UNIVERSITÉ DE GENÈVE

LSMS

Integrative data science in mass spectrometry: connecting molecular structures with diverse MS behaviors

Stepan Stepanovic and Gérard Hopfgartner

stepanstepanovic.github.io

Life Science Mass Spectrometry, Geneva

Life Sciences Mass Spectrometry, Department of Inorganic and Analytical Chemistry, University of Geneva
24 Quai Ernest Ansermet, CH-1211 Geneva 4, Switzerland

Outline :

1. From Separation to Structure: Expanding Mass Spectrometry Capabilities from metabolites to peptides

2. Data-Driven Insights into Differential Ion Mobility (DMS): Selectivity & Signal Suppression

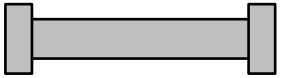
3. Machine Learning in Fragmentation: Predicting CID, EAD & UVPD Behavior

Ion Mobility & DMS Introduction

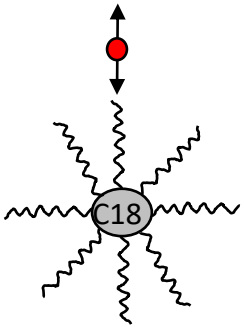
1. Time separation

Liquid phase

(*min*)



Mobile phase



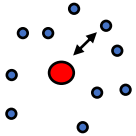
logD, pKa, …

4. Mobility

Gas phase (1 atm.)

(*ms*)

DTIMS
TWIMS
FAIMS
DMS

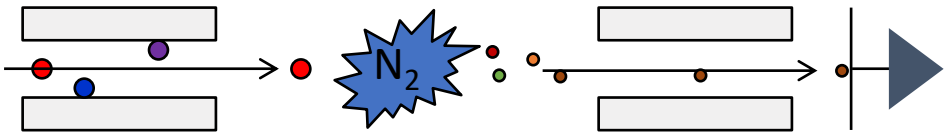


CSS
Shape-to-charge ratio

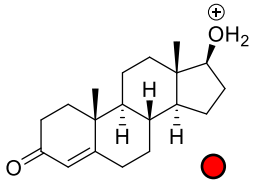
2. Mass separation (*m/z*) 3. Fragment separation (*m/z*)

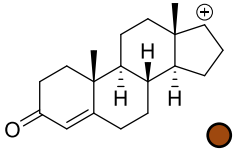
Gas phase (high vacuum)

(*m/z*) (*m/z*)



