AI4Science camp USC-Columbia 7/25/24
Sophya Garashchuk
Chemistry@USC

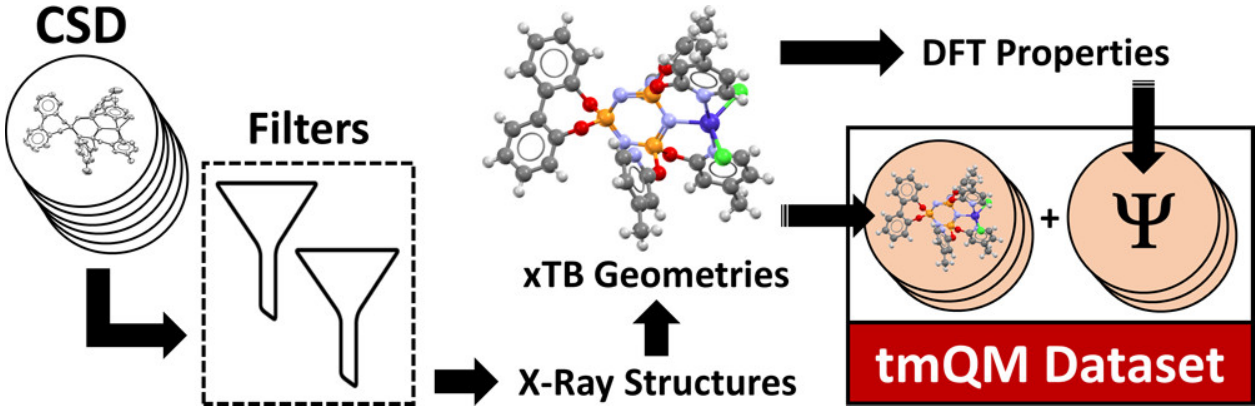
Data science for chemistry
Focus on the molecules

[some background for the rational choice of ML descriptors]

tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes

David Balcells and Bastian Bjerkem Skjelstad

Journal of Chemical Information and Modeling **2020** *60* (12), 6135-6146 DOI: 10.1021/acs.jcim.0c01041



- Quantum chemistry
- Electronic structure
- Machine learning

outline

- Atoms and the periodic table
- Electronic configurations
- Properties of elements as potential descriptors
- Molecular properties
- Chemical bonding
- tmQM dataset

General chemistry textbook by Stephen Lower, Ch.5
https://chem.libretexts.org/Bookshelves/General_Chemistry/Chem1_(Lower)

Atom

Proton ‘p’	m=1836.15 a.u	charge = +1
Electron ‘e’	m = 1.0 a.u.	charge = -1
Neutron ‘n’	m = 1838.68 a.u.	charge = 0

Atomic units:
|e-charge|=1 a.u. = 1.6 e-19 Coulomb e-mass = 1 a.u. = 9.11e-31 kg
Distance a.u. is bohr, 1 bohr = average e-p separation in H-atom
1 bohr = 0.53 Angstrom
1 Angstrom = 10^{-10} meters = 100 picometers [pm]

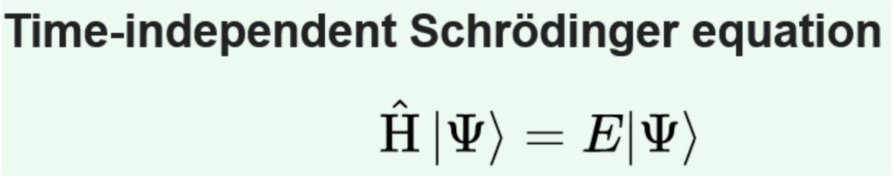
Identity of an atom is defined by the number of protons
Number of neutrons defines an isotope ^{2}H = proton + neutron + electron

Neutral atom: $N_p = N_e$ charge = 0 sodium atom Na
Cation: remove an electron sodium cation Na^+
Anion: add an electron sodium anion Na^-

The typical size of an atom or a molecule is 1-10 Angstrom or 100-100 pm
size of H^+ = 0.00084 pm size of H = 53 pm

Atom is ‘empty’

- At this scale atoms and molecules are governed by quantum mechanics (QM) and in particular electrons act as both particles and waves
- They are governed by a differential equation



Time-independent Schrödinger equation

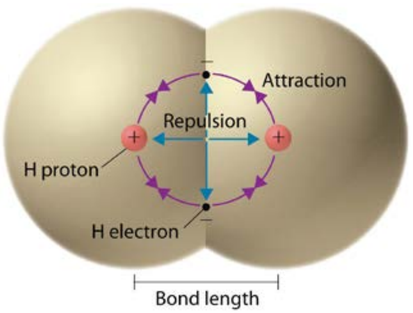
$$\hat{\mathrm{H}}\,|\Psi\rangle = E|\Psi\rangle$$

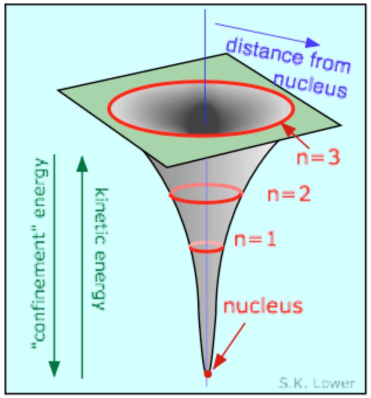
Hard!

Try ML

- Coulomb interaction between the charges
- Electrons are described by stationary probability distributions, **$|\Psi|^2$**, of charge -- *orbitals* -- of certain energy, ***E***
- The energy values are discrete; orbitals have specific shapes
- Electrostatic (Coulomb) interaction between the charges

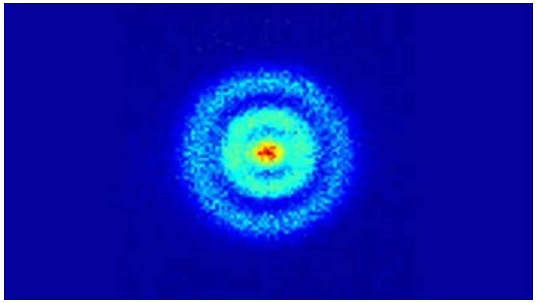
V ~ -1/distance





Hydrogen Atoms under Magnification: Direct Observation of the Nodal Structure of Stark States

