Supporting Information

Prediction of NDMA formation potential using non-target analysis data: a proof of concept

Josep Sanchís[1,2,*], Mira Petrović[1,3], Maria José Farré[1,2,*]

[1.] Catalan Institute for Water Research (ICRA), Scientific and Technological Park of the University of Girona, H2O Building, C/Emili Grahit, 101, E17003, Girona, Spain.

[2.] University of Girona, 17071, Girona, Spain.

[3.] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010, Barcelona, Spain.

* Co-corresponding authors: jsanchis@icra.cat, mjfarre@icra.cat

Text S1. Linear regression models.

The tentative identification of the 11 features is presented in **Table S3**. Six of these tentatively identified substances were probably of anthropogenic origin, including pharmaceutical compounds (those features with quasi-molecular ions m/z 760.5048, 818.5466, 302.20769, 876.5886, and 658.4370) and a fluorosurfactant employed in fire extinguishing foams formulations (m/z 571.0925), while four features (m/z 202.1437, 174.0549, 171.1493, 192.1596) were small molecules with an unclear origin. One feature (m/z 679.473) was not associated to any MS^2 spectrum and hence could not be tentatively identified.

It should be highlighted that such degree of correlation does not necessarily imply that they are directly involved in the formation of NDMA. Apart from the molecule that elutes at t_R = 12.05 (feature #1 in **Table S3**), which presumably contains a dimethylamino group, their tentative structures do not support a potential role as NDMA precursor during chloramination, and this aspect should be assessed in further tests.

Figure S1. FISh annotated MS^2 spectra of the final NDMA precursor candidates. Those MS^2 signals with an atomic composition and presumed structure that was consistent with the composition and structure of the proposed precursor were automatically highlighted in green and annotated.

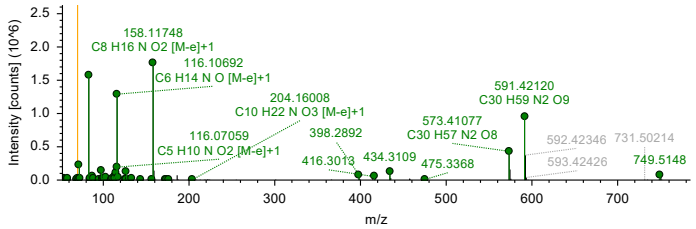
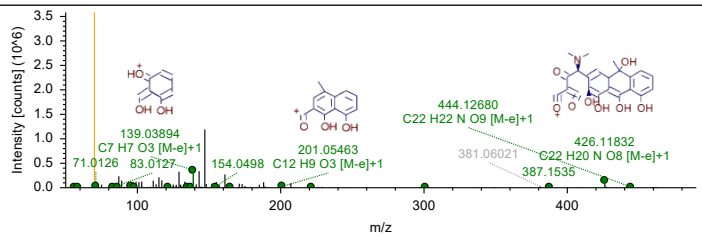
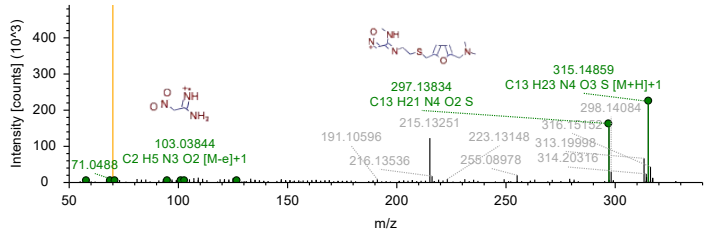
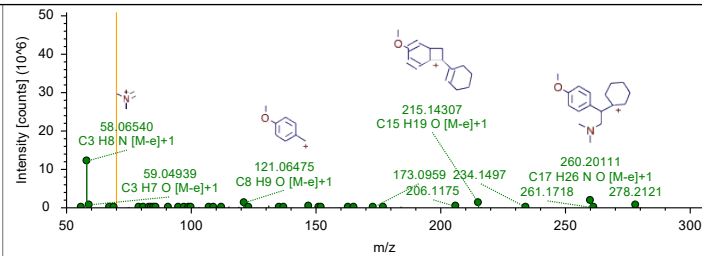
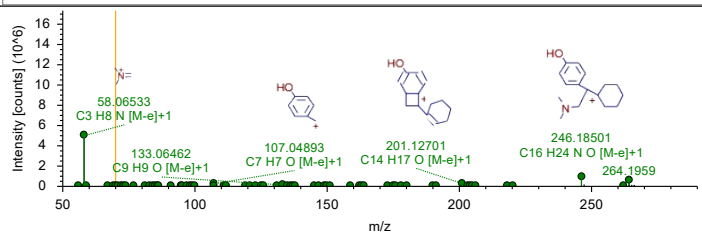
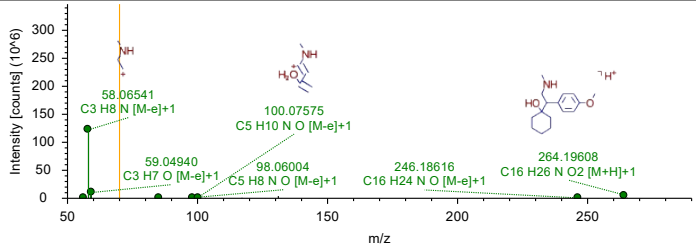
Compound	Spectrum
Azithromycin	
Oxytetracycline	
Ranitidine	
Venlafaxine	
***O*-desmethyl-venlafaxine**	
***N*-desmethyl-venlafaxine**	

