



INCHI ON WIKIPEDIA: WHY MANY COMPOUNDS HAVE MORE THAN ONE INCHI.

Roger Sayle

NextMove Software, Cambridge, UK



DISCLAIMER

- Firstly, thanks to the organizers; Steve, Marc, Evan, Noel and Janelle.
- Admission, NextMove Software makes relatively little use of InChI (but see John Mayfield's talk tomorrow).
- However, one role where I do use InChI is as a contributor/editor to Wikipedia, adding SMILES, InChI and InChI keys to ChemBox and DrugBox.
- Thanks to the developers of Open Babel for their graphical user interface, and Perkin Elmer for ChemDraw.



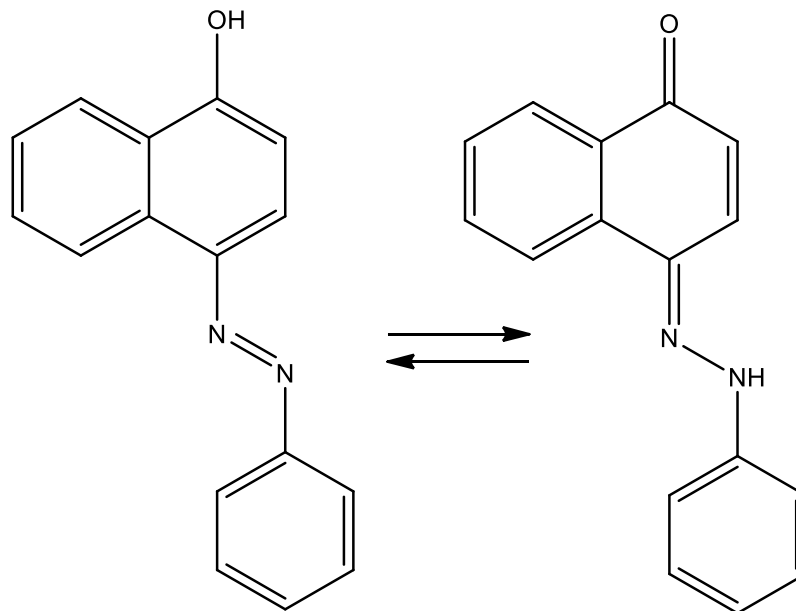
DISCLAIMER DISCLAIMER

- There are two types of errors for when assigning unique identifiers to chemical structures.
 1. When two (or more) distinct chemicals get mapped to the same identifier.
 2. When two (or more) representations of the same chemical get mapped to different identifiers.
- This presentation concerns type 2 errors.
- For the most part, InChI does a pretty good job.



CLASSIC TYPE 2 ERROR: TAUTOMERS

- The world's first tautomer [Laar 1885] contains a 1,7-shift not handled (by default) by standard InChI.



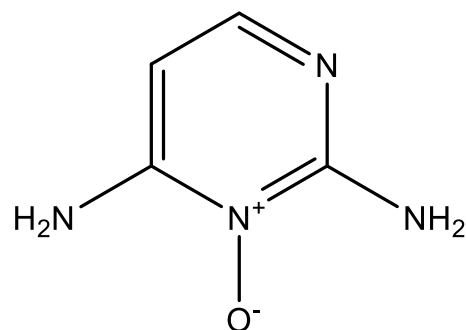
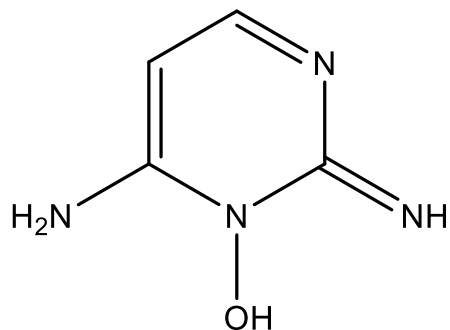
InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H

InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H



KOPEXIL

- An example of a 1,4-tautomer with different InChI.