Intact molecule → Fragments

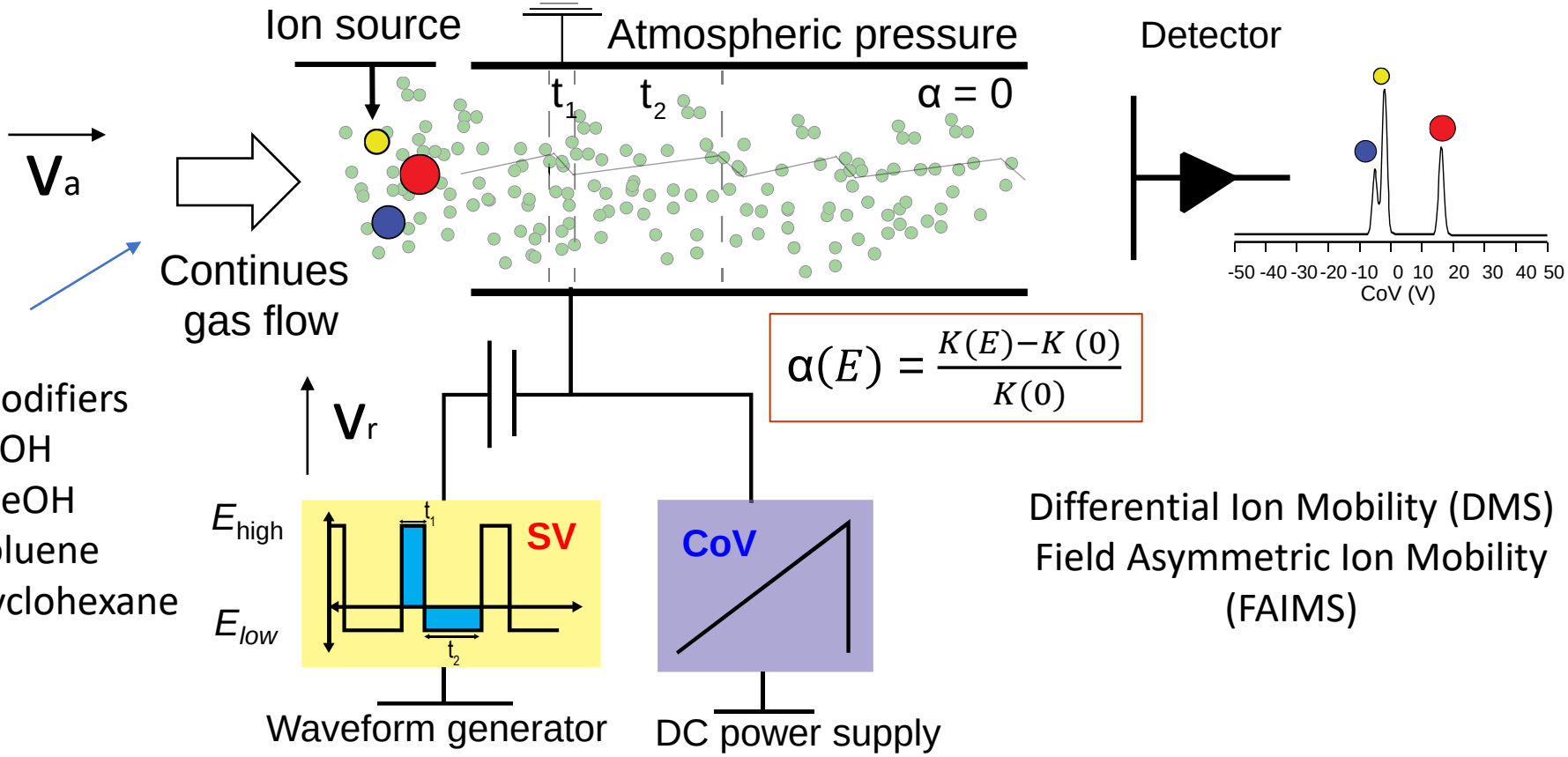




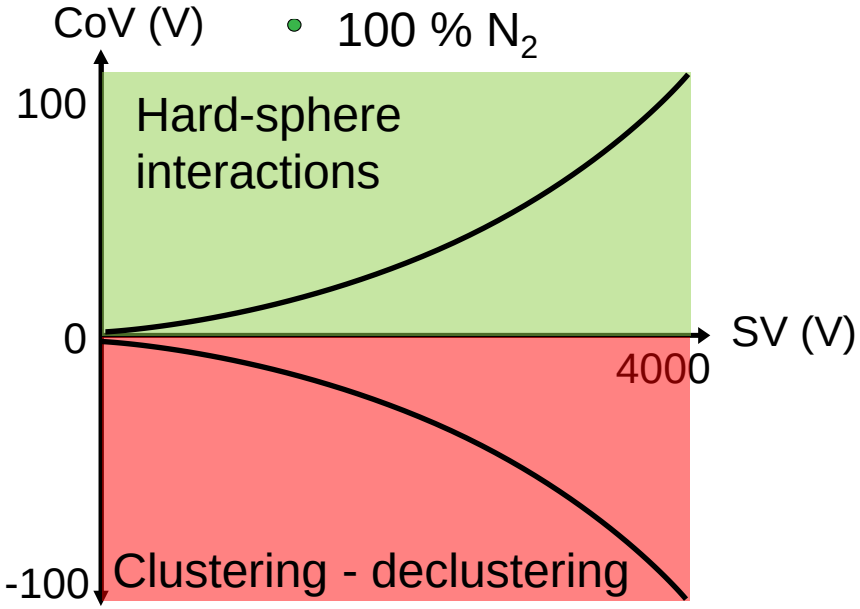
Chemical Formula:
C19H29O2
m/z 289.2162

Chemical structure

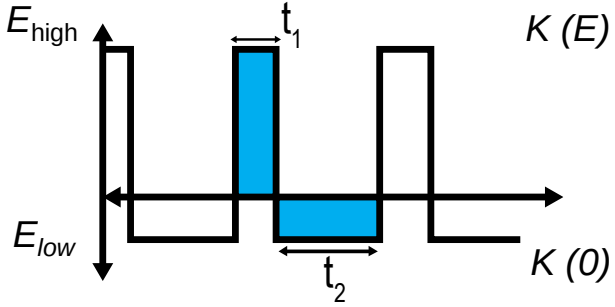
DMS Introduction

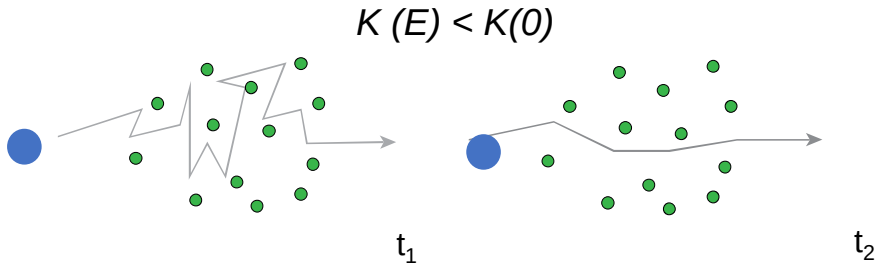


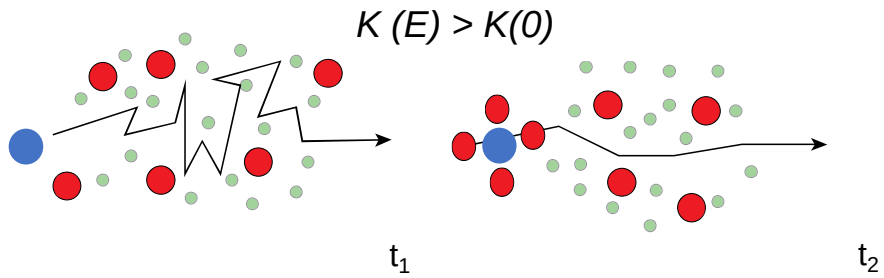
DMS & Modifier Chemistry



Modifier – polar organic solvent
1.5 – 3 % in the N_2

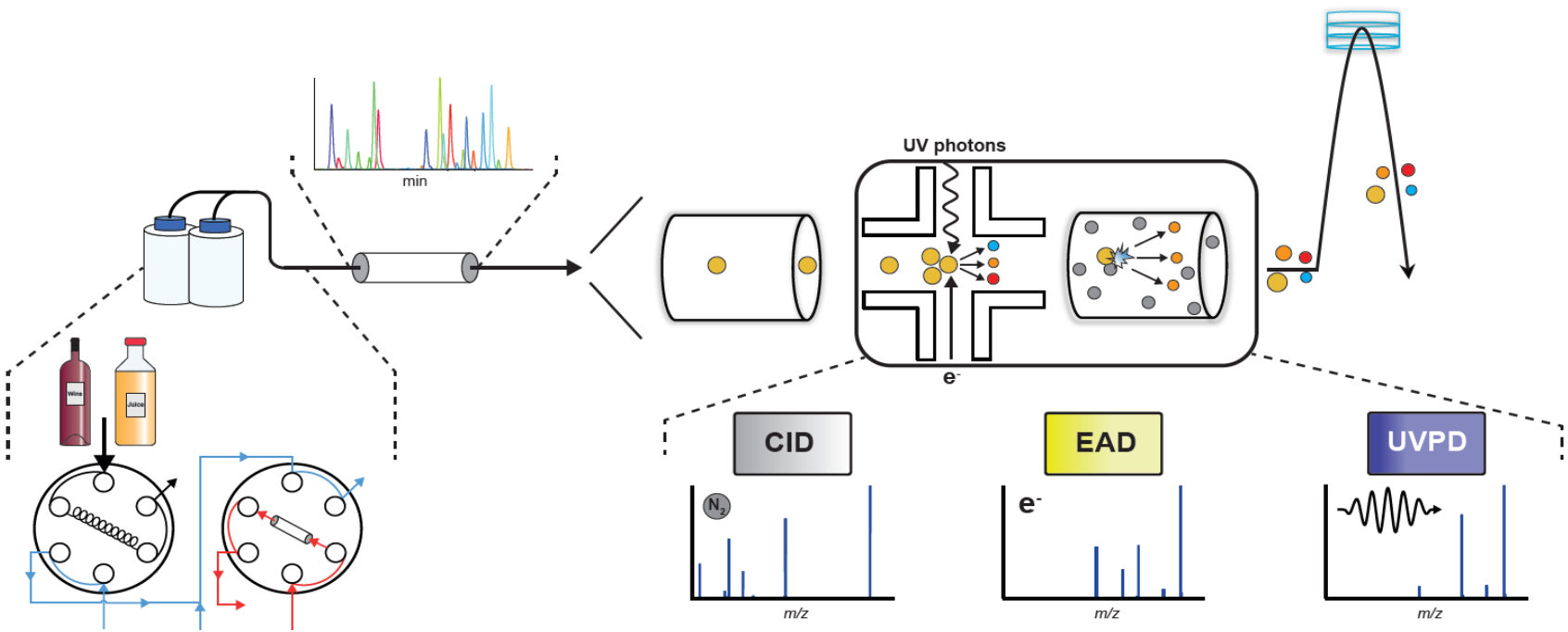






Fragmentation Techniques

(gas, electrons and photons)

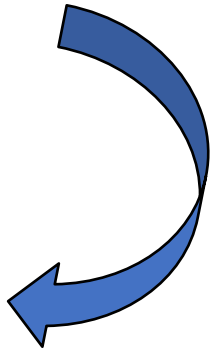


Data-Driven Insights into Differential Ion Mobility (DMS):

Nr.	Mix 50	Formula	RT (min)	CoV (V)						
				SV 3800 V					SV 4000 V	SV 3800 V
				N_2	1.5% Ch	1.5% IPA	0.05% IPA	1.5% EtOH	1.5% Tol	1.5% ACN
1	L-lysine	$C_6H_{14}N_2O_2$	1.09	4	3		-31			-39
2	L-Histidine	$C_6H_9N_3O_2$	1.12	1	-1		-38			-46
3	Carnosine	$C_9H_{14}N_4O_3$	1.13	7	6		-16	-41	-20	-29
4	1-methylhistidine	$C_7H_{11}N_3O_2$	1.14	5	3		-2	-42	-7	
5	Glycerophosphocholine	$C_8H_{20}NO_6P$	1.18	9	8	-21	6	-7	3	-31
6	Homo-L-arginine	$C_7H_{16}N_4O_2$	1.19	7	5		-9	-40	-15	-37
7	L-Glutamine	$C_5H_{10}N_2O_3$	1.19	-1	-3		4		8	
8	L-carnitine	$C_7H_{15}NO_3$	1.21	8	6	-39	3	-19	-4	
9	N-acetylneuraminic acid	$C_{11}H_{19}NO_9$	1.25	7	6	-49	-11	-39	-5	-25
10	Creatinine	$C_4H_7N_3O$	1.25	-10	7		5		3	
11	Trigonelline	$C_7H_7NO_2$	1.26	2	-1	-10	-21	-2	-22	
12	Creatine	$C_4H_9N_3O_2$	1.29	-1	-3		4			
13	L-Proline	$C_5H_9NO_2$	1.30	-10	-11					
14	Homocitrulline	$C_7H_{15}N_3O_3$	1.34	4	2	-12	-31	-2	-42	-43
15	N-acetylputrescine	$C_6H_{14}N_2O$	1.34	3	1	-50	-37			-46
16	L-acetylcarnitine	$C_9H_{17}NO_4$	1.40	10	8	-22	6	-8	3	-39
17	4-guanidinobutanoic acid	$C_5H_{11}N_3O_2$	1.69	3	1		10			-47
18	3-methyladenine	$C_6H_7N_5$	1.76	2	1	-41	0	-41	-7	-30
19	Urocanic acid	$C_6H_6N_2O_2$	1.85	-2	-2	-44	4	-50		-19
20	7-methylguanine	$C_6H_7N_5O$	2.76	1	-1	-32	-10	-25	1	
21	Niacinamide	$C_6H_6N_2O$	2.84	-6	-7	-41				
22	Tyramine	$C_8H_{11}NO$	3.01	-4	-5	-42	-34	-42		-27

50 low molecular weight compounds

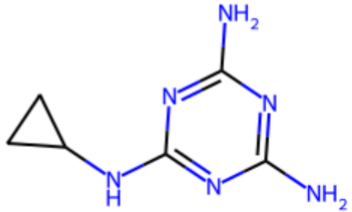
```
import pubchempy as pcp
pcp.get_compounds('L-lysine','name')[0]
```

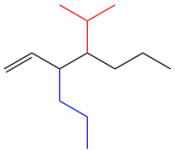


	N2	Ch	IPA	Ch_IPA	EtOH	Tol	ACN	pcp	name	mol	NaN	Gb	MW	logP	RT	logp	TPSA	Ar
0	4	3	NaN	-31.0	NaN	NaN	-39.0	Compound(5962)	L-lysine	<rdkit.Chem.rdchem.Mol object at 0x0000012015A970B0>	3	-34.29	146.190	-0.47270	1.09	-3.0	89.34	0
1	1	-1	NaN	-38.0	NaN	NaN	-46.0	Compound(6274)	L-Histidine	<rdkit.Chem.rdchem.Mol object at 0x0000012015A97150>	3	-31.89	155.157	-0.63590	1.12	-3.2	92.00	1
2	7	6	NaN	-16.0	-41.0	-20.0	-29.0	Compound(439224)	Carnosine	<rdkit.Chem.rdchem.Mol object at 0x0000012015A971F0>	1	-36.07	226.236	-1.12960	1.13	-4.0	121.10	1
3	5	3	NaN	-2.0	-42.0	-7.0	NaN	Compound(92105)	1-methylhistidine	<rdkit.Chem.rdchem.Mol object at 0x0000012015A97290>	2	-28.59	169.184	-0.62550	1.14	-3.3	81.14	1
4	9	8	-21.0	6.0	-7.0	3.0	-31.0	Compound(657272)	Glycerophosphocholine	<rdkit.Chem.rdchem.Mol object at 0x0000012015A97330>	0	-27.83	257.223	-1.45260	1.18	-2.3	99.05	0

Where from mol?

