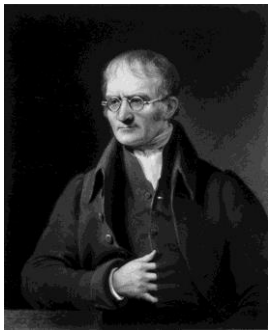
Chemical Structure Representation of Inorganic Salts and Mixtures of Gases: A Newer System of Chemical Philosophy

Roger Sayle

NextMove Software, Cambridge, UK

JOHN DALTON'S LEGACY

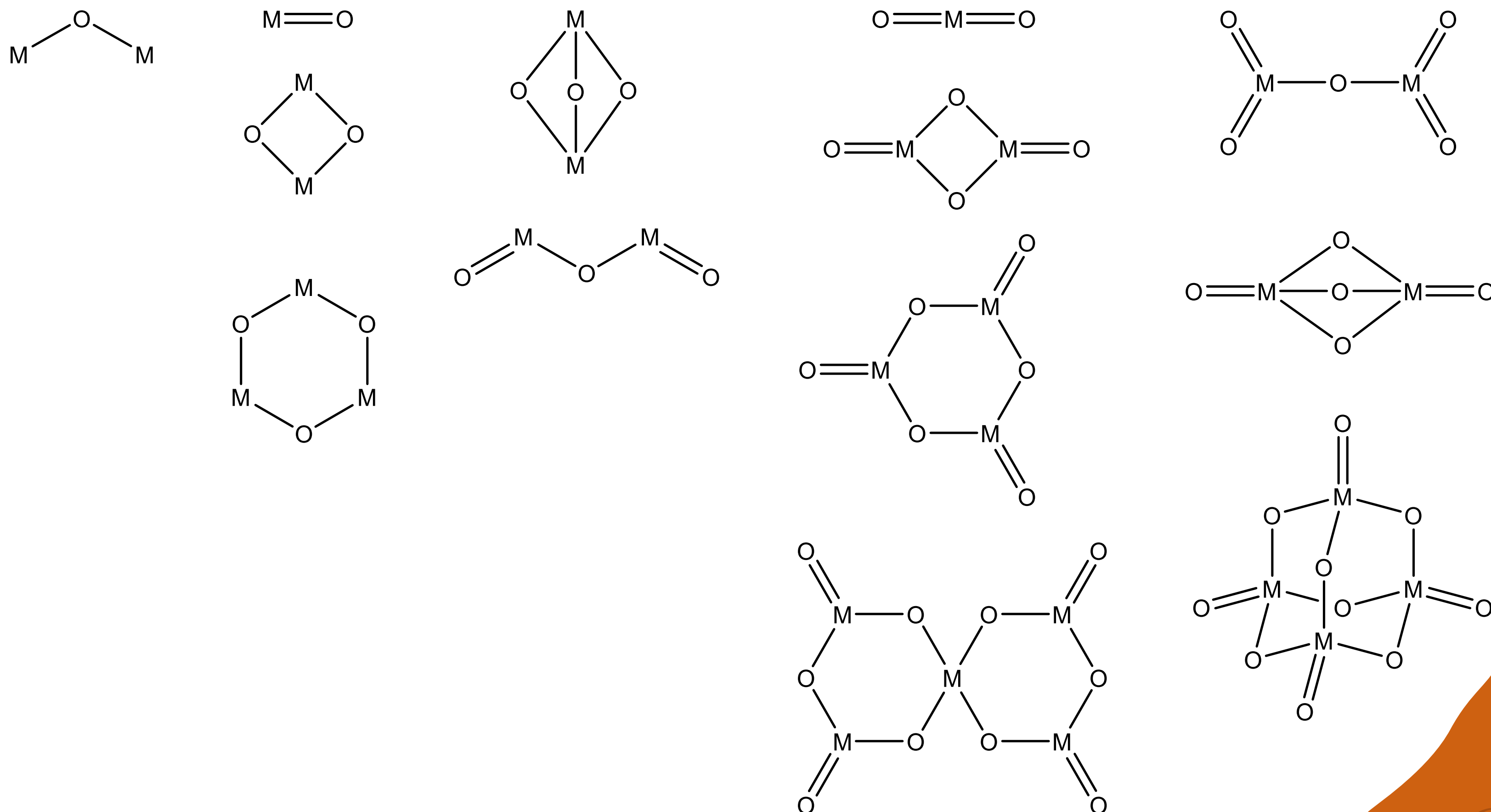


- In 1808, John Dalton published "A New System of Chemical Philosophy", in which he described his atomic theory, based upon the law of multiple proportion that revolutionized/defined molecular chemistry.
- Compounds are composed of atoms in defined whole-number ratios, where all atoms of an element are identical.
- Interestingly, 209 years later, boundary cases of this rule, define the frontiers of cheminformatics.

