe 26th Annual Meeting. 2016. Nantes, France.

108. Verslycke, T., et al., Mysid crustaceans as standard models for the screening and testing of endocrine-disrupting chemicals. Ecotoxicology, 2007. **16**(1): p. 205-19.

109. Gorbi, G., et al., Standardized methods for acute and semichronic toxicity tests with the copepod Acartia tonsa. Environmental Toxicology and Chemistry, 2012. **31**(9): p. 2023-2028.

110. Connors, K.A., et al., Daphnia magna and Ceriodaphnia dubia Have Similar Sensitivity in Standard Acute and Chronic Toxicity Tests. Environ Toxicol Chem, 2022. **41**(1): p. 134-147.

111. Oliveira-Filho, E., et al., Comparison between four and seven-day Ceriodaphnia dubia survival and reproduction test protocols using oil refinary effluent samples. Brazilian Archives of Biology and Technology, 2008. **51**: p. 137-142.

112. Daflon, S.D.A., et al., Toxicity identification and evaluation (TIE) of a petroleum refinery wastewater. J Environ Sci Health A Tox Hazard Subst Environ Eng, 2017. **52**(9): p. 842-848.

113. Truter, J.C., et al., An evaluation of the endocrine disruptive potential of crude oil water accommodated fractions and crude oil contaminated surface water to freshwater organisms using in vitro and in vivo approaches. Environ Toxicol Chem, 2017. **36**(5): p. 1330-1342.

114. Eriyamremu, G.E., et al., Alterations in glutathione reductase, superoxide dismutase, and lipid peroxidation of tadpoles (Xenopus laevis) exposed to Bonny Light crude oil and its fractions. Ecotoxicology and environmental safety, 2007. **71**: p. 284-90.

115. Park, J., et al., Interlaboratory Validation of Toxicity Testing Using the Duckweed Lemna minor Root-Regrowth Test. Biology (Basel), 2021. **11**(1).

116. ISO, ISO 20079:2005. Water quality - Determination of the toxic effect of water constituents and waste water on duckweed (Lemna minor) - Duckweed growth inhibition test. 2005.

117. Liu, J., et al., Bioluminescent nanopaper for rapid screening of toxic substances. Nano Research, 2018. **11**(1): p. 114-125.

118. Banni, M., et al., Assessing the impact of Benzo[a]pyrene on Marine Mussels: Application of a novel targeted low density microarray complementing classical biomarker responses. PLoS One, 2017. **12**(6): p. e0178460.

119. Fischer, M., et al., Repeatability and Reproducibility of the RTgill-W1 Cell Line Assay for Predicting Fish Acute Toxicity. Toxicological Sciences, 2019. **169**(2): p. 353-364.

120. Tanneberger, K., et al., Predicting fish acute toxicity using a fish gill cell line-based toxicity assay. Environ Sci Technol, 2013. **47**(2): p. 1110-9.

121. Natsch, A., et al., Accurate prediction of acute fish toxicity of fragrance chemicals with the RTgill-W1 cell assay. Environmental Toxicology and Chemistry, 2018. **37**(3): p. 931-941.

122. Dayeh, V.R., K. Schirmer, and N.C. Bols, Ammonia-containing industrial effluents, lethal to rainbow trout, induce vacuolisation and Neutral Red uptake in the rainbow trout gill cell line, RTgill-W1. Altern Lab Anim, 2009. **37**(1): p. 77-87.

123. Dayeh, V.R., K. Schirmer, and N.C. Bols, Applying whole-water samples directly to fish cell cultures in order to evaluate the toxicity of industrial effluent. Water Res, 2002. **36**(15): p. 3727-38.

124. Dayeh, V.R., D.H. Lynn, and N.C. Bols, Cytotoxicity of metals common in mining effluent to rainbow trout cell lines and to the ciliated protozoan, Tetrahymena thermophila. Toxicol In Vitro, 2005. **19**(3): p. 399-410.

125. Sansom, B., et al., Rapid assessment of the toxicity of oil sands process-affected waters using fish cell lines. In Vitro Cell Dev Biol Anim, 2013. **49**(1): p. 52-65.

126. Lee, L.E., et al., Fish cell lines as rapid and inexpensive screening and supplemental tools for whole effluent testing. Integr Environ Assess Manag, 2008. **4**(3): p. 372-4.

127. Georgantzopoulou, A., et al., Ecotoxicological Effects of Transformed Silver and Titanium Dioxide Nanoparticles in the Effluent from a Lab-Scale Wastewater Treatment System. Environmental Science & Technology, 2018. **52**(16): p. 9431-9441.

128. Park, J.A., Aquatic ecotoxicological assessment of sediments collected in Bergen using in vitro bioassays based on fish cell lines, in Department of Chemistry, Bioscience and Environmental Engineering. 2019, University of Stavanger.

129. Amaeze, N.H., et al., Cytotoxic and genotoxic responses of the RTgill-W1 fish cells in combination with the yeast oestrogen screen to determine the sediment quality of Lagos lagoon, Nigeria. Mutagenesis, 2015. **30**(1): p. 117-127.

130. Schirmer, K., et al., Ability of 16 priority PAHs to be directly cytotoxic to a cell line from the rainbow trout gill. Toxicology, 1998. **127**(1-3): p. 129-41.

131. Schirmer, K., et al., Use of fish gill cells in culture to evaluate the cytotoxicity and photocytotoxicity of intact and photomodified creosote. Environmental Toxicology and Chemistry, 1999. **18**(6): p. 1277-1288.

132. Schirmer, K., et al., Ability of fractionated petroleum refinery effluent to elicit cyto- and photocytotoxic responses and to induce 7-ethoxyresorufin-O-deethylase activity in fish cell lines. Sci Total Environ, 2001. **271**(1-3): p. 61-78.

133. Dayeh, V., D. Lynn, and N. Bols, Assessment of Metal Toxicity to Protozoa and Rainbow Trout Cells in Culture. 2022.

134. Minghetti, M. and K. Schirmer, Effect of media composition on bioavailability and toxicity of silver and silver nanoparticles in fish intestinal cells (RTgutGC). Nanotoxicology, 2016. **10**(10): p. 1526-1534.

135. Scott, J., B. Jason, and M. Minghetti, Applications of the RTgill-W1 Cell Line for Acute Whole-Effluent Toxicity Testing: In Vitro–In Vivo Correlation and Optimization of Exposure Conditions. Environmental Toxicology and Chemistry, 2020.

136. Bury, N.R., S. Schnell, and C. Hogstrand, Gill cell culture systems as models for aquatic environmental monitoring. J Exp Biol, 2014. **217**(Pt 5): p. 639-50.

137. Sharma, A. and M. Kumar, Evaluation of Environmental Contaminants and Natural Products: A Human Health Perspective. 2019: Bentham Science Publishers.

138. Lee, D.G., et al., Wastewater compounds in urban shallow groundwater wells correspond to exfiltration probabilities of nearby sewers. Water Res, 2015. **85**: p. 467-75.

139. Macova, M., et al., Bioanalytical tools for the evaluation of organic micropollutants during sewage treatment, water recycling and drinking water generation. Water Research, 2011. **45**(14): p. 4238-4247.

140. Macova, M., et al., Monitoring the biological activity of micropollutants during advanced wastewater treatment with ozonation and activated carbon filtration. Water Res, 2010. **44**(2): p. 477-92.

141. Fang, Y.X., et al., Assessment of hormonal activities and genotoxicity of industrial effluents using in vitro bioassays combined with chemical analysis. Environ Toxicol Chem, 2012. **31**(6): p. 1273-82.

142. Tang, J.Y.M., et al., Which chemicals drive biological effects in wastewater and recycled water? Water Research, 2014. **60**: p. 289-299.

143. Reungoat, J., et al., Removal of micropollutants and reduction of biological activity in a full scale reclamation plant using ozonation and activated carbon filtration. Water Research, 2010. **44**(2): p. 625-637.

144. Neale, P.A., et al., Bioanalytical Assessment of the Formation of Disinfection Byproducts in a Drinking Water Treatment Plant. Environmental Science & Technology, 2012. **46**(18): p. 10317-10325.

145. Stalter, D., et al., Point-of-use water filters can effectively remove disinfection by-products and toxicity from chlorinated and chloraminated tap water. Environmental Science: Water Research & Technology, 2016. **2**(5): p. 875-883.