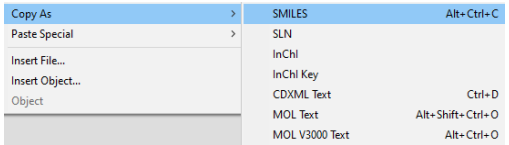
```
comp=pcp.get_compounds('Cyromazine','name')[0]
MolFromSmiles(comp.canonical_smiles)
```

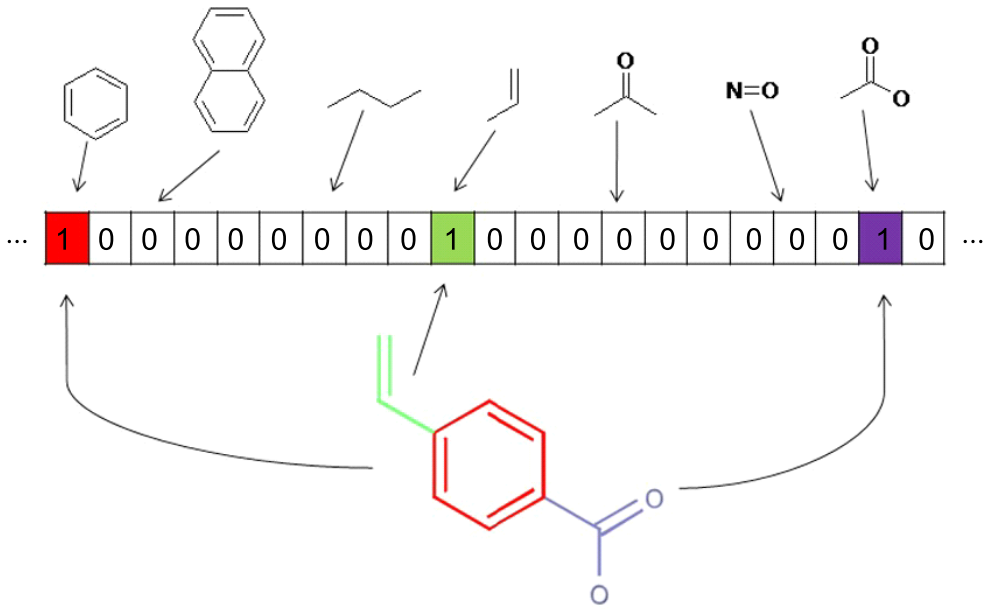




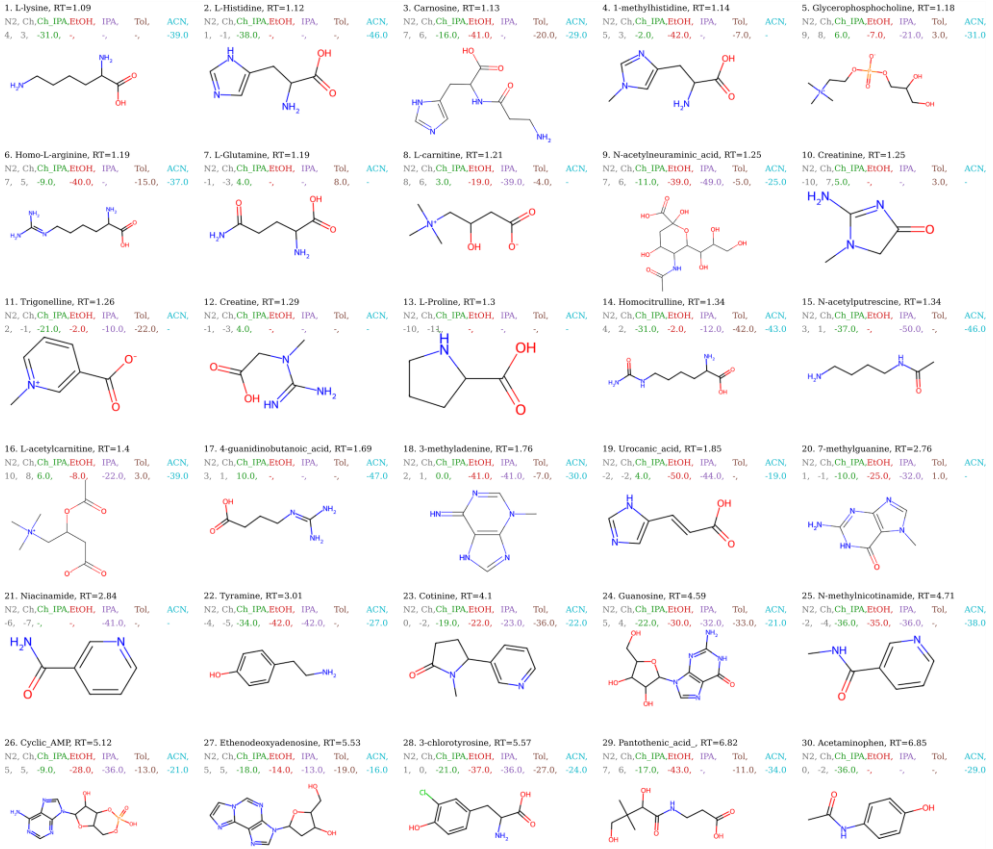
→ C=CC(CCC)C(C(C)C)CCC

How easiest to obtain SMILES?
Draw a structure in ChemDraw:

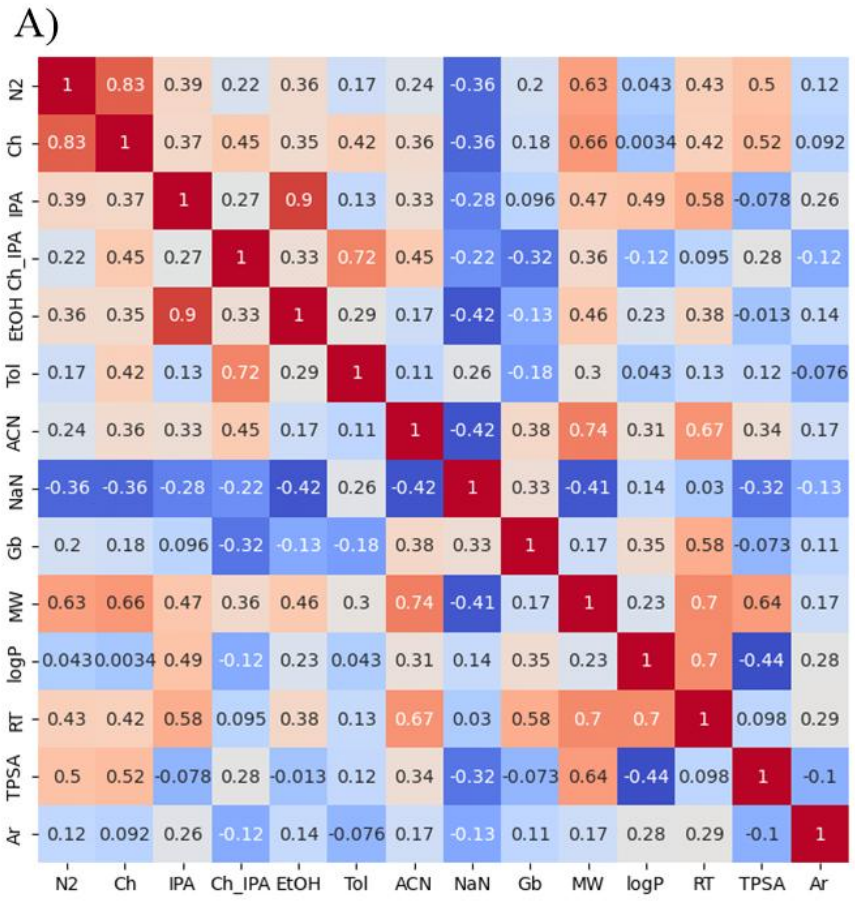


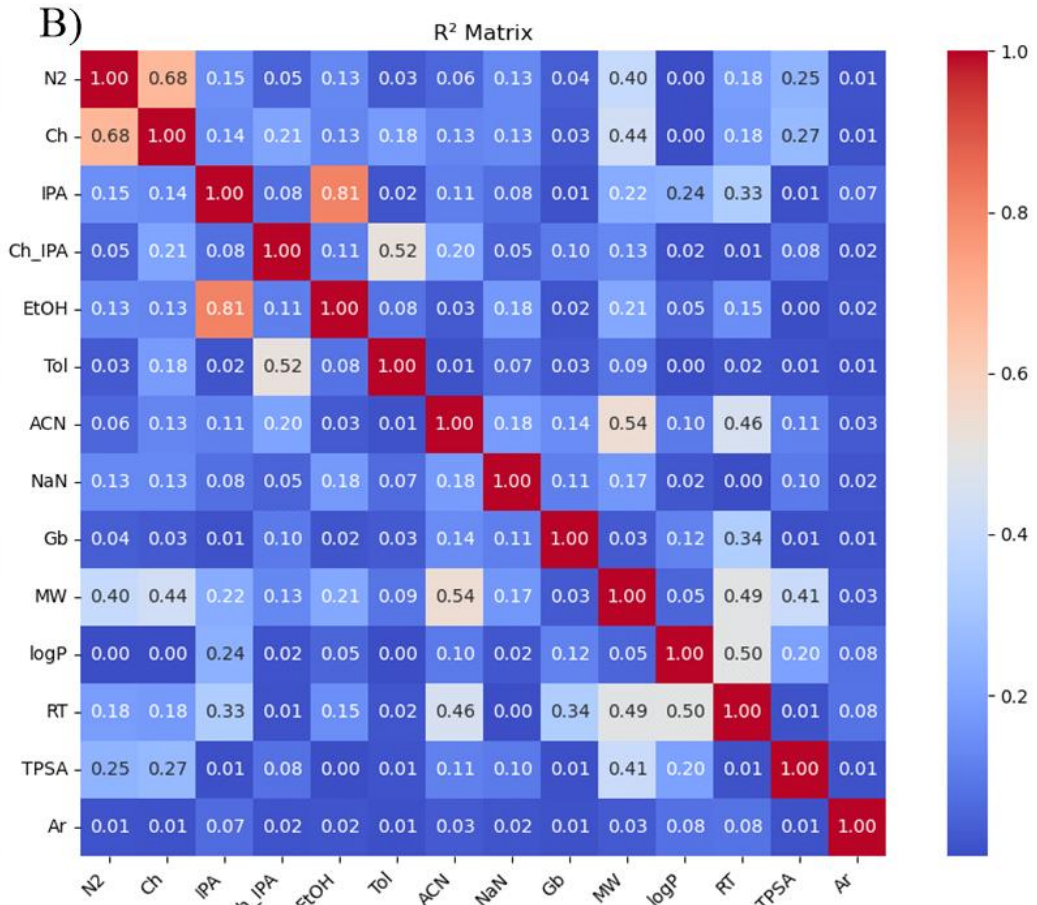


Data-Driven Insights into Differential Ion Mobility (DMS):



Data-Driven Insights into Differential Ion Mobility (DMS):





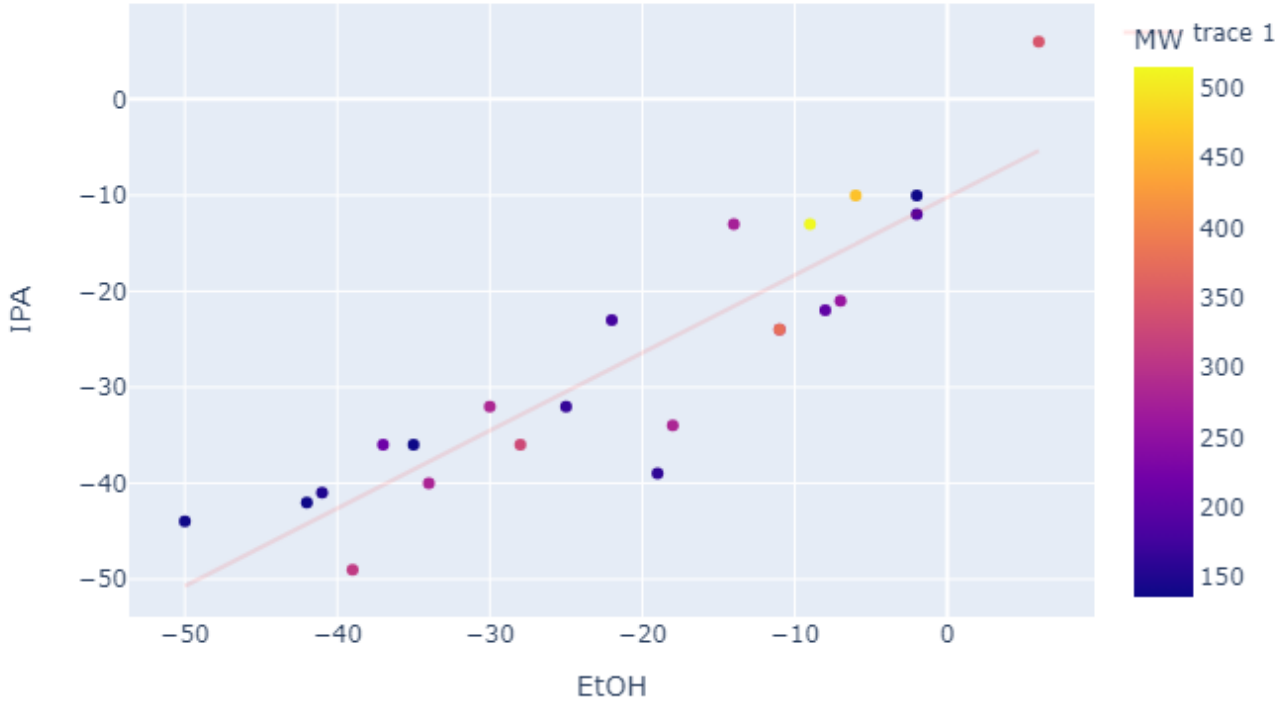
Data-Driven Insights into Differential Ion Mobility (DMS):

- CoV (EtOH) and CoV(IPA) with R^2=0.81
- CoV between non-clustering modifier Cyclohexane and N_2 with R^2=0.68
- MW with CoV(N_2), CoV(cyclohexane), CoV(ACN), RT and TPSA
- RT with logP, MW, CoV(ACN), proton affinity(Gb) and CoV(IPA)

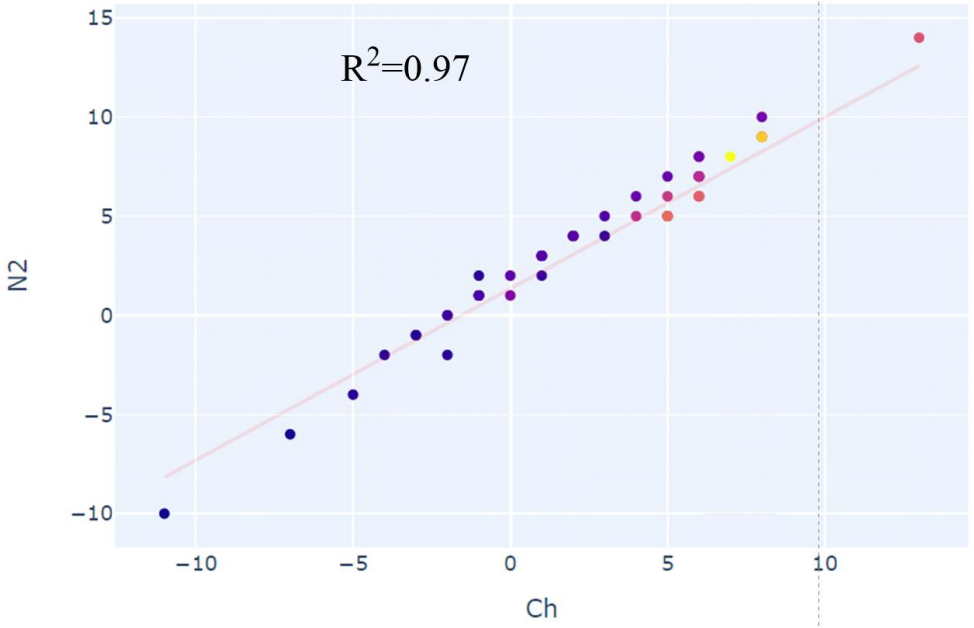
$$RT = 1.72 \cdot logP + 0.035 \cdot MW + 0.198 \cdot Gb + 3.95, \qquad R^2 = 0.9$$

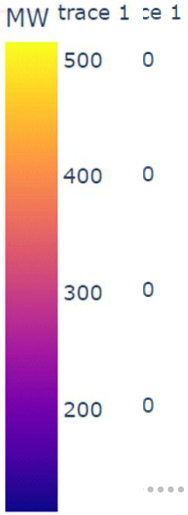
$$RT = 1.64 \cdot logP + 0.042 \cdot MW - 0.004 \cdot CoV(IPA) - 4.1, \qquad R^2 = 0.92$$

Data-Driven Insights into modifier assisted Differential Ion Mobility (DMS)

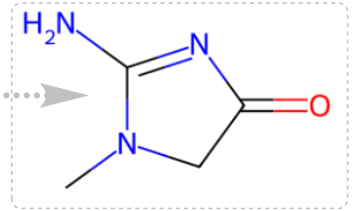


Data-Driven Insights into Differential Ion Mobility (mDMS):

