Maastricht University

Maastricht Science Programme

MSC Coverpage Assignment

Project code:	**PRO3001**
Project Title:	Volatile Organic Compounds: what does your body tell you about your health?
Supervisors:	**Egon Willighagen** **Friederike Ehrhart**
Student Name:	Defne Bilgin, Chidera Chukwumah, Lizzy Godynyuk, Eliis Kalbus, Kunsulu Nurekeyeva, Jannika Oeke, Steve Suthakaran, Olivier Traets, Cristina Trauffler, Jacob Windsor
Student ID nr.:	i6092850, i6081299, i6080733, i6093028, i6085735, i6094284, i6082276, i6066163, i6089414, i6083200
Academic Year:	**2015-2016**
Number of Words:	**5737**
Title Assignment:	Volatile Organic Compounds: a detailed account of identity, origin, activity and pathways.

Date: **28.06.2016**

Volatile Organic Compounds: a detailed account of identity, origin, activity and pathways.

INTRODUCTION (In recent times the advancement of technology in medicine has allowed us to bring forth new and effective treatments to counteract various diseases. In spite of the success of modern treatments, a more effective approach may lie in an earlier stage, the diagnosis. An early diagnosis requires the ability to identify substances that unambiguously correlate with the disease (Saalberg & Wolff, 2016). This would be the detection of indicators – biomarkers – for the onset and development of a disease before it becomes detrimental to the body. Currently the use of diagnostic tools is at the forefront of this new emerging field. One such diagnostic tool, that is the focus of this paper, is metabolomics, which is the study of the metabolites present in all organisms. These metabolites represent the unique chemical fingerprints that the processes in our bodies leave behind. The metabolites can be detected outside of the body as volatile organic compounds (VOCs) by various means. VOCs can be classified into several groups including aromatic compounds, sulfur volatile compounds, nitrogen containing compounds and fatty acids (Schulz & Dickschat, 2007). However, even if the technology and methods are adequate to detect these indicators, it is important to note that the biggest shortcoming is the accurate interpretation of the data.

According to the World Health Organisation's (WHOs) definition, biomarker is "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" (International Programme on Chemical Safety, 2001). The use of VOCs as biomarkers is not a novel approach. These compounds were first used in 1964 by Geldreich and his coworkers to detect coliform bacteria based on the the production of indole, acetoin, pyruvate and 2,3-butanediol in culture media (Geldreich, Kenner & Kabler, 1964). However, in recent years detection of VOCs as biomarkers for fatal diseases such as cancer have made a lot of progress. The findings of research on lung and colorectal cancer show that the presence of various metabolites can distinguish diseased patients from healthy individuals (Arasaradnam et al., 2014; Hakim et al., 2012). For example, several research groups have confirmed 2-butanone and 1-propanol as best suited discriminators in breath for lung cancer (Saalberg & Wolff, 2016).

There are several sources where VOCs can be detected. Mainly, the compounds can be detected in exhaled breath, saliva, blood, milk, skin secretions, urine, and faeces (De Lacy Costello et al., 2014). In research, the most common and efficient method used for testing volatile biomarkers is the analysis of exhaled breath (Sethi, Nanda, & Chakraborty, 2013). Breath tests are replicable and non-invasive, which makes it a patient friendly option, allowing more follow-up studies leading to a higher amount of data, which can be obtained.

This paper aims to expand the previous study of De Lacy Costello et al. (2014) hence will explore the origin and possible bioactivity of VOCs in the dataset provided by a topical review of the volatiles from the healthy human body. This dataset includes 1843 compounds, where 1732 compounds have unique CAS registry numbers. The goal was to link both endogenous and exogenous VOCs to specific pathways, focusing on the compounds that did not have any mappings. From the unmapped dataset, which

incorporated 819 compounds, 100 compounds were researched. As the used dataset included a list of volatiles emanating from healthy individuals, the results from this study will mostly contribute to the knowledge of what is normal, in order to further assess what is abnormal.

Most importantly, VOCs which were found to have pathways can be added to the online platform WikiPathways (Kelder et al., 2011). WikiPathways is an easily accessible portal that supports the notion of open science. Since 2008, biological pathways of 25 species have been archived, adding up to 2300 pathways in total. Consequently, any contribution made to WikiPathways will allow for information exchange between members of the scientific community and improve current knowledge on VOCs. In addition, new entries of VOCs can be made in Wikidata (Wikimedia Foundation, 2016), a similar online, open data platform. Hence, the use of the two platforms together will reinforce a greater depth of knowledge and accessibility.

MATERIALS AND METHODS

RANKER PROGRAMME Many metabolites in the unmapped dataset had very little or no information regarding the function or metabolic pathway(s) readily available through online or physical mediums. This presented the need to filter the metabolites by the amount of data available for it. Thus, a ranking programme was developed that ranks the dataset by the amount of bioassays or biosystems found in PubChem (Kim et al., 2016). The developed programme is publically available under MIT license (Windsor, 2016) .

The software is written using the Python 3 scripting language (Python Software Foundation, 2016) and built upon the Flask web-micro-framework (Ronacher, 2016). The software provides a command-line interface (CLI) for seeding and ranking the data and a web-based interface for viewing the ranked dataset with links to each compound's PubChem page. In order to run the software, the user initially seeds an SQLite (Hipp, 2016) database with CAS registry numbers and the corresponding IUPAC names by running a command that accepts a comma separated value (CSV) file and the desired name for the dataset. Following this, another command is run that uses the representational state transfer (REST) application programming interface (API) for PubChem's power user gateway (PUG) to gather all of the compound identifiers (CIDs) for each CAS registry number and save them to the dataset. Finally, a command is then run which uses PubChem's SDQ agent to gather and save the bioassay and biosystem counts for each CID (Kim et al., 2016). Biosystems in PubChem are groups of molecules that interact in a biological system, this includes biological pathways and diseases (Wang et al., 2013a). Thus, biosystems can be used to identify any metabolic or disease pathways associated with a VOC. When accessing the web-interface the dataset is ranked by the number of biosystems or bioassays. The web-based interface was made publically available through Python Anywhere (Python Anywhere, 2016) so it could be easily accessed by other group members.

Using the ranked list (Windsor, 2016), 100 VOCs with no mappings to WikiPathways were selected. Following this, an extensive search for the origin, biological effect(s), and pathway(s) of each compound was commenced. Multiple databases were utilised to during this process.

DATABASES USED

PUBCHEM AND CHEMSPIDER PubChem and ChemSpider work side by side as databases containing valuable background information on chemical compounds with

ChemSpider being opened to PubChem through patent links. PubChem is a database offered by the national center for biotechnology information (NCBI). As a database of chemical molecules, PubChem contains descriptions of smaller molecules with fewer than 1000 bonds and 1000 atoms (Kim et al., 2016). ChemSpider looks into the chemical structure of each chemical compound helping to provide physical properties of the compound searched and also providing literature references, aiding in piecing together where exactly these compounds may have come from and what their uses are (Pence & Williams, 2010). In order to obtain information names, synonyms and CAS registry numbers were put into the search functions and notes were taken on the relevant information.

CHEMICAL ABSTRACTS SERVICES (CAS) AND SCIFINDER CAS and the SciFinder database were also used to identify and group together VOCs based on their origin and purpose. The database contains 7900 chemicals (Sommerville 1998), and allowed matching of the CAS numbers stated in the unmapped metabolite sub-dataset to the chemical compound in order to check whether the CAS number for the specific compound listed is correct. Once the CAS number or name is typed into the search engine, the database presents a plethora of literature, journals, papers etc. on the compound at hand, taking into account wherever the compound has been mentioned (Sommerville 1998) . In order to refine and reduce the number of results, certain filters are applied such as looking into more subject specific areas – chemistry, biology etc. in order to identify and find out exactly what each individual VOC's primarily purpose is.

HUMAN METABOLOME DATABASE (HMDB) HMDB contains detailed information about every small molecule found in the human body. From a description of the clinical and chemical history of the compound to the molecular biology and biochemistry, this database serves to link up all three areas of interest so that it can be used for applications in metabolomics, biomarker discoveries and clinical chemistry (Wishart et al., 2013). In addition to this, filters can be applied during the search for the metabolite in order to show and group together, where most of the VOCs originated from. For example, whether these metabolites originated in plants or food and so are found in the human body because of what we eat and drink. Toxins, pollutants and drugs are also categorised showing that the compounds detected could be present in the human body due to what people have ingested. In order to locate the source in which it was found, HMDB maps out which fluid in the body the metabolite was discovered in. This includes blood, urine, saliva and cerebrospinal fluid (Wishart et al., 2013).

DATABASE OF CHEMICAL ENTITIES OF BIOLOGICAL INTEREST (ChEBI)

ChEBI is a database for molecular metabolites that focuses on smaller chemical entities. Its major feature is that it permits the relationships between classes of chemical entities to be shown in a structured way – presenting a good starting point for identifying pathways (Hastings et al., 2013).

BIOCYC PATHWAY/GENOME DATABASE

BioCyc is a database with a collection of 7615 genomes and pathways which is able to describe a variety of sequenced genomes. The biochemical pathways as well as the network it forms are described, improving the understanding of many biological systems and aiding in areas such as biotechnology and drug development (Caspi et al., 2016). Biocyc has pathway/genome databases which include Metacyc - which describes a large proportion of metabolic pathways, enzymes and organisms, EcoCyc

(Karp et al., 2014) – which focuses on the E Coli gene and its many pathways and HumanCyc (Romero et al., 2005). HumanCyc provides a collection of human metabolic pathways which is formed through the analysis of the human genome and literature based research. As the VOCs that were being investigated were found on humans, the chosen VOCs were investigated using HumanCyc in order to find pathways it might be involved in, resulting in more information about the specific compound.

KYTO ENCYCLOPEDIA OF GENES AND GENOMES (KEGG) KEGG Pathway is a database that manually maps out pathways and interactions between different molecules in a compound and in network reactions (Kanehisa & Goto, 2000). These networks of reactions include categories such as metabolic reactions, diseases in humans and environmental processes. The main aim was to investigate whether the selected VOCs are mapped out to networks, enabling them to be interpreted on a biological level. Furthermore, some of the compounds have a potential to act as biomarkers of certain human diseases, benefitting identification and treatment possibilities of these diseases in the future.

WIKIPATHWAY ADDITIONS Through the use of the ranker programme an attempt was made to add all the pathways for each VOC to WikiPathways. By ranking the VOCs by pathways the compounds with any biosystems listed on PubChem were quickly identified. Using PubChem's list of biosystems, pathways could be identified through KEGG, BioCyc or literature. Subsequently, new WikiPathway entries could be generated using the Java based creation tool (Pico et al., 2008).

WIKIDATA ADDITIONS Once all data on each VOC was collected a Wikidata entry for every compound was made. All found mappings, such as the CAS registry number, the CID ID, the InChIKey and others were included as identifiers per compound. Furthermore the VOCs were classified as human metabolites, due to their presence either in or on a human being. Information about the origin of the compound, the potential use as a biomarker and pathways were uploaded if applicable and backed up with references. These references are not separately listed in the report but can be found on the wikidata page of the compound of interest. The tool Source MetaData (Manske, 2016) was used in order to create a Wikidata page for each source before including them as references. For those compounds were a Wikidata page was already present novel information was added.