146. Farré, M.J., et al., Bioanalytical and chemical assessment of the disinfection by-product formation potential: role of organic matter. Water Res, 2013. **47**(14): p. 5409-21.

147. Reifferscheid, G. and T. Grummt, Genotoxicity in German Surface Waters - Results of a Collaborative Study. Water, Air, and Soil Pollution, 2000. **123**(1): p. 67-79.

148. Han, Y., et al., Evaluation of genotoxic effects of surface waters using a battery of bioassays indicating different mode of action. Ecotoxicol Environ Saf, 2016. **133**: p. 448-56.

149. Sun, J., et al., Genotoxicity and cytotoxicity reduction of the polluted urban river after ecological restoration: a field-scale study of Jialu River in northern China. Environmental Science and Pollution Research, 2017. **24**(7): p. 6715-6723.

150. Vincent-Hubert, F., et al., Use of low density polyethylene membranes for assessment of genotoxicity of PAHs in the Seine River. Ecotoxicology, 2017. **26**(2): p. 165-172.

151. Ye, Y., et al., Assessing of genotoxicity of 16 centralized source-waters in China by means of the SOS/umu assay and the micronucleus test: initial identification of the potential genotoxicants by use of a GC/MS method and the QSAR Toolbox 3.0. Mutat Res Genet Toxicol Environ Mutagen, 2014. **763**: p. 36-43.

152. Bi, H., et al., Recovery of a freshwater wetland from chemical contamination after an oil spill. Journal of Environmental Monitoring, 2011. **13**(3): p. 713-720.

153. Yamamoto, A., Y. Kohyama, and T. Hanawa, Mutagenicity evaluation of forty-one metal salts by the umu test. Journal of Biomedical Materials Research, 2002. **59**(1): p. 176-183.

154. Prasse, C., et al., Spoilt for choice: A critical review on the chemical and biological assessment of current wastewater treatment technologies. Water Research, 2015. **87**: p. 237-270.

155. Resende, F.A., et al., Mutagenicity of Flavonoids Assayed by Bacterial Reverse Mutation (Ames) Test. Molecules, 2012. **17**(5).

156. Gallampois, C.M., et al., Integrated biological-chemical approach for the isolation and selection of polyaromatic mutagens in surface waters. Anal Bioanal Chem, 2013. **405**(28): p. 9101-12.

157. Gruener, N., Mutagenicity of ozonated, recycled water. Bull Environ Contam Toxicol, 1978. **20**(4): p. 522-6.

158. Kurelec, B., et al., Induction of benzo (a)pyrene monooxygenase in fish and the salmonella test as a tool for detecting mutagenic/carcinogenic xenobiotics in the aquatic environment. Bulletin of Environmental Contamination and Toxicology, 1979. **21**(1): p. 799-807.

159. Wang, D., et al., The influence of alkyl substitution on the in vitro metabolism and mutagenicity of benzo[a]pyrene. Chemico-Biological Interactions, 2022. **363**: p. 110007.

160. Vincent-Hubert, F., et al., Mutagenicity and genotoxicity of suspended particulate matter in the Seine river estuary. Mutat Res, 2012. **741**(1-2): p. 7-12.

161. Iqbal, M., et al., Mutagenicity and cytotoxicity evaluation of photo-catalytically treated petroleum refinery wastewater using an array of bioassays. Chemosphere, 2017. **168**: p. 590-598.

162. Metcalfe, C.D., R.A. Sonstegard, and M.A. Quilliam, Genotoxic activity of particulate material in petroleum refinery effluents. Bull Environ Contam Toxicol, 1985. **35**(2): p. 240-8.

163. Metcalfe, C.D. and R.A. Sonstegard, GENOTOXIC AND CARCINOGENIC ACTIVITY OF OIL REFINERY EFFLUENTS, in Oil in Freshwater: Chemistry, Biology, Countermeasure Technology, J.H. Vandermeulen and S.E. Hrudey, Editors. 1987, Pergamon. p. 319-334.

164. Vargas, V.M., V.E. Motta, and J.A. Henriques, Mutagenic activity detected by the Ames test in river water under the influence of petrochemical industries. Mutat Res, 1993. **319**(1): p. 31-45.

165. Lemos, C.T., et al., Genotoxicity of river water under the influence of petrochemical industrial complexes. Bull Environ Contam Toxicol, 1994. **52**(6): p. 848-55.

166. Khatoon, K. and A. Malik, Cyto-genotoxic potential of petroleum refinery wastewater mixed with domestic sewage used for irrigation of food crops in the vicinity of an oil refinery. Heliyon, 2021. **7**(10): p. e08116.

167. Rastall, A.C., et al., The identification of readily bioavailable pollutants in Lake Shkodra/Skadar using semipermeable membrane devices (SPMDs), bioassays and chemical analysis. Environ Sci Pollut Res Int, 2004. **11**(4): p. 240-53.

168. Brack, W., et al., Effect-directed analysis of mutagens and ethoxyresorufin-O-deethylase inducers in aquatic sediments. Environmental Toxicology and Chemistry, 2005. **24**(10): p. 2445-2458.

169. OECD, Test No. 471: Bacterial Reverse Mutation Test. 1997.

170. Kamber, M., et al., Comparison of the Ames II and traditional Ames test responses with respect to mutagenicity, strain specificities, need for metabolism and correlation with rodent carcinogenicity. Mutagenesis, 2009. **24**(4): p. 359-66.

171. Reifferscheid, G., et al., International round-robin study on the Ames fluctuation test. Environmental and Molecular Mutagenesis, 2012. **53**(3): p. 185-197.

172. Kaur, T., R. Bhardwaj, and S. Arora, Evaluating groundwater for its probable mutagenicity and genotoxicity using in vitro bioanalytical tools. Exposure and Health, 2019. **11**(1): p. 21-31.

173. Umbuzeiro Gde, A., et al., Comparison of the Salmonella/microsome microsuspension assay with the new microplate fluctuation protocol for testing the mutagenicity of environmental samples. Environ Mol Mutagen, 2010. **51**(1): p. 31-8.

174. Rao, S.S., et al., Assessment of the genotoxic potential of pulp mill effluent using bacterial, fish and mammalian assays. Chemosphere, 1995. **31**(6): p. 3553-3566.

175. Alabi, O.A., Comparative chemical analysis, mutagenicity, and genotoxicity of Petroleum refinery wastewater and its contaminated river using prokaryotic and eukaryotic assays. Protoplasma, 2022.

176. Reifferscheid, G., et al., Measurement of genotoxicity in wastewater samples with the in vitro micronucleus test: results of a round-robin study in the context of standardisation according to ISO. Mutat Res, 2008. **649**(1-2): p. 15-27.

177. Fagundes Soares Garcia, A.C., et al., Micronucleus study of the quality and mutagenicity of surface water from a semi-arid region. Journal of Environmental Monitoring, 2011. **13**(12): p. 3329-3335.

178. Nunes, E.A., et al., Genotoxic assessment on river water using different biological systems. Chemosphere, 2011. **84**(1): p. 47-53.

179. Souza, T.S., et al., The Allium cepa bioassay to evaluate landfarming soil, before and after the addition of rice hulls to accelerate organic pollutants biodegradation. Ecotoxicology and Environmental Safety, 2009. **72**(5): p. 1363-1368.

180. Leme, D.M. and M.A. Marin-Morales, Chromosome aberration and micronucleus frequencies in Allium cepa cells exposed to petroleum polluted water—A case study. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 2008. **650**(1): p. 80-86.

181. Gupta, A.K. and M. Ahmad, Assessment of cytotoxic and genotoxic potential of refinery waste effluent using plant, animal and bacterial systems. J Hazard Mater, 2012. **201-202**: p. 92-9.

182. Hoshina, M.M. and M.A. Marin-Morales, Micronucleus and chromosome aberrations induced in onion (Allium cepa) by a petroleum refinery effluent and by river water that receives this effluent. Ecotoxicol Environ Saf, 2009. **72**(8): p. 2090-5.

183. Hara, R.V. and M.A. Marin-Morales, In vitro and in vivo investigation of the genotoxic potential of waters from rivers under the influence of a petroleum refinery (São Paulo State - Brazil). Chemosphere, 2017. **174**: p. 321-330.