Figure S2. Correlation matrix showing the Pearson's correlation coefficients among LC-HRMS features.

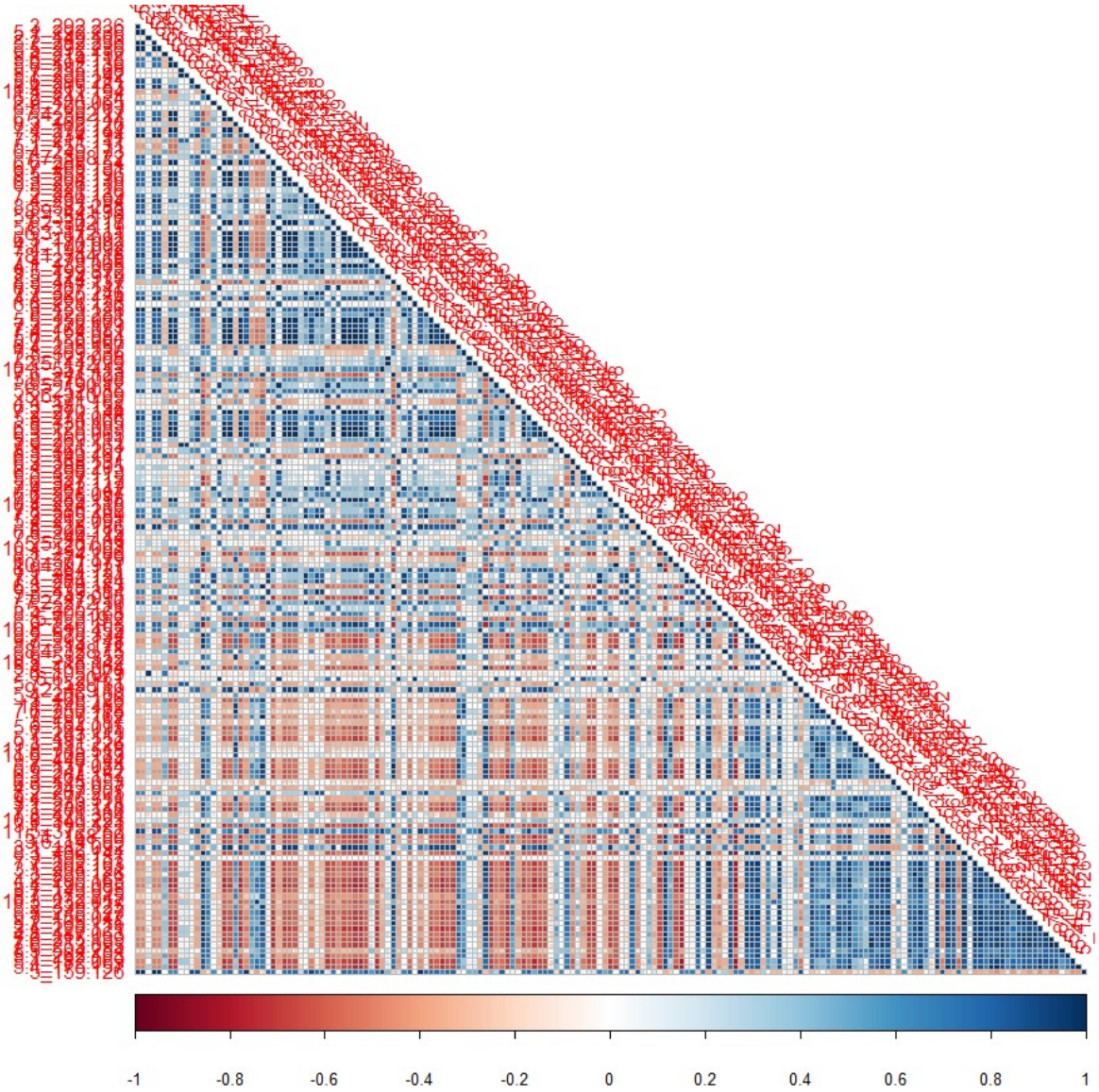


Figure S3. Correlation matrix with the final set of peaks, chosen because of their