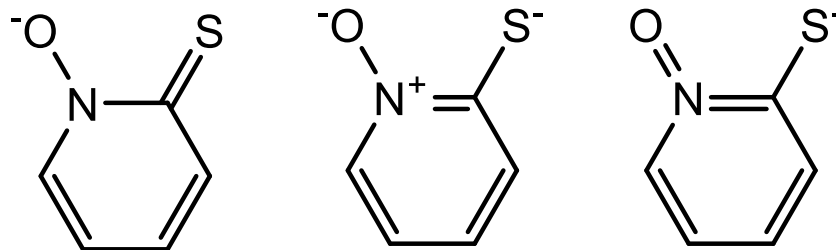


- On1c(N)ccnc1=N
 - InChI=1S/C4H6N4O/c5-3-1-2-7-4(6)8(3)9/h1-2,6,9H,5H2
- [O-][n+1]1c(N)ccnc1N
 - InChI=1S/C4H6N4O/c5-3-1-2-7-4(6)8(3)9/h1-2H,5H2,(H2,6,7)



PYRITHIONE

- Pyrithione mesomers have different InChI

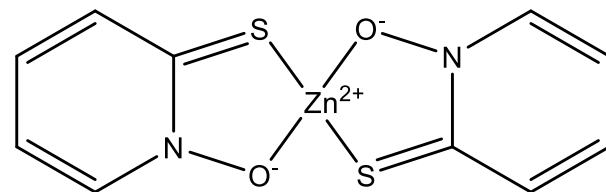


- InChI=1S/C5H4NOS/c7-6-4-2-1-3-5(6)8/h1-4H/q-1
- InChI=1S/C5H5NOS/c7-6-4-2-1-3-5(6)8/h1-4,8H/p-1



POSTERA'S COVID MOONSHOT

- Zinc pyrithione has been proposed (by researchers in Hamburg) as an inhibitor of COV2 main protease.
- Two structure representation families of InChI
 - C1=CC2=[S][Zn+2]3([O-]N2C=C1)[O-]N4C=CC=CC4=[S]3
 - S=c1cccn1O[Zn]On1ccccc1=S
 - S=c1cccn1[O-].[Zn+2].[O-]n1ccccc1=S
 - [O-]n1ccccc1=[S+][Zn][S+]=c1cccn1[O-]
 - O=n1ccccc1[S-].[Zn+2].[S-]c1cccn1=O
 - O=n1ccccc1S[Zn]Sc1cccn1=O
 - [O-][n+]1ccccc1S[Zn]Sc1cccc[n+]1[O-]
- Perhaps why PDB code 6YT8 has been obsoleted by 7B83.




NEUTRAL COMPONENT DUPLICATION

- Duplicated components lead to different InChI
 - Water (O)
 - InChI=1S/H2O/h1H2
 - XLYOFNOQVPJJNP-UHFFFAOYSA-N
 - Wet water (O.O)
 - InChI=1S/2H2O/h2*1H2
 - JEGUKCSWCFPDGT-UHFFFAOYSA-N
 - Dilute water (O.O.O)
 - InChI=1S/3H2O/h3*1H2
 - JLFVIEQMRKMAIT-UHFFFAOYSA-N
- Goodman's Hypothesis: How many InChI keys?



WATERS OF HYDRATION

- One usage where this is useful is in distinguishing hydrates (with different properties such as density, melting point, boiling point, solubility, appearance).

 **CAS**
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

About CAS Contact

Calcium chloride, dihydrate

CAS Registry Number®
10035-04-8

Cl[Ca]Cl

• 2 H₂O

CAS Name
Calcium chloride, dihydrate

Molecular Formula
CaCl₂·2H₂O

Compound Properties

Density (1)
1.86 g/cm³


Source(s)
(1) Leclaire, A.; Acta Crystallographica, Section B: Structural Crystallography and Crystal Chemistry, (1977), B33(5), 1608-10, CAplus