RESULTS

From the 819 unmapped compounds from the original dataset provided by de Lacy Costello et al. (2014), 100 VOCs were further investigated. All information that was found regarding these metabolites was documented in tables and the CAS registry number as well as the name of each compound is included in all tables. *Table 1* from which an excerpt is presented below states the origin of the compound, it's chemical class and where it was measured. The complete list can be found in the *Appendix A*.

The researched VOCs were also listed in a table (*table 2*) including different identifiers for the compound of interest as well as pathway(s) description and disease information if found. An excerpt is presented below and the complete table can be found in *Appendix A*.

Table 1. Excerpt. List of the 100 VOCs from the unmapped dataset showing the compounds origin, chemical class and where it was measured.

CAS	name	Wikidata ID	Origin	chemical class	Faeces	Urine	Breath	Skin	Milk	Blood	Saliva
75-35-4	1,1-dichloroethene	Q161284	food packaging, pollutant	alkene						Bl	
1334-78-7	tolualdehyde ui	Q2439424	polluting agent, derived from petroleum, reguarly detected in air	benzenoid			Br				
500-00-5	p-menth-3-ene	Q24716505	found in herbs and spices: isolated from thyme and peppermint oil;pollutant; related to smoking	cycloalkene			Br				Sa
105-21-5	gamma-heptalactone	Q24730325	possible metabolite used in fungal quorum sensing/xenobiotic, food, flavouring agent	lactone				Sk			Sa
106-55-8	2,5-Dimethylpiperazine	Q22079573	Attractant for the Yellow Fever Mosquito (Aedes aegypti)	heterocyclics				Sk			

Table 2. Excerpt. List of the 100 VOCs from the unmapped dataset showing different identifiers including pathway ID's, pathway description and disease information.

CAS	name	CID	ChEBI	InChI-Key	KEGG Pathway ID	Wiki Pathways ID	HMDB ID	Pathway(s) description	Disease	Wikidata ID
75-35-4	1,1-dichloroethene	6366	34031	LGXVIGDEPROXKC-UHFFFAOYSA-N	ko00980	WP3666		hsa00980 "metabolism of xenobiotics by		Q161284

								CYP450" homo sapiens		
1334-78-7	tolualdehyde ui	10722		BTFQKIATRPGRBS-UHFFFAOYSA-N	ko01220	WP3667			Degrades from o-xylene which is toxic, A human health hazard, can cause fatigue, confusion, headaches, dizziness, and even death.	Q2439424
500-00-5	p-menth-3-ene	10369	88834	YYCPSEFQLGXPCO-UHFFFAOYSA-N			HMDB37213			Q24716505
105-21-5	gamma-heptalactone	7742	89744	VLSVVMPLPMNWBH-UHFFFAOYSA-N			HMDB31681	Phase I biotransformations, non P450 (*Homo sapiens*)	Chronic Granulomatous Disease	Q24730325
106-55-8	2,5-Dimethylpiperazine	7816		NSMWYRLQHIXVAP-UHFFFAOYSA-N						Q22079573

To visualize the information from *table 1*, histograms were designed, comparing the VOCs chemical class (*figure 1*) and their origin (*figure 2*) between the different detection locations.

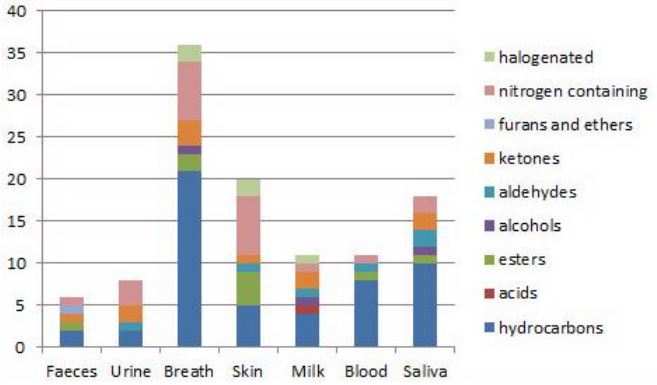

Figure 1. *The number of compounds in each class (hydrocarbons, acids, esters etc.), that have been found in faeces, urine, breath, skin, breast milk and saliva.*

Figure 1 shows that of the total identified and researched compounds, breath was found to have the highest quantity, whereas faeces was lowest. Hydrocarbons were found to be most abundant in all areas of origin, with breath containing 21 separate additions. Ketones, nitrogen containing compounds, and hydrocarbons were shown to be the most widespread throughout the body. Acids were only found in milk.

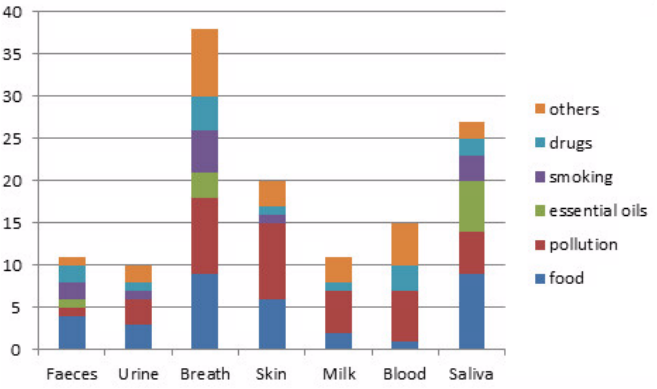

Figure 2. *The different sources of chemical compounds (food, pollution. Essential oils, smoking, drugs and others) found in faeces, urine, breath, skin, milk, blood and saliva.*

The highest compound sources were pollutants, along with food sources, appearing in all endogenous sources in and on the human body (*figure 2*). Drugs were found to appear in all sources to a lesser extent. Its abundance may be attributed to its wide-spread function in the body, and multi-targets. An interesting discovery are compounds found to originate from smoking appearing in urine and faeces. This may be due to single compounds having multiple sources.

From the 100 analyzed VOCs, 70 were not yet in Wikidata. For these compounds an entry was made, including all the information that was found regarding identifiers, origin, potential usage and pathways. *Table 3* lists all new entries with the Wikidata ID, the CAS number of the compound and its name. Below an excerpt of the complete table, which can be found in the *Appendix B*, is shown.

Table 3. Excerpt. List of the 70 new Wikidata entries, stating the Wikidata ID, the CAS registry number and the name of the compound.

Wikidata ID	CAS	Compound Name
Q24702624	583-57-3	1,2-Dimethylcyclohexane
Q24514387	586-63-0	Isoterpinolene
Q24702742	590-66-9	1,1-Dimethylcyclohexane
Q24705698	58-90-2	2,3,4,6-Tetrachlorophenol
Q24705869	764-39-6	(E)-2-Pentenal
Q24705889	84-67-3	m-Tolidine
Q24710331	592-48-3	1,3-Hexadiene
Q24711948	598-25-4	1,1-Dimethylallene
Q24712449	598-94-7	1,1-Dimethylurea
Q24713391	469-92-1	beta-Clovene

Two pathways were identified and added to WikiPathways. *Figure 3* shows the

metabolism of 1,1-dichloroethene by cytochrome P450 (CYP450) in *Homo sapiens* as well as many other organisms. 1,1-dichloroethene was identified in breath and this degradation pathway was found in KEGG. O-tolualdehyde (IUPAC: 2-methylbenzaldehyde) was identified in blood and found to be a metabolite in the o-xylene degradation pathway in many prokaryotes including *Escherichia coli* (*figure 4*).

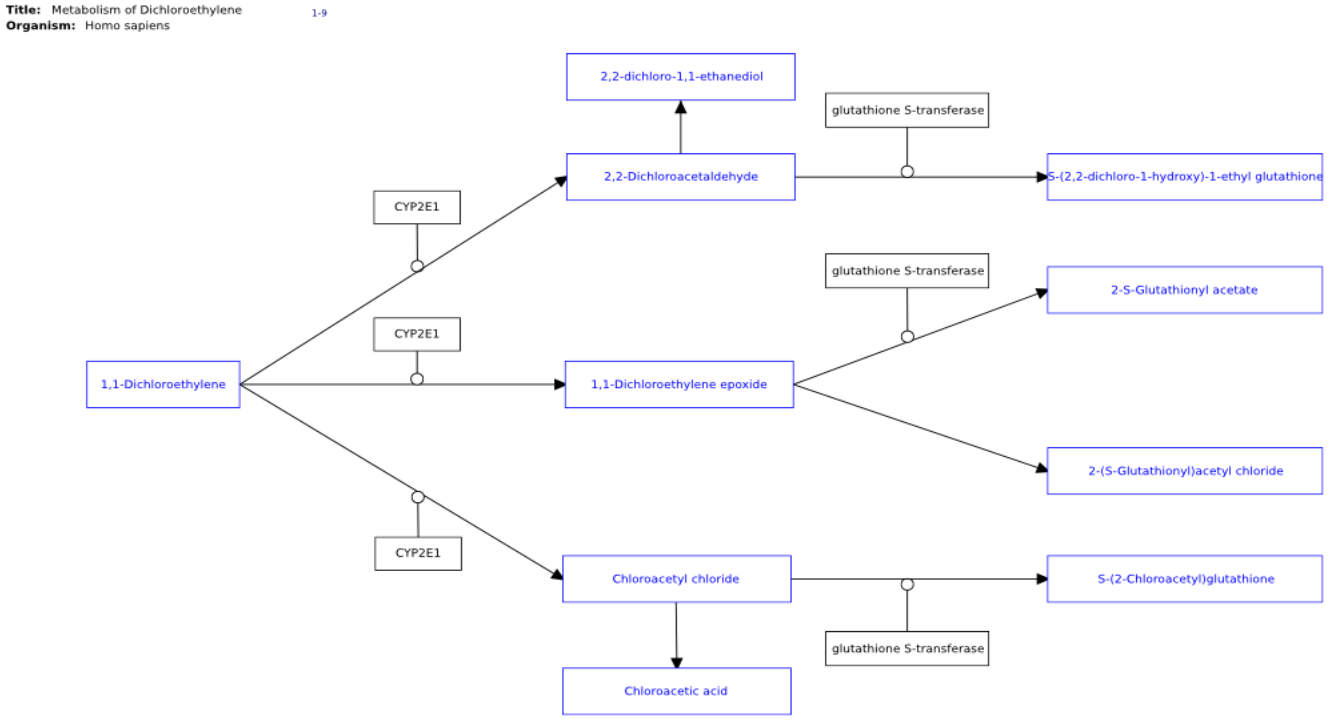

Figure 3. *Metabolism of dichloroethene by CYP450 (Homo sapiens). Source: http://www.wikipathways.org/index.php/Pathway:WP3666*

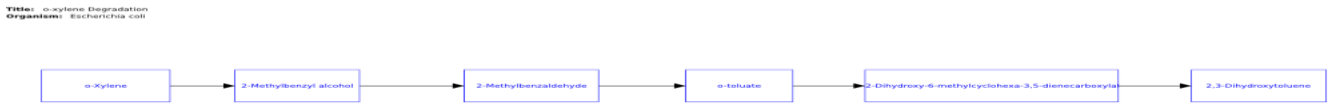

Figure 4. O-xylene degradation (*Escherichia coli*). *Source: http://www.wikipathways.org/index.php/Pathway:WP3671*

DISCUSSION

VOCs IN FAECES Very little is known about how VOCs in faeces could relate to human health and well-being.