184. Brinkmann, M., et al., Genotoxicity of heterocyclic PAHs in the micronucleus assay with the fish liver cell line RTL-W1. PLoS One, 2014. **9**(1): p. e85692.

185. Lah, B., M. Narat, and R. Marin, Monitoring of Genotoxicity in Drinking Water Using in vitro Comet Assay and Ames Test. Food Technology and Biotechnology, 2005. **43**.

186. Zhong, Y., et al., Evaluating the genotoxicity of surface water of Yangzhong City using the Vicia faba micronucleus test and the comet assay. Bull Environ Contam Toxicol, 2001. **67**(2): p. 217-24.

187. Schnurstein, A. and T. Braunbeck, Tail moment versus tail length--application of an in vitro version of the comet assay in biomonitoring for genotoxicity in native surface waters using primary hepatocytes and gill cells from zebrafish (Danio rerio). Ecotoxicol Environ Saf, 2001. **49**(2): p. 187-96.

188. Iji, O., et al., Evaluation of the genotoxic potential of water impacted by acid mine drainage from a coal mine in Mpumalanga, South Africa, using the Ames test and Comet assay. Water S.A, 2021. **47**.

189. Cui, Z., et al., Genotoxicity detection of oil-containing drill cuttings by Comet assay based on a demersal marine fish Mugilogobius chulae. Ecotoxicology and Environmental Safety, 2021. **208**: p. 111655.

190. Walmsley, R. and M. Tate, The GADD45a-GFP greenscreen HC assay. Methods in molecular biology (Clifton, N.J.), 2012. **817**: p. 231-50.

191. Daniel, M., et al., Results of a technology demonstration project to compare rapid aquatic toxicity screening tests in the analysis of industrial effluents. Journal of Environmental Monitoring, 2004. **6**(11): p. 855-865.

192. Akerman, G.S., et al., Gene expression profiles and genetic damage in benzo(a)pyrene diol epoxide-exposed TK6 cells. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2004. **549**(1): p. 43-64.

193. Birrell, L., et al., GADD45a-GFP GreenScreen HC assay results for the ECVAM recommended lists of genotoxic and non-genotoxic chemicals for assessment of new genotoxicity tests. Mutat Res, 2010. **695**(1-2): p. 87-95.

194. van der Linden, S.C., et al., Development of a panel of high-throughput reporter-gene assays to detect genotoxicity and oxidative stress. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 2014. **760**: p. 23-32.

195. Neale, P.A., et al., Application of in vitro bioassays for water quality monitoring in three drinking water treatment plants using different treatment processes including biological treatment, nanofiltration and ozonation coupled with disinfection. Environmental Science: Water Research & Technology, 2020. **6**(9): p. 2444-2453.

196. König, M., et al., Impact of untreated wastewater on a major European river evaluated with a combination of in vitro bioassays and chemical analysis. Environmental Pollution, 2017. **220**: p. 1220-1230.

197. Hebert, A., et al., Bioanalytical assessment of adaptive stress responses in drinking water: A predictive tool to differentiate between micropollutants and disinfection by-products. Water Research, 2018. **132**.

198. Välitalo, P., et al., Effect-based assessment of toxicity removal during wastewater treatment. Water Res, 2017. **126**: p. 153-163.

199. Yeh, R., et al., Bioanalytical and chemical evaluation of disinfection by-products in swimming pool water. Water research, 2014. **59C**: p. 172-184.

200. Escher, B.I., et al., Water quality assessment using the AREc32 reporter gene assay indicative of the oxidative stress response pathway. Journal of Environmental Monitoring, 2012. **14**(11): p. 2877-2885.

201. Neale, P.A., F.D.L. Leusch, and B.I. Escher, What is driving the NF-κB response in environmental water extracts? Chemosphere, 2018. **210**: p. 645-652.

202. Neale, P.A., et al., Assessing the Mixture Effects in In Vitro Bioassays of Chemicals Occurring in Small Agricultural Streams during Rain Events. Environmental Science & Technology, 2020. **54**(13): p. 8280-8290.

203. Müller, M.E., C. Zwiener, and B.I. Escher, Storm Event–Driven Occurrence and Transport of Dissolved and Sorbed Organic Micropollutants and Associated Effects in the Ammer River, Southwestern Germany. Environmental Toxicology and Chemistry, 2021. **40**(1): p. 88-99.

204. Müller, M.E., et al., Combining in vitro reporter gene bioassays with chemical analysis to assess changes in the water quality along the Ammer River, Southwestern Germany. Environmental Sciences Europe, 2018. **30**(1): p. 20.

205. Escher, B.I., et al., Most oxidative stress response in water samples comes from unknown chemicals: the need for effect-based water quality trigger values. Environ Sci Technol, 2013. **47**(13): p. 7002-11.

206. Hashmi, M.A.K., et al., Effect-directed analysis (EDA) of Danube River water sample receiving untreated municipal wastewater from Novi Sad, Serbia. Sci Total Environ, 2018. **624**: p. 1072-1081.

207. Nivala, J., et al., Application of cell-based bioassays to evaluate treatment efficacy of conventional and intensified treatment wetlands. Environmental Science: Water Research & Technology, 2018. **4**(2): p. 206-217.

208. Albergamo, V., et al., Evaluation of reverse osmosis drinking water treatment of riverbank filtrate using bioanalytical tools and non-target screening. Environmental Science: Water Research & Technology, 2020. **6**(1): p. 103-116.

209. Muthusamy, S., C. Peng, and J.C. Ng, Effects of multi-component mixtures of polyaromatic hydrocarbons and heavy metal/loid(s) on Nrf2-antioxidant response element (ARE) pathway in ARE reporter-HepG2 cells. Toxicol Res (Camb), 2016. **5**(4): p. 1160-1171.

210. Escher, B.I. and P.A. Neale, Effect-Based Trigger Values for Mixtures of Chemicals in Surface Water Detected with In Vitro Bioassays. Environmental Toxicology and Chemistry, 2021. **40**(2): p. 487-499.

211. Schroeder, A.L., et al., Environmental surveillance and monitoring—The next frontiers for high-throughput toxicology. Environmental Toxicology and Chemistry, 2016. **35**(3): p. 513-525.

212. Escher, B.I., P.A. Neale, and F.D. Leusch, Effect-based trigger values for in vitro bioassays: Reading across from existing water quality guideline values. Water Res, 2015. **81**: p. 137-48.

213. Lundqvist, J., et al., Innovative drinking water treatment techniques reduce the disinfection-induced oxidative stress and genotoxic activity. Water Research, 2019. **155**: p. 182-192.

214. Lundqvist, J., et al., In vitro bioanalytical evaluation of removal efficiency for bioactive chemicals in Swedish wastewater treatment plants. Scientific Reports, 2019. **9**(1): p. 7166.

215. Pieterse, B., et al., Effect-based assessment of persistent organic pollutant and pesticide dumpsite using mammalian CALUX reporter cell lines. Environ Sci Pollut Res Int, 2015. **22**(19): p. 14442-54.

216. Johann, S., et al., Combining Different In Vitro Bioassays to Evaluate Genotoxicity of Water-Accommodated Fractions from Petroleum Products. Toxics, 2020. **8**(2).

217. Eichbaum, K., et al., In vitro bioassays for detecting dioxin-like activity--application potentials and limits of detection, a review. Sci Total Environ, 2014. **487**: p. 37-48.

218. Vrabie, C.M., M.T.O. Jonker, and A.J. Murk, Specific in vitro toxicity of crude and refined petroleum products. 1. Aryl hydrocarbon receptor—mediated responses. Environmental Toxicology and Chemistry, 2009. **28**(9): p. 1995-2003.

219. Neale, P.A., et al., Solid-phase extraction as sample preparation of water samples for cell-based and other in vitro bioassays. Environmental Science: Processes & Impacts, 2018. **20**(3): p. 493-504.

220. Machala, M., et al., Aryl hydrocarbon receptor-mediated activity of mutagenic polycyclic aromatic hydrocarbons determined using in vitro reporter gene assay. Mutat Res, 2001. **497**(1-2): p. 49-62.

221. Hong, S., et al., Two Years after the Hebei Spirit Oil Spill: Residual Crude-Derived Hydrocarbons and Potential AhR-Mediated Activities in Coastal Sediments. Environmental Science & Technology, 2012. **46**(3): p. 1406-1414.