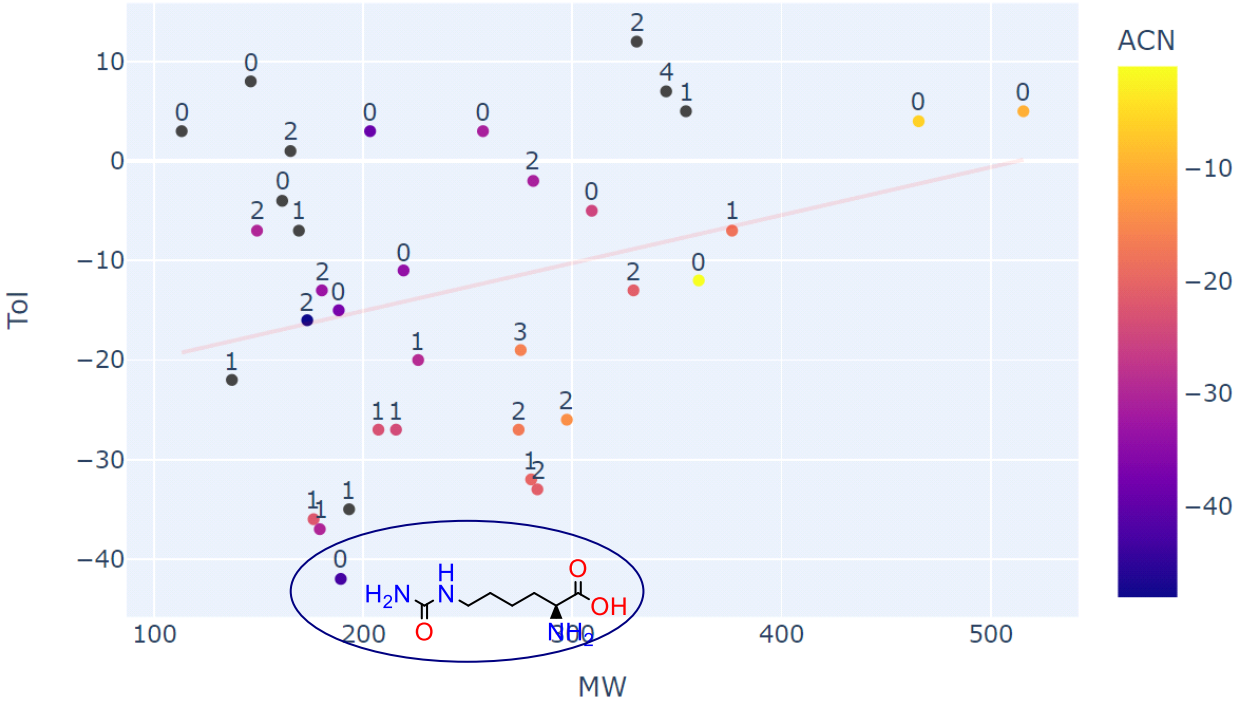




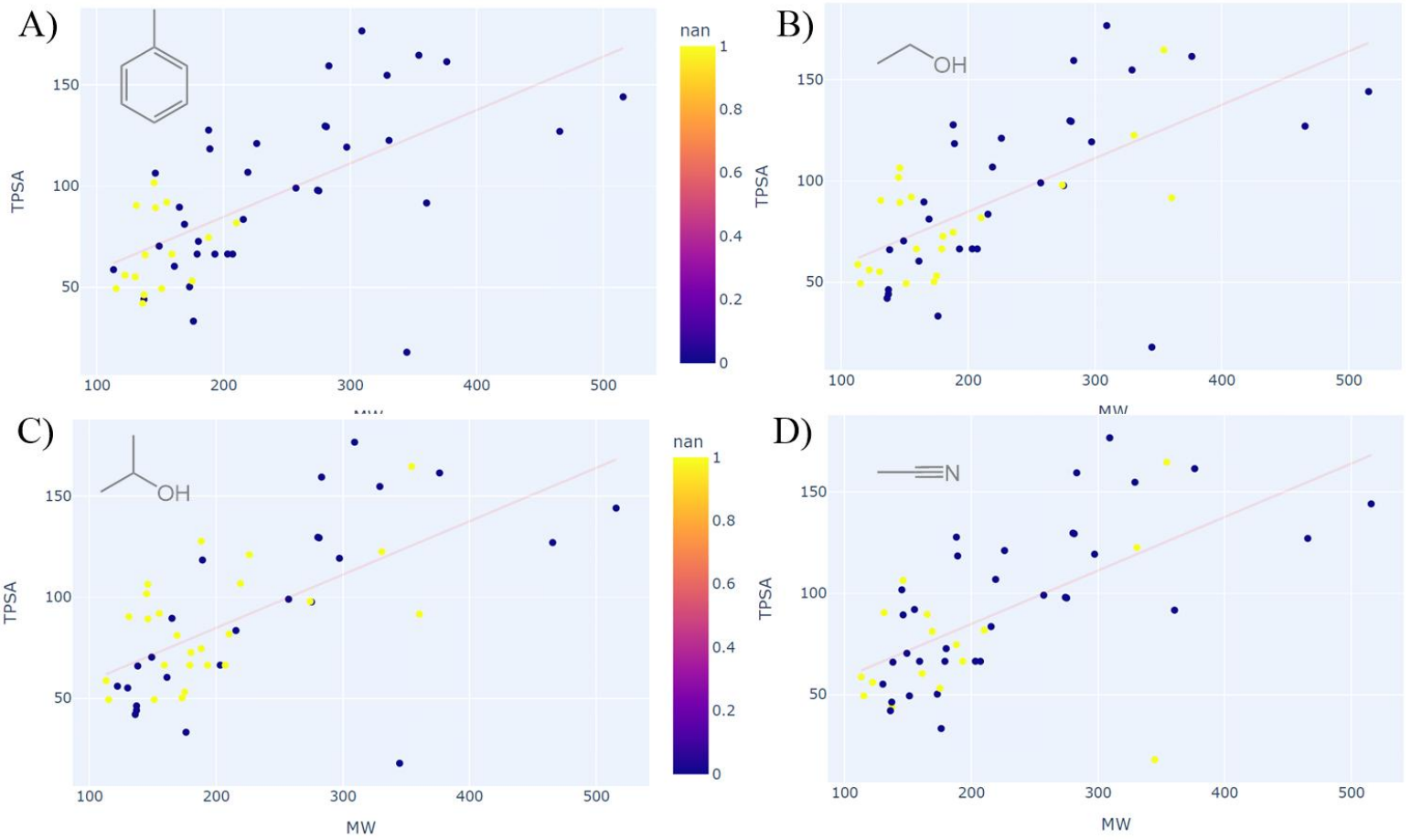
Creatinine



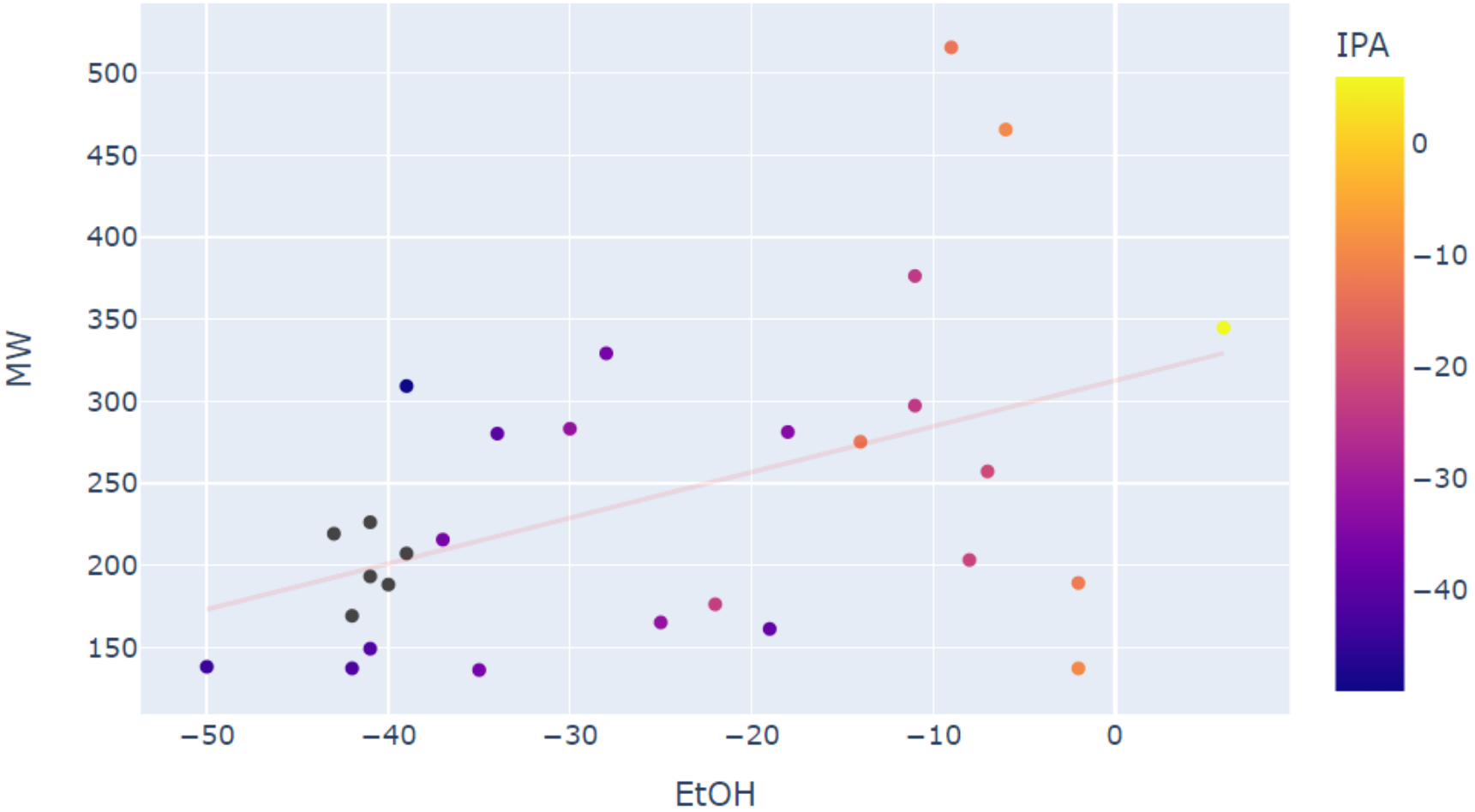
Data-Driven Insights into Differential Ion Mobility (DMS):