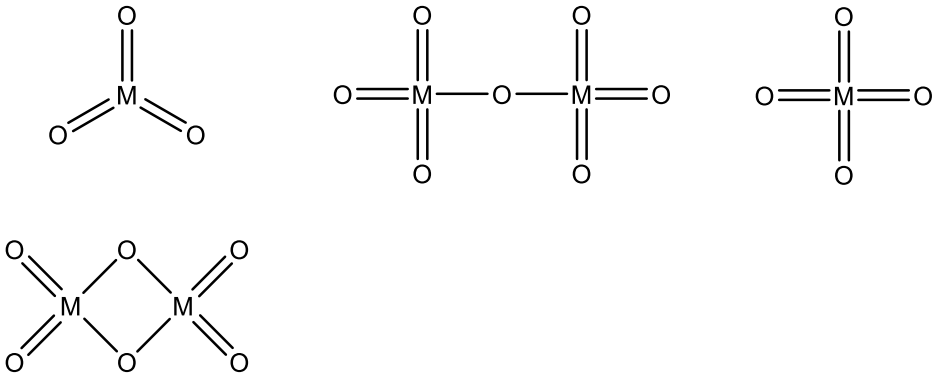
METALLIC OXIDES (AND FRIENDS)

- One problem area for chemical representation are compounds that have no discrete chemical structure but are defined by the ratios of their elements.
- One of Dalton's case studies was on tin oxides, SnO and SnO_2, often represented as O=[Sn], O=[Sn]=O

Representations of Ratios



REPRESENTATIONS OF RATIOS



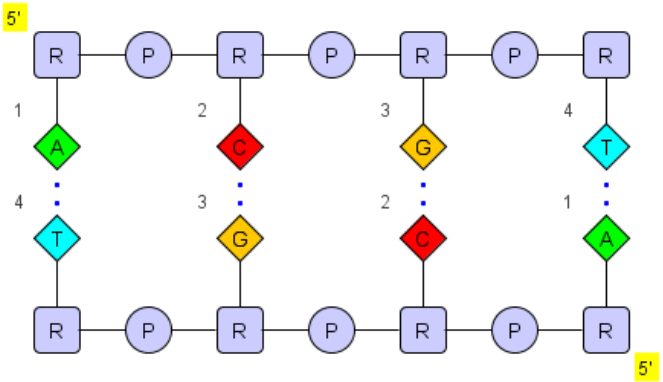
- And these are just the neutral binary metal oxides, there are even more permutations for ions (permanganates, perchlorate) and halides (aluminium chloride) and so on.
- Fortunately, a defining feature of a substance is that it has zero net charge.

CONTRIBUTION #1: DALTON SMILES

- A molecular representation that correctly captures the ratio of elements, but not necessarily connectivity.
  - O=[Si]=O — Silicon Dioxide (c.f. Wikipedia history)
  - O=P(=O)OP(=O)=O — Phosphorus pentoxide
  - [C] — Diamond, Graphene, Fullerenes
  - [C]=[C] — Graphene, Fullerene
- Extension to mixtures, where each component is listed, but not necessarily the relative amounts of each.
  - Cl.O — Hydrochloric acid
  - Cu1OS(=O)(=O)O1.O — Copper(II) sulfate hydrate
  - [Fe].[Cl] — Iron chloride

NEUTRAL COMPONENT DE-DUPLICATION

- Problems with InChI:
  - Water: InChI=1S/H2O/h1H2 `XLYOFNOQVPJJNP-UHFFFAOYSA-N`
  - Wet water: InChI=1S/2H2O/h2*1H2 `JEGUKCSWCFPDGT-UHFFFAOYSA-N`
  - Dilute water: InChI=1S/3H2O/h3*1H2 `JLFVIEQMRKMAIT-UHFFFAOYSA-N`
- Problems with Canonical HELM:
  - RNA1{R(A)P.R(C)P.R(G)P.R(T)}$$$$
  - RNA1{R(A)P.R(C)P.R(G)P.R(T)}|RNA2{R(A)P.R(C)P.R(G)P.R(T)}$$$$



IDENTIFIERS VS. REPRESENTATIONS

- Compounds are composed of atoms in defined whole-number ratios, where all atoms of an element are identical.
- It is this statement that allows us to claim that two compounds (or crystals) are identical, and can be captured by a canonical form or universal identifier.
- Without it, substances or mixtures of arbitrary composition are each unique, and one can only talk of similarity and equivalence, not of equality.