222. Kim, J., et al., Newly Identified AhR-Active Compounds in the Sediments of an Industrial Area Using Effect-Directed Analysis. Environmental Science & Technology, 2019. **53**(17): p. 10043-10052.

223. Jonker, M.T., et al., Weathering and toxicity of marine sediments contaminated with oils and polycyclic aromatic hydrocarbons. Environ Toxicol Chem, 2006. **25**(5): p. 1345-53.

224. Bekki, K., et al., Evaluation of Toxic Activities of Polycyclic Aromatic Hydrocarbon Derivatives Using In Vitro Bioassays. Journal of Health Science, 2009. **55**.

225. de Weert, J., et al., Time integrative sampling properties of Speedisk and silicone rubber passive samplers determined by chemical analysis and in vitro bioassay testing. Chemosphere, 2020. **259**: p. 127498.

226. Pieterse, B., et al., PAH-CALUX, an Optimized Bioassay for AhR-Mediated Hazard Identification of Polycyclic Aromatic Hydrocarbons (PAHs) as Individual Compounds and in Complex Mixtures. Environmental Science & Technology, 2013. **47**(20): p. 11651-11659.

227. Ziccardi, M.H., et al., Application of the luciferase cell culture bioassay for the detection of refined petroleum products. Marine Pollution Bulletin, 2002. **44**(10): p. 983-991.

228. Alygizakis, N., et al., Characterization of wastewater effluents in the Danube River Basin with chemical screening, in vitro bioassays and antibiotic resistant genes analysis. Environment International, 2019. **127**: p. 420-429.

229. Brussaard, C.P.D., et al., Immediate ecotoxicological effects of short-lived oil spills on marine biota. Nature Communications, 2016. **7**(1): p. 11206.

230. Louiz, I., et al., Monitoring of dioxin-like, estrogenic and anti-androgenic activities in sediments of the Bizerta lagoon (Tunisia) by means of in vitro cell-based bioassays: Contribution of low concentrations of polynuclear aromatic hydrocarbons (PAHs). Science of The Total Environment, 2008. **402**(2): p. 318-329.

231. Emelogu, E.S., et al., Investigating the significance of dissolved organic contaminants in aquatic environments: coupling passive sampling with in vitro bioassays. Chemosphere, 2013. **90**(2): p. 210-9.

232. Huuskonen, S., et al., Induction of CYP1A and porphyrin accumulation in fish hepatoma cells (PLHC-1) exposed to sediment or water from a PCB-contaminated lake (Lake Kernaala, Finland). Marine Environmental Research, 1998. **46**(1): p. 379-384.

233. Suares-Rocha, P., et al., Assessment of cytotoxicity and AhR-mediated toxicity in tropical fresh water sediments under the influence of an oil refinery. Environ Sci Pollut Res Int, 2015. **22**(16): p. 12566-75.

234. Behrens, A., et al., Polycyclic aromatic hydrocarbons as inducers of cytochrome P4501A enzyme activity in the rainbow trout liver cell line, RTL-W1, and in primary cultures of rainbow trout hepatocytes. Environ Toxicol Chem, 2001. **20**(3): p. 632-43.

235. Harman, C., E. Farmen, and K.E. Tollefsen, Monitoring North Sea oil production discharges using passive sampling devices coupled with in vitro bioassay techniques. Journal of Environmental Monitoring, 2010. **12**(9): p. 1699-1708.

236. Franco, M.E., G.E. Sutherland, and R. Lavado, Xenobiotic metabolism in the fish hepatic cell lines Hepa-E1 and RTH-149, and the gill cell lines RTgill-W1 and G1B: Biomarkers of CYP450 activity and oxidative stress. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology, 2018. **206-207**: p. 32-40.

237. Leusch, F.D., et al., Analysis of the sensitivity of in vitro bioassays for androgenic, progestagenic, glucocorticoid, thyroid and estrogenic activity: Suitability for drinking and environmental waters. Environ Int, 2017. **99**: p. 120-130.

238. Leusch, F.D., et al., Comparison of five in vitro bioassays to measure estrogenic activity in environmental waters. Environ Sci Technol, 2010. **44**(10): p. 3853-60.

239. French, V.A., et al., Characterisation of microcontaminants in Darwin Harbour, a tropical estuary of northern Australia undergoing rapid development. Sci Total Environ, 2015. **536**: p. 639-647.

240. Vrabie, C.M., et al., Specific in vitro toxicity of crude and refined petroleum products: II. Estrogen (alpha and beta) and androgen receptor-mediated responses in yeast assays. Environ Toxicol Chem, 2010. **29**(7): p. 1529-36.

241. Johann, S., et al., Receptor-mediated estrogenicity of native and chemically dispersed crude oil determined using adapted microscale reporter gene assays. Environment International, 2020. **134**: p. 105320.

242. Tollefsen, K.E., et al., Estrogen receptor (ER) agonists and androgen receptor (AR) antagonists in effluents from Norwegian North Sea oil production platforms. Mar Pollut Bull, 2007. **54**(3): p. 277-83.

243. Jarošová, B., et al., What level of estrogenic activity determined by in vitro assays in municipal waste waters can be considered as safe? Environment International, 2014. **64**: p. 98-109.

244. Li, J., M. Ma, and Z. Wang, In vitro profiling of endocrine disrupting effects of phenols. Toxicol In Vitro, 2010. **24**(1): p. 201-7.

245. Kunz, P.Y., et al., Effect-based tools for monitoring estrogenic mixtures: Evaluation of five in vitro bioassays. Water Research, 2017. **110**: p. 378-388.

246. Sonneveld, E., et al., Comparison of In Vitro and In Vivo Screening Models for Androgenic and Estrogenic Activities. Toxicological Sciences, 2006. **89**(1): p. 173-187.

247. Könemann, S., et al., Effect-based and chemical analytical methods to monitor estrogens under the European Water Framework Directive. TrAC Trends in Analytical Chemistry, 2018. **102**: p. 225-235.

248. Shi, P., et al., Toxicological and chemical insights into representative source and drinking water in eastern China. Environ Pollut, 2018. **233**: p. 35-44.

249. Jia, Y., et al., Effect-based and chemical analyses of agonistic and antagonistic endocrine disruptors in multiple matrices of eutrophic freshwaters. Sci Total Environ, 2019. **651**(Pt 1): p. 1096-1104.

250. Houtman, C.J., et al., High resolution effect-directed analysis of steroid hormone (ant)agonists in surface and wastewater quality monitoring. Environmental Toxicology and Pharmacology, 2020. **80**: p. 103460.

251. Van der Linden, S.C., et al., Detection of multiple hormonal activities in wastewater effluents and surface water, using a panel of steroid receptor CALUX bioassays. Environ Sci Technol, 2008. **42**(15): p. 5814-20.

252. Brand, W., et al., Trigger values for investigation of hormonal activity in drinking water and its sources using CALUX bioassays. Environment International, 2013. **55**: p. 109-118.

253. Leusch, F.D., et al., Assessment of the application of bioanalytical tools as surrogate measure of chemical contaminants in recycled water. Water Res, 2014. **49**: p. 300-15.

254. Bain, P.A., M. Williams, and A. Kumar, Assessment of multiple hormonal activities in wastewater at different stages of treatment. Environ Toxicol Chem, 2014. **33**(10): p. 2297-307.

255. Roberts, J., et al., Tracking multiple modes of endocrine activity in Australia's largest inland sewage treatment plant and effluent- receiving environment using a panel of in vitro bioassays. Environ Toxicol Chem, 2015. **34**(10): p. 2271-81.

256. Houtman, C.J., R. Ten Broek, and A. Brouwer, Steroid hormonal bioactivities, culprit natural and synthetic hormones and other emerging contaminants in waste water measured using bioassays and UPLC-tQ-MS. Sci Total Environ, 2018. **630**: p. 1492-1501.

257. Gehrmann, L., et al., (Anti-)estrogenic and (anti-)androgenic effects in wastewater during advanced treatment: comparison of three in vitro bioassays. Environ Sci Pollut Res Int, 2018. **25**(5): p. 4094-4104.

258. Vrabie, C.M., et al., Specific in vitro toxicity of crude and refined petroleum products: 3. Estrogenic responses in mammalian assays. Environmental Toxicology and Chemistry, 2011. **30**(4): p. 973-980.

