ANNOTATION OF SEMI-POLAR ORGANIC CONTAMINANTS BY USING GAS CHROMATOGRAPHY COUPLED TO MASS SPECTROMETRY AND MACHINE LEARNING

Milka Ljoncheva

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Assoc. Prof. Dr. Tina Kosjek, Jožef Stefan Institute, Ljubljana, Slovenia
Co-Supervisor: Prof. Dr. Sašo Džeroski, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Ester Heath, Chair, Department of Environmental Sciences, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Prof. Dr. Juho Rousu, Member, Institute for Information Technology, Department of Computer Science, Aalto University, 02150 Espoo, Finland
Assoc. Prof. Dr. Nataša Atanasova, Member, Department of Environmental Civil Engineering, Faculty of Civil and Geodetic Engineering, University of Ljubljana, Hajdrihova cesta 22, 1000 Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Milka Ljoncheva

ANNOTATION OF SEMI-POLAR ORGANIC CONTAMINANTS BY USING GAS CHROMATOGRAPHY COUPLED TO MASS SPECTROMETRY AND MACHINE LEARNING

Doctoral Dissertation

IDENTIFIKACIJA DELNO POLARNIH ORGANSKIH ONESNAŽEVAL Z UPORABO PLINSKE KROMATOGRAFIJE, SKLOPLJENE Z MASNO SPEKTROMETRIJO IN STROJNIM UČENJEM

Doktorska disertacija

Supervisor: Assoc. Prof. Dr. Tina Kosjek

Co-Supervisor: Prof. Dr. Sašo Džeroski

Ljubljana, Slovenia, August 2022

Acknowledgments

I express my immense gratitude to my supervisor Prof. dr. Tina Kosjek, and co-supervisor, Prof. dr. Sašo Džeroski. I will always be grateful for your guidance, scientific and personal support, and all the knowledge essential to my scientific achievements. I would also like to acknowledge Prof. dr. Ester Heath, Dr. David Heath, and Dr. Dušan Žigon for the help and support in instrumental issues, all discussion, and ideas.

I sincerely thank Dr. Tomaž Stepišnik for our fruitful collaboration. Building the multidisciplinary bridge between environmental analytical chemistry and computer sciences was challenging for both research groups. I am grateful for all the shared knowledge and support on the cheminformatics section of this journey.

I address my endless acknowledgment to all current co-workers, Ana Kovačič, Žiga Tkalec, Taja Verovšek, Tamara Gajšt, Marija Laimou-Geraniou, Helena Plešnik, Anja Vehar, and former co-workers Marjeta Česen, Jelena Golubovič, Tjaša Gornik, and David Škufca. Thank you for all your collaborative spirit, valuable suggestions, scientific and non-scientific talks, translations, shared coffees, lunches, walks, and workouts. All these will be a loving memory I will carry in my heart.

I express my warmest gratitude to Tome, Gordana, Ana, Gjorgi, and Spase - thank you for sharing all ups and downs on the crossroads of our academic journies that created a lifetime bond. To Daniela, Marija, Dejana, Stanisha, and Jovana, thank you for being kind, supportive, and understanding lifelong friends. To Prof Gjoshe Stefkov, your words of encouragement and support eased this journey significantly.

I express my biggest acknowledgment to my parents, my sister Katerina, brother-in-law Mariyan, grandmother Katica, and uncle Gjorgi, who left during this journey. Thank you for your endless love and support.

I gratefully acknowledge the funding provided through the AdFutura scholarship by the Public Scholarship, Development, Disability, and Maintenance Fund of the Republic of Slovenia and within the projects “Cycling of substances in the environment, mass balances, modeling of environmental processes and risk assessment” (PI-043) and “Knowledge technologies” (P2-0103) by the Slovenian Research Agency.

Finally, I thank the evaluation board – Prof dr. Ester Heath, Prof. dr. Nataša Atanasova, and Prof. dr. Juho Rousu for the corrections that improved the thesis.

Abstract

Contaminants of emerging concern (CECs), representing a subgroup of organic compounds of natural or synthetic origin, and their degradation and transformation products (TPs), with potentially harmful effects on humans, biota, and the environment, are the eco-exposome (EE) constituents of utmost importance. Their identification, quantification, and continued investigation into their environmental behavior significantly increase our knowledge of their impact on the environment. These challenging tasks require the use of state-of-the-art analytical techniques involving gas chromatography (GC) and liquid chromatography (LC) coupled with mass spectrometry (MS).

While LC-MS is nowadays most commonly employed, GC-MS remains a powerful tool that offers reproducible, sensitive, and relatively low-cost identification and quantification of a broad array of structurally diverse compounds. The range of compounds amenable to GC can also be significantly extended through the derivatization of semi-volatile compounds prior to analysis. The most common derivatization method is silylation, which generates trimethylsilyl (TMS) or *tert*-butyl dimethylsilyl (TBDMS) derivatives. These analytical techniques, together with compound databases (DB), mass spectral libraries (MSL), computational workflows, and cheminformatics approaches, provide accurate and reliable compound annotation (CA). In contrast to LC-MS, however, the use of GC-MS analytical platforms in the *de novo* annotation of CECs, the resulting spectral data in the cheminformatics-assisted annotation of CECs using MS data, and the related challenges regarding method optimization and stability are not widely researched.

This thesis investigates the annotation of semi-polar organic contaminants using both GC-MS and machine learning (ML) approaches. The thesis is divided into three parts. The first part addresses the current state-of-art cheminformatics-assisted CA approaches. Here, we define three crucial cheminformatics tasks in eco-exposome annotation (EEA): molecular formula (MF) assignment, compound prioritization, and CA. A novel methodological classification of CA approaches is provided, along with an assessment of their ability to annotate EE constituents.

The second part of the thesis addresses the generation of GC-electron impact ionization (EI)-MS spectral datasets for developing, validating, and evaluating cheminformatics and ML-based CA approaches. A comprehensive dataset of GC-EI-MS spectra of TMS and TBDMS derivatives was derived from the National Institute of Standards and Technology (NIST) 17, Mass Spectral Library [1] and filtered by relevance, molecular weight (M_w), and the quality of the GC-EI-MS spectra. This classification resulted in two training datasets (1) 4,648 GC-EI-MS spectra of TMS derivatives and (2) 1,883 GC-EI-MS spectra of TBDMS derivatives. Further, two test datasets of GC-EI-MS spectra of about 100 TMS and 85 TBDMS derivatives of CEC were generated by using in-house GC-MS analytical methods. This work was followed by applying a supervised ML approach based on Input Output Kernel Regression (IOKR) for the annotation of CEC silyl derivatives by using GC-EI-MS spectra. The IOKR approach correctly ranked 37% and 50% of the tested CEC-TMS derivatives among the top 10 and 20 candidates. The satisfactory identification rates show that the IOKR approach can be successfully employed in reliable and faster CA compared to manual MSL search approaches.

The third part of the thesis investigates silylation procedures, particularly the stability of silyl derivatives of CEC under different storage conditions and their associated measurement uncertainty (MU). We optimized the derivatization conditions of 70 CEC using N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA), N, O-bistrifluoroacetamide (BSTFA) and N, O-bistrifluoroacetamide + 1% trimethylchlorosilane (BSTFA + 1% TMCS) in 36 different temperature and duration experiments. Further, we tested their stability in a solvent and artificial wastewater (AWW) extract under relevant storage conditions (25°C, 4°C, and -18°C) for

up to 20 weeks, along with five cycles of freezing and thawing. Significant stability issues were revealed for TMS derivatives of polyhydroxy compounds and estrogen hormones, in addition to derivatives degraded to $\leq$ 85% of their initial concentration after only two freezing and thaw cycles.

The results of this thesis are gathered in two published papers and two manuscripts submitted for peer review. They highlight the importance of silylation conditions in reliable CEC annotation and quantification and provide insight into the stability profiles of TMS derivatives. In addition, this thesis demonstrates the successful employment of ML and GC-EI-MS in identifying CEC as silyl derivatives for the first time. The performed work resulted in the generation of comprehensive datasets that are publicly available and of interest to the ML community for further development of ML-based CA approaches.

Povzetek

Onesnažila, ki vzbujajo nastajajočo zaskrbljenost (CEC), so tipične organske spojine naravnega ali sintetičnega izvora ter produkti njihove razgradnje in pretvorbe (TP) s potencialno škodljivimi učinki na človeka, bioto in okolje. Te spojine so izjemno pomemben del eko-ekspozoma (EE). Njihova identifikacija in kvantifikacija ter raziskovanje njihovega okoljskega obnašanja bistveno večajo naše poznavanje njihovega vpliva na onesnaževanje okolja. V tem kontekstu je nepogrešljiva uporaba analitskih tehnik, zlasti plinske kromatografije (GC) in tekočinske kromatografije (LC), sklopljene z masno spektrometrijo (MS).

Kljub temu da se najpogosteje uporablja LC-MS, je GC-MS konvencionalni analitični sistem, ki ponuja ponovljivo, občutljivo in razmeroma poceni identifikacijo in kvantifikacijo širokega nabora strukturno različnih spojin. Nabor spojin se dodatno razširi z derivatizacijo pred analizo, najpogosteje s silacijo, pri kateri nastanejo derivati trimetilsilila (TMS) ali tert-butil dimetilsilila (TBDMS). Te analitične tehnike skupaj z bazami podatkov o spojinah (DB), knjižnicami masnih spektrov (MSL), računalniškimi delotoki in pristopi kemoinformatike zagotavljajo natančno in zanesljivo anotacijo spojin (CA).

V nasprotju z LC-MS je uporaba GC-MS pri de novo anotaciji CEC, skupaj z uporabo nastalih spektralnih podatkov v kemoformatično podprti anotaciji CEC ter s tem povezanimi izzivi glede optimizacije in stabilnosti metode, premalo raziskana.

Doktorska disertacija raziskuje označevanje (anotacijo) delno polarnih organskih onesnaževal z uporabo pristopov GC-MS in strojnega učenja (ML). Disertacija je razdeljena na tri dele. Prvi del obravnava trenutno stanje keminformatiskih CA pristopov. Tukaj definiramo tri ključne naloge keminformatike pri anotaciji eko-ekspozoma (EEA): dodelitev molekulske formule (MF), prioritizacija spojin in anotacija spojin (CA). Podana je nova metodološka klasifikacija pristopov CA skupaj z oceno njihove učinkovitosti pri anotaciji komponent EE. Drugi del doktorske disertacije obravnava generiranje spektralnih podatkovnih naborov z GC, in sicer MS za razvoj, validacijo in vrednotenje pristopov CA, ki temeljijo na keminformatiki in zlasti na ML. Obsežen nabor podatkov GC-EI-MS spektrov TMS in TBDMS derivatov je bil pridobljen iz knjižnice masnih spektrov Nacionalnega inštituta za standarde in tehnologijo (NIST) 17 [1], ki smo jih filtrirali za kemijsko pomembnost spojin, molekulsko maso spojin (Mw) in kakovost GC-EI-MS spektrov. Rezultat filtriranja sta dva končna nabora podatkov za učenje z ML pristopi. Prvi je sestavljen iz 4,648 GC-EI-MS spektrov TMS derivatov, drugi pa iz 1,883 GC-EI-MS spektrov TBDMS derivatov. Poleg tega sta bila z uporabo analitičnih metod GC-MS ustvarjena dva nova testna nabora podatkov GC-EI-MS spektrov, s približno 100 TMS in 85 TBDMS derivatov CEC. Temu je sledila uporaba pristopa nadzorovanega ML, ki temelji na regresiji vhodno-izhodnih jeder (ang. Input-Output Kernel Regression, IOKR), za anotacijo sililnih derivatov CEC z uporabo GC-EI-MS spektrov. Pristop IOKR je pravilno uvrstil 37 % oziroma 50 % testiranih CEC-TMS derivatov med 10 najboljših oziroma 20 najboljših kandidatov. Zadovoljive stopnje identifikacije kažejo, da je pristop IOKR mogoče uspešno uporabiti v zanesljivi in hitrejši anotaciji v primerjavi z ročnimi pristopi iskanja v knjižnicah masnih spektrov.

Tretji del doktorske disertacije raziskuje postopke siliranja, predvsem stabilnost sililnih derivatov širokega nabora CEC pri različnih pogojih shranjevanja, in s tem povezano merilno negotovost (MU). Optimizirali smo pogoje siliranja za optimalno učinkovitost derivatizacije s testiranjem učinkovitosti derivatizacije 70 CEC z N-metil-N-(trimetilsilil) trifluoroacetamidom (MSTFA), N, O-bistrifluoroacetamidom (BSTFA) in N, O-bistrifluoroacetamidom + 1 % trimetilklorosilanom (BSTFA + 1 % TMCS) v 36 različnih poskusih z različnimi temperaturami in trajanjem. Poleg tega smo testirali njihovo stabilnost v topilu in ekstraktu umetne odpadne vode (AWW) pri ustreznih pogojih shranjevanja (25 °C, 4 °C in -18 °C) do 20 tednov, skupaj s petimi cikli zamrzovanja in odmrzovanja. Poleg drugih TMS derivatov, za katere je bilo dokazano, da se

razgradijo na $\leq$ 85 % njihove začetne koncentracije po dveh ciklih zamrzovanja in odmrzovanja vzorca, so bile odkrite pomembne težave s stabilnostjo derivatov polihidroksi CEC in estrogenskih hormonov.

Rezultati te doktorske disertacije so zbrani v treh objavljenih člankih in enem rokopisu, ki je bil oddan v recenzijo. Rezultati poudarjajo pomen silacijskih pogojev pri zanesljivi anotaciji in kvantifikaciji CEC ter zagotavljajo vpogled v profile stabilnosti TMS derivatov. Prav tako je v tem doktorskem delu prvič prikazana uspešna uporaba ML in GC-EI-MS pri identifikaciji sililnih derivatov CEC. Rezultat opravljenega dela so celoviti nabori GC-EI-MS podatkov, ki so javno dostopni in zanimivi za ML skupnost za nadaljnji razvoj pristopov anotacije spojin, ki temeljijo na strojnem učenju.

Contents

List of Figures

List of Tables

Abbreviations

Δ^9-THC	. . .	(-)- Δ^9 tetrahydrocannabinol
Δ^9-THCA	. . .	(-)- Δ^9 tetrahydrocannabinolic acid
11-HAD	. . .	11α-hydroxyandrostenedione
11-HT	. . .	11α-hydroxytestosterone
11N9THCA	. . .	(±)-11-nor-9-carboxy- Δ^9-tetrahydrocannabinol
11OHTHC	. . .	(±)-11-hydroxy- Δ^9- tetrahydrocannabinol
17-HP	. . .	17α-hydroxyprogesterone
2AA	. . .	2-anilinophenylacetic acid
22BPF	. . .	2,2'-methylenediphenol
24BPF	. . .	2,4-methanediyldiphenol
3MC	. . .	3-methylcatechol
3M5NC	. . .	3-methyl-5-nitrocatechol
4,4'-BP	. . .	4,4'-biphenol
4-NG	. . .	4-nitroguiaiacol
4-NS	. . .	4-nitrosyringol
4-OP	. . .	4-tert-octylphenol
5-A3B17B	. . .	5- α -androstene-3- β ,17- β -diol
5-AD	. . .	5-androsten-3β, 17β-diol
5-NG	. . .	5-nitroguaiacol
6-HP	. . .	6β-hydroxypregnenolone
6-MAM	. . .	6-monoacetylmorphine
6-NG	. . .	6-nitroguaiacol
8-HQ	. . .	8-hydroquinone
9-HF	. . .	9-hydroxyfluorene
AA	. . .	adipic acid
AWW	. . .	artificial wastewater
ACN	. . .	acetonitrile
AMPH	. . .	(±)-amphetamine
APCI	. . .	atmospheric pressure chemical ionization
API	. . .	atmospheric pressure ionization
APPI	. . .	atmospheric pressure photoionization
ARP	. . .	absolute ranking position
BP-8	. . .	2,2'-dihydroxy-4-methoxybenzophenone
BA	. . .	benzoic acid
BD	. . .	boldenone, 5 α -dihydrotestosterone
BHT	. . .	butylated hydroxytoluene
BPA	. . .	bisphenol A
BPAF	. . .	bisphenol AF
BPAP	. . .	bisphenol AF
BPB	. . .	bisphenol B
BPBP	. . .	bisphenol BP
BPC	. . .	bisphenol C
BPCL	. . .	bisphenol CL
BPE	. . .	bisphenol E
BPF	. . .	bisphenol F
BPFL	. . .	bisphenol FL
BPM	. . .	bisphenol M
BPP	. . .	bisphenol P
BPPH	. . .	bisphenol PH
BPS	. . .	bisphenol S
BPZ	. . .	bisphenol Z
BSA	. . .	N, O-bis(trimethylsilyl)acetamide

BuPb	. . .	butylparaben
BZECG	. . .	benzoylecgonine
BzPb	. . .	benzyl paraben
CA	. . .	citric acid
CA	. . .	compound annotation
CASMI	. . .	Critical Assessment of Small Molecule Identification
CAT	. . .	catechol
CBC	. . .	cannabichromene
CBD	. . .	cannabidiol
CBN	. . .	cannabinol
CBZ	. . .	carbamazepine
CE	. . .	capillary electrophoresis
CEC	. . .	contaminant of emerging concern
CI	. . .	chemical ionization
CID	. . .	collision-induced dissociation
CLA	. . .	clofibric acid
CLP	. . .	clorophene (2-benzyl-4-chlorophenol)
COD	. . .	codeine
CSI	. . .	Compound Structure Identification
Da	. . .	Dalton
DASI	. . .	Database Assisted Structure Identification
DB	. . .	database
DF	. . .	diclofenac
DH-BP	. . .	2,4-dihydroxybenzophenone
DHDPE	. . .	4,4'-dihydroxydiphenyl ether
E1	. . .	estrone
E2	. . .	estradiol
E3	. . .	estriol
ECD	. . .	electron-capture detector
EE	. . .	eco-exposome
EE2	. . .	17 α -ethynyl estradiol
EEA	. . .	eco-exposome annotation
EI	. . .	electron impact ionization
ERY	. . .	erythritol
ET	. . .	etofylline
ESI	. . .	electrospray ionization
EtAc	. . .	ethyl acetate
EtOH	. . .	ethanol
EtPb	. . .	ethylparaben
FID	. . .	flame ionization detector
FPD	. . .	flame photometric detector
FSEA	. . .	fragment set enrichment analysis
FT-ICR	. . .	Fourier transform-ion cyclotron resonance
FWHM	. . .	full width half maximum
GC	. . .	gas chromatography
H-BP	. . .	4-hydroxybenzophenone
HMDS	. . .	hexamethyldisilazide
HPLC	. . .	high-performance liquid chromatography
HPP	. . .	4-cumylphenol
HR/AM-MS	. . .	high resolution/accurate mass - mass spectrometry
IB	. . .	ibuprofen
IBuPb	. . .	isobutylparaben
IM-MS	. . .	ion mobility-mass spectrometry
IOKR	. . .	Input Output Kernel Regression
IPrPb	. . .	isopropylparaben
IT	. . .	ion trap
IT-TOF	. . .	ion trap-time-of-flight
KET	. . .	ketoprofen

LAA	. . .	L-ascorbic acid
LC	. . .	liquid chromatography
L-LEU	. . .	L-leucine
LR-MS	. . .	low resolution-mass spectrometry
L-SER	. . .	L-serine
LTQ	. . .	linear trap quadrupole
L-TYR	. . .	L-tyrosine
MALDI	. . .	matrix-assisted laser desorption/ionization
MAMPH	. . .	methylenedioxymethamphetamine
MCA	. . .	m-coumaric acid
MEC	. . .	mecoprop
MeOH	. . .	methanol
MePb	. . .	methyl paraben
MF	. . .	molecular formula
MFR	. . .	match factor
ML	. . .	machine learning
MORPH	. . .	morphine
MRM	. . .	multiple reaction monitoring
MS	. . .	mass spectrometry
MSA	. . .	N-methyl-N-trimethylsilylacetamide
MS^{all}	. . .	all ion fragmentation
MSD	. . .	mass selective detector
MS/MS	. . .	tandem mass spectrometry
MSL	. . .	mass spectral library
MSTFA	. . .	N-methyl-N-(trimethylsilyl)trifluoroacetamide
MS^n	. . .	multiple stage mass spectrometry
MTBSTFA	. . .	N-tert-butyldimethylsilyl-N-methyltrifluoroacetamide
MU	. . .	measurement uncertainty
M_w	. . .	molecular weight
m/z	. . .	mass-to-charge ratio
NAP	. . .	naproxen
NIST	. . .	National Institute of Standards and Technology
NL	. . .	nylidrin
NMR	. . .	nuclear magnetic resonance
NTS	. . .	non-targeted screening
NX	. . .	nitroxolone
OCA	. . .	o-coumaric acid
PAA	. . .	phenylacetic acid
PCA	. . .	p-coumaric acid
PID	. . .	photoionization detector
ppm	. . .	parts per million
PrPb	. . .	propyl paraben
QA	. . .	quinic acid
Q-Orbitrap	. . .	quadrupole-Orbitrap
QQQ	. . .	triple quadrupole
Q-TOF	. . .	quadrupole-TOF
RES	. . .	resorcinol
RMFR	. . .	reverse match factor
RRF	. . .	relative response factor
RRP	. . .	relative ranking position
R_t	. . .	retention time
RW	. . .	river water
SA	. . .	salicylic acid
S/N	. . .	signal-to-noise ratio
SFA	. . .	sulfanilamide
SHA	. . .	shikimic acid
SIM	. . .	selected ion monitoring
SPE	. . .	solid-phase extraction

SRM	. . .	selected reaction monitoring
SS	. . .	suspect screening
STA		stanolone
SYR	. . .	syringol
T3HC	. . .	trans-3′-hydroxycotinine
TBDMCS	. . .	*tert*-butyldimethylchlorosilane
TBDMS	. . .	*tert*-butyl dimethylsilyl
TCD	. . .	thermal conductivity detector
TCS	. . .	triclosan
TID	. . .	thermionic ionization detector
TLC	. . .	thin-layer chromatography
TMBA	. . .	4,4′-isopropylidenebis(2,6-dimethylphenol)
TMCS	. . .	trimethylchlorosilane
TMS	. . .	trimethylsilyl
TMS-DEA	. . .	trimethylsilyldiethylamine
TMSI	. . .	trimethylsilylimidazole
TOF	. . .	time-of-flight
UA	. . .	urea
UV	. . .	ultraviolet
UHPLC	. . .	ultra-high-performance liquid chromatography
WWE	. . .	wastewater effluent
WWI	. . .	wastewater influent

Glossary

accurate mass	experimentally determined mass of an ion measured to an certain degree of accuracy and precision
atmospheric pressure chemical ionization	ionization technique in which the reactant ions are generated by photoionization of suitable dopant species and subsequent ion/molecule reactions of their molecular ions
atmospheric pressure photoionization	ionization technique for direct ionization of molecules at atmospheric pressure by electron detachment induced photons by forming M^+ ions
background spectrum	the mass spectrum is observed when no analyte is introduced into the mass spectrometer
base peak	a peak in an MS spectrum that has the highest intensity
bond dissociation energy	the energy required to break a bond to generate a fragment
chemical ionization	ionization technique in which neutral molecules fragment by reacting with an excess of ions. The process may involve the transfer of an electron, a proton, or other charged species between the reactants
compound database	an organized online repository of chemical compounds containing one or more of the following: name, CAS number, database ID, chemical structure, molecular descriptors, predicted and experimentally determined physico-chemical properties, toxicity, environmental behavior, MS data, and NMR data, etc.
compound annotation	process of linking a detected mass spectrometric feature with a chemical identity, taking into account the detected chromatographic and spectrometric characteristics
compound identification	process of proving or verifying that the annotated compound is indeed the proposed chemical so that the annotation can be confirmed
contaminant of emerging concern	contaminants, either natural or synthetic chemicals that have been recently (de novo) detected in one or more environmental compartments, present as parent compounds or transformed into new compounds, with potentially harmful effects to humans, biota, and the environment
cross-validation	method for evaluation of predictive models by partitioning the original sample into a training set to train the model and a test set for evaluation
data acquisition (in mass spectrometry)	process of sampling signals that measure a specific sample and converting them into a digital form that can be manipulated by a computer and software

data processing	organizing and manipulating data according to a set of instructions; in MS, it transforms representations of spectrometric signals from their original form into representations that would allow their further analysis (e.g., identification and quantification)
derivatization	a chemical reaction that yields a product that is more volatile and stable and that has improved gas chromatographic behavior over the original substance
detection	collection of compound-specific data by instrumental analysis; in MS, this is R_ts, *m/z* of molecular ions, adducts, and possible fragment ions, the presence and relative abundances of fragment ions and isotopologues
diagnostic ion	product ion whose formation reveals structural or compositional information about its precursor ion
eco-exposome	a segment of the exposome that accounts for all environmental contaminants entering the human body through air, food, water, and dust and endogenous metabolites produced as a response to inflammation, (oxidative) stress, infections, and other natural processes, considering both organism and ecosystem exposure
electron ionization	ionization technique that removes one or more electrons from an atom or molecule through interactions with electrons that are typically accelerated to energies between 10 and 150 eV
electrospray ionization	spray ionization technique in which either cations or anions in solution are transferred to the gas phase via formation and desolvation at atmospheric pressure of a stream of highly charged droplets that result from applying a potential difference between the tip of the electrospray needle containing the solution and a counter electrode
exact mass	the calculated mass of an ion whose elemental formula, isotopic composition, and charge state are known
exposome	the sum of all environmental exposures (including lifestyle factors) from the prenatal period onwards
extracted ion chromatogram	chromatogram created by plotting the intensity of the signal observed at a chosen m/z value or set of values in a series of mass spectra recorded as a function of retention time
fragment ion	product ion that results from the dissociation of a precursor ion
fragmentation	the systematic process of bond breakage in order to remove the excess energy, restoring stability to the resulting ion
fragmentation pattern	the sum of consecutive fragmentation reactions of a chemical structure during ionization, beginning from the parent ion, that results in the formation of more than one fragment ion and neutral losses; it is an ionization technique- and structure (class)-specific process
fragmentation tree	graph representation that models the fragmentation process of a compound, in which each node assigns a molecular formula to a fragment peak arising from a (hypothetical) fragmentation step, and each edge

	represents fragmentation reaction, labeled with the molecular formula of the corresponding loss
hard ionization technique	formation of gas-phase ions accompanied by extensive fragmentation
ion	atomic, molecular, or radical species with a non-zero net electric charge
ionization	process of generation of one or more ions, e.g., by loss of an electron from a neutral molecular entity, by the unimolecular heterolysis of such an entity into two or more ions, or by a heterolytic substitution reaction involving neutral molecules
ion source	the region in a mass spectrometer where gas-phase ions are produced
isomers	compounds with identical elemental composition but different arrangement of atoms in the molecule and different properties
kernel	similarity measure applied on a data instance to map the original non-linear observations into a higher-dimensional space
machine learning	branch of artificial intelligence and computer science that focuses on the use of data and algorithms to imitate the way humans learn, gradually improving its accuracy
mass accuracy	the relative difference between the measured and the theoretical *m/z* value, calculated as 10^6 x $(m/z_{EXP}-m/z_{THEOR})/m/z_{THEOR}$
mass spectral library	an organized collection of mass spectra of different compounds represented as two-dimensional (*m/z* and intensity) peak lists, with or without accompanying metadata related to the compound (e.g., name, molecular descriptors, structure, IDs, and physicochemical properties)
mass spectrometer	an instrument that measures the *m/z* values and abundances of gas-phase ions
mass spectrometry	study of matter through the formation of gas-phase ions that are characterized using mass spectrometers by their mass, charge, structure, and physicochemical properties
mass spectrum	a plot of the relative abundances of ions forming a beam or other collections as a function of their *m/z* values
mass-to-charge (*m/z*) ratio	the unitless ratio of the mass number of the ion to the number of fundamental charges z on the ion
molecular ion	the ion formed by the removal of one or more electrons from a molecule to form a positive ion or the addition of one or more electrons to a molecule to form a negative ion
molecular fingerprint	bit vectors of defined length, where each bit represents the presence or absence of a specific substructure in the chemical structure of a given compound
monoisotopic mass	the exact mass of a compound calculated using the mass of the most abundant isotope of each element

multiple stage mass spectrometry	multiple stages of precursor ion m/z selection followed by product ion detection for successive nth-generation product ions
neutral loss	loss of an uncharged species from an ion during dissociation
nominal mass	the mass of an ion or molecule calculated using the mass of the most abundant isotope of each element rounded to the nearest integer value and equivalent to the sum of the mass numbers of all constituent atoms
precursor ion	an ion that reacts to form particular product ions or undergoes specified neutral losses
resolution	the ability to distinguish two peaks of slightly different m/z expressed as $\Delta m/z$ for a given m/z value
resolving power	m/z value of a particular peak divided by the peak FWHM, i.e., RP= $(m/z)/\Delta m/z$
supervised machine learning	class of machine learning systems and algorithms that uses labeled datasets to train algorithms that classify data or predict outcomes accurately
selected ion monitoring	operation of a mass spectrometer in which the abundances of ions of one or more specific m/z values are recorded rather than the entire mass spectrum
soft ionization technique	formation of gas-phase ions without extensive fragmentation
total ion current	the sum of all separate ion currents carried by the ions of different m/z contributing to a complete mass spectrum or in a specified m/z range of a mass spectrum
tandem mass spectrometry	acquisition and study of the spectra of the product ions or precursor ions of m/z selected ions or precursor ions of a selected neutral loss
total ion current chromatogram	chromatogram created by plotting the total ion current in a series of mass spectra recorded as a function of retention time
unsupervised machine learning	set of machine learning systems and algorithms that use unlabeled datasets to analyze, discover hidden patterns or data groupings without the need for human intervention, and cluster data appropriately

Chapter 1

Introduction

This dissertation combines approaches from environmental analytical chemistry, in particular gas chromatography-mass spectrometry (GC-MS) and machine learning (ML), for annotating contaminants of emerging concern (CECs). In this chapter, we give the background and motivation of this dissertation, followed by the aims of the thesis, its goals, hypotheses, and scientific contributions. The chapter concludes with an overview of the thesis structure.

1.1 Background and Motivation

Exposomics is one of the fastest developing ˈomicsˈ sciences. It arose almost two decades ago as a mixture of environmental research, metabolomics, and toxicology [2], [3]. The focus of its interest is the annotation of the exposome, which comprises all non-genetic factors that influence a phenotype and are responsible for a significant portion of the reisk of chronic diseases [2]. Of particular focus is its subarea, the eco-exposome (EE), studying both external and internal markers of exposure, determining exposures from the point of contact between an external environmental stressor and a receptor inward into the organism and outward to the general environment [4].

Characterization of the EE in environmental exposomics is a highly challenging task due to the immense structural and toxicological diversity of its constituents, along with our limited knowledge of their identity. The standard analytical platforms for selective compound separation are GC and liquid chromatography (LC) coupled with a wide variety of MS analyzers, depending on the compound- and concentration range to which the analysis has to adapt. The output of the mass analyzer is a mass spectrum, which plots the mass to charge (*m/z)* ratios of ions against their intensities. Mass spectra are the essential input for further compound identification, often regarded as a retrieval task. Given an acquired mass spectrum of an unknown compound, the aim is to find a set of candidate compounds from a mass spectral library (MSL) with similar MS spectra. The conventional approach is to match the acquired MS spectrum against reference MSL and rank the matching MS spectra according to their measured spectral similarity to the query MS spectrum. In such a way, the most probable hit, i.e., the MS spectrum with the highest match to the queried MS spectrum, is ranked first. The process is laborious and error-prone, often leading to unreliable compound annotation (CA). Additionally, even now, in the era of their substantial growth in size and comprehensiveness, compound DBs and MSLs lack a significant fraction of EE-relevant chemical information, for example, information on derivatives of CECs.

Semi-volatile and thermolabile CECs are typically analyzed using LC-MS and less frequently by GC-MS. Apart from often requiring an additional time-consuming derivatization step, the main reason for the limited use of GC-MS are the limited opportunities for CA in the case of silyl derivatives. These compounds are poorly represented in compound DBs, and their GC-EI-MS spectra are rarely included in MSLs. Moreover, the cheminformatics CA approaches, mainly the ML-based CA approaches, are almost exclusively developed for use with LC-MS data. Therefore, by recognizing and resolving the limitations of the existing and developing new workflows for ML-based CA, with particular attention to GC-EI-MS data of silyl derivatives, it is possible to improve the applicability of GC-EI-MS data in CA. Additionally, the contribution of in-house

generated datasets of GC-EI-MS data of silyl derivatives to existing MSLs would enrich the corpus of compounds available for CA by direct MSL matching.

In order to support the role of GC-MS analytical methods in CA, research on the factors determining the derivatization efficiency is required. Thus, besides optimization of the derivatization conditions, knowledge about the stability of silyl derivatives under different conditions during sample preparation, storage, and instrumental analysis are needed for their reliable identification and quantification. The currently available scientific literature in this field is scarce, with limited stability data on certain CEC groups, such as estrogen hormones, parabens, and other endocrine disrupting CECs, including benzophenones and bisphenols.

1.2 Purpose of the Dissertation

This dissertation aims to investigate CA for a representative selection of physicochemically and structurally diverse CECs, using GC-EI-MS spectral data and ML approaches. We specifically focus on supervised kernel-based ML approaches. The aim is to gain knowledge of the limiting factors obstructing the everyday use of GC-MS in eco-exposome annotation (EEA). Motivated by the issues above, this work is intended to fill the existing knowledge gap related to the identification/annotation of CECs using GC-MS, including:

- the absence of cheminformatics, and especially ML-based approaches for the annotation of CECs using GC-EI-MS data;
- insufficient standardized GC-EI-MS spectral datasets, specifically of silyl derivatives;
- the limited presence of silyl derivatives in DBs and their spectra in MSLs, which impairs their annotation by using CA approaches;
- the lack of data regarding the stability of silyl derivatives of CECs in versatile matrices (solvent, artificial wastewater (AWW) extract) and storage conditions prior to and during GC-MS analysis.

Based on the identified knowledge gap, this dissertation covers four main topics:

(1) The development of a methodological workflow for generating GC-EI-MS spectral datasets for cheminformatics-assisted CA: The first part of the dissertation (Chapter 3) gives a thorough overview, and novel classification of existing cheminformatics approaches for CA, focusing on ML approaches. It also provides a comparison of their performance in the annotation of EE constituents. Based on existing publications, including comparative studies, we identified the lack of publicly available benchmark spectral datasets, especially GC-EI-MS spectral datasets, as a significant obstacle to a robust evaluation of CA performance for existing and novel cheminformatics approaches [5]. Chapter 4 presents a three-step filtering workflow for generating (or rather curating) GC-EI-MS datasets of silyl (TMS and TBDMS) derivatives from existing MSLs. In the first step, rule-based filtering is applied to exclude chemically illogical compounds, i.e., compounds containing Si atoms that are not the result of typical silylation reactions occurring during derivatization. In the second and third steps, Si-containing compounds with large molecular masses and low-quality GC-EI-MS spectra are eliminated, respectively. In this way, we ensure the generation of GC-EI-MS spectra datasets that adequately serve ML-based CA approaches' training (Chapter 4). Additionally, *de novo* GC-EI-MS spectral datasets of CEC-TMS and CEC-TBDMS derivatives are generated *de novo* by employing in-house GC-MS analytical methods and ensure they can be used to test ML-based CA approaches.

(2) Public accessibility of the generated GC-EI-MS spectral datasets: The aim is to make the GC-EI-MS spectral datasets of silyl derivatives generated in this thesis publicly available. The in-house generated GC-EI-MS spectral datasets, the related metadata, and the metadata of the MSL-derived GC-EI-MS spectral datasets are publicly available in the Mendeley Data Repository.

(3) Application of an ML-based approach for annotating semi-polar CEC silyl derivatives: The current application of ML-based approaches to annotating CECs, as EE constituents are limited and are much less common than in metabolomics and other “omics” fields [5]. Most EEA studies perform CA using workflows consisting of data processing and MF determination by vendor-

specific software and structure elucidation by matching MS spectra against publicly available or vendor-specific MSLs, metadata, and expert knowledge. However, there are no studies of cheminformatics-and specifically ML-based approaches for identifying EE constituents using structural information from GC-EI-MS spectra of their silyl derivatives. For this purpose, a cutting-edge supervised ML approach based on Input Output Kernel Regression (IOKR) [6] is applied for identifying CEC silyl derivatives utilizing structural information inherent to their GC-EI-MS spectra. Achieving satisfactory performance in our study (Section 4.2), further scientific efforts in applying ML-based approaches in the field of EEA are encouraged.

(4) Investigation of derivatization conditions and stability of silyl derivatives of structurally and physicochemically diverse CECs: Reliable ML-based annotation of semi-polar CEC silyl derivatives requires generation of MS data under optimized derivatization conditions with maximum derivatization efficiencies. Also, appropriate knowledge of the behavior of the generated derivatives in the GC-MS system is required. Such knowledge includes insights into their stability profiles during sample storage and GC-MS analysis in different samples. Tens to a few hundreds of CECs are identified and quantified in complex environmental samples during EEA studies. CECs have versatile structures and physicochemical properties and are in different concentration ranges. Studies of optimizing the generic derivatization protocols and examining the stability of a broad range of CEC-silyl derivatives ensure knowledge upon which the behavior of CEC in complex samples would be predicted. To date, few studies have investigated the derivatization of CEC, the stability of the generated TMS derivatives, and the associated measurement uncertainty (MU), but only to a limited selection of CECs and under limited conditions that do not ensure reliable knowledge of the topic [7]–[16]. For this reason, we proposed a methodology for optimizing derivatization conditions for 70 CECs that employs numerous chemometrics tools. The optimized derivatization protocols, along with validated multi-residue GC-MS analytical methods and estimation of MU, are used to investigate the stability of 70 CEC-TMS derivatives in two matrices at the three most common storage temperatures (25°C, 4°C, and -18°C) for up to 20 weeks and during consecutive sample freezing and thawing cycles. Based on the results, CECs whose TMS derivatives are most sensitive to degradation are identified, and the optimum storage conditions are proposed for samples that contain them.

1.3 Aims and Hypotheses

The goals and hypotheses of this dissertation are aligned with its purposes, as described in Section 1.2. The goals of this dissertation are as follows:

- To investigate the stability of silylated derivatives of CECs in relevant matrices (solvent, wastewater effluent (WWE)) and under relevant conditions (storage at room temperature – during preparation and on the autosampler tray, and storage in a refrigerator (4°C) and in the freezer (-18°C)
- To develop a workflow to generate training datasets of GC-EI-MS spectra of silylated derivatives from existing MSLs
- To generate in-house reference datasets of GC-EI-MS spectra of silylated derivatives;
- To apply a ML approach for CA using GC-EI-MS spectra, identify silylated derivatives from CC-EI-MS spectra, and evaluate its performance.

I believe that by investigating the stability of CEC silyl derivatives in relevant matrices and under relevant conditions, we will discover significant patterns of behavior of the CEC-TMS derivatives prior to and during GC-MS analysis. This new knowledge will guide appropriate sample storage and handling of complex environmental samples before and during GC-MS analysis. Consequently, the presence and CA of a potentially broad spectrum of CEC with versatile structural and physicochemical properties will be performed with increased confidence and accuracy. Further, we believe that applying a ML-based approach for CA using GC-EI-MS spectral data of silyl derivatives, which has not been attempted to date, would result in satisfactory performance in terms of annotation accuracy and confidence. In order to evaluate the performance of cheminformatics CA, especially ML-based approaches, we developed a workflow for the generation of training datasets of GC-EI-MS spectra of silyl derivatives from MSLs, which will provide curated datasets of spectra with satisfactory quality. Along with the

training datasets, the intention was to generate de novo in-house reference datasets of GC-EI-MS spectra of silyl derivatives, serving as test datasets for evaluating CA approaches.

Based on the above dissertation goals, the research in this dissertation will test the following hypotheses:

H1: Poor stability and chromatographic behavior (peak shape, R_t) of silylated derivatives can reveal patterns, which can reduce the confidence and accuracy of their identification and quantification of derivatized CECs.

H2: Structurally diverse semi-polar and thermolabile CECs can be successfully identified through their silylated derivatives using their GC-EI-MS spectra in complex mixtures.

H3: An ML approach using training and test datasets of GC-EI-MS spectra of silylated compounds will result in a higher number of correctly identified compounds than a non-ML approach.

We hypothesize that poor stability and chromatographic behavior (peak shape, R_t) of silyl derivatives can reveal patterns in their behavior prior to and during GC-MS analysis, as well as during storage, which can reduce the confidence and accuracy of their identification and quantification. Once discovered and understood, it can be hypothesized that structurally diverse semi-polar and thermolabile CECs can be successfully identified by their silyl derivatives using their GC-EI-MS spectra in complex mixtures. In particular, we hypothesize that an ML-based approach using training and test datasets of GC-EI-MS spectra of silyl derivatives will significantly improve CA compared to non-ML approaches.

1.4 Scientific Contributions

The scientific contributions of the dissertation are as follows:

Contribution 1: A thorough review of the currently available cheminformatics-based CA approaches is provided. For the first time, we defined the three crucial cheminformatics tasks of EEA: MF assignment, compound prioritization, and CA. We also discussed the methodologies employed for each task, emphasizing the last task. CA approaches that utilize structural information inherent to MS data are classified into three classes: direct, indirect, and joint annotation approaches. We also discussed their performance in terms of the ability to annotate EE constituents and have discussed current bottlenecks and future directions to new CA strategies. Finally, performance evaluation protocols are reviewed, identifying the issues currently obstructing their employment in regular EEA workflows.

The publication describing this contribution, included in the thesis, is as follows:

Journal paper: Ljoncheva, M., Stepišnik, T., Kosjek, T., Džeroski, S. (2020) Cheminformatics in MS-based environmental exposomics: current achievements and future directions. *Trends in environmental analytical chemistry*. 28:e00099 2020, ISSN 2214-1588. DOI: 10.1016/j.teac.2020.e00099.

Contribution 2: As part of this doctoral work, we have generated datasets of GC-EI-MS spectra of silyl derivatives, i.e., of TMS and TBDMS derivatives, to develop and evaluate ML-based CA approaches. We have developed a workflow for the generation of MSL-derived datasets of GC-EI-MS spectra that included three filtering steps. Further, we acquired GC-EI-MS spectral datasets in-house to test the ML approaches.

The publications describing this contribution, included in the thesis, are as follows:

Journal paper: Ljoncheva, M., Kosjek, T., Džeroski, S. GC-EI-MS datasets of trimethylsilyl (TMS) and tert-butyl dimethyl silyl (TBDMS) derivatives for development of machine learning-based compound identification approaches, *Data in Brief* (submitted, 4 July 2022)

Conference paper (not included): Ljoncheva, M., Heath, E., Džeroski, S., Kosjek, T. (2018) Generation of a test dataset for machine learning-assisted identification of contaminants of

emerging concern. In: Dežman, Miha (ed.), Proceedings. 10th Jožef Stefan International Postgraduate School Students' Conference and 12th Young Researchers' Day 10th and 11th May 2018, Piran, Slovenia. Ljubljana: Jožef Stefan International Postgraduate School: Jožef Stefan Institute, 21. http://ipssc.mps.si/Proceedings/Proceedings_2018.pdf.

Contribution 3: For the first time MSL-derived and in-house de novo acquired GC-EI-MS spectral datasets are used to investigate the application of the Compound Structure Identification (CSI):IOKR approach to the annotation of CEC silyl derivatives as an alternative to an exhaustive search of MSL, independent of instrumental platform and data processing software. We further investigated the dependence of CSI:IOKR performance on

several factors, including filtering of the training dataset, the overlap between compounds in the training and the test datasets, and the post-acquisition processing of the test dataset.

The publication describing this contribution is the Journal paper: Ljoncheva, M., Kosjek, T., Džeroski, S. Machine learning for identification of silylated derivatives from mass spectra, *Journal of Cheminformatics* (accepted for publication, 31 July 2022)

Contribution 4: An investigation of the factors potentially negatively influencing the use of GC-MS analytical platforms in EEA is provided. It was discovered that non-optimized derivatization conditions and lack of knowledge regarding the stability of the generated TMS derivatives might lead to low accuracy, repeatability, and reliability of results. This work is essential for the reliable identification of chemically and structurally diverse CECs in an aqueous sample without sufficient knowledge of the origin and content. As part of this work, we optimized the derivatization protocol of 70 structurally diverse CECs. Together with custom-developed multi-residual GC-MS methods, the optimized derivatization protocols were used to investigate the stability of the 70 CEC silyl derivatives under the most common storage conditions for up to 20 weeks. The derivatization protocols and GC-MS methods can further investigate CEC's environmental occurrence, fate, and behavior in different environmental aqueous compartments.

The publications describing this contribution, included in the thesis, are as follows:

Journal paper: Ljoncheva, M., Heath, E., Heath D., Džeroski, S., Kosjek, T., Contaminants of emerging concern: silylating procedures, evaluation of the stability of silyl derivatives and associated measurement uncertainty, Environmental Research (submitted, 25 August 2022)

Conference paper (not included): Ljoncheva., M., Heath, E., Džeroski, S., Kosjek, T. (2020) GC-MS analysis of contaminants of emerging concern. In: Jovičević Klug, Patricia (ed.), et al. Book of abstracts. 12th Jožef Stefan International Postgraduate School Students' Conference and 14th Young Researchers' Day, 15th May 2020. Ljubljana: Jožef Stefan International Postgraduate School: Jožef Stefan Institute, 22. http://ipssc.mps.si/BookOfAbstracts.pdf.

1.5 Structure of the Thesis

The remainder of the thesis is structured as follows. Chapter 2 presents the definitions and state-of-art related to CECs, GC-MS and LC-MS analytical techniques, and accompanying derivatization reactions. Chapter 3 presents a review of the state-of-the-art cheminformatics approaches for CA, while Chapters 4 and 5 present the main body of our experimental work, i.e., the datasets of GC-EI-MS generated and their use for CA and the study of the stability of silylated derivatives of CECs.

Chapter 2 presents the necessary background knowledge of CEC as EE constituents, GC-MS and LC-MS analytical techniques, and their use in the task of EEA. Here the concept of CECs as EE constituents is presented and describe EEA's various methodological and instrumental approaches. Further, the thesis focuses on the available and most commonly used GC-MS and LC-MS analytical platforms and describes the concept, benefits, and

bottlenecks of the employment of derivatization prior to GC-MS analysis, with a particular focus on the silylation of semi-volatile organic compounds.

Insights into the task of compound identification based on structural information inherent to MS data are presented in Chapter 3, accompanied by a thorough survey of the existing cheminformatics approaches for this task and a comparison of their performance in terms of accuracy and confidence. This chapter is divided into two sections. First, the task of surveying cheminformatics approaches to CA is introduced, and in the second part, the paper published in the journal *Trends in Environmental Analytical Chemistry* contains our survey.

In Chapter 4, we first present a framework for generation of training and test datasets of GC-EI-MS spectra for the purpose of identification of silyl derivatives of CECs using GC-EI-MS data and ML approaches. This framework includes an efficient filtering approach in MSL-based GC-EI-MS spectral dataset generation that, in three consecutive steps, eliminates GC-EI-MS spectra of **(1)** chemically irrelevant compounds; **(2)** compounds with molecular mass that is over the dynamic linear range (50-1000 *m/z*) of the mass analyzer employed and **(3)** GC-EI-MS spectra with low quality. Second, the CSI:IOKR approach is applied to the task of identification of silyl derivatives of CECs using GC-EI-MS spectral data. We show that the three-step-filtering approach improves the quality of the training data and, thus, identification accuracy. Finally, MSL-curated and experimentally de novo acquired GC-EI-MS datasets are presented as benchmark GC-EI-MS datasets for performance evaluation of existing and novel ML-based compound identification approaches. The chapter is divided into two sections. In the first section, the problem to be addressed is introduced, and in the second, the paper published in the *Journal of Cheminformatics* and the manuscript submitted to *Data in Brief* is presented.

In Chapter 5, we present a multi-residual method for derivatization and GC-MS analysis of the stability of a representative selection of 70 CECs with significant diversity in chemical structure and physicochemical properties. First, chemometrics methods are employed to select the optimal derivatization conditions (temperature and time) by discovering the similarities in compound behavior. Second, we develop and validate a multi-residual analytical method for investigating the stability of 70 CEC-TMS derivatives in a solvent, and artificial wastewater (AWW). The AWW was obtained from a pilot-scale wastewater treatment plant, and the CECs were extracted using solid phase extraction (SPE). Both were analyzed using GC-MS. Finally, the analytical method was used to study the CEC-TMS stability under the following conditions: **(1)** at room temperature (25°C) for one week in solvent and AWW extracts, **(2)** in the refrigerator (4°C), and **(3)** in the freezer (-18°C) for 20 weeks in solvent and four weeks in AWW extract and **(4)** over five freezing and thawing cycles in both matrices. Finally, an estimate of the associated measurement uncertainty (MU) was obtained. This chapter is also divided into two sections: problem description and the manuscript submitted to the journal *Environmental Research*, which addresses the described problem.

We conclude this dissertation in Chapter 6. We first summarize the scientific contributions of the performed work and we then discuss the hypotheses addressed by our research and how they were confirmed. Finally, we describe potential directions for extending the presented work in future research.

Chapter 2

State of the Art

This chapter provides an overview of the state-of-art work related to the research topics this thesis covers: the definition and annotation of CEC as EE constituents, GC-MS analytical platforms, and derivatization prior to GC-MS analysis focusing on silylation. The state of art in cheminformatics-assisted annotation of compounds is presented in Chapter 3.

2.1 Contaminants of Emerging Concern

2.1.1 Definition

CECs are defined as naturally occurring or manmade compounds, which have been recently discovered or are suspected to be present in various environmental compartments that could risk human health [17]. There remains a gap in our knowledge regarding their environmental behavior and toxic effects. Due to their potential toxicity or environmental persistence, they are of potential concern for human health and the environment. Additionally, an already regulated, presumed well-known environmental contaminant, such as parabens and estrogen hormones, can regain “emerging” status as new scientific information becomes available and thus, force regulatory agencies to re-evaluate their norms and guidelines [17]. This broad classification of CEC includes an ever-increasing number of CEC with a broad spectrum of structural, physico-chemical, and toxicological properties, such as pesticides, flame-retardants, surfactants, pharmaceuticals, personal care products, fragrances, plasticizers, algal toxins, cyanotoxins, including their metabolites and environmental TPs, to name few. CECs are yet to be adequately included in environmental monitoring programs since they have been only recently identified, or their toxic effects or environmental impact are not yet well understood.

2.2 Annotation of Contaminants of Emerging Concern

2.2.1 Workflow strategies

Various methodological and instrumental approaches are required to cover CEC's broad and dynamic range, exposure patterns, and biological responses. The following sections discuss the most common methods and instrumental approaches used in EEA.

2.2.1.1 Methodological approaches

In order to keep pace with the wealth of complex samples in EEA, several methodological approaches are developed. The general representation of an exposomics study workflow is

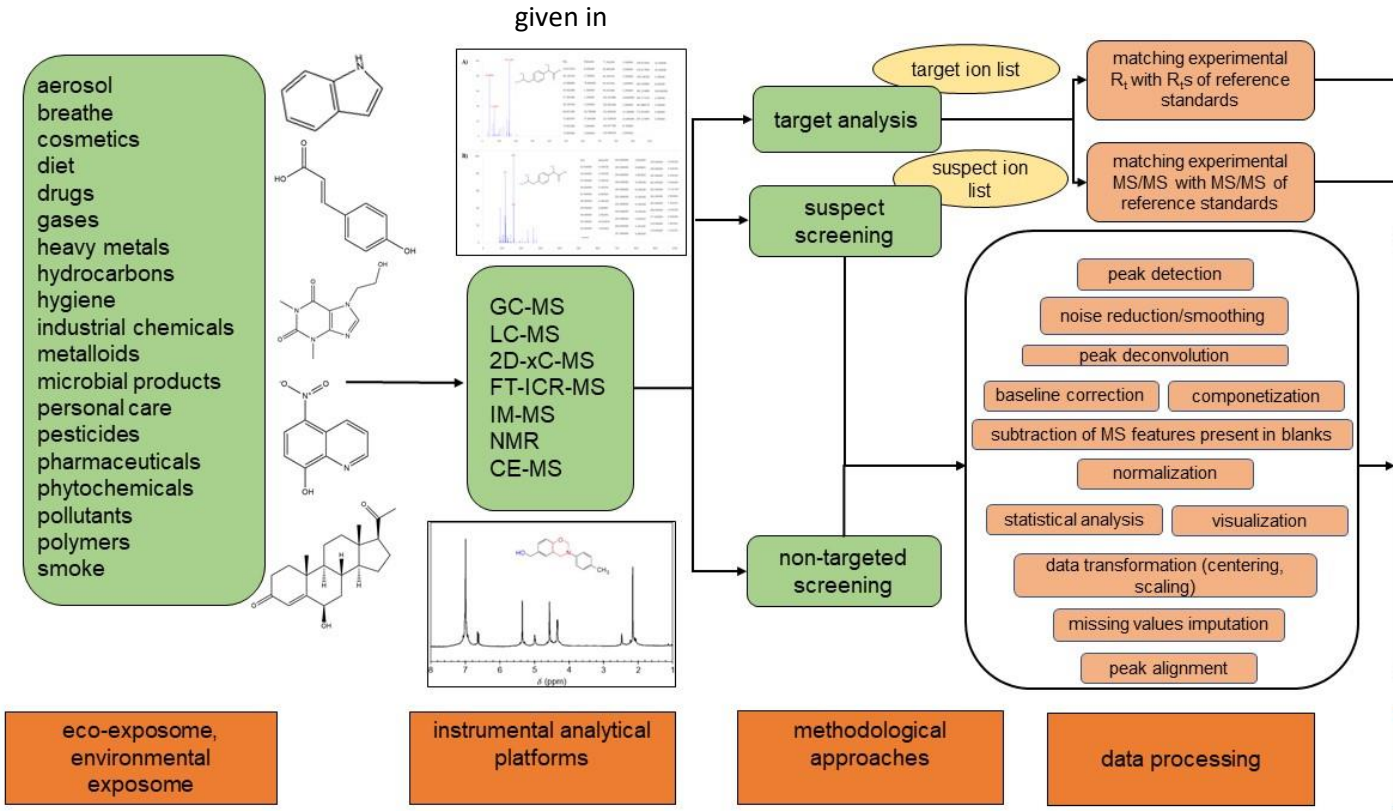


Figure 2.1. All of the compounds constituting the EE are analyzed using one or more of the analytical platforms presented, including GC-MS, LC-MS, ion mobility-mass spectrometry (IM-MS), nuclear magnetic resonance (NMR), and capillary electrophoresis-MS (CE-MS), using either *targeted analysis*, *suspect screening (SS)* or *non-targeted screening (NTS).*

The most straightforward exposomics methodological approach is to investigate the chemical and environmental exposure by monitoring the profile of the most prominent classes of CEC. Such *targeted approaches* are limited to measuring individual compounds or compound classes with a high possibility of missing compounds not on the target list but present in the analyzed sample [18].

Increasing the knowledge of the complex roles of environmental exposures on human health has led to adopting this approach and searching for the presence of compound classes or individual compounds that are expected or even known *a priori*. Such an approach is named SS. Here, a library of compounds suspected to be present is generated by adding information about their exact mass, experimental or predicted retention time (R_t), and isotope pattern in order to decrease the rate of false positives and limit the number of putative annotations.

Comprehensive *NTS* arises when potential EE constituents to be identified are not limited in their number and origin [19]. NTS are qualitative analytical experiments that identify three compound classes' presence. The first class, i.e., “known knowns”, consists of compounds, with known chemical structure, properties, and uses; the second is a class of compounds for which limited information is available in terms of exposure and toxicity. This class is referred to as “known unknowns”. The third compound class consists of compounds whose existence is unknown by the analyst and is commonly referred to as “unknown unknowns”. A typical NTS workflow proceeds in a few consecutive steps. Following sample treatment intended for separating analytes from the matrix and foreign substances, data acquisition is performed, followed by preprocessing to reduce data quantity and complexity. Preprocessing is usually performed through peak detection, accompanied by baseline correction and noise reduction/smoothing, peak deconvolution, annotation or subtraction of MS features present in blanks, componentization via a grouping of isotopes, adducts, multi-charged ions, and in-source fragments of the same compound, usually based on peak shape, intensity and isotopic

correlation and peak alignment across samples. Additional processing with missing values imputation, signal normalization, centering, scaling, and transformation can be performed. Statistical methods are commonly employed for data processing and analysis steps, such as principal component analysis, partial least squares-discriminant analysis, and orthogonal partial least squares [20]. In many cases, ML approaches, such as artificial neural networks, are employed to identify regions of interest in LC-MS data and determine the likelihood of peaks belonging to an authentic compound. Random forest is also employed for data normalization based on quality control samples and missing values imputation [21], [22].

Specific peaks and peak profiles of interest are prioritized for further evaluation from all detected MS features characterized by *m/z*, intensity, and R_t. Here, *data-driven approaches* are used, which evaluate the signal intensity, occurrence frequency in a dataset, and characteristic isotopic pattern of MS features.

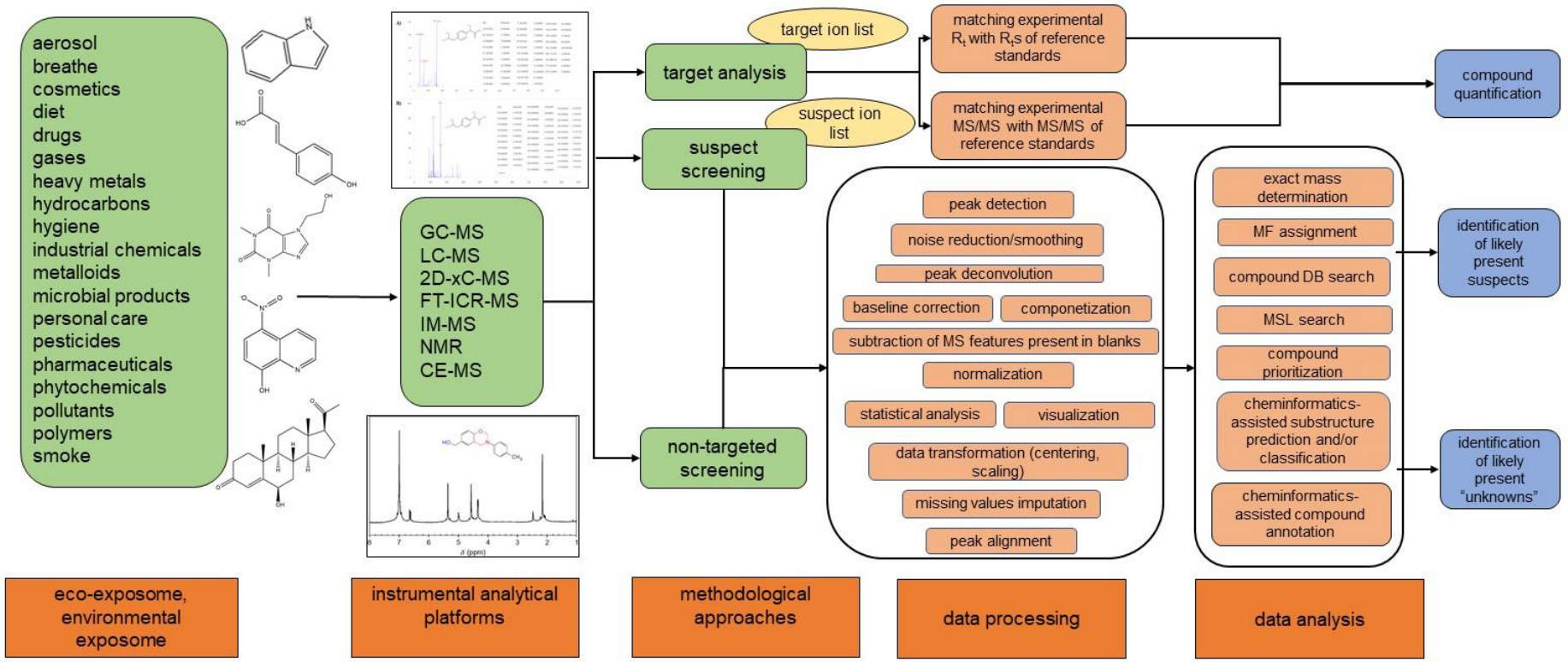


Figure 2.1: General representation of the methodological approaches for annotation of CEC.

They are further prioritized based on functional groups and mass differences, indicating belonging to a homologous series. Alternatively, *experiment-driven approaches* prioritize MS features based on persistence, elimination/formation over process, a reaction-based search of TPs to link masses before and after treatment, and e.g., biological, electrochemical, or oxidative TPs formation. Alternatively, effect-directed prioritization of chromatographic fractions with unknown compounds associated with specific toxic effects is performed [23].

In the next step, the monoisotopic or neutral molecular mass of the prioritized MS features is determined, and in the final step, one or more of the CA tasks are performed: **(1)** MF assignment; **(2)** exact mass or MF-based candidate search in compound DB, followed by candidate prioritization; **(3)** MSL search, followed by candidate prioritization and **(4)** cheminformatics-assisted CA. Analytical equipment vendors already provide many steps in MS data processing. In addition, there is a wide range of cheminformatics tools performing one or more of the preprocessing and CA tasks 1-4. Such tools include Workflow4Metabolomics [10], XCMS [25], MetaboAnalyst [26], MZmine [27], enviMass [28] and Nontarget [29], to name few. Some are in-house software, while others are workflow management systems, only combining existing cheminformatics tools [20]–[22]. Recently, patRoon, an all-steps environmental NTS-specific cheminformatics workflow was developed to overcome the lack of specific functionality and optimizations required for environmental NTS [30].

2.2.1.2 Instrumental techniques

Depending on the specific type of exposure, the exposome can be measured through a wide array of techniques, including remote sensors, questionnaires, geographical information systems, and approaches using chemical technologies. focused on biomonitoring, environmental monitoring, and metabolome investigations. The chemical approaches measure the exposures directly or through early biomarkers using chemical technologies [18]. Such annotation of the EE constituents requires analytical platforms for comprehensive chemical surveillance screening that would provide sufficiently selective and sensitive compound-level data for trace level-small molecules in complex environmental and biological samples [31].

Chromatography-MS analytical platforms are so far the best-established analytical tools for large-scale EE investigations due to their superiority in sensitivity, specificity, and dynamic range [5], [17]–[19], [32]. GC-MS analytical platforms are the oldest and still one of environmental EEA's most powerful analytical techniques [23]. Significant recent innovations include GC x GC-MS, GC- tandem mass spectrometry (MS/MS), and GC-high resolution-accurate mass (HR-AM)/MS, which allow the fulfillment of legal requirements (e.g., detection of organic contaminants in trace concentration ranges (pg/mL) in different matrices) and make data useful in understanding current environmental challenges [23], [31]. As a stand-alone platform, LC-(HR/AM-)MS analytical techniques are a common choice for EEA due to improved mass accuracy and resolving power. They can also serve as an orthogonal analytical technique to GC-MS for unambiguous CA, especially for compounds that cannot be observed using GC-MS due to insufficient ionization or incomplete ionization separation [23]. Today, GC and LC coupled to MS/MS are the analytical platforms of choice for quantitative trace-level target analyses with high sensitivity and broad linear dynamic range. However, they do not offer exact structural information for detecting and identifying “unknowns”. HR-AM/MS platforms provide high-quality mass resolution and high sensitivity in full scan mode, thus enabling the detection of low abundance compounds in complex environmental and biological samples. This ability, in turn, allows confident CA. However, such analytical techniques are expensive, complex and require thorough analytical knowledge.

Methods beyond GC/LC-MS are also used for EEA, offering important orthogonal information [33]. NMR is non-destructive, i.e., sample preserving structure elucidation technique. It has not been widely applied in EEA due to low sensitivity [23], [34]. IM-MS is an increasingly important analytical platform in EEA, offering effective multidimensional separation of ions by size, shape, charge, and thereof resulting gas ion mobility in an electric field in a neutral buffer gas. It provides the separation of enantiomers, chiral stereoisomers, diastereomers, and co-eluting matrix components [35]. Such ions with identical *m/z* are separated by the time the ion takes to traverse a gas-filled cell under the influence of a uniform electric field, named drift time that is directly converted to collision cross-section values, indicating the chemical structure and the 3D conformation of the

ions. Further-depth examination of the chemical exposome employs novel variations of sample preparation and chromatography-MS methods. These include ultra-fractionation, SPE/semi-preparative (ultra) high-performance liquid chromatography (U)HPLC-HRMS, nano/microflow LC-nanoESI, capillary electrophoresis (CE)-MS and ion-exchange chromatography-MS. Each of the aforementioned methods contribute different compound information (e.g. H-NMR:proton position, IM-MS: separation of complex mixtures, resolving ions that may be indistinguishable to MS alone etc.), altogether leading to ubiquitous identification of unknown compounds. However, their employment in extensive EEA investigations is yet to be thoroughly explored.

2.3 Chromatography-Mass Spectrometry in the Annotation of CECs

2.3.1 Chromatographic separation methods

Chromatography encompasses diverse and important methods for separating components in mixtures. In its basic concept, the sample is taken up in a mobile phase, which may be a gas, a liquid, or a supercritical fluid, and further interacts with a stationary phase fixed in a column or on a solid surface. The stationary and mobile phases are selected so that the sample analytes distribute through sorption/desorption process between them to varying degrees, e.g., according to their physicochemical properties. As the fresh mobile phase flows through the plate or column, it carries sample analytes, and a continuous series of interactions between the two phases occur. Analyte partitioning occurs until a state of equilibrium is achieved. The ratio between the molar analyte concentrations in the stationary phase (C_S) and the mobile phase (C_G) is constant. The constant is defined as the distribution constant or partition coefficient (K_C): $K_C = C_S/C_G = (n_S \cdot V_S)/(n_G \cdot V_G)$, where n_S and n_G are the amounts of the compound in stationary and mobile phases, while the V_S and V_G are the volumes of the stationary and mobile phases, respectively. Accordingly, analytes with a high K_C have a higher affinity to the stationary phase. It also means that they are strongly retained and move slower (elute later). Due to the differences in migration rates, sample analytes separate, resulting in different, specific elution times.

When a concentration-sensitive detector is placed at the end of the column, and its signal is plotted as a function of time, a series of peaks is obtained. This plot is termed a chromatogram. Here, peak positions on the time axis, named Rts, are used as one of the parameters for compound identification, whereas peak areas provide a quantitative measure of each compound.

The first, older classification divides chromatographic separations into planar chromatography, where the stationary phase is supported on a plate or paper, thin- layer chromatography (TLC), where compound separation occurs using a thin stationary phase supported by an inert backing, and column chromatography, where the stationary phase is packed in a narrow tube through which the mobile phase is forced. Based on the mobile and stationary phase types, three general categories of chromatographic methods are defined: GC, LC and supercritical fluid chromatography [36].

2.3.1.1 Gas chromatography

The basic operating principle of GC involves volatilizing the sample in a heated inlet of the gas chromatograph, from where it is flushed with an inert gaseous mobile phase, i.e., a carrier gas such as He, Ar, N_2 or H_2. Here, analyte separation occurs by partitioning between the carrier gas and a liquid stationary phase immobilized on the surface of an inert solid packing or the walls of capillary tubing. The separation occurs due to partitioning the analytes between the phases. A carrier gas then transports separate analytes to the detector.

Multiple GC instrumental innovations have appeared since its launch in the 1940s, differing in size, robustness, column type and instrument control, but all have the basic architecture (Figure 2.2). The carrier gas is available in pressurized tanks. A sample is introduced in the system, mostly using an autosampler into an inert inlet at high enough

temperatures to volatilize all analytes. Typically, 0.1-2.0 µL of the sample are injected through a septum into a heated sample port located at the head of the GC column. This process describes the traditional splitless mode, commonly used in trace analyses. Split injections are often used for the analysis of compounds present at higher levels, especially for thermolabile compounds, as it decreases the time spent in the hot injector. Typical split ranges are from 1:10 to 1:100 [37]. Other sample introduction techniques are used for specific sample types, such as the headspace technique, solid-phase microextraction, or direct on-column injection [37].

Once introduced, the gaseous sample is transferred to the head of the GC column, situated in a thermostatted oven. At the initial column temperature, usually 10-15°C below the solvent's boiling point, analytes and solvent condense in a narrow band in the stationary phase. Column temperature control possibilities include an isothermal run at a constant temperature, a temperature-programmed run, where the temperature is increased at a constant rate, and a multilevel run, where the temperature rate is increased at different rates at different times during the GC run. During the run, analytes partition between the carrier gas and the stationary phase until they reach an equilibrium state and elute from the column as sharp peaks. In the final step, analytes arrive at the detector, while the solvent is concentrated and evaporated with increasing column temperature [37], [38].