Out of the 100 compounds analyzed, only 6 were present in faeces. While no pathways related to disease were found, they do play an important role in human metabolism. For example, a methyl ester of 2-methylbutanoic acid is linked to gut microbial action on the proteins (De Lacy Costello et al., 2014). Nitrogen-containing compounds have also been reported and were linked to the diet and dietary supplements; for instance, acetonitrile is a component of antibiotics and vitamins. Alkanes and alkenes were also previously reported to be found in faeces and generally originate from food or naturally occur in plants (Garner et al., 2007). One of the analyzed compounds, 1,5,5-Trimethyl-6-methylene-cyclohexene, for instance, was originated from celery and was also found in water dropwort (Deng et al., 2003). Phenolic compounds (acetophenone) have previously been proven to be products of the metabolism of aromatic amino acids by gut bacteria (De Lacy Costello et al., 2014). Lastly, trans-tetrahydro-2,5-dimethylfuran was found in tobacco, but also as mentioned by Garner et al. it could be related to a disease : 2 (or 3)-methyl furan compounds were found in healthy patients and their absence was positively associated with *Campylobacter* (Garner et al., 2007).

VOCs IN URINE VOCs found in urine cover a wide range of classes of chemical compounds, including hydrocarbons, aldehydes, ketones and nitrogen-containing compounds.

Most of the urine-derived VOCs analyzed in this paper were products of pollution, flavouring agents, food or tobacco-originated. Nevertheless, review of the literature shows that VOCs found in urine have been used in detection of metabolic disorders. Presence of specific compounds in high concentrations have been linked to a metabolic disorder, e.g. Reyes syndrome (excess ammonia) and trimethylaminuria (excess trimethylamine) (Amann et al., 2014). In addition, urine-derived volatiles have also been used to facilitate the detection of bladder, prostate, lung, colorectal, breast and esophageal cancer (Weber et al., 2011; Spanel et al., 1999; Carrola et al., 2011; Silva, Passos & Camara, 2012; Huang et al., 2013). VOCs found in urine were also reported to be efficient in the prediction of ovulation via ammonia and acetone levels (Diskin, Spanel & Smith 2003). In addition to this, sugars and ketones found in urine are used to monitor diabetes in patients and manage symptoms, determining if treatment is successful (Cafasso, 2015). Though not a form of detection for the disorder, it is nevertheless a solution to its management.

VOCs IN BREATH Exhaled human breath contains the largest amount of VOCs, compared to other sources. For a majority of those compounds, it is unclear whether they were produced in the human body endogenously or related to smoking, food and medicine consumption, pollution and etc. (De Lacy Costello et al., 2014). Out of the 100 analyzed compounds form this paper, 36 compounds were found in human breath, with a majority of them being hydrocarbons or nitrogen-containing compounds. Majority of the alkanes, alkenes, dienes and benzenoids found in human breath were proven to be related to smoking and pollution, with only a few of them being food-derived (Filipiak et al., 2012). For example, one of the compounds researched - 1,2,3,5-Tetramethylbenzene - was found in rice products (Maga, 1978). Ketones found in exhaled human breath were mostly related to the diet, as they arise from carbohydrate and fatty acid metabolisms (De Lacy Costello et al., 2014). Many essential oils that naturally occur in plants were detected in breath, possibly deriving from food and cosmetic

products. There were nitrogen-containing compounds detected as well. One of them is 4-Methylpentanenitrile and it was proven to be food-derived, as it was found in honey (Kaskoniene et al., 2008), cooked eggs and other food products (Umano et al., 1978). It should be noted that VOCs found in human breath contain the most compounds that were included in the category named ¨others¨, i.e. not originated from food, drugs, pollution, essential oil or smoking (Fig. 2). Some of them were found to be a part of the composition of pesticides (1,4-dichlorobenzene) and insecticides (1,2-dichlorobenzene and 1,3-dichlorobenzene), while other VOCs were derived from petroleum and gasoline, including 1,1-dimethylcyclohexane and tolualdehyde (See Appendix A, Table 1).

One particular compound found in breath, gamma-heptalactone, was of particular interest. Upon further inspection, Gamma-heptalactone was found to be produced as a form of signaling communication in *Aspergillus nidulans*, known as quorum sensing (Williams et al., 2012). Quorum sensing is the ability of bacteria, as well as fungi, to respond to fluctuations in the signal molecule and to, in return, regulate gene expression (Miller & Bassler, 2001). This particular strain of filamentous fungi can be found in the gut of humans when infected, though it is mostly broken down by the Paraoxonase enzyme, breaking the lactone ring (Yang et al., 2005). Chronic granulomatous disease (CGD) is a disease in which the immune system protects the body from pathogens but cannot remove it, causing inflammation, or granuloma (National Institute of Health, 2016). It has been shown that those suffering from this disease are more susceptible to *Aspergillus nidulans* (Henriet, Verweij & Warris, 2012). In the case of gamma-heptalactone, it can be used as a biomarker to aid in the diagnosis of CGD using VOC data from faeces, where the prevalence of the compound in this particular medium is less likely to be related to other sources producing it. According to HMDB, this compound can also have food origins, which may possibly explain its presence in breath (see Appendix A, Table 1).

When working with VOCs found in exhaled breath, it is important to emphasize the amount of exogenous sources from where these compounds originate. Nevertheless, diagnosis using VOCs found in breath has an advantage: breath can be sampled as often as required and can be analyzed in real-time (De Lacy Costello et al., 2014).

VOCs IN SKIN

20 out of the 100 researched VOCs were found in skin secretions. Most were classified as pollutants, followed by food, other sources, smoking, and drugs, in decreasing order. The largest chemical group was found to be nitrogen-containing, followed by esters, hydrocarbons, halogenated compounds, ketones, and aldehydes. However, many of these compounds are not necessarily endogenous. For example, N,N-Dimethyldodecylamine is found to be used in hair conditioning and as an antistatic (Verheugen, 2006) and was present on skin.

De Lacy Costello et al. (2014) describes skin secretions as varying significantly between patients. Skin type amongst the population may differ due to bacterial flora and gland type, with ‘diet, emotional state, menstrual cycle, age, personal care and many other factors’ playing a role in the odour and VOCs present. As a form of VOC detection, it is the most difficult to use as a diagnostic tool due to the parameters of skin variability.

In spite of their drawbacks skin derived VOCs could play a key role for prevention of certain diseases caused by mosquitos. For instance, the yellow fever mosquito (*Aedes aegypti)* is responsible for spreading multiple diseases such as yellow fever, dengue fever, and zika fever (Center for Disease Control and Prevention, 2016; WHO, 2016). Previous research on VOCs emitted from human skin has reported that the compounds act as attractants for *Aedes aegypti* (Bernier et al.,

2000). For example, a possible attractant can be 2,5-Dimethylpiperazine, which is believed to be of human origin (Bernier et al., 2000). Although the results of the study did not determine a particular attractant, it could indicate in general that some individuals may be more vulnerable to mosquitoes than others.

VOCs IN MILK In this research, from the 100 analyzed VOCs only 12 were found in human milk (*figure 2*). The majority being hydrocarbons and ketones but other chemical classes were found such as halogenated, nitrogen containing, aldehydes, alcohols and acids. Most of these compounds are a result of pollution but the rest of them originate from food, drugs and other unknown sources. Indeed, most of the scientific studies look for the specific compounds transferred from mother to child by looking to mother's exposure to medication, environmental contamination and dietary supplementation (De Lacy Costello et al., 2014).

The use of VOCs in physiological fluids as biomarkers of exposure is known to be very challenging in sample collection and analysis (Blount et al., 2010). There is very little information on VOCs presence in human milk (De Lacy Costello et al., 2014). The literature review has shown that quite a lot of research in the field of VOCs neglects those found in milk. However, breast milk being a potential source of infant exposure to VOCs, measuring VOCs in this specific physiological fluid is a gap in scientific understanding of possible health risks and exposures (Blount et al., 2010).

VOCs IN BLOOD Out of the 100 compounds that were analysed, 15 can be detected in the blood. From these 15, the most prevalent class of compounds was hydrocarbons while the rest included esters, aldehydes and nitrogen containing compounds. Using blood to screen for VOCs has been particularly fruitful for identifying many toxic human carcinogens, one such carcinogen is benzene. 4 compounds that belong to the benzenoid class were found: 2,3,4,6-tetrachlorophenol,1,2-dichlorobenzene,1,3-dichlorobenzene and 1,4-dichlorobenzene. Research has found that long term exposure to benzene has harmful effects on the bone marrow, subsequently decreasing the number of blood cells in the body. Furthermore, it has been associated with the developments of one or more types of leukemia (Snyder, 2012).

VOCs detected in the bloodstream have been found to be an effective biomarker for disease, due to the fact that it is a medium that flows through the whole body (Nieuwenhuijsen, 2015). VOCs are diagnostically useful because they are released into the bloodstream before they are exhaled, making them less likely to be affected by external factors like smoking or diet (Wang et al., 2014). However, blood sampling is an invasive procedure that requires specially trained staff.

VOCs IN SALVIA The largest proportion of VOCs that were identified in the saliva came from food and essential oils, with the largest percentage of compounds being hydrocarbons followed by ketones and nitrogen containing compounds. Most foods that contain hydrocarbons are fats and oils, it is formed in plants using sugars and in animals fats are produced with a combination of sugars and proteins. Hence a diet high in meat, fruits, nuts and dairy may contain a substantial amount of VOCs. In addition to this VOCs that come from pollution and smoking are also present in the saliva.

Saliva contains a range of constituents such as proteins, peptides and a wide variety of VOCs (Soini et al., 2010). These constituents play a vital role in chemical communication as well as for our health due to the presence of antimicrobial proteins (Soini et al., 2010). VOCs found in saliva likely originate from substances that we consumed, leaving traces

in the saliva when the food compounds is broken down.

P-menth-3-ene (CAS: 500-00-5), which is originally found in the essential oils of the aerial parts of plants such as *Artemisia suksdorfii* and *Trachyspermum ammi* Linn. (*T. ammi*), a grassy plant originally from Iran, India and Egypt (Mahmoud et al., 2006). Both plants have medicinal value, assisting in the treatment of arthritis, gastrointestinal problems, headaches and flu like symptoms. In addition to this they are used around the world as an essential cooking spice (Asif, Sultana & Akhtar, 2014). Thus, p-menth-3-ene could be derived from foods or medicinal sources.