259. Scott, P.D., et al., An assessment of endocrine activity in Australian rivers using chemical and in vitro analyses. Environ Sci Pollut Res Int, 2014. **21**(22): p. 12951-67.

260. Brion, F., et al., Monitoring estrogenic activities of waste and surface waters using a novel in vivo zebrafish embryonic (EASZY) assay: Comparison with in vitro cell-based assays and determination of effect-based trigger values. Environment International, 2019. **130**: p. 104896.

261. Huang, G.-Y., et al., Feminization and masculinization of western mosquitofish (Gambusia affinis) observed in rivers impacted by municipal wastewaters. Scientific Reports, 2016. **6**(1): p. 20884.

262. Thomas, K.V., et al., An assessment of in vitro androgenic activity and the identification of environmental androgens in United Kingdom estuaries. Environ Toxicol Chem, 2002. **21**(7): p. 1456-61.

263. Elkayar, K., et al., Passive sampling and in vitro assays to monitor antiandrogens in a river affected by wastewater discharge. Science of The Total Environment, 2022. **804**: p. 150067.

264. Jonker, M.T.O., et al., Synergistic androgenic effect of a petroleum product caused by the joint action of at least three different types of compounds. Chemosphere, 2016. **144**: p. 1142-1147.

265. Thomas, K.V., et al., Identification of in vitro estrogen and androgen receptor agonists in North Sea offshore produced water discharges. Environ Toxicol Chem, 2004. **23**(5): p. 1156-63.

266. Indiveri, P., et al., Analytical methodology for the profiling and characterization of androgen receptor active compounds in human placenta. Reprod Toxicol, 2014. **47**: p. 102-10.

267. Vinggaard, A.M., C. Hnida, and J.C. Larsen, Environmental polycyclic aromatic hydrocarbons affect androgen receptor activation in vitro. Toxicology, 2000. **145**(2-3): p. 173-83.

268. Šauer, P., et al., Do progestins contribute to (anti-)androgenic activities in aquatic environments? Environ Pollut, 2018. **242**(Pt A): p. 417-425.

269. Weiss, J.M., et al., Masking effect of anti-androgens on androgenic activity in European river sediment unveiled by effect-directed analysis. Anal Bioanal Chem, 2009. **394**(5): p. 1385-97.

270. Leusch, F.D.L., et al., Comparison of in vitro and in vivo bioassays to measure thyroid hormone disrupting activity in water extracts. Chemosphere, 2018. **191**: p. 868-875.

271. Couderc, M., et al., Levels and distributions of organic pollutants in subtidal sediments from the Loire estuary: Are there any relationships with TTR-binding activity? Journal of Sea Research, 2016. **118**: p. 59-68.

272. Šauer, P., et al., Bisphenols emerging in Norwegian and Czech aquatic environments show transthyretin binding potency and other less-studied endocrine-disrupting activities. Science of The Total Environment, 2021. **751**: p. 141801.

273. Collet, B., et al., Evaluation of a panel of in vitro methods for assessing thyroid receptor ß and transthyretin transporter disrupting activities. Reproductive Toxicology, 2020. **96**: p. 432-444.

274. Weiss, J.M., et al., Tracing thyroid hormone-disrupting compounds: database compilation and structure-activity evaluation for an effect-directed analysis of sediment. Analytical and Bioanalytical Chemistry, 2015. **407**(19): p. 5625-5634.

275. Montaño, M., et al., Metabolic activation of nonpolar sediment extracts results in enhanced thyroid hormone disrupting potency. Environ Sci Technol, 2013. **47**(15): p. 8878-86.

276. Weiss, J.M., et al., Competitive binding of poly- and perfluorinated compounds to the thyroid hormone transport protein transthyretin. Toxicol Sci, 2009. **109**(2): p. 206-16.

277. Rodrigues, E.T., et al., H9c2(2-1)-based sulforhodamine B assay as a possible alternative in vitro platform to investigate effluent and metals toxicity on fish. Chemosphere, 2021. **275**: p. 130009.

278. Rodrigues, E.T., et al., Rat cardiomyocyte H9c2(2-1)-based sulforhodamine B assay as a promising in vitro method to assess the biological component of effluent toxicity. Journal of Environmental Sciences, 2020. **96**: p. 163-170.

279. Tollefsen, K.-E., et al., Effluents from oil production activities contain chemicals that interfere with normal function of intra- and extra-cellular estrogen binding proteins. Marine Environmental Research, 2006. **62**: p. S191-S194.

280. Leslie, H. and P. Leonards, Determination of potentially bioaccumulating substances (PBS) in whole effluents using the EGOM Liquid-Liquid Extraction Method. OSPAR Intercessional Expert Group on Whole Effluent Assessment, Interlaboratory Study. . 2005.

281. Redman, A.D., et al., PETROTOX: An aquatic toxicity model for petroleum substances. Environmental Toxicology and Chemistry, 2012. **31**(11): p. 2498-2506.

282. Redman, A.D. and T.F. Parkerton, Guidance for improving comparability and relevance of oil toxicity tests. Mar Pollut Bull, 2015. **98**(1-2): p. 156-70.

283. Redman, A.D., et al., A re-evaluation of PETROTOX for predicting acute and chronic toxicity of petroleum substances. Environmental Toxicology and Chemistry, 2017. **36**(8): p. 2245-2252.

284. Hjort, M., et al., Conventional and high resolution chemical characterization to assess refinery effluent treatment performance. Chemosphere, 2021. **278**: p. 130383.

285. Redman, A.D., et al., PETRORISK: A risk assessment framework for petroleum substances. Integrated Environmental Assessment and Management, 2014. **10**(3): p. 437-448.

APPENDIX 1 SUMMARY OF LESSONS LEARNED FROM PREVIOUS WORK

Carere *et al*. [13] Final report from WFD CIS EBM Sub-group (Technical Proposal for Effect-Based Monitoring and Assessment under the WFD).

This work is based on the activity of a specific sub-group for effect-based methods comprising representatives of experts from 9 Member States, Switzerland and several stakeholders in the context of the Common Implementation Strategy (CIS) for the Water Framework Directive. The work examines the possible implementation of effect-based methods for monitoring and assessment in the WFD context, alongside traditional chemical analysis. As part of the work, an inventory of 46 bioassays was collated.

Three criteria were presented for the selection of *in vitro* EBMs to assess the quality of water bodies:

1. Relevance of the EBM. Effects at the molecular level must be mechanistically linked to apical endpoints. This can be done by following the concept of the Adverse Outcome Pathway (AOP), or by using field studies to link the occurrence of adverse effects in the environment to molecular initiating events (MIE) or the presence of compounds known to trigger these MIEs.
2. Maturity of the EBM. EBMs should be standardised according to ISO/CEN, or European national standard organisations (such as DIN, BSI, NEN) or validated via OECD. If no standard is available, the assay's performance should be assessed using (international) interlaboratory trials.
3. "Assessability" of results obtained by the EBM. The results obtained should have the ability to be assessed relative to other values resulting from measurements with the same EBM. Alternatively, the results can be matched against a defined effect-based trigger value (EBT).

Relevant MoAs with developed EBMs for potential implementation in the WFD that were recommended include activation of estrogen receptor (ER), phytotoxicity/PSII inhibition, mutagenicity, dioxin-like effects. Relevant MoAs with need for further research and method development that were recommended include binding to human ER, activation of ER, binding to aryl hydrocarbon receptor (AhR), antimicrobial resistance (AMR) and neurotoxicity (for future). The report also discussed the use of biomarkers, although their limited use thus far and subsequently lower levels of maturity affects their recommendation.

Implications for this work:

- A good battery should include both short- and long-term *in vivo* bioassays comprising three species from different trophic levels and taxonomic groups.
- A battery should have sufficient sensitivity and an overall discriminatory power responding to as many forms of pollution as possible.
- The relevance, maturity/standardisation and result interpretation are imperative criteria that must be considered as part of the evaluations.

Hjort *et al*. [284]. Conventional and high-resolution chemical characterization to assess refinery effluent treatment performance.