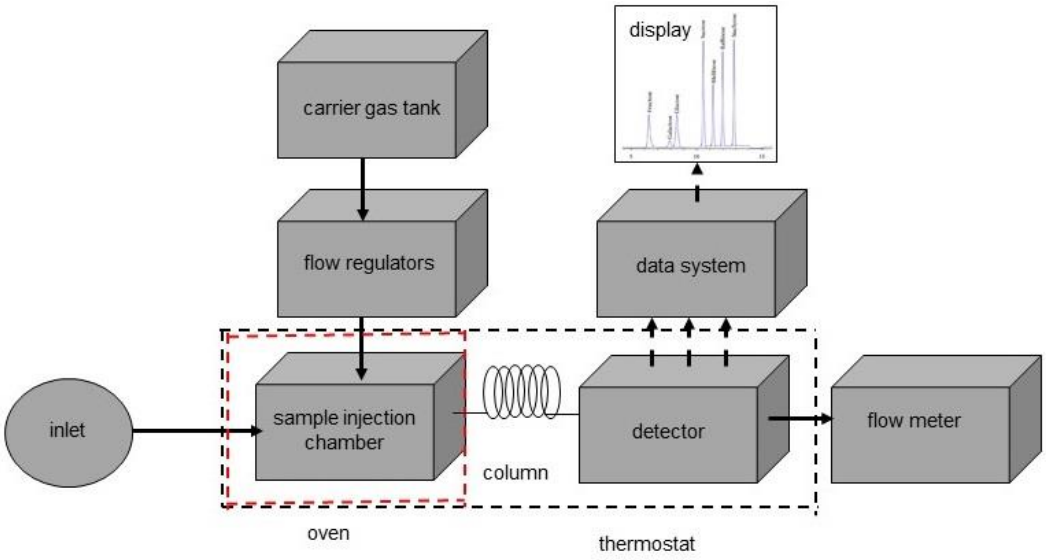


Figure 2.2: Scheme of a typical GC instrument. After introduction through the inlet, the sample is passed to the column oven, where through the sample injection chamber, is introduced to the column and then to the detector.The column oven is connected to carrier gas tank with flow regulators, while the detector is connected to flow meter and data system, which converts the detector signals in chromatograms. (Adapted from [38]).

GC is a suitable technique for separation of volatile and thermally stable analytes. Many other compounds, such as acids, amino acids, amines, saccharides, and steroids, require derivatization to increase their volatility and thermal stability (see Section 2.3.3.1.2) [37].

2.3.1.1.1 Gas chromatography columns

The type of GC columns almost exclusively used nowadays is the fused-silica wall-coated open tubular columns with sufficient physical strength, flexibility, and lower reactivity towards sample analytes. They are usually made of specially purified silica, with a minimum amount of metal oxides and polyamide coating. The choice of stationary phase is crucial to successful separation. It is based on the volatility (preferably low), thermal stability and chemical inertness, and the material's polarity, which must match the analytes' polarities for their successful resolution [38].

Nonpolar analytes, such as alkanes, steroids, and polychlorinated biphenyls, are best separated on 100% poly(dimethylsiloxane), poly(50% n-octyl/50% methyl siloxane) or poly(5% diphenyl/95% dimethyl siloxane) stationary phases. Aldehydes, ketones, and ethers are most efficiently separated on poly(diphenyl/dimethyl siloxane) phases, also on poly(50% n-octyl/50% methyl siloxane) and 100% poly(dimethylsiloxane) poly(cyanopropylphenyl/dimethylsiloxane due to matching polarity. Finally, the choice of stationary phases for polar compounds, such as amines, carboxylic acids, alcohols and diols is the widest, including 1,5-di(2,3-dimethylimidazolium)pentane bis(trifluoromethylsulfonyl)imide; 1,9-di(3-vinylimidazolium)nonane bis(trifluoromethylsulfonyl)imide and 1,12-di(tripropylphosphonium)dodecane bis(trifluoromethylsulfonyl)imide, to name few. Apart from the stationary phase, GC columns differ in length (which is usually 10-60 m), internal diameter (id between 0.10-0.53 mm), and stationary phase thickness (0.1-1.0 μm), strongly influencing column separation efficiency, which further depends on type and velocity of carrier gas and column temperature.

2.3.1.1.2 Gas chromatography detectors

Most commonly used GC detectors include flame ionization detector (FID), electron capture detector (ECD), and MS analyzers, along with less used thermal conductivity detector (TCD), photoionization detector (PID), and flame photometric detector (FPD). Generation of ions by pyrrolization (FID, TID, and FPD), ultraviolet (UV) radiation (PID), or capturing of electrons released by radioactive β-emitter (ECD) is followed by detection through recording of current changes (FID, ECD, TID, PID) in thermal conductivity (TCD) or as the light emitted in characteristic wavelengths (FPD). FID offers universal response to neaf all organic compounds, low limits of detection, high acquisition frequency, wide linearity range and reduced maintenance necessities. In environmental analysis, is it most commonly employed for analysis of small, volatile, non-polar compounds, such as pesticides, polyaromatic hydrocarbons and phthalaes. ECD offers significant response sensitivity for compounds containing halogens and nitro groups, peroxides, and quinines but lower sensitivity to amines, alcohols, and hydrocarbons. ECD is used for pesticide and insecticide analyses in environmental samples with high sensitivity but has a limited linear range. FPD is widely used for analysis of air and water pollutants, pesticides, and coal hydrogenation products [39]. Finally, MS detectors are nowadays most commonly coupled to GC [38], [40]. The architecture and types of mass spectrometers are further discussed in Section 2.3.2.

2.3.1.1.3 Gas chromatography-mass spectrometry analytical techniques

GC-MS is a ubiquitous analytical technique for the identification and quantification of small organic molecules (M_w< 300 Da) in complex matrices. It has been an indispensable analytical method in environmental analysis, forensics, exposomics, medical and biological research, flavor and fragrance industry, food safety, packaging, and many others over the last 40 years. It offers sensitive quantification down to pg/mL level using spectral data acquisition modes, such as selection ion monitoring (SIM). The extensive fragmentation provides unique structural information that, combined with R_t data, leads to unambiguous compound identification [37], [38].

Analysis of environmental matrices and biological fluids often exceeds the separation capacity of a single chromatographic column, producing overloaded chromatograms with co-elution of peaks, and decreased chromatographic resolution. Here, 2D gas chromatography (GC x GC) analytical techniques are successfully employed as a powerful and high-throughput tool for target analysis, SS, and NTS. GC-GC comprises two orthogonal separation mechanisms, combining the use of two GC columns with stationary phases that differ in polarity or chirality, connected with a modulator that transfers the small portions of the eluate from the first (^{1}D) to the second column (^{2}D), preserving the integrity of ^{1}D separation. Due to high data acquisition rates, GC x GC systems are usually hyphenated to TOF mass analyzers. Thus, GC x GC platforms allow shorter run times, lower limits of detection, enhanced resolution and peak capacity, and higher mass selectivity

and sensitivity, which come at higher costs for equipment operation and maintenance [31].

2.3.1.2 Liquid chromatography

In LC, separation occurs in a column with a stationary phase, on which samples are introduced through the mobile phase, consisting of one or more solvents. The analytes are separated by a selective distribution between the mobile and the stationary phase, leaving the column as detectable narrow bands. Here, distribution occurs as a result of one of the following mechanisms: partitioning between the mobile and stationary phase (liquid-liquid chromatography), adsorption (liquid-solid chromatography), ion-exchange, size-exclusion, selective separation of compounds covalently bonded to an affinity ligand and enantioselective separation (chiral chromatography) [41], [42].

The basic configuration of an HPLC instrument is shown in Figure 2.3. The mobile phase solvents are kept in glass reservoirs. The isocratic or gradient flow process introduces the single solvent or solvent mixture through a hydrodynamic pumping system. One µL to 1 mL samples are introduced through an automated sampling loop at controlled temperature and pressures up to 7000 psi to the head of the LC column (usually heated), where data acquisition begins. The system may also contain either a scavenger precolumn between the mobile phase reservoir and injector or a guard column between the injector and the analytical column to prevent contamination and prolong its lifetime. Typical HPLC instruments over time evolved to UHPLC, with decreased particle sizes (sub-2 µm particles) at ultrahigh pressures (up to 20,000 psi), yielding fast analyses and narrow chromatographic peaks [33], [41], [42].

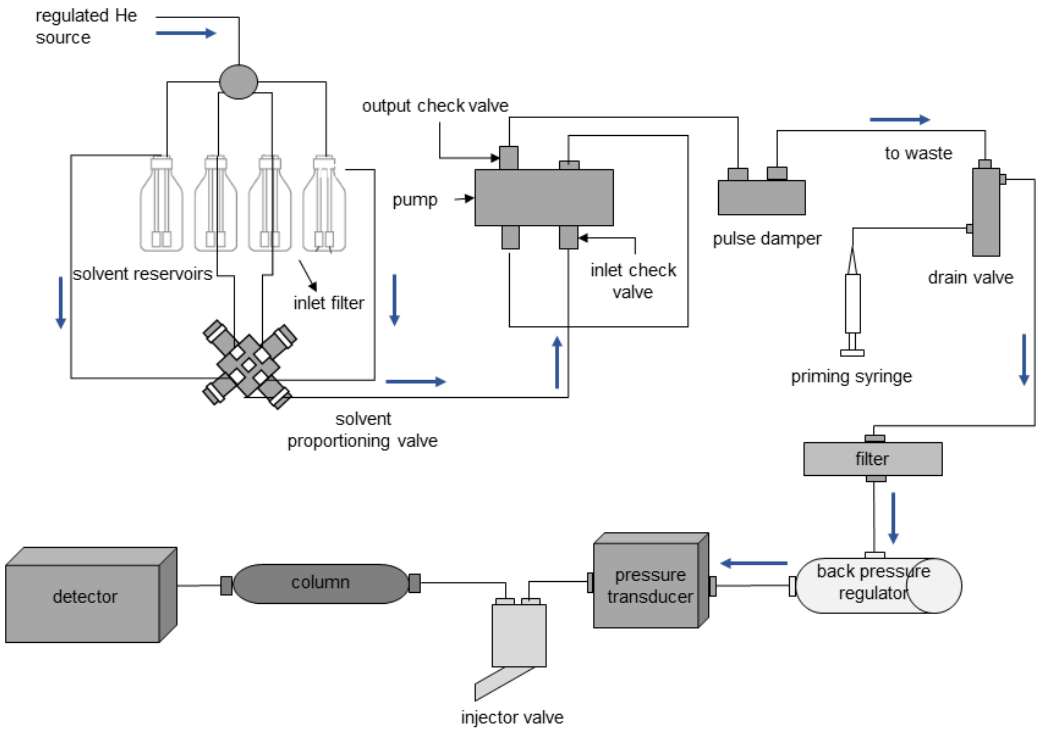


Figure 2.3: Diagram of a typical HPLC instrument.

· 2.3.1.2.1 Liquid chromatography columns

HPLC columns have porous particle packing, with microparticles (d=3-10 µm, <2 µm for UHPLC columns), composed of:

- porous oxides: native silica, alumina, titania, zirconia, porous (graphitized) carbon, hydroxyapatite;
- cross-linked organic polymers: polyacrylamides, polystyrene divinylbenzene, polymethacrylates, polysaccharides, polyvinylalcohols; and
- silica-organo hybrids: polyacrylamide-polystyrene-coated silica composites, silica-organic composites.

Normal-phase separation occurs on polar coatings, such as silica, diol, cyanopropyl-, nitro-, aminopropyl- and dimethylamino- silica stationary phases. Reverse-phase separation is used in ¾ of the cases, on nonpolar stationary phases, such as *n*-octadecyl- (C_{18}), *n*-octyl- (C_8), *n*-butyl- (C_4), *n*-methyl- (C_1) and *n*-triacontyl- (C_{30}) silica, phenyl, phenyl-hexyl or pentafluorophenyl. Enantiomers are separated on chiral stationary phases, such as polysaccharides, human serum albumin, and cyclodextrins. Ion exchange separation of easily ionizable analytes occurs on anion or cation exchange resins. In contrast, size-exclusion, typically of high molecular species, occurs by trapping in the pores of the adsorbent silica- or polymer-based (e.g., sulfonated divinylbenzenes or poly(acrylamides) stationary phases.

2.3.1.2.2 Liquid chromatography detectors

LC detectors are designed with flow cells to measure the analyte concentrations in liquid streams. Most commonly used are:
(1) UV-visible absorption detectors, able to isolate single wavelengths (fixed-wavelength design) or scan over a defined wavelength range (variable wavelength design);
(2) infrared absorption detectors, seldom used due to poor sensitivity;
(3) fluorescence detectors, offering the highest selectivity and sensitivity;
(4) refractive index detectors that are general-purpose detectors, analogous to FID or TCD for GC, but with inferior sensitivity;
(5) evaporation light scattering detector, a newer type, detecting analytes by scattering of light from a laser excitation source;
(6) electrochemical detectors, performing selective reduction or oxidation of analytes and measurement of electrochemically generated current; they are based either on amperometry, voltammetry, coulometry, or conductometry and
(7) MS detectors (further discussed in Section 2.3.2); today, LC-MS and LC-MS/MS are recognized as the ideal merger of separation and detection [41].

2.3.2 Mass spectrometers

2.3.2.1 Basic principles of mass spectrometers

A mass spectrometer is an instrument that generates ions and separates them according to their *m/z* ratios. The principal components of a mass spectrometer are shown in Figure 2.4. The sample is introduced through the inlet and travels first to the ion source. Here, ionization of neutral molecules occurs through electron ejection, electron capture, protonation, deprotonation, adduct formation, or transfer of charged species from a condensed phase to the gas phase. The output of the ion source is a beam of charged ions that are accelerated into the mass analyzer for separation based on their *m/z*. Separated ions are headed to the ion transducer that converts the beam of ions into an electrical signal that can then be processed through the signal processor, stored in a computer, and displayed. All the mass spectrometers' components, except the signal processor and readout (and, in some cases, the inlet), are in a low-pressure vacuum system (10^{-2} to 10^{-7} Pa). Such a system ensures that infrequent collision in the mass spectrometer occurs and produces and maintains free ions and electrons [43].

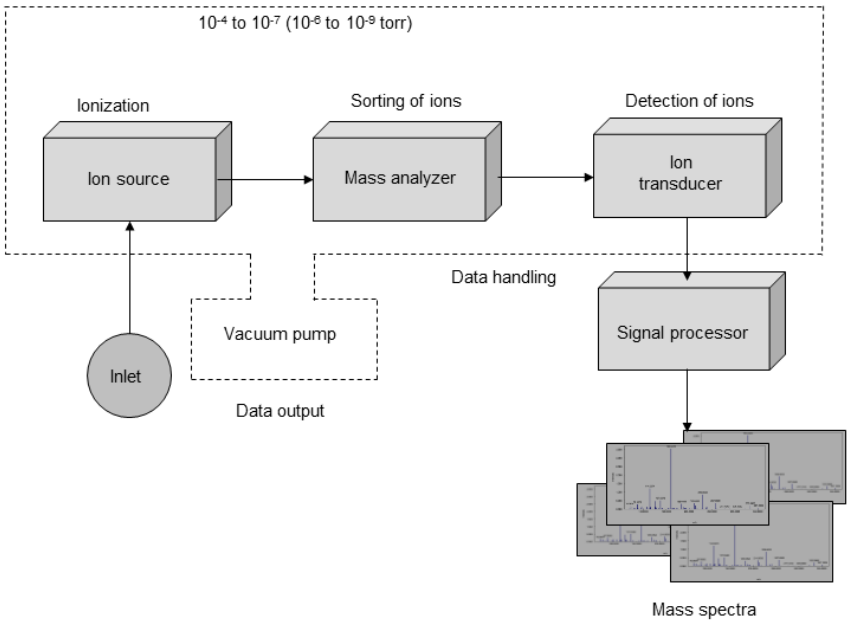


Figure 2.4: Scheme of a mass spectrometer (Adapted from [43]).

2.3.2.2 Types of ion sources and ionization techniques

The most commonly used types of ion sources are [44]:

- Gas-phase ion sources, in which the sample is first vaporized and then ionized by bombardment with electrons, photons, ions, or molecules. EI and chemical ionization (CI) ion sources are only suitable for gas-phase ionization. Such sources are restricted to ionizing thermally stable compounds with boiling points less than 500°C, thus to compounds with molecular masses of less than 10^3 Da.
- Liquid-phase ion sources, where the liquid sample is nebulized into droplets, from which ions are generated at atmospheric pressure; such are the electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI), atmospheric pressure photoionization (APPI) ion sources, applicable to analytes with M_w as large as 10^5 Da.
- Solid-state ion sources, in which analytes are in a solid or viscous fluid, which is then irradiated by energetic particles or photons that desorb ions near the matrix surface, which are extracted by an electric field and focused towards the mass analyzer; such are matrix-assisted laser desorption/ionization (MALDI), plasma desorption and field desorption ion sources.

The ionization techniques are classified as soft and hard ionization techniques. Both are useful in EEA and are frequently combined for orthogonal structural information that would aid compound identification and quantification [43]. A typical hard ionization technique is EI, which involves the interaction of a low-pressure (~10^{-1} Pa) gas cloud with electrons accelerated through a 70eV electric field. This gives enough internal energy to analyte molecules to leave them in highly excited vibrational and rotational energy states, resulting in extensive fragmentation of the molecular ions ($M^{+\cdot}$) while aiming to achieve a lower energy state. This multitude of fragments and their abundances produce a characteristic EI-MS spectrum, referred to as the "EI fingerprint" of the compound [37]. The EI-MS spectrum can provide important information for structural elucidation of unknown compounds, except if fragmentation is so complete that the molecular ion is not detected.

Soft ionization techniques including ESI, CI, APCI, APPI, and MALDI are widely applicable to thermally labile, ionic, high M_W compounds. They impart little residual energy to the analytes, insufficient for extensive fragmentation generating structurally specific ions. They also offer valuable qualitative and quantitative information combined with chromatographic separation.

ESI is the first-choice ionization technique for analysis of small organic compounds and biomolecules (proteins, polypeptides), usually when LC is used for compound separation and thus, the most employed technique in 'omics' sciences. It is based on sample spraying by strong electrical fields, leading to the formation of charged droplets that move towards the oppositely charged electrode, with simultaneous size decrease due to solvent evaporation and break up due to electrostatic repulsion. Finally, ionic species of analytes are desorbed into the gas phase [45].

In CI, that together with EI are most commonly employed ionization techniques when compound separation is performed with GC, ions are generated by the collision of gas analyte molecules with electrons accelerated through a 70eV electric field. The reagent gas, most frequently methane, is ionized by electron impact, and its ionization products (CH_5^+, $C_2H_5^+$ and $C_3H_5^+$) are strong proton donors and generate excited $AH^{\cdot+}$ ions that fragment further [46]. APCI is considered the natural evolution of the CI sources operating under reduced-pressure conditions. Here, low-energy electrons emitted by a radioactive beta source or corona discharge ionize a reagent gas (N_2, O_2, H_2O), which subsequently ionizes the analyte using frequent collisions and several complex ion-molecule reactions [41]. For instance, APPI uses an intense UV light source to ionize the analyte, directly or through a dopant gas, at atmospheric pressure [43]. APCI and APPI are mainly used to analyze medium to low polarity organic compounds and synthetic polymers. The compound range of APPI extends more to less polar species. Simultaneous analysis with ESI is used to detect polar and non-polar analytes in high-throughput screening. Finally, MALDI is most selective to proteins, glycoproteins, and oligonucleotides (M_w>1000 Da). Thus it is seldom used in small organic molecules analysis [43]. It involves the co-crystallization of an analyte and a matrix and further irradiation by a UV or an infrared laser, with simultaneous analyte vaporization and ionization by protonation, deprotonation, or cationization [47].

2.3.2.3 Types of mass analyzers

There are many different designs of mass analyzers available. The choice of the mass analyzer depends on several factors that include, but are not limited to: **(1)** physicochemical properties of the compounds to be analyzed (e.g., chemical structure, M_w, volatility, and polarity); **(2)** the desired *m/z* range to be analyzed; **(3)** the required resolving power and **(4)** the required sensitivity.

According to their configuration, mass analyzers can be classified as single analyzers and tandem/hybrid arrangements, known as MS/MS systems. The most commonly used mass analyzers in environmental exposomics are given in **Error! Reference source not found.**, along with their resolving power at full width half maximum (FWHM), resolution ($\Delta m/z$), mass accuracy, and typical mass ranges.

Tandem mass spectrometry is the "golden standard" in quantitative analysis of complex environmental and biological samples [43]. The basic architecture of MS/MS spectrometers is shown in Figure 2.5. The sample is introduced through the inlet to the ion source, where compound ionization occurs to i.e. "original" or "parent" positively charged ions (ABC^+, $ABCD^+$, $ABCDE^+$). They are introduced to the first mass analyzer, in which ions are sorted and weighed, and the precursor ion ($ABCD^+$) is selected and further introduced to the collision cell. Here, precursor ions are fragmented to product ions (A^+, AB^+, ABC^+) by bombardment with an inert gas (Xe, Ar, etc.). Product ions continue to the second mass analyzer for further fragmentation, after which are detected in the ion transducer. Finally, signals are processed and modificed in the signal processor and represented as MS/MS spectra. The ion source, mass analyzers and ion transducer are in a vacuum system, maintaining pressure of 10^{-6} to 10^{-9} torr.

The process occurring in a tandem mass spectrometer is represented in three steps [48]:

1) generation of ions in the ion source of mass analyzer 1 and selection of the ionic species of interest, i.e., the precursor ion;

2) collision of the precursor ion with molecules or atoms in the interaction cell, either by: spontaneous decomposition; collision-induced dissociation (CID), i.e., ion fragmentation by collisions with chemically inert collision gas (typically argon or nitrogen) at low pressure (~10^3 torr); electron-capture dissociation, where precursor ions capture a low-energy electron to produce an intermediate that rapidly dissociates or photo-induced dissociation, i.e., fragmentation initiated by interaction with an intense laser beam to product ions;

3) mass analysis of product ions based on their *m/z* in mass analyzer 2, with further processing and storage of the digital signals.

The three steps can be separated in space, so each step occurs in different regions. This *tandem-in-space* approach is achievable by a triple quadrupole (QQQ), quadrupole-time-of-flight (Q-TOF), and TOF-TOF mass spectrometers. In the *tandem-in-time* approach, all three steps occur in the same spatial region but in separate time intervals, such that the process can be repeated n times (MS^n). This approach is used in a quadrupole ion trap and Fourier transform ion cyclotron resonance (FT-ICR) [41], [43].

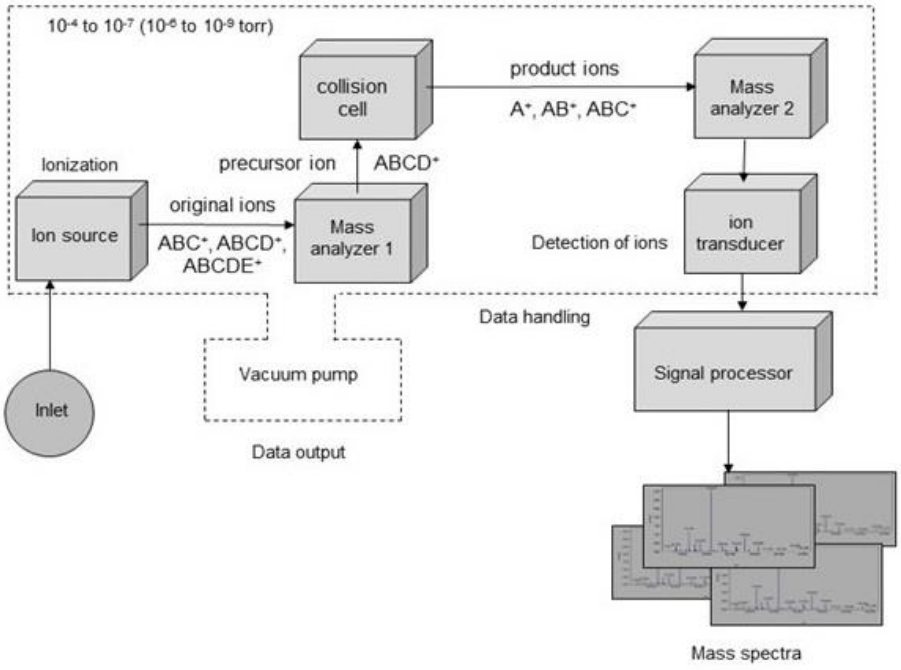


Figure 2.5: Diagram of a tandem mass spectrometer (Adapted from [43]).

Low-resolution (LR) mass analyzers include single quadrupole (Q), QQQ, and ion trap (IT) mass

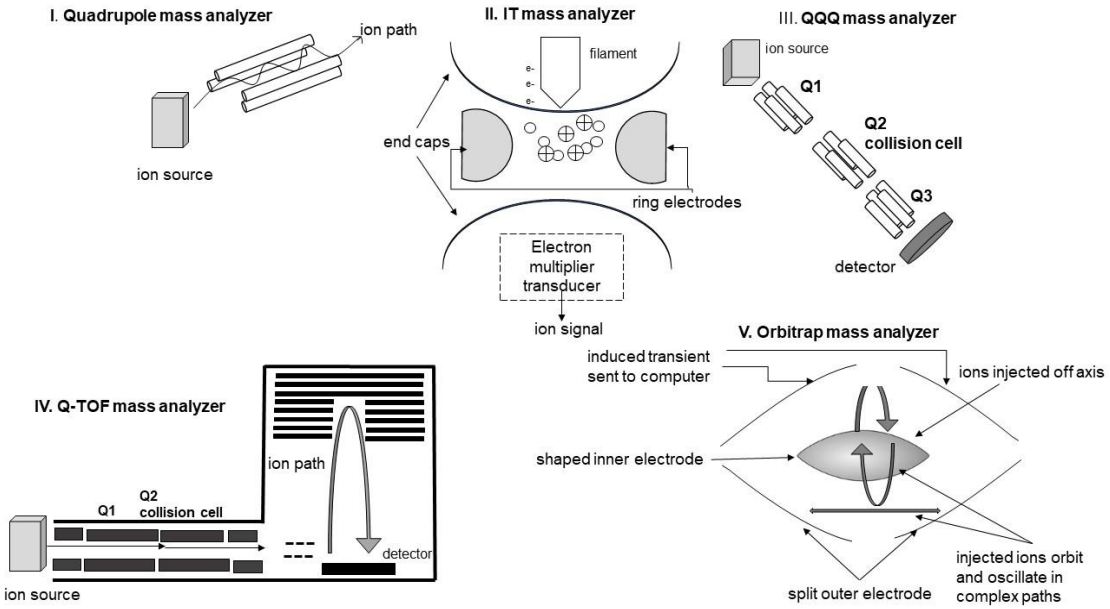


Figure 2.6). LR-MS mass analyzers typically have mass accuracy of ≤ 50 parts per million (ppm), but this is insufficient for determining elemental composition. High mass accuracy (1-5 ppm) is required to exclude candidates with complex elemental compositions (C, H, N, S, O, P, F, Cl, Br and Si) and derive an accurate list of elemental compositions for the measured monoisotopic mass, along with high resolving power [49]. They are compact, easy-to-use, rugged, reliable, benchtop configured instruments with reasonable cost, selectivity, and sensitivity that make the LR-MS the "workhorse" in targeted exposome research. Their low accuracy limits their employment in the identification of unknown compounds. The single quadrupole (

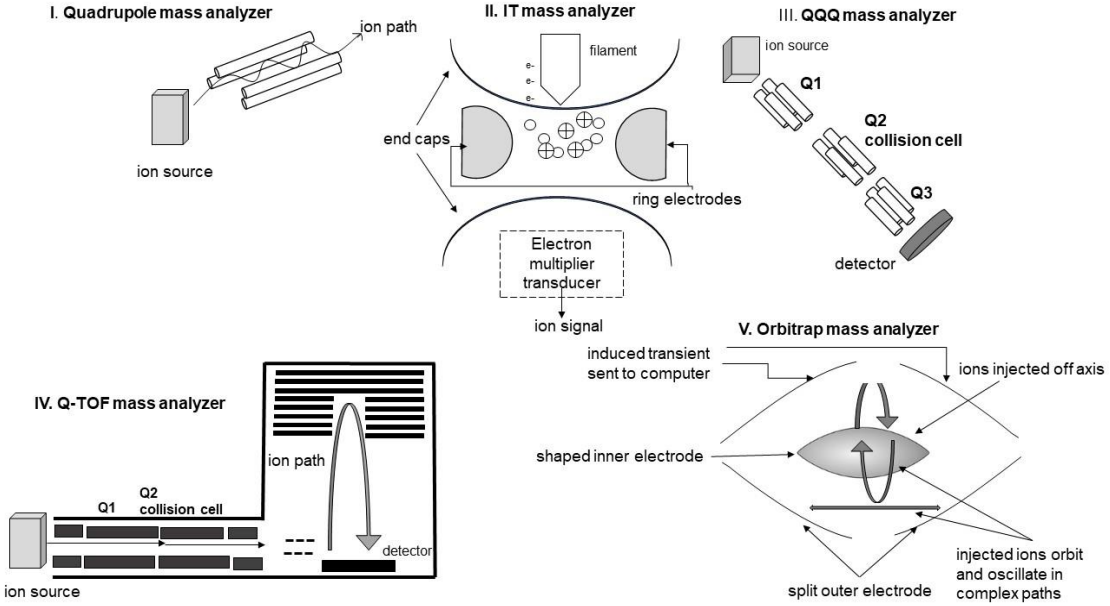


Figure 2.6) is the simplest one, acting as a mass filter, i.e., separating ions according to their *m/z* while radio frequency and direct current potentials are applied on the quadrupole's rods. It is a significantly rugged mass analyzer with minimum requirements for maintenance and excellent stability over long periods that easily interfaces both GC and LC instruments [50]. In IT

analyzers (

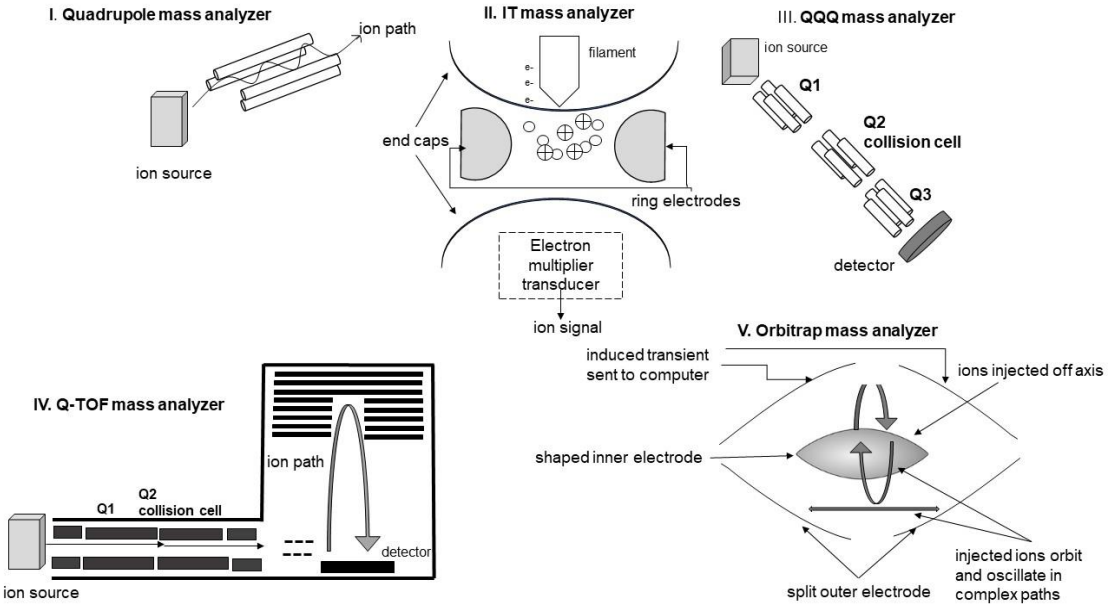


Figure 2.6), gaseous ions are formed by an EI or CI source and are confined, i.e., "trapped" for extended periods by electric and magnetic fields, and selectively ejected axially or radially [43], [50]. It provides excellent sensitivity for product ion measurements but at low resolving power. Also, it has a lower linear dynamic range and precision, with more matrix interferences and worse robustness when compared to QQQ [33].

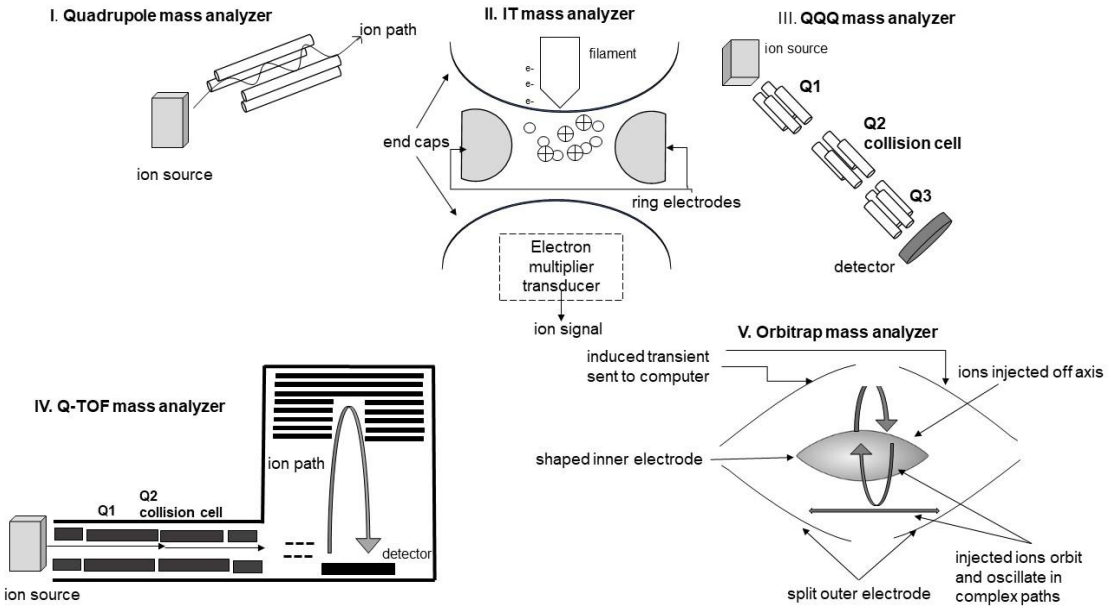


Figure 2.6: Basic architecture of the most commonly used mass analyzers. I. Single quadrupole, adapted from [50]; II. IT analyzer, adapted from [43]; III. QQQ analyzer, adapted from [50]; IV. Q-TOF analyzer, adapted from [50]; V. Orbitrap analyzer. Adapted from [43].

Table 2.1: Most commonly used mass analyzers in EEA and their properties: resolving power at FWHM, mass accuracy and mass range [22], [31].

Mass **analyzer**	**resolving power (FWHM defined at *m*/z)**	**resolution (Δ *m/z*)**	**mass accuracy (ppm)**	**mass range (Da)**
Q	~ 1,000 (m/z 200)	0.5 - 1.0	~ 50	50-1,000
QQQ	5,000 – 10,000 (7,500 at m/z 200)	0.07 - 1.0	~ 50	50-2,000
IT	~1,000 (m/z 508)	0.1 0.5	~50	50-4,000
TOF	2,500 – 22,500 (22,500 at m/z 956)	0.04 - 0.1	5-10	20-500,000
LIT-TOF	10,000 (m/z 1000)	0.1	3-5	50-5,000
Q-TOF	5,000 – 60,000 (maXis 4G, Bruker Daltonics, m/z 1222)	0.02 – 0.05	~ 2-10	50-2,000
Orbitrap	~100,000 (m/z 200)	0.002	< 2	50-4,000
Q-Orbitrap	~150,000 (m/z 200)	0.001	<1-5	50-4,000
LTQ-Orbitrap	>150,000 (m/z 400)	0.001	~ 2-5	50-4,000
LIT-Orbitrap	240,000 (m/z 400)	0.0002	<1-3	50-4,000
FT-ICR	200,000 – 1,000,000 (750,000 at m/z 400)	0.0005	< 1	50-10,000
Q-ICR (SolariX 15T, Bruker Daltonics)	2,500,000 (m/z 400)	0.0002	<0.25	100-10,000

The unit resolution of single quadrupoles and IT is slightly improved in the other LR-MS analyzers, but at the cost of lower ion transmission and, therefore, lower sensitivity [33]. In a QQQ, three quadrupole analyzers are used in sequence (Figure 2.6); the first analyzer (Q1) scans across a range of *m/z* values or selectively filters ions of a selected *m/z,* and Q2 acts as a collision cell to fragment the selected ions from Q1, while Q3 can scan all ions of a certain *m/z* or can be fixed to monitor a particular ion. QQQ can operate in several modes - product ion scan, precursor ion scan, neutral loss scan, selected reaction monitoring (SRM), or multiple reaction monitoring (MRM). In all modes, CID occurs in Q2, while the difference is in *m/z* at which Q1 and Q3 operate. In product ion scan mode, Q1 scans at fixed *m/z* and Q3 scans full *m/z* range. In precursor ion scan mode, Q1 scans in full *m/z* range, while Q3 at fixed *m/z*, while in neutral loss scan mode Q1 also scans in full *m/z* range, while Q3 scans in the range of Q1 minus the *m/z* of the neutral loss of interest (Q3=Q1-Δ*m/z*). Finally, in SRM and MRM, Q1 and Q3 scan both at fixed *m/z* value for the reaction(s) of interest [50]. Therefore, SRM and MRM modes offer high sensitivity, precision, accuracy, and good linear dynamic range, thus becoming the “golden standard” in targeted analysis.

HR/AM-MS analyzers offer high (<5 ppm) or very high (<1 ppm) mass accuracy, high resolving power - up to 1,000,000 at FWHM at defined *m/z* (Table 2.1), wide linear dynamic range, and high sensitivity in full scan mode. HR/AM-MS analyzers are most commonly used for reliable compound identification across a broad mass and

concentration range. They can employ different ionization techniques and data acquisition modes, e.g., data-dependent acquisition, data-independent acquisition, and all ion fragmentation (MS^{all}). Most commonly used HR-AM/MS analyzers include time-of-flight (TOF), Orbitrap, FT-ICR, and hybrid MS configurations, such as quadrupole-Orbitrap (Q-Orbitrap), linear trap-quadrupole (LTQ)-Orbitrap and linear IT (LIT)-Orbitrap, and TOF hybrids.

In its simplest form, the TOF analyzers consist of a flight tube and an acceleration grid that accelerates a "packet" of ions from the ionization source to the MS detector. Here, two ions of different *m/z* accelerated from the ion source with the same kinetic energy and are allowed to drift through a field-free region of the flight tube to arrive in the detector at different times. As the initial kinetic energy (E_k) of the ions and the length of the flight tube (d) are constant, mass is strictly a function of the time it takes for the ions to be detected after initial acceleration[43], [50]. TOF analyzers today also employ reflectron, which reflects the ion path in the direction of the ion source before being detected. This allows for corrections in the slight differences in initial kinetic energies of the ions that may occur during acceleration [51]. TOF analyzers are suitable for separating

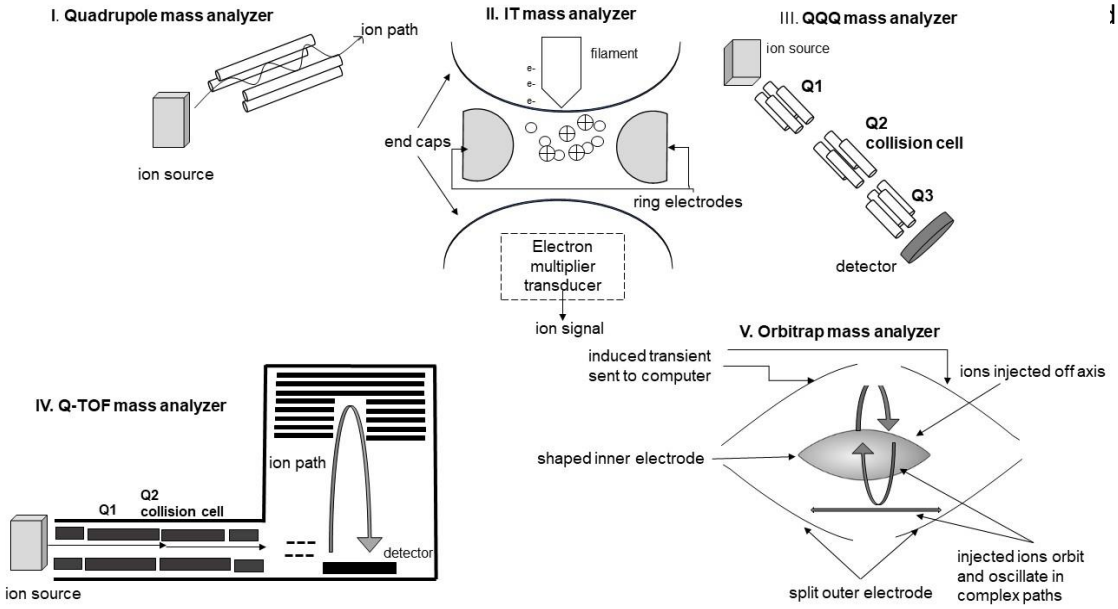


Figure 2.6) are regarded as QQQ in which the third quadrupole is replaced by an orthogonal TOF, such that the ions filtered through the quadrupole are injected orthogonally into the TOF

analyzer as a packet using a set of pusher and puller plates between the two analyzers (

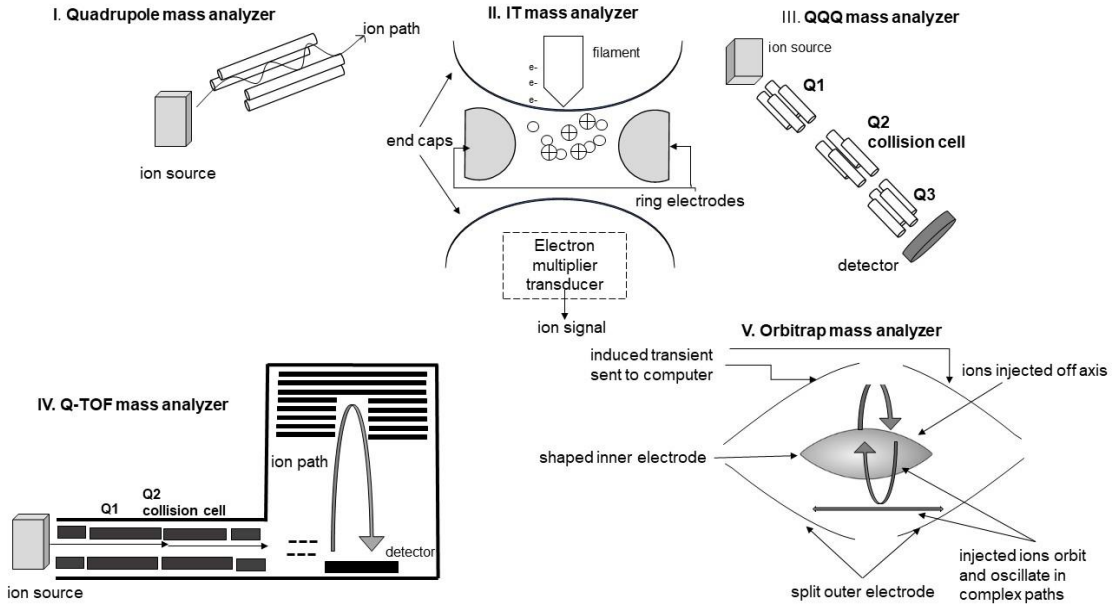


Figure 2.6) [52]. They are popular as they offer high accuracy measurements in both full scan and MS/MS modes. LIT-TOF analyzers perform isolation and dissociation of the parent ions in the LIT, followed by product ion analysis with an orthogonal TOF analyzer [53].

Using FT-ICR analyzers, M_W of analytes are determined by first exciting the ions with a limited frequency sweep of a broad-band radiofrequency field, placing them in a higher cyclotron orbit and allowing them to be detected by measuring their angular (cyclotron) frequency in the fixed magnetic field [50], [54]. FT-ICR analyzers have the highest resolving power and mass accuracy of all mass analyzers and thus perform best at compound identification but at a high cost and slow scan speeds. Similar to FT-ICR analyzers, Orbitrap analyzers use an electric field to induce axial oscillations of the ions around the inner electrode of the electrostatic trap that is proportional to the *m/z* ratio of the injected ions (

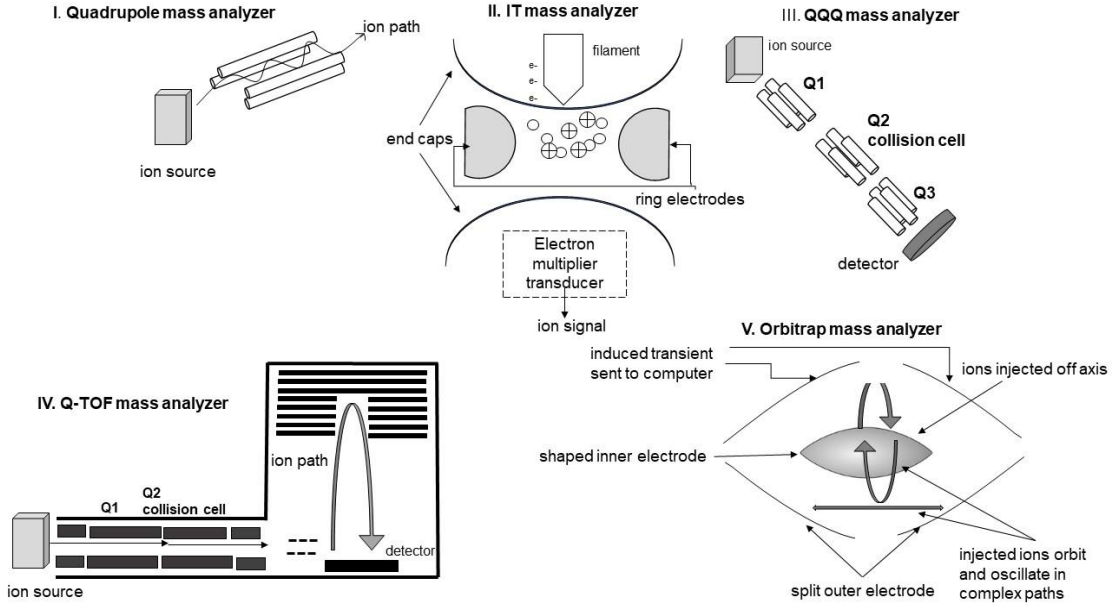


Figure 2.6). Finally, an image current is generated, recorded, and decoded from time to frequency domain by Fourier transform [43], [50], [55]. Due to the high resolving power,

Orbitraps are used as a replacement for FT-ICR analyzers and often have better resolving power at higher *m/z* [56]. The hybrid mass analyzers, which combine an Orbitrap with a quadrupole or a LIT, are Q-Orbitrap, LTQ-Orbitrap, and LIT-Orbitrap and offer higher resolving power and mass accuracy due to the mass filtering prior to orbital trapping.

2.3.2.4 Mass spectra

A mass spectrum can be considered a digitalized ion-current signal that underwent considerable processing before display, normalization, and *m/z* peak assignments. The coordinates for each MS peak are the *m/z* value and the abundance of that ion, represented by a vertical line from their position on a two-dimensional Cartesian coordinate system to the x-axis [37]. The highest peak in the MS spectrum termed the base peak, is often assigned the value of 100, and the intensities of all other peaks are normalized as a percentage of the base-peak height.

The appearance of an MS spectrum for a given compound strongly depends on the

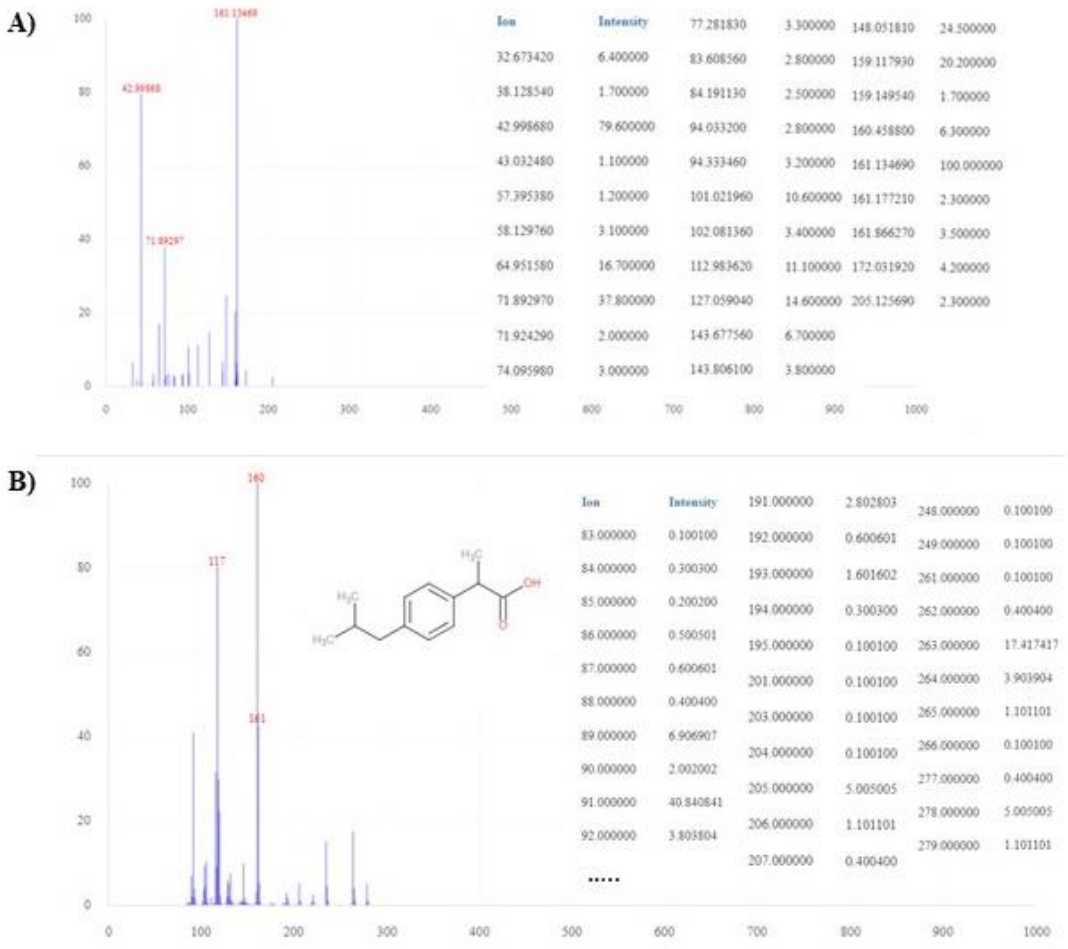


Figure 2.7 illustrates two MS spectra of ibuprofen, with an exact mass of 206.131 Daltons (Da). In order to obtain the first (A) spectrum, ibuprofen was bombarded with an electron stream that led to the formation of the ion $C_{13}H1_7O_2$-, i.e., its deprotonated form [M-H]-, with an exact mass of 205.123 Da. To relax from the excited state after the collision with energetic electrons, the molecular ions further fragment to produce ions of lower masses. For example, because of the loss of the CH_2O group, the major product $C_{12}H1_7$· is

formed, along with other smaller negatively charged fragments. The second spectrum (B) was obtained on GC-EI-TOF by bombarding the molecules with a beam of energetic electrons, resulting in extensive fragmentation. The base peak corresponds to a fragment ion, $C_{12}H_{17}^{+}$ with *m/z* 161; also, other smaller fragments are formed, such as $C_{11}H_{14}^{\cdot+}$ with *m/z* 160, $C_9H_9^{\cdot}$ (*m/z* 117), and $C_7H_7^{+}$ (*m/z* 91), but the molecular ion is absent due to the extensive fragmentation. The ions then exit through the slit of the mass spectrometer, sorted according to their *m/z* ratio, and displayed as an MS spectrum.

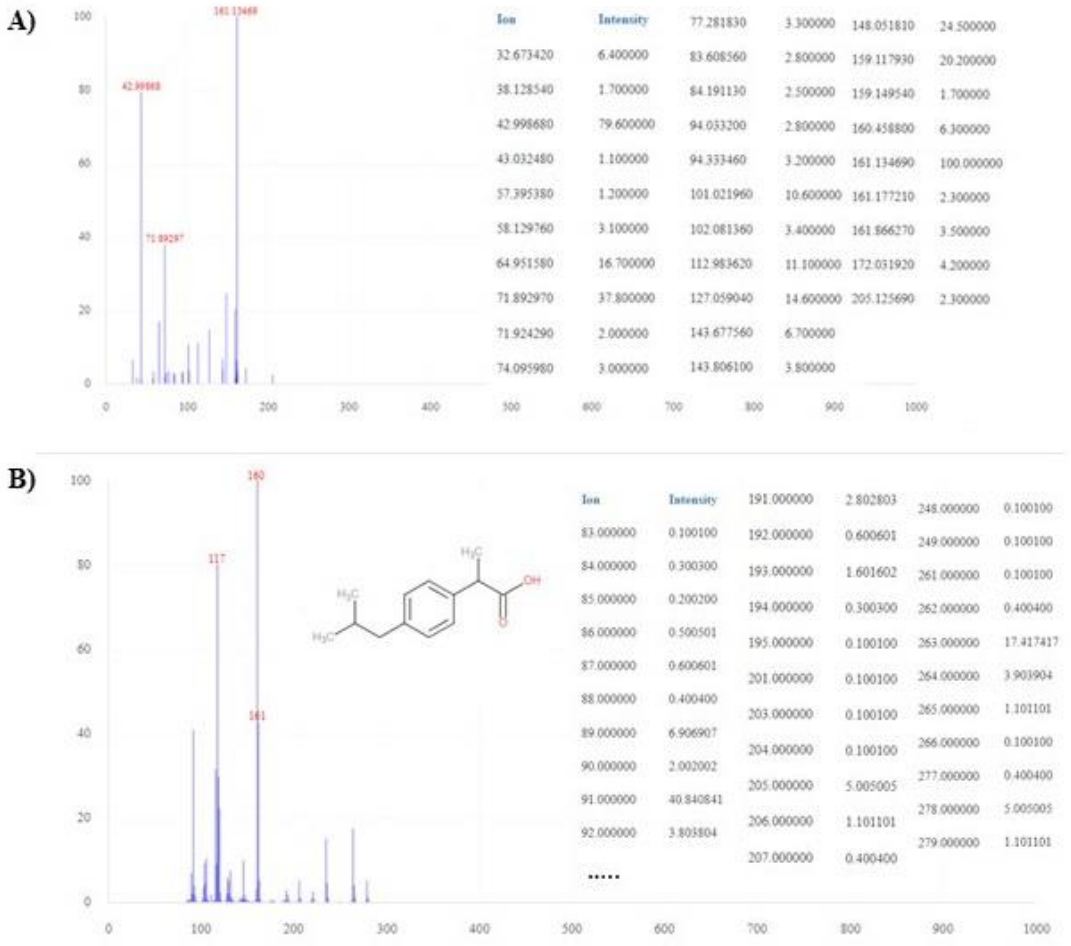


Figure 2.7: Examples of mass spectra from the Fiehn Library with a corresponding structure (left) and peak list (right). A) LC-ESI(-)Q-TOF MS/MS spectrum of ibuprofen; B) GC-EI-TOF MS spectrum of ibuprofen.

2.3.3 Derivatization

The concept of chemical modification of compounds to improve their physicochemical properties and thus amenability to qualitative or quantitative GC/LC-MS analysis has been fully recognized for many decades. Multiple properties are altered by modifying certain functional groups of a compound, such as their polarity, volatility, thermal stability, solubility, and ionization efficiency during MS. In turn, chromatographic behavior is improved, with improved

peak resolution, symmetry, and detector response. Consequently, sensitivity and selectivity of quantification are improved. The generated MS data provides valuable structural information for analyte identification. While the potential of derivatization in LC-MS analysis has been realized only recently [47], [57], its role in GC-MS analysis is well established. The main aims of derivatization prior to chromatography-MS analysis are [47], [58], [59]:

A. **Conferment of volatility**, by increasing the volatility of compounds due to mutual association of molecules through hydrogen or ionic bonds or by decreasing the excessive volatility of compounds (e.g., low M_w amines), and thus separating the compound peak from the solvent front;
B. **Improvement of thermal and catalytic stability:** reducing the number of reactive sites, thus avoiding decomposition and accompanying reactions - decarboxylation, dehydration, formation of cyclic structures, etc.;
C. **Improvement of ionization efficiency**: only for LC-MS of compounds that do not have high proton or cation affinity by improving surface activity (e.g., hydrophobicity) or introducing fixed-charge groups;
D. **Improvement of chromatographic behavior** [58]–[60]:
    i. improving the compatibility with the chromatographic environment. Free carboxylic acids and amines form strong hydrogen bonds with the -SiOH groups from the chromatographic system or sample residues left in the injector or column, resulting in system contamination, peak loss, or tailing caused by irreversible or reversible adsorption, respectively;
    ii. facilitating the separation of closely related analytes;
    iii. improving peak shape and symmetry and preventing peak tailing;
    iv. improving the linearity of chromatographic response, especially in low concentration ranges;
    v. improving selectivity and sensitivity by giving rise to an abundance of parent or high mass ions, thereby improving the signal/noise (S/N) ratio;
    vi. prolongation of R_t out of a region with a “high” background to a less contaminated, higher mass region;
    vii. improving separation of enantiomers by forming diastereomers.
E. **Improvement of spectrometric behavior**:
    i. generation of ions for more sensitive quantification;
    ii. generation of MS spectra with more favorable diagnostic fragmentation patterns for MS-based structural elucidation;
    iii. generation of mass shifts, that provides valuable structural information for determining the number and type of functional groups.

Despite all benefits, introducing the derivatization step prior to chromatographic-MS analyses is often considered the most time-consuming, error-prone step introducing by-product formation, ionization suppression, chromatographic system interferences or contamination and derivatives degradation, thus negatively influencing speed, sensitivity, selectivity, and accuracy of the analyses.

Derivatization prior to LC-MS analysis was introduced in the 1980s to improve separation, sensitivity, selectivity, and overall data quality. Aldehydes, ketones, alcohols, carboxylic acids, amines, and thiols are derivatized using versatile derivatization agents, primarily to achieve improved ionization efficiency, as in the case of alcohols and phenols, to decrease an analyte’s polarity, as in the case of amines, or to prevent oxidation of thiols during sample preparation [57]. Conversely, derivatization prior to GC-MS analysis has been well established for over seven decades, successfully expanding its application in various research fields, such as metabolomics, exposomics, pharmaceutical analysis, and food safety evaluations. Here, the most diversely applicable and popular derivatization methods are based on silylation, acylation, alkylation, and formation of cyclic derivatives, along with less common chiral derivatization [47], [60].

2.3.3.1 Silylation

2.3.3.1.1 Chemical aspects of silylation

Silylation is the simplest, quickest, and most versatile derivatization method for enhancing GC-MS performance [61]. The basic concept relies on the stoichiometric replacement of the active hydrogen atoms bound to electronegative elements (O, N, S, and P) with the electrophilic group $-SiR_1R_2R_3$ by nucleophilic substitution (S_N2). Many silyl groups are tested or employed, such as dialkylsilyl, alkyldimethylsilyl, aryl substituted silyl, aloxydimethylsilyl, and other alkoxy-substituted silyl, cycloorganosilyl, dialkyldimethylsilyl, halocarbon(di)methylsilyl and cyclic silyl groups. Of them, the trialkylsilyl $-Si(CH_3)_3$ and $Si(CH_3)_2C(CH_3)_3$ groups are most commonly used to form TMS and TBDMS derivatives, respectively [62]. The reaction mechanisms are shown in Figure 2.8. Silylation usually gives quantitative yield under relatively mild conditions by adding an excess of silylating agent to a dry sample residue. More vigorous conditions are required for the silylation of more hindered functional groups by proceeding with the reaction at elevated temperatures (40-120℃) and the addition of catalysts (further discussed in Section 2.3.3.1.3), depending on the nature and steric hindrance of the functional groups of interest.

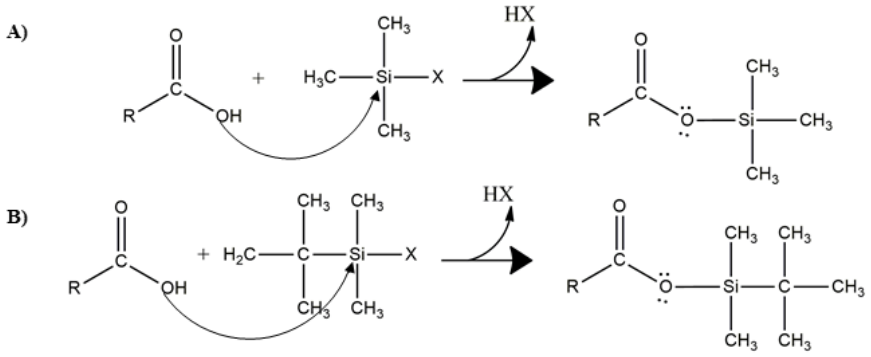


Figure 2.8: Representative examples of silylation mechanisms: A) trimethylsilylation, B) tert-butyldimethylsilylation (X denotes the leaving group).

Silylation is performed in aprotic solvents, dissolving the sample and the formed derivatives [63], such as ethyl acetate (EtAc) and acetonitrile (ACN) [64], dichloromethane, tetrahydrofuran, pyridine, and dimethylformamide [14]. Alternatively, previously dried samples are reconstituted in the silylating agent with/without the presence of a catalytic solvent, such as pyridine, driving the chemical equilibrium towards product formation [65]. However, the use of pyridine can lead to the formation of secondary products and chromatographic anomalies [14]. Silylation requires anhydrous conditions, as even trace amounts of water or protic solvents, such as methanol (MeOH), ethanol (EtOH), and isopropanol, can react with silylating agents and the silyl derivatives. Usually, silylating agents are added in excess so that the amount of water and protic solvent(s) becomes negligible [47], [66]. TMS derivatives are usually formed in sealed vials directly before analysis; however, silylation can also be carried out on solid phase microextraction sorbent, in injector or on-column, although rapid methods are not appropriate for silylation of hindered groups.

The formed silyl derivatives have improved volatility, thermal and catalytic stability, and chromatographic and spectroscopic behavior. Generally, silyl derivatives containing fluorinated groups at the silicon atom are more volatile and mobile in the GC system than the corresponding nonfluorinated analogs and form sharp and symmetric chromatographic peaks [47]. In general, the MS spectra of the TMS derivatives show much more fragmentation than the corresponding TBDMS derivatives due to the ease of elimination of stable radicals from their $M^{\cdot+}$ ions, suppressing competitive decomposition processes [47].

The MS spectra of trialkylsilyl derivatives are very characteristic in that they usually do not exhibit a $M^{\cdot+}$ peak (abundant only for phenylsilyl groups) but a distinctive $[M-15]^+$ peak, corresponding to $-CH_3$ loss in the case of TMS derivatives, or $[M-57]^+$ peak, corresponding to $-C(CH_3)_3$ loss in case of TBDMS derivatives, which frequently is the base peak. In the case of aliphatic alcohols, phenols, and carboxylic acids, this is the very stable siloxonium ion

($R\text{-}O^{+}\text{=}Si(CH_3)_2$), characterized by reasonably intense X+1 and X+2 peaks due to the presence of Si. In the case of amines and amides, the $[M\text{-}15]^{+}$ peak represents a silimmonium ion [37]**.** Other characteristic fragmentation pathways are well established, such as β-cleavage with charge retention on the silicon-containing moiety of silyloxy derivatives, N,N-di-TMS and N,O-di-TMS derivatives, elimination of silyloxy molecule ($R_1R_2R_3SiOH$) from both molecular and product ions and rearrangement reactions [47].

Silylation, however, has several disadvantages. TMS derivatives are susceptible to hydrolysis in the presence of moisture, given in descending stability order: TMS ethers > TMS esters > TMS amines [62]. Thus, their exposure to the atmosphere must be limited [59]. In contrast, the second most commonly used TBDMS derivatives are about 10^4 times more stable to hydrolysis, hydrogenolysis, reduction, and oxidation than the corresponding TMS derivatives, because the bulky TBDMS group protects the remaining compound structure from moisture [47], [61]. Silylation bears few risks, such as undesired degradation, rearrangement reactions, artifact formation, undesirable side reactions of polyfunctional compounds due to incomplete conversion, and unexpected reactions of the silylation agent with other small molecules formed in the course of the reaction (e.g., mineral and organic acids). The formation of by-products is also common that can strongly interfere with chromatographic analyses, causing column contamination and interfering with detectors, thus, affecting the stability of the formed derivatives [47]. Finally, the high M_W of the silyl derivatives, especially of polyfunctional compounds, may surpass the instrument's linear range. In such a case, the derivative may not elute from the GC column even at maximum working temperatures and significantly contaminate the GC-MS system.

2.3.3.1.2 Compounds amenable to silylation

Compounds containing active hydrogen functional groups, such as hydroxyl (-OH), carboxyl (-COOH), primary amine ($-NH_2$), secondary amine ($-R_1R_2\text{=}NH$), or thiol (-SH) are most commonly silylated prior to GC-MS analysis, together with less commonly silylated aldehyde (-CHO) and keto (-CO) functional groups. Alcohols, phenols, carboxylic acids, oximes, sulpho-acids, phosphorous acids, enols, amines, amides, imines, thiols, and thiocarboxylic acids [47]. Compounds containing phosphate (-POH), (-SOH), amide (-NOH), and aldehyde/ketone ($-CH_2RC\text{=}O$) functional groups undergo silylation less readily. BSTFA with potassium acetate, BSTFA, and TMCS, HMDS, and TMCS in pyridine, in the presence of bases or acids, can convert carbonyl groups into enolic tautomers, followed by the formation of enol silyl ethers.

The ease of reaction (Figure 2.9) is generally in the order of alcohols > phenols > carboxylic acids > amines > amides and is higher for primary than for secondary amines [62].

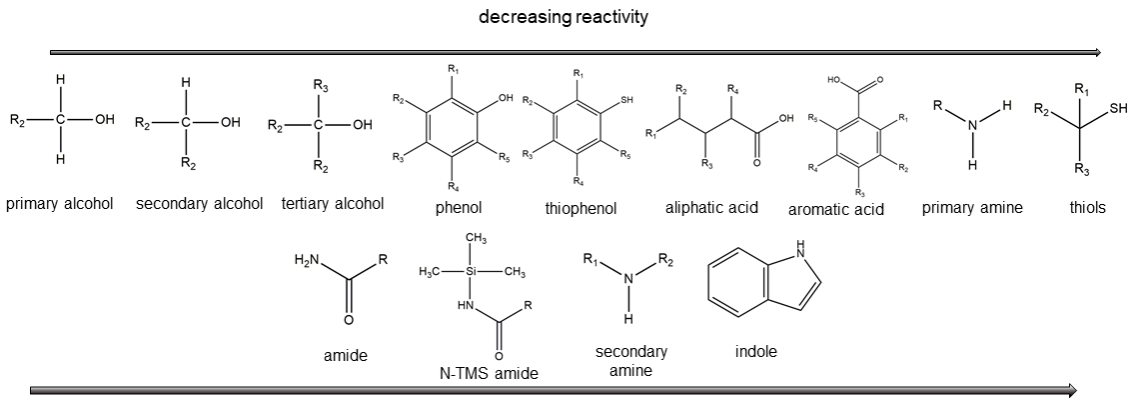
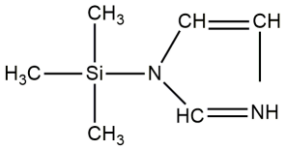
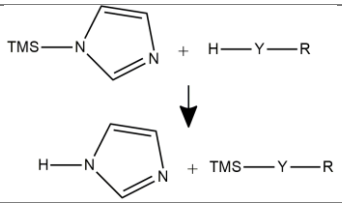
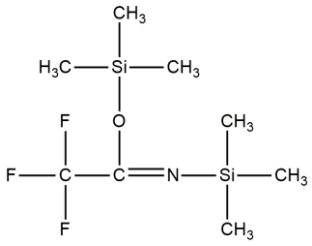
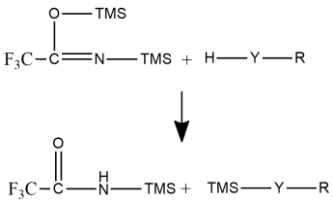
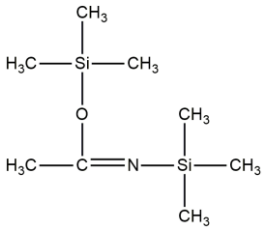
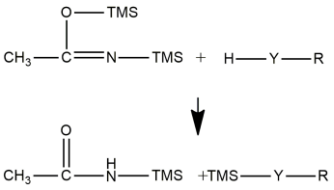
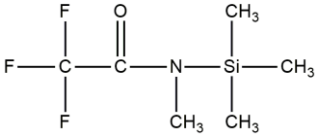
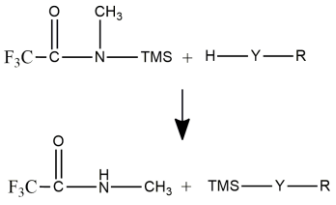
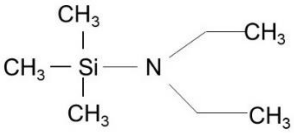
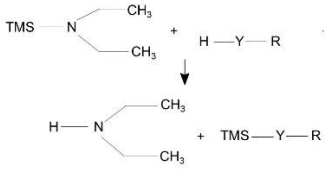


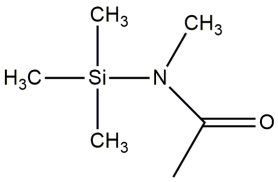
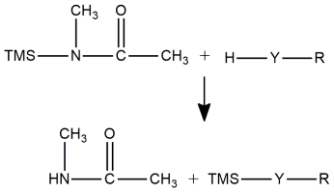
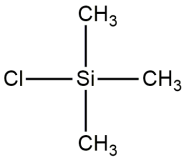
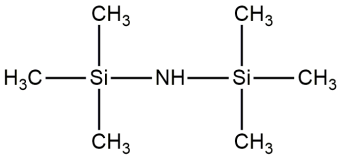
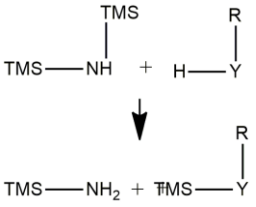
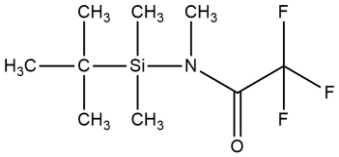
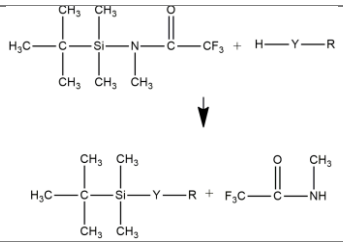
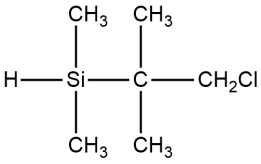
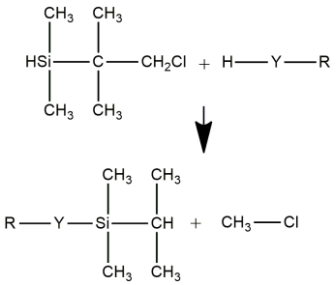
Figure 2.9: Compounds amenable to silylation (listed in the approximate order of decreasing ease of silylation).