WIKIPATHWAY ADDITIONS An attempt was made to add any pathways found for any of the VOCs to WikiPathways in the hope of aiding future research on human metabolites. However, only two pathways corresponding to the present dataset could be added to WikiPathways: 1,1-dichloroethene and o-tolualdehyde. It is unlikely that most of the large number of VOCs in this dataset are not involved in metabolism due to the vast number of human metabolic pathways (Romero et al., 2005).

After ranking the VOCs by the number of biosystems listed on PubChem it became evident that there were a limited number of pathways available for all VOCs; only 8 VOCs with any biosystems were found. Furthemore, extremely high numbers of biosystems (12760 and 4124 for o-tolualdehyde and 3-methylbutanoic acid, respectively) were present in some of these 8 metabolites due to duplicate entries for orthologous pathways, thus leading to misrepresentative ranking. Both of these issues could be rectified with changes to the ranking programme.

In order to increase the amount of pathways found for the dataset other databases could be utilised to gather pathway counts. Reactome is an open source, peer-reviewed and manually-curated database of human pathways and processes which can be queried programmatically (Matthews et al., 2009). Furthemore, the KEGG pathway database is a collection of pathway maps containing, but not limited to, reaction networks for metabolism, cellular processing and human diseases (Kanehisa & Goto, 2000). KEGG may also be queried with a REST API. As well as the aforementioned databases there are numerous pathway databases that could be queried for pathway counts relating to a specific metabolite.

So to give a better representation of the number of pathways available for each metabolite duplicate entries should be removed. NCBI has multiple entries for biosystems of the same name and function since a new entry is created for each species that the pathway is present in (Wang et al., 2013b). Therefore, the number of biosystems for a particular compound can be greatly inflated. In order to remove duplicate entries, biosystems should be grouped by name. However, this may cause further issues since the names of two different pathways could be identical, although this is unlikely.

WikiPathways has a limited species list meaning new entries for many of the identified pathways could not be created (Kutmon et al., 2016). At the time of writing, the creation of a species independent reference pathway is not possible. To this end, a species independent “bacteria reactor” will soon be added to the species list (Ehrhart & Willighagen, 2016). Therefore, WikiPathway entries for metabolites with WikiPathway IDs stated as “pending addition” in *Appendix A* will be created in the near future.

The unidentified tolualdehyde VOC (CAS: 1134-78-7) was estimated to be o-tolualdehyde and the CAS registry number given as such (De Lacy Costello et al., 2014). At the time of writing, o-tolualdehyde has 12760 biosystems associated with it, of which most are duplicate entries for orthologous biosystems. O-tolualdehyde was identified to be a metabolite in the o-xylene degradation pathway in many prokaryotes. Therefore, a

WikiPathway entry was created for *Escherichia coli*. O-tolualdehyde is a primary pollutant derived from petroleum and is regularly detected in ambient air (Obermeyer, Aschmann, Atkinson, & Arey, 2009). Therefore, since the compound was found in breath it is likely that the unidentified tolualdehyde is o-tolualdehyde. Little information was found on any biological effects of o-tolualdehyde. However, o-xylene is reported to be a human health hazard which can cause fatigue confusion, headaches, and death (Robledo-Ortíz et al., 2011). Therefore, degradation of o-xylene by biological processes has been suggested as a sustainable and effective method to remediate contaminated sites (Thayer, 1991). The addition of this degradation pathway to WikiPathways may aid in future research on the removal of this primary pollutant.

A WikiPathway entry was also created for the unidentified dichloroethene compound which was estimated to be 1,1-dichloroethene (De Lacy Costello et al., 2014). PubChem lists 112 biosystems associated to this compound, all of which are for the metabolism of xenobiotics by cytochrome P450 but in different organisms. A KEGG pathway for the metabolism of 1,1-dichloroethene in *Homo sapiens* was identified. Since the compound was identified in a blood sample it is likely that it is metabolised by humans, thus, the estimation of 1,1-dichloroethene is most likely correct. 1,1-dichloroethene is a pollutant and has been found to have a toxic effect on developed lung and cardiac development in animal models (Hardin, Kelman & Brent, 2005; Simmonds et al., 2004). To this end, quantification of the blood concentrations of 1,1-dichloroethene and its metabolites that could indicate a toxic dose would aid in the prevention of any potential adverse effects caused by exposure to the pollutant.

TOWARDS OPEN SCIENCE Open science refers to making the output of publicly funded research widely accessible in digital formats to the scientific community and the general public (OECD, 2015). Open and online tools greatly increase the efficiency of research by drastically lowering the cost and time taken to collect data and extend previous research. The addition of the pathways and disease correlations presented here to open databases, such as WikiPathways and Wikidata, will inevitably contribute towards further research in metabolomics.

Biological pathway analyses and visualizations provide insights and improved understanding of the underlying biological mechanisms in genomics, proteomics and metabolomics (Jennen et al., 2010). WikiPathways provides annotations of all elements in a pathway with links to external databases such as HMDB and PubChem (Pico et al., 2008). Researchers interested in a particular pathway can use WikiPathways for up-to-date information in order to progress their own research.

The semantic web provides another facility to efficiently advance scientific research (Archer, 2016; Waagmeester et al., 2016). Through the linking of related data, from multiple databases, networks of relationships between genes, pathways, bioassays and diseases can be built and efficiently analysed. Machine readability simplifies the task of linking such data and is something that both Wikidata and WikiPathways provides. Open PHACTS is one effort to use semantic web standards to facilitate drug discovery research and by doing so will support open innovation by supporting data sharing and reuse (Williams et al., 2012). The value of open data is clear and as such an effort was made to make all findings present here publically available and machine readable.

CONCLUSION

The research here has gone some way into providing useful information on the origin, pathways and disease relations of many VOCs. However, only a small proportion of

the dataset given by de De Lacy Costello et al. (2014) could be investigated due to a lack of research into the biology and origin of many of the VOCs. An effort was made to aid future research on the human volatolome by adding all found data to open platforms such as WikiPathways and Wikidata.

Currently, there is a large amount of human volatolome data but little information on each VOCs biological importance. Thus, the speed of data interpretation and analysis is the current bottleneck in developing tools for early disease diagnostics through the use of VOCs. Therefore, the results provided here, which are publically available, will improve the rate of discovery of correlations between disease and VOCs.

AUTHOR CONTRIBUTIONS

Defne:	Introduction, VOCs in skin
Chidera:	databases method, VOCs in saliva
Lizzy:	results outline, VOCs in skin and breath
Eliis:	Introduction, editing
Kunsulu:	VOCs in faeces, urine and breath
Jannika:	wikidata methods and results, editing
Steve:	Introduction, VOCs in blood
Olivier:	Conclusion
Cristina:	VOCs in milk
Jacob:	Ranker methods, wikipathways methods, resluts and discussion, towards open science

REFERENCES

Amann, A. et al. (2014). The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *Journal of Breath Research, 8, 034001.* DOI : 10.1088/1752-7155/8/3/034001.

Arasaradnam, R., McFarlane, M., Ryan-Fisher, C., Westenbrink, E., Hodges, P., & Thomas, M. et al. (2014). Detection of Colorectal Cancer (CRC) by Urinary Volatile Organic Compound Analysis. *Plos ONE*, *9*(9), e108750. http://dx.doi.org/10.1371/journal.pone.0108750.

Arbor A, 2016, Uberon Multi Species Anatomy Ontology, Ontobee, University of Michigan Medical School Retrieved from: http://www.ontobee.org/ontology/UBERON?iri=http://purl.obolibrary.org/obo/UBERON_0001836

Archer, P. (2016). W3C data activity - building the web of data. Retrieved from https://www.w3.org/2013/data/

Asif, H.M., Sultana, S., & Akhtar, N. (2014). A panoramic view on phytochemical, nutritional, ethanobotanical uses and pharmacological values of Trachyspermum ammi Linn. *Asian Pacific Journal of Tropical Biomedicine, 4, Supplement 2,* S545-S553. doi: http://dx.doi.org/10.12980/APJTB.4.2014APJTB-2014-0242

Bernier, U. R., Kline, D. L., Barnard, D. R., Schreck, C. E., & Yost, R. A. (2000). Analysis of Human Skin Emanations by Gas Chromatography/Mass Spectrometry. 2. Identification of Volatile Compounds That Are Candidate Attractants for the Yellow Fever Mosquito (Aedes aegypti). Analytical Chemistry, 72(4), 747-756. doi: 10.1021/ac990963k.

Blount et al.. (2010). Methodology for collecting, storing, and analyzing human milk for volatile organic compounds. *Journal of Environmental Monitoring*, 12(1), 1265–1273.

Cafasso, J. (2015). Urine Tests for Diabetes: Glucose Levels and Ketones (P. Pletcher, Ed.). Retrieved June 25, 2016, from http://www.healthline.com/health/type-2-diabetes/urine-tests#Overview1.

Carrola J et al. (2011). Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. *J. Proteome Res.,10,* 221–230.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., . . . Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research, 44(D1), D471-D480. doi: 10.1093/nar/gkv1164.

Centers for Disease Control and Prevention. (2016). *Zika virus*. Retrieved 27 June 2016, from http://www.cdc.gov/zika/transmission/

De Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., … Ratcliffe, N. M. (2014). A review of the volatiles from the healthy human body. *Journal of Breath Research*, *8*(1), 014001. doi:10.1088/1752-7155/8/1/014001.

Deng, C., Song, G., Zheng, X., Hu, Y., & Zhang, X. (2003). Analysis of the volatile constituents of Apium graveolens L. and Oenanthe L. by gas chromatography-mass spectrometry, using headspace solid-phase microextraction. Chromatographia, 57(11), 805-809. doi: 10.1007/bf02491769.

Ehrhart, F., & Willighagen, E. (2016). In Windsor J. (Ed.), *WikiPathways: Species independent bacteria reactor*.

Filipiak W. et al. (2012). Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. *Journal of Breath Research, 6,* 036008.

Garner et al. (2007). Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. *The FASEB Journal, 21(8),* 1675-1688. DOI : 10.1096/fj.06-6927com.

Geldreich, E. E., Kenner, B. A., & Kabler, P. W. (1964). Occurrence of coliforms, fecal coliforms, and streptococci on vegetation and insects.*Applied microbiology*, *12*(1), 63-69.

Gottzein, A. K., Musshoff, F., & Madea, B. (2010). Qualitative screening for volatile organic compounds in human blood using solid-phase microextraction and gas chromatography-mass spectrometry. *Journal of mass spectrometry*, *45*(4), 391-397.

Hakim, M., Broza, Y., Barash, O., Peled, N., Phillips, M., Amann, A., & Haick, H. (2012). Volatile Organic Compounds of Lung Cancer and Possible Biochemical Pathways. *Chemical Reviews*, *112*(11), 5949-5966. http://dx.doi.org/10.1021/cr300174a.