ubiquity, intensity, variability and orthogonality. Their label (in red) is composed by: retention time (in minutes), underscore, and m/z.

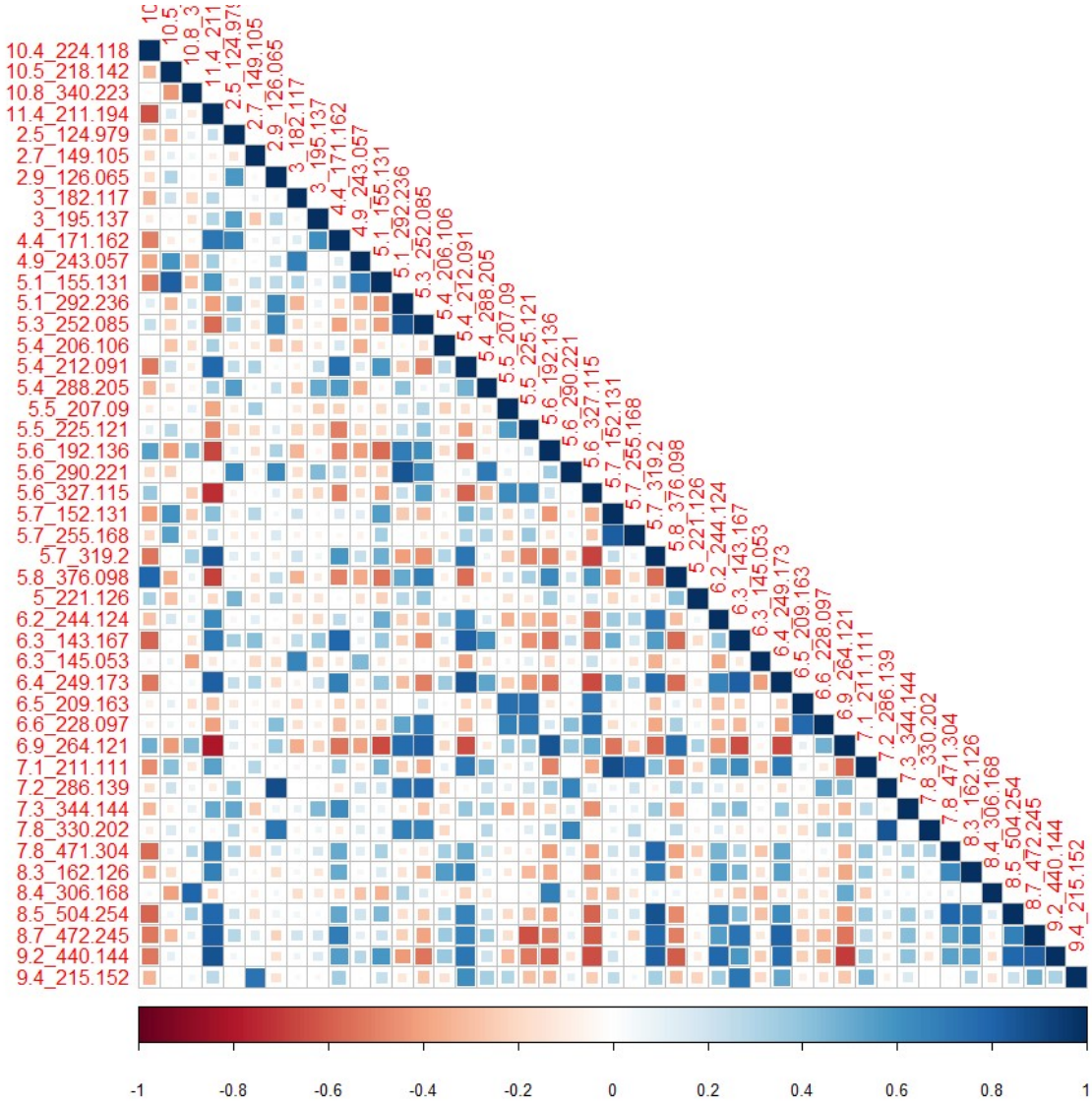


Table S1. Selected parameters of the Compound Discoverer workflow.