A. S. Stodolna et al JPCL 2013

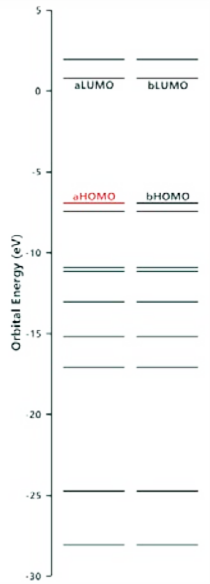


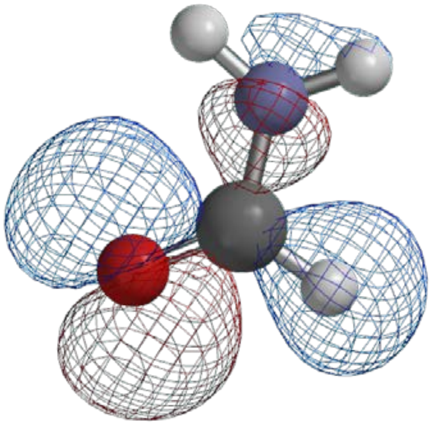
- Electrons ‘occupy’ energy levels
- Atomic or molecular orbitals (MO)
- HOMO = highest occupied MO
- LUMO = lowest occupied MO
- Energy level diagram

Absorption and emission of electromagnetic radiation (light) corresponds to the difference between the energy levels

A ‘material design target’