Hardin, B. D., Kelman, B. J., & Brent, R. L. (2005). Trichloroethylene and dichloroethylene: A critical review of teratogenicity. *Birth Defects Research Part A: Clinical and Molecular Teratology, 73*(12), 931-955.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., . . . Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research, 41(D1), D456-D463. doi: 10.1093/nar/gks1146.

Henriet, S. S., Verweij, P. E., & Warris, A. (2012, July 24). Aspergillus nidulans and Chronic Granulomatous Disease: A Unique Host-Pathogen Interaction. *Journal of Infectious Diseases, 206*(7), 1128-1137. doi:10.1093/infdis/jis473

Hipp, R. (2016). SQLite. https://www.sqlite.org/

Human Metabolome Database. (n.d.). Showing metabocard for Dihydro-5-propyl-2(3H)-furanone (HMDB31681). Retrieved June 22, 2016, from http://www.hmdb.ca/metabolites/HMDB31681

Huang J. et al. (2013). Selected ion flow tube mass spectrometry analysis of volatile metabolites in urine headspace for the profiling of gastro-esophageal cancer. *Anal. Chem. 85,* 3409-3416.

International Programme on Chemical Safety. (2001). Biomarkers In Risk Assessment: Validity And Validation. Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

Jennen, D. G. J., Gaj, S., Giesbertz, P. J., van Delft, Joost H M, Evelo, C. T., & Kleinjans, J. C. S. (2010). Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug Discovery Today, 15*(19-20), 851-858. doi:10.1016/j.drudis.2010.08.002.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27-30.

Karp, P., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., . . . Paulsen, I. (2014). The EcoCyc Database. EcoSal Plus. doi: doi:10.1128/ecosalplus.ESP-0009-2013.

Kaskoniene, V., Venskutonis, P. R., & Ceksteryte, V. (2008). Composition of volatile compounds of honey of various floral origin and beebread collected in Lithuania. Food Chemistry, 111(4), 988-997. doi: doi:10.1016/j.foodchem.2008.05.021.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2011). WikiPathways: building research communities on biological pathways. Nucleic Acids Research. doi: 10.1093/nar/gkr1074.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research, 44*(D1), 1202.

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., . . . Pico, A. R. (2016). WikiPathways: Capturing the full diversity of pathway knowledge. *Nucleic Acids Research, 44*(D1), 488.

Maga, J. A. (1978). Cereal volatiles, a review. Journal of Agricultural and Food Chemistry, 26(1), 175-178. doi: 10.1021/jf60215a055

Mahmoud A, Ahmed A, 2006, *Alpha-pinene-type monoterpenes and other constituents from Artemisa suksdorfii,* Phytochemistry, Pubfacts, Scientific Publication Data

Manske, M. (2016). Source MetaData. from http://tools.wmflabs.org/sourcemd/

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., . . . D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research, 37*(Database issue), 619.

Miller, M. B., & Bassler, B. L. (2001, October). Quorum sensing in bacteria. [Abstract]. *Annu Rev Microbiol., 55*, 165-199. doi:10.1146/annurev.micro.55.1.165

National Institute of Health. (2016, June 21). Chronic granulomatous disease. Retrieved June 27, 2016, from https://ghr.nlm.nih.gov/condition/chronic-granulomatous-disease

Nieuwenhuijsen, M. J. (2015). *Exposure assessment in environmental epidemiology*. Oxford University Press, USA.

Obermeyer, G., Aschmann, S. M., Atkinson, R., & Arey, J. (2009). Carbonyl atmospheric reaction products of aromatic hydrocarbons in ambient air. *Atmospheric Environment, 43*(24), 3736-3744. doi:10.1016/j.atmosenv.2009.04.015

OECD. (2015). *Making open science a reality* Organisation for Economic Co-operation and Development. doi:10.1787/5jrs2f963zs1-en.

Pence, H. E., & Williams, A. (2010). ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education, 87(11), 1123-1124. doi: 10.1021/ed100697w.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., & Evelo, C. (2008). WikiPathways: Pathway editing for the people. *PLoS Biology, 6*(7), e184.

Python Anywhere. (2016). from https://www.pythonanywhere.com

Python Software Foundation. (2016). from https://www.python.org/

Robledo-Ortíz, J. R., Ramírez-Arreola, D. E., Pérez-Fonseca, A. A., Gómez, C., González-Reynoso, O., Ramos-Quirarte, J., & González-Núñez, R. (2011). Benzene, toluene, and o-xylene degradation by free and immobilized P. putida F1 of postconsumer agave-fiber/polymer foamed composites. *International Biodeterioration & Biodegradation, 65*(3), 539-546. doi:10.1016/j.ibiod.2010.12.011

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., & Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology, 6*(1), R2. doi:10.1186/gb-2004-6-1-r2.

Ronacher, A. (2016). Flask. from http://flask.pocoo.org/

Saalberg, Y., & Wolff, M. (2016). VOC breath biomarkers in lung cancer. *Clinica Chimica Acta,459*, 5-9. doi:10.1016/j.cca.2016.05.013.

Sethi, S., Nanda, R., & Chakraborty, T. (2013). Clinical Application of Volatile Organic Compound Analysis for Detecting Infectious Diseases. *Clinical Microbiology Reviews*, *26*(3), 462-475. http://dx.doi.org/10.1128/cmr.00020-13.

Schulz, S., & Dickschat, J. S. (2007). Bacterial volatiles: the smell of small organisms. *Natural product reports*, *24*(4), 814-842.

Silva C.L., Passos M. and Camara J.S. (2011). Investigation of urinary volatile organic metabolites as potential cancer biomarkers by solid-phase microextraction in combination with gas chromatography-mass spectrometry. *Br. J. Cancer,105,* 1894–1904.

Silva C.L., Passos M. and Camara J.S. (2012). Solid phase microextraction, mass spectrometry and metabolomic approaches for detection of potential urinary cancer biomarkers - a powerful strategy for breast cancer diagnosis. *Talanta, 89*, 360-368.

Simmonds, A. C., Reilly, C. A., Baldwin, R. M., Ghanayem, B. I., Lanza, D. L., Yost, G. S., . . . Forkert, P. (2004). Bioactivation of 1,1-dichloromethylene to its epoxide by CYP2E1 and CYP2F enzymes. *Drug Metabolism and Disposition, 32*(9), 1032-1039. Retrieved from http://dmd.aspetjournals.org/content/32/9/1032

Snyder, R. (2012). Leukemia and benzene. *International journal of environmental research and public health*, *9*(8), 2875-2893.

Sohrabi, M., Zhang, K., Ahmetagic, A., Wei, M. Q., & Zhang, L. (2014). Volatile organic compounds as novel markers for the detection of bacterial infections. *Clinical Microbiology: Open Access*, *2014*.

Soini HA, Klouckova I, Wiesler D, Oberzaucher E, Grammer K, Dixon SJ, Xu Y, Brereton RG, Penn DJ, Novotny MV, 2010, *Analysis of Volatile Organic Compounds in Human Saliva by Static Sorptive Extraction, Method and Gas Chromatography-Mass Spectrometry*

Somerville A, 1998, *Scifinder Scholar (by Chemical Abstracts Service)*, Head Science and Engineering Libraries and Chemistry Librarian, Carlson Library, Journal of Chemical Education

Spanel P. et al. (1999). Analysis of formaldehyde in the headspace of urine from bladder and prostate cancer patients using selected ion flow tube mass spectrometry. *Rapid Commun. Mass Spectrom., 13,* 1354–1359.

Thayer, A. (1991). Bioremediation: Innovative technology for cleaning up hazardous waste. *Chemical & Engineering News Archive, 69*(34), 23-44. doi:10.1021/cen-v069n034.p023

Umano, K., Hagi, Y., Shoji, A., & Shibamoto, T. (1990). Volatile compounds formed from cooked whole egg, egg yolk, and egg white. Journal of Agricultural and Food Chemistry, 38(2), 461-464. doi: 10.1021/jf00092a028.

Verheugen, G. (2006). COMMISSION DECISION of 9 February 2006 amending Decision 96/335/EC establishing an inventory and a common nomenclature of ingredients employed in cosmetic products (Text with EEA relevance) (2006/257/EC). Official Journal of the European Union(L97/1).

Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E. L., Evelo, C. T., & Pico, A. R. (2016). Using the semantic web for rapid integration of WikiPathways with other biological online data resources. *PLOS Comput Biol, 12*(6), e1004989. doi:10.1371/journal.pcbi.1004989.

Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., . . . Bryant, S. H. (2013a). PubChem BioAssay: 2014 update. *Nucleic Acids Research, 42*(D1), D1082.

Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., . . . Bryant, S. H. (2013b). PubChem BioAssay: 2014 update. *Nucleic Acids Research, 42*(D1), D1082.

Wang, C., Li, P., Lian, A., Sun, B., Wang, X., Guo, L., ... & Guo, Z. (2014). Blood volatile compounds as biomarkers for colorectal cancer. *Cancer biology & therapy*, *15*(2), 200-206.

Weber, C.M. et al. (2011). Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace. *Analyst, 136,* 359–64.

Wikimedia Foundation. (2016). Wikidata. from https://www.wikidata.org/

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., . . . Mons, B. (2012). Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today, 17*(21–22), 1188-1198. doi:10.1016/j.drudis.2012.05.016.

Williams, H. E., Steele, J. C., Clements, M. O., & Keshavarz, T. (2012, November 28). γ-Heptalactone is an endogenously produced quorum-sensing molecule regulating growth and secondary metabolite production by Aspergillus nidulans. *Appl Microbiol Biotechnol Applied Microbiology and Biotechnology, 96*(3), 773-781. doi:10.1007/s00253-012-4065-5

Windsor, J. (2016). *PubChem ranker* Figshare. doi:10.6084/m9.figshare.3465008.v1

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Scalbert, A. (2013). HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Research, 41(Database issue), D801-D807. doi: 10.1093/nar/gks1065.

World Health Organization. (2016). *The mosquito*. Retrieved 27 June 2016, from http://www.who.int/denguecontrol/mosquito/en/

World Health Organization. (2016). *Yellow fever*. Retrieved 27 June 2016, from http://www.who.int/mediacentre/factsheets/fs100/en/

Yang, F., Wang, L., Wang, J., Dong, Y., Hu, J. Y., & Zhang, L. (2005, July 13). Quorum quenching enzyme activity is widely conserved in the sera of mammalian species. *FEBS Letters, 579*(17), 3713-3717. doi:10.1016/j.febslet.2005.05.060

APPENDIX A

Table 1. List of the 100 VOCs from the unmapped dataset showing the compounds origin, chemical class and where it was measured.