2.3.3.1.3 Silylation reagents

Most commonly used silylation reagents are silyl chlorides, derivatives of acetic or trifluoroacetic acids and silyl derivatives of some amines and amides, including: N-methyl-N-trimethylsilylacetamide (MSA), *N,O*-bis(trimethylsilyl)acetamide (BSA), MSTFA, hexamethyldisilazane (HMDS), BSTFA, TMCS, trimethylsilylimidazole (TMSI), *N,N*-diethyltrimethylsilylamine (TMS-DEA) for TMS derivatization, and N-(tert-Butyldimethylsilyl)-N-methyltrifluoroacetamide (MTBSTFA) and *tert*-butylchlorodimethylsilane (TBDMCS) for TBDMS derivatization. TMCS and TMSI, together with trifluoroacetic acid, are commonly added as catalysts to HMDS, MSA, BSA, MSTFA, and BSTFA, as well as TBDMCS to MTBSTFA. Their chemical structures and their application are given in Table 2.2Table 2.2: Commonly used silyl reagents (*- usually employed as a catalyst) in descending order of their silyl donor activity. Carboxylic acids, unhindered alcohols, including carbohydrates and polyols, and unhindered phenols may be silyl with most of the listed silylation reagents. Also, mixtures of reagents can be used, such as BSTFA + TMSIM + TMCS, BSTFA + TMS-DEA + TMCS or MSTFA + TMCS + TMSIM. Seldom, TMS groups can also be introduced by N-TMS derivatives of piperidine, pyrrolidine, morpholine, N,N''-bis(trimethylsilyl)urea, and N-trimethylsilylacetanilide [47].

Table 2.2: Commonly used silyl reagents (*- usually employed as a catalyst).

Reagent	Chemical structure	Reaction	functional groups
TMSI			all alcohols (including *tert*-OH groups) phenols sugars
BSTFA			carboxylic acids hindered alcohols and phenols amines and diamines nucleosides
BSA			alcohols, phenols aromatics acids +TMCS: amines amino acids
MSTFA			carboxylic acids amines, diamines and amides non-hindered alcohols, phenols, and enols
TMS-DEA			aliphatic acids alcohols

MSA			alcohols polyols sugars amines, amino acids
TMCS*			alcohols carboxylic acids
HMDS			alcohols phenols +TMCS: sugars, carboxyphe nols, acids
MTBST FA			amines and diamines amino acids alcohols carboxylic acids
TBDM CS*			-

2.3.3.2 Other derivatization methods

Other derivatization methods are successfully employed in GC-MS analysis. Such is acylation, a standard method for derivatization of polar compounds, often used as an alternative to silylation [47]. It is most applicable for derivatization of primary and secondary amines and compounds with an alcoholic or phenolic hydroxyl group, producing stable derivatives, but it is not applicable to carboxylic acids, thiocarboxylic acids, sulphonic-, phosphonic- and phosphoric acids [47], [61].

Acyl derivatives are formed by a nucleophilic, electrophilic, or free radical displacement of an active hydrogen atom from OH, -NH_2, -NH, -SH groups by an acyl group (-CO-R). Numerous acylation reagents are used, such as:

- **(a)** acid anhydrides: acetic anhydride, trifluoroacetic anhydride, pentafluoropropionic anhydride, heptafluorobutyric anhydride, pentafluorobenzoyl anhydride;
- **(b)** perfluoroanhydrides and their derivatives, e.g., *N*-hepraflurobutyrylimidazole;
- **(c)** alkyl chloroformates, e.g., pentaflurobenzyl chloroformate;
- **(d)** acyl halides and acyl derivatives (acylimidazoles, acylamides, acylated phenols).

Reactions are carried out in tetrahydrofuran or chloroform in the presence of a basic catalyst, such as pyridine, substituted pyridines, lower tertiary amines, NaOH, or sodium acetate can scavenge a halogen acid or carboxylic acid produced in the course of acylation [47].

A further common derivatization – alkylation (arylation), includes the nucleophilic replacement of an active hydrogen atom by an alkyl or an aryl group. Common reagents for phenols, thiols, carboxylic acids, sulphonic acids, and phosphonic acids include:

- **(a)** *N,N*-dimethylformamide dialkyl acetals:
- **(b)** diazoalkanes homologous to diazomethane (e.g., trimethylsilyldiazomethane);
- **(c)** higher alcohols and halogen-containing alcohols, usually in the presence of Lewis-acid catalysts (e.g., HCl, H_2SO_4, BF_3, BCl_3, PCl_3, $POCl_3$, CF_3COOH, and $C_6H_5SO_3H$);
- **(d)** alkyl (methyl, ethyl, and propyl) bromides or iodides and benzyl bromide and its substituents or fluorinated analogs in the presence of K_2CO_3, BaO, $AgNO_3$, or Ag_2O;
- **(e)** methylation with tetramethylammonium hydroxide or trimethylsulfonium hydroxide under thermal conditions. Permethylation with formaldehyde (in the presence of $NaBH_4$ and H_2SO_4) or CH_3I and NaH is most common for amines, amides, and carbamates [47].

The formation of cyclic derivatives, such as acetals, ketals, esters, cyclic boronates, cyclic siliconides, cyclic carbotanes, or various cyclic derivatives, is less frequent. Compounds containing two or more -OH, -NH, and -SH groups in close proximity (1,2-, 1,3- or 1,4- positions in an alkyl chain or *ortho* position on an aromatic ring) react stereospecifically with appropriate aldehyde or ketone, monoesters of boric acid, methyl/ethyldichlorosilanes or phosgene/thiophosgene, respectively. Chiral separation and quantification of enantiomers rely on converting a racemic mixture of enantiomers to diastereomers using chiral derivatization agents, such as R-(+)-L-phenylethyl isocyanate. Other group-specific derivatization methods are very seldomly used, following the general trend in “omics” sciences to devote research efforts to developing multi-residue screening methods [47].

Chapter 3

Cheminformatics Approaches for MS-Based Compound Annotation

Accurate, confident, and reliable compound annotation (CA) during environmental analysis is crucial for identifying as many eco-exposome (EE) constituents as possible in complex environmental matrices. In the era of intense instrumental and computational development, CA evolves at a surprisingly rapid pace from the conventional approaches of manually matching the acquisition results to compound DB and MSL to cheminformatics approaches at all steps, from experimental design to CA. This chapter aims to overview the most recent developments in cheminformatics-assisted CA approaches, utilizing structural information inherent to MS spectra and their performance. To this end, the chapter proposes a new classification of cheminformatics-based CA approaches and summarizes the results of all recent performance evaluation studies for the first time. The chapter is divided into two sections. We first give the problem description (3.1) and then present the paper published in *Trends in Environmental Analytical Chemistry* (Section 3.2) that addresses the described problem.

3.1 Problem Description

Determining the structural identity of MS features is the bottleneck in conventional non-target screening (NTS) workflow [5], [34], [67] and is even more complicated in eco-exposome annotation (EEA) due to the complexity, diversity, and dynamics of the "chemical space" to be identified. Moreover, the most commonly employed GC-MS and LC-MS analytical platforms in EEA investigations increase the quality and amount of structural data available for CA and the study duration and complexity. Thus, ideal CA approaches have to address all aforementioned challenges. Traditional CA approaches rely on simple compound DB and MSL search. In the first case, the exact mass and MF of the unknown compound are first determined from MS data. They are then used to search for structure candidates across one or more DB, including large repositories (e.g., CAS, PubChem, ChemSpider), more minor, domain-specific DBs (e.g., Human Metabolome Database [68], Toxin and Toxin Target Database [69], Comptox Chemistry Dashboard [70]) or their combination. In the second case, MS data is directly searched against one or more MSL, such as the open access METLIN [71], MassBank [72], MoNa [73], Fiehn Lib [74], and the commercially available NIST Mass Spectral Library [1] and Wiley Library [75], providing means for assigning possible structure annotations to an MS feature. Despite that constant growth in size and scope, MSLs still cover only a tiny fraction of exposomics-relevant chemical information, equivalent to MS data for only 0.6-3.6% of the compounds present in compound DB. This lack of data results from inherent limitations, i.e., limited availability of reference standards and the lack of standardization in the field. Both DB and MSL search approaches are ineffective for EEA, as they cannot provide holistic and comprehensive coverage of the EE-relevant chemical space. Instead, they require manual expert knowledge-based visual inspection of the candidate lists for annotation of "known unknowns" and thus impair the identification of "unknown unknowns". Also, the DB and MSL search approaches are highly susceptible to errors-false positives, especially when

considering the allowed mass error for the queried compound, where the mass or molecular formula (MF) is correct, but the assigned structure is incorrect. They are also susceptible to false negatives, where no candidate compound falls within the allowed mass, MF, or metadata range [32]. As a result, many MS features, i.e., two-dimensional entities, such as an isotope pattern as *m/z* values and elution profile as R_t, correspond to compounds absent from DB and MSL, i.e., are yet to be included [67]. Consequently, most EEA studies have a low yield of correctly annotated compounds, and more than 70% of the MS features remain unidentified [34].

Novel cheminformatics-assisted CA approaches have been introduced and intensively used in many "omics" sciences in the last two decades. Most of them have been developed for metabolite annotation, but they are, in principle, applicable to identifying any small molecules. However, data regarding their use in the CA task using MS spectra is only recently publicly available. However, a few extensive reviews have given a critical insight into the basic principles, limitations, and achievements of CA approaches from the analyst's perspective, i.e., the end-user. Data regarding their use in EEA is based on publications of very few research groups, while the performance evaluations are based on the results from the recent Critical Assessment of Small Molecule Identification (CASMI) contests. In the paper included in this chapter, we address the lack of recent methodological classification and overview of cheminformatics-based CA approaches and their performance in the EEA task. Additionally, we discuss the obstacles to their use in EEA workflows and offer directions for further development, especially towards promoting their use in EEA and using GC-MS spectral data.

3.2 Related Publication

Journal paper

Ljoncheva, M., Stepišnik, T., Kosjek T., Džeroski, S., (2020) Cheminformatics in MS-based environmental exposomics: current achievements and future directions. *Trends in environmental analytical chemistry*. 28:e00099 2020, ISSN 2214-1588. DOI: 10.1016/j.teac.2020.e00099

This publication contains the following contributions:

- An overview of the eco-exposome (EE) concept and the task of eco-exposome annotation (EEA).
- A review of the basic principles of cutting-edge cheminformatics CA approaches by proposing a novel, methodology-based approach classification.
- A critical assessment of the annotation accuracy and confidence of the cheminformatics-assisted CA approaches by reviewing the results of recent performance evaluation studies on EE constituents.
- An identification of the main obstacles to further advancement of the field that should be tackled to improve further the capabilities of cheminformatics CA approaches is performed.

Trends in Environmental Analytical Chemistry 28 (2020) e00099

Contents lists available at ScienceDirect

Trends in Environmental Analytical Chemistry

journal homepage: www.elsevier.com/locate/teac

Cheminformatics in MS-based environmental exposomics: Current achievements and future directions

Milka Ljoncheva[a,c], Tomaž Stepišnik[b,c], Sašo Džeroski[b,c], Tina Kosjek[a,c,*]

[a] Jozef Stefan Institute, Department of Environmental Sciences, Jamova 39, 1000 Ljubljana, Slovenia
[b] Jozef Stefan Institute, Department of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia
[c] Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 4 May 2020
Received in revised form 23 July 2020
Accepted 28 July 2020

Keywords:
Eco-exposome annotation
Mass spectrometry
Molecular formula assignment
Substructure prediction
Structural elucidation
Machine learning

ABSTRACT

Compound annotation using MS/MS data is the major bottleneck in interpretation of mass spectrometry data during non-targeted screening and suspect screening exposomics studies. Apart from compound identification using available databases or mass spectral libraries, the true challenge comes when completely new compounds have to be identified. Along with recent advances in MS instrumentation that set grounds to a new revolutionary age in environmental exposomics, a multitude of cheminformatics annotation approaches has been developed. Herein, we review the basic principles of the cutting-edge cheminformatics MS-based approaches employed in eco-exposome annotation.

We give a solid background discussing the eco-exposome concept in relation to the advances in MS instrumentation, and define the three crucial cheminformatics tasks used in the eco-exposome annotation: molecular formula assignment, compound prioritization and compound annotation. The basic principles of compound annotation are discussed, which are based on three approaches of utilizing structural information inherent to MS data. These involve direct, indirect and joint annotation approaches. We assess their performance through the ability to annotate eco-exposome constituents. We discuss future perspectives and give directions to new annotation strategies and performance evaluation protocols aiming to solve current issues hampering the incorporation of cheminformatics annotation approaches in regular eco-exposome annotation workflows.

1. Introduction

Exposomics is one of the fastest developing 'omics' sciences, arising almost two decades ago as a mixture of environmental research, metabolomics and toxicology. Based on the concept of **exposome**, it investigates the cumulative effects of life-course human exposures, i.e. all non-genetic factors that influence a phenotype and are responsible for a significant portion of chronic disease risk [1]. In its other definition, the exposome represents cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, diet, behavior and endogenous processes [2]. Its extension, the **eco-exposome (EE)** recently became the "hot spot" in exposomics. The term involves both internal and external

Abbreviations: ANN, artificial neural network; AR, absolute rank; BDE, bond dissociation energy; CA, compound annotation; CASMI, Critical Assessment of Small Molecule Identification; CFM, competitve fragmentation modeling; CI, compound identification; CID, collision-induced dissociation; DB, database; DBE, double bond equivalent; DDA, data dependent acquisition; DIA, data independent acquisition; DT, decision tree; EE, eco-esxposome; EEA, eco-exposome annotation; EI, electron impact ionization; ESI, electrospray ionization; FP, fragmentation pattern; FT, fragmentation tree; FT-ICR, Fourier transform ion cyclotrone resonance; FWHM, full width half maximum; HCA, hierarchical clustering analysis; HR/AM-MS, high resolution / accurate mass - mass spectrometry; IM-MS, ion mobility-mass spectrometry; IP, isotope pattern; IM, ion mobility; IOKR, input/output kernel regression; IT, ion trap; KMC, kinetic Monte Carlo; LIT-TOF, linear ion trap-time-of-flight; LSER, linear solvation energy relationship; MF, molecular formula; MFP, molecular fingerprint; MKL, multiple kernel learning; ML, machine learning; MS^{all}, all ion fragmentation; MSL, mass spectral library; MSTR, mass spectral tree; MST, mass spectral tag; MS^n, multistage mass spectrometry; *m/z*, mass-to-charge ratio; NMR, nuclear magnetic resonance; NTS, non-targeted screening; PLSA-DA, partial least squares-discriminant analysis; QSAR, quantitative structure-activity relationship; QSRR, quantitative structure-retention relationship; RDBE, ring double bond equivalent; RF, random forest; RI, retention index; RRP, relative ranking position; R_t, retention time; SML, supervised machine learning; SVM, support vector machine; SS, suspect screening; TMS, trimethylsilyl; UML, unsupervised machine learning; TOF, time-of-flight; Q-TOF, quadrupole-TOF; IT-TOF, ion trap-TOF; Q-Orbitrap, quadrupole-Orbitrap.

* Corresponding author at: Jozef Stefan Institute, Department of Environmental Sciences, Jamova 39, 1000 Ljubljana, Slovenia.
E-mail address: tina.kosjek@ijs.si (T. Kosjek).

https://doi.org/10.1016/j.teac.2020.e00099
2214-1588/

markers of exposure, determining exposures from a point of contact between an external environmental stressor and a receptor inward into the organism and outward to the general environment [3]. Characterization of EE in environmental exposomics is extremely challenging due to the immense structural and toxicological diversity of its constituents, along with our limited knowledge of their identity. Namely, the virtual chemical space is estimated to consist of 10^{20}-10^{200} compounds [4], of which at least 9×10^6 are expected to occur in the environment and show potential harmful effects on human or biota. Of these, only 8.5 % (approximately 762,000 compounds) have been identified and documented in available exposome-specific databases (DBs), such as US EPA's Comptox Chemistry Dashboard [5], ContaminantDB [6], The Toxic Exposome Database (T3DB) [7] and the Exposome Explorer [8]. Consequently, exposomics studies have been shifting from monitoring known chemicals through target analysis, towards identification of new exposome constituents using analytical strategies such as suspect screening (SS) and non-targeted screening (NTS). SS refers to compounds whose molecular formula (MF) and chemical structure can be predicted or speculated *a priori* and thus aims to identify "known unknown" compounds. In contrast, NTS requires no previous knowledge, thus allowing the identification of any compound, including "unknown unknown" compounds.

Reliable CI in high-throughput experiments by either SS or NTS requires state-of-art analytical platforms. Gas chromatography (GC) or liquid chromatography (LC) are used for selective compound separation and are coupled to high resolution / accurate mass - mass spectrometry (HR/AM-MS) analyzers, including Orbitrap, Fourier transform ion cyclotrone resonance (FT-ICR), time-of-flight mass spectrometer (TOF MS), and hybrid MS configurations, such as quadrupole-TOF (Q-TOF), ion-trap-TOF (IT-TOF), quadrupole-Orbitrap (Q-Orbitrap) and others. The high (<5 ppm) or very high (<1 ppm) mass accuracy, high resolving power (up to 1,000,000 at full width half maximum (FWHM) for FT-ICR), wide linear dynamic range and high sensitivity in full scan mode of HR/AM-MS analzyers, enable reliable identification of EE constituents across a wide mass and concentration range. The variety of available hard and soft ionization techniques, together with different acquisition modes such as data dependent acquisition (DDA), data independent acquisition (DIA) and all ion fragmentation (MS^{all}) allow broad opportunities for generation of valuable orthogonal data for the purpose of annotiation of hundreds of exposome constituents in a single sample.

Employment of HR/AM–MS techniques in EEA studies increases data quality, but also study duration and complexity. Yet, these studies often result in a low yield of correctly identified compounds. The widely-accepted reason is that large number of mass spectral features, i.e. two-dimensional chemical entities - an isotope pattern (IP) as *m/z* values and elution profile as retention time (R_t) detected during the initial step of data processing, namely peak detection, alignment and normalization correspond to compounds that are yet to be included in DBs and/or MSLs [9]. Another important reason, however, is that many of the features correspond to "degeneracies". They are represented with redundant signals from same analyte due to in source fragmentation, analyte adduction in the ion source with various charge carriers (H^+, Na^+, K^+ etc.), oligomers formation with coeluting compounds and detection of naturally occurring isotopes (^{13}C, ^{15}N, etc.), artifacts (i.e. features detected due to informatic error which occurred due to baseline fluctuations, electronic noise from the mass spectrometer and poor resolving of compounds by the data-processing method used) of known EE constituents and contaminants (i.e.solvent impurities, plastic leachables, carry overs from previous experimetns etc.) [10]. Their formation is dependent upon sample type and complexity, employed sample extraction and instrumental methods and are more typical for LC–MS/MS analyses. Exhausting investigative efforts for annotation of these features might lead to missannotations. Therefore, a thorough signal classification is performed as step upstream [9]. The task includes identification and removal of features of contaminants and artifacts and grouping of all redundant features of the same analyte using ad hoc annotation rules based on common ion adducts and neutral losses or dynamic rule set created from the combination of observed ions [11]. The cheminformatics methods applied to perform these routine steps determine the complexity of data interpretation for the purpose of compound annotation (CA).

Structural characterization of selected features in EEA is performed following the tasks: 2) MF assignment; 3) candidate prioritization; 4) CA and 5) complete structure elucidation with stereochemical assignments. Each of the tasks 1)-4) is independent and is usually embedded in numerous cheminformatics approaches, separately or in combination. The last task uses reference standards and is preferably performed by at least two complementary analytical techniques (GC/LC-MS, nuclear magnetic resonance (NMR), ion mobility-mass spectrometry (IM-MS)). This review discusses the cheminformatics approaches employed in tasks 2)-4), while overview of cutting-edge data processing tools can be found elsewhere [11,12]. Most cheminformatics approaches have been developed for the purpose of metabolite annotation, but are in principle applicable also to the identification of any small molecules. They utilize structural information inherent to MS and tandem mass (MS/MS) spectra, DBs and mass spectral libraries (MSLs) and versatile metadata.

Data regarding the use of cheminformatics approaches, apart for data processing purposes, is scarce. Aiming to further encourage the employment of these cheminformatics approaches in exposomics, this review gives a critical insight into the basic principles, limitations and achievements of CA approaches from the perspective of an analyst being the end-user, thus omitting their description from a computational perspective. The annotation ability and confidence of these approaches is critically assessed by reviewing the results of the Critical Assessment of Small Molecule Identification (CASMI) contests [13–15]. To enhance further development in the field, we propose several directions for improvement of compound prioritization strategies and standardization of performance evaluation protocols. To the best of our knowledge, our review of the use of cheminformatics in EEA is the first of its kind.

2. Assignment of MF

MF assignment is the first and the most challenging task during CA. Erroneous MF assignments lead to erroneous CA. Considering that many of the CA methods use MFs together with MS data as an additional input (CSI:IOKR, MP-IOKR, SIMPLE, CSI:FingerID, MetExpert, FingerID) [16–21] or as metadata (ChemDistiller, MetFrag, MetFrag v2.2) [22–24], this task becomes even more challenging.

A MF can be assigned to the candidate molecule either with or without prior determination of its accurate mass. Standard deterministic approaches calculate all possible MFs for a given accurate mass, constrained by predefined element type and atom number, mass error window and rules of chemical bonding, such as double bond equivalent (DBE) and the nitrogen rule. This is a computationally intense process which generates long candidate MF lists, often yielding many chemically unreasonable candidate MFs. Another approach is to annotate and group features of the same compound, match experimentally determined *m/z* to the accurate mass of neutral molecules and associated MFs and match the MFs to a reference file of compounds (PUTMEDID-LCMS) [25].

In contrast, the heuristic approaches assign MF to an unknown compound without prior determination of the accurate mass. The most probable MF is selected according to user-predefined thresholds and constraints (see below). This is especially plausible for the spectra with low-intensity or absent molecular ion, which is frequently the case in electron impact (EI)-ionized compounds [26].

1) *IP.* IP is a unique and theoretically predictable set of MS peaks generated by isotopes of the molecular and fragment ions of a molecule. Candidate MFs are ranked by comparing the experimental IP of the unknown compound with the predicted ones (MOLGEN-MS/MS, SIRIUS v1.0, Maximum Colorful Subtree approach) [26–28]. When used as the only constraint, IPs can yield the correct MF assignment only for HR/AM-MS data with resolution > 300,000 FWHM (FT-ICR, Orbitrap) [27]. Combining IP with other constraints eliminates > 95 % of spurious MF assigments [29]. IPs can be combined with information regarding the fragmentation pattern (FP) of a compound. Using MS/MS spectrum, a fragmentation graph can be generated that represents all possible MFs for the fragments of the compound and all possible fragmentation reactions between them. From this graph, the most likely fragmentation reactions are selected, representing a fragmentation tree (FT). FTs are hierarchical organization of ions in a graph representation that models the subsequent series of fragmentation events that occur during MS analysis, resulting with fragments. The root of these trees represents the unfragmented ion, the nodes are labeled with the MFs of the fragments, and the vertices (the direct edges) correspond to neutral losses (i.e. the residues of a fragmentation product that are not detectable by mass spectrometer) [30]. FTs offer more discriminatory power than IPs during MF assignment [31], but the apporaches that employ their computation might have decreased accuracy and longer running times. As the pool of possible fragmentation reactions and consequent fragmentation cascades is massive, a very high number of FTs can be computed, e.g. more than 10^{100} FTs for small molecules with $m/z<$ 150 Da. To avoid this potenital 'combinatorial explosion', restrictions are introduced regarding the number of candidate MFs for which FTs are generated (SIRIUS v3.0, SIRIUS v4.0) [32,33] or the tree depth (that is, the number of consequent fragmentation events explained with FTs) and the intensity of MS/MS peaks to be considered. The combination of FT computation with IP analysis outperforms each of the two methods when used separately (SIRIUS v2.0) [31,33].
2) *Structural information inherent to multistage mass spectrometry (MS^n) data.* Fragmentation cascades during MS^n analysis (HR/AM-MS data to MS^3 level or unit-resolution data from ion trap-mass spectrometry (IT-MS)) can be used for MF assignment through the computation of mass spectral trees (MSTRs). In MSTRs, fragmentation events from each of the MS levels are interconnected in a similar graph hierarchy as in FTs (MAGMa) [34]. Heuristically extracted constraints from the predicted elemental composition of its fragments and neutral losses are used to restrict MSTRs generation and ensure accuracy.
3) *Chemical rules.* LEWIS and SENIOR rules, DBE, ring double bond equivalent (RDBE), the nitrogen rule, and the element ratio check are most frequently applied chemical rules that prevent chemically illogical candidate MFs to be assigned (SIRIUS v1.0, MCS approach) [27,28].
4) *Heuristic rules obtained by statistical examination of large compound DBs.* The Seven Golden Rules [29] and the hydrogen rearrangement rules (MS-FINDER) [35] are most widely used heuristic rules. The Seven Golden Rules filter candidate MFs based on their compliance with rules that introduce restrictions to the number of elements, valence states of each element, IP, elements ratios, heteroatom ratios, sum number of atoms of multiple elements and presence of derivatives [29]. Their application is sufficient to derive the most likely MF from accurate mass and isotopic ratio mass spectral measurements; however their filtering power is dependent upon data diversity and quality [36]. They are able to find candidate structures most similar to the true structure (SIRIUS v3.0) [33], and despite often missing the true candidate, they significantly restrict the number of possible MFs (SIRIUS v1.0, MCS approach, SIRIUS v4.0) [27,28,32]. DB-derived rules of rearrangements of hydrogens during bond cleavages for CNOPS elements are generated based on the classic even-electron rule. The rules are successfully applied in MF assignment from low-energy collision-induced dissociation (CID) mass spectra [35].
5) *Mathematical rules*, detecting spurious MFs by restricting atoms' valences (MOLGEN-MS/MS) [26].
6) *Method-specific dynamic set of heuristic rules.* An approach assigns MFs using MS^n data to compute FTs, and restricts candidate list of MFs by establishing constantly updating rules on numbers of individual atoms to constitute the MF. The repeated application of rules for each precursor-fragment combination produces new constraints that are used in the next cycle to further decrease the list of MF candidates [36]. Other approach learns common losses and their frequencies from MS/MS data for the purpose of selecting the most probable FT, and consequently the most probable MF [33].

3. Compound prioritization

Structural information inherent to MS and MS/MS spectra is often supplemented by various metadata in order to prioritize the exposome-relevant candidates. Most valuable prioritization criteria include:

- ***Chromatographic retention-related criteria.*** Retention index (RI) is a valuable metadata, but it is of lower importance for CA than structural information inherent to MS. Robustness of GC–MS capillary columns allows standardization of R_ts from GC–MS data into library-available RIs, such as Kovats index and Lee RI, to which experimentally derived RIs are compared. As RIs from LC–MS data are dependent upon several parameters, including pH, temperature, buffer, solvent compositions and gradients etc., RI libraries are rarely constructed. Predictive models are based on quantitative structure-retention relationship (QSRR) modeling, which correlates the normalized RIs with compound's most relevant physicochemical properties (log K_{ow}, log D, pKa), linear solvation energy relationship (LSER) [37], 2D molecular descriptors [19,38,39] or their combination [40]. Models learn the correlations for thousands of known compounds and are able to predict R_ts for unknown compounds. Employed models include machine learning (ML) methods, such as multiple linear regression [37], artificial neural networks (ANNs) [19,38], random forest (RF) classificators [39], support vector machines (SVMs) [39] or their combination [40]. They are able to predict RIs for structurally diverse compounds, but have limited use. Their performance depends upon the accuracy of the molecular descriptors and is limited to a single chromatographic setup. Alternatively, retention order can be predicted using SVMs [41]. R_ts determined in one chromatographic setup can be projected to a different one [42]. Although more accurate than QSRR approaches, the projection approach is limited to compounds for which the R_ts have previously been determined. The above-mentioned concepts have merely been integrated with MS-based CA workflows, which is the underlying reason for their limited application in EEA [43].

- ***Energy-related criteria.*** Here, candidate structures are prioritized according to their "energy cost", aiming to eliminate energetically unfavorable candidates. The prioritization includes comparison of steric energy distributions to eliminate sterically hindered candidates [44] and comparison of the energy required to fragment 50 % of a selected precursor ion ($ECOM_{50}$) [38,45].
- ***Data source-related criteria.*** The presence of a compound in DBs and the number of references and patents from scientific literature or monitoring reports indicate the likelihood of being an already identified EE constituent [29,44,46]. The searchable chemical space is also narrowed down by detecting the presence of structurally related substances with common functional groups, fragments, neutral losses and substructures or structures in DBs and MSLs [47], and by SS for homologues [46]. However, these criteria are of no help in the case of "unknown unknowns".
- ***Environmental behavior and toxicity related criteria***. Predicted properties on the bioaccumulation, biotransformation, mutagenicity, carcinogenicity, endocrine disruption and reproductive toxicity of candidate compounds serve as prioritization criteria, in such a way that candidates with properties indicating environmental persistence, toxicity or bioaccumulation are prioritized for further annotation. Numerous prediction models have been shown of great value for EEA [46,48–51], of which the quantitative structure-activity relationship (QSAR) models based on ML methodologies are most commonly exploited. They use experimental or predicted physicochemical ($logK_{ow}$, logD, pKa), quantum chemical (charge, steric energy, electron distribution, spatial disposition) properties and molecular descriptors to predict environmental behavior and toxicological properties of unknown compounds. Other models are the read-across models and models based on literature or data-derived rules. For example, exposure and toxicity predictions were employed for compound prioritization in EEA study on vacuum dust samples [50]. Formation of potential transformation products (TPs) is predicted by models based on pre-defined sets of transformation rules (e.g. enviPath [52], PathPred [53], MINEs [54] and BiotransformerDB [55]). Presence of some of the candidate structures in such predicted TP lists is a strong evidence of their environmental relevance, thus prioritizing them as most plausible annotation during EE investigations, such as in the example of benzotriazole TPs identification [56].
- ***Complementary information.*** Data provided by complementary GC–MS or LC–MS analyses with versatile ionization techniques, two-dimensional GC or LC techniques, orthogonal techniques, such as NMR hyphenated to a GC or LC and IM-MS [57] are used. While the use of NMR in NTS is limited due to concentration-dependent sensitivity, the importance of IM-MS is increasing due to its ability to separate ions with identical *m/z*, based on their 3D molecular structure [38,45].

4. Compound annotation

The multitude of available CA approaches eases the search for 'known unknowns' in MSLs and DBs and navigates towards the annotation of 'unknown unknowns'. Table 1 gives an overview of cutting-edge CA approaches, along with a description of the computational method used, input data, use of DB and MSL, training and test datasets, as well as the type of MS spectra (with regard to the ionization) and the accessibility of the CA approaches. Diverse approaches are used to establish quantitative MS-structure relationships, categorized as **(1)** direct, **(2)** indirect and **(3)** joint approaches. All three groups aim to predict the presence of specific substructures and classify compounds accordingly, or to completely elucidate the compound's structure.

Direct approaches extract structural information directly from MS spectra (Fig. 1,Table 1). **Indirect approaches** (Fig. 2, Table 1) transform mass spectra into FTs [23,24,58–61], MSTRs [34], molecular descriptors [17,19,21] or their combination [16,18,20,21], that indicate presence of structural (atom, type of ring, atom pair, functional group) and/or spectral features. They allow comparison of input MS spectra and candidate structures retrieved from DB search in a "third format", thus preserving structural information from both the structures and the MS spectra. In this view, this "third format" has higher discriminatory power to reflect structural similarity than direct prediction made from MS spectra [13,19,62–64]. **Joint annotation approaches** (Fig. 3, Table 1) combine direct and indirect approaches with metadata in filtering [22] or consensus manner [44,65].

4.1. Direct annotation approaches

4.1.1. Direct substructure prediction and classification approaches

As shown in Fig. 1, direct approaches use MS spectra, sometimes combined with the exact mass to predict the presence of certain substructures or to classify compounds. Classification is performed by locating possible structural neighbors of unknown compounds based on MS/MS spectral similarity in a DB-independent manner. The simplest unsupervised machine learning (UML) approach is used by MS2Analyzer, which combines structural information inherent to product ions and their fragments, neutral losses and isotopic ratios, with literature-derived neutral loss/substructure pairs to detect the presence of same or similar substructures [67]. More advanced ML approaches employ UML [69] and supervised machine learning (SML) [66,68] classification methods to detect the presence of specific substructures and classify unknowns accordingly. MetFamily first extracts spectral patterns of MS/MS features by molecular networking between MS^1 and MS^2 levels and uses hierarchical clustering analysis (HCA) to classify compounds into metabolite families [69]. MSClass determines structural neighbors according to the presence of predefined substructures (spectral features) and classifies them using ANNs [66]. iMet uses RF classifiers to identify structural neighbors for 'unknown unknowns' from an in-house MSL, based on MS/MS and IPs similarity, and mass difference between a compound and the candidate [68]. Additionally, iMet suggests a chemical transformation that is most likely to separate a neighboring candidate from the unknown compound [68].

A recent chemiformatics strategy [70] employs fragment set enrichment analysis (FSEA) for *de novo* compound class recommendation, in cases when no candidates are retrieved from molecular structure DBs. Significant peaks of the unknown MS/MS spectrum are assigned using a fragment set, generated from high-quality mass spectral records extracted from multiple MSLs and annotated using MS-FINDER [35]. The set consists of ontologies of parent-fragment pairs for the most frequrently observed fragments generated in semi-automated manner. Assigned compound class is further checked by the metabolome network linked by MS/MS similarity, ontology term similarity and bioreaction linkage and final check on structure specificity with literature curation.

4.1.2. Direct structure elucidation approaches

These rule-based approaches (Table 1, Fig. 1) include the commercially available MOLGEN-MS [73], Mass Frontier [71] and ACD/MS Fragmenter [72] and the publicly available MS-FINDER [35]. They work by first generating a candidate list from exact mass or MF search in single or multiple DBs [35,71,72]. Then, spectra of candidates are simulated using fragments predicted from a set of fragmentation rules, which are literature-derived (up to 130,000 [73]), user-defined, or automatically learned from experimental data. Predicted spectra are compared against the acquired

Table 1
Cheminformatics annotation approaches. CA- commercially available; NA-not available; PA-publicly available.

Computational tool			Computational method	Input data	Method requirements		**Training dataset**/ Test dataset		Discovering unknown unknowns?	Availability		Ref.
					DB	MSL	data (source)	ionization type		CA	PA	
DIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	MSClass	ANN	MS/MS spectra			NA	EI			NA	[66]
		MS2Analyzer	literature-derived neutral losses	MS/MS spectra			3359 MS/MS / 860 glycosides (MassBank) 12521 MS/MS (cell cultures)	ESI +/-			✓	[67]
		iMet	RF classifiers	QTOF-MS/MS spectra + exact mass		✓ (in-house)	**50 000 pairs of metabolites (825 RPs, KEGG)** D1:148 metabolites D2:100 metabolites D3:31 metabolites (CASMI)	ESI +			✓	[68]
		MetFamily	HCA	MS and MS/MS spectra			(tomato leaf samples)	ESI	(only annotation)		✓	[69]
		FSEA (as part of MS-DIAL v3.0)	MS-FINDER + ontology-based annotation + over-representation analysis	MS/MS spectra	✓		(31 tissues from 12 plant species)	ESI			✓	[70]
	STRUCTURE ELUCIDATION	MassFrontier	literature-derived and user-added	MS/MS spectra M	✓ (in-house)		top 200 most prescribed drugs in USA, 2011	EI, CI and ESI +/-	✓ (if similar to known compounds)	✓		[71]
		ACD/MS Fragmenter	fragmentation rules		✓		100 mass spectra (NIST 05)	EI		✓		[72]
		MS-FINDER	hydrogen rearrangement rules		✓ (in-house)		D1:5036 MS/MS (MassBank) D2:936 MS/MS (plasma)	ESI			✓	[35]
		MOLGEN-MS	Varmuza and NIST classifiers	MF + MS spectra		✓	71 MS spectra of unknown unknowns (groundwater)	EI			NA	[73]
INDIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	Rasche method	FT computation + FT alignment	MS/MS spectra + MF	✓		D1:97 compounds D2:370 compounds D3 44 compounds D4: (poppy extracts)	ESI +/-	✓		NA	[58]
		MS2LDA	ML	MS + MS/MS spectra			D1:1953 MS/MS spectra (MassBank) D2:5670 MS/MS spectra (GNPS) ([62])	ESI +/-	✓		✓	[74]
		GMD algorithm	DTs classification	MS spectra + RI		✓	EI-MS spectra of MeOX and TMS derivatives	EI	✓ (only if present in GMD as unknown)		✓	[75]
	COMBINATORIAL FRAGMENTATION	FT-BLAST	FT computation + FT alignment	MS/MS spectra		✓	D1:97 compounds D2:370 compounds D3:44 compounds D4:(poppy extracts)	ESI +/-			NA	[58]
		FiD	combinatorial search of substructures for product ions	MS/MS spectra			D1:20 amino acids and 17 sugar-phosphates D2:20 amino acids	ESI +/- and MALDI			✓	[76]
		MetFrag	FT generation + heuristics	MS/MS spectra	✓		D1:200 spectra/49 compounds (KEGG DB) D2:510 spectra/102 compounds (search against PubChem)	ESI			✓	[24]
		MetFrag v.2.2			✓		D1: 102 compounds D2:473 MS/MS spectrum/359 compounds	ESI			✓	[23]

Table 1 (*Continued*)

Computational tool			Computational method	Input data	Method requirements		**Training dataset**/ Test dataset		Discovering unknown unknowns?	Availability		Ref.
					DB	MSL	data (source)	ionization type		CA	PA	
DIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	MSClass	ANN	MS/MS spectra			NA	EI			NA	[66]
		MS2Analyzer	literature-derived neutral losses	MS/MS spectra			3359 MS/MS / 860 glycosides (MassBank) 12521 MS/MS (cell cultures)	ESI +/-			✓	[67]
		iMet	RF classifiers	QTOF-MS/MS spectra + exact mass		✓ (in-house)	**50 000 pairs of metabolites (825 RPs, KEGG)** D1:148 metabolites D2:100 metabolites D3:31 metabolites (CASMI)	ESI +			✓	[68]
		MetFamily	HCA	MS and MS/MS spectra			(tomato leaf samples)	ESI	(only annotation)		✓	[69]
		FSEA (as part of MS-DIAL v3.0)	MS-FINDER + ontology-based annotation + over-representation analysis	MS/MS spectra	✓		(31 tissues from 12 plant species)	ESI			✓	[70]
	STRUCTURE ELUCIDATION	MassFrontier	literature-derived and user-added	MS/MS spectra M	✓ (in-house)		top 200 most prescribed drugs in USA, 2011	EI, CI and ESI +/-	✓ (if similar to known compounds)	✓		[71]
		ACD/MS Fragmenter	fragmentation rules		✓		100 mass spectra (NIST 05)	EI		✓		[72]
		MS-FINDER	hydrogen rearrangement rules		✓ (in-house)		D1:5036 MS/MS (MassBank) D2:936 MS/MS (plasma)	ESI			✓	[35]
		MOLGEN-MS	Varmuza and NIST classifiers	MF + MS spectra		✓	71 MS spectra of unknown unknowns (groundwater)	EI			NA	[73]
INDIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	Rasche method	FT computation + FT alignment	MS/MS spectra + MF	✓		D1:97 compounds D2:370 compounds D3 44 compounds D4: (poppy extracts)	ESI +/-	✓		NA	[58]
		MS2LDA	ML	MS + MS/MS spectra			D1:1953 MS/MS spectra (MassBank) D2:5670 MS/MS spectra (GNPS) ([62])	ESI +/-	✓		✓	[74]
		GMD algorithm	DTs classification	MS spectra + RI		✓	EI-MS spectra of MeOX and TMS derivatives	EI	✓ (only if present in GMD as unknown)		✓	[75]
	COMBINATORIAL FRAGMENTATION	FT-BLAST	FT computation + FT alignment	MS/MS spectra		✓	D1:97 compounds D2:370 compounds D3:44 compounds D4:(poppy extracts)	ESI +/-			NA	[58]
		FiD	combinatorial search of substructures for product ions	MS/MS spectra			D1:20 amino acids and 17 sugar-phosphates D2:20 amino acids	ESI +/- and MALDI			✓	[76]
		MetFrag	FT generation + heuristics	MS/MS spectra	✓		D1:200 spectra/49 compounds (KEGG DB) D2:510 spectra/102 compounds (search against PubChem)	ESI			✓	[24]
		MetFrag v.2.2			✓		D1: 102 compounds D2:473 MS/MS spectrum/359 compounds	ESI			✓	[23]
		MAGMa	MST computation	MS^n spectra + candidate list	✓		D1:510 spectra/102 drug-like compounds D2:11 compounds	ESI -			✓	[34]

		MAGMa	MST computation	MS^n spectra + candidate list	✓		D1:510 spectra/102 drug-like compounds D2:11 compounds	ESI -	✓	[34]
	COMBINATORIAL FRAGMENTATION	MolFind	FT computation + heuristics	MSMS spectra + exact mass + RI + $ECOM_{50}$+drift time	✓		35 pharmaceuticals 253 compounds (PubChem)	ESI +/-	✓	[38]
INDIRECT APPROACHES		Hufsky method	FT computation	MS spectra			50 compounds, TMS and PFBO derivatives	EI	✓ (tentative identification)	NA	[59]
		MIDAS	FT computation + heuristic rules	MS/MS spectra	✓		D1:Orbitrap-MS/MS spectra D2:TOF-MS/MS spectra	ESI +/-	✓	[60]
		MIDAS-G	FT computation + EGG				D3:TOF-MS/MS spectra D4:TOF-MS/MS spectra (MassBank)			[61]
	MOLECULAR FINGERPRINT PREDICTION	MetExpert	ANN (Rt prediction) PLS-DA	MF + MS spectra RI	✓		**1782 compounds (GMD + Adam Essential Oil Library)** D1:412 plant metabolites-143 TMS derivatives D2:161 TMS derivatives (MassBank)	EI	✓	[19]
		FingerID	ML (MKL, SVMs)	peak list (MF as optional entry)	✓ ✓		**D1:2755 MS/MS spectra (MassBank) D2:4879 MS/MS spectra (GNPS)** D1:514 MS/MS spectra D2:293 MS/MS spectra D3:403 MS/MS spectra	ESI +/-	✓	[20]
		CSI:FingerID	Combinatorial optimization + ML (IOKR, MKL, SVMs)	MS/MS spectra	✓		**D1:4879 spectra (GNPS) D2:2755 spectra (MassBank)** D1:978 compounds (Metlin) D2:402 compounds (MassBank)	ESI +/-	✓	[18]
					✓		**D1:4138 compounds (GNPS) D2:2120 compounds (Agilent library)** D1:3868 compounds (GNPS) D2:2055 compounds (Agilent library) D3:D1+D2	ESI +/-	✓	[62]
INDIRECT APPROACHES	MOLECULAR FINGERPRINT PREDICTION	CSI:IOKR	ML (IOKR, MKL, SVMs)	MF + MS/MS spectra			D1 from [62]	ESI +/-		[16]
		MP-IOKR					625 compounds [62]	ESI +/-		NA	[21]
		SIMPLE					D2 from [18]	ESI +/-	✓	[17]
		L-SIMPLE	sparse model	MS/MS spectra + MF			D2 from [58]	ESI +/-	✓	
IN SILICO SPECTRAL PREDICTION	DIRECT APPROACHES	CFM-ID v1.0	ML (generic probabilistic model + ANN)	Spectra prediction task: chemical structure CA task: EI-MS or ESI-MS/MS spectra + monoisotopic mass + candidate list (optional)	Spectra prediction task: CA task: ✓ (if no candidate list submitted)	Spectra prediction task: CA task: ✓ (*in silico*)	D1:17324 compounds (NIST14) D2:100 compounds ([24]) D3:20588 compounds (NIST 14)	EI	✓	[77]
		CFM-ID v2.0	ML (generic probabilistic model + ANN)				**D1:1985 peptides (Metlin) D2: 1491 compounds from Metlin** D3:192 compounds (MassBank) D2: 976 compounds (Metlin) D3:192 compounds (MassBank) D4:500 compounds (HMBD)	ESI +/-	✓	[78]
		CFM-ID v3.0	ML (generic probabilistic model + ANN)+ fragmentation rules + metadata				**1000 compounds (drugs, lipids)** 208 ESI-MS/MS spectra of 185 unique compounds (CASMI 2016, [15])	ESI +/-	✓	[79]
		ISIS	ML (KMC, ANN)	chemical structure		✓ (*in silico*)	**22 lipids** 45 lipids	ESI +	✓ (lipids)	NA	[80]
		Quantum chemical modelling	quantum chemical methods, stochastic methods		✓		not evaluated	EI		NA	[81]
	INDIRECT APPROACH	NEIMS	ML (ANN)+mass filtering	MS spectra		✓ (combined	**240 942 EI-MS spectra (NIST 2017, main library)**	EI		✓	[82]

Table 1 (*Continued*)

Computational tool			Computational method	Input data	Method requirements		**Training dataset**/ Test dataset		Discovering unknown unknowns?	Availability		Ref.
					DB	MSL	data (source)	ionization type		CA	PA	
DIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	MSClass	ANN	MS/MS spectra			NA	EI			NA	[66]
		MS2Analyzer	literature-derived neutral losses	MS/MS spectra			3359 MS/MS / 860 glycosides (MassBank) 12521 MS/MS (cell cultures)	ESI +/-			✓	[67]
		iMet	RF classifiers	QTOF-MS/MS spectra + exact mass		✓ (in-house)	**50 000 pairs of metabolites (825 RPs, KEGG)** D1:148 metabolites D2:100 metabolites D3:31 metabolites (CASMI)	ESI +			✓	[68]
		MetFamily	HCA	MS and MS/MS spectra			(tomato leaf samples)	ESI	(only annotation)		✓	[69]
		FSEA (as part of MS-DIAL v3.0)	MS-FINDER + ontology-based annotation + over-representation analysis	MS/MS spectra	✓		(31 tissues from 12 plant species)	ESI			✓	[70]
	STRUCTURE ELUCIDATION	MassFrontier	literature-derived and user-added	MS/MS spectra M	✓ (in-house)		top 200 most prescribed drugs in USA, 2011	EI, CI and ESI +/-	✓ (if similar to known compounds)	✓		[71]
		ACD/MS Fragmenter	fragmentation rules		✓		100 mass spectra (NIST 05)	EI		✓		[72]
		MS-FINDER	hydrogen rearrangement rules		✓ (in-house)		D1:5036 MS/MS (MassBank) D2:936 MS/MS (plasma)	ESI			✓	[35]
		MOLGEN-MS	Varmuza and NIST classifiers	MF + MS spectra		✓	71 MS spectra of unknown unknowns (groundwater)	EI			NA	[73]
INDIRECT APPROACHES	SUBSTRUCTURE PREDICTION AND CLASSIFICATION	Rasche method	FT computation + FT alignment	MS/MS spectra + MF	✓		D1:97 compounds D2:370 compounds D3 44 compounds D4: (poppy extracts)	ESI +/-	✓		NA	[58]
		MS2LDA	ML	MS + MS/MS spectra			D1:1953 MS/MS spectra (MassBank) D2:5670 MS/MS spectra (GNPS) ([62])	ESI +/-	✓		✓	[74]
		GMD algorithm	DTs classification	MS spectra + RI		✓	EI-MS spectra of MeOX and TMS derivatives	EI	✓ (only if present in GMD as unknown)		✓	[75]
	COMBINATORIAL FRAGMENTATION	FT-BLAST	FT computation + FT alignment	MS/MS spectra		✓	D1:97 compounds D2:370 compounds D3:44 compounds D4:(poppy extracts)	ESI +/-			NA	[58]
		FiD	combinatorial search of substructures for product ions	MS/MS spectra			D1:20 amino acids and 17 sugar-phosphates D2:20 amino acids	ESI +/- and MALDI			✓	[76]
		MetFrag	FT generation + heuristics	MS/MS spectra	✓		D1:200 spectra/49 compounds (KEGG DB) D2:510 spectra/102 compounds (search against PubChem)	ESI			✓	[24]
		MetFrag v.2.2			✓		D1: 102 compounds D2:473 MS/MS spectrum/359 compounds	ESI			✓	[23]
		MAGMa	MST computation	MS^n spectra + candidate list	✓		D1:510 spectra/102 drug-like compounds D2:11 compounds	ESI -			✓	[34]

					experimental and *in silico* MS spectra)	EI-MS spectra (NIST 2017, replicates library)				
JOINT APPROACHES	Schymanski method	[26] + [24] + metadata	MS spectra	✓		(river water sample)	EI		NA	[44]
	MetFusion	[24] + MSL search	MS/MS spectra	✓		1099 spectra of metabolites	ESI		✓	[65]
	ChemDistiller	ML (MKL, SVMs) + combinatorial fragmentation + SVM	MS/MS spectra (MF as optional metadata)	✓		**5038 MS/MS spectra (sources: NIST14, MassBank, HMDB)** 1259 MS/MS spectra (NIST14, MassBank, HMDB)	ESI +/-		✓	[22]

spectrum and candidates are scored according to product ion assignment, neutral loss assignment, isotopic ratio [35,72], existence of the compound in an in-house [71] or external DB (PubChem), fragment linkages and hydrogen rearrangement rules [35].

Direct structure elucidation methodologies rely on the high correlation between FP similarity and structural similarity, but as it turns out, structurally similar compounds do not always generate similar fragments. The employed general fragmentation rules are often unable to explain compound-specific fragmentation mechanisms and complex rearrangement reactions, with the exceptions of MS-FINDER [35] and MassFrontier [71]. They are also non-exclusive, i.e. one rule affects other rules and *vice versa*. Importantly, the direct structure elucidation approaches are able to explain only unimolecular fragmentation reactions and are not able to predict peak intensities. Consequently, "barcode spectra" are generated, where the same intensity is assigned to all peaks. Therefore, rule-based fragmenters are principally used for *in silico* fragmentation of compound classes with consistent FPs (i.e. parabens) that follow known fragmentation rules. The interpretation is improved by higher spectral reproducibility and is therefore less challenging for EI-MS than for electrospray ionization -tandem mass spectrometry (ESI-MS/MS) spectra.

4.2. Indirect annotation approaches

4.2.1. Indirect substructure prediction and classification approaches

These approaches (Fig. 2, Table 1) were originally developed for EI–MS spectra [75] and later extended to ESI-MS/MS [58,74]. Compounds are classified according to the presence of specific substructures [74], functional groups [75] or complete fragmentation pathways [58] with [75] or without metadata [58,74]. UML methods predict the presence of substructures without prior knowledge of fragments of interest. One example is MS2LDA [74], which decomposes MS/MS spectra into blocks of co-occurring fragments and neutral losses, named Mass2Motifs. An expert then matches MFs of specific fragments and neutral losses with corresponding substructures. Subsequently, compounds are grouped using HCA according to the presence of substructures. The main disadvantages of using Mass2Motifs involve the need of expert knowledge for their characterization, and the inability to incorporate structural information inherent to peak correlations. The strong correlation of FP similarity and chemical similarity is used, such that FTs (Section 2) are employed to identify structural neighbors of unknown compounds ("Rasche method") or to aid DB/MSL search (FT-BLAST) [58]. In the "Rasche method", FTs of all compounds from the NTS run are aligned in an all-against-all pairwise manner and are grouped to compound classes according to FT similarity by using HCA.

The Golm Metabolome Database (GMD) algorithm is a SML method combining structural information for presence of specific functional groups inherent to EI-MS data (simple spectral features, such as *m/z* and intensity, intensity ratios and neutral losses) with normalized R_t data in mass spectral tags (MSTs). Unknowns are classified according to the presence of multiple functional groups by using decision trees (DTs) [75].

4.2.2. Indirect structure elucidation approaches

Indirect structure elucidation approaches (Fig. 2, Table 1) include combinatorial fragmenters [23,24,34,58–61,76], SML pattern-recognition methods [16–21,62] that use ESI-MS/MS spectra as input, and UML that uses EI-MS spectra [19].

Combinatorial fragmenters, such as FT-BLAST [58], MetFrag [23,24], Hufsky method [59], MIDAS [60] and MIDAS-G [61] rely on FT or MSTR computation (MAGMa) [34] (Section 2). The oldest combinatorial approach, FiD [76], is an exception. It conducts a

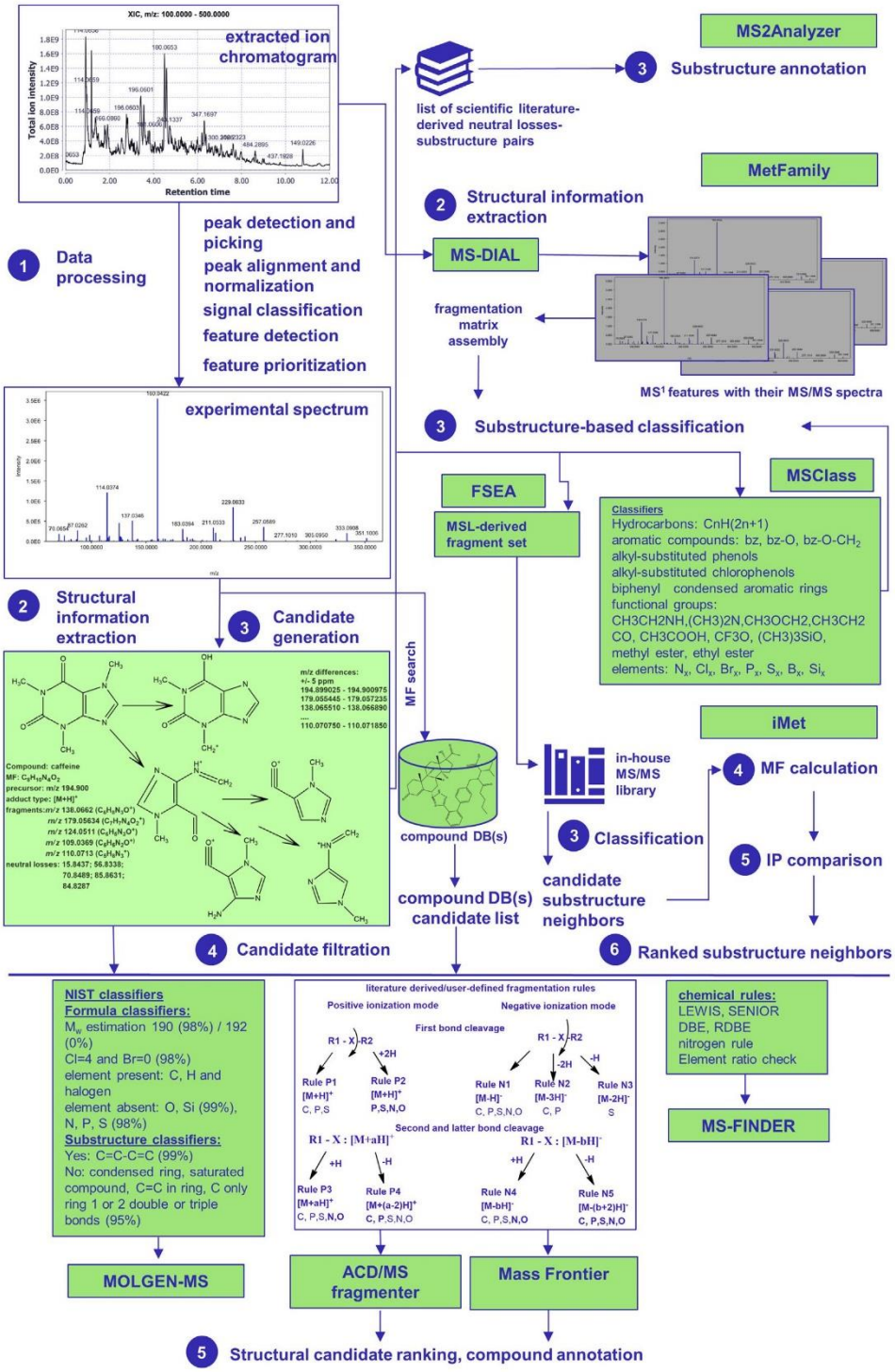


Fig. 1. An overview of direct substructure prediction and classification and structure elucidaiton approaches.

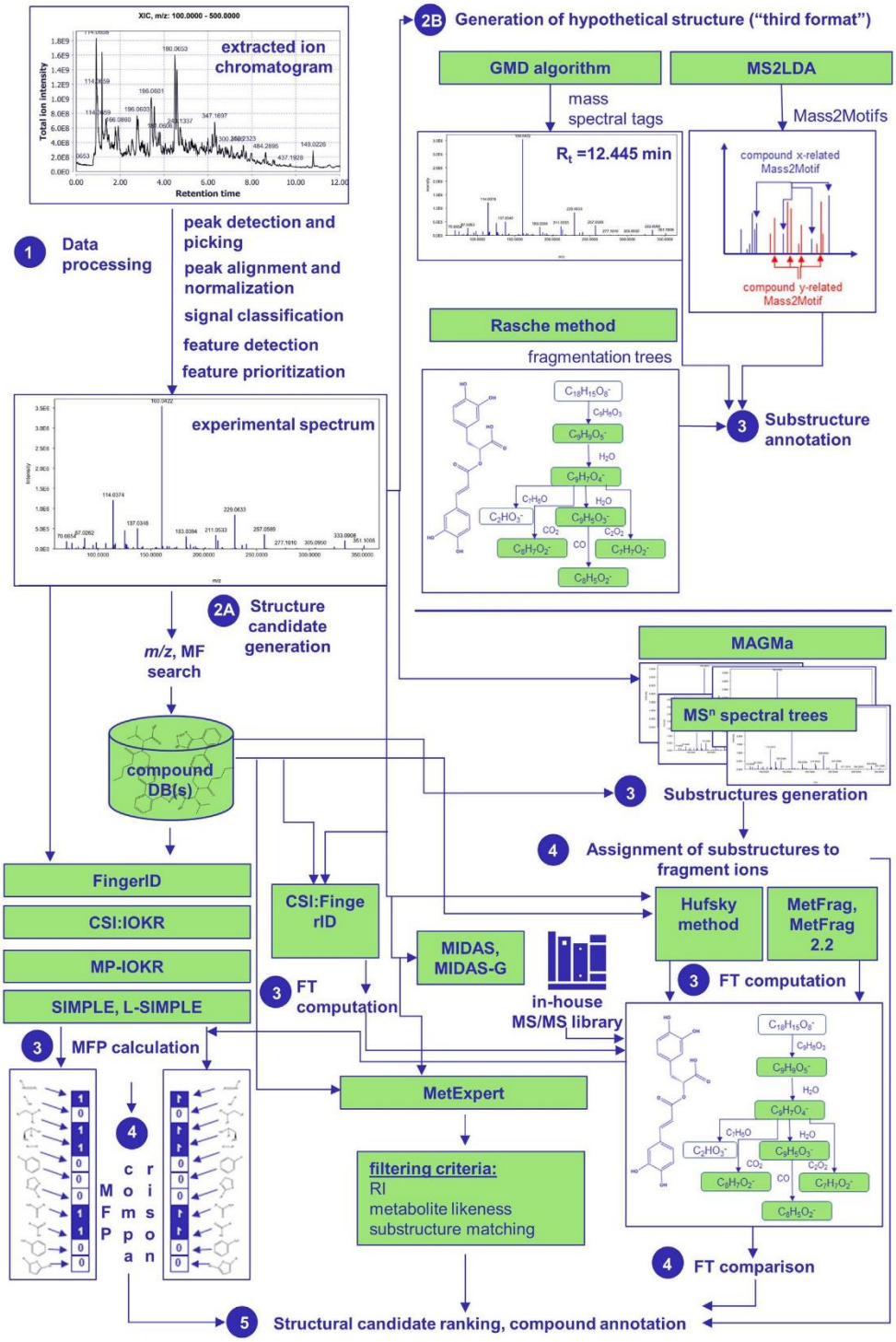


Fig. 2. An overview of indirect substructure prediction and classification and structure elucidaiton approaches.

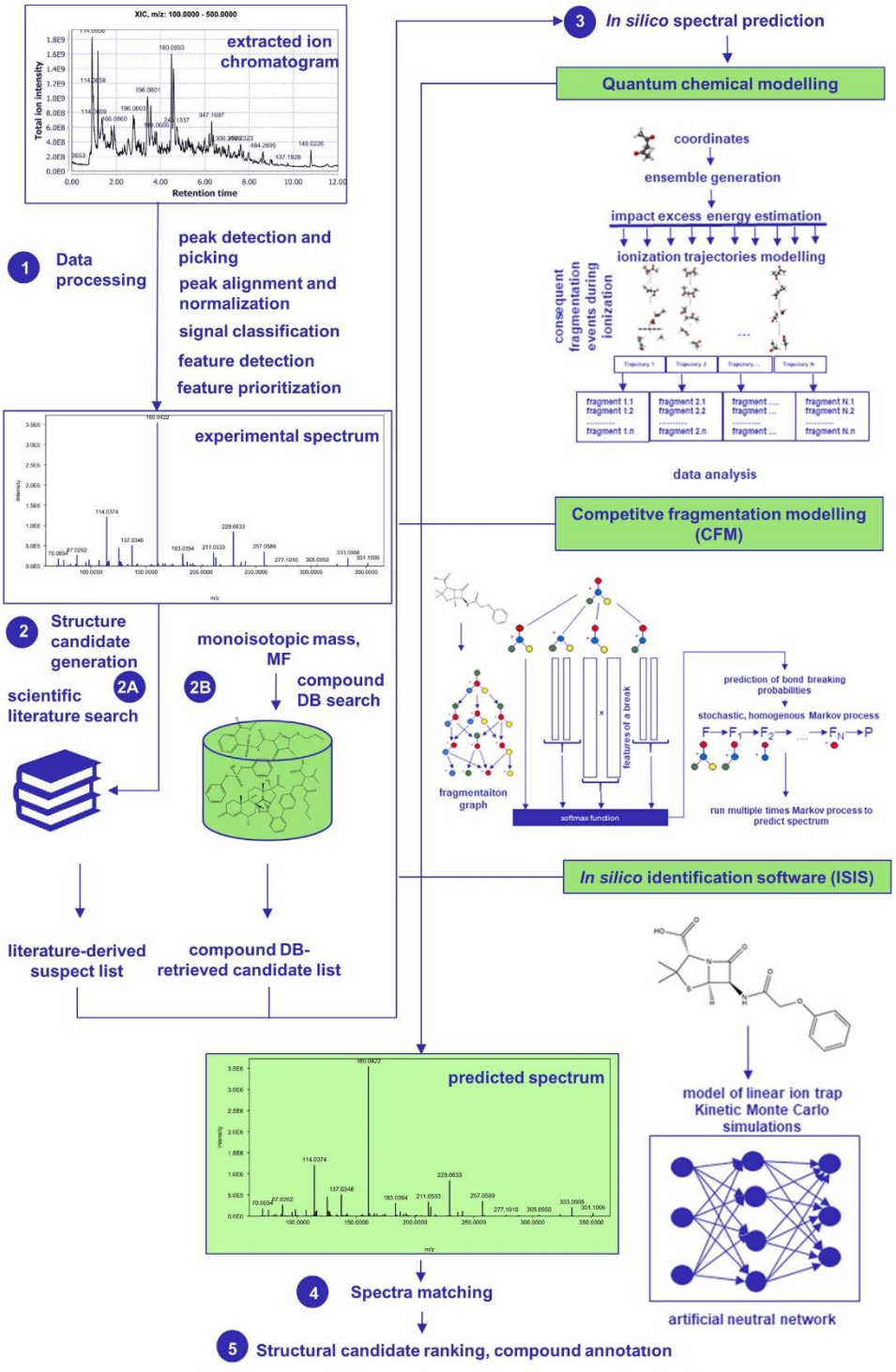


Fig. 3. An overview of direct *in silico* spectral prediction approaches for CA.

combinatorial search over all possible fragmentation pathways that explain the observed MS/MS spectrum and ranks them according to the energetic favorability of the fragmentations.

FTs are computed from MS/MS spectra for *de novo* CA [59], to aid MSL [58] or DB search to accomplish structure elucidation. FTs are computed for a query compound and for candidate compounds, which are retrieved from exact mass or MF search in PubChem, KEGG, Chemspider [23,24] or in-house DBs [60,61]. Candidates are scored by weighting FT edges according to their energetic plausibility, represented as bond dissociation energy (BDE) [23,24,65], that is, the energy required to generate a particular fragment. Nodes are scored based on the actual observation of their precursors' *m/z* and the number of dissociated bonds [60,61] or by using mass deviation and peak intensity [59], and weighted either with equal [60] or different weights [61], indicating equal or different generation possibility. The most plausible FT explains as many experimental peaks as possible and reflects the energetically most favorable fragmentation pathway. This set of approaches requires control over the "combinatorial explosion", which may come at the price of decreased precision and prediction of many unreasonable and energetically unfavorable fragments. Constraints used to decrease the number of candidates (for which FT are to be computed) include limiting the type and number of elements considered [24] and versatile metadata (reference counts, patent information, log K_{ow} [23] and RIs [23,24]). Improved performance of combinatorial fragmenters is reported when RI, $ECOM_{50}$ and drift time filtering of DB-retrieved candidates is performed as a step upstream [38]. More options to control FT generation include restricting the number and intensity of peaks to be considered [34], fragment redundancy and fragment plausibility check and limiting FT depth [23,24,60,61,65]. The drawback of the combinatorial fragmenters is that they omit crucial fragmentation aspects, such as rearrangement reactions on radical sites and charge distribution, which explains their somewhat lower identification performance.

MSTRs (Section 2) generated from multistage HR/AM-MS^n data aid DB-search for structure elucidation. At each MS level of the MSTR, observed peaks are explained by substructures from a hierarchical substructure tree selected according to their ability to explain the observed hierarchical FP. Considering the restrictions made by the assignment at the precursor level, and the consistent annotation on subsequent MS levels, MAGMa successfully prioritizes correct structures for a Pubchem retrieved candidate list [34].

SML approaches include MetExpert [19], FingerID [20], CSI:FingerID [18,62], CSI:IOKR [16], MP-IOKR [21], SIMPLE and L-SIMPLE [17]. They perform CA by learning relationships between MS/MS signals and molecular properties, considering both peak *m/z* and intensity. The presence or absence of specific structural properties from ESI-MS/MS data is detected and represented as bits (at specific positions) in binary vectors named molecular fingerprints (MFPs). MFPs can be predicted solely from MS/MS spectra [17,20] in order to omit the computationally heavy FT generation or from both FTs and MS/MS spectra [28,56]. Although faster, MFPs predicted only from MS/MS spectra extract rich structural information, but without explicitly modeling peak dependencies. Latest approaches [17] consider peak interactions through adding additional kernels (similarity functions encoding the structure of the data) to CSI:FingerID [18], and consequently, achieve higher identification accuracy. In any case, accuracies of all these approaches depend upon descriptiveness of the predicted MFPs. Another problem is the sparsity of MFPs. There are many possible substructures, so the MFPs are long, but each compound usually contains only a small subset of the substructures, and most substructures are only present in a small subset of compounds. This problem can be alleviated by finding and considering only peaks and peak interactions that contribute to improved predictive performance using sparse optimization models [17].

Employment of SVMs as SML classifiers increases the accuracy of MFP prediction, which eliminates the uncertainty rising from their dependency upon MS accuracy, and further improves annotation performance [18,20]. Once predicted, MFPs of a query compound are compared in a pairwise manner to MFPs of candidate compounds with the same MF retrieved from PubChem. Candidates are scored according to the similarity of their MFPs with the predicted MFPs [62]. Identification rates depend on the kernel's ability to describe MFP similarity, type and number of molecular properties included in MFPs and their uniqueness to particular sets of compounds. Often multiple kernels are used and combined with multiple kernel learning (MKL), to make use of as much information as possible.

Latest structured output prediction approaches [16,21] additionally use two different kernel functions in order to encode the similarities in the input (MS/MS spectra) space and the similarities in the output (molecular structures) space, thus omitting the prediction and scoring of MFPs as the intermediate step. MKL is used to learn compound similarities in the input space (MS/MS spectra), and kernels based on MFPs are used for similarity in the output space (molecular structures).

Finally, MFPs can be predicted from EI-MS spectra using partial least squares-discriminant analysis (PLS-DA) [19]. This knowledge is combined with metadata (predicted RIs, MF) and metabolite likeness score to filter candidates, further searched in a chemical space expanded by performing *in silico* trimethylsilyl (TMS) derivatization of existing DBs.