This work is a companion paper to Whale *et al*. [5]. Total Petroleum Hydrocarbons (TPH), benzene, toluene, ethyl benzene, xylenes (BTEX), polycyclic aromatic hydrocarbons (PAHs), (bio) chemical oxygen demand, total nitrogen, total suspended solids and selected metals before, and after, treatment steps were analysed at samples (2015-2016) from 13 European refineries with variable wastewater treatment systems. High-resolution two-dimensional gas chromatography (GCxGC) analysis was used to monitor TPH composition change during

treatment, which was then compared to model predictions by SimpleTreat. While no bioassays were included in the scope of this paper (see Whale *et al*. [5] for this part of the work), the mechanistic oil toxicity model PETROTOX was used to predict toxicity.

The authors found that the European refineries typically operated meeting BAT-AELs, with removal factors of TPH for their whole WWTP varying from 97% to >99.8% between refineries. The removal of low molecular weight PAHs was assumed to be principally due to biotransformation, and somewhat due to evaporation. Compositional shifts were refinery specific, but the largest shifts were found for short chain hydrocarbons (C_9-C_{15}), which were more susceptible to be removed.

The two models used performed well. SimpleTreat successfully, if not slightly conservatively, predicted the shape of the effluent composition; the predicted effluent concentrations were sometimes higher than the measured effluent compositions for the aromatic constituents. In terms of PETROTOX, the reduction in potential hydrocarbon exposure coincided with a decrease in predicted toxicity.

Implications for this work:

- Final effluents from European refineries had low bioavailability; they may produce low response levels in toxicity tests without any form of sample enrichment.

Whale *et al*. [5] Assessment of oil refinery wastewater and effluent integrating bioassays, mechanistic modelling and bioavailability evaluation.

This work is a companion paper to Hjort *et al*. [284]. It assessed the toxicity of refinery wastewater treatment, in the form of mid-treatment (wastewater) and final effluent, before and after XAD®-resin extraction which removes metals and other confounding factors. The bioassays selected to assess the toxicity were the Microtox® test with bioluminescent bacteria *Aliivibrio fischeri*, the 72-hour algal growth inhibition test with *Raphidocelis subcapitata*, the 21-day chronic daphnid test with *Daphnia magna*, and the 5-day fish embryo test with *Danio rerio*. The observed toxicity, or lack thereof, was compared to TPH measurements to estimate exposure levels and biomimetic solid phase microextraction (BE-SPME) to measure bioavailable hydrocarbon concentrations.

The results demonstrated that toxicity in the final treated effluents was negligible, having been reduced following treatment of the effluent. Of the species tested, the bacteria in the Microtox® test had the highest sensitivity. The authors also note that differing sensitivities may not be fully attributable to the species, but also to the exposures achieved within the test systems. For example, the Microtox® test may be the most sensitive due to its rapidity and use of glass cuvettes that minimise losses via adsorption. Further, the results seen in the toxicity tests correlated to the measured bioavailable hydrocarbons using BE-SPME. The authors recommend that BE-SPME can be combined with simple bacterial bioassays to present an attractive screening tool for whole effluent assessments.

Implications for this work:

- Microtox® was quick and simple to implement.
- The sensitivities of the test organisms to oil refinery effluent were as follows: bacteria > daphnids > algae > fish.
- The use of complementary methods, such as BE-SPME can support mechanistic interpretation, and thus strengthen the results, of toxicity bioassays.
- Confounding factors identified during the evaluations may not limit the use of the test; bioassays could be conducted on both raw and XAD® resin-extracted samples.

Wang *et al*. [159] Petroleum refinery effluent contribution to chemical mixture toxic pressure in the environment.

This work assessed the contribution of petroleum refinery effluents to mixture pressure in the receiving environment. The multisubstance potentially affected fraction of species (msPAF), which represents the estimated affected fraction of species at a certain chemical mixture exposure was used as an indicator. msPAF values were calculated for undiluted and receiving water-diluted effluents, using measured chemical concentrations, species sensitivity distributions, and dilutions. The msPAF estimations were compared to the toxicity observed in the bioassays of Whale *et al*. [5], i.e. bacteria (*A. fischeri*) (acute), algae (*R. subcapitata*) (chronic), daphnids (*D. magna*) (chronic), and zebrafish embryos (*D. rerio*) (acute).

In line with other work, the toxicity of refinery effluents in receiving environments was negligible; average msPAF-chronic and msPAF-acute levels of refinery effluents at discharge points were 69% (P50) and 40% (P95), respectively, and levels were reduced substantially <5% downstream. Hydrocarbons (mainly total petroleum hydrocarbons) and inorganics (mainly ammonia) explained at least 85% of the mixture toxic pressure. Although generally low, the calculated msPAFs were explained mostly by the aromatic constituents (C_{10}-C_{15}). Estimated msPAFs at discharge points seemed to be conservative compared to the observed toxicity in bioassays.

Low molecular-weight monoaromatic hydrocarbons and PAHs (generally C10-C15) had higher percentage hazardous unit contributions, likely due to their relatively higher water solubility and thus bioavailability. The contribution of other organics, such as phenols, was negligible. These results support the observation that PAHs and ammonia in refinery effluents are the most likely cause of toxic effects to algae, invertebrates and fish based on toxicity tests.

Implications for this work:

- PAHs and low molecular weight monoaromatic hydrocarbons are the most likely cause of toxicity in refinery effluents, likely due to greater bioavailability, and thus the sensitivity of the evaluated bioassays to these chemical groups should be considered.

Concawe [32] A review of toxicity testing conducted on European refinery effluents in 2010, 2013, 2016, and 2019.

This work summarises the results of water use and effluent quality surveys completed for the 2010, 2013, 2016, and 2019 reporting years. It provides an insight into the types of toxicity tests being carried out in European refineries, their frequency of use, the motivation behind their conduct, and the type of samples (freshwater vs marine/brackish water) being tested. By conducting surveys over multiple years, the report also provides valuable information on the temporal trends seen in toxicity testing of European refinery effluents, perhaps showing trends for future years. The survey encompasses data from refineries across 28 EU countries, as well as Norway, Switzerland, and the UK. Response rate decreased each year, but remained relatively high, with 71% of refineries responding in 2019.

Of the responding refineries in 2019, 37 out of 100 (37%) in 2010, 27 out of 79 (34%) in 2013, 18 out of 76 (24%) in 2016, and 24 out of 65 (37%) in 2019, indicated that at least one of their effluent streams had some form of toxicity test performed. Of these, toxicity testing was predominantly conducted to fulfil permit requirements or to comply with regulatory demands. Annually and monthly testing were the most common monitoring frequencies reported, followed by quarterly. Overall, more sites conducted tests on effluents discharged into salt/brackish environments (58 sites) than into freshwater ones (49 sites), however by 2019, more sites (15 sites) were freshwater than salt/brackish water (9 sites). Tests with *Daphnia magna* were the most frequently reported every survey year, followed by luminescent bacteria (*Vibrio fischeri*) and fish eggs (*Danio rerio*). Tests with benthic invertebrates and algae followed behind. Other less commonly used tests included those with fish, bacteria, plants, genotoxicity and amoeba.

Implications for this work:

- Regulatory demands were the main motivation behind the toxicity testing, it is therefore anticipated that the REF BREF revision will motivate a larger number of refineries to conduct toxicity testing, highlighting the importance of this work.
- There were a larger number of effluents discharged to salt- or brackish water than freshwater, therefore the recommended battery should include *in vivo* tests applicable to saline water as well as freshwater, and consider the need for osmolarity adjustment of *in vitro* tests in the evaluations.
- Bivalves and/or echinoderms were not used.
- The frequency of testing (i.e. most commonly annually and monthly) will have implications for the costs deemed feasible for each test.

Concawe [30] Review of effect-based methods in the context of refinery effluents.

This work reviews and evaluates the use of *in vivo* and *in vitro* effect-based methods for the assessment of refinery effluents. A number of evaluation criteria were selected, similar to that of [27]. These included commercial availability, general validation maturity, sensitivity, robustness, suitability for routine and regulatory applications, interpretation of assay results, relevance of assay outcomes, previous application to environmental samples especially in response to PAHs, applicability to (passive sampler) extracts.

Based on this evaluation, a battery of five tests is presented: a luminescent bacteria assay in 96-well plates, an acute/chronic growth inhibition assay with green algae in glass vessels, a chronic water flea assay in glass vessels, the *umuC* genotoxicity assay in 96-well plates and an AhR activation assay in 96-well plates. The authors also discuss sampling options, ultimately recommending directly sampling effluents with an initial sampling frequency of quarterly in order to account for seasonal variabilities. Composite samples are recommended for native testing, or the use of large volume solid-phase extraction for the *in vitro* assays.