1. Align Retention Times	
Alignment Model	Adaptative Curve
Alignment Fallback	Linear model
Mass Tolerance	5 ppm
2. Detect Compounds	
2.1. General settings	
Mass Tolerance	5 ppm
Intensity Tolerance	30 %
S/N Threshold	3
Min Peak Intensity	10,000 a.u.
Ions	$[M+H]^+$; $[M+K]^+$; $[M+Na]^+$
Elements Counts	$C_{1–66}\,H_{1–126}\,O_{0–27}\,N_{0–25}\,S_{0–8}\,P_{0–6}\,Br_{0–8}\,Cl_{0–11}\,K_{0–1}\,Na_{0–1}$
2.2. Peak Detection	
Max Peak width	0.5
Min # Scans per Peak	5
Min # Isotopes	1
3. Compound Consolidation	
3.1. Compound Consolidation	
Mass Tolerance	5 ppm
RT Tolerance	0.3 min
3.2. Fragmment Data **Selection**	
Preferred Ions	$[M+H]^+$
4. Fill Gaps	
Mass Tolerance	5 ppm
S/N Threshold	3
5. Predict Compositions	
5.1. Precition Settings	
Mass Tolerance	5 ppm
Element Counts	$C_{1–66}\,H_{1–126}\,O_{0–27}\,N_{0–25}\,S_{0–8}\,P_{0–6}\,Br_{0–8}\,Cl_{0–11}\,K_{0–1}\,Na_{0–1}$
RDBE	-1–40
H/C	0.2–3.1
5.2. Pattern Matching	
Intensity Tolerance	30 %
Intensity Threshold	0.1 %
S/N Threshold	3
Min Spectral Fit	30 %
Min Pattern Coverage	90 %
Use Dynamic Recalibration	True
5.3. Fragment Matching	
Use Fragment Matching	True
Mass Tolerance	5 ppm
S/N Threshold	3
6. General Settings	
Isotope Patterns	Cl; Br; S
Mass Tolerance	5 ppm
Intensity Tolerance	30 %
S/N Threshold	3
Min Spectral Fit	0 %

44 **Table S2.** List of NDMA precursors included in the suspect screening, indicating their recovery
45 percentage after PPL-SPE extraction and their NDMA-transformation rate (according to Farré et
46 al. 2016)

	Compound	Class	Recovery rate (%)	NDMA Transformation rate (%)
1	Azithromycin	Antibiotic (macrolide)	107	0.14±0.01
2	Clarithromycin	Antibiotic (macrolide)	118	0.13±0.02
3	Erythromycin	Antibiotic (macrolide)	106	0.059±0.007
4	Roxithromycin	Antibiotic (macrolide)	92.9	0.113±0.001
5	Spiramycin	Antibiotic (macrolide)	58.8	2.6±0.5
6	Tylosin	Antibiotic (macrolide)		0.200±0.006
7	Tetracycline	Antibiotic (tetracycline)	90.7	1.6±0.2
8	Chlorotetracycline	Antibiotic (tetracycline)	96.6	1.7±0.2
9	Oxytetracycline	Antibiotic (tetracycline)	97.3	1.104±0.006
10	Doxycycline	Antibiotic (tetracycline)	90.2	1.3±0.4
11	Citalopram	Antidepressant	69.7	0.31±0.04
12	Ranitidine	Antiacid drug	61.8	50±2
13	Venlafaxine	Antidepressant	115	0.53±0.01
14	*N*-desmethylvenlafaxine	Antidepressant (transformation product)	119	0.025±0.005
15	*O*-desmethylvenlafaxine	Antidepressant	119	1.19±0.02

47

Table S3. Tentative identification of unknowns that correlate linearly with NDMA-FP.