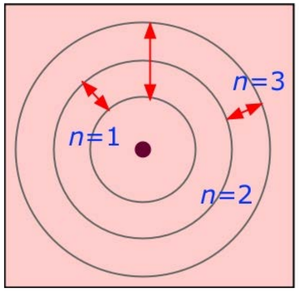
formamide

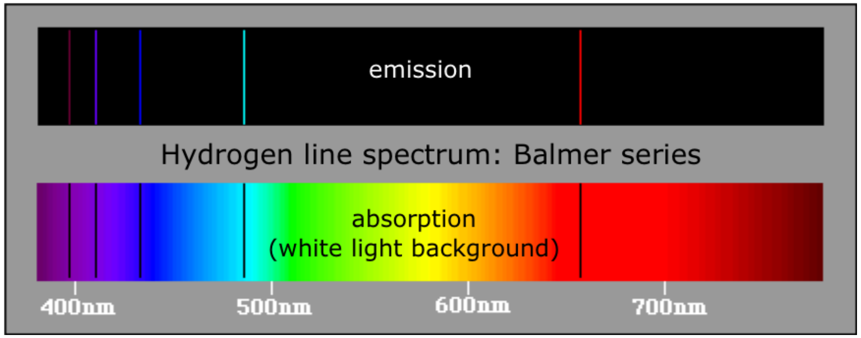




oxygen O ***carbon C***
nitrogen N ***hydrogen H***

H atom





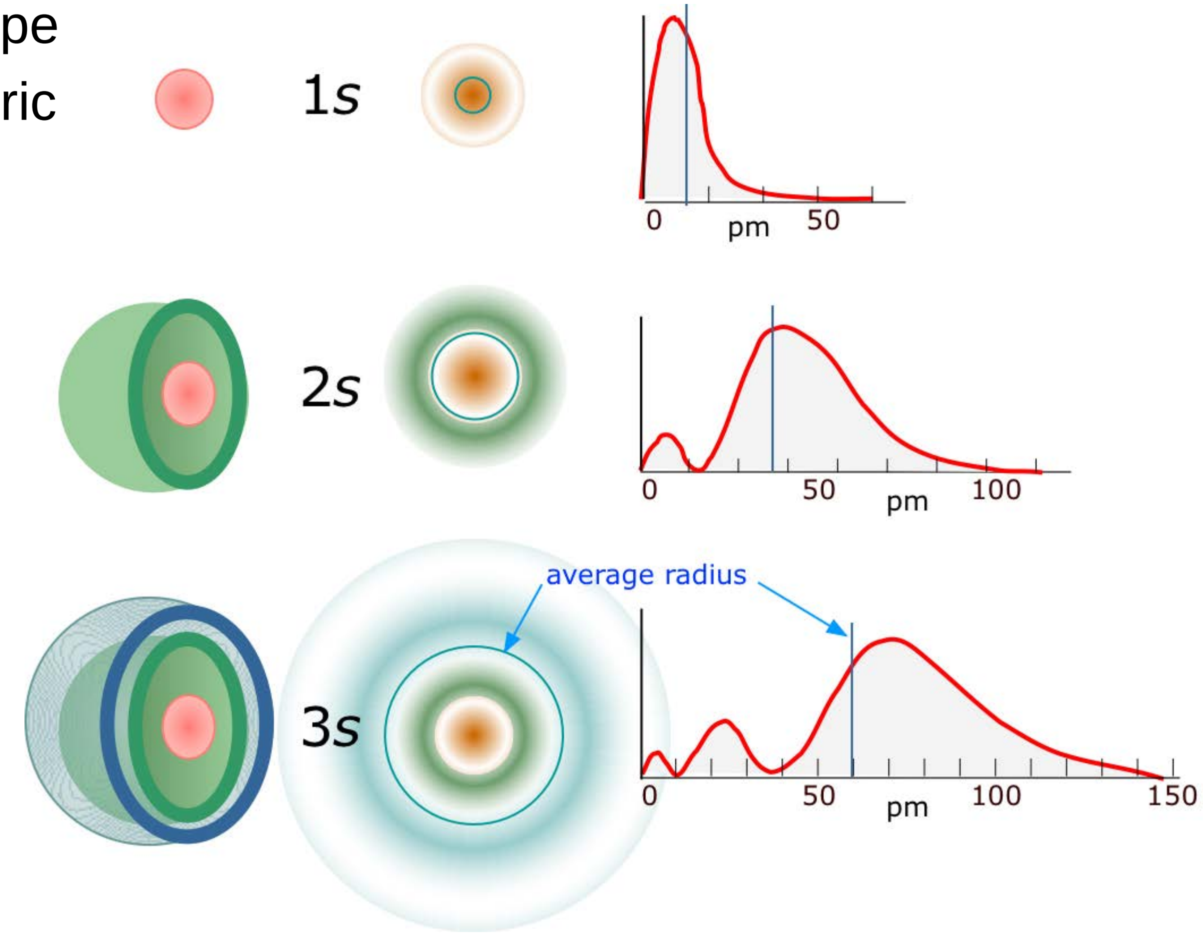
Shapes of atomic orbitals

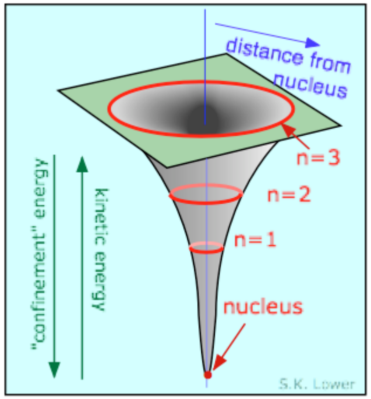
labeled ***ns, np, nd, nf***

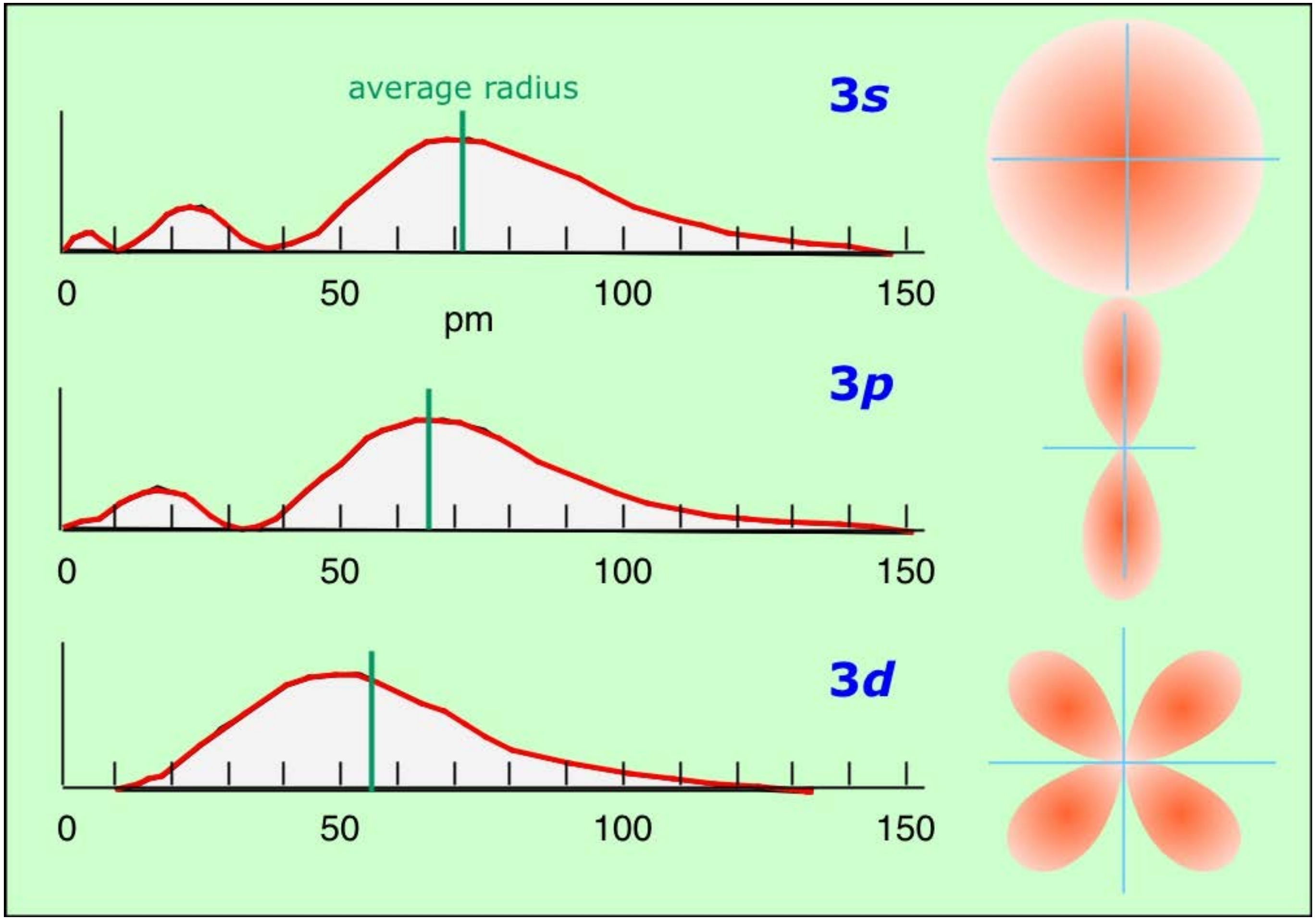
n corresponds to energy E~-1/n^2

s/p/d/f to the orbital shape

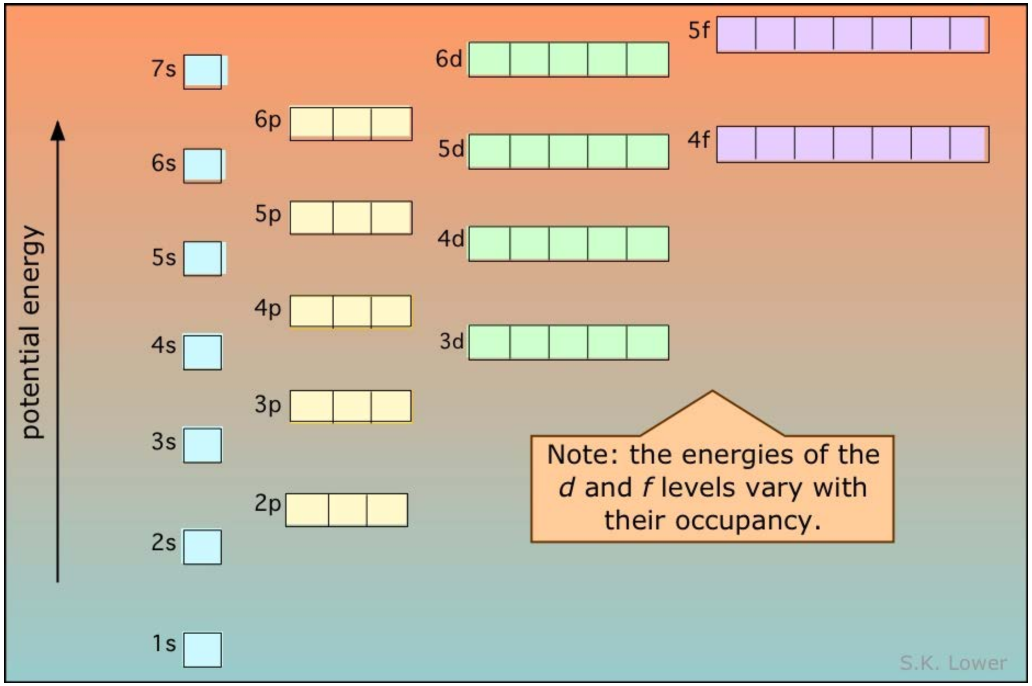
s is spherically symmetric






average radius
3s
0
50
100
150
pm
3p
0
50
100
150
3d
0
50
100
150

The energy level diagram in a multielectron atom



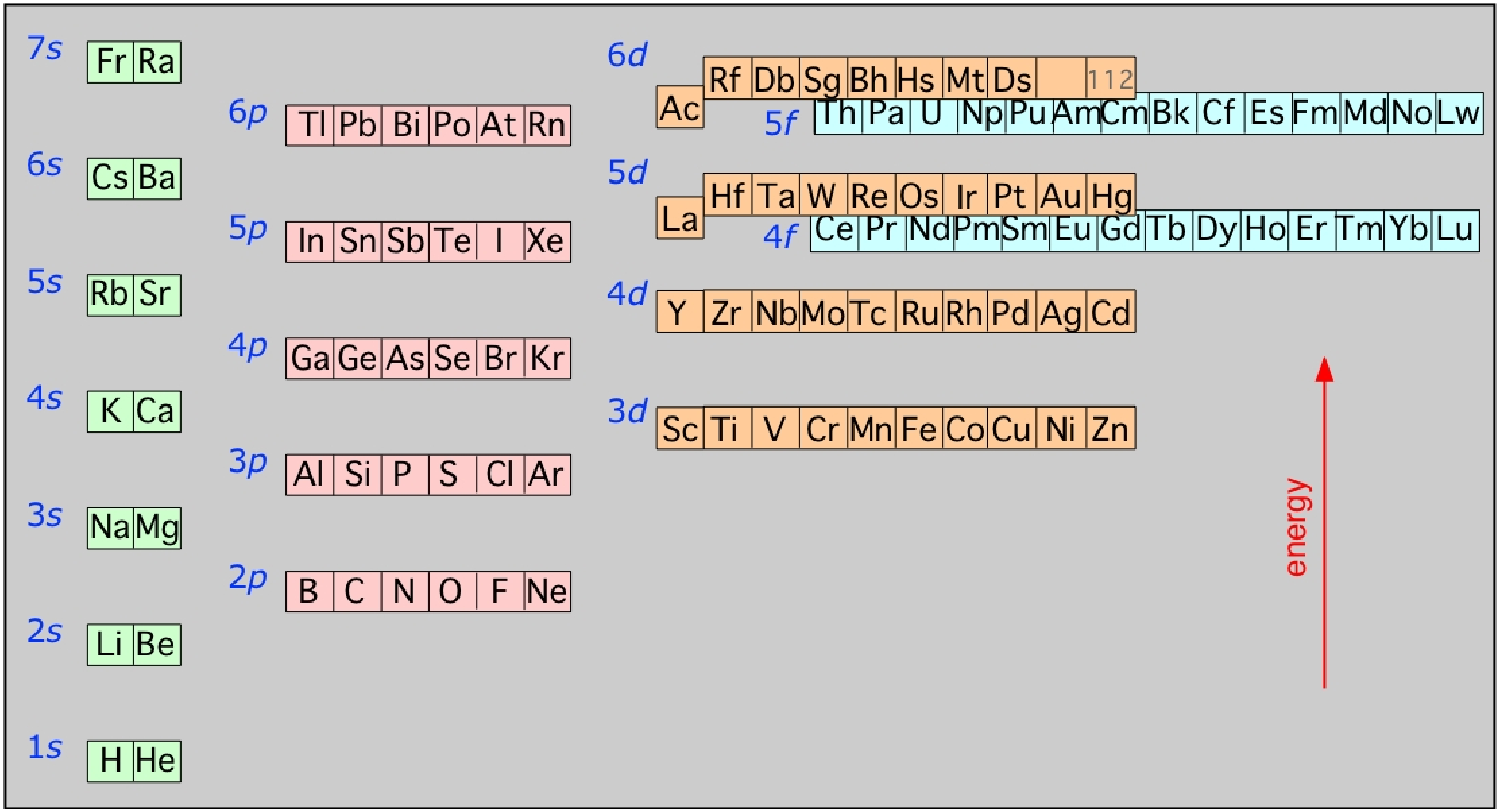
Note: the energies of the *d* and *f* levels vary with their occupancy.

- The electrons fill up the energy levels from bottom up
- Electrons have additional property ‘spin’
- Each level holds up to 2 electrons – one spin up, one spin down
- If the levels are of equal energy, the electron spins will be maximally unpaired
- The electronic structure of the highest energy shell (the highest **n**) -- the valence electrons -- defines chemical bonding
- Electron pairs ‘make’ chemical bonds (spin is important here)

Similar electronic structure in the valence shell →

Similar chemical and physical properties →

Periodic table of elements (Mendeleev 1869)