4.3. In silico *spectral prediction approaches*

In silico spectral prediction approaches learn to predict two-dimensional (*m/z* and intensity) EI-MS [77,81] or ESI-MS/MS [78–80] spectra by simulating fragmentation under ionization in a direct or indirect manner. Existing direct ML approaches (Fig. 3, Table 1) use chemical structures as input data for prediction of two-dimensional (*m/z* and intensity) ESI-MS/MS [78,80] and EI-MS spectra [77] of DB-retrieved candidates or for complete *de novo* generation of EI-MS spectra [81], and annotate compounds by comparing experimental and predicted MS spectra. Alternatively, MFPs calculated from chemical structures are used as input data for faster and more accurate *in silico* prediction of EI-MS spectra (Fig. 4, Table 1) [82].

Using these approaches, CA is performed by matching the experimental MS spectrum to an *in silico* MSL of predicted MS/MS spectra [77,78,80], alone or combined with a rule-based fragmentation model, metadata and chemical similarity check [79], and ranking the list of DB-retrieved candidates. All these approaches are computationally intensive, but they fail to realistically predict peak intensities and intensity ratios, which limits their use in EEA.

4.4. *Joint annotation approaches*

Direct and indirect CA are combined either in a filtering [22] or a consensus approaches [44,65] in order to achieve higher annotation accuracy (Fig. 5, Table 1). The consensus "Schymanski method" [44] combines MOLGEN-MS/MS [26], MetFrag [24], and metadata, such as steric energy calculations, log K_{ow} and RI. MetFusion [65] combines MetFrag [24] with MSL search. ChemDistiller [22] combines MFP [18,62], MetFrag [24] and a competitive fragmentation modelling (CFM)-like approach [78] with MF and metabolite likeness to filter DB candidate lists. In their launching studies (i.e. studies introducing them), these approaches outperformed all single-strategy CA approaches.

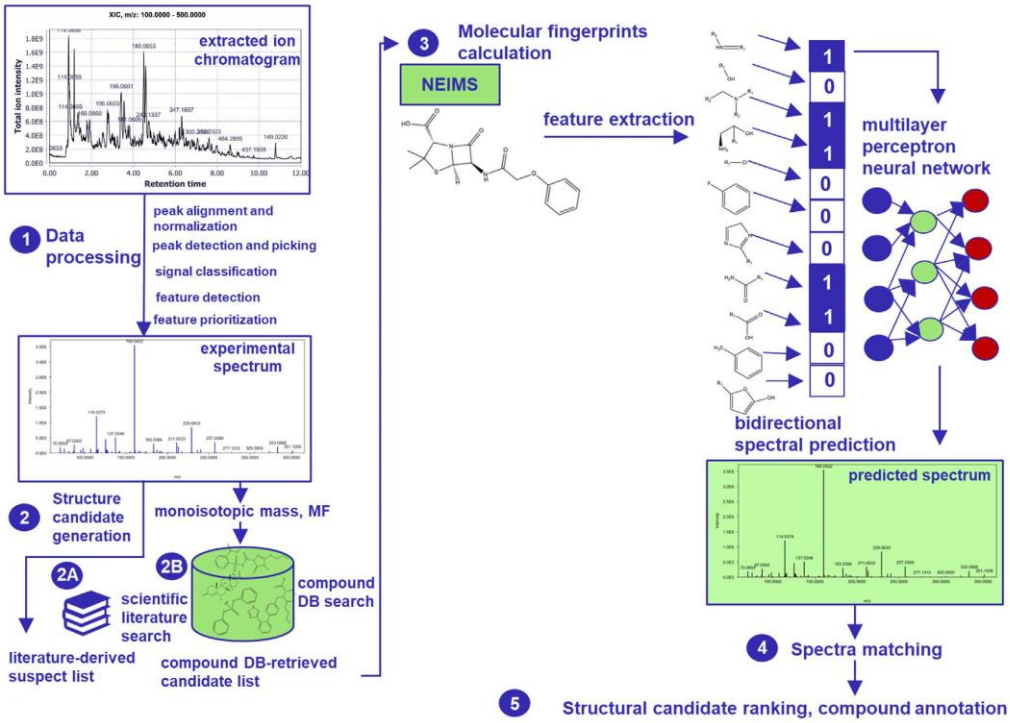


Fig. 4. An overview of indirect *in silico* spectral prediction approach for CA.

5. Employment of CA approaches in EEA

The employment of the CA approaches in EEA studies has so far been less exhaustive as compared to other 'omics' fields. Most of the published EEA studies perform EEA using workflows consisting of data processing and MF determination by vendor-specific software (e.g. Formula Finder function in Peak View™ by Sciex) and structure elucidation by comparing MS/MS spectra against publicly available (e.g. MassBank) or vendor-specific MSLs, metadata and expert knowledge. Thorough literature search revealed that MetFrag is the most frequently employed CA approach for annotation of EE constituents in wastewater [83], river water [84] and groundwater [85]. MetFrag was also used for identification of 72 unique structures in riverbank filtration system, 25 of which were identified in such compartment for the first time [86]. MetFrag has been used alone or in combination with MAGMa for EEA in river water [87], in combination with MOLGEN-MS/MS for EEA in lake sediments [88], or with MOLGEN-MS [44] or MetFusion and Mass Frontier [89] for EEA in river waters. CFM-ID is incorporated in NTS workflow of EPA's Non-Targeted Analysis Collaborative Trial (ENTACT) [90]. It is used to predict MS/MS spectra of compounds from Comptox Chemistry Dashbord, and use this *in silico* MSL for identification of "known knowns" [91].

6. Performance evaluation of cheminformatics approaches: current experience

Upon launch, every CA tool is compared to existing CA approaches. Results from such evaluation studies are variable and seldomly objective due to the lack of standardized evaluation metrics, benchmark datasets, computer codes and executables for automated evaluation. The CASMI contest is so far the only statistically robust comparison of the ability of cutting-edge CA approaches (developed between 2012 and 2017). Its goal was to determine the MF or the correct molecular structure for each challenge, which was an ESI-MS/MS spectrum [13]. Main performance metrics in the initial CASMI contests (2012–2014) included absolute ranking (AR), which is the sum of the number of candidates that have better or equal score as compared to the correct compound [14], relative mean and median rank, mean ranking position and relative ranking position (RRP), reflecting the position of the correct candidate relative to other candidates, and therefore method selectivity [15]. Approaches are then ranked according to the AR of correct solutions. Based on the experience gained between 2012 and 2014, the organizers of the 2016 and 2017 CASMI contests improved ranking metrics, so that the winners were determined according to the number of 'golds'. Gold is awarded to the contestant(s) with the best rank among the candidates submitted for the challenge, so that a winner for each challenge is determined even in the case when no method ranked the correct structure at top 1 [15].

Table 2 shows the number of top 1 rankings, mean AR and mean RRP values for the winning approaches from the CASMI contests 2012–2017. During this period, the winning approaches advanced from manual approaches relying on monoisotopic fragment ion and neutral loss formula analysis or DB search of monoisotopic mass/MF, through pre-processing workflows for automated spectral search, and finally to FT computation and MFP-based ML approaches. Overall, satisfactory performance of the contestant CA approaches is observed, recording constant progress with time. Top 1 ranks of winning approaches for MF assignment increase from 57 % in 2012 to 100 % in 2014, and for the molecular structure assignment from 17.7 % in 2012 to 68.6 % in 2014. The essential aim of CA approaches is to rank the correct structure as high as possible in the candidate list, and in view of their discriminative power, to

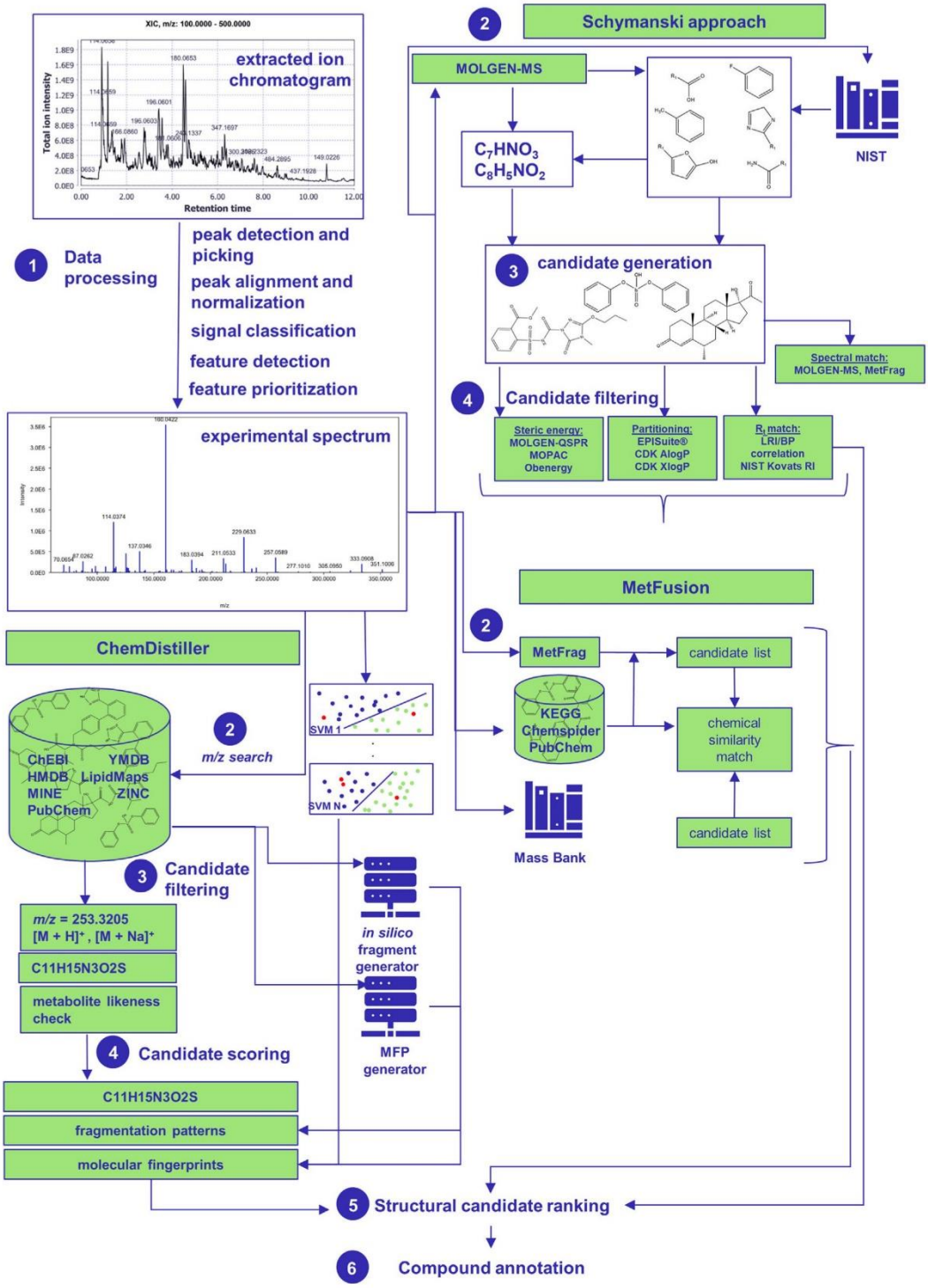


Fig. 5. An overview of joint structure elucidaiton approaches.

Table 2
CASMI contest results from 2012-2017. RRPs are calculated only when total number of candidates ≥2.

Year	2012											
Category	**Best MF LC/MS**			**Best identification LC/MS**			**Best MF GC/MS**			**Best identification GC/MS**		
Winning approach	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP
PUTMEDID-LCMS [25]	8/14	1.00	0.67									
Automated MSL search [14]				3/17	1.00	NA						
MOLGEN-MS [73]							15/16	1.19	0.97	8/20	3.50	0.88
Year	2013											
Category	**Best MF**						**Best structure identification**					
Winning approach	top 1		mean AR		mean RRP		top 1		mean AR		mean RRP	
Manual approach [95]	12/12		1.00		NA		14/15		1.00		0.95	
SIRIUS v2.0 [31]	10/12		1.00		NA							
Year	2014											
Category	**Best MF**						**Best structure identification**					
Winning approach	top 1		mean AR		mean RRP		top 1		mean AR		mean RRP	
MAGMa [34]	40/42		1.05		0.97							
MS-FINDER [35]	36/36		1.00		0.97							
CFM-ID [78]	38/39		1.16		NA		24/35		8.86		0.99	
Year	2016											
Category	**Best structural identification of natural products**						**Best automatic structural identification**					
							***In silico*fragmentation**			***In silico*fragmentation + MD**		
Winning approach	top 1		mean AR		mean RRP		top 1	mean AR	mean RRP	top 1	mean AR	mean RRP
manual approach [96]	14/18		1.22		0.79							
MSL search + MS-FINDER [35] + MetFrag [24] + SIRIUS v3.1.3 [33] + Seven Golden Rules	14/16		5.25		0.99							
CSI:IOKR [16]							62/208	127.34	0.99			
MS-FINDER [35] + MSL search										146/208	6.40	1.00
Year	2017											
Category	**Best structural identification of natural products**			**Best automatic structural identification In silico fragmentation**			**ISF + metadata**			**Best automated candidate ranking**		
Winning approach	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP	top 1	mean AR	mean RRP
CSI:FingerID [18]	11/41	235	0.96	77/239	235	1.00				66/198	658.28	0.91
MS-FINDER [35]+ metadata							91/200	4.33	0.90			

exclude as many false annotations as possible. In that spirit, a constant progress in mean RRPs has been observed for the MF assignment (from 0.67 to 0.97), while the RRPs of structure elucidation approaches have had constantly high values (0.88-1.00, Table 2).

The CASMI 2016 and 2017 results both show better performance of ML approaches when compared to combinatorial fragmenters, which is justified by the higher discriminatory power of spectral features extracted from MS/MS data. However, the latter are essential to cover cases when no training data is available for the ML approaches [15]. Stand-alone approaches in CASMI 2016 correctly identified 17–25 % of the challenges (CSI: FingerID was an exception with 34.4 %), and ranked the correct candidate in the top 10 in more than 49 % of the cases [15]. The post-contest analysis of CASMI 2016 data combined competing approaches with metadata and DB/MSL search. The results showed identification rates improved to 87–93 %, placing 98 % of correct identifications in top 10 rankings, and 59 % in top 1 in the first joint approach [22]. Adding multiple metadata to indirect ML approaches improved top 1 ranks up to 70 %. The importance of metadata is demonstrated, suggesting that once DB/MSL is used as metadata for identification of 'known unknowns', the choice of CA tool is less important [13]. This is further confirmed by CASMI 2017 results, where adding metadata improved top 1 ranks from 32.2 % to 44.6 %, and top 10 ranks from 61.5 % to 94 %.

The CASMI contests in 2016 and 2017 evaluated the performance of many CA approaches crucial for EEA (CSI: IOKR, CSI: FingerID, CFM-ID, MAGMa, MetFrag, MIDAS), with or without adding metadata (ChemSpider ID, R_t, reference counts, biological/environmental relevance) [13,15]. Many others, such as MetFusion, Rasche method, FT-BLAST, ISIS and Hufsky method have never participated. Also, numerous CA approaches, including iMet, ChemDistiller, MetExpert, MetFamily, MS2LDA, L-SIMPLE, MP-IOKR, iMet and SIMPLE were lauched after the last CASMI contest and have therefore not been evaluated. The MF assignment approaches have not been evaluated either.

Being the only platform for reliable comparison of CA approaches to date, CASMI constantly evolved over time. This developement was motivated by changing the designs of contest rules and challenges, in order to simplify the setting, excluding the influence of other processing steps as well as human experts on performance of the approaches. In addition, such designs controlled the variation possibly arising from laboratory preparation, instrumental analysis and data processing during real EEA. These include:

(1) *A priori defining the molecular species in the challenge ESI-MS/MS spectra.* This step is very error-prone and time-consuming in real EEA. *A priori* definition for example allows avoiding the charged species and metal adducts which are frequently formed during LC-HR/AM-MS.
(2) *Using a DB-retrieved candidate list (i.e., pre-defined candidate list, including the best scoring isomer for each candidate from the list) for each challenge in CASMI 2016 and 2017.* This rule indicates that the monoisotopic mass and/or MF of the challenge is already known, which is actually the crucial step in the real

Table 3
Performance of the CASMI 2016 participants for selected challenge ESI-MS/MS spectra of EE constituents. The rank of the correct structure is given in; the rank of the winning approach(es) of Category 2 is given in bold, the rank of the winning approach(es) of Category 3 is given in italics.

Compound name		N-acetylmesalazine	propoxycarbazone	Michler's ketone	medroxyprogesterone	diphenyl phosphate	valsartan
Chemical structure							
PubChem ID		65512	177355	7031	541103	13282	60846
InChIKey		GEFDRROBUCULOD-UHFFFAOYSA-N	JTHMVYBOQLDDIY-UHFFFAOYSA-N	VVBLNCFGVYUYGU-UHFFFAOYSA-N	FRQMUZJSZHZSGN-HBNHAYAOSA-N	ASMQGLCHMVWBQR-UHFFFAOYSA-N	ACWBQPMHZXGDFX-QFIPXVFZSA-N
COMPETING APPORACHES	MS-FINDER [35]	**40.5**	8.0 **1.0**	72.5	**1.0**	**1.0**	**1.0**
	CSI:FingerID [62]	-	- **1.0**	**1.0**	**1.0**	-	**1.0**
	MAGMa+ [35]	86.0	**1.0** **1.0**	1809.5	270.0	3.5	1.5
	CFM-ID [78] basic model/ retrained model	63.0/14.0	1170.5/1170.5 **1.0**/2.0	615.5/2.0	32.0/**1.0**	136.0/136.0	**1.0**
	CSI:IOKR [16]	75.0	972.0 15.0	3101.0	*1.0*	216.0	*1.0*
	MS-FINDER [35]+MSL search	18.0	*1.0* *1.0*	*1.0*	*1.0*	*3.0*	*1.0*

EEA. This way, the performance evaluation eliminates the impact of data source and is in fact limited to the ability to rank "known unknowns".

(3) *High overlap between training datasets of participating ML approaches and CASMI 2016 challenges.* Performance of ML approaches on data not present in the training dataset reflects actual SS and NTS conditions. Their somewhat overestimated success in CASMI 2016 [12,60,72] was observed, when considering the significantly lower identification accuracy (top 1 ranks) for the compounds absent from the training dataset, for both positive (12.3 %, 21.3 % and 34.7 % for CFM-ID, CSI:IOKR and CSI:FingerID, respectively) and negative (12.8 % for CFM-ID and 3.0 % for CSI:IOKR) ESI-MS/MS spectra. There was however no decrease for the top 10 ranks. As follows, CASMI organizers considered providing traning dataset for ML participants, with all CASMI challenges excluded from the dataset [15].

A general overview of CASMI results shows that accurate and confident CA depends not only on method's capability, but also on the type and quality of input data, size, specificity and comprehensiveness of the searched DBs and MSLs [19,39,40], and the used metadata. The examples, where CA approaches fail to correctly identify a compound can be generalized as: **(1)** compounds whose functional groups are prone to intramolecular rearrangements (i.e. nitro groups); **(2)** MS/MS spectra in which the molecular ion peak is absent or misidentified and in turn an erroneous exact mass and MF are assigned; **(3)** MS/MS spectra dominated by trivial neutral loss peak (e.g. Br loss, isopropyl group loss) and **(4)** in the case of constitutional isomers. Overall, it is seen that CA approaches are still in their early development and are yet to introduce independent practical use in annotation of the "known unknowns".

Given the lack of a standardized identification confidence assignment system, most recent EA studies use the 5-level system of Schymanski et al. [92], according to which all available CA approaches achieve confidence level 2 or level 3. At level 2, the probable structure is confirmed by MSL match (level 2a) or diagnostic evidence, i.e. diagnostic MS/MS fragments and ionization behavior (level 2b). At level 3, only tentative candidate(s) are identified by class and presence of specific substructures and substituents. Identification levels defined by Metabolomics Standards Initiative (MSI) [93] can also be used. New confidence classification systems [94] are yet to be employed in EEA.

6.1. The CASMI 2016 on EE constituents

We examined the performance of the CASMI 2016 participants on selected challenges that represent MS/MS spectra of well-known anthropogenic compounds, many of which are known EE constituents (Table 3). Detailed insight into the results showed that the performance of CA approaches depends on the input MS/MS spectra, but is also compound- and methodology-specific. For example, N-acetylmesalazine is human metabolite of the drug mesalazine, which gives the ESI(-) MS/MS spectrum that all approaches struggled with. As a rule, the ESI(-) ionization gives less informative MS spectra, which was also evident in this example, where there were only two intense peaks in the MS/MS spectrum, the deprotonated molecule and the cleavage of carboxyl group. Another issue is using an inappropriate ionization mode, which in turn again produces an uninformative MS/MS spectrum. This was shown in the case of porpoxycarbazone, a widely used pesticide usually analyzed using ESI(+) ionization, as it contains several moieties prone to protonation. Hence, its ESI(-) ionization was weak and it was ranked very low by ML approaches, whereas it was correctly identified by almost all approaches when using the ESI(+) MS/MS spectrum.

For some of the examples there was none of the evaluated approaches, even including metada, that would perform well. Such example is the Michler's ketone, an intermediate in the production of dyes and pigments, that has a symmetric structure. This compound yields the ESI(+) MS/MS spectrum dominated by a peak which does not represent the precursor ion. Here, drastic performance improvement was noticed when the CFM-ID approach [78] was retrained on METLIN and NIST data, which was unfortunately not the case with another symmetrical compound, diphenyl phosphate (Table 3). This could be a potential sign for overfitting or simple data fluctuations [15]. The diphenyl phosphate, an industrial chemical used in production of coating products and adhesives, is however a very rare example of where the combinatorial fragmenters were most accurate.

ML approaches perform well for compounds with MS/MS spectrum typical for a compound class. Such example is medroxyprogesterone, a steroid hormone frequently detected in surface and wastewaters, with a typical steroid spectrum, where all ML approaches performed well (Table 3). Comaprative performance is recorded for annotation of 10-azabenzo[a]pyrene, mutagenic azapolycyclic aromatic hydrocarbon and lauric isopropanolamide, a surfactant in cosmetics (data not shown). Interestingly, all approaches ranked among top 3 only the correct structures of long established EE constituents, including pesticides (flufenoxuron, ethion, pyrazophos, pirimiphos-methtyl, triadimenol, fenthion, tris (2-chloroethyl)phosphate) and drugs (captopril, candersartan, valsartan, loperamide and clotrimazole), described by multiple citations, MS/MS spectra in MSLs and class-specific FPs.

7. Future advancements and conclusions

To further improve the capabilities of cheminformatics approaches in EEA, advancements should tackle the following obstacles:

- Annotation of common degeneracies present in most thoroughly investigated samples during EEA (wastewater effluent, river water, sea water, drinking water, human serum, blood, urine, saliva etc.) is not performed to date. Generation of a public DB for their open reposition and use, following recent initiatives in metabolomics [10], would reduce initial MS data to a subset of features corresponding to unknown small molecules to be identified using cheminfromatics approaches.
- The higher number of entries in available molecular structure DBs and the tendency of faster DB enlargement (corresponding to the constantly increasing number of EE constituents) as compared to MSLs, suggests further development of DB-based CA approaches, rather than MSL-based ones. The danger of redundant results originating from the existence of duplicate and/or low-quality entries in DBs should be avoided by using compound DBs of high quality. Following the model of the recently implemented spectral identifier SPectraL Hash (SPLASH), the duplicates can be removed and structural representations and identifiers in DB entries can be unified.
- Poor performance of MSL-based approaches arising from low-quality spectra could be improved by implementing a spectral quality filtering system, based on statistically determined acceptable variability of monoisotopic and isotope peaks' intensities and isotope intensity ratios. Each spectrum should be annotated with SPLASH, so that the same spectra from different MSLs can be matched.
- Limited searchable chemical space of existing DBs and MSLs for EE constituents is the reason that many compounds remain unidentified during EA ($\geq$70 %). We propose its expansion by generating *in silico* compound DBs built of compounds predicted by simulating transformation and degradation reactions of known and persistent EE constituents from exisiting DBs using cheminformatics approaches for prediction of transformation and degradation pathways [52–55] or by their *in silico* chemical modification (i.e. derivatization), as performed in [19]. Performance evaluation of existing CA approaches should then be conducted on these *in silico* created DBs and MSLs.
- For certain compound classes, still unsatisfactory CA performance can be improved by joining many existing CA approaches and metadata in consensus approaches. Potential solutions include combination of a direct and indirect annotation approach, or each of them with a classification approach, DB search, heuristics, and filtering approaches for retrospective analysis of results from each CA tool for confirmation of presence of structurally similar compounds (TPs, impurities). Multiple metadata addition (e.g. RI, drift time) from complementary analytical techniques (GC/LC–MS, NRM, IM-MS) to existing CA approaches is also shown to be beneficial [19,38].
- At the moment, there exists no common platform that would enable one to follow the development of CA approaches, familiarize oneself with their properties and select the most applicable tool for a selected type of compound and data. In this respect, a single repository of all MF assignment and CA approaches should be built. We believe this would stimulate their use, initiate their harmonization, improve their accessibility and ensure their appropriate use and the reproducibility of their results.
- There is a lack of standardized methods to evaluate the performance of CA approaches. This should be solved by the establishment of a standardized system of evaluation metrics that should include relevant parameters (specificity, sensitivity, positive predictive value, negative predictive value, accuracy and false discovery rate). Benchmark datasets of high-quality GC-MS and LC-MS/MS spectra acquired on different analytical platforms, i.e. "gold spectral dataset standard" should be made publicly available, as it was done for ML-based R_t prediction recently [97].

This review, along with numerous recent evaluations [13–15] confirms the importance of cheminformatics in EEA. Current achievements confirm that cheminformatics approaches are able to increase identification confidence for "known unknowns" and dictate directions for annotation of "unknown unknowns". Beneficial HR/AM-MS instrumentation advancements, coupling GC- or LC-HR/AM-MS to IM-MS or NMR in a hybrid approach to resolve isomer structures, employment of well-established DBs and MSLs and implementation of established R_t prediction models will significantly contribute to the improvement of CA performance.

Further improvement of ML methods is the key to EE characterization. They are a starting point for changing the priorities in CA: to extract as much knowledge as possible from the MS data without the necessity to identify the exact structure. In that way, development of cheminformatics approaches will be relieved from the expectations to replace the use of experts' competences, experience and intuition along the whole CA process and final identity confirmation using reference standards. Cheminformatics will respond to the challenge of EEA only by considering the host of drawbacks and by directing improvement towards their alleviaiton. In that way, the immense efforts invested in their development will be outweighed by their ability of thorough EE characterization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Slovenian Research Agency (Program Groups P1-0143 and P2-0103) and the European Union's Horizon 2020 research and innovation programme HBM4EU under Grant Agreement No. 733032. M.L. is funded by the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia (contract no. 11011-85/2016).

References

[1] C.P. Wild, Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology, Cancer Epidemiol. Biomarkers Prev. 14 (2005) 1847–1850, doi:http://dx.doi.org/10.1158/1055-9965.EPI-05-0456.
[2] G.W. Miller, D.P. Jones, The nature of nurture: refining the definition of the exposome, Toxicol. Sci. 137 (2014) 1–2, doi:http://dx.doi.org/10.1093/toxsci/kft251.
[3] N.R. Council, Exposure Science in the 21st Century: A Vision and a Strategy, The National Academies Press, Washington, DC, 2012, doi:http://dx.doi.org/10.17226/13507.
[4] B.L. Milman, I.K. Zhurkovich, The chemical space for non-target analysis, TrAC Trends Anal. Chem. 97 (2017) 179–187, doi:http://dx.doi.org/10.1016/j.trac.2017.09.013.
[5] A.J. Williams, C.M. Grulke, J. Edwards, A.D. McEachran, K. Mansouri, N.C. Baker, G. Patlewicz, I. Shah, J.F. Wambaugh, R.S. Judson, A.M. Richard, The comptox chemistry dashboard: a community data resource for environmental chemistry, J. Cheminformatics. 9 (2017) 61, doi:http://dx.doi.org/10.1186/s13321-017-0247-6.
[6] ContaminantDB, (2019). https://contaminantdb.ca/ (accessed July 23, 2020).
[7] D. Wishart, D. Arndt, A. Pon, T. Sajed, A.C. Guo, Y. Djoumbou, C. Knox, M. Wilson, Y. Liang, J. Grant, Y. Liu, S.A. Goldansaz, S.M. Rappaport, T3DB: the toxic exposome database, Nucleic Acids Res. 43 (2015) D928–D934, doi:http://dx.doi.org/10.1093/nar/gku1004.
[8] V. Neveu, G. Nicolas, R.M. Salek, D.S. Wishart, A. Scalbert, Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk, Nucleic Acids Res. (2019), doi:http://dx.doi.org/10.1093/nar/gkz1009 gkz1009.
[9] M. Sindelar, G.J. Patti, Chemical discovery in the era of metabolomics, J. Am. Chem. Soc. 142 (2020) 9097–9105, doi:http://dx.doi.org/10.1021/jacs.9b13198.
[10] N.G. Mahieu, G.J. Patti, Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites, Anal. Chem. 89 (2017) 10397–10406, doi:http://dx.doi.org/10.1021/acs.analchem.7b02380.
[11] X. Domingo-Almenara, J.R. Montenegro-Burke, H.P. Benton, G. Siuzdak, Annotation: a computational solution for streamlining metabolomics analysis, Anal. Chem. 90 (2018) 480–489, doi:http://dx.doi.org/10.1021/acs.analchem.7b03929.
[12] B.B. Misra, S. Mohapatra, Tools and resources for metabolomics research community: a 2017–2018 update, Electrophoresis. 40 (2019) 227–246, doi:http://dx.doi.org/10.1002/elps.201800428.
[13] I. Blaženović, T. Kind, H. Torbašinović, S. Obrenović, S.S. Mehta, H. Tsugawa, T. Wermuth, N. Schauer, M. Jahn, R. Biedendieck, D. Jahn, O. Fiehn, Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy, J. Cheminformatics. 9 (2017) 32, doi:http://dx.doi.org/10.1186/s13321-017-0219-x.
[14] E.L. Schymanski, S. Neumann, CASMI: And the Winner is . . . , Metabolites 3 (2013) 412–439, doi:http://dx.doi.org/10.3390/metabo3020412.
[15] E.L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, S. Neumann, Critical assessment of small molecule identification 2016: automated methods, J. Cheminformatics. 9 (2017) 22, doi:http://dx.doi.org/10.1186/s13321-017-0207-1.
[16] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, J. Rousu, Fast metabolite identification with input output kernel regression, Bioinformatics. 32 (2016) i28–i36, doi:http://dx.doi.org/10.1093/bioinformatics/btw246.
[17] D.H. Nguyen, C.H. Nguyen, H. Mamitsuka, SIMPLE: sparse interaction model over peaks of molecules for fast, interpretable metabolite identification from tandem mass spectra, Bioinformatics. 34 (2018) i323–i332, doi:http://dx.doi.org/10.1093/bioinformatics/bty252.
[18] H. Shen, K. Dührkop, S. Böcker, J. Rousu, Metabolite identification through multiple kernel learning on fragmentation trees, Bioinformatics. 30 (2014) i157–i164, doi:http://dx.doi.org/10.1093/bioinformatics/btu275.
[19] F. Qiu, Z. Lei, L.W. Sumner, MetExpert: An expert system to enhance gas chromatography-mass spectrometry-based metabolite identifications, Anal. Chim. Acta. 1037 (2018) 316–326, doi:http://dx.doi.org/10.1016/j.aca.2018.03.052.
[20] M. Heinonen, H. Shen, N. Zamboni, J. Rousu, Metabolite identification and molecular fingerprint prediction through machine learning, Bioinformatics. 28 (2012) 2333–2341, doi:http://dx.doi.org/10.1093/bioinformatics/bts437.
[21] C. Brouard, E. Bach, S. Bocker, J. Rousu, Magnitude-preserving ranking for structured outputs, Proc. Mach. Learn. Res. 77 (2017) 407–422.
[22] I. Laponogov, N. Sadawi, D. Galea, R. Mirnezami, K.A. Veselkov, ChemDistiller: an engine for metabolite annotation in mass spectrometry, Bioinformatics. 34 (2018) 2096–2102, doi:http://dx.doi.org/10.1093/bioinformatics/bty080.
[23] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminformatics. 8 (2016) 3, doi:http://dx.doi.org/10.1186/s13321-016-0115-9.
[24] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinformatics. 11 (2010) 148, doi:http://dx.doi.org/10.1186/1471-2105-11-148.
[25] M. Brown, D.C. Wedge, R. Goodacre, D.B. Kell, P.N. Baker, L.C. Kenny, M.A. Mamas, L. Neyses, W.B. Dunn, Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, Bioinformatics. 27 (2011) 1108–1112, doi:http://dx.doi.org/10.1093/bioinformatics/btr079.
[26] M. Meringer, S. Reinker, J. Zhang, A. Muller, MS/MS data improves automated determination of molecular formulas by mass spectrometry, Commun. Math. Comput. Chem. 65 (2011) 259–290.
[27] S. Böcker, M.C. Letzel, Z. Lipták, A. Pervukhin, SIRIUS: decomposing isotope patterns for metabolite identification, Bioinformatics. 25 (2009) 218–224, doi:http://dx.doi.org/10.1093/bioinformatics/btn603.
[28] S. Böcker, F. Rasche, Towards de novo identification of metabolites by analyzing tandem mass spectra, Bioinformatics. 24 (2008) i49–i55, doi:http://dx.doi.org/10.1093/bioinformatics/btn270.
[29] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics. 8 (2007) 105, doi:http://dx.doi.org/10.1186/1471-2105-8-105.
[30] F. Rasche, A. Svatoš, R.K. Maddula, C. Böttcher, S. Böcker, Computing fragmentation trees from tandem mass spectrometry data, Anal. Chem. 83 (2011) 1243–1251, doi:http://dx.doi.org/10.1021/ac101825k.
[31] K. Dührkop, F. Hufsky, S. Böcker, Molecular formula identification using isotope pattern analysis and calculation of fragmentation trees, Mass Spectrom. 3 (2014) S0037, doi:http://dx.doi.org/10.5702/massspectrometry.S0037.
[32] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, Nat. Methods. 16 (2019) 299–302, doi:http://dx.doi.org/10.1038/s41592-019-0344-8.
[33] S. Böcker, K. Dührkop, Fragmentation trees reloaded, J. Cheminformatics. 8 (2016), doi:http://dx.doi.org/10.1186/s13321-016-0116-8.
[34] L. Ridder, J.J.J. van der Hooft, S. Verhoeven, R.C.H. de Vos, R. van Schaik, J. Vervoort, Substructure-based annotation of high-resolution multistage MSn spectral trees, Rapid Commun. Mass Spectrom. 26 (2012) 2461–2471, doi:http://dx.doi.org/10.1002/rcm6364.
[35] H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, T. Cajka, K. Saito, O. Fiehn, M. Arita, Hydrogen rearrangement rules: computational ms/ms fragmentation and structure elucidation using ms-finder software, Anal. Chem. 88 (2016) 7946–7958, doi:http://dx.doi.org/10.1021/acs.analchem.6b00770.
[36] M. Rojas-Chertó, P.T. Kasper, E.L. Willighagen, R.J. Vreeken, T. Hankemeier, T.H. Reijmers, Elemental composition determination based on MSn, Bioinformatics. 27 (2011) 2376–2383, doi:http://dx.doi.org/10.1093/bioinformatics/btr409.
[37] D.J. Creek, A. Jankevics, R. Breitling, D.G. Watson, M.P. Barrett, K.E.V. Burgess, Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: improved metabolite identification by retention time prediction, Anal. Chem. 83 (2011) 8703–8710, doi:http://dx.doi.org/10.1021/ac2021823.
[38] L.C. Menikarachchi, S. Cawley, D.W. Hill, L.M. Hall, L. Hall, S. Lai, J. Wilder, D.F. Grant, MolFind: a software package enabling hplc/ms-based identification of unknown chemical structures, Anal. Chem. 84 (2012) 9388–9394, doi:http://dx.doi.org/10.1021/ac302048x.
[39] A.M. Wolfer, S. Lozano, T. Umbdenstock, V. Croixmarie, A. Arrault, P. Vayer, UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling, Metabolomics. 12 (2016) 8, doi:http://dx.doi.org/10.1007/s11306-015-0888-2.
[40] P. Bonini, T. Kind, H. Tsugawa, D.K. Barupal, O. Fiehn, Retip: retention time prediction for compound annotation in untargeted metabolomics, Anal. Chem. 92 (2020) 7515–7522, doi:http://dx.doi.org/10.1021/acs.analchem.9b05765.
[41] E. Bach, S. Szedmak, C. Brouard, S. Böcker, J. Rousu, Liquid-chromatography retention order prediction for metabolite identification, Bioinformatics. 34 (2018) i875–i883, doi:http://dx.doi.org/10.1093/bioinformatics/bty590.
[42] J. Stanstrup, S. Neumann, U. Vrhovšek, PredRet: prediction of retention time by direct mapping between multiple chromatographic Systems, Anal. Chem. 87 (2015) 9421–9428, doi:http://dx.doi.org/10.1021/acs.analchem.5b02287.
[43] M.C. Campos-Mañas, P. Plaza-Bolaños, A.B. Martínez-Piernas, J.A. Sánchez-Pérez, A. Agüera, Determination of pesticide levels in wastewater from an agro-food industry: target, suspect and transformation product analysis, Chemosphere. 232 (2019) 152–163, doi:http://dx.doi.org/10.1016/j.chemosphere.2019.05.147.
[44] E.L. Schymanski, C.M.J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack, Consensus structure elucidation combining gc/ei-ms, structure generation, and calculated properties, Anal. Chem. 84 (2012) 3287–3295, doi:http://dx.doi.org/10.1021/ac203471y.

[45] L.C. Menikarachchi, R. Dubey, D.W. Hill, D.N. Brush, D.F. Grant, Development of database assisted structure identification (dasi) methods for nontargeted metabolomics, Metabolites. 6 (2016) 17, doi:http://dx.doi.org/10.3390/metabo6020017.

[46] W. Brack, S. Ait-Aissa, R.M. Burgess, W. Busch, N. Creusot, C. Di Paolo, B.I. Escher, L. Mark Hewitt, K. Hilscherova, J. Hollender, H. Hollert, W. Jonker, J. Kool, M. Lamoree, M. Muschket, S. Neumann, P. Rostkowski, C. Ruttkies, J. Schollee, E.L. Schymanski, T. Schulze, T.-B. Seiler, A.J. Tindall, G. De Aragão Umbuzeiro, B. Vrana, M. Krauss, Effect-directed analysis supporting monitoring of aquatic environments — An in-depth overview, Sci. Total Environ. 544 (2016) 1073–1118, doi:http://dx.doi.org/10.1016/j.scitotenv.2015.11.102.

[47] J.M. Mitchell, T.W.-M. Fan, A.N. Lane, H.N.B. Moseley, Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics, Front. Genet. 5 (2014), doi:http://dx.doi.org/10.3389/fgene.2014.00237.

[48] G.J. Myatt, E. Ahlberg, Y. Akahori, D. Allen, A. Amberg, L.T. Anger, A. Aptula, S. Auerbach, L. Beilke, P. Bellion, R. Benigni, J. Bercu, E.D. Booth, D. Bower, A. Brigo, N. Burden, Z. Cammerer, M.T.D. Cronin, K.P. Cross, L. Custer, M. Dettwiler, K. Dobo, K.A. Ford, M.C. Fortin, S.E. Gad-McDonald, N. Gellatly, V. Gervais, K.P. Glover, S. Glowienke, J. Van Gompel, S. Gutsell, B. Hardy, J.S. Harvey, J. Hillegass, M. Honma, J.-H. Hsieh, C.-W. Hsu, K. Hughes, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.O. Kenyon, M.T. Kim, N.L. Kruhlak, S.A. Kulkarni, K. Kümmerer, P. Leavitt, B. Majer, S. Masten, S. Miller, J. Moser, M. Mumtaz, W. Muster, L. Neilson, T.I. Oprea, G. Patlewicz, A. Paulino, E. Lo Piparo, M. Powley, D.P. Quigley, M.V. Reddy, A.-N. Richarz, P. Ruiz, B. Schilter, R. Serafimova, W. Simpson, L. Stavitskaya, R. Stidl, D. Suarez-Rodriguez, D.T. Szabo, A. Teasdale, A. Trejo-Martin, J.-P. Valentin, A. Vuorinen, B.A. Wall, P. Watts, A.T. White, J. Wichard, K.L. Witt, A. Woolley, D. Woolley, C. Zwickl, C. Hasselgren, In silico toxicology protocols, Regul. Toxicol. Pharmacol. 96 (2018) 1–17, doi:http://dx.doi.org/10.1016/j.yrtph.2018.04.014.

[49] M. Nendza, S. Gabbert, R. Kühne, A. Lombardo, A. Roncaglioni, E. Benfenati, R. Benigni, C. Bossa, S. Strempel, M. Scheringer, A. Fernández, R. Rallo, F. Giralt, S. Dimitrov, O. Mekenyan, F. Bringezu, G. Schüürmann, A comparative survey of chemistry-driven in silico methods to identify hazardous substances under REACH, Regul. Toxicol. Pharmacol. 66 (2013) 301–314, doi:http://dx.doi.org/10.1016/j.yrtph.2013.05.007.

[50] J.E. Rager, M.J. Strynar, S. Liang, R.L. McMahen, A.M. Richard, C.M. Grulke, J.F. Wambaugh, K.K. Isaacs, R. Judson, A.J. Williams, J.R. Sobus, Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, Environ. Int. 88 (2016) 269–280, doi:http://dx.doi.org/10.1016/j.envint.2015.12.008.

[51] H.P. Singer, A.E. Wössner, C.S. McArdell, K. Fenner, Rapid screening for exposure to “non-target” pharmaceuticals from wastewater effluents by combining hrms-based suspect screening and exposure modeling, Environ. Sci. Technol. 50 (2016) 6698–6707, doi:http://dx.doi.org/10.1021/acs.est.5b03332.

[52] J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D. Latino, S. Kramer, K. Fenner, enviPath – The environmental contaminant biotransformation pathway resource, Nucleic Acids Res. 44 (2016) D502–D508, doi:http://dx.doi.org/10.1093/nar/gkv1229.

[53] Y. Moriya, D. Shigemizu, M. Hattori, T. Tokimatsu, M. Kotera, S. Goto, M. Kanehisa, PathPred: an enzyme-catalyzed metabolic pathway prediction server, Nucleic Acids Res. 38 (2010) W138–W143, doi:http://dx.doi.org/10.1093/nar/gkq318.

[54] J.G. Jeffryes, R.L. Colastani, M. Elbadawi-Sidhu, T. Kind, T.D. Niehaus, L.J. Broadbelt, A.D. Hanson, O. Fiehn, K.E.J. Tyo, C.S. Henry, MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics, J. Cheminformatics. 7 (2015), doi:http://dx.doi.org/10.1186/s13321-015-0087-1.

[55] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D.S. Wishart, BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification, J. Cheminformatics. 11 (2019) 2, doi:http://dx.doi.org/10.1186/s13321-018-0324-5.

[56] S. Huntscha, T.B. Hofstetter, E.L. Schymanski, S. Spahr, J. Hollender, Biotransformation of Benzotriazoles: insights from transformation product identification and compound-specific isotope analysis, Environ. Sci. Technol. 48 (2014) 4435–4443, doi:http://dx.doi.org/10.1021/es405694z.

[57] T. Mairinger, T.J. Causon, S. Hann, The potential of ion mobility–mass spectrometry for non-targeted metabolomics, Curr. Opin. Chem. Biol. 42 (2018) 9–15, doi:http://dx.doi.org/10.1016/j.cbpa.2017.10.015.

[58] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, S. Böcker, Identifying the unknowns by aligning fragmentation Trees, Anal. Chem. 84 (2012) 3417–3426, doi:http://dx.doi.org/10.1021/ac300304u.

[59] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, S. Böcker, De novo analysis of electron impact mass spectra using fragmentation trees, Anal. Chim. Acta. 739 (2012) 67–76, doi:http://dx.doi.org/10.1016/j.aca.2012.06.021.

[60] Y. Wang, G. Kora, B.P. Bowen, C. Pan, MIDAS: a database-searching algorithm for metabolite identification in metabolomics, Anal. Chem. 86 (2014) 9496–9503, doi:http://dx.doi.org/10.1021/ac5014783.

[61] Y. Wang, X. Wang, X. Zeng, MIDAS-G: a computational platform for investigating fragmentation rules of tandem mass spectrometry in metabolomics, Metabolomics. 13 (2017) 116, doi:http://dx.doi.org/10.1007/s11306-017-1258-z.

[62] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, Searching molecular structure databases with tandem mass spectra using CSI:FingerID, Proc. Natl. Acad. Sci. 112 (2015) 12580–12585, doi:http://dx.doi.org/10.1073/pnas.1509788112.

[63] F. Hufsky, S. Böcker, Mining molecular structure databases: identification of small molecules based on fragmentation mass spectrometry data, Mass Spectrom, Rev. 36 (2017) 624–633, doi:http://dx.doi.org/10.1002/mas.21489.

[64] F. Hufsky, K. Scheubert, S. Böcker, Computational mass spectrometry for small-molecule fragmentation, TrAC Trends Anal. Chem. 53 (2014) 41–48, doi:http://dx.doi.org/10.1016/j.trac.2013.09.008.

[65] M. Gerlich, S. Neumann, MetFusion: integration of compound identification strategies, J. Mass Spectrom. 48 (2013) 291–298, doi:http://dx.doi.org/10.1002/jms.3123.

[66] K. Varmuza, W. Werther, Mass spectral classifiers for supporting systematic structure elucidation, J. Chem. Inf. Comput. Sci. 36 (1996) 323–333, doi:http://dx.doi.org/10.1021/ci9501406.

[67] Y. Ma, T. Kind, D. Yang, C. Leon, O. Fiehn, MS2Analyzer: a software for small molecule substructure annotations from accurate tandem mass spectra, Anal. Chem. 86 (2014) 10724–10731, doi:http://dx.doi.org/10.1021/ac502818e.

[68] A. Aguilar-Mogas, M. Sales-Pardo, M. Navarro, R. Guimerà, O. Yanes, iMet: a network-based computational tool to assist in the annotation of metabolites from tandem mass spectra, Anal. Chem. 89 (2017) 3474–3482, doi:http://dx.doi.org/10.1021/acs.analchem.6b04512.

[69] H. Treutler, H. Tsugawa, A. Porzel, K. Gorzolka, A. Tissier, S. Neumann, G.U. Balcke, Discovering regulated metabolite families in untargeted metabolomics studies, Anal. Chem. 88 (2016) 8082–8090, doi:http://dx.doi.org/10.1021/acs.analchem.6b01569.

[70] H. Tsugawa, R. Nakabayashi, T. Mori, Y. Yamada, M. Takahashi, A. Rai, R. Sugiyama, H. Yamamoto, T. Nakaya, M. Yamazaki, R. Kooke, J.A. Bac-Molenaar, N. Oztolan-Erol, J.J.B. Keurentjes, M. Arita, K. Saito, A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms, Nat. Methods. 16 (2019) 295–298, doi:http://dx.doi.org/10.1038/s41592-019-0358-2.

[71] Mass Frontier Software, Thermo Fischer Scientific Inc, (2018).

[72] ACD/MS Fragmenter, ACD/Labs, (2018).

[73] E.L. Schymanski, C. Meinert, M. Meringer, W. Brack, The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis, Anal. Chim. Acta. 615 (2008) 136–147, doi:http://dx.doi.org/10.1016/j.aca.2008.03.060.

[74] J.J.J. van der Hooft, J. Wandy, M.P. Barrett, K.E.V. Burgess, S. Rogers, Topic modeling for untargeted substructure exploration in metabolomics, Proc. Natl. Acad. Sci. 113 (2016) 13738–13743, doi:http://dx.doi.org/10.1073/pnas.1608041113.

[75] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, Decision tree supported substructure prediction of metabolites from GC-MS profiles, Metabolomics 6 (2010) 322–333, doi:http://dx.doi.org/10.1007/s11306-010-0198-7.

[76] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R.A. Ketola, J. Rousu, FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data, rapid commun, Mass Spectrom. 22 (2008) 3043–3052, doi:http://dx.doi.org/10.1002/rcm.3701.

[77] F. Allen, A. Pon, R. Greiner, D. Wishart, Computational prediction of electron ionization mass spectra to assist in gc/ms compound identification, Anal. Chem. 88 (2016) 7689–7697, doi:http://dx.doi.org/10.1021/acs.analchem.6b01622.

[78] F. Allen, R. Greiner, D. Wishart, Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Metabolomics. 11 (2015) 98–110, doi:http://dx.doi.org/10.1007/s11306-014-0676-4.

[79] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, D.S. Wishart, CFM-ID 3.0: significantly improved esi-ms/ms prediction and compound identification, Metabolites. 9 (2019), doi:http://dx.doi.org/10.3390/metabo9040072.

[80] L.J. Kangas, T.O. Metz, G. Isaac, B.T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R.R. Lewis, J.H. Miller, In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids, Bioinformatics. 28 (2012) 1705–1713, doi:http://dx.doi.org/10.1093/bioinformatics/bts194.

[81] S. Grimme, Towards first principles calculation of electron impact mass spectra of molecules, Angew. Chem. Int. Ed. 52 (2013) 6306–6312, doi:http://dx.doi.org/10.1002/anie.201300158.

[82] J.N. Wei, D. Belanger, R.P. Adams, D. Sculley, Rapid prediction of electron–ionization mass spectrometry using neural networks, ACS Cent. Sci. 5 (2019) 700–708, doi:http://dx.doi.org/10.1021/acscentsci.9b00085.

[83] C. Hug, N. Ulrich, T. Schulze, W. Brack, M. Krauss, Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening, Environ. Pollut. 184 (2014) 25–32, doi:http://dx.doi.org/10.1016/j.envpol.2013.07.048.

[84] M. Ruff, M.S. Mueller, M. Loos, H.P. Singer, Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds, Water Res. 87 (2015) 145–154, doi:http://dx.doi.org/10.1016/j.watres.2015.09.017.

[85] New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS | Elsevier Enhanced Reader, (n.d.). https://doi.org/10.1016/j.watres.2019.114972.

[86] V. Albergamo, J.E. Schollée, E.L. Schymanski, R. Helmus, H. Timmer, J. Hollender, P. de Voogt, Nontarget screening reveals time trends of polar micropollutants in a riverbank filtration system, Environ. Sci. Technol. 53 (2019) 7584–7594, doi:http://dx.doi.org/10.1021/acs.est.9b01750.
[87] A. Yamamoto, N. Matsumoto, H. Kawasaki, R. Arakawa, Identification of anthropogenic compounds in urban environments and evaluation of automated methods for reading fragmentation—a case of river water, Mass Spectrom. 5 (2016), doi:http://dx.doi.org/10.5702/massspectrometry.A0045.
[88] A.C. Chiaia-Hernandez, E.L. Schymanski, P. Kumar, H.P. Singer, J. Hollender, Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments, Anal. Bioanal. Chem. 406 (2014) 7323–7335, doi:http://dx.doi.org/10.1007/s00216-014-8166-0.
[89] E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibáñez, T. Portolés, R. de Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipaničev, P. Rostkowski, J. Hollender, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, Anal. Bioanal. Chem. 407 (2015) 6237–6255, doi:http://dx.doi.org/10.1007/s00216-015-8681-7.
[90] A. Chao, H. Al-Ghoul, A.D. McEachran, I. Balabin, T. Transue, T. Cathey, J.N. Grossman, R.R. Singh, E.M. Ulrich, A.J. Williams, J.R. Sobus, In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples, Anal. Bioanal. Chem. 412 (2020) 1303–1315, doi:http://dx.doi.org/10.1007/s00216-019-02351-7.
[91] A.D. McEachran, I. Balabin, T. Cathey, T.R. Transue, H. Al-Ghoul, C. Grulke, J.R. Sobus, A.J. Williams, Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns, Sci. Data. 6 (2019) 141, doi:http://dx.doi.org/10.1038/s41597-019-0145-z.
[92] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: communicating confidence, Environ. Sci. Technol. 48 (2014) 2097–2098, doi:http://dx.doi.org/10.1021/es5002105.
[93] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.-M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis, Metabolomics. 3 (2007) 211–221, doi:http://dx.doi.org/10.1007/s11306-007-0082-2.
[94] B. Rochat, Proposed confidence scale and id score in the identification of known-unknown compounds using high resolution ms data, J. Am. Soc. Mass Spectrom. 28 (2017) 709–723, doi:http://dx.doi.org/10.1007/s13361-016-1556-0.
[95] A.G. Newsome, D. Nikolic, CASMI 2013: identification of small molecules by tandem mass spectrometry combined with database and literature mining, Mass Spectrom. 3 (2014) S0034, doi:http://dx.doi.org/10.5702/massspectrometry.S0034.
[96] D. Nikolić, CASMI 2016: A manual approach for dereplication of natural products using tandem mass spectrometry, Phytochem. Lett. 21 (2017) 292–296, doi:http://dx.doi.org/10.1016/j.phytol.2017.01.006.
[97] X. Domingo-Almenara, C. Guijas, E. Billings, J.R. Montenegro-Burke, W. Uritboonthai, A.E. Aisporna, E. Chen, H.P. Benton, G. Siuzdak, The METLIN small molecule dataset for machine learning-based retention time prediction, Nat. Commun. 10 (2019), doi:http://dx.doi.org/10.1038/s41467-019-13680-7.

Chapter 4

Machine Learning in the Annotation of CEC Silyl Derivatives

Herein are presented novel methodological workflows for the generation of datasets of GC-EI-MS spectra of silyl (TMS and TBDMS) derivatives of CEC for the purpose of training and testing of ML-based CA approaches. A three-step workflow is proposed for the generation of spectral datasets of silyl derivatives from mass spectral libraries (MSLs). The three steps correspond to filtering criteria that ensure the presence of good quality GC-EI-MS spectra of chemically logical silyl derivatives, generated as described in Section 2.3.3.1. Next, we generate two test GC-EI-MS datasets of CEC-silyl derivatives using GC-MS acquisition that follows optimized derivatization protocols. Finally, a ML approach based on IOKR methodology for CA of CEC is applied using GC-EI-MS spectral data of their silyl derivatives and achieves good CA accuracy. The CA performance of the ML approach is then compared to the performance of a non-ML approach, and finally, the CA of CEC-TMS derivatives in complex environmental matrices is carried out. The results confirm the second (**H2**) and third (**H3**) hypotheses. In particular, using their GC-EI-MS spectra, structurally diverse semi-polar and thermolabile CEC are successfully identified through their silyl derivatives. Also, it is shown that an ML approach using training and test datasets of silylated compound GC-EI-MS spectra provides more accurate CA than non-ML approaches.

The chapter is divided into four sections: problem description (Section 4.1), two journal papers that address the described problem (Section 4.2), a comparison of the CSI:IOKR approach with a non-ML CA approach (Section **Error! Reference source not found.**), and CEC-TMS identification in complex matrices (Section 4.4).

4.1 Problem Description

Chapter 3 provides a survey of the recent trends in cheminformatics-based CA in the last two decades and points out numerous cheminformatics, and especially ML-based approaches for CA, using MS spectral data [5], [34], [76], [77]. It also highlights the lack of ML-based approaches for CA via GC-EI-MS data. This chapter also presents the challenges in the field to be tackled by future advancements in cheminformatics approaches in EEA. In this context, it is directed toward the following:

- generating publicly available benchmark datasets of high-quality GC-MS spectra that, apart from establishing a standardized system of evaluation metrics, would significantly contribute to the standardization of methods for performance evaluation of CA approaches.

Every novel CA approach is typically compared to other available CA approaches. Such comparison is made using user-defined datasets, evaluation metrics, and executables for automated evaluation, which give variable and seldomly objective results. The only statistically robust comparison of cheminformatics-assisted CA approaches, the 2012-2017 CASMI contests, was performed on datasets of LC-ESI-MS spectra. An exception is the 2012 CASMI contest, where GC-EI-MS datasets were assigned challenges for best MF assignment and compound identification. Moreover, MS spectra of TMS and TBDMS derivatives are poorly represented in compound DB and MSL. This under-representation is potentially one of the issues hampering the employment of cutting-edge CA approaches in CEC annotation through the GC-EI-MS spectra

of their silyl derivatives. We believe that, by generating and publishing datasets of GC-EI-MS spectra and providing them as stand-alone MSL or as a subset of already available MSL, we would encourage the employment of cheminformatics, specifically ML-based CA approaches.

- Stimulating further development of DB-based CA approaches, rather than MSL-based ones, and their use in CA based on GC-EI-MS spectral data.

Out of the 37 cheminformatics-based CA approaches in the CASMI contests [5], 20 approaches were DB-based: ACD/MS Fragmenter [78], Mass Frontier [79], MS FINDER [80], the Rasche method of compound identification by aligning fragmentation trees [81], MetFrag v.1.0-v.3.0 [82]–[84], MAGMa [85], MolFind [35], Database Assisted Structure Identification (DASI) [86], MIDAS [87] and MIDAS-G [88], fragment set enrichment analysis (FSEA) [89], MetExpert [90], FingerID [91], and all the IOKR-based approaches [6], [92]–[95], SIMPLE and L-SIMPLE [76] and DeepEI [96]. Of these, MS-FINDER [80], MetFrag [82], and CSI:IOKR [6] were the best performing approaches in the CASMI 2016 contest, while MS-FINDER [80] + metadata and CSI:FingerID [92] were best performing approaches in the CASMI 2017 contest. Interestingly, only 12 approaches, out of all 37 evaluated approaches were designed and tested on GC-EI-MS data: MSClass [97], MassFrontier [79], ACD/MS Fragmenter [78], MOLGEN-MS [98], Golm Metabolome Database algorithm [99], Hufsky method [100], MetExpert [90], DeepEI [96], CFM-ID v.1.0 [101] and the Quantum Chemistry Electron Ionization Mass Spectra approaches [102]–[104]. The latter three [102]–[104] were not properly used in real-throughput analyses due to their long computation times. The current cutting-edge methodologies, based on IOKR, are ML models trained and tested exclusively on LC-ESI-MS/MS data and are yet to be challenged against GC-EI-MS data. We opt to achieve satisfactory CA performance by adapting and applying IOKR approaches to GC-EI-MS spectral data. The reason is that under EI compounds fragment following predictable and thoroughly studied fragmentation patters, resulting in highly reproducible EI spectra suitable for CEC annotation [64].

This chapter also addresses two aforementioned issues - the generation of publicly available benchmark datasets of high-quality GC-MS spectra and the stimulation of further development of DB-based CA approaches. The first issue is addressed by developing methodological workflows for generation of MSL-derived GC-EI-MS spectral datasets and *de novo* generation of GC-EI-MS datasets of silyl derivatives of a representative selection of CECs. Both are intended to serve as benchmark datasets for the development, training, validation, and performance evaluation of ML-based CA approaches. The second issue is addressed by applying the IOKR-based ML approach to annotating CEC-silyl derivatives. Finally, the performance of the IOKR-based ML approach with the performance of manual MSL search is compared, and the annotation of CEC-TMS derivatives in complex environmental matrices is performed.

4.2 Related Publications

Publication 1

Journal paper

Ljoncheva, M., Kosjek, T., Džeroski, S. GC-EI-MS datasets of trimethylsilyl (TMS) and *tert*-butyl dimethyl silyl (TBDMS) derivatives for the development of machine-learning based compound identification approaches, *Data in Brief* (submitted 4 July 2022).

This publication contains the following contributions:

- A methodology, i.e., a three-step filtering approach for generating GC-EI-MS spectral datasets of silyl derivatives from MSL, ensures reliable and accurate CA.
- GC-EI-MS spectral datasets of CEC TMS and TBDMS derivatives are given along with their metadata in universal ready-to-use formats for further cheminformatics-based processing. The GC-EI-MS spectral datasets are publicly available in the Mendeley Data Repository at the following link: https://data.mendeley.com/datasets/j3z5bmvmnd/3. They are free to use in performance evaluation and validation of other existing ML-based CA approaches and develop and optimize novel ML-based CA approaches.

- The generated GC-EI-MS spectral datasets are of significant value for the environmental and exposomics research communities, as they can be used as a stand-alone MSL or joined with other MSL in the task of eco exposome annotation (EEA).

GC-EI-MS datasets of trimethylsilyl (TMS) and *tert*-butyl dimethyl silyl (TBDMS) derivatives for development of machine learning-based compound identification approaches

Milka Ljoncheva [†,‡], Tina Kosjek [†,‡], Sašo Džeroski [ʃ,‡,*]

Affiliations

[†] Jozef Stefan Institute, Department of Environmental Sciences, Jamova 39, 1000 Ljubljana, Slovenia

[ʃ] Jozef Stefan Institute, Department of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia

[‡] Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia

[*]Corresponding author.

Corresponding author's email address

Tel: +386 14773217

E-mail: saso.dzeroski@ijs.si

Keywords

silylation, derivative, identification, machine learning, GC-MS, mass spectrometry

Abstract

In the field of environment and health studies, recent trends have focused on the identification of contaminants of emerging concern (CEC). This is a complex, challenging task, as resources, such as compound databases (DBs) and mass spectral libraries (MSLs) concerning these compounds are very poor. This is particularly true for semi polar organic contaminants that have to be derivatized prior to gas chromatography-mass spectrometry (GC-MS) analysis with electron impact ionization (EI), for which it is barely possible to find any records. In particular, there is a severe lack of datasets of GC-EI-MS spectra generated and made publicly available for the purpose of development, validation and performance evaluation of cheminformatics-assisted compound structure identification (CSI) approaches, including novel cutting-edge machine learning (ML)-based approaches [1].

We set out to fill this gap and support the machine learning-assisted compound identification, thus aiding cheminformatics-assisted identification of silylated derivatives in GC-MS laboratories working in the field of environment and health. To this end, we have generated 12 datasets of GC-EI-MS spectra, six of which contain GC-EI-MS spectra of trimethylsilyl (TMS) and six GC-EI-MS spectra of *tert*-butyldimethylsilyl (TBDMS) derivatives. Four of these datasets, named testing datasets, contain mass spectra acquired by the authors: They are available in full, together with corresponding metadata. Eight datasets, named training datasets, were derived from mass spectra in the NIST 17 Mass Spectral Library: For these, we have only made the metadata publicly available, due to licensing reasons.

For each type of derivative, two testing datasets are generated by acquiring and processing GC-EI-MS spectra, such that they include raw and processed GC-EI-MS spectra of TMS and TBDMS derivatives of CECs, along with their corresponding metadata. The metadata contains IUPAC name, exact mass, molecular formula, InChI, InChIKey, SMILES and PubChemID, of each CEC and CEC-TMS or CEC-TBDMS derivative, where available. Eight GC-EI-MS training datasets are generated by using the National Institute of Standards and Technology (NIST)/U.S. Environmental Protection Agency (EPA)/National Institute of Health (NIH) 17 Mass Spectral Library. For each derivative type (TMS and TBDMS), four datasets are presented, each corresponding to an original dataset obtained from NIST/EPA/NIH 17 and three variants thereof, obtained after each of the filtering steps of the procedure described below. Only the metadata about the training datasets are available, describing the corresponding NIST/EPA/NIH 17 entries: These include the compound name, CAS Registry number, InChIKey, exact mass, M_w, NIST number and ID number.

The datasets we present here were used to train and test predictive models for the identification of silylated derivatives, built with ML approaches [4]. The models were built by using data curated from the NIST Mass Spectral Library 17 [2] as input to the machine learning approach of CSI:Output Kernel Regression (CSI:OKR) [3]. Data from the NIST Mass Spectral Library 17 are commercially available from the National Institute of Standards and Technology (NIST)/U.S. Environmental Protection Agency (EPA)/National Institute of Health (NIH) and thus cannot be made publicly available. This highlights the need for publicly available GC-EI-MS spectra, which we address by releasing in full the four testing datasets.

Specifications table

Subject	Analytical chemistry, Omics: General
Specific subject area	Generation of mass spectral datasets for testing and training of ML-based CSI approaches using mass spectra
Type of data	Raw and Table
How the data were acquired	The mass spectra in the testing datasets were acquired using Agilent 7890B/5977A series GC-MSD (Agilent Technologies, USA), in electron impact ionization (EI) mode. Chromatographic separation was achieved on Agilent DB-5MS UI fused-silica capillary column (30m x 0.25mm x 0.25 µm; Agilent Technologies, USA). Data was processed using Mass Hunter Quantitative Analysis v.B.07 (Agilent Technologies, USA). The training datasets were generated from the NIST/EPA/NIH Mass Spectral Library 17 [2] using the accompanying NIST Mass Spectral Search Program (version 2.3) and LIB2NIST converter (NIST, 2011).
Data format	The testing GC-EI-MS spectral datasets are given in .txt format and the accompanying metadata is given in .xlsx format. The metadata about the training GC-EI-MS spectral datasets is given in .xlsx format.
Description of data collection	The GC-EI-MS spectral datasets, that we used for testing ML-based CSI models, were generated in the full scan range of *m/z* 50-800 amu for the TMS derivatives and *m/z* 50-1000 amu for the TBDMS derivatives. Raw instrument data was reduced to two-dimensional peak lists (*m/z*, abundance) using Mass Hunter Qualitative Analysis v.B.07 (Agilent Technologies, USA), in which background subtraction was also performed. The GC-EI-MS spectral datasets intended for training of ML-based CSI models were generated from the NIST/EPA/NIH Mass Spectral Library 17: Constrained search was first performed using the NIST Mass Spectral Search Program (version 2.3), followed by further processing according to the step-wise procedure described below.
Data source location	· *Institution: Jožef Stefan Institute, Department of Environmental Sciences* · *City/Town/Region: Ljubljana* · *Country: Slovenia*
Data accessibility	Repository name: Mendeley Data Data identification number: doi: 10.17632/j3z5bmvmnd.3 Direct URL to data: https://data.mendeley.com/datasets/j3z5bmvmnd/3

Value of the data

- The generated testing GC-EI-MS datasets provide a comprehensive collection of GC-EI-MS spectra of TMS and TBDMS derivatives of structurally and chemically diverse environmental contaminants, given along with their metadata in universal ready-to-use formats for further cheminformatics-based processing.
- The generated testing GC-EI-MS datasets are of value for the environmental and exposomics researchers, as well as for the CSI and ML communities, interested in the development of new CSI tools.
- Few datasets of mass spectra are publicly available: This is especially true for GC-EI-MS spectra, which makes the generated data even more valuable.
- Both the testing and the training data can be further used on their own or as part of larger datasets, for training, testing and validation in the development of novel CSI approaches, for challenging existing approaches, and for performance comparison of novel and existing CSI, especially ML-based approaches.
- The data can be used as a stand-alone database (or joined with other in-house databases of GC-EI-MS spectra), serving as valuable reference during suspect screening and non-targeted environmental analysis.

1. Data description

NIST/EPA/NIH 17 Mass Spectral Library [2] was used to generate training GCI-EI-MS spectral datasets of TMS and TBDMS derivatives, which are then used for building ML-based CSI approaches. As NIST/EPA/NIH 17 Mass Spectral Library [2] is commercially available and licensed under the United States Department of Commerce Copyright, the training GC-EI-MS spectral datasets themselves cannot be made publicly available. Instead, we provided metadata files of each dataset, containing the name, InChIKey, CAS Registry number, exact mass, Mw, NIST number and ID number for each GC-EI-MS spectrum. For each derivative type (TMS and TBDMS), four GC-EI-MS datasets were generated – the first ones (TMS_0.1 and TBDMS_0.1) containing the GC-EI-MS spectra initially extracted from NIST/EPA/NIH 17 Mass Spectral Library ("Metadata_training_TMS_0.1", "Metadata_ training_TBDMS_0.1"), followed by TMS_1.3 and TBDMS_1.3 resulting from the first filtering step of the approach described below ("Metadata_ training_TMS_1.3", "Metadata_ training_TBDMS_1.3"), TMS_2.3 and TBDMS_2.3, resulting from the second filtering step ("Metadata_ training_TMS_2.3", "Metadata_ training_TBDMS_2.3"), and TMS_3.3 and TBDMS_3.3, resulting from the final, third filtering step ("Metadata_training_TMS_3.3", "Metadata_ training_TBDMS_3.3"). Using the given metadata and the described procedure (), the training GC-EI-MS spectral datasets can be reconstructed.

Standard solutions of selected environmental contaminants (104 compounds), listed in Table 1, were used for generating the TMS and TBDMS test dataset. The presented data consists of .txt data files for each of the four MS datasets (Test dataset_TMS_RAW, Test dataset_TMS_BS, Test dataset_TBDMS_RAW and Test dataset_TBDMS_BS), in which each of the GC-EI-MS spectra is recorded with the compound name, InChIKey, molecular weight (M_w), molecular formula (MF), CAS Registry number and list of peaks represented as *m/z* and intensities. Metadata of the TMS/TBDMS derivatives and the corresponding parent CEC are given in .xlsx files containing the IUPAC name, exact mass, molecular formula, InChI, InChIKey, SMILES and PubChem ID, when available ("Metadata_test_TMS derivatives.xlsx" and "Metadata_test_TBDMS derivatives.xlsx"). The four datasets were used to test predictive models for identification of silylated derivatives, built with ML approaches.