Other Names and Identifiers

InChI
InChI=1S/Ca.2ClH.H2O/h;2*1H;1H2/q+2;;;/p-2

InChIKey
YMIFCOGYMQTQBP-UHFFFAOYSA-L

SMILES
O.Cl[Ca]Cl

 **CAS**
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

About CAS Contact

Calcium chloride, hexahydrate

CAS Registry Number®
7774-34-7

Cl[Ca]Cl

• 6 H₂O

CAS Name
Calcium chloride, hexahydrate

Molecular Formula
CaCl₂·6H₂O

Compound Properties

Melting Point (1)
299 °C

Source(s)
(1) Akatsu, Eiko; Analytica Chimica Acta, (1971), 55(2), 333-40, CAplus

Other Names and Identifiers

InChI
InChI=1S/Ca.2ClH.H2O/h;2*1H;1H2/q+2;;;/p-2

InChIKey
YMIFCOGYMQTQBP-UHFFFAOYSA-L

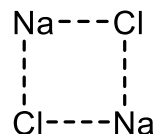
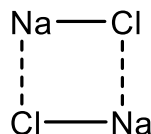
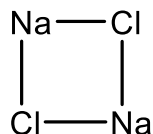
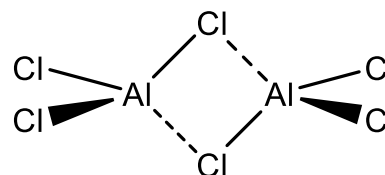
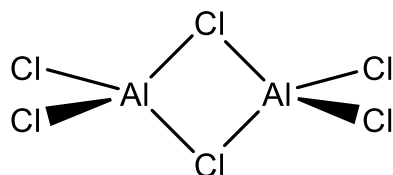
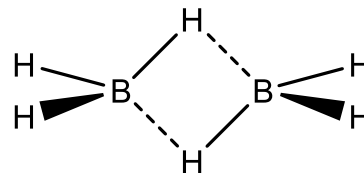
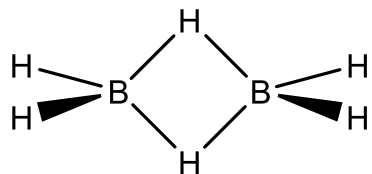
SMILES
O.Cl[Ca]Cl

Canonical SMILES



HOMODIMERS, POLYMERS & LATTICES

- This behavior complicates handling of homodimers.



DEMOCRITOS' CORROLARY

- “Nothing exists except atoms, organic connectivity and net charge, all else is opinion”.
- The molecular formula, including hydrogen count and net charge is important.
- Ionization (and some tautomerism) is a dilute aqueous phenomenon.
- A challenge in inorganic representation is whether bonds undergo heterolytic or homolytic cleavage, which effects how charges distribute over fragments.



WHEN INCHI WORKS WELL

- Silver diammine fluoride (SDF)

- F[Ag].N.N

CID129689514 (ChEMBL)

- F[Ag-2]([NH3+])[NH3+]

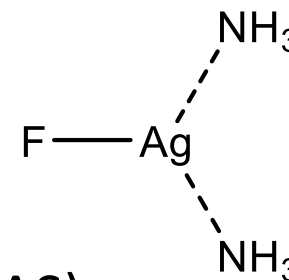
- F[Ag]([NH3])[NH3]

- [F-].[Ag+].N.N

CID161820

- [NH3][Ag+][NH3].[F-]

CID5461019 (CAS)



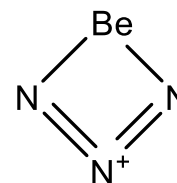
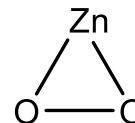
- InChI=1S/Ag.FH.2H3N/h;1H;2*1H3/q+1;;;/p-1

- Incorrect in NIH ChemIDPlus (but MF is OK!?).



WHEN INCHI WORKS WELL #2

- Hydrogen Chloride InChI=1S/ClH/h1H
 - Cl
 - [H+].[Cl-]
- Zinc Peroxide InChI=1S/O2.Zn/c1-2;/q-2;+2
 - [Zn]1OO1
 - [Zn+2].[O-][O-]
- Beryllium Azide InChI=1S/Be.2N3/c;2*1-3-2/q+2;2*-1
 - [N-]=[N+]=N[Be]N=[N+]=[N-]
 - [N-]=[N+]=[N-].[Be+2].[N-]=[N+]=[N-]
 - [Be]1N=[N+]=N1.[N-]=[N+]=[N-]