7*s* Fr Ra

6*d* Rf Db Sg Bh Hs Mt Ds 112

6*p* Tl Pb Bi Po At Rn

Ac

5*f* Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lw

6*s* Cs Ba

5*d* Hf Ta W Re Os Ir Pt Au Hg

La

4*f* Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu

5*p* In Sn Sb Te I Xe

5*s* Rb Sr

4*d* Y Zr Nb Mo Tc Ru Rh Pd Ag Cd

4*p* Ga Ge As Se Br Kr

4*s* K Ca

3*d* Sc Ti V Cr Mn Fe Co Cu Ni Zn

3*p* Al Si P S Cl Ar

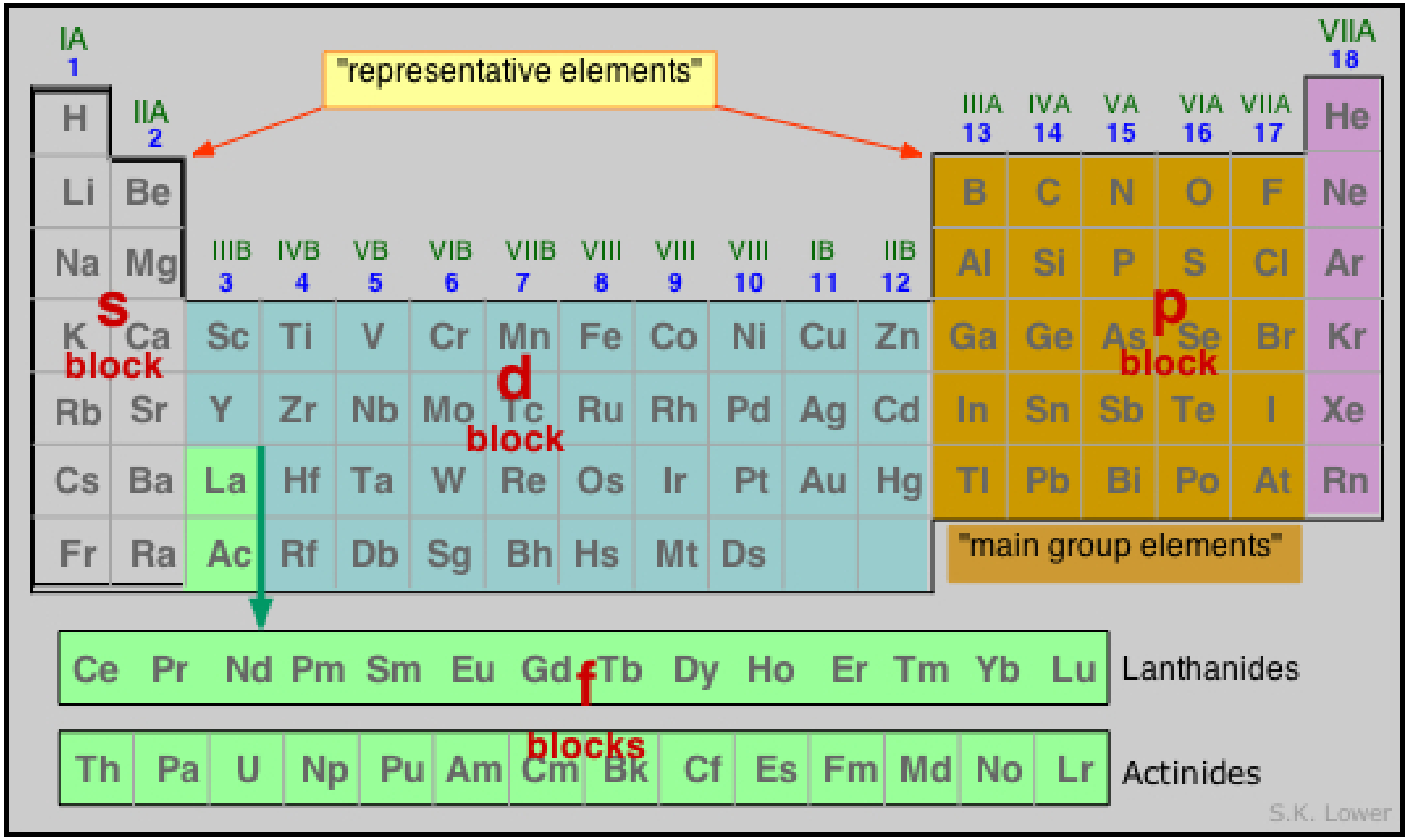
3*s* Na Mg

2*p* B C N O F Ne

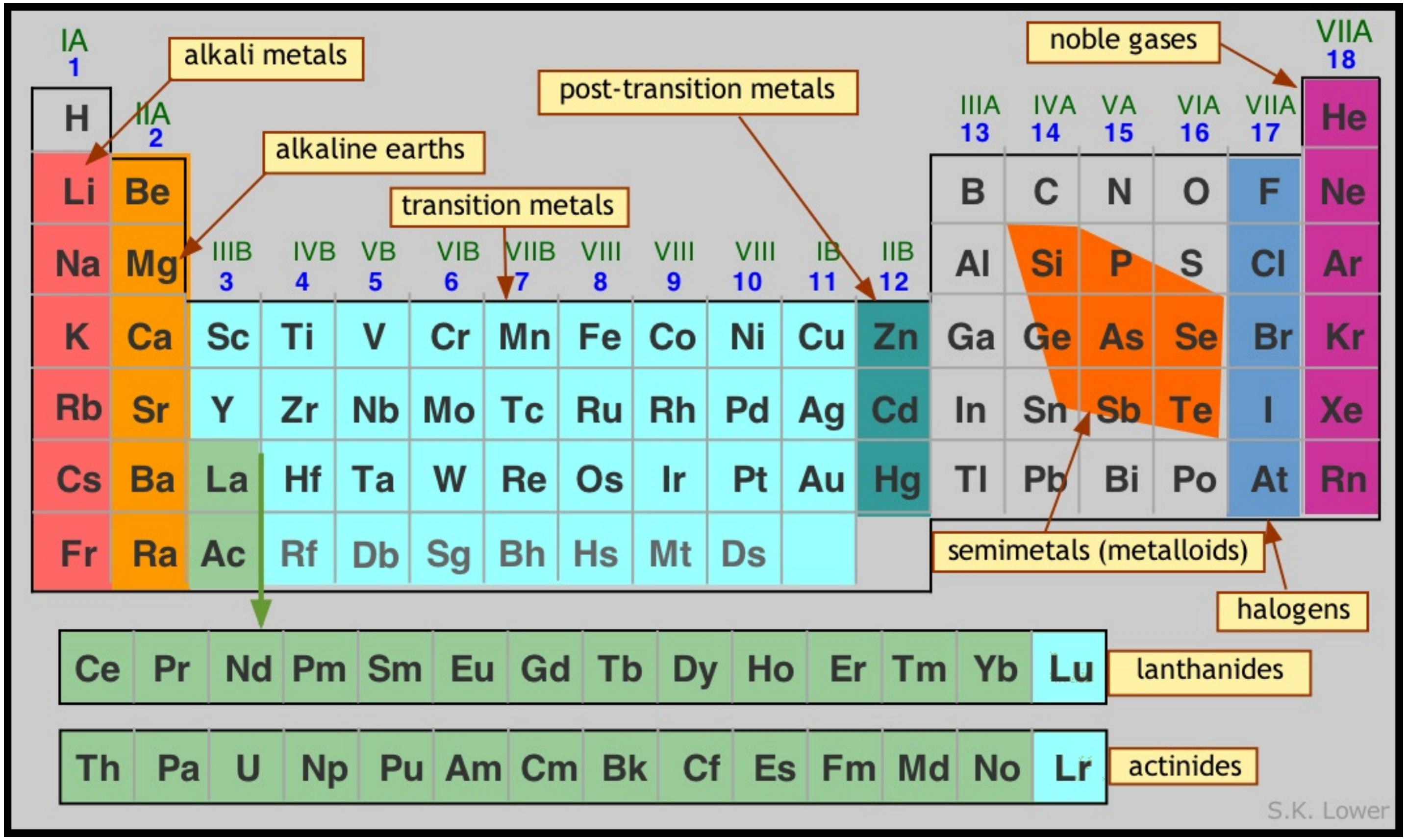
2*s* Li Be

1*s* H He

Periods = rows Groups = columns


IA
1
VIIA
18
"representative elements"
H
IIA
2
IIIA
13
IVA
14
VA
15
VIA
16
VIIA
17
He
Li Be
B C N O F Ne
Na Mg
IIIB
3
IVB
4
VB
5
VIB
6
VIIB
7
VIII
8
VIII
9
VIII
10
IB
11
IIB
12
Al Si P S Cl Ar
s
block
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
p
block
d
block
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba La Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Ac Rf Db Sg Bh Hs Mt Ds
"main group elements"
Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Lanthanides
f
blocks
Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
Actinides
S.K. Lower

- The non-metallic elements occur only in the *p*-block;
- The *d*-block elements contain the so-called transition elements;
- The *f*-block elements go in between Groups 3 and 4 of the *d*-block.