METAL ALLOYS

- **AdmiraltyBrass**
  - Cu 69 %
  - Zn 30 %
  - Sn 1 %
- **RollsRoyceTurbineAlloy1**
  - Ni 29.2-37 %mass
  - Co 29.2-37 %mass
  - Cr 10-16 %mass
  - Al 4-6 %mass
  - Zr 0.04-0.07 %mass

ATMOSPHERIC COMPOSITION

- **Air**
  - Nitrogen 78.084 %v
  - Oxygen 20.964 %v
  - Argon 0.9340 %v
  - Carbon dioxide 0.04 %v
  - Neon 0.001818 %v
  - Helium 0.000524 %v
  - Methane 0.00018 %v
  - Krypton 0.000114 %v
  - Hydrogen 0.000055 %v
- **Martian Atmosphere**
  - Carbon dioxide 95.97 %v
  - Argon 1.93 %v
  - Nitrogen 1.89 %v
  - Oxygen 0.146 %v
  - Carbon monoxide 0.0557 %v

SEA WATER COMPOSITION

- **SeaWater**
  - Water 1 liter
  - Salts 41.953 g
    - NaCl 58.490 %
    - MgCl2.6H20 26.460 %
    - Na2SO4 9.750 %
    - CaCl2 2.765 %
    - KCl 1.645 %
    - NaHCO3 0.477 %
    - KBr 0.238 %
    - H3BO3 0.071 %
    - SrCl2.6H20 0.095 %
    - NaF 0.007 %

WORDS HAVE MEANING (OR IS IT AI?)

- A personal observation from reading Dalton's work is his use of (the English) language to describe the chemistry that he observed.
  - Chapter 5: Compounds of two elements
  - Section 13: Metallic oxides
  - Oxide of zinc
  - Oxides of iron
  - Metallic alloys

209 YEARS, 20.9 YEARS OR 2.09 YEARS?

- Predicting the future is notoriously difficult…
- Clearly, there's an increasing need for chemical structures and scientific information to be captured electronically.
- In the computer age, efforts have been to codify structures as "computer readable" connection tables and line notations.
- My prediction is that computers will soon use natural language in the same way as human scientists.

NAMED MIXTURE EXAMPLES

- Hydrochloric acid
- Formalin
- Ice water
- Benzene (liquid)
- Benzene (solid)
- n-Butyllithium solution, 2.5M in hexanes
- Nitric acid and sulfuric acid
- Fuming sulfuric acid (oleum)

ONTOGREP SUPPORTED QUERIES

- compounds of two elements
- binary compounds
- metal oxide
- branched alkanes
- nitrogen containing heterocycles
- carbon-containing inorganic compounds
- zinc compounds
- neutral mixtures
- polyspiro ring systems
- atropisomers
- lewis acids
- C20H25N3O

REACTION INCHI EXAMPLE

Esterification of acetic acid with ethanol to acetic acid and water catalyzed by sulfuric acid:

RInChI=0.03.1S
/C2H4O2/c1-2(3)4/h1H3,(H,3,4)!
C2H6O/c1-2-3/h3H,2H2,1H3
<>
C4H8O2/c1-3-6-4(2)5/h3H2,1-2H3!
H2O/h1H2
<>
H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)
/d+

- "<>" separates reactants, products, and agents (= catalysts, solvents, etc.)
- "!" separates components within these groups
- alphabetical order of components within groups
- /d+ layer describes the direction of the reaction
- ("RInChI=0.03.1S" version identifier)

MInChI EXAMPLE #1

1.7M t-Butyllithium in Pentane:

MInChI=0.00.0S/
C4H9.Li/c1-4(2)3;/h1-3H3;/q-1;+1
&
C5H12/c1-3-5-4-2/h3-5H2,1-2H3
/n{1&2}
/g{17mr-1;}