Data-Driven Insights into Differential Ion Mobility (DMS):



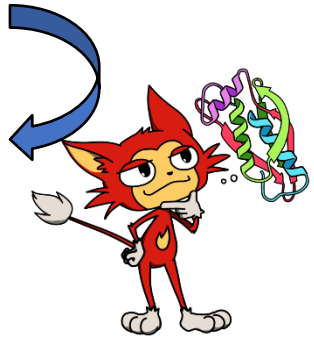
Data-Driven Insights into Differential Ion Mobility (DMS):



Signal Suppression for Peptides

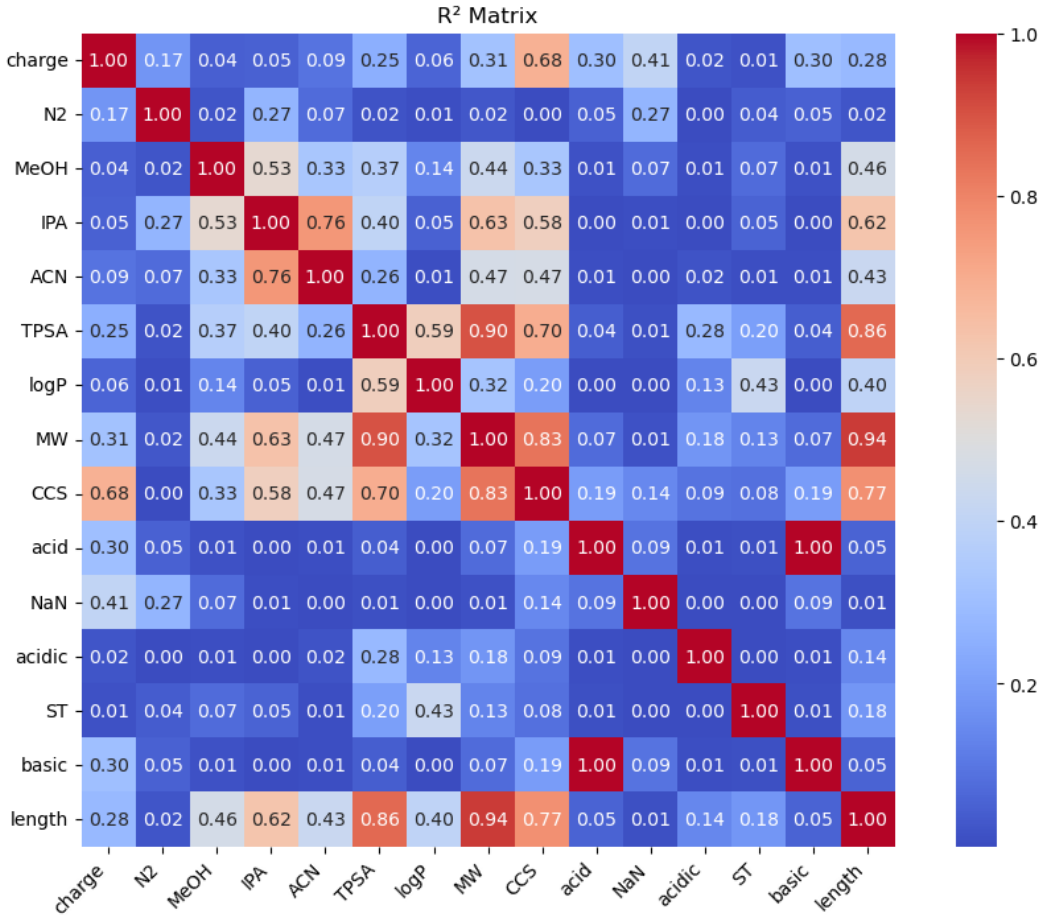
185 peptides with 6-20 AAs

Peptide sequence	charge	N2	MeOH	IPA	ACN	TPSA	logP	MW	css	acid	NaN	ST	basic	acidic	length
VCNQIEFLNTEFK	2	9.4	6.1	-4.6	-0.6	666	-7	1583.8	421	1	0	1	1	2	13
QAEELIQQEHADQAEIR	3	14.7				983	-14	2007.0	544	2	1	0	2	5	17
DNFDIAEGVR	2	11.0	3.8	-17.1	-7.2	545	-8	1134.5	348	1	0	0	1	3	10
VNYNFEDETVR	2	10.9	4.4	-11.3	-4.2	658	-9	1384.6	386	1	0	1	1	3	11
HGFLEGR	2	18.9	4.8	-13.3	-4.6	370	-6	814.4	314	2	0	0	2	1	7



name	charge	N2	MeOH	IPA	ACN	mol	TPSA	logP	MW	CCS	acid	NaN	acidic	ST	basic	length	m.z
VCNQIEFLNTEFK	2	9.38	6.09	-4.57	-0.61	<rdkit.Chem.rdchem.Mol object at 0x0000012015B4E200>	665.88	-6.7986	1583.765377	421.225525	1	0	2	1	1	13	3167.530754
ELIQQEHADQAEIR	3	14.66	NaN	NaN	NaN	<rdkit.Chem.rdchem.Mol object at 0x0000012015B4E2A0>	982.97	-13.5393	2006.965742	543.888428	2	1	5	0	2	17	6020.897226
DNFDIAEGVR	2	10.96	3.80	-17.14	-7.22	<rdkit.Chem.rdchem.Mol object at 0x0000012015B4E340>	545.47	-8.2260	1134.530550	347.680145	1	0	3	0	1	10	2269.061100
VNYNFEDETVR	2	10.89	4.40	-11.32	-4.20	<rdkit.Chem.rdchem.Mol object at 0x0000012015B4E3E0>	658.12	-9.1880	1384.625907	386.323029	1	0	3	1	1	11	2769.251814
HGFLEGR	2	18.89	4.75	-13.30	-4.63	<rdkit.Chem.rdchem.Mol object at 0x0000012015B4E480>	370.41	-6.1648	814.408585	313.631165	2	0	1	0	2	7	1628.817170