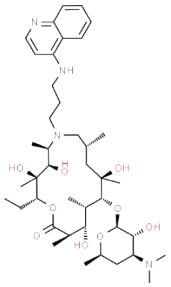
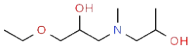
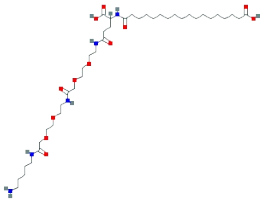
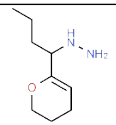
	$[M+H]^+$	Tentative molecular formula	Tentative structure	InChIKey	FISh Score	Normalised MetFrag score
1	760.5048	$C_{41}H_{68}N_4O_9$		LZFSZAPEFRELPV-RQLWGDOHSA-N	26.57	1.0
2	192.1596	$C_9H_{21}NO_3$		WSNUPCBHNGDUHE-UHFFFAOYSA-N	33.33	0.9937
3	818.5466	$C_{40}H_{75}N_5O_{12}$		YLXOZZYIEPNANX-UHFFFAOYSA-N	N/A	1.0
4	171.1493	$C_9H_{18}N_2O$		ATHWGISAQYNYIX-UHFFFAOYSA-N	40.24	1.0

Table S3b. Tentative identification of unknowns that correlate linearly with NDMA-FP.

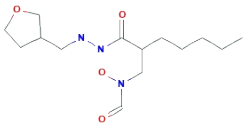
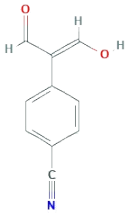
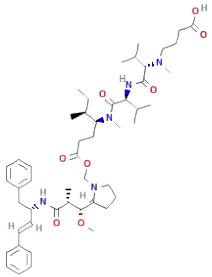
5	302.2077	$C_{14}H_{27}N_3O_4$		MMSUKIWOSVRIHW-UHFFFAOYSA-N	N/A	1.0
6	174.0549	$C_{10}H_7NO_2$		IPSUJMNCUUPWRX-POHAHGRESA-N	N/A	1.0
7	876.5886	$C_{50}H_{77}N_5O_8$		LHRNBHOKOXBXBK-QUKMVLLYSA-N	N/A	0.9827

Table S3c. Tentative identification of unknowns that correlate linearly with NDMA-FP.

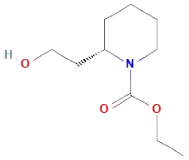
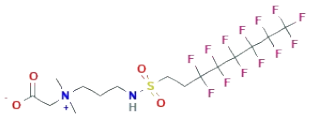
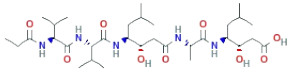
8	202.1437	$C_{10}H_{19}NO_3$		SEKRRKCEGKJUHF-VIFPVBQESA-N	N/A	1.0
9	571.0925	$C_{15}H_{19}F_{13}N_2O_4S$		OKOCIUJVPQKDLL-UHFFFAOYSA-N	N/A	0.7982
10	679.473	Not assessed	No MS^2 spectra was recorded for this peak. Tentative identification was not possible.			
11	658.4370	$C_{32}H_{59}N_5O_9$		HUOUXPWOUNLCOX-IWIWXMQLSA-N	36.72	0.45

Table S4. Confusion matrixes obtained for *k*-nn models after a 10-fold cross validation.

k	True Positives (%)	True Negatives (%)	False Positives (%)	False Negatives (%)	Accuracy (%)	MCC	F_1	FOR (%)
1	57 ± 11	38 ± 18	3.3 ± 7.0	1.7 ± 5.3	95 ± 8	0.897	0.958	5.0 ± 16
2	50 ± 24	40 ± 16	1.7 ± 5.3	8.3 ± 14	90 ± 14	0.806	0.909	11 ± 17
3	57 ± 16	38 ± 18	3.3 ± 7.0	1.7 ± 5.2	95 ± 8	0.897	0.958	2.5 ± 8
4	53 ± 13	38 ± 16	3.3 ± 4.0	5.0 ± 8.1	92 ± 9	0.830	0.928	11 ± 18
5	52 ± 18	35 ± 12	6.7 ± 8.6	6.7 ± 11	87 ± 13	0.726	0.886	12 ± 20
6	50 ± 19	37 ± 17	3.3 ± 7.0	10 ± 14	87 ± 17	0.736	0.882	20 ± 32
7	52 ± 20	40 ± 12	1.7 ± 5.3	40 ± 12	92± 12	0.836	0.925	9.8 ± 16
8	43 ± 16	32 ± 15	10 ± 14	17 ± 18	73 ± 18	0.468	0.758	28 ± 29
9	48 ± 17	27 ± 16	15 ± 23	8.3 ± 14	75 ± 23	0.509	0.806	16 ± 25
10	50 ± 20	23 ± 16	16 ± 22	15 ± 15	73 ± 18	0.356	0.759	38 ± 35