CAS	name	wikidata ID	Origin	chemial class	Faeces	Urine	Breath	Skin	Milk	Blood	Saliva
583-57-3	1,2-Dimethylcyclohexane	Q24702624	essential oil, org pollutant	alkane			Br				
586-63-0	Isoterpinolene	Q24514387	essential oil, turpentine	alkene							Sa
590-66-9	1,1-Dimethylcyclohexane	Q24702742	essential oil, gasoline, pollution (air)	alkane			Br				
58-90-2	2,3,4,6-tetrachlorophenol	Q24705698	pollutant, wood preservative	benzenoid, phenol						Bl	
25321-22-6	mix of 1,2-; 1,3-; 1,4-dichlorobenzene	Q1209403		benzenoid					M		
95-50-1	1,2-dichlorbenzene	Q2609815	insecticide, pollutant (declared deadly by EPA), pharmaceuticals, disinfectants, rubbers, plastics and electric goods	benzenoid			Br			Bl	
541-73-1	1,3-dichlorobenzene	Q2216854	insecticide, medicine, dyes, pollutant	benzenoid			Br			Bl	
106-46-7	1,4-dichlorobenzene	Q161529	pollutant, pesticide, deodorant	benzenoid		U	Br		M	Bl	
25323-30-2	Dichloroethene ui	Q1209399		alkene					M		
75-35-4	1,1-dichloroethene	Q161284	food packaging, pollutant	alkene						Bl	
540-59-	1,2-dichloroethene	Q161475	pollutant	alkene						Bl	

0											
1576-87-0 and 764-39-6	(E)-2-pentenal	Q24705869	flavor (food stuff), insecticide (pollutant)	aldehyde					M		
90-17-5	alpha-(trichloromethyl)benzyl acetate; roseacetol	Q22052707	fragrance (pollutant?)	ketone							Sa
84-67-3	4-(4-amino-2-methylphenyl)-3-methylaniline	Q24705889		benzenoid							
95-96-5	3,6-dimethyl-1,4-dioxane-2,5-dione; lactide	Q421313		ester			Br				
592-48-3	1,3-hexadiene	Q24710331	air pollution, biomass burning and traffic emmisions	alkene							
591-24-2	3-methyl-cyclohexanone	Q22138333	essential oil, flavouring agent	ketone							
598-25-4	1,1-dimethylallene	Q24711948	found in cigarette smoke	alkene			Br				
598-94-7	1,1-dimethylurea	Q24712449	a possible metabolite of a pesticide that inhibits photosynthesis	diamide		U					
469-92-1	beta-clovene	Q24713391	essential oil, occurs in plants	sesquiterpenoid							Sa
499-99-0	isolimonene	Q22079607	occurs in citruses, cosmetic contaminant	terpene							Sa
541-28-6	3-methyl-1-iodobutane	Q24716059	related to emissions, air/water pollutant	haloalkane					M		
500-00-	p-menth-3-ene	Q24716505	found in herbs and spices:	cycloalkene			Br				Sa

5			isolated from thyme and peppermint oil;pollutant; related to smoking								
0544-10-5	1-chloro-hexane	Q24716558		haloalkane				Sk			
473-91-6	1,2,3-trimethyl-1-cyclopentene	Q24734266		cycloalkene							
624-80-6	1-ethylhydrazine	Q24734328	Urethral tract; antiasthmatics, kidney, dermatological disorders	alkyl hydrazines			Br				
57084-17-0	γ-Undecalactone	Q24734376	Flavouring agent, tabacco,	aldehyde		U		Sk			Sa
75-05-8	acetonitrile	Q21761558	found in: coal tar, volcanoe's smoke, perfumes, some medicines: antibiotics, insulin and vitamins, chemical solvent.	Nitriles	F		Br			Bl	Sa
96-86-2	acetophenone	Q375112	found in: drug, food, water and ambient air	Ketones	F	U	Br	Sk	M		Sa
111-15-9	2-Ethoxyethyl acetate	Q209376	industrial solvent	ether			Br				
637-50-3	1-phenyl-1-propene	Q2020228		styrene			Br				
638-28-8	2-chlorohexane	Q24734466		haloalkane			Br				
100-45-8	4-cyanocyclohexene	Q24730219	food packaging (adhesive), pollutant	cycloalkene				Sk			
105-21-5	gamma-heptalactone	Q24730325	possible metabolite used in fungal quorum sensing/xenobiotic, food,	lactone				Sk			Sa

			flavouring agent								
105-42-0	4-methyl-2-hexanone	Q24730663	bacterial synthesis from diet	ketone			Br				
78-80-8	isopropenylacetylene	Q22077076	cigarette adhesive, coating component for food & beverage containers	alkyne							
108-96-3	4-hydroxypyridine	Q24730939	Plant metabolite, food	pyridine							
108-00-9	N,N-dimethyl-1,2-ethanediamine	Q24731206	unknown	amine				Sk			
2390-94-5	trans-tetrahydro-2,5-dimethylfuran	Q24734527	compound found in tobacco, related to healthy gut, xenobiotic	furan	F						
868-57-5	2-methylbutanoic acid methyl ester	Q24734848	Found in fruits. Flavor component of apple & strawberry	fatty acid ester	F						
1731-92-6	heptadecanoic acid methyl ester	Q24735110	Intermediate in waste cooking oil transesterification - pollutant	ester				Sk			
2080-89-9	3-ethyl-1,3-hexadiene	Q24735189	Food product - canned salmon, green beans, fish oil	alkene				Sk			
645-62-5	2-ethyl-2-hexenal	Q24735451	Used as an insecticide, intermediate for chemical synthesis, and warning odorant	aldehydes						Bl	Sa
692-24-0	2-methyl-trans-3-hexene	Q24735526		alkene			Br				
111-06-08	butyl palmitate; hexadecanoic acid butyl ester; N-Butyl palmitate	Q24735748	solvents/deluents for cosmetic, flavoring substance for food	ester				Sk			

111-67-1	2-octene (both trans and cis)	Q24735790	flavouring substance for food	acyclic olefins	F		Br	Sk		Bl	
13389-42-9	trans-2-octene; (2E)-2-Octene	Q24735900	flavouring substance for food	acyclic olefins			Br				Sa
7642-04-08	cis-2-octene; (2Z)-2-Octene	Q24736444	flavouring substance for food	acyclic olefins			Br				Sa
112-18-5	N,N-Dimethyldodecylamine; N,N-Dimethyl-1-dodecanamine	Q24736495	antistatic/hair conditioning	amine				Sk			
112-52-7	1-chlorodecane, Lauryl chloride	Q24736563	pollutant from plastic recycling	chlorinated hydrocarbon				Sk			
109-01-03	1-methylpiperazine	Q24736602		amine				Sk			
112-69-6	N,N-Dimethyl-1-hexadecanamine;	Q24736646	air pollution	tetiary amine				Sk			
222-51-5	dibenzopentaphene ; Dibenzo[c,m]pentaphene	Q24736718		alkene						Bl	
503-28-6	Dimethyldiazene; azomethane	Q22053947	found in essential oil	amine			Br				
1333-41-1	Methylpyridine; picoline	Q2092388	pollutant	methylpyridines		U					
109-06-08	2-methylpyridine	Q2216745	pollutant (oil shale, coal)	methylpyridines				Sk			Sa
693-89-0	1-methyl-1-cyclopentene	Q24736866		Alkenes			Br				
693-97-0	5-methylisothiazole	Q24736902		thiazole, N and S containing		U					
758-16-	dimethylthioformamide	Q24736934		amide			Br				

7											
57295-30-4	3-hydroxy-4-methylbenzaldehyde	Q24755156	used in fuel processing, disinfecctants, surfactants, wood preservatives, and intermediates in the manufacture of pesticides and resins	benzenoid				Sk			
1334-78-7	tolualdehyde ui	Q2439424	polluting agent, derived from petroleum, reguarly detected in air	benzenoid			Br				
35915-22-1	Methylbutanoic acid ui (3-methylbutanoic acid)	Q415536		acid					M		
39276-09-0	Furaldehyde ui (2-furaldehyde)	Q24755507		benzenoid					M		
471-84-1	Aplha-fenchene	Q24715140	endogenous, food, plants, perfumes, shampoos, detergents.	monoterpene							Sa
26952-23-8	Dichloropropene	Q24755666	Agricuultural soil fumigant	amine					M		
695-34-1	4-methylpyridinamine (2-amino-4-methylpyridine)	Q24755764		benzenoid				Sk			
5131-66-8	1-butoxy-2-propanol	Q24762634	Solvent	alcohol			Br				
74381-40-1	isobutyric acid	Q415062	Solvent, odour	acid			Br				
2442-49-1	Methyl tetracosanoate	Q24762827	Wood tar, bioactive molecule of costus pictus	ester				Sk			
624-42-0	6-methyl-3-heptanone	Q24762941	Flavoring agents	ketone		U	Br				
625-27-	2-methyl-2-pentene	Q24763010	ampule opener'	alkene			Br				

4											
763-69-9	Ethyl 3-ethoxypropionate	Q24763201	Solvent for cleaning and degreasing	ester			Br				
625-28-5	2-methylbutane secondary mononitrile	Q24763442		nitrile			Br				
625-54-7	2-ethoxypropane	Q24764257		alkane			Br				
1731-88-0	methyl tridecanoate	Q24764359	Flavouring agent	ester							
4403-61-6	2-cyano-2-butene	Q24764474		alkene		U					
1077-16-3	Hexylbenzene	Q24764675		alkene							Sa
27137-41-3	methylfuran	-		furan							
628-61-5	1-methylheptylchloride / 2-chlorooctane	Q24765603		haloakane			Br				
1551-06-0	2-ethylpyrrole	Q24765696		amine			Br				
2306-89-0	Decyl octanoate	Q24765784	Antimicrobial pesticide	ester					M		
63072-44-6	Methyl pentanone ui	Q223076		ketone					M		
0135-01-03	1,2-diethylbenzene	Q2429316		alkene			Br				
502-26-1	4-Octadecanolide	Q24816483		ester						Bl	
141-93-5	1,3-diethylbenzene	Q24816743		alkene			Br				

109-49-9	1-hexen-5-one	Q24816828	Air pollutant (mold growth detection), compound found in tobacco leaf	ketone			Br				
109-68-2	2-pentene	Q22051121	compound found in tobacco leaf	alkene			Br				
96-38-8	5-Methyl-1,3-cyclopentadiene	Q22077211	found in milk after ultrasound treatment (sanitization)	diene							
96-39-9	1-Methyl-1,3-cyclopentadiene	Q5094303	Pollutant, found in petrochemical plants	diene			Br				
106-55-8	2,5-Dimethylpiperazine	Q22079573	Attractant for the Yellow Fever Mosquito (Aedes aegypti)	heterocyclics				Sk			
110-98-5	1,1'-Oxydi-2-propanol	Q22829651	Found in marker pens and textile colorant	diol							
112-75-4	N,N-Dimethyltetradecylamine	Q24817433	Pollutant (drinking water)	amine				Sk			
124-28-7	N,N-Dimethyloctadecylamine	Q24817440	Contaminant (sea)	amine				Sk			
144-19-4	2,2,4-Trimethyl-1,3-pentanediol	Q24817446	Pollutant (drinking water)	diol					M		
460-01-05	2,6-Dimethyl-1,3,5,7-octatetraene	Q24817453	Emitted from creosotebush (Larrea tridentata), volatile oil in the rhizomes and radixes of Notopterygium incisum.	alkene							Sa
513-23-5	isothujol	Q24817455	Volatile aroma (citrus fruits), essential oil in flower heads of Chrysanthemumindicum L. from China	alcohol							Sa
514-14-7	2,7,7-trimethylbicyclo[2.2.1]hept-2-ene	Q24817463	Volatile oil in the rhizomes and radixes of Notopterygium incisum.	alkene							Sa

514-95-4	1,5,5-Trimethyl-6-methylene-cyclohexene	Q24819025	Found in celery and water dropwort and is an essential oil from Wedelia trilobata (L.) Hitchc	alkene	F						
527-53-7	1,2,3,5-Tetramethylbenzene	Q24819115	Found in rice products	alkene			Br				
542-54-1	4-Methylpentanenitrile	Q24819169	Found in honey, cooked egg and other foods	nitrile			Br				

Table 2. List of the 100 VOCs from the unmapped dataset showing different identifiers including pathway ID's, pathway description and disease information.