Signal Suppression for Peptides



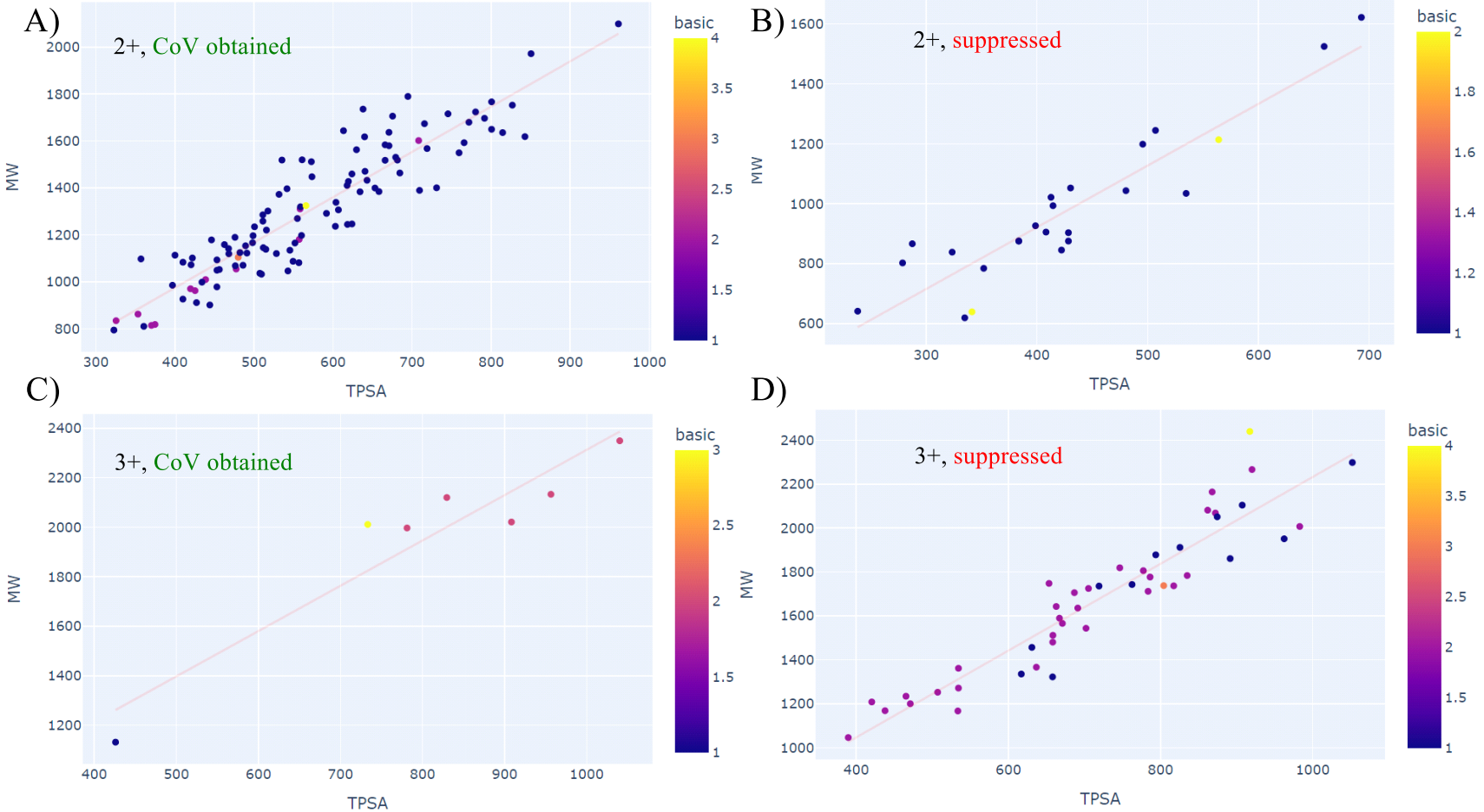
- CSS, MW, charge, TPSA and peptide length are all intercorrelated
- MW is very nicely correlated with the CoV for all polar modifiers
- $R^2 \approx 0$ for CoV(N_2) vs MW/CCS/ peptide length
- CoV(IPA) is correlated with both MeOH(R^2=0.53) and ACN (R^2=0.76)
- The absence of CoV values is nicely correlated with charge (R^2=0.41)

Signal Suppression for Peptides

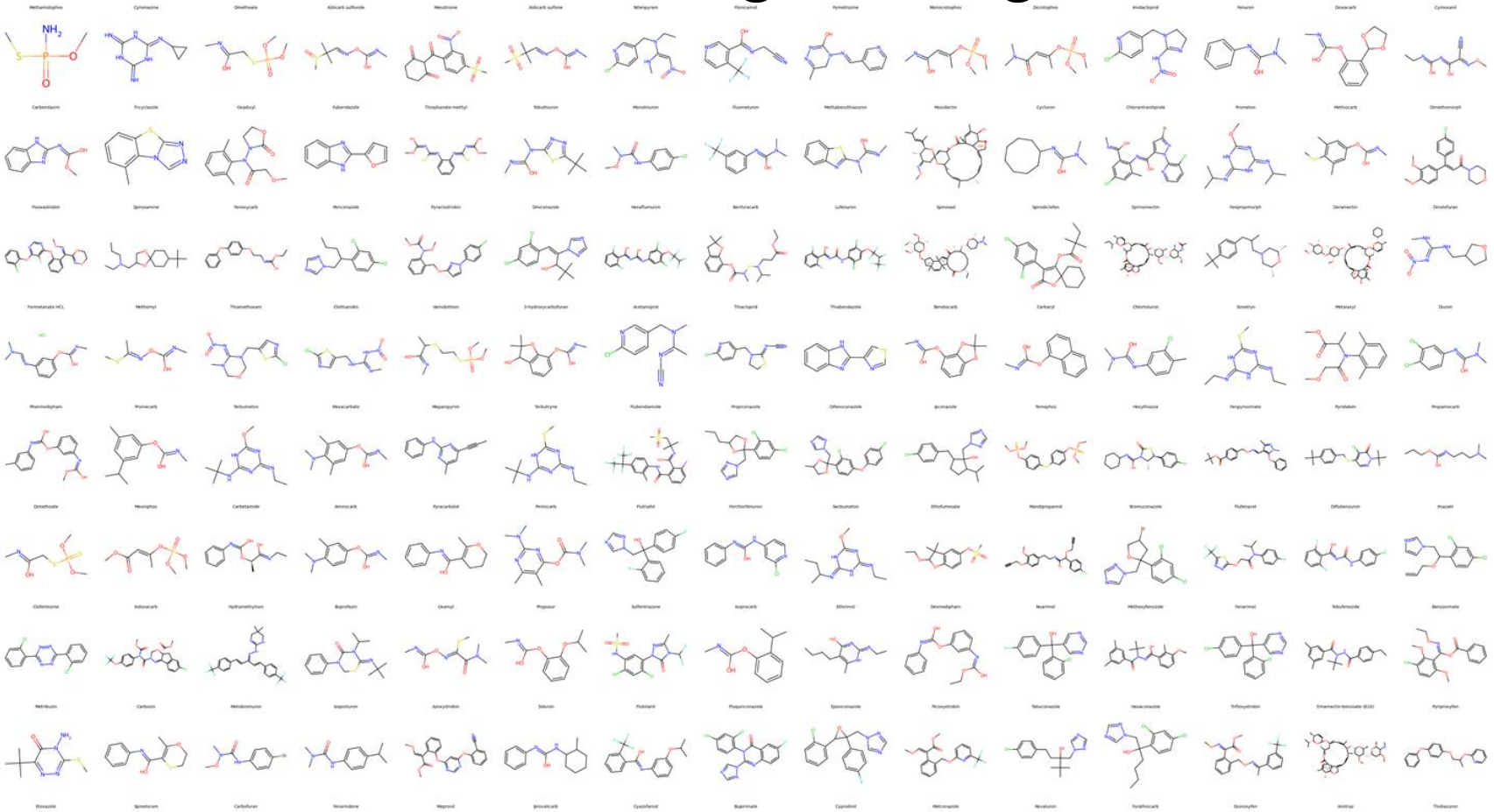
CoV ranges: N_2(21.72V to 7.19V), MeOH (9.51V to -6.93V), IPA (-1.28V to -23.63V) and ACN (1.69V to -10.81V)

- The features used are: peptide charge, CoV(N_2), TPSA, MW, number of basic AAs(base), number of Ser + Thr residuer (ST), number of acidic sidechains (acid), and length of peptide.
- Scaled and 5-fold cross validation
- Grid hyperparameter search for Support Vector, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression and Ridge classifiers is performed and compaired with “Logistic Regression CV”
- Accuracy of 0.89, precision of 0.93, recall of 0.81, F1 score of 0.87, and an AUC-ROC of 0.90.
- Feature importance: coefficient analysis, permutation feature importance, SHAP Values (peptide charge and three strongly related features describing peptide size: MW, TPSA and peptide length) and recursive feature elimination (RFE) - basic amino acids count.

Signal Suppression for Peptides



Machine Learning in Fragmentation



Machine Learning in Fragmentation

```
((5310, 10),
CID                          2832
EAD                          2301
UVPD simultaneous trapping    177
```

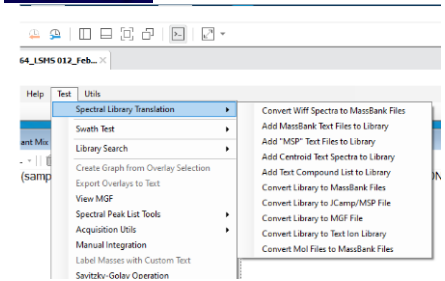
.wiff or .wiff2 → mzML

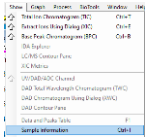
```
In [418]: romain
```

Out[418]:

	Source	name	Fragmentation mode	method	spectra	Retention	precursor	new_data	mol	prec
0	27	Methamidophos	EAD	0	17	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
1	27	Methamidophos	EAD	3	19	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
2	27	Methamidophos	EAD	5	40	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
3	27	Methamidophos	EAD	7	38	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
4	27	Methamidophos	EAD	9	36	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
5	27	Methamidophos	EAD	11	29	2.6	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3
6	27	Methamidophos	EAD	13	39	2.61	142.0	4	<rdkit.Chem.rdchem.Mol object at 0x000002409B512520>	3