WHEN THINGS GO WRONG

- Sodium Chloride

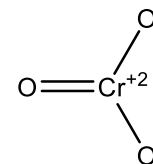
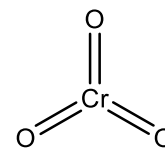
- [Na].[Cl] InChI=1S/Cl.Na
- [Na+].[Cl-] InChI=1S/ClH.Na/h1H;/q;+1/p-1

- Lithium Oxide

- [Li]O[Li] InChI=1S/2Li.O
- [Li+].[Li+].[O-2] InChI=1S/2Li.O/q2*+1;-2

- Chromium(IV) oxide

- O=[Cr](=O)=O InChI=1S/Cr.3O
- O=[Cr+2]([O-])[O-] InChI=1S/Cr.3O/q+2;;2*-1



- NH₃·BF₃

- N.FB(F)F InChI=1S/BF3.H3N/c2-1(3)4;/h;1H3
- [NH3+][B-](F)(F)F InChI=1S/BF3H3N/c2-1(3,4)5/h5H3



WHEN THINGS GO WRONG #2

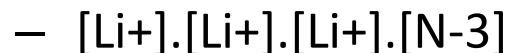
- Lithium nitride



InChI=1S/3Li.N



InChI=1S/3Li.N/q;;+1;-1

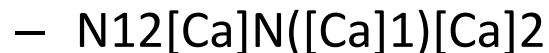


InChI=1S/3Li.N/q3*+1;-3

- Calcium nitride



InChI=1S/3Ca.2N



InChI=1S/3Ca.2N



InChI=1S/3Ca.2N/q;;+2;2*-1



InChI=1S/3Ca.2N/q3*+2;2*-3



WHEN THINGS GO WRONG #3

- Boron nitride

- B#N InChI=1S/BN/c1-2
- B1=NB=N1 InChI=1S/B2N2/c1-3-2-4-1
- B1=NB=NB=N1 InChI=1S/B3N3/c1-4-2-6-3-5-1

- Aluminium chloride

- Cl[Al](Cl)Cl InChI=1S/Al.3ClH/h;3*1H/q+3;;;/p-3
- Cl[Al]1(Cl)[Cl][Al]([Cl]1)(Cl)Cl
InChI=1S/2Al.4ClH.2Cl/h;;4*1H;;/q2*+2;;;;;/p-4
- Cl[Al-]1(Cl)[Cl+][Al-]([Cl+]1)(Cl)Cl
InChI=1S/2Al.4ClH.2Cl/h;;4*1H;;/q2*+1;;;;;2*+1/p-4



WHEN THINGS GO WRONG #4

- Collins reagent

- c1ccncc1.c1ccncc1.O=[Cr](=O)=O

- InChI=1S/2C5H5N.Cr.3O/c2*1-2-4-6-5-3-1;;;;;/h2*1-5H;;;;;

- c1cccc[n+]1[Cr](=O)([O-])([O-])[n+]1cccc1

- InChI=1S/2C5H5N.Cr.3O/c2*1-2-4-6-5-3-1;;;;;/h2*1-5H;;;;;/q;;+2;;2*-1

- Pyridinium Chlorochromate (PCC)

- [nH+]1cccc1.[O-][Cr](=O)(=O)Cl

- InChI=1S/C5H5N.ClH.Cr.3O/c1-2-4-6-5-3-1;;;;;/h1-5H;1H;;;;;/q;;+1;;;-1

- n1cccc1.O[Cr](=O)(=O)Cl

- InChI=1S/C5H5N.ClH.Cr.H2O.2O/c1-2-4-6-5-3-1;;;;;/h1-5H;1H;;1H2;;/q;;+2;;;/p-2

- Magnus' Green Salt

- [NH3+][Pt-2]([NH3+])([NH3+])[NH3+].Cl[Pt-2](Cl)(Cl)Cl

- InChI=1S/4ClH.4H3N.2Pt/h4*1H;4*1H3;;/q;;;;;;;2*+2/p-4

- [NH3][Pt]([NH3])([NH3])([NH3])[Pt](Cl)(Cl)(Cl)Cl









- InChI=1S/4ClH.4H3N.2Pt/h4*1H;4*1H3;;/q;;;;;;;+4/p-4



WIKIPEDIA'S PRAGMATIC SOLUTION

- Wikipedia's ChemBox and DrugBox templates support multiple InChIs, InChI keys and SMILES.