The predictive models for CSI of silylated derivatives were built by ML approaches from training datasets of GC-EI-MS spectra of TMS and TBDMS derivatives, which are not publicly available, as they were curated from the commercially available NSIT/EPA/NIH Mass Spectral Library 17 [2], licensed under the United States Department of Commerce Copyright. NIST's end-user's license for the NIST 17 MSL restrict its use to a single computer that is not accessible by more than one person. While the training datasets themselves cannot be made publicly available, we make available the corresponding metadata: With licensed access to NIST MSL 17, they can be used to reconstruct the training datasets, by following the workflow summarized below and described by Ljoncheva et al. [4].

The ML approach used to build models for the identification of silylated derivatives from these data [4] was the approach titled Compound Structure Identification-Input:Output Kernel Regression (CSI:OKR) [3].

2. Experimental design, materials and methods

2.1 Experimental design and generation of training datasets

Initial versions of the TMS and TBDMS datasets (TMS_0.1 and TBDMS_0.1) were generated by extracting all GC-EI-MS spectra of TMS, resp. TBDMS, derivatives of small molecules from the NIST/EPA/NIH 17 Mass Spectral Library [2]. The first constrained search for GC-EI-MS TMS spectra, using the constraints *name fragment: trimethylsilyl* and *elements allowed: Si*, resulted in a collection of 9958 entries, while for GC-EI-MS TBDMS spectra, the constraints *name fragment: tertbutyldimethylsilyl* and *elements allowed: Si, r*esulted in an initial dataset of 2238 entries. Entries were extracted in .msp file format and subsequently converted to .txt format, using the LIB2NIST conversion tool (NIST 2011). Each GC-EI-MS entry included the compound name, InChIKey, MF, Mw, exact mass, CAS number, NIST ID and MS peak list. The GC-EI-MS spectra of TMS/TBDMS derivatives with erroneous metadata (name, molecular formula, InChIKey) that did not correspond to the analyzed compound were excluded from the dataset.

The TMS/TBDMS GI-EI-MS spectral datasets were further filtered using a three-step spectral filtering process, including:

1) Exclusion of chemical irregularities. The GC-EI-MS spectra of TMS, resp. TBDMS derivatives of compounds not susceptible to derivatization, defined by the absence of functional group(s) amenable to silylation, were filtered out. The functional groups amenable to silylation are those containing an active hydrogen, i.e., carboxyl, hydroxyl, amine and thiol.

2) Exclusion of high-molecular mass TMS, resp. TBDMS derivatives. The GC-EI-MS spectra of TMS, resp. TBDMS derivatives of CEC with molecular mass ≥ *m/z* 1000 were eliminated, since, as such, they are above the working linear range of the GC-MS instruments.

3) Exclusion of insufficient-quality GC-EI-MS spectra. The following GC-EI-MS spectra were excluded:
- GC-EI-MS spectra not acquired at the upper *m/z* of at least Mw of the derivative + 10 amu;
- GC-EI-MS spectra that do not contain both the molecular ion $[M]^+$ peak and at least one of the isotope peaks, such as the ^{13}C isotope peak;
- GC-EI-MS spectra that contain neither peaks of fragment ions specific for TMS groups (*m/z* 73, 147, 221 and 295, corresponding to one, two, three and four TMS groups, respectively) nor for TBDMS groups (*m/z* 115, 230 and 345, corresponding to one, two and three TBDMS groups, respectively) and
- GC-EI-MS spectra not containing at least five fragment ion peaks.

As a result, the final version of the TMS dataset consists of 4648 TMS GC-EI-MS spectra, while the final version of the TBDMS dataset consists of 1883 GC-EI-MS spectra. For each of the GC-EI-MS spectra in the final TMS and TBMDS datasets, the *m/z* range was between 50 *m/z* to M_w of the derivative ± 10 amu. For data parsing, all ion fragments with intensity 0 were removed from the refined TMS and TBDMS datasets.

2.2 Chemical analysis for generation of test datasets

2.2.1 Chemicals and materials

From the in-house pool of reference standards, 104 CEC were selected as environmentally relevant, according to the criteria of the Regulation (EC) No.1907/2006 of the European Parliament and the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Annex III [5]. These are listed in Table 1.

Table 1. CEC for generation of the TMS and TBDMS datasets. *-CECs that have two TMS derivatives; **-CECs that have two TBDMS derivatives.

CEC included in TMS datasets only	
Nitroxoline	shikimic acid
17α-hydroxyprogesterone	meso-erythritol
6β-hydroxypregnenolone	amphetamine
5-androstene-3β, 17β-diol	6-nitroguaiacol
5α-dehydrotestosterone (boldenone)	estriol
11α-hydroxytestosterone	codeine
11α-hydroxyandrostenedione	L-tyrosine
methamphetamine	L-ascorbic acid
nylidrin	cannabidiolic acid
(-)-quinic acid	bisphenol FL
	butylated hydroxytoluene
CEC included in TBDMS datasets only	
4,6-dinitroguaiacol	mycophenolic acid
CEC included in both TMS and TBDMS datasets	
bisphenol A	2-benzyl-4-chlorophenol
benzoic acid	citric acid monohydrate
mecoprop	4-cumylphenol
4,4'-biphenol	2,4-dihydroxybenzophenone
4,4'-dihydroxydiphenyl ether	estrone
4,4′-isopropylidenebis(2,6-dimethylphenol)	17β-estradiol
2,4'-dihydroxydiphenylmethane (24BPF)	m-coumaric acid
bisphenol AF	p-coumaric acid
bisphenol AP	o-coumaric acid
bisphenol C	triclosan
bisphenol E	(+)-cannabidiol
bisphenol F	cannabinol
bisphenol M	cannabichromene
bisphenol BP	morphine
bisphenol P	6-monoacetylmorpine
bisphenol S	carbamazepine
bisphenol Z	isopropylparaben
2,2′-methylenediphenol (22BPF)	bisphenol B
dihydrotestosterone (stanolone)	17α-ethynyl estradiol
(±)-11-hydroxy-Δ^9- tetrahydrocannabinol	4-hydroxybenzophenone
(±)-11-nor-9-carboxy-Δ^9-tetrahydrocannabinol	2,2'-dihydroxy-4-methoxybenzophenone (BP-8)
sulfanilamide	clofibric acid
adipic acid	ibuprofen
4-tert-octylphenol	naproxen
9-hydroxyfluorene	ketoprofen
L-leucine	diclofenac
L-serine	methylparaben
(-)-Δ9 tetrahydrocannabinol	ethylparaben

(-)-Δ9 tetrahydrocannabinolic acid	propylparaben
trans-3'-hydroxycotinine	butylparaben
benzoylecgonine	isobutylparaben
bisphenol CL	benzylparaben
bisphenol PH	4-nonylphenol
8-hydroxyquinoline	phenylacetic acid
2-anilinophenylacetic acid	resorcinol
4-nitroguaiacol	salicylic acid
5-nitroguaiacol	urea
catechol	4-nitrocatechol
3-methylcatechol	syringol
3-methyl-5-nitrocatechol	4-nitrosyringol
	etofylline

The selected compounds had to satisfy at least three of the following five criteria: **1) Positioning**: the compound is present in the US EPA CCD [11], the most comprehensive repository of EE constituents; **2) Persistence**: compound's half-life in fresh or estuarine water is > 40 days; **3) Bioaccumulation**: BAF and/or BCF > 2000, or in absence of such data, $\log Ko_w \geq 5.0$; **4) Mobility**: compound's water solubility is ≥ 0.15 mg/L and log Koc is ≤ 4.0, i.e. between -10.0 and 4.0; and **5) EcoToxicity**: long-term no-observed-effect concentration (NOEC) for marine or freshwater organisms is < 0.01 mg/L., following. Further details of the selection procedure are given by Ljoncheva et al. [4].

Individual stock solutions of each CEC at a concentration of approximately 150 µg/mL were prepared in acetonitrile (ACN), ethyl acetate (EtAc) or methanol (MeOH), depending on CEC solubility. Of them, individual working solutions (IWS) at concentration 1 µg/mL were prepared and used within 7 days. The CEC included in each GC-EI-MS dataset are listed in Table 1.

2.2.2 Derivatization and analysis

For each CEC, to 500 µL of IWS, 470 µL EtAc and 30 µL derivatization agent was added; for generation of TMS derivatives, 30 µL N, O-bis trifluoroacetamide with 1% trimethylchlorosilane (BSTFA + 1% TMCS), for TBDMS derivatives N-*tert*-butyldimethylsilyl-N-methyltrifluoroacetamide (MTBSTFA) with 1% TMCS (MTBSTFA + 1% TMCS). For CECs whose IWS are in ACN or MeOH, the 500 µL IWS was dried under N_2 flow, reconstituted in 970 µL EtAc, to which 30 µL of the appropriate derivatization agent were added at predefined reaction temperature and duration. 106 TMS derivatives of 102 CEC (4 CEC resulted in two TMS derivatives, namely salicylic acid, dihydrotestosterone (stanolone), sulfanilamide and 5α-androsten-3β, 17β-diol) and 85 TBDMS derivatives of 83 CEC, two with two TBDMS derivatives each, namely sulfanilamide and L-serine) were generated, with molecular weights of derivatives ranging up to 650 amu.

GC-EI-MS spectra were acquired on Agilent 7890B/5977A series GC-MSD (Agilent Technologies, USA). Separation was achieved on Agilent DB-5MS UI fused-silica capillary column (30m x 0.25mm x 0.25 µm; Agilent Technologies, USA). He of 99.99999% purity at the flow rate of 1.2 mL/min was used as a carrier gas. The manifold, ion source and transfer line temperatures were set at 230°C, 150°C and 250°C, respectively. Injections (1 µL) were performed in the splitless mode. Depending upon compound properties, one of the following column oven temperature programs was used: (1) initial temperature 70 ºC (held 1 min), ramped at 15 ºC/min

to 280 ºC (held 1 min); total runtime: 16 min; (2) initial temperature 70 ºC (held 1 min), ramped at 20 ºC/min to 240 ºC (held 1 min), at 12 ºC to 310 ºC (held 2 min); total runtime: 18.3 min; (3) initial temperature 70 ºC (held 1 min), ramped at 20 ºC/min to 240 ºC (held 1 min), at 12 ºC to 310 ºC (held 4 min); total runtime: 20.3 min. The MSD was operated in EI ionization mode (70 eV) by scanning over the mass range of *m/z* 50-800 amu for TMS derivatives and *m/z* 50-1000 amu for TBDMS derivatives. In-between the acquisitions of the derivatized standards, EtAc was run as the solvent check to assess potential background interferences and was used for background subtraction as a part of the post-acquisition processing of the GC-EI-MS spectra.

2.3 Data processing

GC-EI-MS data acquisition resulted in the generation of multiple (≥15) GC-EI-MS spectra for most of the TMS and TBDMS derivatives. Exceptions are the L-ascorbic acid TMS, L-leucine TMS and L-serine TMS, with three GC-EI-MS spectra each, and the TBDMS derivatives of L-serine, 4-nitroguaiacol, 5-nitroguaiacol, catechol, 3-methylcatechol, 3-methyl-5-nitrocatechol, syringol, 4-nitrosyringol, 4-nitrocatechol, p-coumaric acid, m-coumaric acid, o-coumaric acid, mycophenolic acid, 4,6-dinitroguaiacol, etofylline and urea, with one GC-EI-MS spectrum in the test TBDMS datasets for each. All GC-EI-MS spectra were processed using Mass Hunter Qualitative Analysis v B.07 (Agilent Technologies, USA) that reduced raw instrument data to two-dimensional peak lists (*m/z*, abundance), exported in .txt format. This software was also used to perform background subtraction, in order to remove constantly present background signals, such as *m/z* 149 as a typical phtalate interference, *m/z* 282, *m/z* 256 and *m/z* 284 for oleic, palmitic and stearic acid, and *m/z* 207, *m/z* 281 and *m/z* 327 of common polysiloxanes resulting from GC column stationary phase degradation. Their presence was confirmed *a priori* in the multiple EtAc solvent runs acquired between the acquisitions of CEC silyl derivatives, and was used for background subtraction.

The final test GC-EI-MS datasets, both raw and noise background subtracted, of TMS and TBDMS derivatives were compiled as .txt file that included MFs, InChIKey strings, Mw and two-dimensional peak lists. Molecular stereochemistry was not considered, since stereoisomers are not readily distinguished by MS.

Ethics statements

The authors declare that the manuscript meets all the rules and conditions described in the “Ethics in publishing” section standards (https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors). The work did not include any investigations involving animal experiments, human participants and data collected from social media platforms.

The training GC-EI-MS spectral datasets were curated from the commercially available NIST/EPA NIH 17 Mass Spectral Library. Explicit permission to release the metadata about the training datasets was obtained by the authors from NIST. Due to NIST’s individual license, restricting the use to a single computer that is not accessible by more than one person, the training datasets cannot be made available to the public by the authors. However, with licensed access to the NIST 17 MSL, the training data can be reconstructed from the available metadata files by following the detailed description of data preparation, given above.

CRediT author statement

Milka Ljoncheva: Investigation, Formal Analysis, Data curation, Software, Writing-original draft, Writing-review and editing; **Tina Kosjek:** Conceptualization, Resources, Validation, Supervision, Writing-review and editing; **Sašo Džeroski:** Conceptualization, Resources, Validation, Supervision, Writing-review and editing.

Acknowledgments

The authors acknowledge the instrumental and computational resources provided by the Jožef Stefan Institute. This work was supported by the Slovenian Research Agency (ARRS through the programs P1-0143 and P2-0103). M.L. was funded by the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia (contract no. 11011-85/2016).

Declaration of interests

Please tick the appropriate statement below and declare any financial interests/personal relationships which may affect your work in the box below.

✓The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Please declare any financial interests/personal relationships which may be considered as potential competing interests here.

References

[1] M. Ljoncheva, T. Stepišnik, S. Džeroski, T. Kosjek, Cheminformatics in MS-based environmental exposomics: Current achievements and future directions, Trends in Environmental Analytical Chemistry. 28 (2020) e00099. https://doi.org/10.1016/j.teac.2020.e00099.

[2] National Institute of Standards and Technology, NIST/EPA/NIH Mass Spectral Library 2017, Wiley.Com. (2017). https://www.wiley.com/en-ai/NIST+EPA+NIH+Mass+Spectral+Library+2017-p-9781119750291 (accessed June 10, 2021).

[3] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, J. Rousu, Fast metabolite identification with Input Output Kernel Regression, Bioinformatics. 32 (2016) i28–i36. https://doi.org/10.1093/bioinformatics/btw246.

[4] M. Ljoncheva, T. Stepišnik, T. Kosjek, S. Džeroski, Machine learning for identification of silylated derivatives from mass spectra, (submitted for publication, currently under review).

[5] European Commsion, Regulation (EC) No.1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Official Journal of the European Communities. 396 (2021) 1–552.

Publication 2

Journal paper

Ljoncheva, M., Stepišnik T., Kosjek, T., Džeroski, S. Machine learning for identification of silylated derivatives from mass spectra, *Journal of Cheminformatics* (accepted, 31 July 2022)

This publication contains the following contributions:

- ML-based CSI:IOKR approach has been used to identify CEC silyl derivatives from their GC-EI-MS spectra.
- The impact of several factors on the identification performance of CSI:IOKR has been investigated. The factors include the use of filtering approaches in generating MSL-derived spectral datasets, the overlap between training and testing datasets, and the post-acquisition processing of the test dataset.
- Confirmation that the identification rates of the IOKR-based approaches largely depend on the quality of the training MS spectral dataset, with a crucial improvement of identification rates achieved by expert-driven curation of training datasets.
- Confirmation that the performance of the IOKR model is not influenced by the size of the compound's candidate set.
- Confirmation that the IOKR approach has identification rates for TMS derivatives outside the training dataset comparable to the rates for those inside the training dataset.
- Confirmation that the IOKR approach can be successfully employed in CA of CEC using the GC-EI-MS spectra of their silyl derivatives as a viable alternative to MSL(s) search.

Metadata of the article that will be visualized in OnlineFirst

ArticleTitle	Machine learning for identification of silylated derivatives from mass spectra	
Article Sub-Title		
Article CopyRight	The Author(s) (This will be the copyright line in the final PDF)	
Journal Name	Journal of Cheminformatics	
Corresponding Author	FamilyName	**Džeroski**
	Particle	
	Given Name	**Sašo**
	Suffix	
	Division	Department of Knowledge Technologies
	Organization	Jozef Stefan Institute
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Division	
	Organization	Jozef Stefan International Postgraduate School
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Phone	
	Fax	
	Email	saso.dzeroski@ijs.si
	URL	
	ORCID	
Author	FamilyName	**Ljoncheva**
	Particle	
	Given Name	**Milka**
	Suffix	
	Division	Department of Environmental Sciences
	Organization	Jozef Stefan Institute
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Division	
	Organization	Jozef Stefan International Postgraduate School
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Phone	
	Fax	
	Email	
	URL	
	ORCID	
Author	FamilyName	**Stepišnik**
	Particle	
	Given Name	**Tomaž**
	Suffix	
	Division	Department of Knowledge Technologies
	Organization	Jozef Stefan Institute
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Division	
	Organization	Jozef Stefan International Postgraduate School
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Phone	
	Fax	
	Email	
	URL	
	ORCID	

Author	Family Name	**Kosjek**
	Particle	
	Given Name	**Tina**
	Suffix	
	Division	Department of Environmental Sciences
	Organization	Jozef Stefan Institute
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Division	
	Organization	Jozef Stefan International Postgraduate School
	Address	Jamova 39, 1000, Ljubljana, Slovenia
	Phone	
	Fax	
	Email	
	URL	
	ORCID	

Schedule	Received	13 May 2022
	Revised	
	Accepted	31 Jul 2022

Abstract	*Motivation:* Compound structure identification is using increasingly more sophisticated computational tools, among which machine learning tools are a recent addition that quickly gains in importance. These tools, of which the method titled Compound Structure Identification:Input Output Kernel Regression (CSI:IOKR) is an excellent example, have been used to elucidate compound structure from mass spectral (MS) data with significant accuracy, confidence and speed. They have, however, largely focused on data coming from liquid chromatography coupled to tandem mass spectrometry (LC–MS). Gas chromatography coupled to mass spectrometry (GC–MS) is an alternative which offers several advantages as compared to LC–MS, including higher data reproducibility. Of special importance is the substantial compound coverage offered by GC–MS, further expanded by derivatization procedures, such as silylation, which can improve the volatility, thermal stability and chromatographic peak shape of semi-volatile analytes. Despite these advantages and the increasing size of compound databases and MS libraries, GC–MS data have not yet been used by machine learning approaches to compound structure identification. *Results:* This study presents a successful application of the CSI:IOKR machine learning method for the identification of environmental contaminants from GC–MS spectra. We use CSI:IOKR as an alternative to exhaustive search of MS libraries, independent of instrumental platform and data processing software. We use a comprehensive dataset of GC–MS spectra of trimethylsilyl derivatives and their molecular structures, derived from a large commercially available MS library, to train a model that maps between spectra and molecular structures. We test the learned model on a different dataset of GC–MS spectra of trimethylsilyl derivatives of environmental contaminants, generated in-house and made publicly available. The results show that 37% (resp. 50%) of the tested compounds are correctly ranked among the top 10 (resp. 20) candidate compounds suggested by the model. Even though spectral comparisons with reference standards or de novo structural elucidations are neccessary to validate the predictions, machine learning provides efficient candidate prioritization and reduction of the time spent for compound annotation.

Keywords (separated by '-')	Silylation - Derivative - Identification - Machine learning - Mass spectrometry - Molecular fingerprint - Prediction
Footnote Information	The online version contains supplementary material available at https://doi.org/10.1186/s13321-022-00636-1.

Ljoncheva *et al. Journal of Cheminformatics* _####################_
https://doi.org/10.1186/s13321-022-00636-1

Journal of Cheminformatics

RESEARCH Open Access

Machine learning for identification of silylated derivatives from mass spectra

Milka Ljoncheva[1,3], Tomaž Stepišnik[2,3], Tina Kosjek[1,3] and Sašo Džeroski[2,3*]

Abstract

Motivation: Compound structure identification is using increasingly more sophisticated computational tools, among which machine learning tools are a recent addition that quickly gains in importance. These tools, of which the method titled Compound Structure Identification:Input Output Kernel Regression (CSI:IOKR) is an excellent example, have been used to elucidate compound structure from mass spectral (MS) data with significant accuracy, confidence and speed. They have, however, largely focused on data coming from liquid chromatography coupled to tandem mass spectrometry (LC–MS).

Gas chromatography coupled to mass spectrometry (GC–MS) is an alternative which offers several advantages as compared to LC–MS, including higher data reproducibility. Of special importance is the substantial compound coverage offered by GC–MS, further expanded by derivatization procedures, such as silylation, which can improve the volatility, thermal stability and chromatographic peak shape of semi-volatile analytes. Despite these advantages and the increasing size of compound databases and MS libraries, GC–MS data have not yet been used by machine learning approaches to compound structure identification.

Results: This study presents a successful application of the CSI:IOKR machine learning method for the identification of environmental contaminants from GC–MS spectra. We use CSI:IOKR as an alternative to exhaustive search of MS libraries, independent of instrumental platform and data processing software. We use a comprehensive dataset of GC–MS spectra of trimethylsilyl derivatives and their molecular structures, derived from a large commercially available MS library, to train a model that maps between spectra and molecular structures. We test the learned model on a different dataset of GC–MS spectra of trimethylsilyl derivatives of environmental contaminants, generated in-house and made publicly available. The results show that 37% (resp. 50%) of the tested compounds are correctly ranked among the top 10 (resp. 20) candidate compounds suggested by the model. Even though spectral comparisons with reference standards or de novo structural elucidations are neccessary to validate the predictions, machine learning provides efficient candidate prioritization and reduction of the time spent for compound annotation.

Keywords: Silylation, Derivative, Identification, Machine learning, Mass spectrometry, Molecular fingerprint, Prediction

Introduction

Growing awareness of the environmental impact on human health has increased interest into the environmental chemical space of the human exposome, that consists of the multitude of structurally and toxicologically diverse synthetic and naturally occurring compounds [1–3]. This has turned the annotation of CEC into a task of utmost importance [4–6], as it can

*Correspondence: saso.dzeroski@ijs.si

[2] Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Full list of author information is available at the end of the article

Journal : BMCTwo 13321	Dispatch : 9-8-2022	Pages : 20
Article No : 636	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

provide valuable knowledge about their identity, accumulation, degradation and transformation patterns, exposure pathways and toxicity. Among the multitude of chemical, biological and toxicity estimation methods, chromatography coupled to MS methods has become the essential analytical tool for thorough CEC annotation. Employment of strongly consolidated, targeted, suspect screening and non-targeted screening strategies requires the use of data processing software, cheminformatics tools, ever-growing compound databases (DBs), MS libraries (MSLs) and computational MS workflows for assignment of chemical identities to MS signals.

In its beginning, MS-based high throughput exposome exploration involved manual determination of compound's molecular weight, computation of a molecular formula (MF) and then search against data repositories for canddiates. Different data resources have been used for this purpose, including user-generated specified suspect lists (e.g. [7, 8]), specialized lists compiled by, e.g., the US EPA's Distributed Structure-Searchable Toxicity (DSSTox) database [9] and environmental communities such as the NORMAN Network [10]. Medium-sized DBs contain tens to hundreds of thousands of compounds (e.g., US EPA's Comptox Chemistry Dashboard (CCD) [11], ContaminantDB [12], the Toxin and Toxin Target Database (T3DB) [13], the Exposome Explorer [14]), while the most comprehensive chemical repositiories, such as PubChem [15] and Chemspider [16] can contain over 100 million compounds. The latter are the most frequently exploited sources. They offer an exceptionally wide chemical space, hence a simple exact mass or MF search rapidly turns into a non-target identification challenge, often with hundreds to thousands of hits [7, 17]. Later, MSLs were introduced to obtain rapid tentative identifications at relatively high confidence [18]. Many MSLs either contain predominantly LC–MS data (e.g., the Human Metabolome Database (HMDB) 4.0 [19], METLIN [20], MassBank [21], mzCloud [22]), GC–MS data (e.g., the Golm Metabolome Database (GMD) [23], the Fiehn Library [24]), or both (e.g., National Institute of Standards and Technology (NIST) Mass Spectral Library [25] and Wiley Registry [26]). Compounds are identified by comparing experimentally acquired and reference MSL spectra using versatile spectral similarity functions. Yet even nowadays, in the era of their substantial increase in size and comprehensiveness, MSLs cover only a fraction of the exposomics-relevant chemical information, as inclusion of newly identified CEC is inherently limited by the availability of reference standards, the relative youth and the lack of their standardization [27]. This coverage is even poorer for silyl derivatives, with very few MSLs [23–25] containing their MS spectra.

In the last decade, compound structure identification (CSI) based on compound DB and MSL has been replaced by numerous cheminformatics methods [28]. These methods perform CSI by either determining the exact mass or MF, by using a predefined exact mass or MF, or by converting the structural information inherent to MS data, including the presence of specific substructures, functional groups or complete fragmentation pathways, into a computationally more convenient "third format". Here, "third format" representation of the structural information contained in MS spectra includes more computationally manageable formats, such as fragmentation trees, mass spectral trees (for multi-stage MS data, MS^n), and molecular fingerprints (MFP), all which include structural information that can be extracted from an MS spectrum and further processed. Based on this third format, the cheminformatics approaches perform exhaustive interrogation/search of MSLs or compound DBs to create candidate sets, from which, according to (sub)structural similarity (possibly accompanied with other criteria, such as chromatographic behaviour, energy, data source, environmental behavior and toxicity related criteria and/or complementary information), most probable candidates are prioritized and ranked [28, 29]. Among these approaches, those based on machine learning have offered highest accuracy, confidence and speed in performing the CSI task [7, 29, 30].

Revolutionary breakthroughs in the technological development of GC/LC coupled to MS (GC–MS and LC–MS, respectively), especially high resolution/accurate mass—mass spectrometry (HR/AM-MS), allow for measuring hundreds to thousands of chemical features, represented by MS signals, in a single complex sample [6, 31]. LC–MS analytical platforms are considered "the golden standard" in exposomics research, shadowing the GC–MS analytical platforms. Despite offering highly efficient, sensitive and reproducible analysis with relatively modest cost and substantial compound coverage, GC–MS is a somewhat underestimated source of valuable complementary analytical data in CEC annotation [32]. The ultimately predominant ionisation method for the acquisition of GC–MS spectra is electron impact (EI) ionisation, along with the less frequently used chemical ionization. The great reproducibility of EI spectra, following predictable and thoroughly studied fragmentation patterns and broad internal energy distribution, promises highly accurate, yet not thoroughly explored, instrument-independent data for CSI. Even less explored is the identification of semi-volatile and thermolabile compounds using the MS data of their silylated derivatives, mainly trimethylsilyl (TMS) or *tert*-butyl dimethylsilyl (TBDMS) derivatives. While being useful in greatly enhancing the compounds' chromatographic and mass spectrometric

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

characteristics, the derivatization may complicate peak annotations due to sometimes incomplete derivatization processes, with formation of multiple and/or partially derivatized compounds. Moreover, TMS and TBDMS derivatives and their MS spectra are poorly represented in compound DBs and MS libraries (MSLs), and, accordingly, they are not readily identified using the traditional CSI approaches of compound DB(s) and/or MSL(s) search. While cheminformatics CSI approaches are expected to solve this task as well, they have been almost exclusively developed and tested using electrospray LC-(ESI)-MS/MS data and are yet to be challenged against GC-EI-MS data.

This paper presents the first application of a machine learning approach, named Compound Structure Identification:Input Output Kernel Regression (CSI:IOKR),for the identification of CEC silyl derivatives using GC-EI-MS spectra. First, we generate two unique collections of GC-EI-MS spectra of TMS derivatives: a collection curated from the NIST 17 Mass Spectral Library that is used to train a model with CSI:IOKR and a collection of GC-EI-MS spectra experimentally acquired in our laboratory that is used to test the model. Second, we evaluate the performance of the CSI:IOKR model in identifying CEC silyl derivatives. Note that we have generated our own test data (thus using different sources for the training and testing data) for two reasons: (1) to maximize the size of the training data, and (2) to obtain better estimates of the performance of the model in its intended use scenario, i.e., for identification of CEC compounds throught their silyl derivatives, on unseen data. We also investigate how identification performance depends on several factors, including the filtering of the training dataset, the overlap between compounds in the training and the test datasets, and the post-acquisition processing of the test dataset. The CSI:IOKR approach reaches satisfactory identification performance for TMS derivatives, both within and outside the training dataset, indicating its potential for use in GC–MS based annotation of contaminants.

Related work

The field of cheminformatics-assisted compound structure identification (CSI) has grown intensively over the last two decades, developing three groups of approaches (Fig. 1). The simplest ones are direct approaches, such as Mass Frontier [33], AC/MS Fragmenter [34], MOLGEN-MS [35] and MS-FINDER [36], that extract and use structural information directly from the MS spectra, represented as a set of *m/z* values of molecular ions, relative abundances of isotopologues, given the MF or fragment ions. Indirect approaches include the combinatorial fragmentation methods, e.g. FiD [37], MetFrag 2.2 [38], MAGMa [39], MolFind [40] and MIDAS [41]. Approaches from the third group, including MetExpert [42], FingerID [43], CSI:FingerID [44], CSI:IOKR [45],

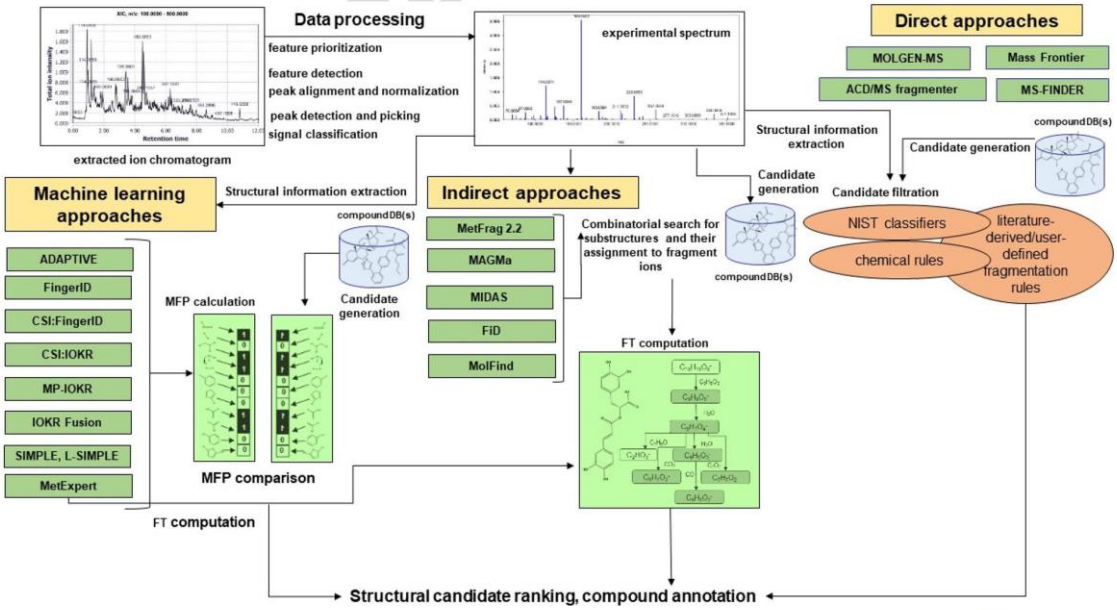


Fig. 1 An overview of direct, indirect and joint approaches for compound structure identification. Adapted from Ljoncheva et al. [28]

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

magnitude-preserving IOKR (MP-IOKR) [46], IOKRFusion [47], SIMPLE and L-SIMPLE [48] and ADAPTIVE [49], rely on the use of machine learning (ML). The third group utilizes the alternative concept of in silico spectral prediction, i.e. prediction of two dimensional (*m/z* and intensity) EI-MS (CFM-ID [50], NEIMS [51]) or ESI–MS/MS spectra (CFM-ID [52], ISIS [53]) by simulating fragmentation for a defined compound candidate set and performing CSI by comparing the measured and the in silico predicted MS/MS spectra [28]. The cutting-edge CSI approaches are more thoroughly described in recent reviews [28, 30].

In their core, the indirect "thrid-format" ML approaches transform the MS structural information into "third formats", such as MFP, molecular descriptors, or their combination, that have higher discriminatory power to reflect structural similarity and therefore lead to more accurate and confident compound structure identification. The ice-breaking ML-based approach is FingerID [43], that in a first step uses the probability product kernel (PPK) [54] directly computed from MS spectra and runs support vector machines to perform MFP predictions. In the second step, it ranks candidates from DB-derived sets according to their similarity to the predicted MFP. This method is mainly based on the information from the individual spectral peaks and ignores their interactions. The follow-up approach, CSI:FingerID [44], uses MS spectra and fragmentation trees to calculate multiple kernels combined via multiple kernel learning [55], resulting in improved predictive performance. Its disadvantage is in the long running times due to the "one-at-a-time" spectrum processing approach and computationally heavy conversions of MS spectra into fragmentation trees. The CSI:IOKR approach [45] learns mappings from MS spectra to MFP using multiple input kernels to encode similarities in the input space (MS spectra) and output kernels for encoding similarities in the output space (MFP). It predicts all components of a MFP simultaneously, resulting in a faster one-step approach. Further efforts to preserve the discrepancy between compounds in the input space, and between candidates in the output space, as well as incorporate candidate ranking information in the learning phase resulted in the development of MP-IOKR [46], with improved compound identification accuracy as compared to CSI:IOKR. The latest method in the IOKR series, IOKRFusion [47] is a score aggregation method that cobmines 60 IOKR models and 60 IOKR reverse models that learn the mapping of molecular structures into the MS/MS feature space rather than the output feature space. Finally, MFP are combined with ML prediction of retention indices and compound substructures, in silico derivatization of DBs, and metabolite-likeness evaluation in the MetExpert approach [56].

The ultimate ML-based "thrid-format" approaches exchange either the fixed, redundant MFP with novel, non-reduntant, data-driven and specific molecular vectors (ADAPTIVE [49]) or multiple kernels with a simpler prediction function, incorporating peak interactions (SIMPLE [48]). The first method combines the learning of a mapping from structures to molecular vectors utilizing message passing neural network with IOKR-based learning of the mapping from MS spectra to molecular vectors. The second method offers performance comparable to that of kernel-based methods at higher prediction speed that is proportional to the number of peaks in the queried spectrum, unlike all aforementioned kernel-based methods [30].

The most recent Critical Assesment of Small Molecule Identification (CASMI) contests (2016 [57] and 2017 [58]) identified the ML-based approaches CSI:FingerID [44], CSI:IOKR [45] and CFM-ID [59] as the most accurate compound structure identification tools, ranking as top1 and among the top10 in 17 and 34.4% (for CSI:IOKR) and more than 49% of the challenges, respectively. The challenges used LC–ESI–MS/MS spectra of reference standards. Despite their expansive development and excellent performance, the ML-based compound structure identification tools have been seldomly used in CEC research [28]: they have been used in few LC-(ESI)-MS/MS-based studies [8, 60–65], but no GC-EI-MS-based studies. In fact, only three approaches, including MetExpert [42], CFM-ID [50] and NEIMS [51] have been specifically developed to handle GC-EI-MS data, among which only the first one performs the CSI task on GC-EI-MS spectra of TMS and methoxy/TMS derivatives.

Materials and methods

Generation of the training dataset

The NIST 17 Mass Spectral Library [66] was selected as reference MSL for the generation of our training dataset. NIST 17 is the most comprehensive selection of GC-EI-MS spectra, containing 306,622 GC-EI-MS spectra of 267,376 compounds. Two of the NIST 17 libraries were searched; the main spectral library (*mainlib*), with 267,376 GC-EI-MS spectra and the replicate library (*replib*), with 39,246 GC-EI-MS spectra that are independent replicates of spectra of compounds contained in *mainlib*. *Replib* is a collection of noisier spectra as compared to *mainlib,* which reflect normally occuring experimental and instrumental response variations and make the training dataset more informative.

The spectral search was perfomed by using the NIST MS Search Program v.2.3 (NIST, 2017), with two constraints: *name fragment: trimethysilyl* and *elements allowed: Si.* The GC-EI-MS spectra were extracted in.msp file format and subsequently converted into.txt

format using the LIB2NIST conversion tool (NIST, 2011), saving the following data for each extracted GC-EI-MS spectrum: name, InChIKey, MF, M_w, exact mass, CAS number, NIST ID and MS peak list.

The originaly extracted NIST 17 entries were filtered by using the three-step approach shown in Fig. 2. The first step involved manual inspection of the spectra to retain only the Si-containing compounds generated as a result of the silylation reaction. The GC-EI-MS spectra of chemically irrelevant Si-containing compounds were removed from the dataset. Here, we set the following structural categories for exclusion:

1) Structures with Si–Si bonds (siloxanes);
2) Structures with C-Si bonds;
3) Structures with O-Si bonds other than hydroxyl/carboxyl-TMS derivatives;
4) Structures with N-Si bonds other than primary/secondary amine-TMS derivatives;
5) Structures with N–O-Si and N–N-Si bonds;
6) Structures with S–Si bonds other than thiol-TMS derivatives;
7) Structures with P-/(O/N/S)-Si bonds;
8) TMS derivatives generated as a result of a rearrangement derivatization reaction;
9) TBDMS derivatives;
10) Mixed TMS and TBDMS derivatives;
11) TMS derivatization agents;
12) TMS derivatives of inorganic compounds and
13) TMS derivatives that contain heavy metals.

As a part of the first data filtering step, we also removed erroneous NIST 17 entries, i.e., those GC-EI-MS spectral entries whose names and structures did not correspond. In the second step, GC-EI-MS spectra of TMS derivatives with $m/z \geq 1{,}000$ Da were removed, since such high molecular masses are above the working linear range of most of the mass analyzers used in GC–MS platforms. As a final data filtration step, we used four basic criteria to ensure baseline spectral quality. GC-EI-MS spectra were excluded unless they complied with all of the requirements below:

– GC-EI-MS spectra have to be acquired at the upper m/z of at least Mw of the derivative + 10 amu;
– GC-EI-MS spectra have to contain the molecular ion [M] + peak and at least one of the isotope peaks, such as the 13C isotope peak;
– GC-EI-MS spectra have to contain peaks of fragment ion specific for TMS groups (m/z 73, 147, 221 and 295, corresponding to one, two, three and four TMS groups, respectively) and

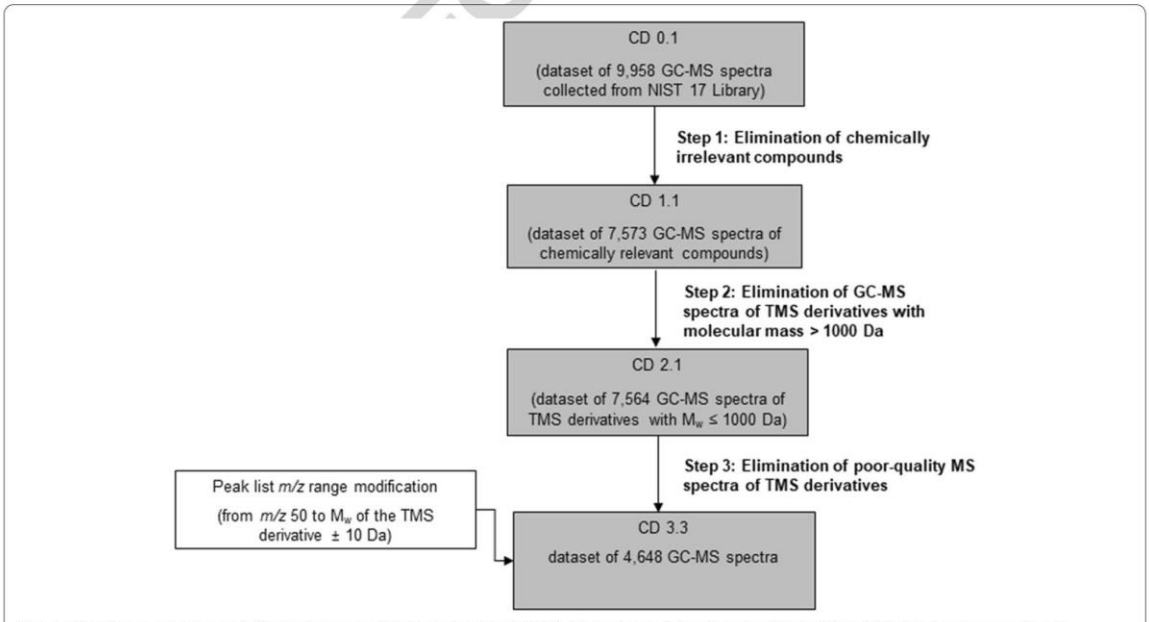


Fig. 2 The three-step spectral filtering process. CD 0.1 is the initial GC-EI-MS training dataset curated from NIST, CD 1.1 is the step 1-refined GC-EI-MS dataset, CD 2.1 is the step 1 and 2-refined GC-EI-MS dataset, while CD 3.3 is the final GC-EI-MS dataset that has gone through the three steps of spectral filtering

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

– GC-EI-MS spectra have to contain at least five fragment ion peaks.

Generation of the test dataset

Chemicals and reagents

From the in-house pool of reference standards, we selected 129 compounds with potential environmental relevance and at least one functional group amenable to TMS derivatization. Preliminary derivatization experiments showed that 100 compounds out of 129 could get successfully derivatized. The list and the basic description of the selected reference standards and other chemicals and reagents used in this study is given in Additional file 1. The compounds are of anthropogenic origin and are potentially bioactive CECs. In order to verify their environmental relevance, the componds were searched against CCD [11], followed by predicting their environmental properties. US EPA's Toxicity Estimation Software Tool (T.E.S.T.) [67] was used to predict the common toxicity endopoints: 96 h fathead minnow LC_{50}, developmental toxicity and estrogen receptor binding affinity. The Estimation Programs Interface (EPI) Suite™ v.4.11 [68] was used to predict the log carbon–water partitioning coefficient (log K_{OC}), log octanol–water partitioning coefficient (log K_{ow}), water solubility, bioaccumulation factor, bioconcentration factor, biotransformation half-life, half-life in river and half-life in lake, for each of the compounds. To be considered for the test dataset, a compound had to fulfill at least three of the following five criteria, established in accordance with the Regulation (EC) No 1907/2006 of the European Parliament and the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Annex XIII [69]:

1) Positioning (R): the compound is present in the US EPA CCD [11], the most comprehensive repository of EE constituents;
2) Persistence (P): compound's half-life in fresh or estuarine water is >40 days;
3) Bioaccumulation (B): bioaccumulation factor and/or bioconcentration factor >2000, or in absence of such data, $\log K_{ow} \geq 5.0$;
4) Mobility (M): compound's water solubility is ≥0.15 mg/L and log K_{oc} is ≤4.0, i.e. between −10.0 and 4.0 and
5) EcoToxicity (T): long-term no-observed-effect concentration (NOEC) for marine or freshwater organisms is <0.01 mg/L. Here, instead of NOEC, chronic acquatic toxicity (mg/L) for fish, daphnid, and green algae is considered, calculated as the geometric mean of NOEC and lowest observed effect concentration (LOEC).

The results of the Comptox-T.E.S.T and EPI Suite™ predictions are given in Additional file 2.

Silylation

The individual SSS of each compound at the concentration of approximatelly 150 µg/mL were prepared in EtAc, MeOH or ACN, depending on the solubility of the reference compound (Table 1). The SSSs were kept at +4 °C and were diluted to prepare working solutions (WSs) at the concentration of 1 µg/mL, which were used within 7 days. TMS derivatives were prepared individually, by mixing 150 µL of a WS with 30 µL of a derivatization agent (MSTFA, BSTFA or BSTFA + 1% TMCS, depending on the derivatization yield determined during the preliminary derivatization experiments). For compounds dissolved in MeOH, the solvent was removed under gentle steam of N_2 prior to the addition of the derivatization agent, which was folowed by reconstitution in 150 µL EtAc and vortexing for 1 min. Derivatization conditions (temperature, time) were selected based on prior optimization, so that compounds were derivatized under either of the following conditions: (1) at 60 °C for 45 min; (2) at 70 °C for 90 min or (3) at 70 °C for 45 min.

GC-EI-MS spectra acquisition and dataset compilation

GC-EI-MS spectra were acquired on Agilent 7890B/5977A series GC-MSD (Agilent, USA). Separation was achieved on Agilent DB-5MS UI fused-silica capillary column (30 m × 0.25 mm × 0.25 µm; Agilent, USA). He of 99.99999% purity at the flow rate of 1.2 mL/min was used as a carrier gas. The manifold, ion source and transfer line temperatures were set at 230 °C, 150 °C and 250 °C, respectively. Injections (1 µL) were performed in the splitless mode. Depending upon compound properties, one of the following column oven temperature programs was used: (1) initial temperature 70 °C (held 1 min), ramped at 15 °C/min to 280 °C (held 1 min); total runtime: 16 min; (2) initial temperature 70 °C (held 1 min), ramped at 20 °C/min to 240 °C (held 1 min), at 12 °C to 310 °C (held 2 min); total runtime: 18.3 min and (3) initial temperature 70 °C (held 1 min), ramped at 20 °C/min to 240 °C (held 1 min), at 12 °C to 310 °C (held 4 min); total runtime: 20.3 min. The MSD was operated in EI ionization mode (70 eV) by scanning over the mass range of *m/z* 50–800 amu. Mass Hunter Qualitative Analysis v B.07 (Agilent, USA) was used to reduce raw instrument data to two-dimensional peak lists (*m/z*, abundance) and to perform background subtraction (BS).

In-between the acquisitions of the derivatized standards, EtAc was run as the solvent check to assess

Table 1 Optimized derivatization and acquisition conditions for CEC-TMS derivatives from the test dataset

CEC (abbreviations are provided in Additional file 4)			Dissolved in	Derivatization agent and conditions	GC oven programme (see "GC-EI-MS spectra acquisition and dataset compilation" section)
SA LAA SHA	CBD QA	AMP MAMP	MeOH	MSTFA, 60 °C, 45 min	(1)
CBC	Δ^9-THC	CBN		BSTFA + 1% TCMS, 70 °C, 90 min	(2)
11N9THC T3HC 110HTHC 6-MAM	BZECG LLEU COD	LSER MORPH ERY		BSTFA + 1% TCMS, 70 °C, 45 min	(3)
BA PrPb MePb IBuPb	EtPb BuPb IPrPb	TCS IB BzPb	EtAC	MSTFA, 60 °C, 45 min	(1)
RES CBZ CLA DF 9-HF E1	HPP E2 4-NP E3 NAP EE2	DH-BP BP-8 4,4′-BP SFA KET		BSTFA + 1% TCMS, 70 °C, 90 min	(2)
22BPF BPBP 3M5NC BPAF BPPH 4NC BPF BPFL SYE BPE DHDPE 4-NS BPA PAA PCA BPC 2AA	CA MCA BPB CLP OCA BP26DM AA 17HP BPCL 8-HQ 5AD BPZ 4-OP BD BPS CBDA 6HP	11HT BPAP 4-NG 11HAD H-BP 5-NG ST BHT 6-NG 2APA BPM CAT ET BPP 3MC NX		BSTFA + 1% TCMS, 70 °C, 45 min	(3)
UA			ACN	MSTFA, 60 °C, 45 min	(1)
Δ^9-THCA-A				BSTFA + 1% TMCS, 70 °C, 90 min	(2)
LTYR				BSTFA + 1% TMCS, 70 °C, 45 min	(3)

potential background interferences, carryover and sample contamination and was used for background subtraction as a part of the post-acquisiton processing of the GC-EI-MS spectra. The test GC-EI-MS dataset was compiled as.txt file that included MF, InChIKey strings, M_w and two-dimensional peak lists. Molecular stereochemistry was not considered, since stereoisomers are not readily distinguished by MS.

GC-EI-MS spectral similarity analysis and selection

For each TMS-derivative, multiple ($\geq$15) GC-EI-MS spectra were generated for the experimental dataset. In order to estimate spectral reproducibility (and therefore the "interchangeability") of the GC-EI-MS spectra of a TMS derivative, the cosine similarity was calculated. An R script was written to read GC-EI-MS spectra, perform binning in 1.0 Da bins and an intensity N-dimensional vector is constructed in which element v_i corresponds to the average peak intensity of all peaks within the bin. The cosine similarity c between spectra v and u was calculated as the dot product of the two vectors divided by the product of their norms (Eq. 1):

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

$$c = \frac{\sum_i v_i u_i}{\|v\| \|u\|} \quad (1)$$

giving values between 0 and 1, with 0 indicating that the spectra share no common peaks and 1 indicating that the spectra are identical. Cosine similarity is calculated in all-against-all manner in both the RAW and BS experimental datasets. The influence of background subtraction on spectral reproducibility and similarity was explored by calculating the cosine similarity between each raw GC-EI-MS spectrum in the RAW dataset and its corresponding background-subtracted GC-EI-MS spectrum in the BS dataset for each TMS derivative. The results are visualized as a separate cosine similarity measure matrix for each TMS derivative. Finally, in order to explore mass spectral similarities of the TMS derivatives included, a consensus spectrum was built from all binned GC-EI-MS spectra for each TMS derivative. Clustering was then performed of the consensus spectra for both RAW and BS datasets using the distance matrices of all against all consensus RAW, and respective BS spectra.

CSI:IOKR protocol

Identification of TMS derivatives was performed by using a simplified version of CSI:IOKR [45]. The workflow is given in Fig. 3. CSI:IOKR is a kernel-based method, where a kernel function is a positive semi-definite function that measures similarity between two elements [45]. In our study, input kernels measure similarity between MS spectra, while output kernels measure similarity between molecular propeties represented as MFPs. PPK [54] was used as an input kernel, and the linear kernel calculated on MFPs was used as an output kernel. The PPK kernel is computed from MS spectra, by modelling each peak in a spectrum as a normal distribution with two dimensions: *m/z* and intensity, and modelling an MS spectrum as a mixture of normal distributions. The PPK kernel is evaluated by integrating the product between the two corresponding mixture distributions [45]. The kernels were centralized and normalized. The strength of regularization for IOKR was determined with internal cross-validation on the training dataset, as proposed by Brouard et al. [45].

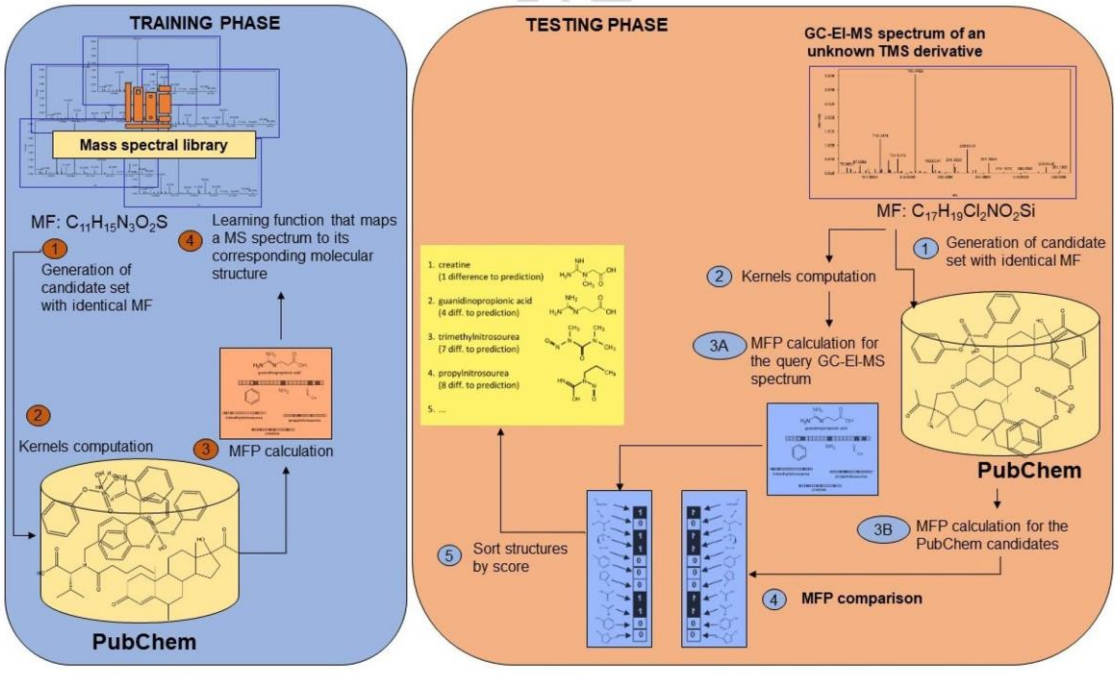


Fig. 3 Workflow of CSI:IOKR protocol we employ, based on Dührkop et al. [72] and Brouard et al. [45]. The model is trained on a set of MFPs generated from GC-EI-MS data from the NIST 17 MSL. PPK kernel is computed for each pair of compounds and MFPs are calucalted for each compound from the training NIST 17 MSL dataset in preparation for learning the kernel-based model. The model is tested using GC-EI-MS data of (unknown) TMS derivartives, by calculating the MFPs of candidate compounds with identical MFs, comparing these MFPs with the MFP predicted by the IOKR kernel-based model from the GC-EI mass spectrum and producing and ordering the list of candidate structures by the similarity of their MFPs with the predicted one

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

In the pre-image step, we assume the MFs of the TMS derivatives of compounds corresponding to the GC-EI-MS spectra from the test dataset to be known: This is certainly true if the GS-EI-MS spectra are generated for testing purposes, as in our case, but note that the MF corresponding to a given MS can be also obtained by using software such as SIRIUS [70] We use these MFs to generate candidate set of compounds from PubChem [15] with a MF identical to the MF for any test TMS derivative. The InChIKeY strings of PubChem candidates are retrieved by submitting queries to PubChem's Power User Gateway (PUG) through extensible markup language (XML) and further stored for MFP calculation. For each challenge GC-EI-MS spectrum of a TMS derivative and PubChem candidate, four types of MFPs were calculated and concatenated by using the Chemistry Development Kit (CDK) [71]: substructure fingerprints (307 molecular properties), MACCS fingerprints (166 molecular properties), PubChem (CACTVS) fingerprints (881 molecular properties) and Klekota-Roth fingerprints (4,860 molecular properties), giving 6,214 molecular properties in total. Of these, 3,215 molecular properties were removed, as they were either duplicates or were constant through the entire training dataset. This resulted in 2,999 bit-long vectors describing the structures of the TMS derivatives.

We used IOKR for model learning on both the raw and the curated datasets (CD 0.1 and CD 3.3, respectively; see Fig. 2). We used the learned models to make predictions for the two test datasets, of raw (RAW) and background-subtracted spectra (BS). All experiments were performed on a computer with a 2.7 GHz Intel Core processor. The computer code was written in Phyton and MATLAB.

Results and discussion

Generation of the training dataset

Using the NIST MS Search Software, the initial traning dataset of GC-EI-MS spectra (CD 0.1) was generated, consisting of 9,958 GC-EI-MS spectra (Fig. 2). In the first step, the GC-EI-MS spectra of chemically irrelevant compounds were removed. These compounds contained in their chemical structures Si atom(s) that were not part of a TMS group, but belonged to one of the structural categories for exclusion (see "Generation of the training dataset" section.). This resulted in the removal of 2,385 GC-EI-MS spectra (24%), yielding the refined dataset (CD 1.1) of 7,573 GC-EI-MS spectra. The remaining collection of GC-EI-MS spectra comprises compounds consisting of the 11 most typical elements in organic chemistry: C, H, N, O, P, S, Br, I, F, Cl and Si [73, 74]. Further 9 GC-EI-MS spectra of high-mass TMS derivatives ($M_w > 1000$) and 2,925 GC-EI-MS spectra of insufficient quality were removed in the second and third filtration step, respectively. The final traning dataset, CD 3.3, consists of 4,648 GC-EI-MS spectra (of 3,948 TMS derivatives), which is 47% of the initial CD 0.1 dataset. After the third filtering step, a final modification in which the *m/z* range was set to *m/z* 50 up to $M_w + 10$ Da was made to the 4,648 spectra remaining in the final version of the training dataset.

Generation of the test dataset

The predictions, the criteria and the results from the environmental evaluation of the compounds considered for the generation of the test dataset are described in detail in Additional file 2. The evaluation of the 100 compounds selected for generating the test dataset of GC-EI-MS spectra (see Additional file 3) revealed significant environmental relevance for the majority of the test compounds. Briefly, 96 compounds meet at least three RPMBT classification criteria (see "Chemicals and reagents" section.), while four compounds (3-methyl-5-nitrocatechiol (3M5NC), 4-nitrosyringol (4NS), 6-hydroxypregnenolone (6HP) and 11-hydroxytestosterone (11HT)) do not, though according to the Regulation EC 1907/2006 [69], they can be considered as persistent, mobile and toxic compounds (Additional file 3).

The derivatization experiments resulted in the formation of 104 TMS derivatives with M_w ranges from 182 to 575 Da (see Additional file 4). The optimised derivatization and acquisition conditions can be found in Table 1. During the acquisition, no significant sample contamination or carryover was detected. Baseline subtraction was still performed to remove constantly present background signals, such as those originating from common GC–MS contaminants, e.g., *m/z* 149 as a typical phtalate interference, *m/z* 282, *m/z* 256 and *m/z* 284 for oleic, palmitic and stearic acid, and *m/z* 207, *m/z* 281 and *m/z* 327 of common polysiloxanes resulting from GC column stationary phase degradation. The raw GC-EI-MS spectra were assigned to the RAW test dataset. After background subtraction, the resulting spectra were assigned to the BS test dataset.

GC-EI-MS spectral similarity analysis and selection

The most widely used, reliable and accurate way of comparing MS spectra is to quantify the fraction of shared peaks by using cosine-based similarity scores that rely on multiplying the intensities of matching peaks [75]. When multiple EI-MS spectra of the same compound are acquired, it is necessary to understand whether each particular MS spectrum should be taken into account, and if not, which one(s) should. To validate the hypothesis that GC-EI-MS spectra of the same compound (here, TMS derivative) are highly reproducible/similar, we performed an all-against-all cosine similarity comparation within the RAW and BS experimental dataset. While the estabilshed

Table 2 The identification accuracies of CSI:IOKR on different training and test datasets

Training dataset	Test dataset	Presence of the test compounds in training dataset	Number of test compounds	Missing n (%)	Top 1 n (%)	Top 10 n (%)	Top 20 n (%)	$\overline{ARP}$	$\overline{RRP}$
CD 0.1	RAW	Yes	63	9 (14.3)	1 (1.6)	10 (15.9)	18 (28.6)	59.8	0.79
CD 0.1	RAW	No	41	23 (56.1)	2 (4.9)	9 (22.0)	16 (39.0)	24.7	0.69
CD 0.1	RAW	Merged	104	32 (31.8)	3 (2.9)	19 (18.3)	34 (32.7)	52.0	0.77
CD 0.1	BS	Yes	62	8 (12.9)	1 (1.6)	10 (16.1)	18 (29.0)	60.0	0.79
CD 0.1	BS	No	41	23 (56.0)	2 (4.9)	9 (22.0)	16 (39.0)	24.9	0.72
CD 0.1	BS	Merged	103	31 (30.1)	3 (2.9)	19 (18.5)	34 (33.0)	52.2	0.77
CD 3.3	RAW	Yes	63	9 (14.3)	7 (11.1)	25 (39.7)	37 (58.7)	23.8	0.37
CD 3.3	RAW	No	41	23 (56.1)	4 (9.8)	14 (34.2)	16 (39.0)	11.3	0.35
CD 3.3	RAW	Merged	104	32 (30.8)	11 (10.6)	39 (37.5)	53 (51.0)	21.0	0.36
CD 3.3	BS	Yes	62	8 (12.9)	4 (6.5)	24 (38.7)	36 (58.1)	26.2	0.39
CD 3.3	BS	No	41	23 (56.1)	5 (12.2)	14 (34.2 5)	16 (39.0)	11.0	0.36
CD 3.3	BS	Merged	103	31 (30.1)	9 (8.7)	38 (36.9)	52 (50.5)	22.8	0.38

For each experimental setup, the total number of CEC-TMS derivatives, the number (n) and percentage (%) of missing CEC-TMS derivatives, and CEC-TMS derivatives correctly ranked in the top 1, 10 and 20 hits (top k accuracies), average absolute ranking position ($\overline{ARP}$) and average relative ranking position ($\overline{RRP}$) are given.

cosine similarity threshold value is 0.50, the minimum cosine similarity for most of the TMS derivative pairs was higher than 0.95 (Table 2). There are very few TMS derivatives for which a pair of spectra existed either in RAW 2-anilinophenylacetic acid-TMS (2APA-TMS), BS (cannabidiolic acid TMS (CBDA-TMS), nitroxoline TMS (NX-TMS)) or in both experimental datasets (L-tyrosine TMS (LTYR-TMS), salicylic acid TMS (SA-TMS)) that yielded a minimum cosine similarity factor below 0.50. Despite these few observed discrepances, we kept all the GC-EI-MS spectra of these TMS derivatives in the experimental datasets.

Further, for 2APA-TMS, 17α-ethinyl estradiol TMS (EE2-TMS), estriol TMS (E3-TMS), NX-TMS, LTYR-TMS, L-leucine (LLEU-TMS) and L-serine TMS (LSER-TMS) the minimum cosine similarity between a pair of RAW and BS MS spectra was below 0.50. Moreover, for the latter two TMS derivatives also the maximum cosine similarity factor did not exceed 0.50. Such values indicate that significant changes in MS spectra occur when background subtraction is performed. An example TMS derivative with highly reproducible spectra is given in Fig. 4A, together with an example TMS derivative where GC-EI-MS spectra are less reproducible (Fig. 4B), where green color indicates high cosine similarity (0.99–1.00), yellow color indicates medium cosine similarity (0.51–0.98) and red color indicates low cosine similarity (below 0.50).

Still, the reproducibility of GC-EI-MS spectra of TMS derivatives is overall satisfactory. Any of the acquired GC-EI-MS spectra of each TMS derivative can thus be used to test the CSI:IOKR model. This is clearly visible in Additional file 5, where very few TMS derivatives have pairs of GC-EI-MS spectra of low similarity, i.e. factor below 0.50. Despite these few observed discrepances, we kept all the GC-EI-MS spectra of these TMS derivatives in the experimental datasets.

CSI: IOKR

The protocol

CSI:IOKR was used to identify CECs from GC-EI-MS spectra of their TMS derivatives. While many different kernels have been proposed in the literature [43–46, 48], it is well known that kernel-based supervised ML methods have computational complexity issues, particularly when using complex kernels. They can have high predictive performance at the price of a heavy computational load. Lead by this knowledge, we used two simple kernels, namely the PPK as input and the linear kernel as output kernel. The PPK is computed from a spectrum by modeling each peak in the MS as Gaussian distribution, where the *m/z* ratio and intensity represent the dimensions, and modeling the whole spectrum as a mixture of normal distributions. All-against-all matching is performed by integrating the product between the two corresponding distribution mixtures. This kernel is shown to be superior to simple peak and loss matching kernels computed directly from the spectra (without the knowledge of fragmentation trees) [43, 44]. Among the 24 input kernels of the CSI:IOKR model, PPK was one of the best performing kernels and was assigned the highest weight in the ALIGNF approach of Brouard et al. [45]. The linear kernel was selected as output kernel based on the evaluation results of Brouard et al. [45], where it performed comparably to the polynomial kernel and insignificantly worse than the Gaussian kernel (30.02% vs. 30.66% with

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

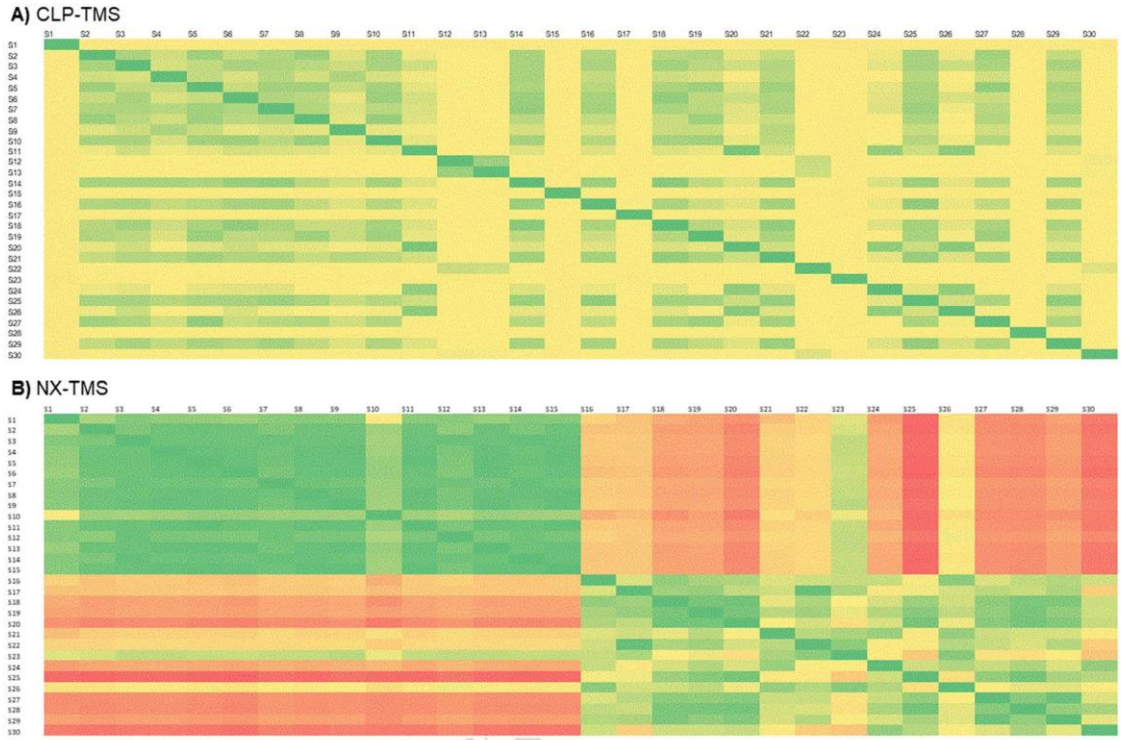


Fig. 4 Cosine similarity matrices for multiple spectra of **A** CLP-TMS RAW and **B** NX-TMS RAW. Green color indicates high cosine similarity (0.99–1.00), yellow color indicates medium cosine similarity (0.51–0.98) and red color indicates low cosine similarity (below 0.50)

the UNIMKL approach, 28.54% vs. 29.78% with the ALIGNF approach). PPK as the input and linear kernel as the output kernel were also the best performing kernels in the IOKRFusion method [47].

The performance of IOKR with the two selected kernels was evaluated on each of the test sets. The identification accuracy was evaluated by using three metrics: (1) the top-k aaccuracy, that corresponds to the percentage of test TMS derivatives for which the correct structural candidate is found among the *k* top ranked candidates; (2) the average absolute ranking position ($\overline{\text{ARP}}$), the average of ARP values for all CEC-TMS,defined as the number of candidates with better ranking than the correct compound plus 1 and (3) the average relative ranking position ($\overline{\text{RRP}}$), of RRP values for all CEC-TMS [76], calculated as:

$$\text{RRP} = \frac{1}{2}\left(1 + \frac{\text{BC} - \text{WC}}{\text{TC} - 1}\right) \quad (2)$$

where BC denotes the number of candidates that are better scored than the correct candidate, WC denotes the number of candidates which are ranked lower, i.e., worse than the correct candidate and TC denotes the total number of candidates. The $\overline{\text{RRP}}$ ranges from 0 to 1, with $\overline{\text{RRP}} = 0$ if the correct candidate is ranked first and $\overline{\text{RRP}} = 1$ if the correct candadate is ranked last. For each IOKR run, the TMS derivatives missing from the PubChem canddidates pool were refered to as "missing".

Performance results

The results of evaluating the performance of CSI:IOKR are gathered in Table 2. First, we investigated whether the filtering of the training dataset and the post-acquisition processing of the test dataset affected the performance. The spectral filtering of the training dataset involved the steps illustrated in Fig. 2, whereas the post-acquisition processing only involved baseline subtraction. As evident in Table 2 and Fig. 5, lower performance was achieved when using the unfiltered NIST GC-EI-MS dataset (CD 0.1) in the learning phase, for both test datasets. 2–fourfold increase of the top-*k* accuracies was observed when the 3-step filtered NIST GC-EI-MS dataset (CD 3.3) was used to train the model (instead of CD 0.1). Also, the $\overline{\text{ARP}}$ and $\overline{\text{RRP}}$ improved twofold with the CD 3.3 dataset. For example, the $\overline{\text{ARP}}$ of the correct TMS derivative was

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

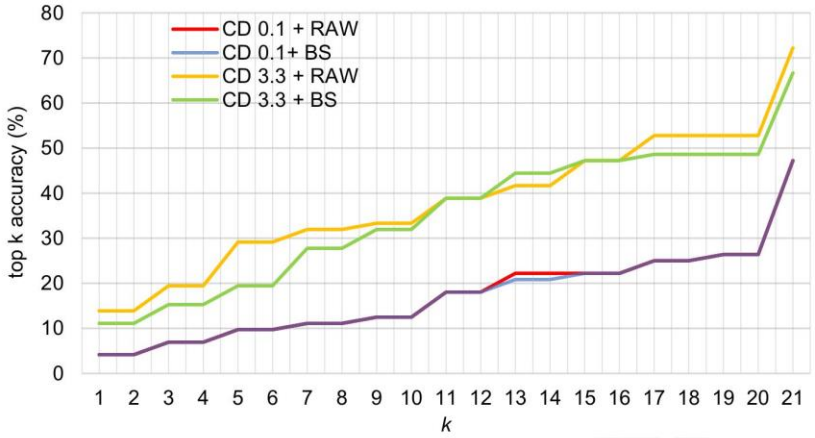


Fig. 5 Plot of top-k accuracy for CSI:IOKR with different training and test datasets. CD 0.1 + RAW (red line); CD 0.1 + BS (blue line); CD 3.3 + RAW (yellow line) and (CD 3.3 + BS (green line)

31 positions and 29 positions higher for the RAW and BS datasets, respectively. As evident in Fig. 5, very subtle differences of less than 2% appeared between performance on the RAW and BS test datasets in all experiments, slightly favoring the RAW test dataset, especially when CD 3.3 was used to train the model. However, the $\overline{RRP}$ values were comparable for both the RAW and the BS test sets with both the CD 0.1 and the CD 3.3 training sets, confirming that this baseline subtraction is not important for the identification task. Therefore, we consider that the CSI:IOKR model performs best when trained using the CD 3.3 training dataset and tested on RAW test dataset. Thus, furhter evaluation of the CSI:IOKR performance is done based on the results from CD 3.3 + RAW.