- alphabetical order of components
- "&" separates components
- "{}" denotes mixture groups
- "/n" layer indexes components (e.g., order)
- "/g" layer notates concentration (symbols detailed separately)
- ("MInChI=0.00.0S" version identifier)

MInChI EXAMPLE #2

37% wt. Formaldehyde in Water with 10-15% Methanol:

MInChI=0.00.0S/
CH2O/c1-2/h1H2&
CH4O/c1-2/h2H,1H3&
H2O/h1H2
/n{1&2&3}
/g{37wf-2&10-15vf-2&}

- alphabetical order of components
- "&" separates components
- "{}" denotes mixture groups
- "/n" layer indexes components (e.g., order)
- "/g" layer notates concentration (symbols detailed separately)
- ("MInChI=0.00.0S" version identifier)

MINCHI EXAMPLE #3

25:24:1 (v/v) Phenol:Chloroform:Isoamyl Alcohol
with 10mM Tris, pH 8.0, and 1 mM EDTA:

MInChI=0.00.0S/
CHCl3/c2-1(3)4/h1H&
C4H11NO3/c5-4(1-6,2-7)3-8/h6-8H,1-3,5H2&
C5H12O/c1-5(2)3-4-6/h5-6H,3-4H2,1-2H3&
C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H&
C10H16N2O8/c13-7(14)3-11(4-8(15)16)1-2-12(5-9(17)18)6-10(19)20/h1-6H2,(H,13,14)(H,15,16)(H,17,18)(H,19,20)&
H2O/h1H2
/n{1&3&4}{2&5&6}
/g{24vp&1vp&25vp}{1mr-3&1mr-2&}
/pH8.0

- alphabetical order of components
- "&" separates components
- "{}" denotes mixture groups
- /g layer notates concentration (symbols detailed separately)
- ("MInChI=0.00.0S" version identifier)

PEPTIDE NOMENCLATURE

- PEPTIDE1{H.E.L.M}$$$$
- L-histidyl-L-alpha-glutamyl-L-leucyl-L-methionine
- His-Glu-Leu-Met-OH

- acetyl-casokefamide
- PEPTIDE1{[ac].Y.[dA].F.[dA].Y.[am]}$$$$
- N-acetyl-L-tyrosyl-D-alanyl-L-phenylalanyl-D-alanyl-L-tyrosinamide
- Ac-Tyr-D-Ala-Phe-D-Ala-Tyr-NH2

RECENT EXAMPLE

- Earlier this week have been discussing the nomenclature to used for non-standard amino acids.
- Under discussion is the use of three-letter codes Dap vs Dpr, for 3-aminoalanine.
- Both appear in the literature, so both should be read.
- Ideally, a system should also support H-Ala(NH2)-OH.
- In PubChem synonyms Dap is 6 times more common that Dpr, and on Google Dap is 14 times more often.
- Naturally, the Pistoia Alliance chose to use Dpr!?

CCDC CSD INORGANIC EXAMPLE

- catena(Tetra-aqua-tetrakis(mu!2$-formato-O,O')-bis(formato-O)-di-barium-copper)
- (mu!2$-2,5-bis((Phenylimino)methyl)benzene-1,4-diyl-C,C',N,N')-bis(eta$5!-pentamethylcyclopentadienyl)-dichloro-di-iridium

INN ANTIBODY NOMENCLATURE

- immunoglobulin G1-kappa, anti-[Homo sapiens SLAMF7 (SLAM family member 7, CD2 subset 1, CS1, CD2-like receptor-activating cytotoxic cells, CRACC, 19A24, CD319)], humanized and chimeric monoclonal antibody antibody conjugated to auristatin E; gamma1 heavy chain (1-447) [humanized VH (Homo sapiens IGHV3-7*01(91.80%) -(IGHD) -IGHJ4*01 L123>T (112)) [8.8.10] (1-117) -Homo sapiens IGHG1*03v, G1m3>G1m17, nG1m1 (CH1 R120>K (214) (118-215), hinge (216-230), CH2 (231-340), CH3 E12(366), M14 (368) (341-445), CHS (446-447) (118-447)], (220-220')-disulfide with kappa light chain chimeric (1'-220') [Mus musculus V-KAPPA (IGKV1-110*01 (93.00%) -IGKJ4*01) [11.3.10] (1'-113') -Homo sapiens IGKC*01, Km3 A45.1 (159), V101 (197) (114'-220')]; dimer (226-226'':229-229'')bisdisulfide; conjugated, on an average of 3 cysteinyl, to monomethylauristatin E (MMAE), via a cleavable maleimidocaproyl-valyl-citrullinyl-paminobenzyloxycarbonyl (mc-val-cit-PABC) type linker

CONCLUSIONS

- Dalton's observations provided the insight that allow us to determine when two molecules are the same (one-to-one).
- Understanding when these rules don't apply can be useful for describing substances as similar (one-to-many).
- Natural language can be used to specify both precise and generic/ambiguous names (resp.)