Kopexil

	
Names	
IUPAC names	
2,3-Dihydro-3-hydroxy-2-imino-4-pyrimidinamine	
2,4-Diaminopyrimidine 3- <i>N</i> -oxide	
Other names	
Aminexil	
Identifiers	
CAS Number	113275-13-1  ✓ 74638-76-9  (pyridine oxide tautomer) ✓
3D model (JSmol)	Interactive image 
ChemSpider	10445922 
EC Number	616-121-2
PubChem  CID	10197687 
UNII	1756681479 
InChI	[hide]
InChI=1S/C4H6N4O/c5-3-1-2-7-4(6)8(3)9/h1-2H,5H2,(H2,6,7) Key: SGHQFNHCCOBUKB-UHFFFAOYSA-N	
InChI=1S/C4H6N4O/c5-3-1-2-7-4(6)8(3)9/h1-2,6,9H,5H2 Key: YTKGAYFHUZTLCI-UHFFFAOYSA-N	
SMILES	[hide]
c1cnc([n+](c1N)[O-])N	

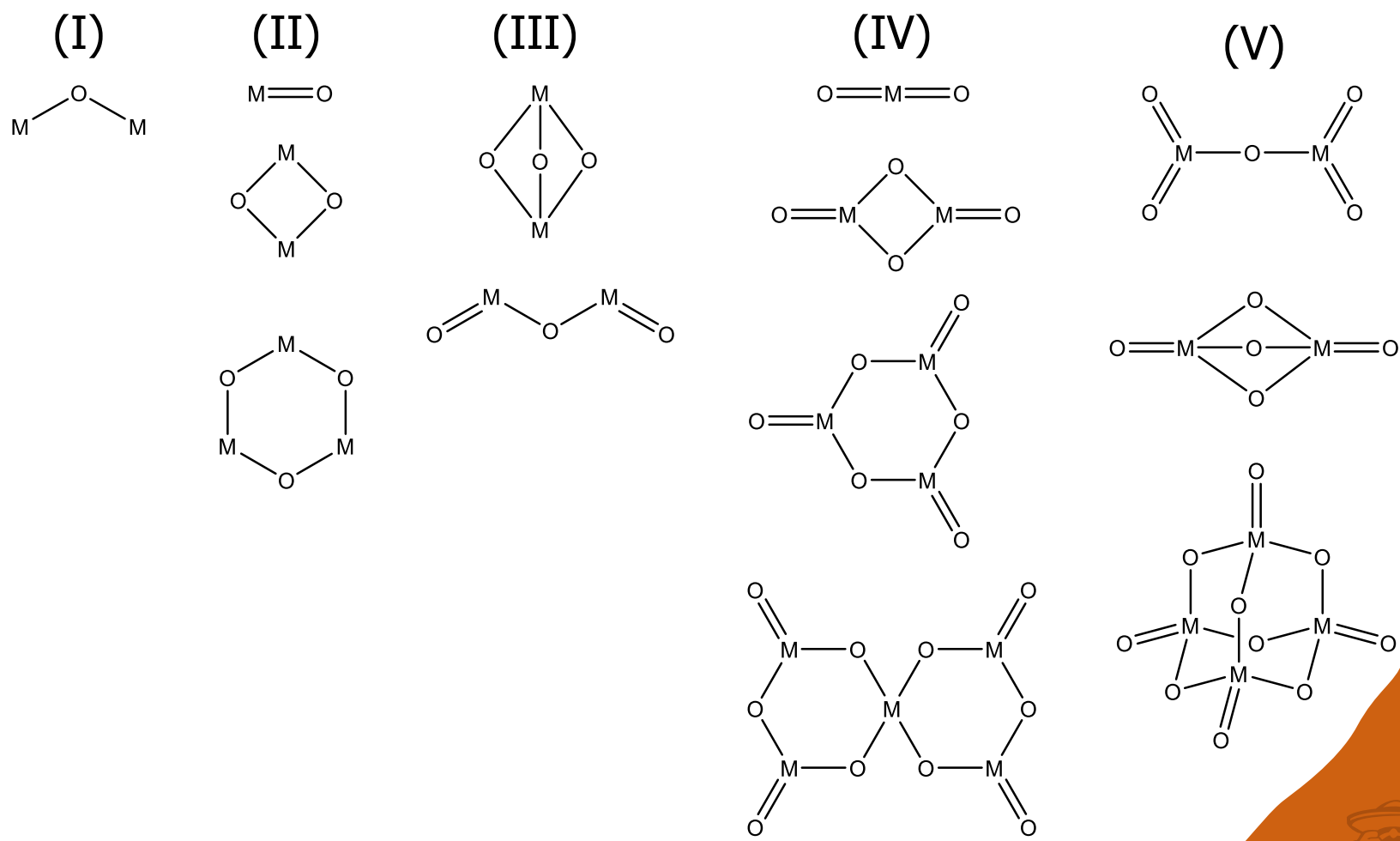


PERSONAL BEST PRACTICES

- Prefer molecular over ionic representations (both).
 - “Humpty dumpty” principle: It’s easier to break things, than to put them back together again.
- Prefer uncharged over zwitterionic representations.
 - “Born-Oppenheimer” may also be implemented algorithmically on InChIs (during registration/search).
- Prefer canonical forms of metal oxides, metal halides and metal chalcogenides.
- Capture both monomeric and dimeric/polymeric forms.

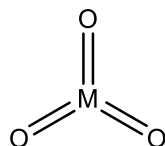


METAL OXIDE EQUIVALENT FORMS #1

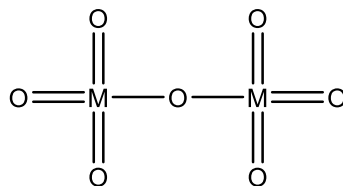


METAL OXIDE EQUIVALENT FORMS #2

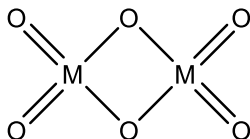