For each experimental setup, the total number of CEC-TMS derivatives, the number (n) and percentage (%) of missing CEC-TMS derivatives, and CEC-TMS derivatives correctly ranked in the top 1, 10 and 20 hits (top k accuracies), average absolute ranking position ($\overline{ARP}$) and average relative ranking position ($\overline{RRP}$) are given.

Further, we compared the perfomance of CSI:IOKR for two subgroups of TMS derivatives from the test set, i.e., those with GC-EI-MS spectra within and outside the training dataset (»presence in traning dataset« Yes/No, Table 2). The results show better identification performance for the GC-EI-MS spectra that were part of the traning dataset for the CD 3.3 dataset. The differences in performance are small and their direction is unclear for the CD 0.1 training set, expecially for the top 1 metric. The underlying reason may be that the size of the candidate sets was typically much lower for the group of TMS derivatives that were not part of the training dataset, reflecting the high number of TMS derivatives that are not part of PubChem.

With this in mind, we investigated the relation between candidate set size and identification performance. The distribution of candidate sets sizes is presented in Fig. 6. The maximum size of a candidate set was less than 400, while the majority candidate sets (about 50%) consisted of 0–25 candidates (Fig. 6A). According to the results (Fig. 6B), the difficulty of the identification task does not seem to strongly depend on the size of the candidate set, as the method is able to correctly identify a significant proportion of test compounds within the top 1 and top 10 candidates, even for larger candidate sets [45]. For 32 challenges from the test set, their corresponding candidate sets did not contain the correct compound.

Relating the number of candidates within each PubChem candidate set with the percentage of candidates ranked higher than the correct compound (Fig. 7) did not reveal any specific pattern, regardless of whether the TMS derivatives had their spectra within or outside of the training dataset. The results indicate that the influence of the size of the PubChem candidate sets on the identification accuracy is negligible. That is, the CSI:IOKR model, in a percentage-wise manner, does not perform worse with larger candidate sets. However, this may not yield satisfactory performance when the correct compound is, for example, ranked at position 100 among 1000 candidates. In this case, the percentage is good, while the rank itself s not.

In order to investigate the ability of CSI:IOKR to identify particular groups of TMS derivatives, we divided the latter into 6 structural TMS classes, based on the moiety that the TMS group was attached to (Additional file 6).

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

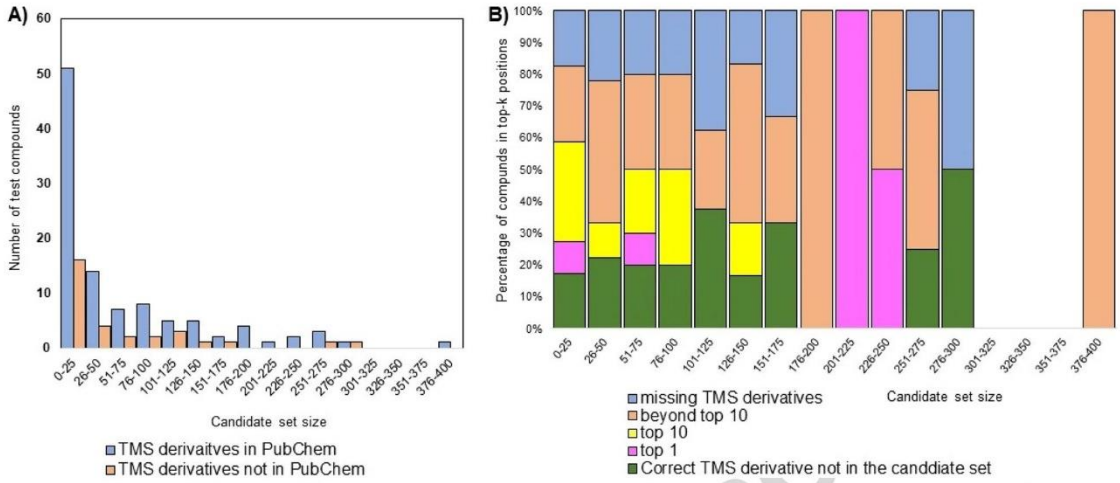


Fig. 6 **A**: Distribution of PubChem candidate set sizes for the test set spectra. Blue color represents the TMS derivatives that are in PubChem, and orange color represents the TMS derivatives that are not in PubChem **B**: Distribution of TMS derivatives from test dataset; the y-azis represents the percentage of correct TMS derivatves not present in their canddiate set (green color); percentage of correct TMS derivatives that are ranked top 1 (pink color); percentage of correct TMS derivatives ranked in the top 10 (yellow color); percentage of TMS derivatives that are ranked below the top 10 positions (orange color) and percentage of missing TMS derivaitves, i.e. TMS derivaitves that are not in the PubChem candidate set (blue color) among the test TMS derivatives with candidate set sizes in each size bin, while the x-axis represents the candidate set size bins

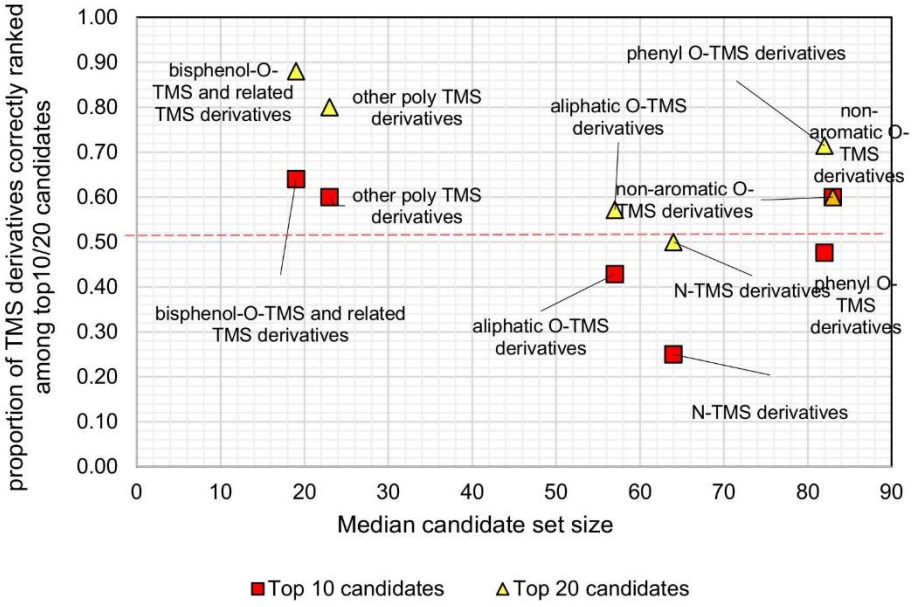


Fig. 7 The percentage of candidates ranked higher than the correct compound versus the size of each PubChem candidate set. Red triangles represent the percentage of TMS derivatives that are not present in the training dataset; the yellow squares represent the oercentage of TMS derivatives that are present in the training dataset

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | LE | TYPESET
MS Code : | CP | DISK

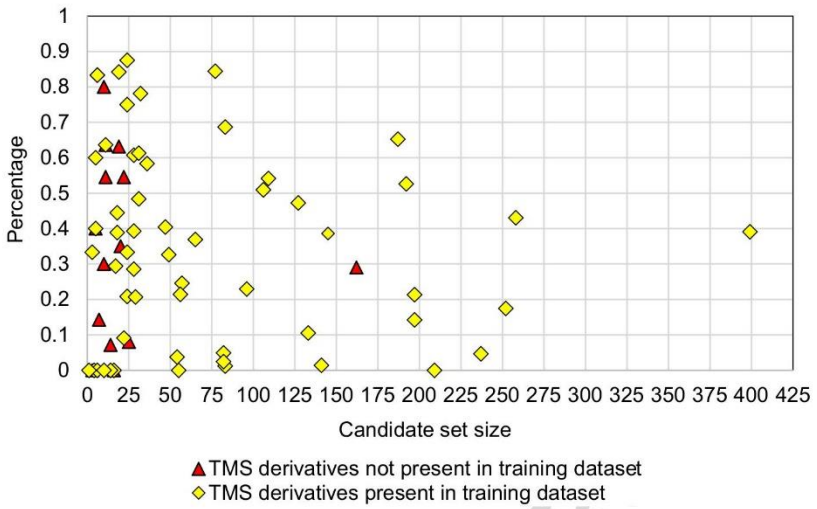


Fig. 8 The median number of candidates in the candidate sets of different classes of TMS derivatives plotted against the proportion of challenge TMS derivatives within the clusters correctly ranked within the top 10 candidates (red squares) and top 20 candidates (yellow triangles)

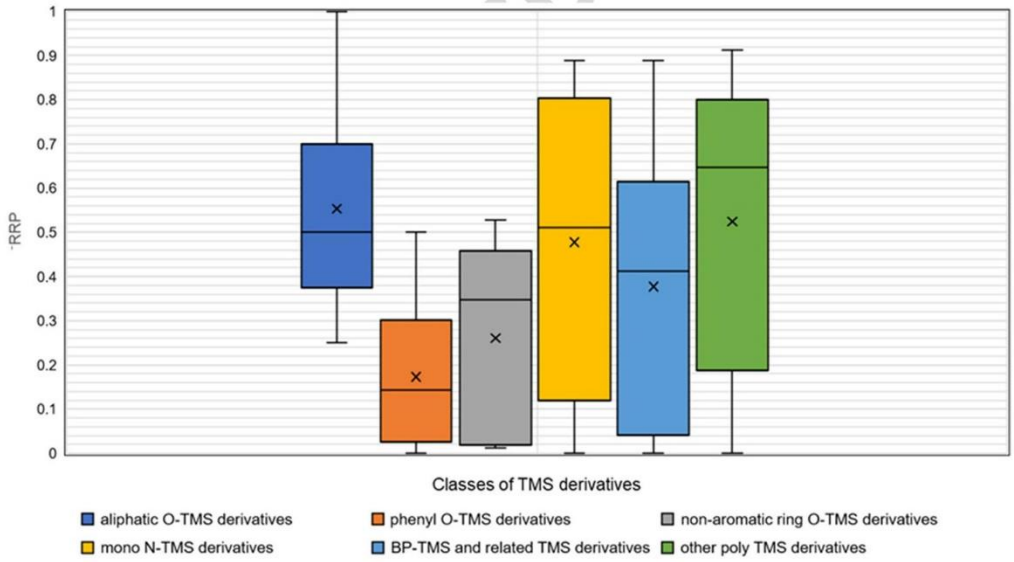


Fig. 9 Box plot representing the distribution of the $\overline{RRP}$ of the challenge compounds in each class of TMS derivatives. Dark blue plot represents the aliphatic O-TMS derivatives, orange plot represents phenyl O-TMS derivatives, grey plot represents the non-aromatic ring O-TMS derivatives, the yellow plot represents the mono NT-MS derivatives, the light blue plot represents the bisphenol-TMSs and related TMS derivatives, while the green plot represents the other poly TMS derivatives

For each TMS class, the median number of candidates in all candidate sets in the class was plotted against the proportion of TMS derivatives for which the correct candidate was ranked among top 10 and top 20 candidates (Fig. 8) and average $\overline{RRP}$ (Fig. 9). The TMS derivatives for which the correct candidate was absent from the corresponding candidate sets were omitted.

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

For all TMS classes, CSI:IOKR performs satisfactorily both in terms of the proportion of TMS derivatives correctly ranked among the top10/20 candidates and in terms of the $\overline{RRP}$ of the challenge TMS derivatives. Except for aliphatic O-TMS derivarives and N-TMS derivatives, $\geq$50% of the correct TMS derivatives are ranked among the top 10 candidates. Especially good ranking scores are achieved for the poly TMS derivatives, i.e., bisphenol O-TMS derivatives and related TMS derivatives, and the other poly TMS derivatives, including mixed O,N-TMS and N-TMS derivatives, that have highest M_w and lowest median candidate size,which may partially contribute to their relatively good ranking. Namely, the correct CEC-TMS was ranked on average positions 10.68 and 19.50, respectively, while the average PubChem candidate set size was 22.04 and 28.60, respectively, which is 2–5 times lower than the values for the other TMS classes. Also evident from Fig. 8 is that CSI:IOKR performs solidly for phenyl O-TMS and non-aromatic O-TMS derivatives, which yield relatively high average candidate set sizes (108.43 and 120.67, respectively, data not shown). Despite that, their ranking scores are satisfactory, as well as their average RRP. The class of non-aromatic O-TMS derivatives contains 5 CEC-TMS derivatives, and thus the number of CEC-TMS derivatives is not representative, so that solid conclusions can be extracted. On the other hand, the phenyl O-TMS class is represented by 21 CEC-TMS, with low average ranking position (19.14), but hihg average PubChem candidate set size (108.43). Here, a factor that may positively contribute to the good ranking of some structural classes is the specificity of the fragmentation patterns, leading to uniqueness of its GC-EI-MS spectrum, which is responsible for the good ranking, independent of the sizr of the PubChem candaidate set. Finally, $\overline{RRP}$ is >0.50 or close to 0.50 (the threshold of satisfactory accuracy) for all TMS classes, except for phenyl-O-TMS derivatives (data not shown).

Clustering of MS spectra for the RAW (Fig. 10A) and the BS dataset (Fig. 10B) revealed 6 and 4 clusters, respectively. The RRP and proportion of TMS derivatives ranked among top 10/20 candidates differed significantly between the clusters of TMS derivatives with significant MS spectral similarity. The median candidate sizes for all clusters (except for cluster 3) were <35 candidates. For all of them (except for cluster 6, where the top 10 ratio is 0.4444), top 10 and top 20 ratios of >0.55 were achieved (Fig. 11A). $\overline{RRP}$ values vary significantly within all clusters, with average $\overline{RRP}$ <0.60 and clusters 2 and 5 having the lowest average $\overline{RRP}$ s (0.264 and 0.208) (Fig. 11B).

Legend: 1: BPAF-TMS; 2: DHBP-TMS; 3: 2APA-TMS; 4: 3M5NC-TMS; 5: CLP-TMS; 6: 4MC-TMS; 7: 4,4-BP-TMS; 8: HPP-TMS; 9: HB-P-TMS; 10: 4NC-TMS; 11: 4NG-TMS; 12: 4NS-TMS; 13: 4NP-TMS; 14: 4OP-TMS; 15: 5A3B17B-TMS; 16: 5NG-TMS; 17: 6HP-MS; 18: 6MAM-TMS; 19: 6NG-TMS; 20: 8HQ-TMS; 21: 9HF-TMS; 22: 11AHA-TMS; 23: 11AHT-TMS; 24: 11OHTHC-TMS; 25: 11N9THC-TMS; 26: E2-TMS; 27: EE2-TMS; 28: 17HPT-MS; 29: AA-TMS; 30: AMPH-TMS; 31: PAA-TMS; 32: BA-TMS; 33: BZECG-TMS; 34: BzPb-TMS; 35: 22BPF-TMS; 36: 24BPF-TMS; 37: BPA-TMS; 38: BPAP-TMS; 39: BPB-TMS; 40: BPBP-TMS; 41: BPC-TMS; 42: BPCL-TMS; 43: BPE-TMS; 44: BPF-TMS; 45: BPFL-TMS; 46: BPM-TMS; 47: BPP-TMS; 48: BPPH-TMS; 49: BPS-TMS; 50: BPZ-TMS; 51: BD-TMS; 52: BP26DM-TMS; 53: BuPb-TMS; 54: BHT-TMS; 55: CBC-TMS; 56: CBD-TMS; 57: CBDA-TMS; 58: CBN-TMS; 59: CBZ-TMS; 60: CAT-TMS; 61: CA-TMS; 62: CLA-TMS; 63: COD-TMS; 64: THC-TMS; 65: THCA-TMS; 66: DF-TMS; 67: BP8-TMS; 68: ERY-TMS; 69: E3-TMS; 70: E1-TMS; 71: EtPb-TMS; 72: ET-TMS; 73: IB-TMS; 74: IbUPb-TMS; 75: IPrPb-TMS; 76: LLEU-TMS; 77: LAA-TMS; 78: LLEU-TMS; 79: LSER-TMS; 80: LTYR-TMS; 81: MCA-TMS; 82: MAMPH-TMS; 83: MePb-TMS; 84: MORPH-TMS; 85: NAP-TMS; 86: NTX-TMS; 87: OCA-TMS; 88:PCA-TMS; 89: PrPb-TMS; 90: QA-TMS; 91: RES-TMS; 92: SA-2TMS; 93: SA-TMS; 94: SHA-TMS; 95: STA-2TMS; 96: STA-TMS; 97: SFA-2TMS; 98: SFA-TMS; 99: SYR-TMS; 100: T3HC-TMS; 101: TCS-TMS; 102: DHDPE-TMS; 103: UA-TMS.

Overall, the performance of the CSI:IOKR model for identification of TMS derivatives using GC-EI-MS spectra is somewhat lower as compared to its performance on a benchmark dataset, represented by 4,138 LC–ESI–MS/MS spectra from the Global Natural Products Social (GNPS) library [45]. This might be due to the smaller size of our test dataset or the type of input data (LC–ESI–MS/MS vs. GC-EI-MS). Interestingly, CSI:IOKR in our study resulted in identical median ARP as MetExpert for TMS derivatives, with slightly lower top 1 (11% vs. 13%) and remarkably better top 15 accuracy (63% vs. 52%).

Conclusions and further perspectives

The rate, volume and variety of compounds being introduced to the environment continues to expand exponentially. Consequently, many research groups and regulatory agencies are developing computational and high-throughput approaches for CEC annotation. As ML-based approaches are the future of CEC annotation, exploiting the perspectives for their further use is of utmost importance. Here we show that ML approaches, which have been predominantly used to annotate CEC from LC–MS data, can also be used to address the task of annotating TMS derivatives of CECs from GC–MS data. More specifically, this study shows that CSI:IOKR can be successfully employed for the annotation of TMS

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

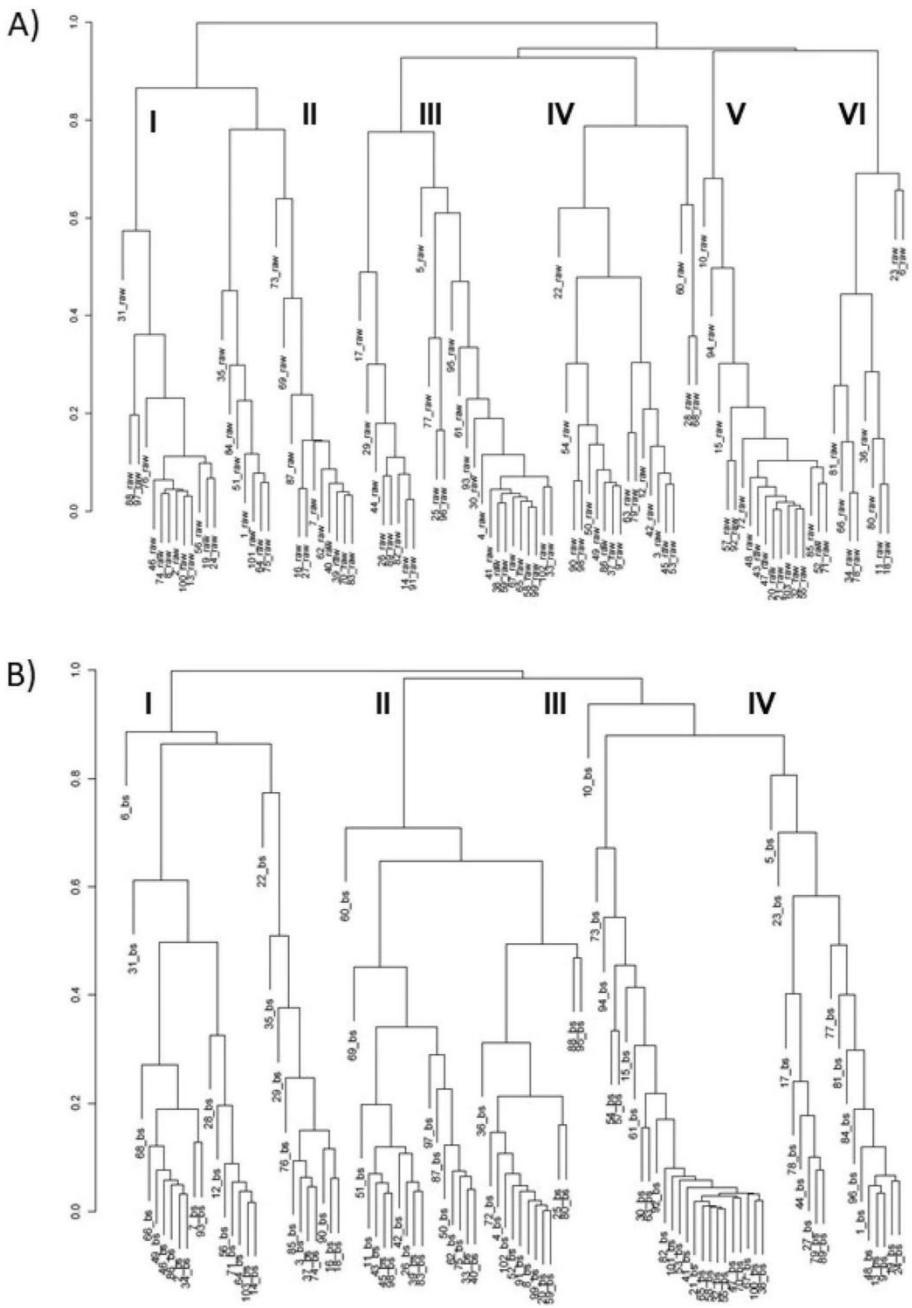


Fig. 10 **A** Clustering of consensus RAW and **B** BS experimental dataset based on similarirty of MS behavior and properties

derivatives of CEC from GC-EI-MS data. This presents a viable alternative to MSL search independent of an instrumental platform and data processing software.

Importantly, this study shows that expert curation of spectral datasets crucially improves the identification performance of ML-based approaches. Furthermore,

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | LE | TYPESET
MS Code : | CP | DISK

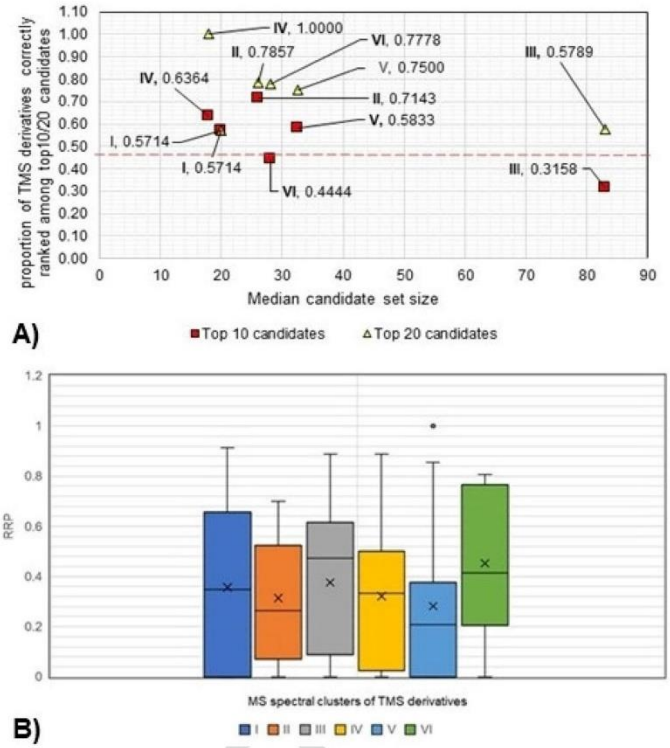


Fig. 11 **A** The median number of candidates in the candidate sets of different clusters of TMS derivatives (RAW dataset) plotted against the proportion of challenge TMS derivatives within the cluster correctly ranked within the top 10 candidates (red squares) and top 20 candidates (yellow triangles). **B** Box plot representing the different RRP values for the challenge compounds in each cluster of TMS derivatives from RAW dataset (bottom): dark blue plot—I, orange plot—II, grey plot—III, yellow plot—IV, light blue plot—V and green plot—VI

CSI:IOKR is useful in the identification of CEC that have been previously characterized (i.e., known unknowns that are currently in compound DBs) but whose GC-EI-MS spectra are not included in MSLs, thus increasing our knowledge on the composition of environmental samples. While spectral comparisons with reference standards or de novo structural elucidations might be required to validate the predictions, CSI:IOKR provides an efficient approach to prioritize candidates and reduces the time spent for compound annotation.

As further work, we propose a few straightforward extensions of this research that could be potentially successful and useful in enhancing the employment of the CSI:IOKR method in GC–MS-based CECs annotation. Instead of the PubChem repository, middle-sized compound DBs of particular value to the environmental science and toxicology communities, such as the US EPA's CCD [11], can be used. These compound DBs were proven to have higher potential in compound structure identification and exposure risk assessment over large repositories, such as ChemSpider [16] and PubChem [15, 17]. Moreover, the potential of CSI:IOKR could be further exploited on GC-EI-MS spectral data of TBDMS derivatives.

However, the ultimate challenge for IOKR would be the identification of the underivatized (parent) compounds using the GC-EI-MS spectra of their silyl derivative. The employment of IOKR and other IOKR-based methods would be significantly encouraged by their implementation within existing and upcoming CA frameworks. Besides CSI:IOKR, it would be very beneficial if other IOKR approaches [46, 47] and other cutting-edge ML-based methods [48, 49] are also challenged against identifying CECs using GC-EI-MS spectra. In that spirit, we would like to encourage the use of the generated GC-EI-MS datasets as benchmark datasets for further evaluation and improvement of ML-based approaches in GC–MS-based compound annotation.

Abbreviations

11HT: 11-Hydroxytestosterone; 2-APA: 2-Anilinophenylacetic acid; 3M5NC: 3-Methyl-5-nitrocatechol; 4NS: 4-Nitrosyringol; 6HP: 6-Hydroxypregnenolone;

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

For all TMS classes, CSI:IOKR performs satisfactorily both in terms of the proportion of TMS derivatives correctly ranked among the top10/20 candidates and in terms of the $\overline{RRP}$ of the challenge TMS derivatives. Except for aliphatic O-TMS derivarives and N-TMS derivatives, $\geq$50% of the correct TMS derivatives are ranked among the top 10 candidates. Especially good ranking scores are achieved for the poly TMS derivatives, i.e., bisphenol O-TMS derivatives and related TMS derivatives, and the other poly TMS derivatives, including mixed O,N-TMS and N-TMS derivatives, that have highest M_w and lowest median candidate size,which may partially contribute to their relatively good ranking. Namely, the correct CEC-TMS was ranked on average positions 10.68 and 19.50, respectively, while the average PubChem candidate set size was 22.04 and 28.60, respectively, which is 2–5 times lower than the values for the other TMS classes. Also evident from Fig. 8 is that CSI:IOKR performs solidly for phenyl O-TMS and non-aromatic O-TMS derivatives, which yield relatively high average candidate set sizes (108.43 and 120.67, respectively, data not shown). Despite that, their ranking scores are satisfactory, as well as their average RRP. The class of non-aromatic O-TMS derivatives contains 5 CEC-TMS derivatives, and thus the number of CEC-TMS derivatives is not representative, so that solid conclusions can be extracted. On the other hand, the phenyl O-TMS class is represented by 21 CEC-TMS, with low average ranking position (19.14), but hihg average PubChem candidate set size (108.43). Here, a factor that may positively contribute to the good ranking of some structural classes is the specificity of the fragmentation patterns, leading to uniqueness of its GC-EI-MS spectrum, which is responsible for the good ranking, independent of the sizr of the PubChem candaidate set. Finally, $\overline{RRP}$ is >0.50 or close to 0.50 (the threshold of satisfactory accuracy) for all TMS classes, except for phenyl-O-TMS derivatives (data not shown).

Clustering of MS spectra for the RAW (Fig. 10A) and the BS dataset (Fig. 10B) revealed 6 and 4 clusters, respectively. The RRP and proportion of TMS derivatives ranked among top 10/20 candidates differed significantly between the clusters of TMS derivatives with significant MS spectral similarity. The median candidate sizes for all clusters (except for cluster 3) were <35 candidates. For all of them (except for cluster 6, where the top 10 ratio is 0.4444), top 10 and top 20 ratios of >0.55 were achieved (Fig. 11A). $\overline{RRP}$ values vary significantly within all clusters, with average $\overline{RRP}$ <0.60 and clusters 2 and 5 having the lowest average $\overline{RRP}$ s (0.264 and 0.208) (Fig. 11B).

Legend: 1: BPAF-TMS; 2: DHBP-TMS; 3: 2APA-TMS; 4: 3M5NC-TMS; 5: CLP-TMS; 6: 4MC-TMS; 7: 4,4-BP-TMS; 8: HPP-TMS; 9: HB-P-TMS; 10: 4NC-TMS; 11: 4NG-TMS; 12: 4NS-TMS; 13: 4NP-TMS; 14: 4OP-TMS; 15: 5A3B17B-TMS; 16: 5NG-TMS; 17: 6HP-MS; 18: 6MAM-TMS; 19: 6NG-TMS; 20: 8HQ-TMS; 21: 9HF-TMS; 22: 11AHA-TMS; 23: 11AHT-TMS; 24: 11OHTHC-TMS; 25: 11N9THC-TMS; 26: E2-TMS; 27: EE2-TMS; 28: 17HPT-MS; 29: AA-TMS; 30: AMPH-TMS; 31: PAA-TMS; 32: BA-TMS; 33: BZECG-TMS; 34: BzPb-TMS; 35: 22BPF-TMS; 36: 24BPF-TMS; 37: BPA-TMS; 38: BPAP-TMS; 39: BPB-TMS; 40: BPBP-TMS; 41: BPC-TMS; 42: BPCL-TMS; 43: BPE-TMS; 44: BPF-TMS; 45: BPFL-TMS; 46: BPM-TMS; 47: BPP-TMS; 48: BPPH-TMS; 49: BPS-TMS; 50: BPZ-TMS; 51: BD-TMS; 52: BP26DM-TMS; 53: BuPb-TMS; 54: BHT-TMS; 55: CBC-TMS; 56: CBD-TMS; 57: CBDA-TMS; 58: CBN-TMS; 59: CBZ-TMS; 60: CAT-TMS; 61: CA-TMS; 62: CLA-TMS; 63: COD-TMS; 64: THC-TMS; 65: THCA-TMS; 66: DF-TMS; 67: BP8-TMS; 68: ERY-TMS; 69: E3-TMS; 70: E1-TMS; 71: EtPb-TMS; 72: ET-TMS; 73: IB-TMS; 74: IbUPb-TMS; 75: IPrPb-TMS; 76: LLEU-TMS; 77: LAA-TMS; 78: LLEU-TMS; 79: LSER-TMS; 80: LTYR-TMS; 81: MCA-TMS; 82: MAMPH-TMS; 83: MePb-TMS; 84: MORPH-TMS; 85: NAP-TMS; 86: NTX-TMS; 87: OCA-TMS; 88:PCA-TMS; 89: PrPb-TMS; 90: QA-TMS; 91: RES-TMS; 92: SA-2TMS; 93: SA-TMS; 94: SHA-TMS; 95: STA-2TMS; 96: STA-TMS; 97: SFA-2TMS; 98: SFA-TMS; 99: SYR-TMS; 100: T3HC-TMS; 101: TCS-TMS; 102: DHDPE-TMS; 103: UA-TMS.

Overall, the performance of the CSI:IOKR model for identification of TMS derivatives using GC-EI-MS spectra is somewhat lower as compared to its performance on a benchmark dataset, represented by 4,138 LC–ESI–MS/MS spectra from the Global Natural Products Social (GNPS) library [45]. This might be due to the smaller size of our test dataset or the type of input data (LC–ESI–MS/MS vs. GC-EI-MS). Interestingly, CSI:IOKR in our study resulted in identical median ARP as MetExpert for TMS derivatives, with slightly lower top 1 (11% vs. 13%) and remarkably better top 15 accuracy (63% vs. 52%).

Conclusions and further perspectives

The rate, volume and variety of compounds being introduced to the environment continues to expand exponentially. Consequently, many research groups and regulatory agencies are developing computational and high-throughput approaches for CEC annotation. As ML-based approaches are the future of CEC annotation, exploiting the perspectives for their further use is of utmost importance. Here we show that ML approaches, which have been predominantly used to annotate CEC from LC–MS data, can also be used to address the task of annotating TMS derivatives of CECs from GC–MS data. More specifically, this study shows that CSI:IOKR can be successfully employed for the annotation of TMS

Journal : BMCTwo 13321 | Dispatch : 9-8-2022 | Pages : 20
Article No : 636 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

15. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47(D1):D1102–D1109
16. Pence HE, Williams A (2010) ChemSpider: an online chemical information resource. J Chem Educ 87(11):1123–1124
17. McEachran AD, Sobus JR, Williams AJ (2017) Identifying known unknowns using the US EPA's CompTox chemistry dashboard. Anal Bioanal Chem 409(7):1729–35. https://doi.org/10.1007/s00216-016-0139-z
18. Stein S (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. Anal Chem 84(17):7274–7282
19. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R et al (2018) HMDB 40: the human metabolome database for 2018. Nucleic Acids Res 46(D1):D608-17
20. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G et al (2018) METLIN: a technology platform for identifying knowns and unknowns. Anal Chem 90(5):3156–64. https://doi.org/10.1021/acs.analchem.7b04424
21. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K et al (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45(7):703–14. https://doi.org/10.1002/jms.1777
22. mzCloud—Advanced mass spectral database. 2021. https://www.mzcloud.org/Accessed 10 Jun 2021.
23. Hummel J, Selbig J, Walther D, Kopka J (2007) The golm metabolome database: a database for GC-MS based metabolite profiling. In: metabolomics a powerful tool in systems biology. Springer, Berlin
24. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S et al (2009) FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. Anal Chem 81(24):10038–48. https://doi.org/10.1021/ac901952
25. National Institute of Standards and Technology. NIST/EPA/NIH Mass Spectral Library. Wiley.com. 2020. https://www.wiley.com/en-ai/NIST+EPA+NIH+Mass+Spectral+Library+2020-p-9781119750291
26. Wiley Registry of Mass Spectral Data, 12th Edition. Wiley science solutions. 2021. https://sciencesolutions.wiley.com/solutions/technique/gc-ms/wiley-registry-of-mass-spectral-data-12th-edition/. Accessed 6 Aug 2021.
27. Oberacher H, Sasse M, Antignac JP, Guitton Y, Debrauwer L, Jamin EL et al (2020) A European proposal for quality control and quality assurance of tandem mass spectral libraries. Environ Sci Eur 32(1):43
28. Ljoncheva M, Stepišnik T, Džeroski S, Kosjek T (2020) Cheminformatics in MS-based environmental exposomics: current achievements and future directions. Trends Environ Anal Chem 28:e00099
29. Blaženović I, Kind T, Ji J, Fiehn O (2018) Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. Metabolites 8(2):31
30. Nguyen DH, Nguyen CH, Mamitsuka H (2018) Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. Brief Bioinform 20(6):2028–43
31. Andra SS, Austin C, Patel D, Dolios G, Awawda M, Arora M (2017) Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome. Environ Int 100:32–61
32. Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O (2016) Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. Trennds Anal Chem 78:23–35
33. Mass Frontier™. Spectral interpretation software. 2021. https://www.thermofisher.com/order/catalog/product/OPTON-30920. Accessed 11 Jun 2021.
34. ACD/MS Fragmenter. Advanced Chemistry Labs, Toronto, Canada. 2020. https://www.acdlabs.com/products/adh/ms/ms_frag/. Accessed 20 Jul 2020.
35. Schymanski EL, Meinert C, Meringer M, Brack W (2008) The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis. Anal Chim Acta 615(2):136–47
36. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T et al (2016) Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Anal Chem 88(16):7946–7958
37. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA et al (2008) FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass Spectrom 22(19):3043–3052
38. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminformatics 8(1):1–16
39. Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MSn spectral trees. Rapid Commun Mass Spectrom 26(20):2461–71. https://doi.org/10.1002/rcm.6364
40. Menikarachchi LC, Cawley S, Hill DW, Hall LM, Hall L, Lai S et al (2012) MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures. Anal Chem 84(21):9388–9394
41. Wang Y, Kora G, Bowen BP, Pan C (2014) MIDAS: a database-searching algorithm for metabolite identification in metabolomics. Anal Chem 86(19):9496–9503
42. Qiu F, Lei Z, Sumner LW (2018) MetExpert: an expert system to enhance gas chromatography-mass spectrometry-based metabolite identifications. Anal Chim Acta 11(1037):316–326
43. Heinonen M, Shen H, Zamboni N, Rousu J (2012) Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics 28(18):2333–2341
44. Shen H, Dührkop K, Böcker S, Rousu J (2014) Metabolite identification through multiple kernel learning on fragmentation trees. Bioinformatics 30(12):i157–i164
45. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J (2016) Fast metabolite identification with Input output kernel regression. Bioinformatics 32(12):i28-36
46. Brouard C, Bach E, Bocker S, Rousu J (2017) Magnitude-preserving ranking for structured outputs. Proc Mach Learn Res 77:407–22
47. Brouard C, Bassé A, d'Alché-Buc F, Rousu J (2019) Improved small molecule identification through learning combinations of kernel regression models. Metabolites 9(8):160
48. Nguyen DH, Nguyen CH, Mamitsuka H (2018) SIMPLE: sparse interaction model over peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra. Bioinformatics 34(13):i323–i332
49. Nguyen DH, Nguyen CH, Mamitsuka H (2019) ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. Bioinformatics 35(14):i164-72
50. Allen F, Pon A, Greiner R, Wishart D (2016) Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. Anal Chem 88(15):7689–7697
51. Wei JN, Belanger D, Adams RP, Sculley D (2019) Rapid prediction of electron-ionization mass spectrometry using neural networks. ACS Cent Sci 5(4):700–8. https://doi.org/10.1021/acscentsci.9b00085
52. Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, Arndt D et al (2019) CFM-ID 30: significantly improved ESI-MS/MS prediction and compound identification. Metabolites 9(4):72
53. Kangas LJ, Metz TO, Isaac G, Schrom BT, Ginovska-Pangovska B, Wang L et al (2012) In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. Bioinformatics 28(13):1705–1713
54. Jebara T, Kondor R, Howard A (2004) Probability product kernels. J Mach Learn Res 5:819–844
55. Gonen M, Alpaydın E, Tr BE, Tr BE (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
56. Koo I, Kim S, Shi B, Lokriewicz P, Song M, McClain C et al (2016) Elder: a compound identification tool for gas chromatography mass spectrometry data. J Chromatogr A 1448:107–114
57. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K et al (2017) Critical assessment of small molecule identification 2016: automated methods. J Cheminformatics 9(1):22
58. Critical assessment of small molecule identification. 2021. http://www.casmi-contest.org/2017/index.shtml. Accessed 19 Jun 2021.
59. Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 11(1):98–110
60. Hug C, Ulrich N, Schulze T, Brack W, Krauss M (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environ Pollut 184:25–32

Journal : BMCTwo 13321	Dispatch : 9-8-2022	Pages : 20
Article No : 636	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

61. Ruff M, Mueller MS, Loos M, Singer HP (2015) Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry—identification of unknown sources and compounds. Water Res 87:145–54
62. Kiefer K, Müller A, Singer H, Hollender J (2019) New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS. Water Res 165:114972
63. Albergamo V, Schollée JE, Schymanski EL, Helmus R, Timmer H, Hollender J et al (2019) Nontarget screening reveals time trends of polar micropollutants in a riverbank filtration system. Environ Sci Technol 53(13):7584–94
64. Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA et al (2014) Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ Sci Technol 48(3):1811–8. https://doi.org/10.1021/es4044374
65. Moschet C, Piazzoli A, Singer H, Hollender J (2013) Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. Anal Chem 85(21):10312–20. https://doi.org/10.1021/ac4021598
66. National Institute of Standards and Technology. NIST/EPA/NIH Mass Spectral Library 2017. Wiley.com. 2017. https://www.wiley.com/en-ai/NIST+EPA+NIH+Mass+Spectral+Library+2017-p-9781119750291.
67. US EPA O. Toxicity estimation software tool (TEST). 2015. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test. Accessed 11 Aug 2021.
68. Epa US (2021) Estimation programs interface suite™ for Microsoft® windows. United States Environmental Protection Agency, Washington
69. European Commsion (2021) Regulation (EC) No.1907/2006 of the European Parliament and of the Council on the registration, evaluation, authorisation and restriction of chemicals (REACH). Off J Eur Communities 396:1–552
70. Dührkop K, Hufsky F, Böcker S (2014) Molecular formula identification using isotope pattern analysis and calculation of fragmentation trees. Mass Spectrom 3:S0037–S0037
71. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N et al (2017) The Chemistry development kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics 9(1):33. https://doi.org/10.1186/s13321-017-0220-4
72. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci 112(41):12580–12585
73. Meringer M, Reinker S, Zhang J, Muller A (2011) MS/MS data improves automated determination of molecular formulas by mass spectrometry. Commun Math Comput Chem 65:259–90
74. Kind T, Fiehn O (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8(1):105
75. Stein SE, Scott DR (1994) Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom 5(9):859–66. https://doi.org/10.1016/1044-0305%2894%2987009-8
76. Kerber A, Meringer M, Rücker C (2006) CASE via MS: ranking structure candidates by mass spectra. Croat Chem Acta 79(3):449–64

Publisher's Note

Journal : BMCTwo 13321	Dispatch : 9-8-2022	Pages : 20
Article No : 636	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

Journal:	**13321**
Article:	**636**

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details Required	Author's Response
AQ1	References: As per pubmed findings, citation details [DOI] for Reference [5] have been inserted. Kindly check and confirm the inserted details.	
AQ2	References: As per pubmed findings, citation details [Page] for Reference [52] have been inserted. Kindly check and confirm the inserted details.	
AQ3	References: As per pubmed findings, citation details [Page, vo;ume and issue no] for Reference [76] have been inserted. Kindly check and confirm the inserted details.	
AQ4	References: Citation details for Reference [25, 66] are incomplete. Please supply the [Accessed date] of this reference. Otherwise, kindly advise us on how to proceed.	
AQ5	References: Citation details for Reference [12, 22, 26, 33, 34] are incomplete. Please supply the missing details of this reference. Otherwise, kindly advise us on how to proceed.	
AQ6	Author contributions: Journal standard instruction requires the statement "All authors read and approved the final manuscript." in the Author contributions section. This was inserted at the end of the paragraph of the said section. Please check if appropriate.	

4.3 Comparison of Machine learning-based with non-machine learning-based CA

This section compares the CSI:IOKR approach's performance for identifying CEC-TMS derivatives through their GC-EI-MS spectra to a non-ML-based identification approach. The latter matches the acquired GC-EI-MS spectra of CEC-TMS from the test datasets RAW and BS described in the two papers in Section 4.2 to the NIST 17 MSL [1]. The resulting GC-EI-MS spectra were first extracted as mzXML files from the Agilent Mass Hunter Qualitative Analysis software v.B.07.00 and matched against NIST 17 MSL [1], using the MS Search Program v.2.3.

The results of the NIST 17 MSL [1] matching of the RAW and BS GC-EI-MS spectra of 106 CEC-TMS derivatives are given in Table 4.1.1. They are represented by ranking the correct CEC-TMS derivatives in the top 100 library hits, the match factor (MFR), and the reverse match factor (RMFR). The MFR for the unknown and the library spectrum is calculated by directly matching peak *m/z* values and relative intensities. The RMFR is a match factor for the query spectrum and the library spectrum, ignoring any peaks in the unknown not in the library spectrum. Both scores are derived from a modified cosine of the angle between the spectra (normalized dot product). A perfect MFR result would have a value of 999, whereas two spectra with no m/z in common would have a value of zero. As a general guide, 900 or higher MFR and RMFR values mean an excellent match; 800-900 is a good match, 700-800 is a fair match, and $\leq$ 600 is a poor match [105].

For almost a third (28 %) of the studied CEC-TMS, there was no corresponding GC-EI-MS spectrum in the NIST 17 MSL [1]. Matching the GC-EI-MS spectra for the remaining 72 % CEC-TMS from the TMS RAW test dataset against NIST 17 MSL [1] (Table 4.1: NIST 17 MSL search results for TMS RAW and BS GC-EI-MS test dataset.) resulted in ranking the correct CEC-TMS at #1 for 98 CEC-TMS derivatives; thus 89.61% of the CEC-TMS derivatives were ranked first. 99.05 % of the CEC-TMS were ranked among the top 20 hits, and 97.17% were among the top 10 hits. The average absolute ranking position (ARP) and the average relative ranking position (RRP) of the correct CEC-TMS, calculated according to the equations given in the second paper in Section 4.2, are 2.36 and 0.014, respectively. The average MFR and RMFR are 882.325 (308-990) and 904.857 (322-922), respectively.

For many of the CEC-TMS derivatives whose GC-EI-MS spectra are not present in NIST 17 MSL [1], TMS derivatives of structurally similar compounds were ranked in the top hits. For example, for both 5-NG-TMS and 6-NG-TMS, the first hit is 2-methoxy-5-nitrophenol TMS (MFR 986/RMFR 993 for 5-NG-TMS and 716/759 for 6-NG-TMS). For 6-HP TMS, the #1 hit is pregn-5-en-20-one, 3.17-bis[(trimethylsilyl)oxy]-, O-methyl, which has the identical sterol ring backbone. Further, 11-HT-TMS is #5 and #1 hit for 11-HAD-TMS and 17-HP-TMS, respectively. Also, BPA-2TMS, the correct compound, is ranked #2, while at #1 is 3,4'-isopropylidenediphenol, bis (trimethylsilyl) ether, which has the two -OH groups at m- and p- positions at the two phenyl rings, instead of p-/p- in BPA. ERY-4TMS is ranked #2 (835/864), while #1 is 1,2,3-butanetriol 3TMS (842/886). The same is true for E3-3TMS, ranked #2 (945/948), while #1 is the stereoisomer 16β, 17β-estriol 3TMS (968/971). For IbuPb-TMS, not included in NIST 17 MSL [1], the GC-EI-MS spectrum of BuPb-TMS is ranked #1 (956/956). Another example is BP-8-TMS, for which {4-methoxy-2-[trimethylsilyl)oxy]phenyl}{2-[(trimethylsilyloxy]phenyl)methanone is ranked #1 (969/969). The top hit has the two TMS groups attached to the ortho position of the two phenyl rings. Finally, for DH-BP 2TMS, which is not in NIST 17 MSL [1], the GC-EI-MS spectrum of 2,4'-dihydroxybenzophenone 2TMS is ranked #1 (851/914), and the one of 2,2'-dihydroxybenzophenone 2TMS is ranked #2 (796/865).

For other CEC-TMS derivatives whose GC-EI-MS spectra are not included in NIST 17 MSL [1], most top 10 hits have an identical MF to the queried derivatives. Examples are MePb-TMS, for which the top 9 hits share the MF ($C_{11}H_{16}O_3Si$), and EtPb-TMS, whose top 4 hits had the MF ($C_{12}H_{18}O_3Si_3$), including acetovanillone-TMS, acetoisovanillone-TMS, and ethyl vanillate-TMS. In contrast, none of the top 10 hits shares the correct MF for KET-TMS ($C_{19}H_{22}O_3Si$). Interestingly, for one CEC-TMS derivative, 9-HF-TMS, underivatized compounds structurally similar to the parent CEC appeared among the top 10 hits. Such are 9-benzylfluorene ranked #4 (554/650), and 9H-fluorene-9-carboxylic acid, ethyl ester, ranked 6 (524/871).

For some of the CEC-TMS whose GC-EI-MS spectra are not included in NIST 17 MSL [1], the top 10 hits are chemically illogical Si-containing compounds with different MFs and low MFR/RMFR. Such is MEC-TMS, where the match scores of the top 10 hits ranged from 516 to 563, and the #1 hit is a compound with a C-Si bound. Also, for CBDA-3TMS, many inappropriate matches are found between the top 10 hits, for example, siloxanes or even arsenic-containing compounds. Likewise, for BPBP-2TMS that is not part of NIST 17 MSL [1], the top 20 matches do not include TMS derivatives of any other bisphenols present. Also, for CBC-TMS, none of the cannabinoids included in NIST 17 MSL [1] is listed in the top 100 hit list, except Δ^9-THC-TMS, ranked #53 (438/447).

For 4,4'-BP-2TMS, the resulting hit list is also unsatisfactory. Although its GC-EI-MS spectra are in the library, the correct compound is ranked #78 (308/322). Interestingly, for the compound with three stereoisomers, coumaric acid (o-, m- and p-), only one of all three stereoisomers is ranked #1 (m- 901/960), while the o- isomer is ranked #12 (638/651), and the p-isomer is ranked #13 (640/660). For OCA-2TMS, the p- isomer is ranked #1 (936/968) and *vice versa,* i.e., the o-isomer is ranked #1 for PCA-2TMS (981/985), with identical values for both RAW and BS spectra.

Matching the background-subtracted GC-EI-MS spectra of CEC-TMS against NIST 17 MSL [1] did not result in any crucial changes in the CSI performance (Table 4.1). The ARP of the correct CEC-TMS remained the same – 2.36, with average MFR and RMFR of 882 (308-990) and 903 (332-992), respectively.

Still ranked #1, but with lower MFR/RMFR are the GC-EI-MS spectra of TMS derivatives of STA, NX, L-TYR, EE2, and CBN, while BPCL-2TMS has only lower MFR for BS. The MFR/RMFR are higher for the BS GC-EI-MS spectra of the TMS derivatives of SA, PAA, MAMPH, COD, BHT, BZECG, 4,4'-BP, and AMP (778/852 vs. 879/948) and 6-MAM (681/682 vs. 780/781). Only the RAW GC-EI-MS spectrum of 3-MT-2TMS is ranked #1, and the BS spectrum is ranked #2, as the positional isomer 4-MT-2TMS is ranked #1 (986/991).

Comparison of the performance of the CSI:IOKR approach and the NIST 17 MSL [1] manual search approach in terms of identification accuracy reveals higher top 1, top 10, and top 20 rankings and average ARP for the NIST 17 MSL manual search, compared to CSI:IOKR. The average RRP for the NIST 17 MSL manual search was 25.7 times lower than for CSI:IOKR. Lower RRP values indicate better rankings [106], meaning that compounds are more correctly ranked in the case of the NIST 17 MSL manual search approach. However, the results of the NIST 17 MSL manual search include only the TMS derivatives whose GC-EI-MS spectra are present in the MSL, while the CSI:IOKR approach performs CA for all TMS derivatives, not considering the presence/absence of their GC-EI-MS spectra in the NIST 17 MSL. Despite this issue, the NIST 17 MSL manual search approach has more disadvantages. The MS Search Program v.2.3 allows import of spectra in limited file formats (.mzXML, .mzData, .MSP, .JDX, .SDF, .MOL, .KCF, .DTA, .PKL), while the Agilent Mass Hunter Qualitative Analysis v.B.07.00 allows export of MS spectra in other file formats (.xls, .txt, .csv, .xml, .mzData, .emf and .bmp). Therefore, GC-EI-MS spectra must first be extracted from Mass Hunter Qualitative Analysis v.B.07.00 as .mzData files and then imported into NIST MS Search Program v.2.3. As the NIST MS Search Program does not allow batch import and search of .mzData MS files, a search was performed manually, one by one for each GC-EI-MS spectra.

Table 4.1: NIST 17 MSL search results for TMS RAW and BS GC-EI-MS test dataset.

CEC-TMS	CEC-TMS present in NIST 17 MSL	ARP		RRP		MFR		RMFR	
		RAW	BS	RAW	BS	RAW	BS	RAW	BS
2AA-TMS	no	-		-		-		-	
3M5NC- TMS	no	-		-		-		-	
3-MC-TMS	yes	1	1	0	0	923	921	926	924
4,4'-BP-2TMS	yes	1	1	0	0	826	841	971	972
HPP-TMS	yes	1	1	0	0	963	967	965	968
H-BP-TMS	yes	1	1	0	0	921	921	969	969
4-NC-TMS	no	-		-		-		-	
4-NG-TMS	no	-		-		-		-	
4-NS-TMS	no	-		-		-		-	
4-NP-TMS	no	-		-		-		-	
4-OP-TMS	yes	1	1	0	0	894	894	991	991
5-AD-TMS	yes	1	1	0	0	823	823	826	826
5-AD-2TMS	yes	1	1	0	0	970	970	975	975
5-NG-TMS	no	-		-		-		-	
6HP-TMS	no	-		-		-		-	

6-MAM-TMS	yes	2	2	0.01	0.01	681	780	682	781
6-NG-TMS	no	-	-	-		-		-	
8-HQ-TMS	yes	1	1	0	0	929	930	958	959
9-HF-TMS	no	-		-		-		-	
11-HA-TMS	no	-		-		-		-	
11HT-TMS	no	-		-		-		-	
11-OH-THC-TMS	yes	1	1	0	0	891	891	892	892
11N9CTHC-TMS	yes	1	1	0	0	751	751	786	786
EE2-TMS	yes	1	1	0	0	762	588	770	572
17HP-TMS	no	-		-		-		-	
E2-2TMS	yes	1	1	0	0	975	975	981	981
22BPF-2TMS	yes	1	1	0	0	971	971	972	972
24BPF-2TMS	yes	1	1	0	0	963	963	971	971
AA-2TMS	yes	1	1	0	0	985	985	988	988
AMPH-TMS	yes	1	1	0	0	778	879	852	948
BA-TMS	yes	1	1	0	0	967	967	983	983
BP-8 2TMS	no	-		-		-		-	
BZECG-TMS	yes	1	1	0	0	751	776	904	921
BzPb-TMS	yes	1	1	0	0	888	888	913	913
BD-TMS	yes	1	1	0	0	955	955	963	963
TMBA-2TMS	yes	1	1	0	0	954	954	956	956
BPA-2TMS	yes	2	2	0.01	0.01	968	968	969	969

BPAF-2TMS	yes	1	1	0	0	926	926	935	935
BPAP-2TMS	no	-		-		-		-	
BPB-2TMS	yes	1	1	0	0	933	933	934	934
BPBP-2TMS	no	-	-	-		-		-	
BPCL-2TMS	yes	1	1	0	0	935	925	939	939
BPC-2TMS	yes	1	1	0	0	951	951	951	951
BPE-2TMS	yes	1	1	0	0	949	949	952	952
BPF-2TMS	yes	1	1	0	0	941	941	943	943
BPFL-2TMS	no	-		-		-		-	
BPM-2TMS	no	-		-		-		-	
BPP-2TMS	no	-		-		-		-	
BPPH-2TMS	no	-		-		-		-	
BPS-2TMS	no	-		-		-		-	
BPZ-2TMS	yes	1	1	0	0	924	924	926	926
BHT-TMS	yes	1	1	0	0	742	818	776	882
BuPb-TMS	yes	1	1	0	0	965	965	973	973
CBC-TMS	no	-		-		-		-	
CBD-TMS	yes	1	1	0	0	812	791	837	842
CBDA-3TMS	no	-		-		-		-	
CBN-TMS	yes	1	1	0	0	959	959	963	963
CBZ-TMS	yes	1	1	0	0	869	869	868	898
CAT-TMS	yes	1	1	0	0	977	977	977	977
CLP-TMS	yes	1	1	0	0	949	949	952	952
CA-4TMS	yes	1	1	0	0	965	965	967	967
CLA-TMS	yes	1	1	0	0	969	969	984	984
COD-TMS	yes	1	1	0	0	799	855	801	857

DH-BP-TMS	no	-		-		-		-	
DHDPE-2TMS	yes	78	78	0.77	0.77	308	307	322	321
DF-TMS	yes	1	1	0	0	940	890	961	919
ERY-TMS	yes	2	2	0.01	0.01	835	835	864	863
E3-3TMS	yes	2	2	0.01	0.01	945	945	948	948
E1-TMS	yes	1	1	0	0	939	939	946	946
EtPb-TMS	no	-		-		-		-	
ET-TMS	yes	1	1	0	0	955	965	968	968
IB-TMS	yes	1	1	0	0	982	982	992	992
IBuPb-TMS	no	-		-		-		-	
IprPb-TMS	yes	1	1	0	0	968	969	973	974
KET-TMS	no	-		-		-		-	
L-AA-4TMS	yes	1	1	0	0	961	961	961	961
L-LEU-TMS	yes	1	1	0	0	931	931	961	961
L-SER-3TMS	yes	1	1	0	0	969	969	981	981
L-TYR-3TMS	yes	1	1	0	0	629	616	686	674
MCA-2TMS	yes	1	1	0	0	901	901	960	960
OCA-2TMS	yes	12	12	0.11	0.11	638	638	651	651
PCA-2TMS	yes	13	13	0.12	0.12	640	640	660	660
MEC-TMS	no	-		-		-		-	
MAMPH-TMS	yes	1	1	0	0	893	910	933	944
MePb-TMS	yes	1	1	0	0	990	990	992	992
MORPH-2TMS	yes	1	1	0	0	772	772	774	774
NAP-TMS	yes	1	1	0	0	960	960	992	992
NX-TMS	yes	1	1	0	0	900	829	947	932
NL-2TMS	yes	1	1	0	0	987	987	990	990
PAA-TMS	yes	1	1	0	0	955	983	973	985

Prpb-TMS	yes	1	1	0	0	971	971	971	971
QA-5TMS	yes	1	1	0	0	861	861	934	934
RES-TMS	yes	1	1	0	0	980	980	985	985
SA-TMS	yes	1	1	0	0	864	863	923	933
SA-2TMS	yes	1	1	0	0	956	956	992	992
SHA-4TMS	yes	1	1	0	0	673	673	707	707
STA-TMS	yes	1	1	0	0	987	974	988	977
STA-2TMS	yes	1	1	0	0	839	839	875	875
SFA-TMS	yes	1	1	0	0	764	764	934	934
SFA-2TMS	yes	1	1	0	0	625	625	672	672
SYR-TMS	yes	2	2	0.01	0.01	982	981	988	988
Δ^9-THC-TMS	yes	1	1	0	0	931	931	936	936
Δ^9-THCA-TMS	yes	1	1	0	0	669	669	670	670
T3HC-TMS	yes	1	1	0	0	838	838	862	862
TCS-TMS	no	-		-		-		-	
UA-2TMS	yes	1	1	0	0	947	947	961	961

The time required for manual GC-EI-MS spectra extraction, NIST 17 MSL search, and exporting and analyzing the results was approximately 18 hours for each TMS RAW and BS spectral dataset. Compared to CSI:IOKR, for the same dataset, it requires approximately 50 times more of the analysts' time, which would be the most time-consuming step during high-throughput analyses.

4.4 Identification of CEC-TMS Derivatives in Complex Environmental Matrices

In order to evaluate the ability to identify CEC-TMS derivatives in complex environmental samples, an identification exercise was carried out of a representative sub-selection of 56 CEC-TMS from the TMS RAW and BS test datasets. In this frame, four different complex environmental samples were collected: grab samples of wastewater (WW) influent (WWI) and WW effluent (WWE) from a municipal wastewater treatment plant and grab samples from two rivers (A and B) in central Slovenia. Each 200 mL sample was spiked with 600 µL 1.2 µg/mL group working solution of the selected CEC. The samples were analyzed using the procedure described in the manuscript *Ljoncheva M., Heath E., Heath D., Džeorski S., Kosjek T., Contaminants of emerging concern: silylating procedures, evaluation of the stability of silyl derivatives and associated measurement uncertainty* (Section 1045.2). This includes solid phase extraction (SPE) on Oasis HLB Prime cartridges, derivatization of 0.5 mL sample with 30 µL N,O-bis trifluoroacetamide (BSTFA) + 1% trimethylchlorosilane (TMCS) at 70°C for 45 min, and GC-MSD analysis in full scan mode. The samples were prepared in triplicate (n=3), and blanks were prepared to monitor the influence of the matrix effect. The resulting GC-EI-MS spectra were extracted as .mzXML files from the Agilent Mass Hunter Qualitative Analysis v.B.07.00 software and matched against the NIST 17 MSL [1], using the MS Search Program v.2.3.

The NIST 17 MSL search metrics sum is given in Table 4.3, while the detailed results of the NIST 17 MSL matching are given in **Error! Reference source not found.**. CEC-TMS derivatives whose GC-EI-MS spectra are not included in NIST 17 MSL [1] and CEC-TMS derivatives, which were not among the top 100 hits, were excluded. This exclusion resulted in 30.36%, 28.57%, 28.57%, and 37.50% of CEC-TMS missing in river water (RW) A, RW B, WWE, and WWI, respectively. For the remaining, we calculated the top 1, top 10, average ARP, average RRP, average MFR, and average RMFR. The average values of ARP and RRP indicate the most accurate and confident compound structure identification (CSI), in descending order, in RW A, RW B, WWE, and WWI (**Error! Reference source not found.**). Average MFR and RMFR were highest for RW A, WWE, RW B, and WWI, all in the range of a fair match (700-800). Generally, the ARP and RRP values in the case of surface waters were better than in solvent (Table 4.1), while for wastewaters, they were lower. Finally, of the remaining CEC-TMS derivatives after the exclusion, 82.05%, 87.50%, 72.50%, and 57.14% were ranked first in RW A, RW B, WWE, and WWI, respectively. All CEC-TMS derivatives were ranked among the top 10 hits for RW A and RW B, while 95.00% and 88.57% were ranked among the top 10 for WWE and WWI.

The results were unsatisfactory for some of the CEC-TMS derivatives, whose spectra are included in NIST 17 MSL [1], as the spectra of the correct CEC-TMS were not in the top 100 hits. The same was true for NAP-TMS in RW A in WWI, 11OHTHC-TMS, and 11N9THC-TMS in WWI and DHDPE-2TMS in the four matrices. The GC-EI-MS spectrum of DHDPE-2TMS was ranked #78 when both RAW and BS test GC-EI-MS spectra were searched against NIST 17 MSL [1]; therefore, low identification accuracy was expected. For SFA-TMS, the GC-EI-MS spectrum was not among the top 100 hits in RW, while no peak was observed in WWE and WWI, while SFA-TMS had a low chromatographic response and significant peak tailing and was therefore excluded from further stability studies (Section 5.2).

According to the presented results, manual matching of acquired GC-EI-MS spectra of CEC-TMS derivatives to the NIST 17 MSL [1], despite being time-consuming, is an efficient approach for their identification in complex environmental samples. Careful examination of the top 10 hits must be performed since, in some cases, they contain the correct match

or structurally similar TMS derivatives, which would further ease the structural elucidation of the queried CEC-TMS derivatives.

Table 4.2: Metrics of NIST 17 MSL search in environmental matrices. Top 1, top 10, and top 20 are expressed as the percentage (%) of the remaining CEC-TMS derivatives (total – total missing).

Matrix type/parameter	not in NIST 17 MSL	not in top 100 hits	total missing (%)	top 1 (%)	top 10 (%)	top 20 (%)	ARP	RRP	MFR	RMFR
River A	14	3	17 (30.36%)	32 (82.05%)	39 (100%)	39 (100.00%)	1.316	<0.001	835.160	871.421
River B	14	2	16 (28.57%)	35 (72.50%)	40 (100%)	40 (100.00%)	1.821	0.008	798.615	825.231
WWE	14	1	16 (28.57%)	29 (72.50%)	38 (95.00%)	39 (97.50%)	3.921	0.029	801.605	836.051
WWI	14	4	21 (37.50%)	35 (57.14%)	31 (88.57%)	32 (91.43%)	5.794	0.042	709.59	763.176

Table 4.3: Results of manual NIST 17 MSL identification of CEC-TMS from complex environmental matrices, N/A - GC-EI-MS spectra not found.

CEC-TMS	River A				River B				WWE				WWI			
	ARP	RRP	MFR	RMFR	ARP	RRP	MFR	RMFR	ARP	RRP	MFR	RMFR	ARP	RRP	MFR	RMFR
BA-TMS	4	0.03	591	680	1	0	703	788	12	0.12	555	646	1	0	737	816
RES-2TMS	1	0	938	951	1	0	939	952	1	0	905	919	2	0.01	926	945
MePb-TMS	1	0	703	762	1	0	938	955	2	0.01	660	714	2	0.01	640	748
8-HQ-TMS	1	0	846	917	1	0	774	875	1	0	647	793	27	0.27	454	477
EtPb-TMS	not in NIST 17															
IPrPb-TMS	3	0.02	790	813	3	0.02	790	813	3	0.02	803	812	2	0.01	852	905
IB-TMS	2	0.01	652	905	1	0	798	874	10	0.09	674	798	46	0.45	455	489
MEC-TMS	not in NIST 17															
4-OP TMS	5	0.04	791	986	5	0.04	791	986	6	0.05	752	992	16	0.15	714	983
PrPb-TMS	1	0	718	762	1	0	887	912	1	0	917	943	1	0	887	941

iBuPb-TMS	not in NIST 17															
BuPb-TMS	1	0	768	818	1	0	928	952	1	0	958	970	1	0	789	857
9-HF-TMS	not in NIST 17															
HPP-TMS	1	0	970	977	1	0	966	971	1	0	982	986	1	0	900	921
CLP-TMS	1	0	956	965	1	0	946	960	1	0	966	975	1	0	723	779
4-NP-TMS	not in NIST 17															
22BPF-2TMS	1	0	953	954	1	0	937	937	1	0	930	931	2	0.01	840	850
BPAF-2TMS	1	0	949	961	1	0	945	956	1	0	965	975	1	0	712	743
H-BP-TMS	1	0	902	935	1	0	858	906	1	0	954	973	1	0	624	640
24BPF-2TMS	1	0	981	983	1	0	974	977	1	0	984	987	1	0	929	936
NAP-TMS	not among top first 100 hits				25	0.24	563	497	76	0.75	665	543	not among top 100 hits			
TCS-2TMS	not in NIST 17															
DHDPE-2TMS	not among top first 100 hits				not among top first 100 hits				not among top first 100 hits				not among top first 100 hits			
DH-BP-2TMS	1	0	890	909	1	0	844	880	1	0	816	895	1	0	772	818
4,4'-BP-2TMS	2	0.01	852	962	1	0	718	735	2	0.01	841	866	4	0.03	701	889
BPF-2TMS	1	0	940	953	1	0	595	630	1	0	895	915	64	0.63	522	584
SFA-2TMS	not among top first 100 hits				not among top first 100 hits				N/A				N/A			
BzPb-TMS	1	0	983	988	1	0	879	788	5	0.04	784	885	N/A			
BPE-2TMS	1	0	935	947	1	0	891	907	1	0	854	874	1	0	860	885
BPA-2TMS	1	0	969	973	1	0	973	975	1	0	980	980	2	0.01	598	636
BP-8-TMS	not in NIST 17															
CBD-2TMS	1	0	866	895	1	0	793	843	1	0	793	843	1	0	731	785
CBZ-TMS	1	0	967	985	3	0.02	635	774	5	0.04	404	568	N/A			
BPC-2TMS	1	0	975	981	1	0	941	949	1	0	944	956	1	0	826	861
BPB-2TMS	1	0	972	980	1	0	935	958	1	0	978	983	1	0	882	936
CBC-2TMS	not in NIST 17															
Δ^9-THC-TMS	1	0	850	863	1	0	721	735	1	0	897	909	1	0	573	639
TMBA-2TMS	1	0	883	911	1	0	794	825	1	0	931	945	2	0.01	757	847

BPCL-2TMS	1	0	957	964	1	0	933	939	1	0	970	975	1	0	858	875
CBN-2TMS	1	0	922	946	1	0	940	964	1	0	940	964	1	0	697	790
COD-TMS	1	0	489	524	1	0	568	556	1	0	317	354	3	0.02	177	192
BPZ-2TMS	1	0	972	973	1	0	906	910	1	0	875	878	1	0	903	923
6-MAM-TMS	1	0	718	720	1	0	528	534	2	0.01	640	657	3	0.02	551	574
11OHTHC-2TMS	1	0	810	824	1	0	624	645	1	0	793	800	not among top 100 hits			
E1-2TMS	1	0	886	900	1	0	873	893	1	0	816	826	1	0	821	850
Δ^9-THCA-2TMS	2	0.01	310	340	1	0	426	450	1	0	385	417	not among top 100 hits			
BPS-2TMS	not in NIST 17															
E2-2TMS	1	0	748	760	1	0	593	613	1	0	711	732	2	0.01	658	724
BPAP-2TMS	not in NIST 17															
EE2-TMS	1	0	782	793	1	0	712	727	1	0	713	728	1	0	541	582
11NOR9THC-2TMS	1	0	552	654	1	0	587	643	1	0	670	699	1	0	516	528
BPM-2TMS	not in NIST 17															
BPP-2TMS	not in NIST 17															
BPBP-2TMS	not in NIST 17															
BPPH-2TMS	not in NIST 17															
BPFL-2TMS	not in NIST 17															

Chapter 5

Chemometrics-Based Evaluation of the Stability of TMS Derivatives and Related Issues

This chapter presents the optimization of derivatization procedures, chemometrics-based evaluation of the stability of TMS derivatives, and evaluation of the associated MU. The chapter is divided into two sections. First, the problem is introduced, and then the work that addresses it is presented. The results have been submitted for publication to the journal *Environmental Research*. The work presented in this Chapter confirms the **H1** hypothesis of this dissertation and that if samples are not stored correctly, poor stability and chromatographic behavior, including peak shape and R_t of silylated derivatives, can arise. This problem can have a significant impact on the confidence and accuracy of their identification and quantification,

5.1 Problem Description

Analytical efforts in EEA mainly focus on the characterization of the chemical exposome in terms of occurrence, fate, and effect on CEC in versatile environmental compartments. As previously discussed, GC-MS, LC-MS, and NMR are the most commonly employed analytical techniques [18] that can successfully handle the chemical diversity of CEC to be identified. LC-MS platforms have become more readily available in the last decade, providing validated and reliable quantification even at ultra-trace levels. LC-MS is preferred as it does not require special derivatization. Due to its sufficient mass accuracy and resolving power at lower cost, GC-MS remains a widely employed instrumental method in EEA [11]. It helps address the ever-growing demand to provide thorough chemical information of as many analytes of interest as possible [31]. They are appropriate analytical tools for various anthropogenic pollutants [107] and offer greatly reproducible structural information inherent to EI spectra that follows predictable and thoroughly studied fragmentation patterns and broad internal energy distribution [64].