ACKNOWLEDGEMENTS

- The team at NextMove Software
  - John Mayfield
  - Noel O'Boyle
  - Daniel Lowe
- And the Cheminformatics Community
  - Leah McEwen
  - Philip Skinner
  - Evan Bolton
  - Greg Landrum
  - Ian Bruno
- And many thanks for your time!

MOTIVATIONAL EXAMPLE

- In 2011, ILSAC (the International Lubricant Standardization and Approval Committee) introduced GF-5, a new standard (test) for motor oils.
- This specification introduced a new test to check whether the oil prevents ice blockages forming in engines at freezing temperatures, from a small amount of water contamination.
- The test itself is ATSM D7563-10: Standard Test Method for the Evaluation of the Ability of Engine Oil to Emulsify Water and Simulated ED85 Fuel.

RELEVANT LITERATURE

- P. Patel, C. Puckett, D. George and K. Nass, "**Effect of Viscosity Index Improvers in Ethanol/Gasoline/Water Emulsions formed with E25 and E85 in Passenger Car Motor Oils**", *SAE International Journal of Fuels and Lubricants*, 3(2):938-945, 2010. doi:10.4271/2010-01-2258

FORMULATION QSAR MODELING

	API	KV100 mm^2/s	KV40	Visc. Index	Pour Point	Flash Point	Sulfur Mass%	$\%C_A$	$\%C_N$	$\%C_P$
BaseOil1	3	4.2	19.4	123	-15.0	214	0.0008	0	22.4	77.6
BaseOil2	3	7.6	45.6	133	-12.5	240	0.001	0	20.4	79.6
BaseOil3	3	3.1	12.4	104	-32.5	194	<0.01	0	31.1	69.9
BaseOil4	1	4.6	24.4	99	-20.0	228	0.48	3.4	30.1	66.5
BaseOil5	1	7.6	55.1	99	-12.5	256	0.62	3.2	30.7	66.1
BaseOil6	1	11.3	101.6	97	-10.0	262	0.67	2.9	29.7	67.4
BaseOil7	3	5.0	23.7	146	-20.0	232	<0.01	0	7	93

FORMULATION REPRESENTATION

- **Formulation1**
  - BaseOil1 74.1 %mass
  - BaseOil3 6.00 %mass
  - Glycerine monooleate 9.05 %mass Additive
  - VIImprover2 10.50 %mass ViscosityIndexImprover
  - AntiFoamingAgent1 0.04 %mass AntiFoamingAgent

FORMULATION MODELING

- **Formulation2**
  - BaseOil1 64.01 %mass
  - BaseOil4 10.00 %mass
  - BaseOil5 10.00 %mass
  - Glycerine monooleate 0.90 %mass
  - GF5Package1 9.05 %mass
  - VIImprover 6.00 %mass
  - AntiFoamingAgent1 0.04 %mass
- **Formulation3**
  - BaseOil1 54.01 %mass
  - BaseOil4 20.00 %mass
  - BaseOil5 10.00 %mass
  - Glycerine monooleate 0.90 %mass
  - GFPackage1 9.05 %mass
  - VIImprover1 6.00 %mass
  - AntiFoamingAgent1 0.04 %mass

FORMULATION ASSAY RESULTS

Calculated	Formulation1	Formulation2	Formulation3
Base S%	0.00	0.13	0.19
Base CA%	0.0	0.8	1.2
Base CN%	23.0	24.3	25.2
Base CP%	77.0	74.9	73.6
Experimental	**Formulation1**	**Formultation2**	**Formulation3**
Viscosity Grade	0W-20	5W-30	5W-30
KV100	8.7	10.5	10.6
D7563 @ 0°C	No Sep.	No Sep.	No Sep.
D7563 @ 25°C	Separation	Separation	No Sep.

Formulations 2 and 3 are “matched pairs” that differ in the results Of ASTM D7563 @25° after 24 hours; an important observation.