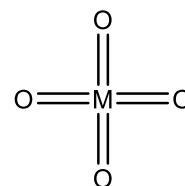
(VI)



(VII)



(VIII)



- And these are just the neutral binary metal oxides, there are even more permutations for ions (permanganates, perchlorate) and halides (aluminium chloride) and so on.



BLACK ADDER QUOTE

- “The path of my life is strewn with cowpats from the devil’s own satanic herd”.



ACKNOWLEDGEMENTS

- The team at NextMove Software
 - John Mayfield, Richard Gowers, Ingvar Lagerstadt
- And Alumni (Daniel Lowe, Noel O'Boyle)
- And the InChI/Cheminformatics Community (including)
 - Evan Bolton
 - Greg Landrum
 - Marc Nicklaus
 - Alex Clark
 - Jonathan Goodman
 - Andrew Dalke
 - Nick Tomkinson
 - Colin Batchelor
 - Philip Skinner
 - Ben Bracke
 - Pierre Morieux
 - Phil McHale
- And many thanks for your time!





THE NATURE OF THE CHEMICAL BOND

- Ionization energy (first)
 - $X \rightarrow X^+ + e^-$ $[\text{Na}] \rightarrow [\text{Na}^+]$ requires 496 kJ/mol
- Electron affinity (first)
 - $X + e^- \rightarrow X^-$ $[\text{Cl}] \rightarrow [\text{Cl}^-]$ produces 349 kJ/mol
 - $[\text{Na}].[\text{Cl}] \rightarrow [\text{Na}^+].[\text{Cl}^-]$ requires 147 kJ/mol
- Enthalpy of hydration
 - $[\text{Na}^+]$ produces 406 kJ/mol
 - $[\text{Cl}^-]$ produces 363 kJ/mol
 - $[\text{Na}^+].[\text{Cl}^-]$ produces 770 kJ/mol