GC-MS analytical systems carry several limitations when it comes to EEA, among which most important are the inability to cover polar or semi-polar analytes and the frequent generation of GC-EI-MS spectra of extensive fragmentation, with the parent ion missing [18]. Derivatization, in particular silylation, in conjunction with GC-MS and GC-MS/MS, provides a highly sensitive and selective method for semi-volatile and thermolabile compounds. The result is improved volatility, thermal and catalytical stability, and chromatographic and spectrometric behavior in peak shape and response [47]. As discussed in Section 2.3.3.1.2, such compounds have moieties with acidic protons, such as -OH, -COOH, NH_2, and -SH, as well as -CHO and -CO groups [108]. Silylation shifts compound peaks, especially of low M_w, from lower mass, noisy to less contaminated,

higher-mass chromatographic regions, empty of interfering peaks, which results in separation that is more satisfactory. TMS derivatives have improved ionization efficiencies and undergo more extensive fragmentation during EI than parent compounds [109]. The resulting GC-EI-MS spectra contain significantly more structural information, more suitable for cheminformatics-assisted CA.

For high-throughput annotation of CEC using GC-MS analytical platforms, the introduction of a rapid, mild-condition, single-step derivatization reaction is required. This step would result in a high and reproducible yield of stable silyl derivatives. In the best case, by-products that would interfere with the chromatogram or harm the GC column should not be generated [47], [59]. In order to achieve this for structurally diverse CECs, thorough optimization of the derivatization conditions according to the reaction kinetics of the investigated CEC is required. Optimization includes the selection of derivatization agent, derivatization agent-to-sample ratio, and derivatization conditions, such as temperature and duration. Although of great importance, the aforementioned issues affecting the stability of CEC-TMS derivatives have not been thoroughly studied in a representative CEC collection. Instead, method development in CEC identification and quantification studies usually does not focus on optimizing derivatization conditions and investigating the stability of the resulting silyl derivatives.

Upon derivatization, silyl derivatives are susceptible to hydrolysis, potentially resulting in irreversible degradation and transformation during sample preparation, storage, and analysis, indicating unsatisfactory stability. The irreversible degradation and transformation may negatively influence the accuracy, linearity, reproducibility, and measurement uncertainty of their quantification, finally leading to false information regarding their presence and occurrence in different environmental compartments.

Chapter 5 also investigates the effect of the most important variables, i.e., the type of silylating agent, derivatization temperature, and time, on derivatization efficiency and yield. Their optimization was performed using chemometrics-based tools, such as contour plotting of relative response factors (RRF), principal component analysis, and hierarchical clustering analysis. The representative selection of 70 CEC included active pharmaceutical ingredients, steroid hormones, illicit drugs, their metabolites, cannabinoids, preservatives, UV filters, plasticizers, and other industrial chemicals. Stability was studied in EtAc, and AWW extracts for 20 weeks at 25°C, 4°C, and -18°C and during five consecutive freeze/thaw cycles at -18℃. Finally, we estimated the measurement uncertainty (MU) during the stability examination in order to determine the influence of MU on the results.

5.2 Related Publication

Submitted manuscript

Ljoncheva, M., Heath, E., Heath D., Džeroski, S., Kosjek, T., Contaminants of emerging concern: silylating procedures, evaluation of the stability of silyl derivatives and associated measurement uncertainty, *Environmental Research* (submitted, 25 August 2022)

This publication describes the following scientific contributions:

- Confirmation that the derivatization conditions for a representative selection of CEC with significant structural and physicochemical diversity can be efficiently optimized by using chemometrics tools.
- The TMS derivatives of polyhydroxy CEC and estrogen hormones exhibit the highest instability among the 70 CEC-TMS studied.
- Confirmation that environmental samples analyzed for the presence and quantification of numerous chemodiverse CEC using silylation and GC-MS are safe to be stored at -18℃ for up to 20 weeks in solvent and AWW extracts.
- Confirmation that environmental samples analyzed for the presence and quantification of certain CEC via their TMS derivatives should not be frozen and thawed more than two times to maintain losses of TMS derivatives below 20%.

- It is possible to use a measurement uncertainty scheme for stability analyses that can be used in general, not only for the specific compounds we consider.
- Confirmation that CEC-TMS that had concentration deviations during stability studies could not be attributed to the measurement uncertainty.

Contaminants of emerging concern: silylating procedures, evaluation of the stability of silyl derivatives and associated measurement uncertainty

Ljoncheva M.[§,ʃ], Heath E.[§,ʃ], Heath D.[ʃ], Džeroski S.[ς], Kosjek T.[§,ʃ,*]

[ʃ] *Jožef Stefan Institute, Department of Environmental Sciences, Jamova cesta 39, 1000 Ljubljana, Slovenia*

[§] *Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*

[ς] *Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia*

*Corresponding author.

Tel: +386 14773288

E-mail: tina.kosjek@ijs.si

ABSTRACT

Analyte range of gas chromatography-mass spectrometry (GC-MS), widely used in environmental analysis, can be significantly broadened by derivatization. Silyl derivatives have improved volatility and thermal stability, chromatographic and mass spectrometric behaviours, and thus detection, structural elucidation and quantification. However, silylation use is often hindered by the stability of generated derivatives and the need to optimize silylation conditions. In this study, we optimized the derivatization conditions for 70 selected contaminants of emerging concern (CEC) using chemometrics approaches. N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA), N, O-bis(trimethylsilyl)trifluoroacetamide (BSTFA) and BSTFA + 1% trimethylchlorosilane (TMCS) were investigated, among which the latter gave the best yield. CEC were grouped in three derivatization protocols: 60°C/45 min, 70°C/90 min, and 70°C/45 min. The short- and long-term stability of the CEC-trimethylsilyl (TMS) derivatives were examined in a solvent and artificial wastewater (AWW) extract at 25˚C, 4˚C and -18˚C, and during repeated five freeze-thaw (F/T) cycles, at two concentration levels: 100 µg/L and 1000 µg/L. Except for TMS derivatives of shikimic acid (SHA), quinic acid (QA) and sulphanilamide (SFA), the remaining derivatized compounds were stable in solvent (EtAc) for 28 days. In AWW extract, TMS derivatives of citric acid (CA), 17β-estradiol (E2), estriol (E3) and 17α-ethinyl estradiol (EE2) were unstable at 25˚C and 4˚C. Within up to 20 weeks, only the TMS derivatives of CA, meso erhythritol (ERY) and bisphenol BP (BPBP) were unstable. The most significant hydrolytic breakdown was observed during repeated F/T cycles. After three cycles, ≤20% of the initial concentration of six and nine CEC-TMS derivatives had degraded in solvent and AWW extracts, respectively. According to the deep statistical comparison (DSC) approach, the most prominent degradation was observed for TMS derivatives of E2, CA 9-hydroxyfluorene (9-HF), estrone (E1) and trans-3'-hydroxycotinine (T3HC) in a solvent; E2, CA, 9HF, E3 and E1 in AWW extracts and ERY, E2, CA, 9HF and E1 in both matrices. Finally, the sample amount of CEC accounted for most of the measurement uncertainty (MU).

Keywords: contaminants of emerging concern, trimethylsilyl derivatives, optimization, derivatization, stability, measurement uncertainty

Funding sources: This work was supported by the Slovenian Research Agency (J1-6744, Program Groups P1-0143 and P2-0103). M.L. is funded by the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia (Contract No. 11011-85/2016).

Abbreviations: Δ^9-THC, Δ^9-tetrahydrocannabinol; 11N9THC, (±)-11-nor-9-carboxy-Δ^9-tetrahydrocannabinol; 11OHTHC, (±)-11-hydroxy-Δ^9- tetrahydrocannabinol; 22BPF, 2,2′-methylenediphenol; 24BPF, 4,4;-BP, 4,4′-biphenol; 4',4'-methanediyldiphenol; 4-NP, 4-nonylphenol; 4-OP, 4-*tert* octylphenol; 6-MAM, 6-monoacetylmorphine; 8HQ, 8-hydroxyquinoline; 9-HF, 9-hydroxyfluorene; AA, adipic acid; ACN, acetonitrile; AUC, area under the curve; AWW, artificial wastewater; BA, benzoic acid; BP-8, 2,2'-dihydroxy-4-methoxybenzophenone; BPA, bisphenol A; BPAF, bisphenol AF; BPAP, bisphenol AP; BPB, bisphenol B; BPBP, bisphenol BP; BPC, bisphenol C; BPCL, bisphenol CL; BPE, bisphenol E; BPF, bisphenol F; BPFL, bisphenol FL; BPM, bisphenol M; BPP, bisphenol P; BPPH, bisphenol PH; BPS, bipshenol S; BPZ, bisphenol Z; BSA, N,O-bis(trimethylsilyl)acetamide; BSTFA, N,O-bistrifluoroacetamide; BuPb, butylparaben; BZECG, beznoylecgonine; BzPb, benzylparaben; CA, citric acid; CBC, cannabichromene; CBD, cannabidiol; CBN, cannabinol; CBZ, carbamazepine; CEC, contaminant of emerging concern; CLA, clofibric acid; CLP, 2-benzyl-4-chlorophenol; COD, codeine; DF, diclofenac; DH-BP, 2,4-dihydroxybenzophenone; DHDPE, 4,4'-dihydroxydiphenyl ether; DSC, deep statistical comparison; E1, estrone; E2, 17β-estradiol; E3, estirol; EE2, 17α-ethinylestradiol; EI, electron impact ionization; ERY, meso-erythritol; ESI, electrospray ionization; EtAc, ethyl acetate; EtPb, ethylparaben; F/T, freeze/thaw; GC, gas chromatography; H-BP, 4-hydroxybenzophenone; HCA, hierarchical clustering analysis; HMDS, hexamethyldisilazide; HPP, 4-cumylphenol; IB, ibuprofen; IbuPb, isobutylparaben; IS, internal standard; KET, ketoprofen; LC, liquid chromatography; MDL, method detection limit; MEC, mecoprop; MeOH, methanol; MePb, methylparaben; MORPH, morphine; MQL, method quantification limit; MS, mass spectrometry; MSD, mass selective detector; MSL, mass spectral libraries; MSTFA, N-methyl-N-(trimethylsilyl)trifluoroacetamide; MTBSFTA, N-*tert*-butyldimethylsilyl-N-methyltrifluoroacetamide; MU, measurement uncertainty; MW, molecular weight; NAP, naproxen; PAA, phenylacetic acid; PC, principal component; PCA, principal component analysis; PrPb, propylparaben; QA, quinic acid; QC, quality control; RES, resorcinol; RRF, relative response factor; R_t, retention time; r_s, relative standard uncertainty; SFA, sulfanilamide; SHA, shikimic acid; SIM, selected ion monitoring; SPE, solid-phase extraction; T3HC, trans-3'-hydroxycotinine; TBDMS, tert-butyldimethylsilyl; TBDMCS, *tert*-butyldimethylchlorosilane; TCS, triclosan; THCA, Δ^9-tetrahydrocannabinolic acid; TMBA, 4,4′-isopropylidenebis(2,6-dimethylphenol; TMCS, trimethylchlorosilane; TMIS, trimethyliodosilane; TMPAH, trimethylphenyl ammonium hydroxide; TMS, trimethylsilyl; TMS-DEA, trimethylsilyldiethylamine; u, combined standard uncertainty; U, relative expanded uncertainty.

88 **1. INTRODUCTION**

89 Investigating the chemical diversity of compounds of emerging concern (CEC) in multiple environmental
90 compartments requires versatile analytical platforms. To date, gas and liquid chromatography coupled with mass
91 spectrometry (GC-MS and LC-MS) are the methods of choice [1], and although LC-MS has become more prevalent,
92 GC-MS remains a widely used and cost-effective analytical platform with substantial compound coverage [2].
93 Moreover, it offers excellent reproducible structural information inherent to electron impact (EI) ionization that follows
94 predictable and well-studied fragmentation patterns. Such information can be used for structural elucidation in
95 suspect screening, non-targeted workflows, and sensitive and selective high-throughput quantification when using
96 targeted approaches [2–4]. The conventional compound range of the GC-MS analytical platforms is based on volatile
97 and thermostable compounds. However, this range can be broadened using derivatization to semi-volatile and
98 thermolabile compounds, including many CEC containing active hydrogen functional groups: hydroxyl (-OH), carboxyl
99 (-COOH), primary amine (-NH_2), secondary amine (-R_1R_2NH), aldehyde (-CHO) or thiol (-SH). Typical derivatization
100 includes acylation, alkylation, esterification, and silylation, the latter being the most commonly used derivatizing
101 method due to availability, low cost and low sample destruction.

102 During silylation, the active H atoms bind to electronegative elements (O, N, S or P), which are replaced with
103 either trimethylsilyl (TMS) (-$Si(CH_3)_3$) or a tert-butyldimethylsilyl (TBDMS) (-$SiCH_2CH_2C(CH_3)_3$) group, thus reducing
104 the number of reactive sites and in turn, the possibility of hydrogen bonding. Silylation must occur in polar aprotic
105 solvents, such as ethyl acetate (EtAc), acetonitrile (ACN) and acetone. Protic solvents, such as water, methanol
106 (MeOH), ethanol (EtOH), formic acid and acetic acid, are unsuitable as they react with the silylating agents.

107 Among the silylation reagents available, hexamethyldisilazane (HMDS), N, O-bis(trimethylsilyl) acetamide (BSA),
108 N-(trimethylsilyl)diethylamine (TMS-DEA), MSTFA and BSTFA are most popular. They react readily with most
109 alcohols, phenols, carboxylic acids, amino acids, saccharides and indoles [5]. In many cases, a catalyst is added to
110 increase their TMS donor potentials, such as trifluoroacetic acid (TFA) for HMDS and 1% trimethylchlorosilane
111 (TMCS) and N-trimethylsilylimidazole (TMSI) for MSTFA and BSTFA. Active H atoms can also be replaced with a
112 TBDMS group, using N-*tert*-butyldimethylsilyl-N-methyltrifluoroacetamide (MTBSTFA) as a silylation agent. Under
113 more vigorous conditions, higher M_W TBDMS derivatives are generated, which are 1,000-10,000 times more stable to
114 hydrolysis than TMS derivatives [5–7]. However, TBDMS derivatization leads to “dirty” extracts, poor peak resolution
115 and poor MS spectra for parabens, 4-nonylphenol (4-NP) and 4-*tert* octylphenol (4-OP) [8], benzophenone-type UV-
116 filters [9], and completely fails to derivatize aliphatic -OH and -NH_2 groups, as in drugs of abuse [10].

117 Silylation improves a compound’s volatility and thermal stability [6,11], chromatographic and mass spectrometric
118 behavior regarding peak shape and ionization. These lead to better sensitivity, selectivity and linearity. Furthermore,

119 silylation prevents compound loss by preventing undesirable interactions with silanol groups in the injector liner or the
120 column [9]. Silylation also shifts compound peaks, especially those with lower M_W [9], to higher-mass
121 chromatographic regions, with less noise and interfering peaks, resulting in satisfactory separation and MS spectra
122 with significant structural information [5,8]. Finally, silylation provides additional structural information on parent
123 compounds, as TMS derivatives have improved ionization efficiency and undergo more extensive fragmentation
124 during EI than parent compounds, allyl, isopropyl and TBDMS derivatives [9]. Richer fragmentation patterns have a
125 higher diagnostic value for confident compound identification and sensitive quantification [3,6,8,10].

126 Silylation is considered a laborious and time-intensive step that adds complexity and uncertainty to the analytical
127 workflow. It is also limited to CEC, whose silyl derivatives' M_W are within the linear range of the mass analyzer (up to
128 650 Da). Silyl derivatives with M_W >800 Da will not elute from the GC column even at its maximum working
129 temperature [6] and can permanently contaminate the GC-MS system. Different by-products, partial derivatives and
130 artefacts can result from the silylation of polyfunctional compounds. Together with compounds with functional groups
131 of different stereochemistry, it can result in multiple chromatographic peaks [6]. Susceptibility to hydrolysis of silyl
132 derivatives may lead to irreversible degradation and transformation during sample preparation, storage and analysis.
133 The result is a decrease in analytical reproducibility, linearity and accuracy. Moreover, it can lead to false negatives
134 when identifying unknown CEC. The stability of silyl derivatives and the influence of the issues mentioned above are
135 not well researched since method development usually overlooks the optimization of derivatization and investigation
136 of the stability of the resulting derivatives.

137 This study investigates relevant and practical aspects of silylation prior to GC-MS analysis by investigating the
138 effects of the most critical variables: silylation agent, derivatization temperature and time, on derivatization yield of 70
139 CEC with a broad range of physicochemical properties, including pharmaceuticals, steroid hormones, illicit drugs and
140 their metabolites, cannabinoids, preservatives, UV-filters, plasticizers and other industrial chemicals. The stability of
141 TMS derivatives was assessed in solvent (EtAc) and AWW extracts over 20 weeks under relevant storage conditions
142 (-18°C, 4°C and 25°C) and during five consecutive freeze-thaw (F/T) cycles. Finally, the MU was estimated as a
143 significant factor affecting result interpretation.

144

145 ## 2. MATERIALS AND METHODS

146 ### 2.1. Standards and reagents

147 The initial compound selection included 103 CEC, thirty-three (see Supplementary material SI-II) which had
148 inconsistent, low responses or distorted peak shapes (with extensive tailing and low *m/z* ions in the MS spectra) and
149 accordingly, were considered unsuitable for robust quantification and were excluded from the CEC selection. For

150 some CEC, mixtures of mono- and di-substituted TMS derivatives were formed under all silylation conditions.
151 Information regarding the analytical standards of the remaining 70 CEC, other reagents, solvents and materials used
152 are given in the Supplementary material (SI-I). Analytical standards include: benzoic acid (BA), phenylacetic acid
153 (PAA), resorcinol (RES), methylparaben (MePb), ERY, adipic acid (AA), salicylic acid (SA), 8-HQ, ethylparaben
154 (EtPb), clofibric acid (CLA), isopropyl paraben (iPrPB), ibuprofen (IB), mecoprop (MEC), 4-OP, propylparaben (PrPb),
155 isobutylparaben (iBuPb), butylparaben (BuPb), shikimic acid (SHA), CA, 9-HF, (-)-quinic acid (QA), T3HC, 4-
156 cumylphenol (HPP), 2-benzyl-4-chlorophenol (CLP), 4-NP, 2,2′-methylenediphenol (22BPF), bisphenol AF (BPAF), 4-
157 hydroxybenzophenone (H-BP), ',4'-methanediyldiphenol (24BPF), naproxen (NAP), triclosan (TCS), 4,4'-
158 dihydroxydiphenyl ether (DHDPE), 2,4-dihydroxybenzophenone (DH-BP), 4,4′-biphenol (4',4'-BP), bisphenol F (BPF),
159 bisphenol E (BPE), benzylparaben (BzPb), ketoprofen (KET), bisphenol A (BPA), SFA, 2,2'-dihydroxy-4-
160 methoxybenzophenone (BP-8), cannabidiol (CBD), carbamazepine (CBZ), bisphenol C (BPC), bisphenol B (BPB),
161 diclofenac (DF), benzoylecgonine (BZECG), cannabichromene (CBC), Δ^9-tetrahydrocannabinol (Δ^9-THC), 4,4′-
162 isopropylidenebis(2,6-dimethylphenol) (TMBA), bisphenol Cl (BPCL), cannabinol (CBN), codeine (COD), morphine
163 (MORPH), BPZ, 6-monoacetylmorphine (6-MAM), (±)-11-hydroxy-Δ^9-tetrahydrocannabinol (11OHTHC), E1, Δ^9-
164 tetrahydrocannabinolic acid (THCA), bisphenol S (BPS), E2, bisphenol AP (BPAP), EE2, (±)-11-nor-9-carboxy-Δ^9-
165 tetrahydrocannabinol (11N9THC), bisphenol M (BPM), E3, bisphenol P (BPP), BPBP, bisphenol PH (BPPH) and
166 bisphenol FL (BPFL). The IUPAC names, common names, abbreviations, CAS numbers and M_W of CEC and their
167 TMS derivatives are given in Table SI-I.

168 **2.2 Preparation of standard solutions**

169 Individual stock solutions (150 µg/mL) were prepared in EtAc, MeOH or ACN and stored in the dark at 4°C.
170 These stock solutions were then used to prepare a series of working standards (0.12 µg/mL and 1.2 µg/mL) for
171 stability testing in solvent and AWW extract (0.5 µg/mL and 5.0 µg/mL). To optimize the derivatization reaction, we
172 prepared a standard solution (1.2 µg/mL) of each CEC and 19 µg/mL of methadone to serve as an internal standard
173 (IS), as it cannot be silylated and has reproducible peak intensities. Stability was investigated at two concentrations
174 (100 µg/L and 1000 µg/L) in two matrices, i.e. solvent and AWW extracts. The CEC were divided into three groups
175 based on the optimal silylation conditions (see 3.1.2). Group standards were then used to prepare both samples and
176 quality control (QC) samples. The calibrants and IS solutions were stored at 4°C and used to prepare the calibration
177 curve at each time point within the stability study. Stock solutions of individual IS (Supplementary material SI-I) were
178 prepared at a concentration of 100 µg/mL for each IS, except for $^{13}C_6$-EtPb, $^{13}C_6$-PrPb, CBD-d_3, E2-d_5 (50 µg/mL) for
179 Group 1, CBD-d_3 and E2-d_5 (75 µg/mL) for Group 2 and 11OH-Δ^9-THC-d_3 (75 µg/mL) for Group 3. The stock solutions

were then used to prepare IS mixtures for stability testing in solvent (2.28 µg/mL) and in AWW extract (2.8 µg/mL) used in the stability experiments.

2.3 Optimization of silylation

The effect of **(1)** silylating agent: MSTFA, BSTFA and BSTFA + 1%TMCS; **(2)** derivatization time: 30min, 45 min, 60 min and 90 min, and **(3)** derivatization temperature: 60°C, 70°C and 80°C, was assessed to achieve the highest derivatization yield for each CEC. In total, 36 experiments were conducted (Table 1). In each experiment, 150 µL of the standard mixture (1.2 µg/mL) and 30 µL of the silylating agent were transferred into a 250 µL insert in an amber glass GC vial (experiments 1-36). Once derivatization was complete, samples were cooled to room temperature (10-15 min), spiked with 10 µL of 19 µg/mL of methadone, and analyzed by GC-MS. Each experiment was conducted in triplicate. Derivatization was then evaluated regarding reaction yield, conversion of polar compounds and by-product formation. Silylation completeness, and silylation efficiency, were estimated using the average relative response factor (RRF) as the average ratio between the area under the curve (AUC) of the CEC-TMS and AUC of the IS ($AUC_{CEC-TMS}/AUC_{IS}$) of three parallels. Higher RRFs indicate better silylation yield, while conversion of polar compounds and by-product formation were evaluated by visual inspection of the chromatograms. Once the optimal silylating procedure was obtained, the repeatability of derivatization was assessed by derivatizing the standard mixtures at two concentration levels (100 µg/L and 1000 µg/L) in triplicate. Repeatability was recorded as the relative standard deviation (RSD) of the RRF.

2.4 Sample preparation

Samples for the solvent-based stability experiments were performed by transferring 150 µL of each standard mixture (group 1 to 3) into a 250 µL insert in an amber glass GC vial. The mixture was then derivatized following the optimized protocol and stored under the specified experimental conditions described in Figure 1. At each sampling point, 150 µL of the IS mixture (2.28 µg/mL) were derivatized separately by adding 30 µL of the silylation agent following the optimal derivatization protocol. Afterwards, 10 µL of separately derivatized IS solution were added to each sample. The stability experiments were performed using artificial wastewater (AWW) effluent from an in-house built bench-scale wastewater treatment plant. Samples (200 mL) of AWW were collected and filtered sequentially through a glass-microfiber (0.5 µm, Machery Nagel, Düeren, Germany) and a 0.45 µm cellulose nitrate membrane filter (Sartorius Stedim Biotech GmbH, Göttingen, Germany). The filtrate was then extracted by solid-phase extraction (SPE) using an Oasis HLB Prime cartridge (60mg, 3cc) on a Supelco Vacuum Manifold (Bellefonte, USA). After loading, the sorbent was dried under vacuum (-10 mmHg, 30 min) and eluted with EtAc (3 x 0.6 mL). The solvent was

211 evaporated to 0.5 mL under N_2 (10-15 psi) at 40ºC. One-hundred and forty µL of the standard mixture and 30 µL
212 silylation agent were added, and the contents were derivatized. The derivatized extracts were stored as described in
213 Figure 1. Prior to GC-MS, samples were spiked with 30 µL of a separately derivatized IS mixture (150 µL standard
214 mixture of IS at 2.8 µg/mL + 30 µL silylation agent). Samples were prepared identically for the F/T experiment in
215 solvent and AWW extracts. All samples were prepared in triplicate (n=3), stored and further processed as described
216 in 2.5.

217 **2.5 GC-MS analysis**

218 Samples were analyzed using an Agilent 7890B/5977A series GC-MSD (Agilent, USA). Chromatographic
219 separation was achieved on an Agilent DB-5MS UI fused-silica capillary column (30 m x 0.25 mm x 0.25 µm; Agilent,
220 USA). Helium (99.99999% purity) was used as the carrier gas at a 1.2 mL/min flow rate. The manifold, ion source,
221 and transfer line temperatures were 230°C, 150°C and 250°C, respectively. The column oven temperature programs
222 for the three CEC-TMS groups are given in Supplementary material (SI-III). Injections (1 µL) were performed in
223 splitless mode. Ionization was achieved at 70 eV, and the mass spectrometer was operated in full scan mode
224 covering a mass range (*m/z*) of 50-600 amu. The GC peaks were identified during initial optimization experiments
225 based on their retention times and characteristic MS spectra. Data processing was performed using Mass Hunter
226 Quantitative Analysis v B.07.00 (Agilent, USA).

227 **2.6 Method performance evaluation**

228 The method was validated using a least square regression analysis of a matrix-matched 10-point calibration
229 curve: 5, 10, 25, 50, 75, 100, 250, 500, 750 and 1000 µg/L. Matrix-matched calibration was achieved by spiking each
230 SPE extract with a known concentration of CEC, after which 30 µL silylating agent and 10 µL of the separately
231 derivatized ISs mixture were added. Method performance parameters, including linearity, accuracy, precision
232 (method and instrument repeatability), sensitivity, method detection limits (MDL) and method quantification limits
233 (MQL), were determined for each CEC in solvent and AWW.

234 **2.7 Stability of TMS derivatives**

235 The stability over short-, long-term and five F/T cycles of the CEC-TMS derivatives was investigated in solvent
236 and AWW extracts at two concentration levels: 100 µg/L (low level, LL) and 1000 µg/L (high level, HL) according to
237 the experimental design (Figure 1). Experimental conditions were selected based on standard practices, i.e., short-
238 term storage, i.e., one week on the autosampler rack or in the refrigerator (Exp. 1-4: 25°C and 4°C) and long-term
239 storage (20 weeks) in a refrigerator or freezer (Exp. 5-8: 4°C and -18°C).

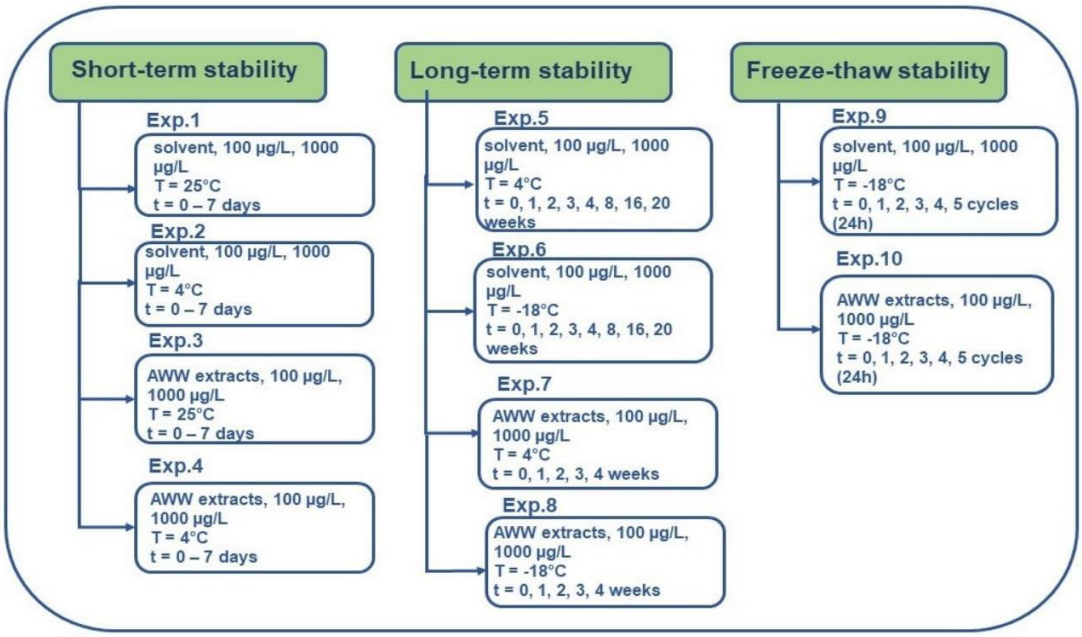


Figure 1. Flow chart showing the experimental design for testing the stability of CEC-TMS derivatives.

Short-term stability was tested for one week at room temperature (25°C) and in the refrigerator (4°C) at 8-time points: immediately after derivatization (t=0) and every 24 h for 7 days. For long-term stability tests, seven samples were collected immediately after derivatization (t=0), then every 7th day for the first 4 weeks, followed by every 28th day up to week 20. All samples were kept in the dark during storage.

The F/T experiments (Exp. 9-10) involved five consecutive F/T cycles within a 24h period in both solvent and AWW extracts. Each freeze-thaw cycle consisted of sample storage in the freezer at -18ºC for 2h, thawing to room temperature (25ºC) for 0.5-1h. Samples were analyzed by GC-MS immediately after derivatization (t=0) and after each freeze-thaw cycle.

At each time point (short-, long-term and F/T stability experiments), QC samples were prepared for both concentration levels in both matrices by freshly spiking and derivatizing samples before analysis. All samples were prepared and analyzed in triplicate, except QC samples prepared and analyzed in duplicate. The stability of CEC-TMS derivatives was evaluated based on the average relative concentration of the CEC-TMS at a specific time point, expressed as a percentage of the initial concentration (100% at t=0).

253 **2.8 Data analysis and visualization**

254 Statistical analysis and data visualization were performed using ClustVis [13] and the R programming language in
255 the R Studio environment [14]. Contour plots, principal component analysis (PCA) and hierarchical cluster analysis
256 (HCA) of the normalized RRF were used to analyze the optimization data. A statistical comparison of two or more
257 sample datasets was performed to assess the influence of temperature, storage time and matrix on the stability of the
258 CEC-TMS derivatives. For this, the Shapiro-Wilk test was used for checking data normality and comparing
259 distributions between samples and QC samples and Levene's test for determining homoscedasticity of variance of
260 the samples and QC samples. The influence of storage temperature on stability was checked using paired t-test for
261 paired samples and Wilcoxon signed-rank test (two-sided) for unpaired samples. Statistical significance was
262 evaluated at the $\alpha = 0.05$ level for all tests. The CEC-TMS derivatives' stability was ranked using the deep statistical
263 comparison (DSC) ranking approach, which combines multiple criteria decision analysis and deep statistical rankings
264 [15]. In DSC, the CEC-TMS derivatives are ranked in descending order according to their relative concentrations,
265 such that the CEC-TMS derivative ranked #1 shows the most prominent deviation towards low concentrations, i.e.,
266 the lowest stability.

267

268 **2.9 Evaluation of measurement uncertainty (MU)**

269 Measurement uncertainty (MU) was estimated using a bottom-up approach following the Guide to the Expression
270 of Measurement Uncertainty [16]. The uncertainty budget was generated by combining the contributions of
271 repeatability, instrumental precision and amount of CEC in the sample. The uncertainties of individual components
272 and relative standard uncertainty (r_s) were calculated according to the sources and the degree of freedom for the r_s.
273 The r_ss were combined by the root-sum-of-squares method to give relative combined standard uncertainty (u). The
274 degrees of freedom were also calculated. The relative expanded uncertainty (U) was determined by multiplying the r_s
275 with an appropriate coverage factor (k=1.96).

276

277 **3. RESULTS AND DISCUSSION**

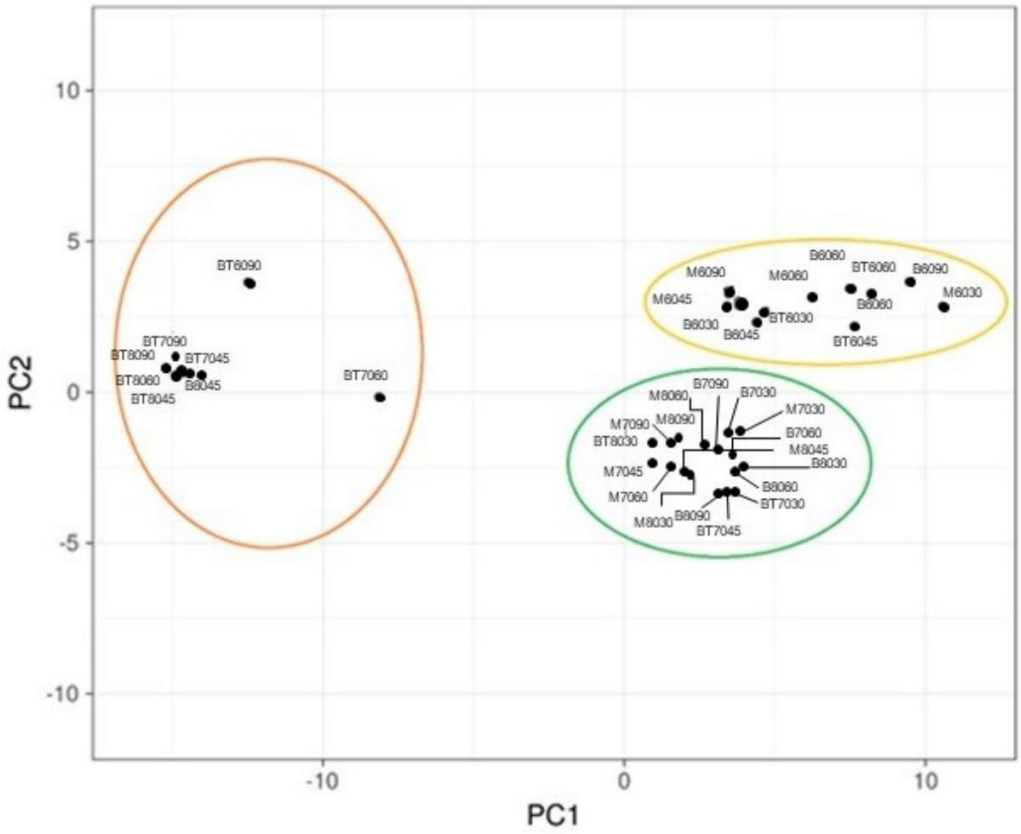
278 **3.1 GC-MS optimization**

279 After determining the R_t of the individual CEC-TMS in full scan mode, the GC-MS method was used for optimizing
280 silylation. For the stability studies, three GC-MS methods in selected ion monitoring (SIM) mode were developed for
281 each of the three CEC-TMS groups. One quantification (Q1) and two qualification ions (Q2, Q3) were selected per

CEC-TMS, where one was the molecular ion, and all three have a signal-to-noise (S/N) ratio ≥ 3:1 [17]. The GC-MS conditions and the corresponding R_t are given in the Supplementary material (SI-III). The R_t, SIM ions and selected surrogate standard for the final GC-MS methods are given in Supplementary material (SI-III, Table SI-II, Table SI-III and Table SI-IV for CEC-TMS groups 1, 2 and 3, respectively).

3.2 Optimization of silylation

Numerous silylating agents have been reported in the literature, with or without a catalyst, employed for silylation prior to GC-MS analysis, among which the three most frequently reported, MSTFA, BSTFA and BSTFA + 1% TMCS, were selected in the present study. They react differently towards hindered functional groups when used with a catalyst. Excess silylating agents and temperatures > 60˚C are generally known to favor the production of TMS derivatives, especially for unreactive functional groups [7]. Here, we avoided temperatures > 90˚C due to the risk of evaporation of the silylating agent solvent and degradation of the parent CEC and TMS derivatives. Also, reaction times > 90 min were not investigated, as they are not acceptable for high-throughput GC-MS analysis.



294 **Figure 2.** Two-dimensional PCA plot based on RRF values obtained under 36 optimization experiments.

295 The RRF was used to evaluate silylation results, as they provide a more reliable comparative analysis than
296 evaluation by comparing peak area [17] or response coefficients [18].

297 Preliminary experiments did not reveal the presence of underivatized compounds; therefore, silylation was
298 considered complete in all investigated experimental set-ups. All three agents reacted well with -OH and -COOH
299 groups of the selected CEC, with significant response variation for CEC with hindered -OH groups, $-NH_2$ and $-CONH_2$
300 groups. The formation of by-products (N-methyltrifluoroacetamide and trifluoroacetamide) of MSTFA was observed
301 due to its higher volatility, although they usually elute with the solvent front (i.e. at temperatures ≤70˚C). Additionally,
302 for MSTFA, distorted peak shapes with front tailing were observed, especially for MEC-TMS, most probably due to
303 the stronger interaction between MSTFA and the column phase when compared to BSTFA and BSTFA + 1% TMCS
304 [19].

305 Hierarchical cluster analysis (HCA) of the silylation optimization data (Supplementary material, SI-III, Figure SI-II)
306 revealed three major clusters, which were confirmed by PCA (Figure2), which gives the loadings and score plots for
307 the variables (the experiments) and the objects (CEC-TMS derivatives) in two-dimensional space. The first two PC
308 account for 95.9% of the total variance (PC1 for 87.8% and PC2 for 8.8% of the total variance). The results indicate
309 that one experiment per cluster should be selected; therefore, a closer investigation of the effects of each CEC-TMS
310 was performed.

311 Contour plotting was performed for each CEC-TMS derivative (Supplementary material, SI-IV, Figure SI-III)
312 by plotting time (x-axis) against temperature (y-axis). The responses of the CEC-TMS derivatives, represented as
313 RRF, clearly indicate the following general findings:

314 **1)** A 30 min reaction time proved unsatisfactory, especially when BSTFA and BSTFA + 1% TMCS were used.

315 **2)** Derivatization efficiency for all CEC with MSTFA decreases with increasing temperature, which is especially
316 pronounced for paraben-TMS derivatives agreeing with the data published in the literature [20].

317 **3)** BSTFA + 1% TMCS was superior to MSTFA for most of the CEC-TMS derivatives, both in RRF and response
318 repeatability (<10% RSD for all CEC), which confirms literature findings that BSTFA is a more reactive silyl donor
319 than MSTFA [5]. Also, all tested temperatures proved appropriate.

320 **4)** Comparable RRF were generated for more than half of the generated CEC-TMS derivatives with BSTFA and
321 BSTFA + 1% TMCS. Although unexpected, this is explained by the low TMS donor strength of TMCS in neutral
322 solvents, such as EtAc [3].

5) BSTFA + 1% TMCS is most appropriate for the silylation of hindered -OH groups, compounds with more than one functional group and functional groups in different steric environments [5].

The CEC with sterically unhindered -OH groups are readily derivatized. Their responses are satisfactory under all examined conditions, with the highest derivatization yields with BSTFA + 1% TMCS. This observation includes bisphenols (BPF, BPAF, BPE, BPA, BPC, BPB, BPCL, BPZ, BPS, BPAP, BPM, BPP, BPPH, BPFL, 22BPF, 24BPF and TMBA), other endocrine disrupting compounds (MePb, EtPb, IPrPb, PrPb, IBuPb, BuPb, BzPb, 4-OP, 4-NP, TCS, RES, 4,4'-BP, 8-HQ, CLP, HPP, DHDPE), cannabinoids (CBC, Δ^9-THC, CBN), illicit drug metabolites (6-MAM, 11OH-THC) and 9-HF. BSTFA + 1% TMCS was optimal for silylation of endocrine disruptive CEC (MePb, EtPb, IPrPb, PrPb, IBuPb, BuPb, 4-OP, 4-NP, TCS, BzPb, BPA and other bisphenols). This is supported by other studies. where BSTFA + 1% TMCS was found to be more efficient than BSTFA, BSA, acetic anhydride and pentafluoropropionic anhydride [20].

Compounds with significantly hindered -OH groups, such as illicit drugs T3HC, COD, and MORPH, show no preferences when either MSTFA or BSTFA are used, except for lower responses observed under extreme conditions (80˚C, >70 min), and higher responses at ≤70˚C/40-60 min with BSTFA + 1% TMCS. In one study, silylation at 70°C/30 min with BSTFA or BSTFA + 1% TMCS was shown to give satisfactory responses for Δ^9-THC, 11OHTHC, 11NTHCA, COD, MORPH, BZECG, and 6-MAM, although more vigorous conditions were also reported, e.g. BSTFA + 1% TMCS for 20 min or 30 min at 100°C for BZECG, COD, MORPH and 6-MAM [21]. Also, the hindered -OH groups of Δ^9-THCA and 11N9THC require more vigorous conditions, which induce their decarboxylation. In our case, the best response was achieved with BSTFA + 1% TMCS at 70˚C/40-60 min.

Carboxyl groups are less reactive than -OH, but any strong silyl donor is suitable for silylation, although BSTFA + 1% TMCS is most frequently used [5,6]. For unhindered -COOH groups, such as in BA, PAA, AA, MEC, BZECG, IB, NAP, CLA, KET and DF, the highest responses were obtained at ≤70˚C/45-60 min with BSTFA + 1% TMCS, but gradually decreased with increasing temperature and time, except for CLA and PAA. Responses of PAA and ERY were only affected by temperature – in all three plots, the temperature regions up to 65˚C show the best reaction yield (Figure SI-I).

The compounds with multiple -OH groups (ERY) and -OH and -COOH groups (SA, SHA, QA, and CA) are higher in polarity, and their TMS derivatives are unstable and strongly fragmenting [22]. Similar silylation efficiencies were observed at multiple conditions, but the best response was achieved with BSTFA + 1% TMCS at 70°C/30 min for AA, CA, ERY and SHA [23]. However, responses of multiple TMS derivatives of SA (SA-TMS and SA-2TMS) were observed under all investigated conditions, either due to incomplete derivatization or degradation. CEC

353 Catalysts such as TMCS promote partial enolization of the keto group of estrogen hormones (E1, E2, E3, EE2),
354 with the consecutive formation of enol-TMS ethers [5]; however, such an effect was not observed for E1. The other
355 sterically hindered -OH groups required higher temperature and a longer derivatization time, such as the C-17α-OH
356 group, to which access is hindered by the C_{19}-C_{20} triple bond (Table SI-1). For EE2, under many set-ups, the
357 formation of both E1-TMS and EE2-TMS was observed, as previously reported [24,25]. Here, derivatization with
358 BSTFA + 1% TMCS (70°C/30 min) resulted in 38-42% of EE2 silylated to E1-TMS and 34% to EE2-TMS. Under the
359 most vigorous investigated conditions (80-90˚C/90 min), only EE2-TMS was generated, which is in line with a
360 previous study using BSTFA + 1% TMCS and EtAc [26]. The formation of mono- and di-TMS-EE2 with BSTFA and
361 EtAc was also reported [27]. The formation of 16-epiandrosterone as a degradation product of E3 was observed
362 when using MSTFA. Another study also suggests higher responses of E1-TMS, E2-TMS, EE2-TMS and DF-TMS
363 with MSTFA, following the order: 60°C > 70°C > 80°C >>22°C [28], which we observed only for E2-TMS. Although
364 optimal derivatization yield was obtained for estrogen hormones with BSTFA + 1% TMCS at 70°C, better than at
365 30°C or 80°C [29], others reported no difference between silylation at 60ºC and 90ºC and between 30 and 120 min
366 [26]. In this study, we selected BSTFA + 1% TMCS at 70˚C/90min as the optimum derivatization protocol for
367 estrogen hormones.
368 Replacement of both protons in primary amines requires more vigorous conditions, such as BSTFA + 1% TMCS,
369 60-80ºC/0.5-2.0 h [6]. 40% (*v/v*) BSTFA/MTBSTFA at 80°C/2h failed to derivatize aliphatic -OH and -NH_2 groups.
370 However, MSTFA successfully derivatized AMP, MAMP, BZECG, COD, MORPH, Δ^9-THC and 11N9THCA under
371 much milder conditions (20% *v/v*, 60°C/1h) [10]. SFA-TMS and CBZ-TMS follow a pattern similar to that observed for
372 QA, SHA, PAA and ERY, with a significant decrease in response for CBZ under extreme conditions. According to the
373 results, three silylation protocols were established:
374 **Protocol 1:** CEC silylated with BSTFA + 1% TMCS at 60°C for 45 min: BA, MePb, SA, EtPb, PrPb, IprPb, IB, MEC,
375 BuPb, IBuPb, SHA, QA, TCS, BzPb and CBD
376 **Protocol 2:** CEC silylated with BSTFA + 1% TMCS at 70°C for 90 min: RES, CLA, 9-HF, HPP, 4-NP, NAP, DH-BP,
377 4,4'-BP, KET, SFA, BP8, CBZ, DF, CBC, Δ^9-THC, CBN, THCA, E1, E2, EE2 and E3
378 **Protocol 3:** CEC silylated with BSTFA + 1% TMCS at 70°C for 45 min: PAA, ERY, AA, 8-HQ, 4-OP, CA, T3HC, CLP,
379 22BPF, BPAF, H-BP, 24BPF, DHDPE, BPF, BPE, BPA, BPC, BPB, BZECG, TMBA, BPCL, COD, MORPH, BPZ,
380 6MAM, 11OHTHC, BPS, BPAP, 11N9THC, BPM, BPP, BPBP, BPPH and BPFL.

381 **3.3 Method performance evaluation**

382 The matrix-matched method performance evaluation was studied in the linear range (5 – 1000 µg/L) for all
383 CEC. First, the assumption of homogeneity of variances was tested using Levene's test for each CEC-TMS at three

384 levels: low (1st calibration point, LOQ; n=3), medium (6th calibration point, 100 µg/L; n=5) and high (10th calibration
385 point, 1000 µg/L, n=5). Accordingly, two calibration curves were constructed for each CEC-TMS: low level (5-100
386 µg/L) and high level (100 – 1000 µg/L). The method performance parameters for the investigated compounds are
387 given in Supplementary material (SI-IV, Table SI-V and SI-VI).

388 Linearity (coefficient of determination, R^2) ranged from 0.99-1.00 in solvent and 0.99 to 1.00 in AWW
389 extracts for all compounds at both concentration levels. The method accuracy, expressed as [(experimental
390 value/spiked value) * 100] (n=3, at the 3rd, 6th and 10th calibration point), was satisfactory at low (88.68-107.92%),
391 medium (97.81-111.33%) and high level (95.14-103.58%) in solvent and at low (93.55 – 106.72%), medium (97.77 -
392 106.22%) and high level (95.95 - 102.19%) in AWW extracts. The instrumental repeatability expressed as the RSD of
393 three consecutive injections of the same sample at the 6th and 10th calibration points were ≤1.43% (low level) and
394 ≤1.27% (high level) in solvent and ≤1.74% (low level) and 0.02-0.53% (high level) in AWW extracts. Method
395 repeatability (RSD of three replicates at the 3rd, 6th, and 10th calibration points was 0.03-5.75% (low level), ≤1.58%
396 (medium level) and ≤0.59% (high level) in solvent and ≤8.99% (low level), 0.02-0.73% (medium level) and ≤0.18%
397 (high level) in the AWW extracts. Method's sensitivity, which is expressed as the slope of the calibration curves, was
398 also considered satisfactory at both levels for solvent, i.e., ≤0.19 (low) and ≤0.24 (high) and for the AWW extracts,
399 i.e., ≤0.06 (low) and ≤0.07 (high). The MDL and MQL were expressed as 3- and 10-times SD of the lowest calibration
400 point at which satisfactory precision (<15%) and accuracy (90-110%) were achieved, divided by the slope of the
401 calibration curve. In solvent, MDL ranged from 0.01 to 0.52 µg/L and MQL from 0.01 to 1.73 µg/L. The MDL in AWW
402 extracts ranged from 0.32 µg/L and MQL up to 1.06 µg/L.

403 **3.4 Stability of CEC-TMS derivatives**

404 We observed that SHA-TMS, QA-TMS and SFA-TMS did not show sufficiently linear response and were excluded
405 from the stability experiments. In solvent, ERY-TMS increased up to 859.60% of the initial concentration at 25°C,
406 1104.56% at 4°C and 1238.73% at -18°C and therefore was also excluded. The most probable reason for these
407 observed irregularities is that the parent compounds, such as organic acids (SHA, QA) and sugars (ERY), form
408 unstable poly-TMS derivatives that are subjected to intensive fragmentation, which results in numerous nonspecific
409 *m/z* fragments in their MS spectra [22]. Also, compounds with sulphonamide and amine functional groups are the
410 least stable TMS derivatives [30], and the formation of mono-TMS derivatives, such as SFA-TMS, was observed only
411 during the initial optimization experiments but, due to poor peak shape, could not be quantified.

3.4.1 Short-term and long-term stability

Over 28 days, most of the CEC-TMS derivatives remained stable in solvent at 25°C (FigureSI-IIIA), at 4°C (Figure IIIB) and -18°C (Figure IIIC). The highest overall variability was observed at 25°C (Figure 3A), where 64 of the 67 CEC-TMS derivatives remained stable, i.e. within 85-115% of the initial concentration. Only 8HQ-TMS and BPZ-2TMS varied by up to ~140% and ~185% on days 3-6 but remained stable (-4.03 to +11.13% and -2.70 to +13.11%) at other time points. For 8HQ-TMS, a significant difference was observed between the samples and QC only at HL (p_L=0.736, p_H=0.024), while for BPZ-2TMS, no significant differences were observed (p_L= 0.726, p_H=0.071). At individual time points for days 1-7 BA-TMS (-27.78%), MePb-TMS (-28.20%), MEC-TMS (-20.34%), iBuPb-TMS (-17.45%), BzPb-TMS (-25.89%), KET-TMS (-21.64%), PAA-TMS (+24.34%), 11N9THC-TMS (-26.83%), E1-TMS (+26.12%) and E2-TMS (-29.78%) deviated from the acceptable stability range (±15% of the initial concentration). However, these observed deviations did not significantly impact their overall stability profiles. Finally, degradation was not observed for any of the 67 CEC-TMS derivatives tested.

At 4°C (Figure SI-IIIB), most (64/67) of the CEC-TMS derivatives remained stable over the 28 days. The exceptions were SHA-TMS, QA-TMSA and SFA-TMS. Here, CA-TMS had a low starting concentration, which significantly varied over time (~70.93 - 180.00%), while the concentration of ERY-TMS continuously increased (data not shown). Fewer deviations of the individual time point at days 1-7 beyond the lower (MePb-TMS: -29.26% to -17.20%, EtPb-TMS: -18.36% to -15.85%, IPrPb-TMS:-17.70%, IB-TMS: -21.10%; PrPb-TMS: -28.40% to -17.14%; 4,4'-BP-TMS: -28.29% and KET-TMS: -20.57%) or the upper (8HQ-TMS: 20.41-38.97% and BPZ-2TMS: 26.90-33.96%) limit of the established stability range (85-115%) were observed compared to the results at 25˚C.

At -18˚C (FigureSI-IIIC), 66 of the 67 CEC-TMS derivatives were stable in the solvent during short- and long-term experiments. Only ERY-TMS exhibited an extreme increase in concentration, ≤486.23% and 627.44% at LL and HL, respectively. Also, although some individual time points were out of the stability range for certain CEC-TMS derivatives (DH-BP, Δ^9-THCA, E2, 8-HQ, CA, BPAP and 11N9THC), it did not influence the overall stability profiles.

In AWW extracts, at 25˚C and 4˚C, E2-TMS and CA-TMS exhibited extreme concentration increases (data omitted from Figure SI-IVA and IVB), with >200% for E2-TMS at LL and HL, potentially due to matrix effect, and for CA at LL at day 7. EE2-TMS and E3-3TMS exhibited degradation at both temperatures, although only at the LL for E3-3TMS. Here, approx. 80% of CEC remained on day 28 at 25˚C and 4˚C. EE2-TMS degraded continuously during the 28 days; by day 28, 60% and 35% remained (25˚C and 29% and 7% at 4˚C) at the HL and LL, respectively. Here, the degradation of EE2-TMS was not accompanied by a notable increase in the amount of E1-TMS.

Long-term stability in a solvent at 4˚C (Figure SI-IIIB) and -18˚C (Figure SI-IIIC) showed far fewer deviations outside the established stability range (85-115%). The only remarkable deviations were low concentrations of 10

443 CEC-TMS derivatives (of CLA, E3, 22BPF, 24BPF, BPC, BPB, BZECG, MORPH, BPFL and COD) at week 16 at 4˚C,
444 without significantly affecting the overall long-term stability profile. Concentrations of ERY-TMS and CA-TMS
445 continued increasing long-term due to continuous formation of TMS derivatives, for ERY-TMS at both temperatures
446 and concentration levels, while for CA-TMS only at LL, at 4˚C. Degradation was observed only for BPBP-2TMS at
447 4˚C, starting at week 3; by week 20, 32.86% and 30.69% of the initial concentration had been degraded. Such
448 degradation, however, was not observed at -18˚C. Additionally, no deviations and instabilities were observed in AWW
449 extracts during storage for up to 20 weeks at 4˚C and -18˚C (Figure SI-IV).

450 The reasons for the observed variations in CEC-TMS derivatives observed within and out of the stability range
451 (85-115%) are likely the result of **(1)** the constant and continuous formation and degradation of TMS derivatives due
452 to the reversible nature of the silylation reaction; **(2)** incomplete compound derivatization (<100%); **(3)** significant
453 variation in peak areas due to the presence of an additional amount of TMS donor (BSTFA + 1% TMCS) originating
454 from the IS solution when spiked in the samples [19] and **(4)** the presence of humidity in the headspace above
455 sample in the vial, which during consequent heating and cooling down, condensates and “utilizes” the silylation agent.
456 Where degradation is observed, the most likely effect is hydrolysis due to the lack of entirely anhydrous conditions,
457 which are difficult to achieve during the analysis of environmental samples. Such samples have a significant matrix
458 burden, which traps water and can further promote hydrolysis of the TMS derivatives.

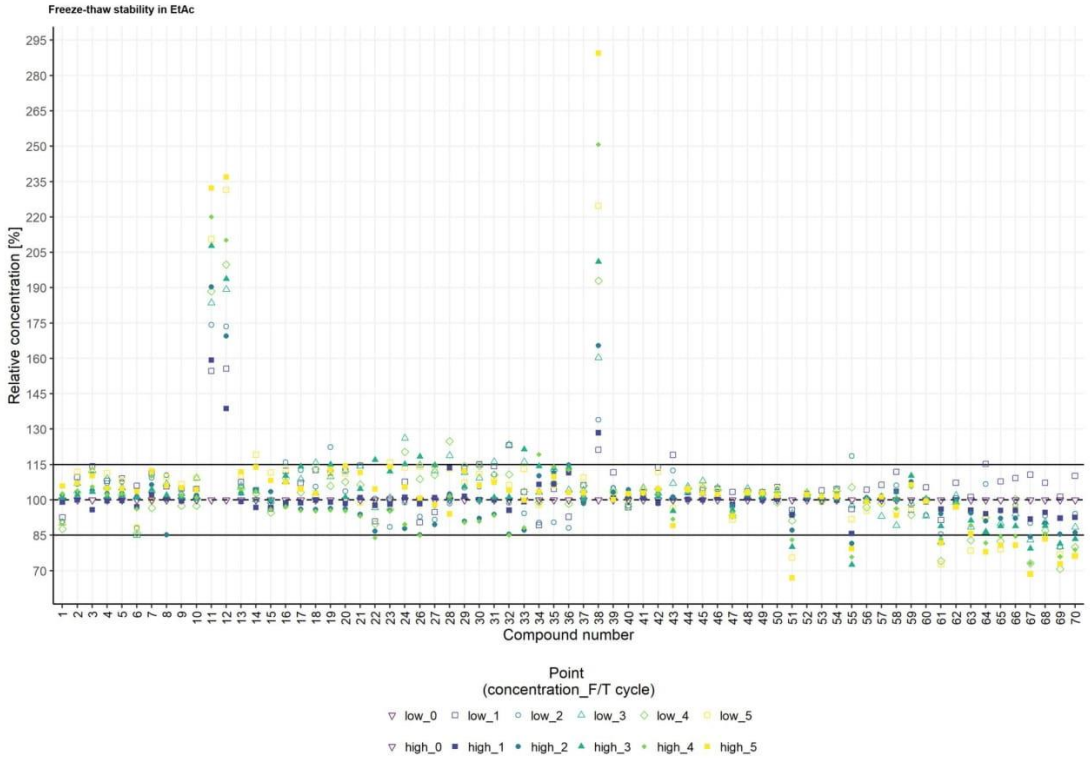
459 The limited number of studies confirm the stability of TMS derivatives of Δ^9-THCA, COD, BZECG and
460 11N9THCA at -18°C in EtAc for 7 days [10], and TMS derivatives of E1, E2, EE2 and E3 in BSTFA and pyridine for 2
461 days [29]**.** Also, TMS derivatives of 4-NP, E1, E2, EE2, 4-OP, and BPA in *n*-hexane were stable for up to 4 days at
462 25˚C after removing the BSTFA and pyridine [31], while DH-BP and BP-8 TMS derivatives were stable for 6 h at 4°C
463 [32] and up to 1 month at 4°C (RSD ≤ 4%), respectively [9]. The TMS derivatives of parabens ((MePb, EtPb, IPrPb,
464 PrPb, IBuPb, BuPb, BzPb), 4-NP and 4-OP were stable for up to 2 days at 25˚C and at least 1 week at 4°C (RSD
465 <5%) [8]. A gradual decrease during 48h of storage at 25˚C of COD-TMS and MORPH-TMS was observed in
466 methylene chloride after derivatization with BSTFA + 1% TMCS [33]. Finally, E1, E2, EE2, 4-NP, 4-OP and BPA from
467 AWW extracts were most stable for 48h in pyridine and hexane after silylation with BSTFA + 1% TMCS at 70°C/30
468 min, with or without pyridine. Beyond 48h, their stability was more variable and less satisfactory [25].

469

470 ### 3.4.2 Freeze-thaw stability

471 During the five F/T cycles in solvent (Figure 3) and AWW extracts (Figure 5), concentrations of TMS
472 derivatives of SHA, QA and ERY increased to 200-300% at both levels, and all of the CEC-TMS derivatives were
473 stable.

The sample/QC response ratio of TMS derivatives of SHA, QA and ERY increased as F/T cycles increased, indicating that changes occurred during freezing and thawing. From the remaining 67 CEC-TMS derivatives, nine:



BPE-2TMS, BZECG-TMS, 6- MAM-TMS, BPS-2TMS, 11N9THC-2TMS, BPM-2TMS, BPP-2TMS, BPPH-2TMS and BPFL-2TMS showed significant degradation, with ~66.90-81.60% remaining after the fifth cycle. Degradation was also more prominent at HL (Figure 4), indicating its potentially concentration-dependent nature. For most sensitive CEC-TMS at LL – BPPH-2TMS, BPP-2TMS and 6MAM-TMS, >85% of initial concentration remains after two F/T cycles. In addition, 65-80% of the initial concentration remains of BPPH-2TMS, BPFL-2TMS, BPP-2TMS, BPM-2TMS, BPE-2TMS and 6MAM-TMS after the fifth F/T cycle. At HL, <85% of BZECG-TMS, BPP-2TMS, BPPH-2TMS, BPE-2TMS and BPFL-2TMS remain after three F/T cycles, while after the five F/T cycles, 65-85% of the initial nine TMS derivatives remain.

Figure 3. F/T stability of CEC-TMS in the solvent at LL (100 µg/L) and HL (1000 µg/L). TMS derivatives are numbered 1- 70 according to Table SI-I in Supplementary material.

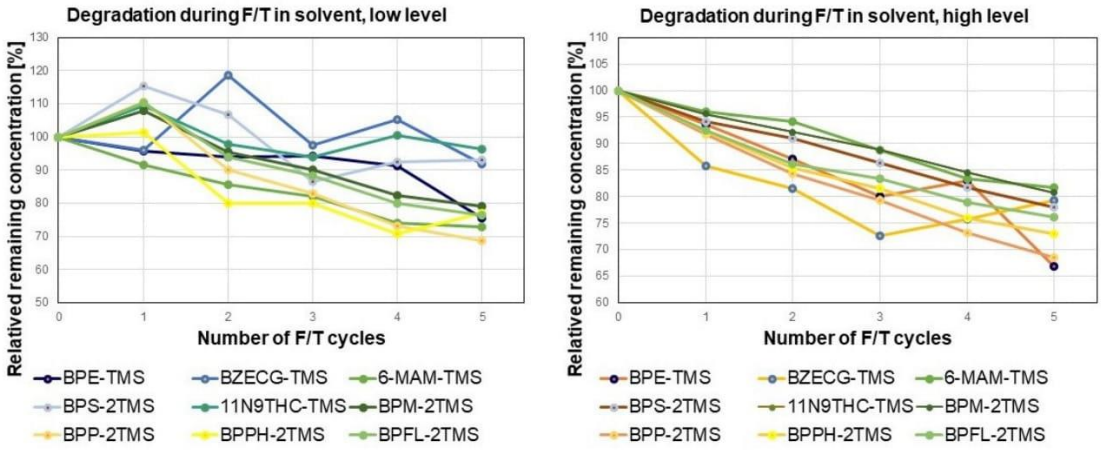
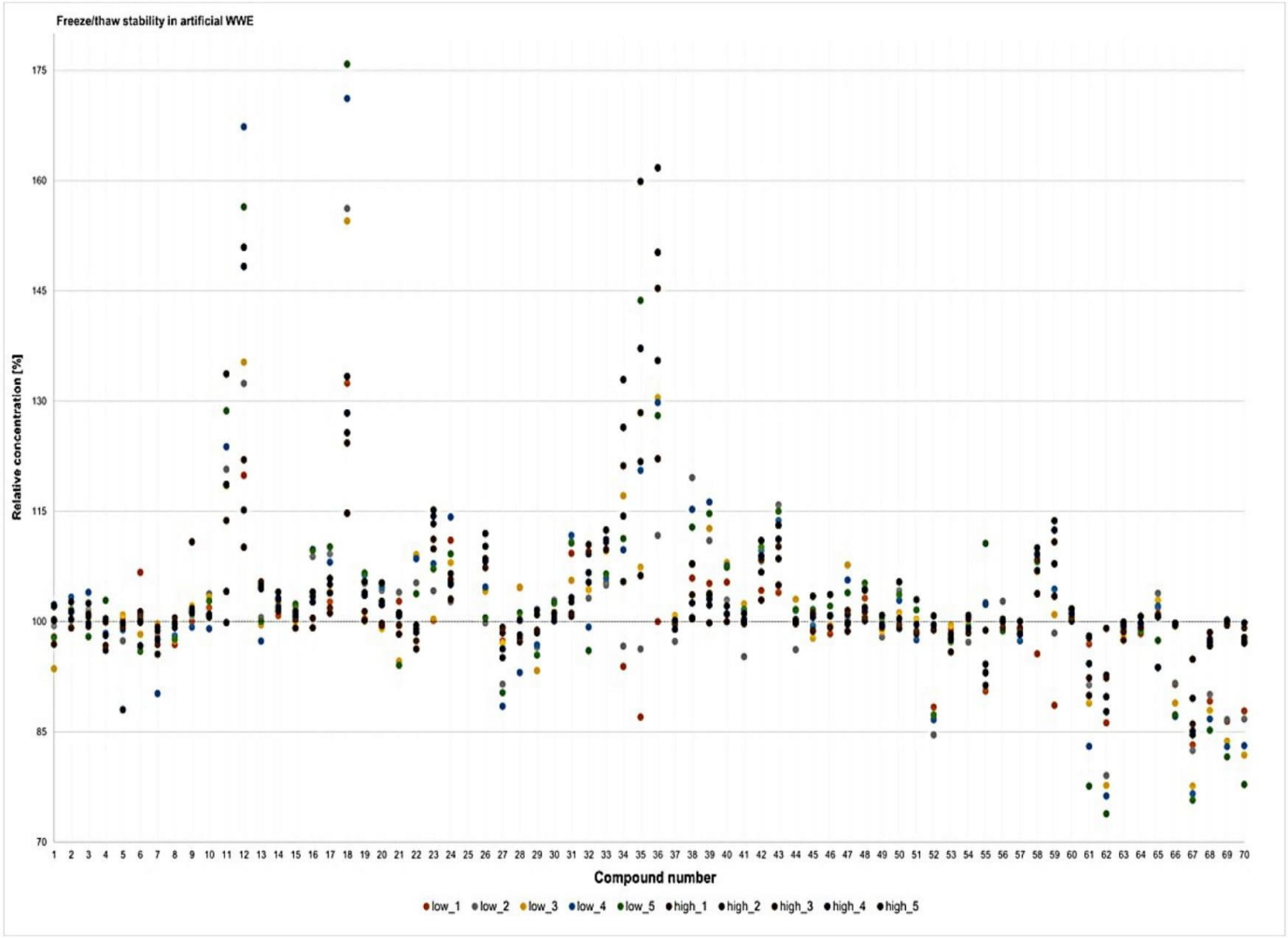
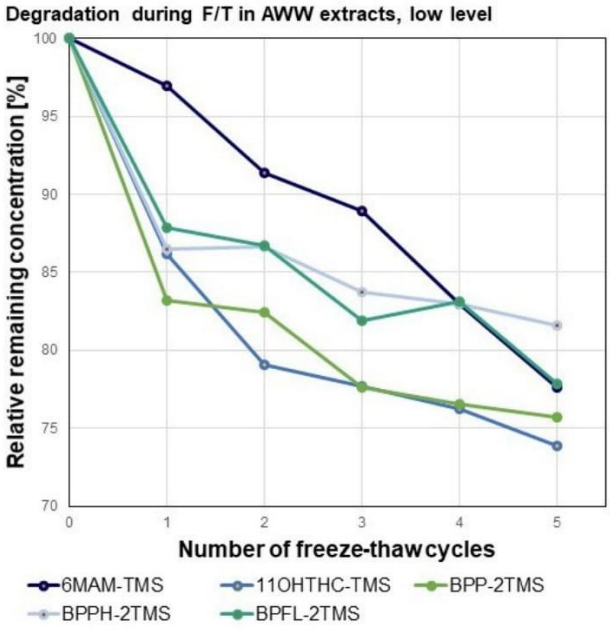


Figure 4. Degradation of unstable CEC-TMS derivatives during F/T in the solvent; A) LL; B) HL.

In AWW extract (Figure 5), CEC-TMS of SHA, QA, 9-HF, E2, EE2 and E3 showed deviations at high concentrations after the fifth F/T cycle but lower when compared to freezing and thawing in solvent. Here, the sample/QC response ratios for SHA-TMS, 9HF-TMS, E2-TMS and EE2-TMS increase (the latter two only at HL), but QA-TMS and E3-TMS remain within the ±15% range. Therefore, for the first four CEC-TMS, changes occur exclusively due to freezing and thawing, while for QA-TMS and E3-TMS, identical changes also occur in the freshly prepared QC samples, indicating that their concentration also decreases. From the remaining 64 CEC-TMS, five (11OHTHC-TMS, BPP-2TMS, BPPH-2TMS and BPFL-2TMS) exhibited degradation only at the LL, with <85% remaining after the 3rd F/T cycle and together with 6MAM-TMS, 73.86-81.60% remaining after the 5th F/T cycle (Figure 6).



498 **Figure 5.** F/T stability of CEC-TMS in AWW extracts. CEC-TMS derivatives are numbered 1- 70 according to Table SI-I in Supplementary Material.



499 **Figure 6.** Degradation of unstable CEC-TMS derivatives during F/T in AWW extracts, LL.