IA 1	IIA 2	IIIB 3	IVB 4	VB 5	VIB 6	VIIB 7	VIII 8	VIII 9	VIII 10	IB 11	IIB 12	IIIA 13	IVA 14	VA 15	VIA 16	VIIA 17	VIIA 18
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt	Ds								

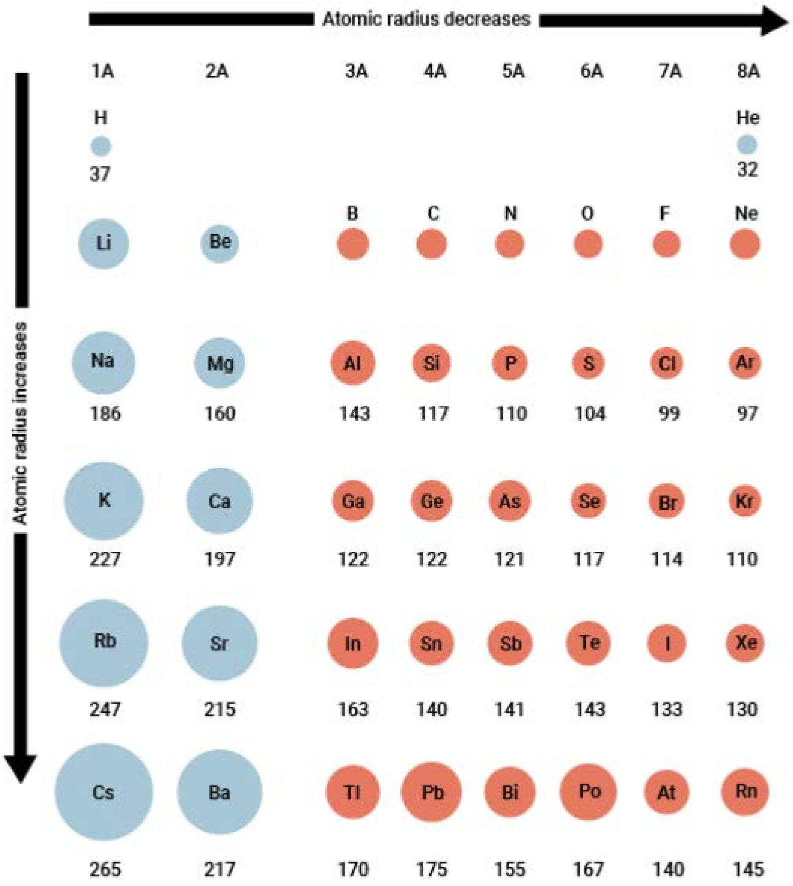
Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	lanthanides
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	actinides

Transition metals have ***d*** and ***s*** energy levels closely spaced → multiple electronic configurations are possible → rich chemistry (catalysis, metalloenzymes)

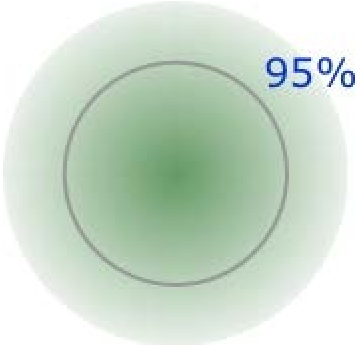
Sizes of atoms and ions ~ 70-400 pm

In molecules both atomic sizes and charges are not well-defined

But this size could be important when ‘measuring bonds’ in ML



1A	2A	3A	4A	5A	6A	7A	8A
H 37							He 32
Li	Be	B	C	N	O	F	Ne
Na 186	Mg 160	Al 143	Si 117	P 110	S 104	Cl 99	Ar 97
K 227	Ca 197	Ga 122	Ge 122	As 121	Se 117	Br 114	Kr 110
Rb 247	Sr 215	In 163	Sn 140	Sb 141	Te 143	I 133	Xe 130
Cs 265	Ba 217	Tl 170	Pb 175	Bi 155	Po 167	At 140	Rn 145



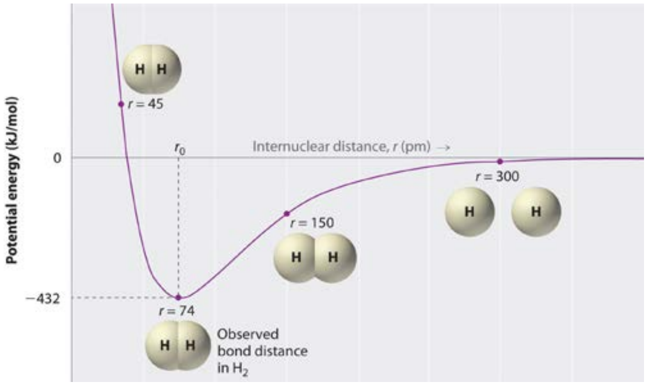
https://www.breakingatom.com/learn-the-periodic-table/atomic-radius-of-elements

Chemical bonds

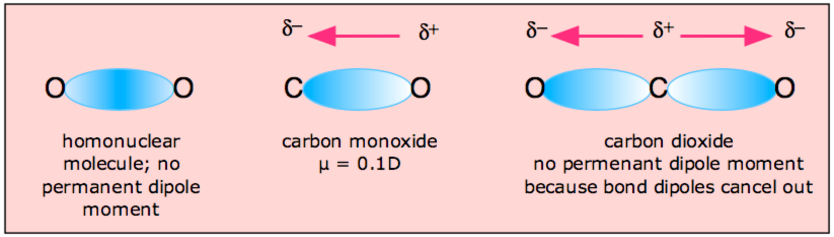
Minimum energy defines the bond distance

Shorter bonds are usually stronger

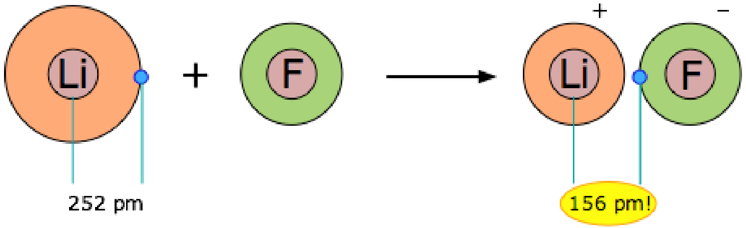
Bonding = (most often) shared electron pairs



Covalent bond (polar/non-polar)

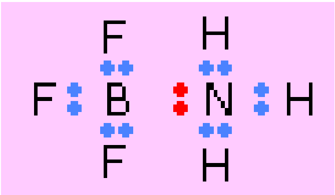

O O
homonuclear molecule; no permanent dipole moment
δ− δ+
C O
carbon monoxide
μ = 0.1D
δ− δ+ δ−
O C O
carbon dioxide
no permanent dipole moment because bond dipoles cancel out

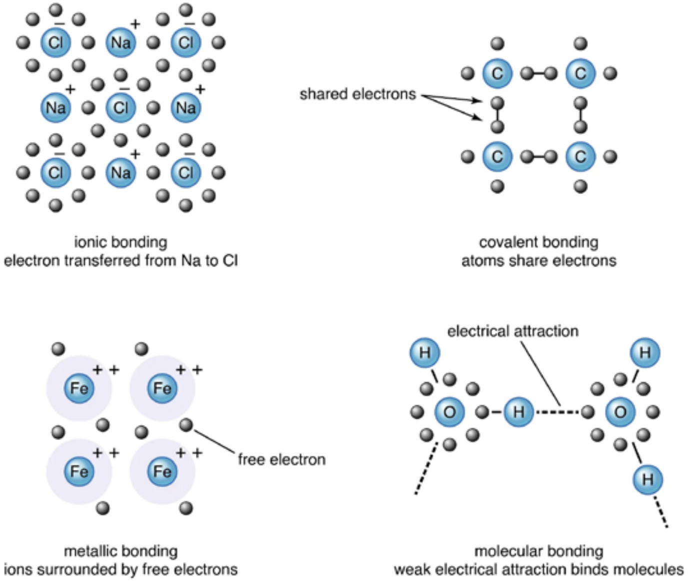
Ionic bond



Dative or coordinate covalent bond:

both electrons come from one atom





© Encyclopædia Britannica, Inc.