Implications for this work:

- The evaluation criteria cover a wide range of useful considerations that should be represented in this work too.
- A notable exception from the evaluation was fish-based assays (including fish embryo tests), therefore there is worth in evaluating them in this work.
- A good battery should capture a broad range of different compounds eliciting effects via different modes of action.
- Passive sampling devices may need several configurations to cover the various chemicals, therefore permanent automated samplers may be more suitable.

APPENDIX 2 DETAILED EVALUATION TABLES FOR *IN VIVO* AND *IN VITRO* TOXICITY TESTS

Table A2-1: Evaluation of *in vivo* tests.

	Toxicity to *Allivibrio fischeri* e.g. MicroTox®	Multi-species microbial toxicity	Algal growth inhibition	Combined algal assay	*Daphnia* immobilisation	*Daphnia* reproduction	*Daphnia magna* metabolic activity	Amphipod (*Corophium volutator*) lethality	Fish embryo toxicity test (FET)	Zebrafish qFET	Zebrafish Toxarray	Bivalve embryo development	Echinoderm embryo development	Nematode growth and reproduction
Acute or chronic	Acute	Acute	Chronic	Acute	Acute	Chronic	Acute	Acute	Acute	Acute	Acute	Chronic	Chronic	Chronic
Marine samples, freshwater samples, or both	Both	Both	Both	Freshwater	Freshwater	Freshwater	Freshwater	Marine and brackish	Freshwater	Freshwater	Freshwater	Both	Marine	Freshwater
Trophic level	Decomposers	Decomposers	Primary producers	Primary producers	Primary consumers	Primary consumers	Primary consumers	Primary consumers	Secondary consumers	Secondary consumers	Secondary consumers	Secondary consumers	Secondary consumers	Secondary consumers
Test organism	Bacteria	Bacteria, yeast and fungi	Algae	Algae		Crustacean	Crustacean	Crustacean	Fish embryo	Fish embryo	Fish embryo	Mollusc	Echinoderm	Nematode
Species	*Aliivibrio fischeri*	Multiple	Multiple	*Raphidocelis subcapitata*	*Daphnia*	*Daphnia*	*Daphnia*	*Corophium*	*Danio rerio*	*Danio rerio*	*Danio rerio*	Mussels or oysters	Urchins	*Caenorhabditis elegans*
Endpoint	Cytotoxicity	Cytotoxicity	Growth inhibition	Growth inhibition and photosystem II inhibition	Immobilisation	Inhibition of reproduction	Inhibition of metabolism	Lethality	Developmental toxicity	Developmental toxicity	Genotoxicity	Developmental toxicity	Developmental toxicity	Inhibition of growth and reproduction
Application of test in a regulatory context	Yes (IED)	Limited	Yes (IED)	Limited	Yes (IED)	Yes	No	Yes (OSPAR, but sediment only)	Yes (IED)	Yes	No	Yes	Yes	Yes
Application of test in European refinery effluents	Yes	No	Yes	No	Yes	No	No	No	Yes (ISO 15088)	Yes (ISO 15088)	No	No	No	Yes (benthic invertebrate)
Commercial availability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Standardisation	ISO 11348-3	No	ISO 8692 (freshwater) and 10253 (marine)	ISO in prep	ISO 6341	OECD 211	No	ISO 16712 ASTM E1367-99	ISO 15088	ISO 15088	No	ISO 17244	ASTM E1563-21a	ISO 10872
Validation maturity (based on [33])	High (routine lab)	High (routine lab)	High (routine lab)	Medium (expert lab)	High (routine lab)	High (routine lab)	Medium (expert lab)	High (routine lab)	High (routine lab)	High (routine lab)	Low (research lab)	High (routine lab)	Medium (expert lab)	High (routine lab)
Potency of effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sensitivity	High	High	High	Variable	High	Variable		Variable	Variable	Variable	Variable	High	High	Variable
Use in environmental water samples	Very common	Common	Very common	Occasional	Very common	Very common	Rare	Common	Very common	Very common	Occasional	Very common	Common	Occasional

	Toxicity to *Allivibrio fischeri* e.g. MicroTox®	Multi-species microbial toxicity	Algal growth inhibition	Combined algal assay	*Daphnia* immobilisation	*Daphnia* reproduction	*Daphnia magna* metabolic activity	Amphipod (*Corophium volutator*) lethality	Fish embryo toxicity test (FET)	Zebrafish qFET	Zebrafish Toxarray	Bivalve embryo development	Echinoderm embryo development	Nematode growth and reproduction
Response to refinery effluent constituents	Yes B(a)P, benzo(b)fluoranthene, naphthenic acids, class-fractionated oils.	Yes PAHs and crude oil.	Yes B(a)P, benzo(b)fluoranthene, cracked gas oil .	Yes Oil.	Yes B(a)P, benzo(b)fluoranthene, PAHs, cracked gas oil .	Yes Diesel, biodiesel, cracked gas oil		Yes Oil	Yes B(a)P, benzo(b)fluoranthene, cracked gas oil	Yes (weak)		Yes Phenanthrene, naphthalene, pyrene, fluorene	Yes Phenanthrene , naphthalene, pyrene, fluorene , fluoranthene, crude oil	Yes (weak)
Response reported in test with refinery effluents	Yes		Yes		Yes	Yes (weak)		Yes	Yes (weak)	Yes		Yes		Yes
Applicability to passive sampler extracts	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	Yes
Time to run test	Results within a day; ≤30 m	Results within a day; 18-24 h	Results within a week; 72 h	Results within a week; 24 h	Results within a week; 48 h	Results within a week; 7 d	Results within a day; 1 h 15 m	Results within 2 weeks; 10 d	Results within a week; 96 h	Results within a week; ≤120 h	Results within a week; ≤120 h	Results within a week; 24 h	Results within a week; 48 h	Results within a week; ≤96 h
Trigger values proposed in academic literature	0.05 Toxic Units [38]; 1.2 mg baseline toxicant equivalent /L [39]		0.05 TU [38]; 116.5 ng Diuron equivalent /L [39] [13]	0.07 µg diuron equivalents /L [39]; 130 (growth) and 73.7 (PSII) ng Diuron equivalent /L [39] [13]	0.05 TU [38]; 15 ng Chlorpyrifos equivalents /L [39]	0.05 TU [38]			0.183 mg BPA equivalent /L [39] [13]					
Bioassay category	2	Assume 2	2	1	1	Assume 2	Assume 2	Assume 2	2	2	Assume 2	Assume 2	Assume 2	Assume 2
Chemical analysis replacement	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No
Cost per sample (€)	<100 to <1,000 (survey)	1,000 to <3,000 (survey)	<100 to < 10,000 (survey)	Unknown	>100 to <10,000 (survey)	>1,000 to >10,000 (survey)	Unknown	>100 to <10,000 (survey)	>1,000 to >10,000 (survey); <200 [28]	Unknown	Unknown	>1,000 to <10,000 € (survey)	>1,000 to <10,000 € (survey)	>1,000 to <3,000 € (survey)
Relevance of effects to refinery effluents	High	High	High	High	High	High	Medium	Medium	High	High	High	High	High	Medium
Confounding factors	COD, colour, turbidity	COD, colour, turbidity	COD, ammonium, organic matter, alkalinity, hardness, nutrient concentrations, species used	COD, ammonium, organic matter, alkalinity, hardness, nutrient concentrations, species used	pH, oxygen content	pH, oxygen content		Salinity, oxygen concentration, ammonium	pH, oxygen content, plastic apparatus	pH, oxygen content, plastic apparatus		Osmotic stress	Osmotic stress	

Table A2-2: Evaluation of *in vitro* tests.