CAS	name	CID	ChEBI	InChI-Key	KEGG Pathway ID	BioCyc pathway ID	Wiki Pathways ID	HMDB ID	Pathway(s) description	Disease	Wiki data ID
583-57-3	1,2-Dimethylcyclohexane	11416		KVZJLSYJROEPSQ-UHFFFAOYSA-N							Q24702624
586-63-0	Isoterpinolene	102443	88840	CIPXOBMYVWRNLL-UHFFFAOYSA-N				HMDB61938			Q24514387
590-66-9	1,1-Dimethylcyclohexane	11549		QEGNUYASOUJEHD-UHFFFAOYSA-N							Q24702742
58-90-2	2,3,4,6-tetrachlorophenol	6028		VGVRPFIJEJYOFN-UHFFFAOYSA-N							Q24705698
25321-22-6	mix of 1,2-; 1,3-; 1,4-dichlorobenzene		23697								Q1209403

95-50-1	1,2-dichlorbenzene	7239	35290	RFFLAFLAYFXFSW-UHFFFAOYSA-N		PWY-6090	Pending addition		Some species of pseudomonas can utilize it as their sole carbon source		Q2609815
541-73-1	1,3-dichlorobenzene	10943	36693	ZPQOPVIELGIULI-UHFFFAOYSA-N				HMDB59855			Q2216854
106-46-7	1,4-dichlorobenzene	4685	28618	OCJBOOLMMGQPQU-UHFFFAOYSA-N				HMDB41971			Q161529
25323-30-2	Dichloroethene ui										Q1209399
75-35-4	1,1-dichloroethene	6366	34031	LGXVIGDEPROXKC-UHFFFAOYSA-N	ko00980		WP3666		hsa00980 "metabolism of xenobiotics by CYP450" homo sapiens		Q161284
540-59-0	1,2-dichloroethene	10900		KFUSEUYYWQURPO-UHFFFAOYSA-N							Q161475
1576-87-0 and 764-39-6	(E)-2-pentenal	5364752		DTCCTIQRPGSLPT-ONEGZZNKSA-N		PWY-6786	Pending addition		Described in NCBI Biosystem (BSID: 545702)		Q24705869
90-17-5	alpha-(trichloromethyl)benzyl acetate; roseacetol	7007	88586	JKRWZLOCPLZZEI-UHFFFAOYSA-N							Q22052707

84-67-3	4-(4-amino-2-methylphenyl)-3-methylaniline	66537		QYIMZXITLDTULQ-UHFFFAOYSA-N						Toxic to developed lung and cardiac development	Q24705889
95-96-5	3,6-dimethyl-1,4-dioxane-2,5-dione; lactide	7272		JJTUDXZGHPGLLC-UHFFFAOYSA-N							Q421313
592-48-3	1,3-hexadiene	11602		AHAREKHAZNPPMI-UHFFFAOYSA-N							Q24710331
591-24-2	3-methyl-cyclohexanone	11567		UJBOOUHRTQVGRU-UHFFFAOYSA-N							Q22138333
598-25-4	1,1-dimethylallene	11714		PAKGDPSCXSUALC-UHFFFAOYSA-N							Q24711948
598-94-7	1,1-dimethylurea	11737		YBBLOADPFWKNGS-UHFFFAOYSA-N							Q24712449
469-92-1	beta-clovene	590565	88631	NRJSJVYGFMVJAR-UHFFFAOYSA-N				HMDB61840			Q24713391
499-99-0	isolimonene	521268	90014	TWCNAXRPQBLSNO-UHFFFAOYSA-N				HMDB61793			Q22079607
541-28-6	3-methyl-1-iodobutane	10924		BUZZUHJODKQYTF-UHFFFAOYSA-N							Q24716059
500-00-5	p-menth-3-ene	10369	88834	YYCPSEFQLGXPCO-UHFFFAOYSA-N				HMDB37213			Q24716505

0544-10-5	1-chloro-hexane	10992		MLRVZFYXUZQSRU-UHFFFAOYSA-N					mtf00361 Chlorocyclohexane and chlorobenzene degradation; mtf00625 Chloroalkane and chloroalkene degradation ;mtf01100 Metabolic pathways ;mtf01120 Microbial metabolism in diverse environments		Q24716558
473-91-6	1,2,3-trimethyl-1-cyclopentene	136316		YNEQKUWRFVQFPF-UHFFFAOYSA-N							Q24734266
624-80-6	1-ethylhydrazine	29062		WHRIKZCFRVTHJH-UHFFFAOYSA-N					Can be turned into a toxic compound by CUP2E1, therefor carginogenic.		Q24734328
57084-17-0	γ-Undecalactone	7714	37581	PHXATPHONSXBIL-UHFFFAOYSA-N							Q24734376

75-05-8	acetonitrile	6342	38472	WEVYAHXRMPXWCK-UHFFFAOYSA-N				HMDB61869			Q21761558
96-86-2	acetophenone	7410	27632	KWOLFJPFCHCOCG-UHFFFAOYSA-N				HMDB33910	ethylbenzene degradation pathway		Q375112
111-15-9	2-Ethoxyethyl acetate	8095		SVONRAPFKPVNKG-UHFFFAOYSA-N							Q209376
637-50-3	1-phenyl-1-propene	252325		QROGIFZRVHSFLM-QHHAFSJGSA-N							Q2020228
638-28-8	2-chlorohexane	12521		GLCIPJOIEVLTPR-UHFFFAOYSA-N							Q24734466
100-45-8	4-cyanocyclohexene	66013		GYBNBQFUPDFFQX-UHFFFAOYSA-N							Q24730219
105-21-5	gamma-heptalactone	7742	89744	VLSVVMPLPMNWBH-UHFFFAOYSA-N				HMDB31681	Phase I biotransformations, non P450 (Homo sapiens)	Chronic Granulomatous Disease	Q24730325
105-42-0	4-methyl-2-hexanone	7754		XUPXMIAWKPTZLZ-UHFFFAOYSA-N						Non-alcoholic Fatty Liver Diseases (disputed)	Q24730663
78-80-8	isopropenylacetylene	62323		BOFLDKIFLIFLJA-UHFFFAOYSA-N							Q22077076

108-96-3	4-hydroxypyridine	12290	87614	GCNTZFIIOFTKIY-UHFFFAOYSA-N							Q24730939
108-00-9	N,N-dimethyl-1,2-ethanediamine	66053		DILRJUIACXKSQE-UHFFFAOYSA-N						potential to indicate health of patient. If present, may indicate patient is free of Ulcerative colitis, campylobacter jejuni, Clostridium difficile	Q24731206
2390-94-5	trans-tetrahydro-2,5-dimethylfuran	7567603		OXMIDRBAFOEOQT-UHFFFAOYSA-N							Q24734527
868-57-5	2-methylbutanoic acid methyl ester	13357	88538	OCWLYWIFNDCWRZ-UHFFFAOYSA-N				HMDB29762			Q24734848
1731-92-6	heptadecanoic acid methyl ester	15609		HUEBIMLTDXKIPR-UHFFFAOYSA-N							Q24735110
2080-89-9	3-ethyl-1,3-hexadiene	6433321		XTEHSUDXCMUZEH-FPLPWBNLSA-N							Q24735189
645-62-5	2-ethyl-2-hexenal	12582	88838	PYLMCYQHBRSDND-UHFFFAOYSA-N							Q24735451

692-24-0	2-methyl-trans-3-hexene	5352548		IQANHWBWTVLDTP-AATRIKPKSA-N							Q24735526
111-06-08	butyl palmitate; hexadecanoic acid butyl ester; N-Butyl palmitate	8090		GLYJVQDYLFAUFC-UHFFFAOYSA-N							Q24735748
111-67-1	2-octene (both trans and cis)	8126		ILPBINAXDRFYPL-UHFFFAOYSA-N							Q24735790
13389-42-9	trans-2-octene; (2E)-2-Octene	5364448		ILPBINAXDRFYPL-HWKANZROSA-N				HMDB61904			Q24735900
7642-04-08	cis-2-octene; (2Z)-2-Octene	5356796		ILPBINAXDRFYPL-HYXAFXHYSA-N				HMDB61903			Q24736444
112-18-5	N,N-Dimethyldodecylamine; N,N-Dimethyl-1-dodecanamine	8168		YWFWDNVOPHGWMX-UHFFFAOYSA-N							Q24736495
112-52-7	1-chlorodecane, Lauryl chloride	8192		YAYNEUUHHLGGAH-UHFFFAOYSA-N							Q24736563
109-01-03	1-methylpiperazine	53167		PVOAHINGSUIXLS-UHFFFAOYSA-N							Q24736602
112-69-6	N,N-Dimethyl-1-hexadecanamine;	16221		NHLUVTZJQOJKCC-UHFFFAOYSA-N							Q24736646
222-51-5	dibenzopentaphene ; Dibenzo[c,m]pentaphene	136042		RQCIQCDYYKLUBE-UHFFFAOYSA-N							Q24736718
503-28-6	Dimethyldiazene; azomethane	10421		JCCAVOLDXDEODY-UHFFFAOYSA-N							Q22053947
1333-41-1	Methylpyridine; picoline										Q2092388

109-06-08	2-methylpyridine	7975	50415	BSKHPKMHTQYZBB-UHFFFAOYSA-N				HMDB61888			Q2216745
693-89-0	1-methyl-1-cyclopentene	12746		ATQUFXWBVZUTKO-UHFFFAOYSA-N							Q24736866
693-97-0	5-methylisothiazole	136500		LBBKWEDRPDGXPM-UHFFFAOYSA-N							Q24736902
758-16-7	dimethylthioformamide	69794		SKECXRFZFFAANN-UHFFFAOYSA-N							Q24736934
57295-30-4	3-hydroxy-4-methylbenzaldehyde	585182		DHVJHJQBQKKPNB-UHFFFAOYSA-N		PWY-7698	Pending addition		Degraded by several pseudomonas species.		Q24755156