Metallic bonding

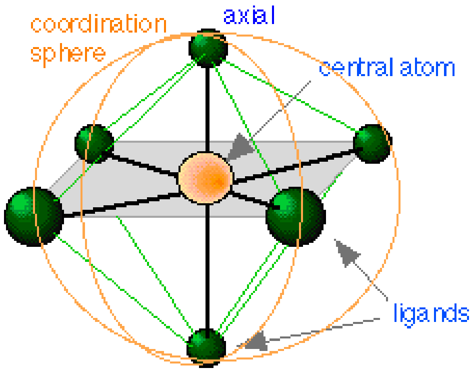
shared electrons

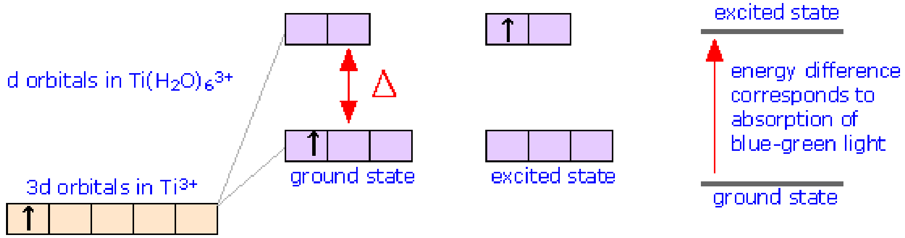
Hydrogen bond

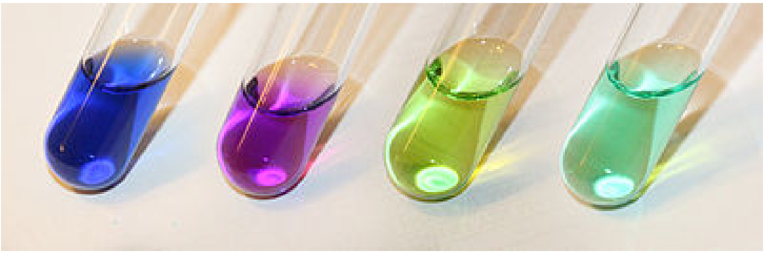
partial charges

ML descriptor: number of bonds

Coordination complexes; transition metals




d orbitals in $Ti(H_2O)_6^{3+}$
3d orbitals in Ti^{3+}
Δ
ground state
excited state
excited state
energy difference corresponds to absorption of blue-green light
ground state



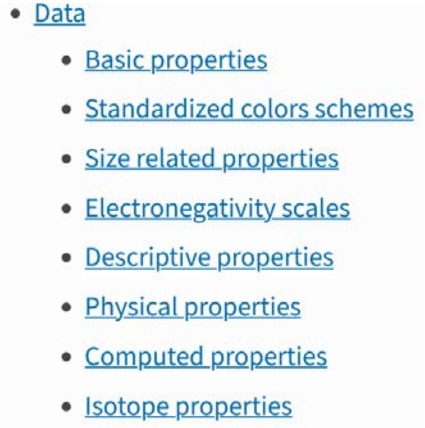
Fog: Color of various Ni(II) complexes in aqueous solution.

From left to right, hexaamminenickel(II), tris(ethylenediamine)nickel(II), tetrachloronickelate(II) and hexaaquanickel(II)