500 **3.4.3 Statistical analysis**

501 A comparison of the stability profiles of the investigated CEC-TMS derivatives to the responses of the QC
502 samples in solvent (SI-VI, Table SI-VII) showed that 22.4% and 25.6% of the TMS derivatives had statistically
503 significant differences in their stability profiles at LL and HL, respectively ($p<0.05$), during 20 weeks at 25˚C, of which
504 three CEC-TMS at both levels. At 4˚C, 44% of the CEC-TMS derivatives at each level, including 14 CEC-TMS at both
505 levels, had statistically significant differences in stability profiles. Slightly fewer TMS derivatives (22.4% and 35.8%)
506 exhibited statistically significant differences at -18˚C. Interestingly, the TMS derivatives of parabens differed from the
507 QC samples under all investigated conditions, indicating that although changes occur during storage, they do not
508 influence their general stability profile. In AWW extracts (Table SI-VIII), the findings were similar, with 71.60% and
509 46.30% of the TMS derivatives showing a statistically significant difference between the samples and QC samples at
510 LL and HL at 25˚C and 64.20% and 50.80% at LL and HL at 4˚C, respectively; while 21 and 22 of TMS derivatives
511 demonstrated stability profiles differing significantly from the QC responses at both levels.

512 Inter-level comparison at the same temperature and matrix (Table SI-IX) showed statistically significant
513 differences in stability profiles in solvent for most CEC-TMS derivatives in descending order at -18˚C, 4˚C and 25˚C, 4
514 of which 4,4'-BP-TMS, T3HC-TMS, TMBA-TMS and MORPH-TMS at all investigated conditions. In AWW extracts,
515 significant differences are observed between the two levels for more CEC-TMS derivatives at 4˚C than at 25˚C, eight

of which (TMS derivatives of IBuPb, BuPb, BzPb, 4OP, CA, CLP, BPAF and BPC) at both 25˚C and 4˚C. The stability profiles of CLP-TMS and TMS derivatives of 4,4'-BP, CA, BPAF, BPZ and BPS differed at HL and LL at 25˚C and 4˚C in both matrices. Generally, the stability of the CEC-TMS derivatives is more concentration-dependent at lower temperatures when compared to room temperature, independent of the matrix.

An inter-matrix comparison at 4˚C and 25˚C (Table SI-X) shows more significant differences in stability profiles in solvent and AWW extracts at LL than at HL. For 11 CEC-TMS derivatives, statistically significant difference was observed at LL. At both levels, differences were observed for THCA-TMS, PAA-TMS and BPBP-TMS at 25˚C and CBZ-TMS, THCA-TMS, 8HQ-TMS and PAA-TMS at 4˚C. In AWW extracts, differences were observed at 4˚C and 25˚C only for 9-HF-TMS, 4,4;-BP-TMS and CBN-TMS at LL and CBN-TMS and EE2-TMS at HL. More differences were observed when comparing stability profiles in the solvent at -18˚C and 4˚C, while improved stability was observed at -18˚C for 22 and 18 CEC-TMS derivatives, respectively.

The 67 CEC-TMS derivatives were also ranked using the DSC approach according to their overall stability in a solvent, AWW extract and both matrices (Table SI-XII, Figure SI-V), excluding SHA-TMS, QA-TMS and SFA-TMS, as previously stated (see 3.4). The five CEC-TMS derivatives that showed the most prominent deviation at LL, i.e. highest degradation, are, from highest to lowest: E2-TMS > CA-TMS > 9-HF-TMS > E1-TMS > T3HC-TMS in the solvent; E2-TMS > CA-TMS > 9-HF-TMS > E3-TMS > E1-TMS in AWW extracts and ERY-TMS > E2-TMS > CA-TMS > 9-HF-TMS > E1-TMS in both matrices. These are the TMS derivatives of estrogen hormones (E1, E2, E3) and polyhydroxy CEC (CA, ERY), but also 9-HF which has the $Si(CH_3)_3$ group attached to the -OH group of the bridging carbon between two benzene rings. 9-HF is known to oxidize to 9-fluorenone (ketone) reversibly; therefore, 9HF-TMS hydrolysis and conversion of 9HF to 9-fluorenone is the probable reaction explaining the observed instability. The five CEC-TMS derivatives that showed the most prominent deviation at HL are 6-MAM-TMS, 11N9THC-TMS, DF-TMS, 11OHTHC-TMS and EE2-TMS in solvent and DF-TMS, 6-MAM-TMS, EE2-TMS, BPP-TMS and 11OHTHC-TMS in AWW extracts. Both matrices were TMS derivatives of 6-MAM, 11OH-THC, 11N9THC, BZECG and DF. However, deviations in their stability profiles are unlikely due to additional formation of TMS derivatives during storage, as evidence of uncomplete derivatization was not observed during optimization experiments. Therefore, TMS derivatives of the natural estrogen hormones (E1, E2, EE2 and E3) and poly-hydroxy CEC (ERY and CA) are most prone to degradation during short-term storage and heating in solvent and AWW extracts.

3.4.4 Estimation of MU

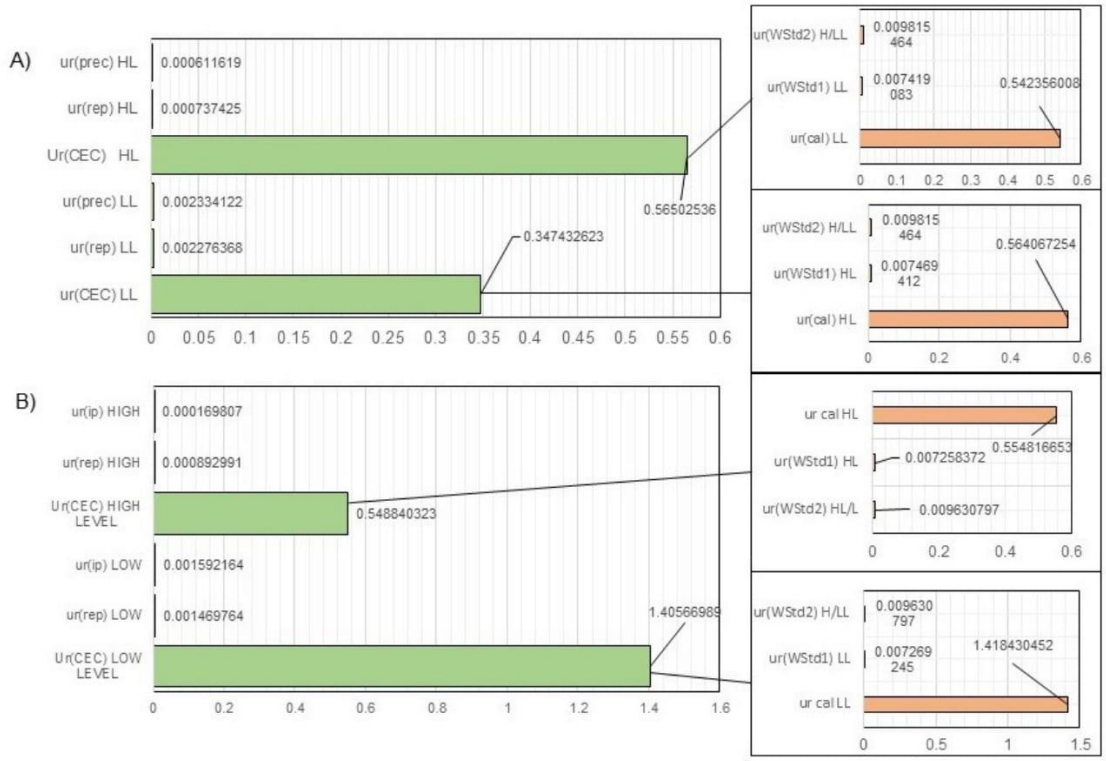
The MU was evaluated using the bottom-up approach to investigate whether the deviations of CEC-TMS derivatives observed during the stability experiments can be attributed to the associated MU. Here, the combined standard uncertainty uc(y) associated with a result (y) is determined from the estimated standard deviation

associated with each input xi, the standard uncertainty u(xi). The relative combined standard uncertainty, $u_c(y)$, was calculated as:

$$uc(y)=\sqrt[2]{\zeta\zeta} \quad (1)$$

by combining the contributions of repeatability (u_{rep}) and instrument precision (u_{ip}), and the amount of CEC in the sample ($u_{c(CEC)}$). The latter combines individual contributions of the r_s associated with the working standard solution of CEC (ur(WStd1)), r_s associated with the working solution of IS (ur(WStd2)) and the r_s associated with the calibration curve (ur(cal)). Further, ur(WStd1) includes the contributions of the r_s associated with the stock standard solution (ur(SStd1)), r_s associated with the volume (ur(V)) and r_s associated with the use of pipettes (u(pipette)). The purity and the expanded uncertainties of the reference standards of CEC and IS, the expanded uncertainties of the analytical balance, volumetric flasks of 10 mL, 25 mL and 100 mL and pipettes of 10 µL and 200 µL were based on manufacturers' certificates of analyses and certificates of calibration, respectively. Supplementary material (SI-VI) provides the equations employed to calculate each contribution value and further details of the individual contributions. For the two concentration levels, $u_c(y) = y\ u_c(y)$ and the expanded uncertainty, $U(y) = k\ u_c(y)$, were calculated for three determinations per sample. Expanded uncertainty was calculated using a coverage factor k=1.96, as the effective degrees of freedom resulted in $v_{eff} > 10$, providing a level of confidence of approximately 95%.

The results from MU estimation for solvent and AWW extract stability studies are summarized in Supplementary material (Table SI-IX and Table SI-X,) and represented in Figure 7. In solvent, expanded uncertainty was highest for ERY-TMS (L:2.40% H: 19.94%), E3-3TMS (L:4.16,% H:12.50%) and CA-TMS (L: 0.48%, H:1.97%), while higher results were obtained in AWW extracts, for CLA-TMS (L:2.43%, H:2.68%), PAA-TMS (L:4.12%, H:1.29%), MORPH-2TMS (L:4.70%, H:2.93%), ERY-TMS (L:7.61%, H:15.45%), BA-TMS (L:22.95%, H:14.81%) and CA-TMS (L:106.73%, H:0.83%). For solvent (Figure 7A) and AWW extract (Figure 7B) stability, the CEC-TMS concentration contributes most significantly to the overall combined standard uncertainty. Within u(CEC), the uncertainty of the calibration curve had the highest contribution to the uc(y) for all CEC-TMS, in both matrices and at both concentration levels. However, the concentration deviations observed at some sampling points for some CEC-TMS derivatives were not within the estimated MU, indicating that such deviations cannot be attributed to MU.



572

573 **Figure 7.** Histogram showing the average contribution of the uncertainty components. A) solvent stability, B) AWW extracts stability.

574 ## 4. Conclusions

575 In this study, we established three optimal silylation protocols for 70 CEC using BSTFA + 1% TMCS as a
576 silylation agent, which are (1) 60˚C/30 min, (2) 70˚C/45 min and (3) 70˚C/90 min. The stability of the generated CEC-
577 TMS derivatives was investigated under relevant storage conditions and F/T stability via ten different experiments.
578 The majority of CEC-TMS were stable in solvent at 25˚C, 4˚C and -18˚C for 28 days, such that instabilities were
579 observed for 8HQ-TMS, ERY-TMS and BPZ-TMS at 25˚C, for ERY-TMS and CA-TMS at 4˚C and ERY-TMS at 18˚C.
580 During 20 weeks, most CEC-TMS remained stable at 4˚C and -18˚C, such that only ERY-TMS, CA-TMS and BPBP-
581 2TMS exhibited stability issues at 4˚C and ERY-TMS at -18˚C. In AWW extracts, 63 of the 67 CEC-TMS derivatives
582 were stable at 25˚C and 4˚C for 28 days. E2-TMS and CA-TMS exhibited unexpected concentration increases, while
583 EE2-TMS and E3-3TMS were significantly degraded at both temperatures.

584 Overall, stability should be considered when storing GC samples containing TMS derivatives of 8HQ, ERY,
585 BPZ and CA TMS derivatives for up to 28 days, and ERY-TMS, CA-TMS and BPBP-TMS at 4˚C and ERY-TMS at -
586 18˚C, for up to 20 weeks in a solvent. The same is true when considering E2, CA, EE2 and E3 TMS derivatives in

587 AWW samples stored at 4˚C and -18˚C. Based on the overall stability ranking, we recommend storing samples tested
588 for the presence of a chemodiverse pool of CEC using trimethylsilylation and GC-MS at -18˚C for up to 20 weeks to
589 ensure the stability of their TMS derivatives.

590 During sample freezing and thawing, nine and six CEC-TMS derivatives were degraded during five F/T
591 cycles in solvent and AWW extracts, respectively. The most sensitive CEC-TMS derivatives to freezing and thawing
592 were BPPH-2TMS, 6MAM-2TMS and BPP-2TMS at LL and BZECG-TMS, BPP-2TMS, BPPH-2TMS, BPE-2TMS and
593 BPFL-2TMS at HL in a solvent. A maximum of two F/T cycles in solvent are allowed to maintain ≥80% of the initial
594 CEC-TMS concentration. The same recommendation is valid for AWW extracts containing 6-MAM-TMS, 11OHTHC-
595 TMS, BPP-TMS, BPPH-TMS, and BPFL-TMS to maintain ≥85% of the initial TMS derivatized compounds. MU
596 estimation showed that uncertainty of calibration accuracy is the most significant source of uncertainty. However,
597 concentration changes during stability experiments could not be attributed solely to MU for any investigated CEC-
598 TMS derivatives.

599 This study provides valuable findings about the effect of the silylation conditions on silylation efficiency and it
600 significantly contributes to a better understanding of the generation and stability of TMS derivatives of CEC during
601 GC-MS multi-residue analyses. The stability of CEC-TMS should also be examined in other complex matrices to
602 ensure their reliable and robust quantification. In addition, the stability of other, less frequently used derivatives that,
603 together with TMS derivatives, broaden the CEC coverage should also be investigated.

604

605 **Author Statement**

606 **Milka Ljoncheva:** Investigation, Formal analysis, Data Curation, Visualization, Writing-Original Draft; **Tina Kosjek:**
607 Conceptualization, Methodology, Writing-Review and Editing, Supervision, Methodology validation, Funding
608 acquisition; **Sašo Džeroski:** Funding acquisition, Writing-Review and Editing; **Ester Heath**: Methodology validation,
609 Writing-Review and Editing; **David Heath:** Writing-Review and Editing.

610

611 **Acknowledgements**

612 This work was supported by the Slovenian Research Agency (J1-6744, Program Groups P1-0143 and P2-0103). M.L.
613 is funded by the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia
614 (Contract No. 11011-85/2016). The authors would like to thank Dr Tome Eftimov for the help with the implementation
615 of the DSC approach.

616 **Conflict of interest**

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this manuscript includes SI-I to SI-VI and Table SI-I to Table SI-XIV.

References:

[1] M. Pourchet, L. Debrauwer, J. Klanova, E.J. Price, A. Covaci, N. Caballero-Casero, H. Oberacher, M. Lamoree, A. Damont, F. Fenaille, J. Vlaanderen, J. Meijer, M. Krauss, D. Sarigiannis, R. Barouki, B. Le Bizec, J.-P. Antignac, Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: From promises to challenges and harmonisation issues, Environment International. 139 (2020) 105545. https://doi.org/10.1016/j.envint.2020.105545.

[2] B. Gruber, F. David, P. Sandra, Capillary gas chromatography-mass spectrometry: Current trends and perspectives, TrAC Trends in Analytical Chemistry. 124 (2020) 115475. https://doi.org/10.1016/j.trac.2019.04.007.

[3] F.J. Santos, M.T. Galceran, Modern developments in gas chromatography–mass spectrometry-based environmental analysis, Journal of Chromatography A. 1000 (2003) 125–151. https://doi.org/10.1016/S0021-9673(03)00305-4.

[4] B.P. Gumbi, B. Moodley, G. Birungi, P.G. Ndungu, Target, Suspect and Non-Target Screening of Silylated Derivatives of Polar Compounds Based on Single Ion Monitoring GC-MS, IJERPH. 16 (2019) 4022. https://doi.org/10.3390/ijerph16204022.

[5] C.F. Poole, Trialkylsilyl derivatives (other than TMS) for Gas Chromatography and Mass Spectrometry, Journal of Chromatography A. 17 (1979) 115–123. https://doi.org/10.1016/j.chroma.2013.01.097.

[6] J.M. Halket, V.G. Zaikin, Derivatization in Mass Spectrometry—1. Silylation, Eur J Mass Spectrom (Chichester). 9 (2003) 1–21. https://doi.org/10.1255/ejms.527.

[7] F. Orata, Derivatization Reactions and Reagents for Gas Chromatography Analysis, in: M. Ali Mohd (Ed.), Advanced Gas Chromatography - Progress in Agricultural, Biomedical and Industrial Applications, InTech, 2012. https://doi.org/10.5772/33098.

[8] R.A. Pérez, B. Albero, E. Miguel, C. Sánchez-Brunete, Determination of parabens and endocrine-disrupting alkylphenols in soil by gas chromatography–mass spectrometry following matrix solid-phase dispersion or in-column microwave-assisted extraction: a comparative study, Anal Bioanal Chem. 402 (2012) 2347–2357. https://doi.org/10.1007/s00216-011-5248-0.

[9] C. Sánchez-Brunete, E. Miguel, B. Albero, J.L. Tadeo, Analysis of salicylate and benzophenone-type UV filters in soils and sediments by simultaneous extraction cleanup and gas chromatography–mass spectrometry, Journal of Chromatography A. 1218 (2011) 4291–4298. https://doi.org/10.1016/j.chroma.2011.05.030.

[10] I. González-Mariño, J.B. Quintana, I. Rodríguez, R. Cela, Determination of drugs of abuse in water by solid-phase extraction, derivatisation and gas chromatography–ion trap-tandem mass spectrometry, Journal of Chromatography A. 1217 (2010) 1748–1760. https://doi.org/10.1016/j.chroma.2010.01.046.

[11] D.-L. Lin, S.-M. Wang, C.-H. Wu, B.-G. Chen, R.H. Liu, Chemical derivatization for the analysis of drugs by GC-MS - A conceptual review, Journal of Food and Drug Analysis. 16 (2008) 1–10. https://doi.org/10.38212/2224-6614.2373.

[12] Minitab Statistical Software v19, (2020). https://www.minitab.com/en-us/ (accessed August 5, 2020).

[13] T. Metsalu, J. Vilo, ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap, Nucleic Acids Res. 43 (2015) W566–W570. https://doi.org/10.1093/nar/gkv468.

[14] R Core Team, R: A language and environment for statistical computing, (2020). https://www.R-project.org/.

[15] T. Eftimov, P. Korošec, B. Koroušić Seljak, Data-Driven Preference-Based Deep Statistical Ranking for Comparing Multi-objective Optimization Algorithms, in: P. Korošec, N. Melab, E.-G. Talbi (Eds.), Bioinspired Optimization Methods and Their Applications, Springer International Publishing, Cham, 2018: pp. 138–150. https://doi.org/10.1007/978-3-319-91641-5_12.

[16] Joint, Committee for Guides in Metrology (JCGM/WG 1)., Evaluation of measurement data-Guide to the expression of uncertainty in measurement, (2008). https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf (accessed September 28, 2020).

[17] H. Miyagawa, T. Bamba, Comparison of sequential derivatization with concurrent methods for GC/MS-based metabolomics, Journal of Bioscience and Bioengineering. 127 (2019) 160–168. https://doi.org/10.1016/j.jbiosc.2018.07.015.

[18] M. Phélippé, R. Coat, C. Le Bras, L. Perrochaud, E. Peyretaillade, D. Kucma, A. Arhaliass, G. Thouand, G. Cogne, O. Gonçalves, Characterization of an easy-to-use method for the routine analysis of the central metabolism using an

669 affordable low-resolution GC–MS system: application to Arthrospira platensis, Anal Bioanal Chem. 410 (2018) 1341–
670 1361. https://doi.org/10.1007/s00216-017-0776-x.
671 [19] Park, Sun Young, 김훈식, Lee, Jun-Gae, Hong, Jongki, Chemical Derivatization of Catecholamines for Gas
672 Chromatography-Mass Spectrometry, Bulletin of the Korean Chemical Society. 30 (2009) 1497–1504.
673 https://doi.org/10.5012/BKCS.2009.30.7.1497.
674 [20] A. Azzouz, E. Ballesteros, Trace analysis of endocrine disrupting compounds in environmental water samples by use of
675 solid-phase extraction and gas chromatography with mass spectrometry detection, Journal of Chromatography A. 1360
676 (2014) 248–257. https://doi.org/10.1016/j.chroma.2014.07.059.
677 [21] B. Molnár, I. Molnár-Perl, The role of alkylsilyl derivatization techniques in the analysis of illicit drugs by gas
678 chromatography, Microchemical Journal. 118 (2015) 101–109. https://doi.org/10.1016/j.microc.2014.08.014.
679 [22] C. Schummer, O. Delhomme, B. Appenzeller, R. Wennig, M. Millet, Comparison of MTBSTFA and BSTFA in derivatization
680 reactions of polar compounds prior to GC/MS analysis, Talanta. 77 (2009) 1473–1482.
681 https://doi.org/10.1016/j.talanta.2008.09.043.
682 [23] K.K. Pasikanti, P.C. Ho, E.C.Y. Chan, Development and validation of a gas chromatography/mass spectrometry
683 metabonomic platform for the global profiling of urinary metabolites, Rapid Commun. Mass Spectrom. 22 (2008) 2984–
684 2992. https://doi.org/10.1002/rcm.3699.
685 [24] A. Shareef, M.J. Angove, J.D. Wells, Optimization of silylation using N-methyl-N-(trimethylsilyl)-trifluoroacetamide, N,O-
686 bis-(trimethylsilyl)-trifluoroacetamide and N-(tert-butyldimethylsilyl)-N-methyltrifluoroacetamide for the determination
687 of the estrogens estrone and 17α-ethinylestradiol by gas chromatography–mass spectrometry, Journal of
688 Chromatography A. 1108 (2006) 121–128. https://doi.org/10.1016/j.chroma.2005.12.098.
689 [25] Z.L. Zhang, A. Hibberd, J.L. Zhou, Optimisation of derivatisation for the analysis of estrogenic compounds in water by
690 solid-phase extraction gas chromatography–mass spectrometry, Analytica Chimica Acta. 577 (2006) 52–61.
691 https://doi.org/10.1016/j.aca.2006.06.029.
692 [26] N. Migowska, M. Caban, P. Stepnowski, J. Kumirska, Simultaneous analysis of non-steroidal anti-inflammatory drugs and
693 estrogenic hormones in water and wastewater samples using gas chromatography–mass spectrometry and gas
694 chromatography with electron capture detection, Science of The Total Environment. 441 (2012) 77–88.
695 https://doi.org/10.1016/j.scitotenv.2012.09.043.
696 [27] Z. Yu, S. Peldszus, P.M. Huck, Optimizing gas chromatographic–mass spectrometric analysis of selected pharmaceuticals
697 and endocrine-disrupting substances in water using factorial experimental design, Journal of Chromatography A. 1148
698 (2007) 65–77. https://doi.org/10.1016/j.chroma.2007.02.047.
699 [28] M. Česen, E. Heath, Disk-based solid phase extraction for the determination of diclofenac and steroidal estrogens E1, E2
700 and EE2 listed in the WFD watch list by GC–MS, Science of The Total Environment. 590–591 (2017) 832–837.
701 https://doi.org/10.1016/j.scitotenv.2017.02.222.
702 [29] B. Huang, X.-J. Pan, J.-L. Liu, K. Fang, Y. Wang, J.-P. Gao, Hydroxyl Group Derivatization of Steroid Environmental
703 Endocrine Disrupting Chemicals, Chinese Journal of Analytical Chemistry. 37 (2009) 1651–1656.
704 https://doi.org/10.1016/S1872-2040(08)60145-0.
705 [30] I. Jardine, K. Blau and G. S. King (editors). Handbook of derivatives for chromatography. Heyden, London, 1977
706 (Reprinted with corrections 1978), 1978. http://doi.wiley.com/10.1002/bms.1200051011 (accessed July 30, 2020).
707 [31] A. Hibberd, K. Maskaoui, Z. Zhang, J. Zhou, An improved method for the simultaneous analysis of phenolic and steroidal
708 estrogens in water and sediment, Talanta. 77 (2009) 1315–1321. https://doi.org/10.1016/j.talanta.2008.09.006.
709 [32] I. Tarazona, A. Chisvert, Z. León, A. Salvador, Determination of hydroxylated benzophenone UV filters in sea water
710 samples by dispersive liquid–liquid microextraction followed by gas chromatography–mass spectrometry, Journal of
711 Chromatography A. 1217 (2010) 4771–4778. https://doi.org/10.1016/j.chroma.2010.05.047.
712 [33] B.H. Chen, E.H. Taylor, A.A. Pappas, Comparison of Derivatives for Determination of Codeine and Morphine by Gas
713 Chromatography/Mass Spectrometry, Journal of Analytical Toxicology. 14 (1990) 12–17.
714 https://doi.org/10.1093/jat/14.1.12.
715

Chapter 6

Conclusions

This thesis has explored multiple aspects of using GC-MS and machine learning (ML) in annotating semi-polar organic contaminants of emerging concern (CECs) through their silyl derivatives. Chapter 6 provides a summary of the scientific contributions of the thesis. It then revisits our hypotheses from Chapter 1 and discusses to which degree they were confirmed and presented in Chapters 3, 4, and 5. The thesis is concluded with an outline of several possible directions for further work.

6.1 Scientific Contributions of the Thesis

The thesis presents contributions to solving three problems related to the use of GC-MS analytical platforms and ML-based compound annotation (CA) approaches in eco-exposome annotation (EEA), particularly the annotation of semi-polar organic CEC.

1. The development of a methodological workflow and the generation of GC-EI-MS spectral datasets for cheminformatics-assisted CA: In Chapter 4, we proposed methods for the generation of training and test datasets of GC-EI-MS spectra for ML-based CA. The training data derived by MSL search and testing data (acquired analytically) were used with the ML-based CSI:IOKR approach to annotate CEC through their silyl derivatives from GC-EI-MS spectra. Notably, the generated GC-EI-MS datasets and the accompanying metadata are now publicly available on the Mendeley Data Repository. The intention is to encourage their further use as benchmark datasets for developing, validating, and evaluating existing and novel cheminformatics, particularly ML-based CA approaches. In such a way, we intend to encourage their use in GC-MS-based EEA.

2. *Application of a ML-based approach for annotating semi-polar CEC through their silyl derivatives:* We adapted the CSI:IOKR approach [6] and used it for CA based on GC-EI-MS spectral data. Here, the CSI:IOKR approach was trained on different NIST 17 MSL-derived GC-EI-MS spectra, obtained after three filtering steps, and tested on raw and background-subtracted GC-EI-MS spectra. The highest identification rates were achieved by training the CSI:IOKR model using the finally filtered GC-EI-MS training dataset and the raw GC-EI-MS dataset. The correct solution in the top 1, top 10, and top 20 hits in 10.6%, 37.5%, and 51.0% of the cases, respectively, were achieved by training CSI:IOKR with the GC-EI-MS training dataset obtained after three filtering steps and the raw testing GC-EI-MS dataset. Careful curation of training datasets of (GC-EI-)MS spectra, involving analytical knowledge in selecting the compound and spectra, is crucial for accurate and reliable ML-based CA. CSI:IOKR represents an efficient approach to prioritizing candidates and shortens the time required for CA compared to manual MSL search by a factor of 50. To our knowledge, this study is the first to investigate the application of sophisticated ML-based CA approaches for CEC identification utilizing GC-EI-MS spectra.

3. *Investigation of derivatization conditions and stability of silyl derivatives of structurally and physicochemically diverse CECs:* detection and recognition of the importance of derivatization reactions prior to GC-MS analysis and the stability of the generated CEC-TMS derivatives on the overall determination and quantification performance is vital. In such a way, it is possible to enhance the accuracy and confidence of CEC identification and quantification. Chapter 5 introduced a methodology for optimizing derivatization using chemometrics tools, such as contour plotting, principal component analysis (PCA), and hierarchical clustering analysis (HCA). It also presents a robust analytical procedure for quantification of 70 CEC using solid-phase extraction (SPE), followed by trimethylsilylation and GC-MS analysis. Method validation proved the

method's appropriateness for studying the stability of CEC-TMS derivatives in solvent and artificial wastewater (AWW) extracts. The method can also be successfully employed in quantifying selected CECs in aqueous environmental compartments. Significant stability patterns were discovered, some of which revealed stability irregularities of certain CECs, especially after consecutive freezing and thawing of the sample. In particular, TMS derivatives of polyhydroxy CECs (ERY, CA) and estrogen hormones (E2, E3, and EE2) showed the most irregularities. ERY-TMS and CA-TMS were unstable in a solvent at 25°C and 4°C when stored for up to 28 days. E2-TMS and CA-TMS exhibited unexpected increases in concentration, while EE2-TMS and E3-TMS degraded significantly in AWW extracts stored at 25°C and 4°C for 20 weeks. For the first time, nine and six CEC-TMS derivatives were identified that degrade during three freezing and thawing cycles in solvent and AWW extracts, respectively. Based on the generated knowledge, for most tested CEC-TMS derivatives, it is recommended to store samples (solvent and AWW extract) at -18°C for up to 20 weeks. After more than two cycles of sample freezing and thawing, performing analysis should be strictly avoided.

6.2 Research Hypotheses and Their Confirmation

According to the dissertation goals, the research in this dissertation tested the following hypotheses:

H1: Poor stability and chromatographic behavior (peak shape, R_t) of silylated derivatives can reveal patterns, which can reduce confidence and accuracy of their identification and quantification.

In Section 1.3, it was hypothesized that poor stability and chromatographic behavior (peak shape, R_t) of silyl derivatives of semi-polar CEC could reveal patterns in their behavior prior to, during GC-MS analysis, and during storage, which can reduce the confidence and accuracy of their identification and quantification (**H1**). Reliable knowledge of the stability and associated chromatographic behavior patterns of silyl derivatives of semi-polar CEC should be gained through investigation in different experimental conditions and matrices, using optimized derivatization procedures and validated GC-MS analytical method with satisfactory accuracy, linearity, and repeatability. These issues are addressed in Section 5.2, which presents a workflow for the optimization of derivatization conditions and stability evaluation for a representative selection of semi-polar CEC silyl derivatives. For all 70 CEC, BSTFA + 1% TMCS gave the highest derivatization yield, while according to the derivatization conditions, CEC were grouped into three derivatization protocols. Regarding the stability of the generated CEC-TMS derivatives, the degradation of several CEC-TMS derivatives was identified, either estrogen hormones or polyhydroxy CEC ($\geq$ 3 -OH groups). Stability irregularities were observed during storage of ERY and CA TMS derivatives at 4°C and ERY at -18°C in a solvent. TMS derivatives of EE2 and E3 significantly degraded at 25°C and 4°C in AWW extracts, while TMS derivatives of the compounds CA and E2 exhibited unexpected concentration increases. Noticeable degradation was also observed during freezing and thawing, such that nine and five CEC-TMS derivatives degraded in solvent and AWW extracts, respectively, to $\leq$ 85% of the initial concentration after two freeze and thaw cycles. Finally, MU estimation showed that the uncertainty of the calibration curve is the most significant source of uncertainty. However, observed deviations in CEC-TMS concentrations during the stability studies are attributed solely to their unsatisfactory stability but not to MU. Based on the findings above, the **H1** hypothesis was confirmed.

H2: Structurally diverse semi-polar and thermolabile CEC can be successfully identified by their silylated derivatives using their GC-EI-MS spectra in complex mixtures.

The second hypothesis (**H2**) was that structurally diverse semi-polar and thermolabile CEC could be successfully identified by their silyl derivatives using their GC-EI-MS spectra in complex mixtures. Structurally diverse CEC silyl derivatives in different environmental matrices were identified (Section 4.4), including river water (RW), wastewater effluent

(WWE), and wastewater influent (WWI). On a subset of 56 CEC-TMS derivatives, we performed NIST 17 MSL search and excluded CEC-TMS whose GC-EI-MS spectra were not in NIST 17 MSL and were not among the top 100 hits. Of the remaining CEC-TMS derivatives, it was possible to correctly identify (as top 1 hit) 82.05%, 87.50%, 72.50%, and 57.14% of the CEC-TMS whose spectra are present in NIST 17 MSL [1] in RW A, RW B, WWE, and WWI, respectively. Based on the findings above, hypothesis **H2** was confirmed.

H3: ML approach using training and test datasets of GC-EI-MS spectra of silylated compounds will result in a higher number of correctly identified compounds than non-ML approach.

The third hypothesis (**H3**) was that using a ML-based approach with training and test datasets of GC-EI-MS spectra of silyl compounds and CA would significantly improve, especially compared to non-ML approaches. We addressed the issues of generating GC-EI-MS datasets and applying the IOKR-based approach in annotating semi-polar CEC as silyl derivatives in Section 4.2. It shows that the IOKR approach correctly ranks 37% and 50% of the tested CEC among the top 10 and top 20 candidates, respectively, and thus provides efficient candidate prioritization and reduced time required for CA. Additionally, we compared the performance of the IOKR-based approach with a non-ML approach, i.e., with manual MSL search in Section **Error! Reference source not found.**. A manual search of NIST 17 MSL [1] ranked 96.10% of the CEC-TMS present in NIST 17 MSL in the top 10 and 89.61% in the top 1 hits. Despite the high identification rates, the manual NIST 17 MSL search requires conversion of the MS spectral data to various data formats, potentially leading to the loss of critical structural information for CA. Additionally, the manual non-ML approach requires more (approximately 50-times more) of the analysts' time to accomplish the same task compared to the ML approach. Based on these findings, the third hypothesis was confirmed.

6.3 Further Work

The results obtained in this thesis can be extended by employing the presented approaches to become valuable tools for the task of thorough eco-exposome investigation. Due to the broadness of the derivatization-SPE-GC-MS approach employed in stability investigation, it can be further upgraded to an automated analytical pipeline. Together with the cheminformatics advances and minimal human effort, the parent CEC would be identified from complex environmental samples in a "vial-to-list" approach. The developed approach can be further used for the investigation of the stability of other silyl derivatives of CEC, such as TBDMS derivatives. Used together with TMS silylation, TBDMS silylation can broaden the CEC coverage of GC-MS analytical methods for suspect and non-target screening. Additionally, stability profiles of silyl derivatives of structurally and physicochemically diverse CEC selections should be examined in other complex matrices in order to discover stability patterns important for accurate and confident annotation during eco-exposome annotation (EEA) studies.

In this thesis, we showed that the IOKR-based ML approach could be very successful in candidate prioritization and CA of CEC silyl derivatives. Further work should focus on fully automating all steps of collecting the required data through a user-friendly GUI interface that would include modules for model training, model testing, and data analyses. In that way, the CSI:IOKR approach would be successfully used by environmental and exposomics analysts with no prior ML and programming knowledge. Also, apart from the CSI:IOKR approach, other ML CA approaches based on IOKR methodologies, such as MP-IOKR [94], IOKRFusion [95], as well as SIMPLE [110], and ADAPTIVE [111], should be used in GC-MS-based CEC annotation. Unsupervised ML, which is expected to discover potentially meaningful substructures from MS data (also known as substructure prediction and classification, discussed in the paper in Section 3.2), should also be considered in future work. Finally, a combination of direct and indirect annotation approaches should be considered, each with a classification approach, DB search, heuristics, and filtering approaches. They would serve for retrospective analysis of results from each CA tool for

confirmation of the presence of structurally similar compounds. The data insufficiencies in compound DB and MSL in regards to EE data, and especially in regard to the presence of silyl derivatives of CEC must be overcome for further implementation of the cheminformatics tools in EEA, especially ML approaches, including the approach we used in this thesis, that employ DBs for candidate retrieval.

Finally, we note the lack of standardized methods for performance evaluation of compound identification approaches along relevant metrics/parameters (e.g., specificity, sensitivity, positive predictive value, negative predictive value, accuracy, and false discovery rate) and the lack of benchmark datasets of high-quality GC-MS and LC-MS/MS spectra acquired on different analytical platforms as a significant limitation in the use of MS-based cheminformatics CA. In that spirit, we have provided the cheminformatics community with a valuable collection of datasets of GC-EI-MS spectra of TMS and TBDMS derivatives and have made significant effort to encourage the cheminformatics and EE communities toward their use. A worthwhile direction for future work is the generation of a cheminformatics platform that will integrate all open access cheminformatics approaches/software for compound identification. It should also allow analyst-friendly parameter optimization, performance evaluation, and results analysis using the versatile collection of GC-EI-MS and LC-MS datasets curated at the research group for Organic Analysis using a standardized set of evaluation metrics. We believe that the cheminformatics, exposomics, and metabolomics communities will significantly benefit from the implementation of the findings of this doctoral study.

References

[1] National Institute of Standards and Technology, "NIST/EPA/NIH Mass Spectral Library 2017," *Wiley.com*, 2017. https://www.wiley.com/en-ai/NIST+EPA+NIH+Mass+Spectral+Library+2017-p-9781119750291 (accessed Aug. 15, 2022).

[2] C. P. Wild, "Complementing the Genome with an 'Exposome': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 8, pp. 1847–1850, Aug. 2005, doi: 10.1158/1055-9965.EPI-05-0456.

[3] C. P. Wild, "The exposome: from concept to utility," *Int. J. Epidemiol.*, vol. 41, no. 1, pp. 24–32, Feb. 2012, doi: 10.1093/ije/dyr236.

[4] N. R. Council, "Exposure Science in the 21st Century: A Vision and a Strategy," vol. 23, no. 1, Jan. 2013, doi: https://doi.org/10.1038/jes.2012.109.

[5] M. Ljoncheva, T. Stepišnik, S. Džeroski, and T. Kosjek, "Cheminformatics in MS-based environmental exposomics: Current achievements and future directions," *Trends in Environmental Analytical Chemistry*, vol. 28, p. e00099, Dec. 2020, doi: 10.1016/j.teac.2020.e00099.

[6] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, and J. Rousu, "Fast metabolite identification with Input Output Kernel Regression," *Bioinformatics*, vol. 32, no. 12, pp. i28–i36, Jun. 2016, doi: 10.1093/bioinformatics/btw246.

[7] A. Kovačič *et al.*, "Stability, biological treatment and UV photolysis of 18 bisphenols under laboratory conditions," *Environmental Research*, vol. 179, p. 108738, Dec. 2019, doi: 10.1016/j.envres.2019.108738.

[8] M. Česen, D. Heath, M. Krivec, J. Košmrlj, T. Kosjek, and E. Heath, "Seasonal and spatial variations in the occurrence, mass loadings and removal of compounds of emerging concern in the Slovene aqueous environment and environmental risk assessment," *Environmental Pollution*, vol. 242, pp. 143–154, Nov. 2018, doi: 10.1016/j.envpol.2018.06.052.

[9] I. González-Mariño, J. B. Quintana, I. Rodríguez, and R. Cela, "Determination of drugs of abuse in water by solid-phase extraction, derivatisation and gas chromatography–ion trap-tandem mass spectrometry," *Journal of Chromatography A*, vol. 1217, no. 11, pp. 1748–1760, Mar. 2010, doi: 10.1016/j.chroma.2010.01.046.

[10] A. Azzouz and E. Ballesteros, "Trace analysis of endocrine disrupting compounds in environmental water samples by use of solid-phase extraction and gas chromatography with mass spectrometry detection," *Journal of Chromatography A*, vol. 1360, pp. 248–257, Sep. 2014, doi: 10.1016/j.chroma.2014.07.059.

[11] B. Molnár and I. Molnár-Perl, "The role of alkylsilyl derivatization techniques in the analysis of illicit drugs by gas chromatography," *Microchemical Journal*, vol. 118, pp. 101–109, Jan. 2015, doi: 10.1016/j.microc.2014.08.014.

[12] C. Schummer, O. Delhomme, B. Appenzeller, R. Wennig, and M. Millet, "Comparison of MTBSTFA and BSTFA in derivatization reactions of polar compounds prior to GC/MS analysis," *Talanta*, vol. 77, no. 4, pp. 1473–1482, Feb. 2009, doi: 10.1016/j.talanta.2008.09.043.

[13] K. K. Pasikanti, P. C. Ho, and E. C. Y. Chan, "Development and validation of a gas chromatography/mass spectrometry metabonomic platform for the global profiling of urinary metabolites," *Rapid Commun. Mass Spectrom.*, vol. 22, no. 19, pp. 2984–2992, Oct. 2008, doi: 10.1002/rcm.3699.

[14] A. Shareef, M. J. Angove, and J. D. Wells, “Optimization of silylation using N-methyl-N-(trimethylsilyl)-trifluoroacetamide, N,O-bis-(trimethylsilyl)-trifluoroacetamide and N-(tert-butyldimethylsilyl)-N-methyltrifluoroacetamide for the determination of the estrogens estrone and 17α-ethinylestradiol by gas chromatography–mass spectrometry,” *Journal of Chromatography A*, vol. 1108, no. 1, pp. 121–128, Mar. 2006, doi: 10.1016/j.chroma.2005.12.098.

[15] Z. L. Zhang, A. Hibberd, and J. L. Zhou, “Optimisation of derivatisation for the analysis of estrogenic compounds in water by solid-phase extraction gas chromatography–mass spectrometry,” *Analytica Chimica Acta*, vol. 577, no. 1, pp. 52–61, Sep. 2006, doi: 10.1016/j.aca.2006.06.029.

[16] N. Migowska, M. Caban, P. Stepnowski, and J. Kumirska, “Simultaneous analysis of non-steroidal anti-inflammatory drugs and estrogenic hormones in water and wastewater samples using gas chromatography–mass spectrometry and gas chromatography with electron capture detection,” *Science of The Total Environment*, vol. 441, pp. 77–88, Dec. 2012, doi: 10.1016/j.scitotenv.2012.09.043.

[17] S. Sauvé and M. Desrosiers, “A review of what is an emerging contaminant,” *Chemistry Central Journal*, vol. 8, no. 1, p. 15, Dec. 2014, doi: 10.1186/1752-153X-8-15.

[18] J. Xue, Y. Lai, C.-W. Liu, and H. Ru, “Towards Mass Spectrometry-Based Chemical Exposome: Current Approaches, Challenges, and Future Directions,” *Toxics*, vol. 7, no. 3, p. 41, Aug. 2019, doi: 10.3390/toxics7030041.

[19] B. L. Milman and I. K. Zhurkovich, “The chemical space for non-target analysis,” *TrAC Trends in Analytical Chemistry*, vol. 97, pp. 179–187, Dec. 2017, doi: 10.1016/j.trac.2017.09.013.

[20] R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, and C. Steinbeck, “Navigating freely-available software tools for metabolomics analysis,” *Metabolomics*, vol. 13, no. 9, p. 106, Sep. 2017, doi: 10.1007/s11306-017-1242-7.

[21] U. W. Liebal, A. N. T. Phan, M. Sudhakar, K. Raman, and L. M. Blank, “Machine Learning Applications for Mass Spectrometry-Based Metabolomics,” *Metabolites*, vol. 10, no. 6, p. 243, Jun. 2020, doi: 10.3390/metabo10060243.

[22] K. O’Shea and B. B. Misra, “Software tools, databases and resources in metabolomics: updates from 2018 to 2019,” *Metabolomics*, vol. 16, no. 3, p. 36, Mar. 2020, doi: 10.1007/s11306-020-01657-3.

[23] J. Hollender, E. L. Schymanski, H. P. Singer, and P. L. Ferguson, “Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?,” *Environ. Sci. Technol.*, vol. 51, no. 20, pp. 11505–11512, Oct. 2017, doi: 10.1021/acs.est.7b02184.

[24] M. Pétéra *et al.*, “Workflow4Metabolomics: an international computing infrastructure for Metabolomics,” presented at the Galaxy Community Conference (GCC2019), Freiburg, Germany, 2019.

[25] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak, “XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification,” *Anal. Chem.*, vol. 78, no. 3, pp. 779–787, Feb. 2006, doi: 10.1021/ac051437y.

[26] Z. Pang *et al.*, “MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights,” *Nucleic Acids Research*, vol. 49, no. W1, pp. W388–W396, Jul. 2021, doi: 10.1093/nar/gkab382.

[27] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, “MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data,” *BMC Bioinformatics*, vol. 11, no. 1, p. 395, Dec. 2010, doi: 10.1186/1471-2105-11-395.

[28] M. Loos, “enviMass version 3.1.” Zenodo, Mar. 23, 2016. doi: 10.5281/zenodo.48164.

[29] “nontarget-package: Detecting Isotope, Adduct and Homologue Relations in LC-MS... in nontarget: Detecting Isotope, Adduct and Homologue Relations in LC-MS Data.” https://rdrr.io/cran/nontarget/man/nontarget-package.html (accessed Aug. 02, 2022).

[30] R. Helmus, T. L. ter Laak, A. P. van Wezel, P. de Voogt, and E. L. Schymanski, “patRoon: open source software platform for environmental mass spectrometry based non-target screening,” *J Cheminform*, vol. 13, no. 1, p. 1, Dec. 2021, doi: 10.1186/s13321-020-00477-w.

[31] C. Martins, C. P. Costa, and S. M. Rocha, “Multidimensional gas chromatography for environmental exposure measurement,” in *Multidimensional Analytical Techniques in Environmental Research*, Elsevier, 2020, pp. 209–229. doi: 10.1016/B978-0-12-818896-5.00008-9.

[32] G. J. Getzinger and P. L. Ferguson, “Illuminating the Exposome with High-resolution Accurate-mass Mass Spectrometry and Non-targeted analysis,” *Current Opinion in Environmental Science & Health*, vol. 15, pp. 49–56, Jun. 2020, doi: 10.1016/j.coesh.2020.05.005.

[33] M. Holčapek, R. Jirásko, and M. Lísa, “Recent developments in liquid chromatography–mass spectrometry and related techniques,” *Journal of Chromatography A*, vol. 1259, pp. 3–15, Oct. 2012, doi: 10.1016/j.chroma.2012.08.072.

[34] I. Blaženović, T. Kind, J. Ji, and O. Fiehn, “Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics,” *Metabolites*, vol. 8, no. 2, p. 31, May 2018, doi: 10.3390/metabo8020031.

[35] L. C. Menikarachchi *et al.*, “MolFind: A Software Package Enabling HPLC/MS-Based Identification of Unknown Chemical Structures,” *Analytical Chemistry*, vol. 84, no. 21, pp. 9388–9394, Nov. 2012, doi: 10.1021/ac302048x.

[36] D. A. Skoog, J. F. Holler, and S. R. Crouch, “An Introduction to Chromatographic Separations,” in *Principles of Instrumental Analysis*, 7th ed., USA: Cengage Learning, 2016, pp. 696–719.

[37] D. O. Sparkman, Z. E. Penton, and F. G. Kitson, *Gas Chromatography and Mass Spectrometry: A Practical Guide*, 2nd ed. USA: Elsevier, 2011.

[38] D. A. Skoog, J. F. Holler, and S. R. Crouch, “Gas Chromatography,” in *Principles of Instrumental Analysis*, 7th ed., USA: Cengage Learning, 2016, pp. 720–745.

[39] C. F. Poole, “Ionization-based detectors for gas chromatography,” *Journal of Chromatography A*, vol. 1421, pp. 137–153, Nov. 2015, doi: 10.1016/j.chroma.2015.02.061.

[40] G. Guiochon and C. L. Guillemin, “Methodology: Gas Chromatographic Instrumentation and Detectors for Gas Chromatography,” in *Quantitative Gas Chromatography For Laboratory Analyses and On-Line Process Control*, Elsevier Science, 1988, pp. 393–480.

[41] S. Fanali, P. R. Haddad, C. F. Poole, P. Schoenmakers, and D. Lloyd, *Liquid Chromatography: Fundamentals and Instrumentation*, 1st ed. USA: Elsevier, 2013.

[42] D. A. Skoog, J. F. Holler, and S. R. Crouch, “High-Performance Liquid Chromatography,” in *Principles of Instrumental Analysis*, 7th ed., USA: Cengage Learning, 2016, pp. 746–781.

[43] D. A. Skoog, J. F. Holler, and S. R. Crouch, “Molecular Mass Spectrometry,” in *Principles of Instrumental Analysis*, 7th ed., USA: Cengage Learning, 2016, pp. 501–553.

[44] E. de Hoffman and V. Stroobant, *Mass Spectrometry: Principles and Applications*, 3rd ed. West Sussex, England: John Wiley&Sons Ltd, 2007.

[45] P. Kebarle and U. H. Verkerk, “Electrospray: From ions in solution to ions in the gas phase, what we know now,” *Mass Spectrometry Reviews*, vol. 28, no. 6, pp. 898–917, 2009, doi: 10.1002/mas.20247.

[46] A. G. Harrison, *Chemical Ionization Mass Spectrometry*, 2nd ed. Florida, USA: CRC Press, 2000.

[47] V. Zaikin and J. M. Halket, *A Handbook of Derivatives for Mass Spectrometry*. West Sussex, England: IM Publications, 2009.

[48] R. E. March and J. F. Todd, *Quadrupole ion trap mass spectrometry*, 2nd ed. New Jersey, USA: Wiley-Interscience, 2005.

[49] T. Kind and O. Fiehn, “Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm,” *BMC Bioinformatics*, vol. 7, no. 1, p. 234, Dec. 2006, doi: 10.1186/1471-2105-7-234.

[50] A. M. Haag, “Mass Analyzers and Mass Spectrometers,” in *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*, vol. 919, H. Mirzaei and M. Carrasco, Eds. Cham: Springer International Publishing, 2016, pp. 157–169. doi: 10.1007/978-3-319-41448-5_7.

[51] B. A. Mamyrin, V. I. Karataev, D. V. Shmikk, and V. A. Zagulin, “The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution,” *Zh. Eksp. Teor. Fiz*, vol. 64, pp. 82–89.

[52] I. V. Chernushevich, A. V. Loboda, and B. A. Thomson, “An introduction to quadrupole–time-of-flight mass spectrometry,” *Journal of Mass Spectrometry*, vol. 36, no. 8, pp. 849–865, 2001, doi: 10.1002/jms.207.

[53] J. M. Campbell, B. A. Collings, and D. J. Douglas, “A new linear ion trap time-of-flight system with tandem mass spectrometry capabilities,” *Rapid Communications in Mass Spectrometry*, vol. 12, no. 20, pp. 1463–1474, 1998, doi: 10.1002/(SICI)1097-0231(19981030)12:20<1463::AID-RCM357>3.0.CO;2-H.

[54] M. B. Comisarow and A. G. Marshall, “Fourier transform ion cyclotron resonance spectroscopy,” *Chemical Physics Letters*, vol. 25, pp. 282–283, 1974, doi: https://doi.org/10.1016/0009-2614(74)89137-2.

[55] A. Makarov, “Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis,” *Anal. Chem.*, vol. 72, no. 6, pp. 1156–1162, Mar. 2000, doi: 10.1021/ac991131p.

[56] R. A. Zubarev and A. Makarov, “Orbitrap Mass Spectrometry,” *Anal. Chem.*, vol. 85, no. 11, pp. 5288–5296, Jun. 2013, doi: 10.1021/ac4001223.

[57] B.-L. Qi, P. Liu, Q.-Y. Wang, W.-J. Cai, B.-F. Yuan, and Y.-Q. Feng, “Derivatization for liquid chromatography-mass spectrometry,” *TrAC Trends in Analytical Chemistry*, vol. 59, pp. 121–132, Jul. 2014, doi: 10.1016/j.trac.2014.03.013.

[58] C. W. Brooks, C. G. Edmonds, and S. J. Gaskell, “Derivatives suitable for GC-MS,” *Chemistry and Physics of Lipids*, vol. 21, no. 4, pp. 403–416, 1978, doi: doi:10.1016/0009-3084(78)90049-x.

[59] D.-L. Lin, S.-M. Wang, C.-H. Wu, B.-G. Chen, and R. H. Liu, “Chemical derivatization for the analysis of drugs by GC-MS - A conceptual review,” *Journal of Food and Drug Analysis*, vol. 16, no. 1, pp. 1–10, 2008, doi: 10.38212/2224-6614.2373.

[60] J. Drozd, “Chemical derivatization in gas chromatography,” *Journal of Chromatography A*, vol. 113, no. 3, pp. 303–356, 1975, doi: doi:10.1016/s0021-9673(00)95303-2.

[61] H. Kataoka, “Gas Chromatography of Amines as Various Derivatives,” in *Quantitation of Amino Acids and Amines by Chromatography*, 1st ed., vol. 70, USA: Elsevier, 2005, pp. 364–404.

[62] K. Blau and J. M. Halket, *Handbook of Derivatives for Chromatography*. USA: Wiley, 1993.

[63] J. M. Halket and V. G. Zaikin, “Derivatization in Mass Spectrometry—1. Silylation,” *Eur J Mass Spectrom (Chichester)*, vol. 9, no. 1, pp. 1–21, Feb. 2003, doi: 10.1255/ejms.527.

[64] B. P. Gumbi, B. Moodley, G. Birungi, and P. G. Ndungu, “Target, Suspect and Non-Target Screening of Silylated Derivatives of Polar Compounds Based on Single Ion Monitoring GC-MS,” *IJERPH*, vol. 16, no. 20, p. 4022, Oct. 2019, doi: 10.3390/ijerph16204022.

[65] F. Orata, “Derivatization Reactions and Reagents for Gas Chromatography Analysis,” in *Advanced Gas Chromatography - Progress in Agricultural, Biomedical and Industrial Applications*, M. Ali Mohd, Ed. InTech, 2012. doi: 10.5772/33098.

[66] C. F. Poole, “Trialkylsilyl derivatives (other than TMS) for Gas Chromatography and Mass Spectrometry,” *Journal of Chromatography A*, vol. 17, no. 3, pp. 115–123, 1979, doi: 10.1016/j.chroma.2013.01.097.

[67] M. Sindelar and G. J. Patti, “Chemical Discovery in the Era of Metabolomics,” *J. Am. Chem. Soc.*, vol. 142, no. 20, pp. 9097–9105, May 2020, doi: 10.1021/jacs.9b13198.

[68] D. S. Wishart *et al.*, “HMDB 4.0: the human metabolome database for 2018,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D608–D617, Jan. 2018, doi: 10.1093/nar/gkx1089.

[69] D. Wishart *et al.*, “T3DB: the toxic exposome database,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D928–D934, Jan. 2015, doi: 10.1093/nar/gku1004.

[70] A. J. Williams *et al.*, "The CompTox Chemistry Dashboard: a community data resource for environmental chemistry," *J Cheminform*, vol. 9, Nov. 2017, doi: 10.1186/s13321-017-0247-6.

[71] C. Guijas *et al.*, "METLIN: A Technology Platform for Identifying Knowns and Unknowns," *Anal. Chem.*, vol. 90, no. 5, pp. 3156–3164, Mar. 2018, doi: 10.1021/acs.analchem.7b04424.

[72] H. Horai *et al.*, "MassBank: a public repository for sharing mass spectral data for life sciences," *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, 2010, doi: 10.1002/jms.1777.

[73] "MassBank of North America." https://mona.fiehnlab.ucdavis.edu/ (accessed Aug. 02, 2022).

[74] T. Kind *et al.*, "FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry," *Anal. Chem.*, vol. 81, no. 24, pp. 10038–10048, Dec. 2009, doi: 10.1021/ac9019522.

[75] "Wiley Registry of Mass Spectral Data, 12th Edition," *Wiley Science Solutions*. https://sciencesolutions.wiley.com/solutions/technique/gc-ms/wiley-registry-of-mass-spectral-data-12th-edition/ (accessed Aug. 06, 2022).

[76] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka, "Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches," *Brief Bioinform*, vol. 20, no. 6, pp. 2028–2043, Aug. 2018, doi: 10.1093/bib/bby066.

[77] E. L. Schymanski *et al.*, "Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challneges," *Environmental science: Process and Impacts*, vol. 21, pp. 1426–1445, 2019, doi: DOI: 10.1039/c9em00068b.

[78] "ACD/MS Fragmenter." Advanced Chemistry Labs, Toronto, Canada, 2018. Accessed: Jul. 23, 2022. [Online]. Available: https://www.acdlabs.com/products/adh/ms/ms_frag/

[79] "Mass Frontier™ Spectral Interpretation Software." Accessed: Aug. 11, 2022. [Online]. Available: https://www.thermofisher.com/order/catalog/product/OPTON-30920

[80] H. Tsugawa *et al.*, "Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software," *Anal. Chem.*, vol. 88, no. 16, pp. 7946–7958, Aug. 2016, doi: 10.1021/acs.analchem.6b00770.

[81] F. Rasche *et al.*, "Identifying the Unknowns by Aligning Fragmentation Trees," *Analytical Chemistry*, vol. 84, no. 7, pp. 3417–3426, Apr. 2012, doi: 10.1021/ac300304u.

[82] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann, "In silico fragmentation for computer assisted identification of metabolite mass spectra," *BMC Bioinformatics*, vol. 11, Mar. 2010, doi: 10.1186/1471-2105-11-148.

[83] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann, "MetFrag relaunched: incorporating strategies beyond in silico fragmentation," *J Cheminform*, vol. 8, Jan. 2016, doi: 10.1186/s13321-016-0115-9.

[84] C. Ruttkies, S. Neumann, and S. Posch, "Improving MetFrag with statistical learning of fragment annotations," *BMC Bioinformatics*, vol. 20, Dec. 2019, doi: 10.1186/s12859-019-2954-7.

[85] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, and J. Vervoort, "Substructure-based annotation of high-resolution multistage MSn spectral trees," *Rapid Commun. Mass Spectrom.*, vol. 26, no. 20, pp. 2461–2471, Oct. 2012, doi: 10.1002/rcm.6364.

[86] L. C. Menikarachchi, R. Dubey, D. W. Hill, D. N. Brush, and D. F. Grant, "Development of Database Assisted Structure Identification (DASI) Methods for Nontargeted Metabolomics," *Metabolites*, vol. 6, no. 17, Jun. 2016, doi: 10.3390/metabo6020017.

[87] Y. Wang, G. Kora, B. P. Bowen, and C. Pan, “MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics,” *Anal. Chem.*, vol. 86, no. 19, pp. 9496–9503, Oct. 2014, doi: 10.1021/ac5014783.

[88] Y. Wang, X. Wang, and X. Zeng, “MIDAS-G: a computational platform for investigating fragmentation rules of tandem mass spectrometry in metabolomics,” *Metabolomics*, vol. 13, no. 10, p. 116, Aug. 2017, doi: 10.1007/s11306-017-1258-z.

[89] H. Tsugawa *et al.*, “A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms,” *Nat Methods*, vol. 16, no. 4, pp. 295–298, Apr. 2019, doi: 10.1038/s41592-019-0358-2.

[90] F. Qiu, Z. Lei, and L. W. Sumner, “MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications,” *Analytica Chimica Acta*, vol. 1037, pp. 316–326, Dec. 2018, doi: 10.1016/j.aca.2018.03.052.

[91] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu, “Metabolite identification and molecular fingerprint prediction through machine learning,” *Bioinformatics*, vol. 28, no. 18, pp. 2333–2341, Sep. 2012, doi: 10.1093/bioinformatics/bts437.

[92] H. Shen, K. Dührkop, S. Böcker, and J. Rousu, “Metabolite identification through multiple kernel learning on fragmentation trees,” *Bioinformatics*, vol. 30, no. 12, pp. i157–i164, Jun. 2014, doi: 10.1093/bioinformatics/btu275.

[93] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker, “Searching molecular structure databases with tandem mass spectra using CSI:FingerID,” *Proc Natl Acad Sci USA*, vol. 112, no. 41, pp. 12580–12585, Oct. 2015, doi: 10.1073/pnas.1509788112.

[94] C. Brouard, E. Bach, S. Bocker, and J. Rousu, “Magnitude-Preserving Ranking for Structured Outputs,” in *Proceedings of the Ninth Asian Conference on Machine Learning, PMLR*, 2017, vol. 77, pp. 407–422.

[95] C. Brouard, A. Bassé, F. d’Alché-Buc, and J. Rousu, “Improved Small Molecule Identification through Learning Combinations of Kernel Regression Models,” *Metabolites*, vol. 9, no. 8, Aug. 2019, doi: 10.3390/metabo9080160.

[96] H. Ji, Y. Xu, H. Lu, and Z. Zhang, “Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification,” *Anal. Chem.*, vol. 91, no. 9, pp. 5629–5637, May 2019, doi: 10.1021/acs.analchem.8b05405.

[97] K. Varmuza and W. Werther, “Mass Spectral Classifiers for Supporting Systematic Structure Elucidation,” *J. Chem. Inf. Comput. Sci.*, vol. 36, no. 2, pp. 323–333, Jan. 1996, doi: 10.1021/ci9501406.

[98] E. L. Schymanski, C. Meinert, M. Meringer, and W. Brack, “The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis,” *Analytica Chimica Acta*, vol. 615, no. 2, pp. 136–147, May 2008, doi: 10.1016/j.aca.2008.03.060.

[99] J. Hummel, N. Strehmel, J. Selbig, D. Walther, and J. Kopka, “Decision tree supported substructure prediction of metabolites from GC-MS profiles,” *Metabolomics*, vol. 6, no. 2, pp. 322–333, Jun. 2010, doi: 10.1007/s11306-010-0198-7.

[100] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, and S. Böcker, “De novo analysis of electron impact mass spectra using fragmentation trees,” *Analytica Chimica Acta*, vol. 739, pp. 67–76, Aug. 2012, doi: 10.1016/j.aca.2012.06.021.

[101] F. Allen, A. Pon, R. Greiner, and D. Wishart, “Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification,” *Anal. Chem.*, vol. 88, no. 15, pp. 7689–7697, Aug. 2016, doi: 10.1021/acs.analchem.6b01622.

[102] S. Grimme, “Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules,” *Angewandte Chemie International Edition*, vol. 52, no. 24, pp. 6306–6312, 2013, doi: 10.1002/anie.201300158.

[103] C. A. Bauer and S. Grimme, “How to Compute Electron Ionization Mass Spectra from First Principles,” *J. Phys. Chem. A*, vol. 120, no. 21, pp. 3755–3766, Jun. 2016, doi: 10.1021/acs.jpca.6b02907.

[104] V. Ásgeirsson, C. A. Bauer, and S. Grimme, "Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules," *Chem. Sci.*, vol. 8, no. 7, pp. 4879–4895, 2017, doi: 10.1039/C7SC00601B.

[105] S. E. Stein *et al.*, "NIST/EPA/NIH Mass Spectral Library (NIST 17) and NIST Mass Spectral Search Program (Version 2.3) User's Guide." Apr. 2017. Accessed: Aug. 15, 2022. [Online]. Available: https://chemdata.nist.gov/dokuwiki/lib/exe/fetch.php?media=chemdata:nist17:nistms_ver23man.pdf

[106] A. Kerber, M. Meringer, and C. Rücker, "CASE via MS: Ranking Structure Candidates by Mass Spectra," *Croat. Chem. Acta*, vol. 79, no. 3, pp. 449–464, 2006.

[107] A. T. Lebedev, *Comprehensive Environmental Mass Spectrometry*. Hertfordshire, United Kingdom: ILM Publications, 2012.

[108] Z. Lai and O. Fiehn, "Mass spectral fragmentation of trimethylsilylated small molecules," *Mass Spec Rev*, vol. 37, no. 3, pp. 245–257, May 2018, doi: 10.1002/mas.21518.

[109] C. Sánchez-Brunete, E. Miguel, B. Albero, and J. L. Tadeo, "Analysis of salicylate and benzophenone-type UV filters in soils and sediments by simultaneous extraction cleanup and gas chromatography–mass spectrometry," *Journal of Chromatography A*, vol. 1218, no. 28, pp. 4291–4298, Jul. 2011, doi: 10.1016/j.chroma.2011.05.030.

[110] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka, "SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra," *Bioinformatics*, vol. 34, no. 13, pp. i323–i332, Jul. 2018, doi: 10.1093/bioinformatics/bty252.

[111] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka, "ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra," *Bioinformatics*, vol. 35, no. 14, pp. i164–i172, Jul. 2019, doi: 10.1093/bioinformatics/btz319.

Bibliography

Publications Related to the Thesis

Journal Articles

Ljoncheva, M., Stepišnik, T., Džeroski, S., Kosjek, T. (2020). Cheminformatics in MS-based environmental exposomics: current achievements and future directions. *Trends in environmental analytical chemistry*. 28:e00099, ISSN 2214-1588. DOI: 10.1016/j.teac.2020.e00099.

Ljoncheva M., Stepišnik T., Kosjek T., Džeroski S. (2022) Machine learning for identification of silyl derivatives from mass spectra, *Journal of Cheminformatics* (accepted 31 July 2022).

Ljoncheva M., Kosjek T., Džeroski S., GC-EI-MS datasets of trimethylsilyl (TMS) and tert-butyl dimethyl silyl (TBDMS) derivatives for development of machine learning-based compound identification approaches, *Data in Brief* (submitted 4 July 2022).

Ljoncheva M., Heath E., Heath D., Džeroski S., Kosjek T., Contaminants of emerging concern: silylating procedures, evaluation of the stability of silyl derivatives and associated measurement uncertainty *Environmental Research* (submitted 25 August 2022).

Conference Papers

Ljoncheva, M., Heath, E., Džeroski, S., Kosjek, T. (2018) Generation of a test dataset for machine learning-assisted identification of contaminants of emerging concern. In: *Dežman, Miha (ed.), et al. Proceedings. 10th Jožef Stefan International Postgraduate School Students' Conference and 12th Young Researchers' Day 10th and 11th May 2018, Piran, Slovenia. Ljubljana: International Postgraduate School: Jožef Stefan Institute*, Str. 21. http://ipssc.mps.si/Proceedings/Proceedings_2018.pdf.

Ljoncheva, M., Heath, E., Džeroski, S., Kosjek, T. (2020) GC-MS analysis of contaminants of emerging concern. In: Jovičević Klug, Patricia (ed.), et al. *Book of abstracts. 12th Jožef Stefan International Postgraduate School Students' Conference and 14th Young Researchers' Day, 15th May 2020. Ljubljana: Jožef Stefan International Postgraduate School: Jožef Stefan Institute, Str. 22. http://ipssc.mps.si/BookOfAbstracts.pdf*.

Gerasimoska, T., Ljoncheva, M., Simjanoska, M. (2021). MSL-ST: development of mass spectral library search tool to enhance compound identification. V: LORENZ, Ronny (ur.), FRED, Ana (ur.), GAMBOA, Hugo (ur.). *BIOSTEC 2021: online proceedings. 14h International Joint Conference on Biomedical Engineering Systems and Technologies, February 11-13. [S. l.]: SciTePreess.* Str. 195-203. Biodevices, volume 3. ISBN 978-989-758-490-9. https://www.scitepress.org/Papers/2021/104241/104241.pdf.

Datasets

Ljoncheva, M., Kosjek, T., Džeroski, S. (2022). "GC-EI-MS datasets of trimethylsilyl (TMS) and tert-butyl dimethylsilyl (TBDMS) derivatives for development of machine learning-based compound identification approaches", Mendeley Data, V1, doi: 10.17632/j3z5bmvmnd.3

Other Publications

Ljoncheva, M., Kosjek, T., Isidori, M., Heath, E. (2020). 5-fluorouracil and its prodrug capecitabine: occurrence, fate and effects in the environment. In: HEATH, Ester (ed.), et al. Fate and effects of anticancer drugs in the environment. Cham: Springer, pp. 331-375. ISBN 978-3-030-21047-2, ISBN 978-3-030-21048-9. https://link.springer.com/content/pdf/10.1007%2F978-3-030-21048-9.pdf.

Biography

Milka Ljoncheva was born November 4, 1991, in Strumica, North Macedonia. From 2010-2015, she was enrolled in an integrated bachelor's and master's studies, study program Master of Pharmacy at the Faculty of Pharmacy, Ss. Cyril and Methodius University of Skopje, North Macedonia. During her studies, Milka was a scholar of the Ministry of Education and Science of the Government of the Republic of North Macedonia, receiving the National scholarship for talented bachelor students (2011-2015) and scholar of Trajce Mukaetov Foundation of Alkaloid AD, Skopje, North Macedonia (2012-2015). From 2014-2015, she worked as a laboratory and teaching assistant at the Department of Pharmaceutical Technology, Faculty of Pharmacy, Ss. Cyril and Methodius University of Skopje, North Macedonia. In July 2015, she finished her integrated studies with an overall GPA of 9.24, defending the master thesis entitled “Challenges in the formulation of PLGA-PCL nanoparticles for targeted drug delivery of SN-382 under Prof Katerina Gorachinova.

Between 09/2015 and 02/2017, she worked as a pharmacist; meanwhile, on 06/2016, receiving a pharmaceutical license from the Pharmaceutical Chamber of North Macedonia. In 2017, she became a scholar of the Ad Futura Scholarship for Postgraduate studies of Nationals of Western Balkan Countries in Slovenia by the Public Scholarship, Development, Disability, and Maintenance Fund of the Republic of Slovenia. Since February 2017, she has been enrolled as a student in the study program Ecotechnologies at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, and as a Ph.D. fellow at the group for Organic Analysis, Department of Environmental Sciences, Jožef Stefan Institute, Ljubljana, Slovenia. In 2018, Milka was part of the 8th Regional Biocamp organized by Lek in Ljubljana, Slovenia. As of today, Milka is an active member of numerous organizations, including the BioCamp Alumni Club (06/2020-present), Metabolomics society (11/2017-present), Réseau Francophone de Métabolomique et Fluxomique (RFMF) (11/2017-present), the Macedonian Pharmaceutical Association (MPA) (07/2015- present) and the Pharmaceutical Chamber of North Macedonia (07/2016-present). She has published two scientific papers (two in revision), one book chapter and presented her work at three international conferences.