In vitro methods	*umuC*	Ames	Ames II/ fluctuation	Micronucleus	Comet	hGADD45 activation	p53-pathway activation	AhR activation	PAH-specific AhR activation	EROD	AREc32 activation	Nrf2-pathway activation	Cytotoxicity in RT Gill-W1	(Arxula-)Yeast Estrogen Screen ((A-)YES)	ER activation	Yeast Androgen Screen (YAS)	AR activation	TTR binding
Test organism	Bacteria (modified *Salmonella typhimurium*)	Bacteria (recombinant *S. typhimurium*)	Bacteria (recombinant *S.typhimurium*)	Hamster cell line	Human, other mammalian and fish cell lines	Yeast or human cell line	Human cell line	Rat liver cell line	Rat liver cell line	Fish cell line	Human cell line	Human cell line	Rainbow trout cell line	Yeast (recombinant *Saccharomyces cerevisiae* or *Arxula adeninivorans*)	Human cell line	Yeast (recombinant *S. cerevisiae*)	Human bone marrow or hamster cell lines	Any cells containing transthyretin
Mode of action	Genotoxicity (DNA damage response)	Genotoxicity (gene mutations)	Genotoxicity (gene mutations)	Genotoxicity (chromosomal mutations)	Genotoxicity (DNA damage response)	Genotoxicity (DNA damage response)	Genotoxicity (DNA damage response)	Metabolism (AhR receptor agonists)	Metabolism (AhR receptor agonists)	Metabolism (Cytochrome P450 1A activity)	Oxidative stress (induction of Antioxidant Response Element pathway)	Oxidative stress (induction of ARE pathway)	Cytotoxicity	Endocrine disruption ((anti)estrogenic activity)	ED ((anti)estrogenic activity)	ED ((anti)androgenic activity)	ED ((anti)androgenic activity)	ED (displacement of T4)
Application of test in a regulatory context	Yes													Yes		Yes		
Commercial availability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Standardisation	ISO 13829	ISO 16240	ISO 11350	ISO 21427	OECD 489	No	No	ISO in prep (ISO/CD 24295)	No	No (ISO/TS 23893-2 for *in vivo*)	ISO in prep	No	ISO 21115	ISO 19040-1 (YES) or ISO 19040-2 (A-YES)	ISO 19040-3	ISO in prep	OECD 458 (hamster cell line)	No
Validation maturity (based on [33])	High (routine lab)	High (routine lab)	High (routine lab)	High (routine lab)	High (routine lab)	Medium (expert lab)	Medium (expert lab)	High (routine lab)	Medium (expert lab)	Medium (expert lab)	Medium (expert lab)	Medium (expert lab)	Medium (expert lab)	High (routine lab)	High (routine lab)	High (routine lab)	Medium (expert lab)	Low (research lab)
Potency of effect	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sensitivity	Variable	Variable	Variable	High	High	High	High	High	High	High	High	Variable	Variable	Variable	High	Variable	High	Variable
Use in environmental water samples	Very common	Very common	Very common	Common	Common	Occasional	Occasional	Common	Occasional	Common	Common	Occasional	Occasional	Common	Common	Common	Common (human) Occasional (hamster)	Rare

In vitro methods	*umuC*	Ames	Ames II/ fluctuation	Micronucleus	Comet	hGADD45 activation	p53-pathway activation	AhR activation	PAH-specific AhR activation	EROD	AREc32 activation	Nrf2-pathway activation	Cytotoxicity in RT Gill-W1	(Arxula-)Yeast Estrogen Screen ((A-)YES)	ER activation	Yeast Androgen Screen (YAS)	AR activation	TTR binding
Response to refinery effluent constituents	Yes Metabolically activated B(a)P, other PAHs, fuel oil residues	Yes Metabolically activated B(a)P, class-fractionated oils	Yes Metabolically activated B(a)P, other effluent constituents	Yes B(a)P, benzo(b)fluoranthene, other effluent constituents	Yes Oil-containing drill cuttings	Yes Metabolically activated B(a)P	Yes Metabolically activated B(a)P	Yes B(a)P, benzo(b)fluoranthene, crude oil, other refined petroleum products	Yes B(a)P-like compounds	Yes B(a)P, 3-methylcholanthrene, benzo(b)fluoranthene, chrysene and benzo(a)anthracene	Yes B(a)P, benzo(b)fluoranthene	Yes Crude oil	Yes 2- and 3-ring PAHs, creosote	Yes Crude oils, PAHs	Yes B(a)P, benzo(b)fluoranthene, crude and refined oils	Yes Oil SARA fractions, crude oil	Yes B(a)P	Yes PAH quinones, hydroxides
Response reported in test with refinery effluents		Yes	Yes	Yes				Yes		Yes				Yes		Yes		
Applicability to passive sampler extracts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time to run test	Results within a day; 2 h	Results within a week; 48 h	Results within a week; 48 h	Results within a week	Results within a week	Results within a week	Results within a week	Results within a week	Results within a day; 6 h	Results within a day; 3 h	Results within a week	Results within a week	Results within a week; up to 72 h	Results within a week; 48 or 72 h	Results within a week; 24 h	Results within a week; 72 h	Results within a week; 24 h	Results within a day
Trigger values proposed in literature	0.005 genoTU [38]	0.005 genoTU [38]	0.005 genoTU [38]	0.005 genoTU [38]	0.005 genoTU [38]	0.005 genoTU [38]	0.005 (geno)TU [38] [54]	16.2-50 pg TCDD equivalents /L [38] [54] [72]	6.21-150 ng B(a)P equivalents /L [38] [54] [39]		156 µg - 1.4 mg Dichlorvos equivalents/ L [39] [210]; EBT-Effective concentration induction ratio (EC-IR4) 1.5 = 6 REF [200]	10 µg curcumin equivalents /L [38]; 26 µg Dichlorvos equivalents /L [39]	0.05 TU [38]	0.2-0.4 ng E2 equivalents /L [243]. For the A-YES, 0.4-0.56 ng E2 equivalents /L [39] [13]	0.1-0.5 ng E2 equivalents /L [72] [259] [260] [243] [39] [38] [13]		25 mg flutamide equivalents /L [38]	49-58 ng T4 equivalents /L [13, 39]
Bioassay category	Assume 2	Assume 2	Assume 2	Assume 2	Assume 2	Assume 2	Assume 2	2	2	2	2	2	Assume 2	1	1	1	1	1
Chemical analysis replacement	No	No	No	No	No	No	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes
Cost per sample (€)	<100 [34]	100-1,000[34]; 3,000 - <10,000 (survey)	100-1,000 [34]; <200 [28]	100-1,000 [34]; <200 [28]; ≥10,000 (survey)	<100 [34]		100-1,000 [34]	100-1,000[34]; <200 [28]; 300 - <1,000 (survey)	<200 [28]	<200 [28]; 300 - <1,000 (survey)	<100 [34]	<100 [34]; 3,000 - <10,000 (survey)		100-1,000 [34]; <200 [28]; 1,000 - <3,000 (survey)	100-1,000 [34]; <200 [28]; 1,000 - <3,000 (survey)	100-1,000 [34]; <200 [28]; 1,000 - <3,000 (survey)	100-1,000 [34]; <200 [28]	

In vitro methods	*umuC*	Ames	Ames II/ fluctuation	Micronucleus	Comet	hGADD45 activation	p53-pathway activation	AhR activation	PAH-specific AhR activation	EROD	AREc32 activation	Nrf2-pathway activation	Cytotoxicity in RT Gill-W1	(Arxula-)Yeast Estrogen Screen ((A-)YES)	ER activation	Yeast Androgen Screen (YAS)	AR activation	TTR binding
Relevance of effects to refinery effluents	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	High	Medium	Medium	Medium	Medium	Medium	Low	Low	Medium
Confounding factors	Strain used	Strain used, microbial contamination, extraction method	Strain used, extraction method	Cell line used, subjective evaluation	Cell line used, cytotoxicity	Particulate matter, bacterial contamination, cytotoxic effects	Cytotoxicity	AhR antagonists, cytotoxic effects, exposure time	AhR antagonists, cytotoxic effects	Protein concentration; Food; Inhibiting substances; Cytotoxic effects	Cytotoxic effects	Cytotoxic effects, cell line used	Metabolic activation, water solubility, lipophilicity, osmotic stress	Anti-estrogenic substances, receptor used, chlorination	Anti-estrogenic substances, receptor used, chlorination	Anti-androgenic substances, cytotoxicity	Anti-androgenic substances, cytotoxicity	

Concawe
Environmental Science
for European Refining

Concawe
Boulevard du Souverain 165
B-1160 Brussels
Belgium

Tel: +32-2-566 91 60
Fax: +32-2-566 91 81
e-mail: info@concawe.org
http://www.concawe.eu

ISBN 978-2-87567-171-4