1334-78-7	tolualdehyde ui	10722		BTFQKIATRPGRBS-UHFFFAOYSA-N	ko01220		WP3667			Degrades from o-xylene which is toxic, A human health hazard, can cause fatigue, confusion, headaches, dizziness, and even death.	Q2439424
35915-22-1	Methylbutanoic acid ui (3-methylbutanoic acid)	10430		GWYFCOCPABKNJV-UHFFFAOYSA-N			Pending addition				Q415536
39276-09-0	Furaldehyde ui (2-furaldehyde)	7362		HYBBIBNJHNGZAN-UHFFFAOYSA-N	ko00365		Pending addition		Many degradation pathways in gut microflora.	LD50 in rats of 220-300 mg/kg. Dermal irritant	Q24755507

471-84-1	Aplha-fenchene	28930		XCPQUQHBVVXMRQ-UHFFFAOYSA-N		PWY-6449	Pending addition	HMDB35625	Enzymes for degradation found in leaf extracts of fennel. Pathway described in NCBI Biosystem (BSID: 545571)		Q24715140
26952-23-8	Dichloropropene	11245		ZAIDIVBQUMFXEC-UHFFFAOYSA-N					Pathway described in NCBI Biosystem (BSID: 545568)		Q24755666

695-34-1	4-methylpyridinamine (2-amino-4-methylpyridine)	1533		ORLGLBZRQYOWNA-UHFFFAOYSA-N						Competetively inhibits the catalytic activity of NO synthase (NOS II) enzyme. Could be of great value to treat chronic inflammatory diseases	Q24755764
5131-66-8	1-butoxy-2-propanol	21210		RWNUSVWFHDHRCJ-UHFFFAOYSA-N							Q24762634
74381-40-1	isobutyric acid	6590		KQNPFQTWMSNSAP-UHFFFAOYSA-N							Q415062
2442-49-1	Methyl tetracosanoate	75546		XUDJZDNUVZHSKZ-UHFFFAOYSA-N							Q24762827
624-42-0	6-methyl-3-heptanone	129891		FGPMBONEKHYAHO-UHFFFAOYSA-N							Q24762941
625-27-4	2-methyl-2-pentene	12243		JMMZCWZIJXAGKW-UHFFFAOYSA-N							Q24763010

763-69-9	Ethyl 3-ethoxypropionate	12989		BHXIWUJLHYHGSJ-UHFFFAOYSA-N					Can be degraded by melanised fungi, can be used by C. sphaerospermum (fungus) as a sole carbon and energy source		Q24763201
625-28-5	2-methylbutane secondary mononitrile			QHDRKFYEGYYIIK-UHFFFAOYSA-N							Q24763442
625-54-7	2-ethoxypropane	12256		XSJVWZAETSBXKU-UHFFFAOYSA-N							Q24764257
1731-88-0	methyl tridecanoate	15608		JNDDPBOKWCBQSM-UHFFFAOYSA-N					metabolic intermediate of the oxidation of 1-tetradecene		Q24764359
4403-61-6	2-cyano-2-butene	91586									Q24764474
1077-16-3	Hexylbenzene	14109		LTEQMZWBSYACLV-UHFFFAOYSA-N					Degradation of sec-hexylbenzene and its metabolites by a biofilm-forming yeast Trichosporon asahii B1		Q24764675

27137-41-3	methylfuran										-
628-61-5	1-methylheptylchloride / 2-chlorooctane	12347		HKDCIIMOALDWHF-UHFFFAOYSA-N							Q2476 5603
1551-06-0	2-ethylpyrrole	137075		XRPDDDRNQJNHLQ-UHFFFAOYSA-N							Q2476 5696
2306-89-0	Decyl octanoate	75319		WVWRBUIUZMBLNI-UHFFFAOYSA-N							Q2476 5784
63072-44-6	Methyl pentanone ui	SID: 135154 403		PFCHFHIRKBAQGU-UHFFFAOYSA-N							Q2230 76
0135-01-03	1,2-diethylbenzene	8657		KVNYFPKFSJIPBJ-UHFFFAOYSA-N							Q2429 316
502-26-1	4-Octadecanolide	141998		GYDWWIHJZSCRGV-UHFFFAOYSA-N							Q2481 6483
141-93-5	1,3-diethylbenzene	8864		AFZZYIJIWUTJFO-UHFFFAOYSA-N							Q2481 6743
109-49-9	1-hexen-5-one	7989		RNDVGJZUHCKENF-UHFFFAOYSA-N							Q2481 6828
109-68-2	2-pentene	12585		QMMOXUPEWRXHJS-UHFFFAOYSA-N							Q2205 1121
96-38-8	5-Methyl-1,3-cyclopentadiene	25512		QVRBGKYLLCLCHL-UHFFFAOYSA-N							Q2207 7211
96-39-9	1-Methyl-1,3-cyclopentadiene	66775		NFWSQSCIDYBUOU-UHFFFAOYSA-N							Q5094 303

106-55-8	2,5-Dimethylpiperazine	7816		NSMWYRLQHIXVAP-UHFFFAOYSA-N							Q22079573
110-98-5	1,1'-Oxydi-2-propanol	8087		AZUXKVXMJOIAOF-UHFFFAOYSA-N							Q22829651
112-75-4	N,N-Dimethyltetradecylamine	8211		SFBHPFQSSDCYSL-UHFFFAOYSA-N							Q24817433
124-28-7	N,N-Dimethyloctadecylamine	15365		NAPSCFZYZVSQHF-UHFFFAOYSA-N							Q24817440
144-19-4	2,2,4-Trimethyl-1,3-pentanediol	8946		JCTXKRPTIMZBJT-UHFFFAOYSA-N							Q24817446
460-01-05	2,6-Dimethyl-1,3,5,7-octatetraene	5368451	87600	HPZWSJQQCJZBBG-LQPGMRSMSA-N							Q24817453
513-23-5	isothujol	55250274		SUNFOPSZDVOZOS-UHFFFAOYSA-N							Q24817455
514-14-7	2,7,7-trimethylbicyclo[2.2.1]hept-2-ene	564720		QWEFTWKQGYFNTF-UHFFFAOYSA-N							Q24817463
514-95-4	1,5,5-Trimethyl-6-methylene-cyclohexene	578237		FMXKKHBXBBAQBC-UHFFFAOYSA-N							Q24819025
527-53-7	1,2,3,5-Tetramethylbenzene	10695		BFIMMTCNYPIMRN-UHFFFAOYSA-N							Q24819115
542-54-1	4-Methylpentanenitrile	10956		DUJMVKJJUANUMQ-UHFFFAOYSA-N							Q24819169

APPENDIX B

Table 3. List of the 70 new Wikidata entries, stating the Wikidata ID, the CAS registry number and the name of the compound.

Wikidata ID	CAS	compound name
Q24702624	583-57-3	1,2-Dimethylcyclohexane
Q24514387	586-63-0	Isoterpinolene
Q24702742	590-66-9	1,1-Dimethylcyclohexane
Q24705698	58-90-2	2,3,4,6-Tetrachlorophenol
Q24705869	764-39-6	(E)-2-Pentenal
Q24705889	84-67-3	m-Tolidine
Q24710331	592-48-3	1,3-Hexadiene
Q24711948	598-25-4	1,1-Dimethylallene
Q24712449	598-94-7	1,1-Dimethylurea
Q24713391	469-92-1	beta-Clovene
Q24715140	471-84-1	alpha-Fenchene
Q24716059	541-28-6	1-iodo-3-methylbutane
Q24716505	500-00-5	p-Menth-3-ene
Q24716558	544-10-5	Chlorohexane
Q24730219	100-45-8	4-cyanocyclohexene
Q24730325	105-21-5	gamma-heptalactone
Q24730663	105-42-0	4-methyl-2-hexanone
Q24730939	108-96-3	4(1H)-pyridinone
Q24731206	108-00-9	N,N-dimethyl-1,2-ethanediamine
Q24734266	473-91-6	1,2,3-trimethyl-1-cyclopentene
Q24734328	624-80-6	1-ethylhydrazine
Q24734376	104-67-6	gamma-undecalactone
Q24734466	638-28-8	2-chlorohexane
Q24734527	2390-94-5	trans-tetrahydro-2,5-dimethylfuran
Q24734848	868-57-5	2-methylburyrate
Q24735110	1731-92-6	heptadecanoic acid methyl ester
Q24735189	2080-89-9	3-ethyl-1,3-hexadiene
Q24735451	645-62-5	2-ethyl-2-hexenal
Q24735526	692-24-0	2-methyl-trans-3-hexene
Q24735748	111-06-08	butyl palmitate
Q24735790	111-67-1	2-octene
Q24735900	13389-42-9	trans-2-octene
Q24736444	7642-04-08	cis-2-octene

Q24736495	112-18-5	N,N-Dimethyldodecylamine
Q24736563	112-52-7	1-chlorodecane
Q24736602	109-01-03	1-methylpiperazine
Q24736646	112-69-6	N,N-Dimethyl-1-hexadecanamine
Q24736718	222-51-5	dibenzopentaphene
Q24736866	693-89-0	1-methyl-1-cyclopentene
Q24736902	693-97-0	5-methylisothiazole
Q24736934	758-16-7	dimethylthioformamide
Q24755156	57295-30-4	3-hydroxy-4-methylbenzaldehyde
Q24755507	39276-09-0	2-furaldehyde
Q24755666	26952-23-8	dichloropropene
Q24755764	695-34-1	4-methylpyridinamine
Q24762634	5131-66-8	1-butoxy-2-propanol
Q24762827	2442-49-1	methyl tetracosanoate
Q24762941	624-42-0	6-methyl-3-heptanone
Q24763010	625-27-4	2-methyl-2-pentene
Q24763201	763-69-9	ethyl 3-ethoxypropionate
Q24763442	625-28-5	3-methylbutyronitrile
Q24764257	625-54-7	2-ethoxypropane
Q24764359	1731-88-0	methyl tridecanoate
Q24764474	4403-61-6	2-cyano-2-butene
Q24764675	1077-16-3	Hexylbenzene
Q24765603	628-61-5	2-chlorooctane
Q24765696	1551-06-0	2-ethylpyrrole
Q24765784	2306-89-0	decyl octanoate
Q24816483	502-26-1	4-Octadecanolide
Q24816743	141-93-5	1,3-Diethylbenzene
Q24816828	109-49-9	1-hexen-5-one
Q24817433	112-75-4	N,N-Dimethyltetradecylamine
Q24817440	124-28-7	N,N-Dimethyloctadecylamine
Q24817446	144-19-4	2,2,4-Trimethyl-1,3-pentanediol
Q24817453	460-01-05	2,6-Dimethyl-1,3,5,7-octatetraene
Q24817455	513-23-5	Isothujol
Q24817463	514-14-7	2,7,7-trimethylbicyclo[2.2.1]hept-2-ene
Q24819025	514-95-4	1,5,5-Trimethyl-6-methylene-cyclohexene
Q24819115	527-53-7	1,2,3,5-Tetramethylbenzene
Q24819169	542-54-1	4-Methylpentanenitrile