PV_2024_Lyle





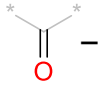
```
Compound ID, Group Name, Precursor Ion (Da), TOF Start Mass (Da), TOF Stop Mass (Da), Accumulation Time (sec), DP (V), CE (V), CE Spread (V), Retentio
Methamidophos 0, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 0.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 3, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 3.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 5, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 5.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 7, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 7.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 9, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 9.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 11, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 11.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 13, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 13.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 15, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 15.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 17, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 17.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 19, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 19.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 21, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 21.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 23, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 23.000, 100.00, 150.00, 20.00, 4, True, True, True, True
Methamidophos 25, Group 1, 142.00, 50.00, 160.00, 0.06, 80.00, 10.00, 0.000, 2.61, 8, EAD, 25.000, 100.00, 150.00, 20.00, 4, True, True, True, True
```

```
Filename          Experiment      CH$NAME:          CH$FORMULA:     AC$MASS_SPECTROMETRY: FRAGMENTATION_TYPE
MS24GE_20230331_RGIR_A027         '2-14   Methamidophos   C2H8NO2PS       EAD
MS24GE_20230331_RGIR_A027         15-27   Cycromazine     C6H10N6 EAD
MS24GE_20230331_RGIR_A027         28-40   Omethoate       C5H12NO4PS      EAD
MS24GE_20230331_RGIR_A027         41-53   Aldicarbsulfoxide       C7H14N2O3S      EAD
MS24GE_20230331_RGIR_A027         54-66   Mesotrione      C14H13NO7S      EAD
MS24GE_20230331_RGIR_A063         2-17    Methamidophos   C2H8NO2PS       CID
MS24GE_20230331_RGIR_A063         18-33   Cycromazine     C6H10N6 CID
MS24GE_20230331_RGIR_A063         34-49   Omethoate       C5H12NO4PS      CID
MS24GE_20230331_RGIR_A063         50-65   Aldicarbsulfoxide       C7H14N2O3S      CID
MS24GE_20230331_RGIR_A063         66-81   Mesotrione      C14H13NO7S      CID
```

Machine Learning in Fragmentation

When compared UVPD spectra and CID=15eV spectra, unique UVPD peaks are detected in 101 analytes while all peaks are also appearing in CID for 76 compounds

$$z = -0.32 - 1.16 \cdot$$ 

$$+1.36 \cdot$$ 

$$- 0.74 \cdot$$ 
$$+0.99 \cdot$$ 

X= N, O
$$- 0.61 \cdot$$ 

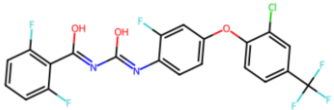
$$+1.13 \cdot$$ 

Accuracy, Recall, Precision, F1 Score, ROC AUC ~ 0.71-0.75

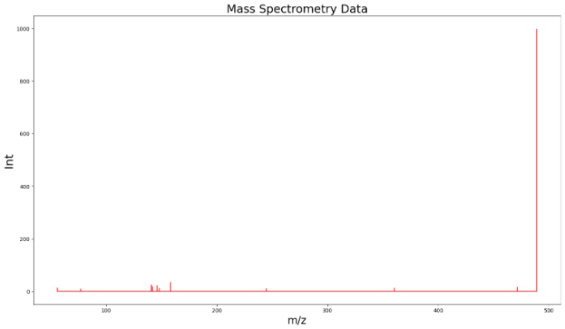
When compared EAD vs combined CID spectra, we obtain redox active groups like Ar-Cl, OH, C=O…

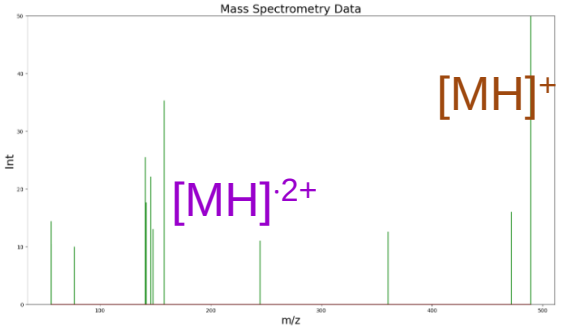
Machine Learning in Fragmentation

Flufenoxuron

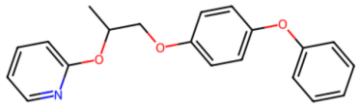


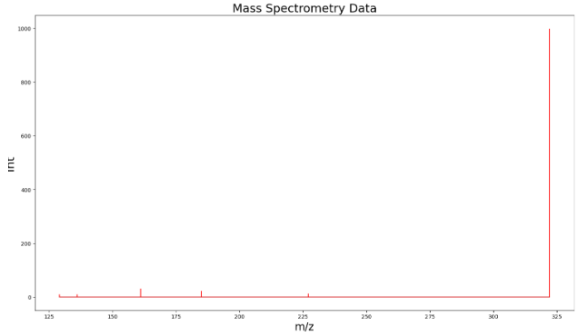
```
 55.93319674,  14.42087619
 56.01057416,  10.44082037
 77.03664484,  10.01075867
141.01435327,  25.5049377
141.92585956,  17.62305146
145.96692073,  22.1430342
147.96399259,  13.0421924
158.03997604,  35.31088094
244.5186681 ,  10.9993784
360.17573899,  12.58667325
471.402279  ,  16.04845947
489.0432555 , 999.
```

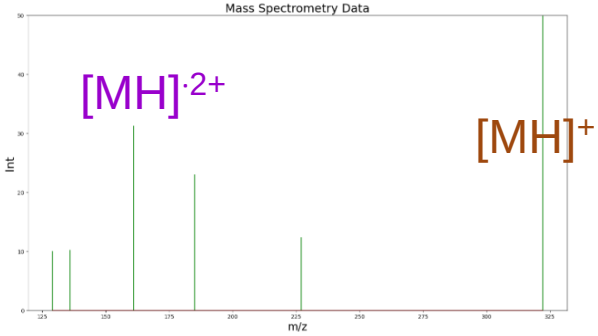




Pyriproxyfen








```
129.06838176,  10.01776756
136.07399316,  10.26311161
161.07095407,  31.26317211
185.05879575,  23.03818201
227.10420166,  12.35915076
322.14341426, 999.
```

Machine Learning in Fragmentation

1) We need to obtain [MH]+ (protonate) all 170+ molecules

2) Optimize [MH]+ with some high level of theory, we can then extract HOMO.

3) Freq on [MH]+ and opt/freq on [MH]·2+ in order to get Gibbs free energy estimates

	HOMO	DG	DE	M	name	MH2+
0	-9.1235	253.395647	255.505021	238.26	3-Hydroxycarbofuran	0
1	-9.2668	264.387689	266.965166	223.69	Acetamiprid	0
2	-10.1526	275.187186	277.290134	211.28	Acibenzolar-S-methyl	0
3	-8.5537	228.327642	229.141453	400.53	Alanycarb	0
4	-10.4435	268.341313	273.507669	223.27	Aldicarb_sulfone	0
5	-9.2910	261.977048	264.077247	207.27	Aldicarb_sulfoxide	0
6	-9.5434	264.179670	266.852214	228.34	Ametryn	0
7	-7.8825	231.528035	232.122758	209.27	Aminocarb	0
8	-8.1596	216.596601	224.097324	294.42	Amitraz	1
9	-8.4604	233.461342	234.505801	404.40	Azoxystrobin	0

Model	Train Accuracy	Train Recall	Train Precision	Train F1 Score	Train ROC AUC	Test Accuracy	Test Recall	Test Precision	Test F1 Score	Test ROC AUC
Random Features	0.62	0.62	0.68	0.59	0.58	0.52	0.52	0.53	0.48	0.42
ΔG Feature	0.93	0.93	0.93	0.93	0.95	0.92	0.78	0.84	0.81	0.81
HOMO Feature	0.69	0.69	0.69	0.69	0.79	0.72	0.67	0.6	0.6	0.85

Conclusions

- ✓ Data science reveals key correlations between modifiers and molecular properties in DMS, identifying two major signal suppression mechanisms for small molecules.
- ✓ Machine learning links peptide properties to proton transfer-induced MS signal suppression, enhancing our understanding of DMS behavior and improving peptide analysis.
- ✓ Overlapping between CID, EAD, and UVPD fragmentation techniques for small molecule pesticides is clearly related to the FGs.
- ✓ There is a significant correlations between the chemistry of small molecules and their molecular MS/MS behavior:

- unique UVPD peaks ~ absorption chromophores
- unique EAD peaks ~ redox FGs
- observation of [MH]·2+ in EAD spectra ~ EAD energy and [MH]·2+ instability

Acknowledgments

David Ruskic
Eliane Kuehn
Charlotte Jacquet
Piotr Sosnowski
Lysi Ekmekciu
Patrick Mueller
Maria Girard
Xiaobo Tian
Quentin Rouchon
Gisela Gonzalez
Bandar Alghanem
Romain Giraud

Ron Bonner

FNSNF
Swiss National Science Foundation

SCIEX

Enhanced Pesticide Screening in Wines and Juices by Column-Switching Liquid Chromatography-Tandem Mass Spectrometry Using Multiple Activation Methods
, Authors: Romain Giraud, J.C. Yves Le Blanc, Mircea Guna, Gérard Hopfgartner
Version 1 posted 20 January 2025

Extract found to contain search terms

Yves Le Blanc2, Mircea Guna2, Gérard Hopfgartner1* 1Life Sciences Mass Spectrometry, Department of Inorganic and Analytical Chemistry, University of Geneva, 24 Quai Ernest Ansermet, CH-1205, Geneva 4, Switzerland 2SCIEX, Toronto, Ontario, Canada LC-MS/MS, column switching, pesticides, juice, wine, general screening, CID-EAD-UVPD ORCID: Romain Giraud: 0009-0006-8180-2874 ORCID: Gérard **Hopfgartner** : 0000-0002-9087-606X *Corresponding author: Gérard **Hopfgartner**