Wealth of info in the periodic table module

mendeleev 0.17.0

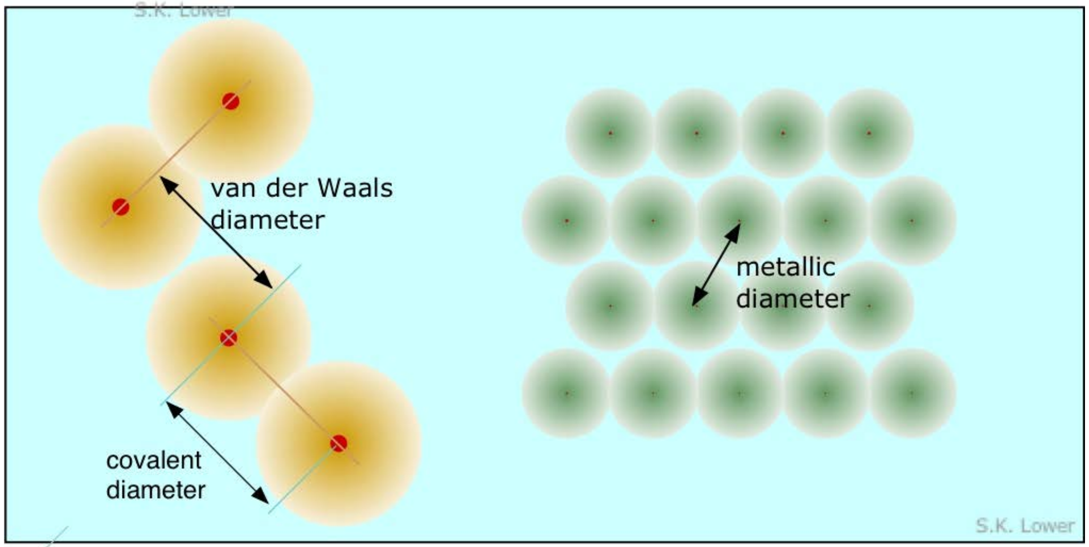
pip install mendeleev

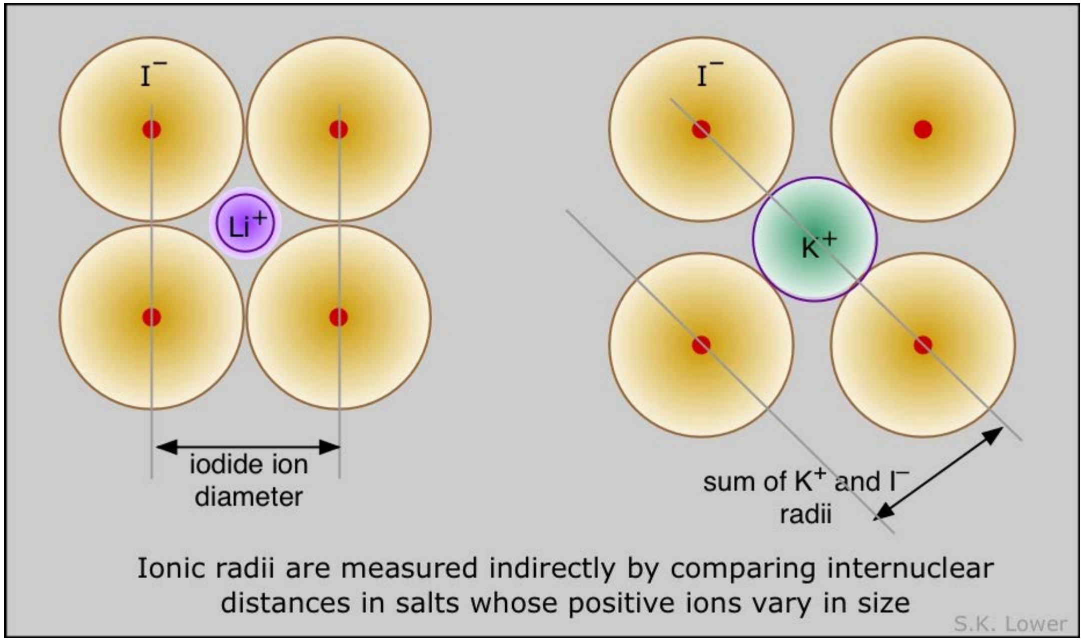


- Data
  - Basic properties
  - Standardized colors schemes
  - Size related properties
  - Electronegativity scales
  - Descriptive properties
  - Physical properties
  - Computed properties
  - Isotope properties

ML descriptor: size

- may correlate with bonding strength, charge …
- Choices



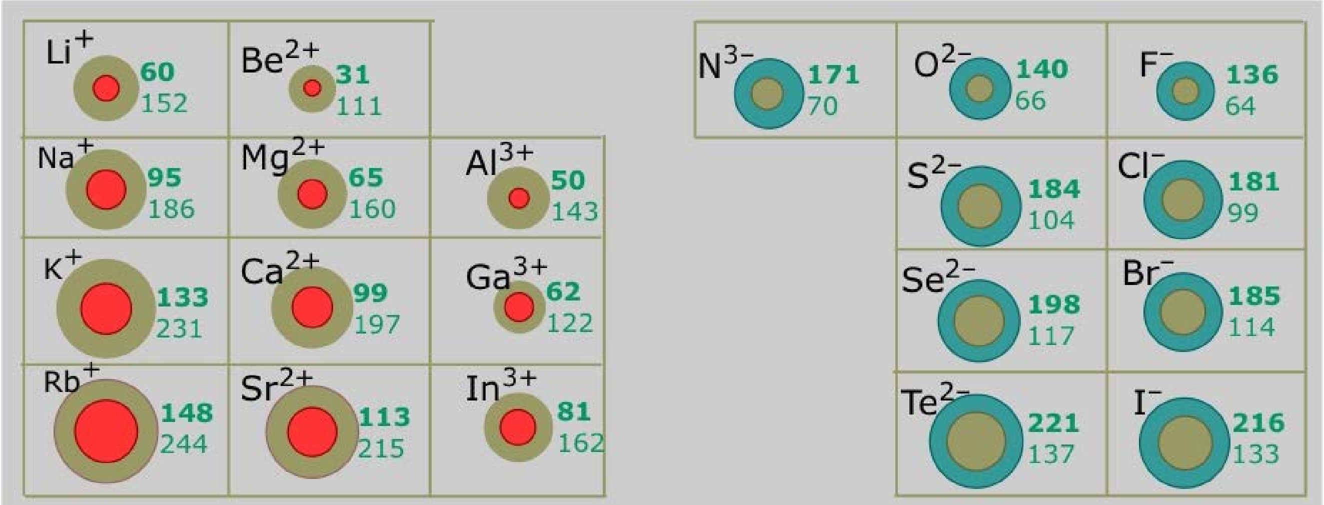


Ionic radii are measured indirectly by comparing internuclear distances in salts whose positive ions vary in size

A positive ion < neutral atom because less e- electron-electron repulsion

The ionic radius of Fe^{2+} is 76 pm, while that of Fe^{3+} is 65 pm

Negative ions > neutral atom because extra electron increases *electron-electron repulsion* which results in a general expansion of the atom



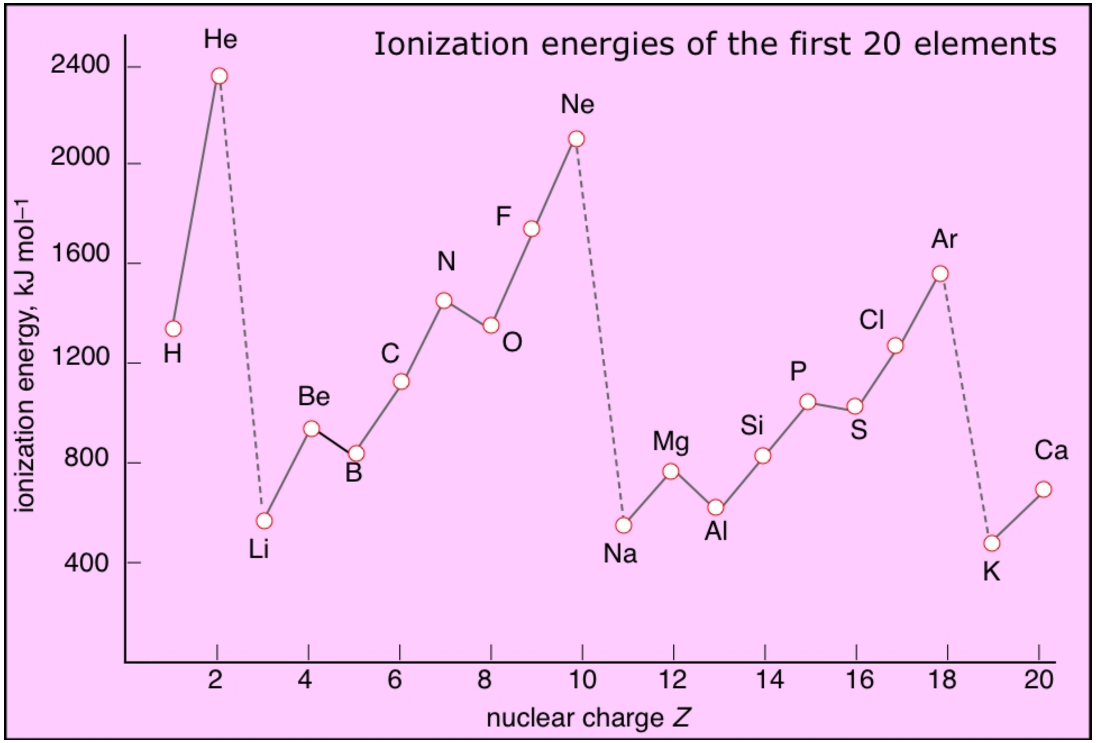
S.K. Lower

Ionic radii

Ions are colored red and blue; parent atoms brown.
Radii are in picometers.

ML descriptor: Ionization Energy

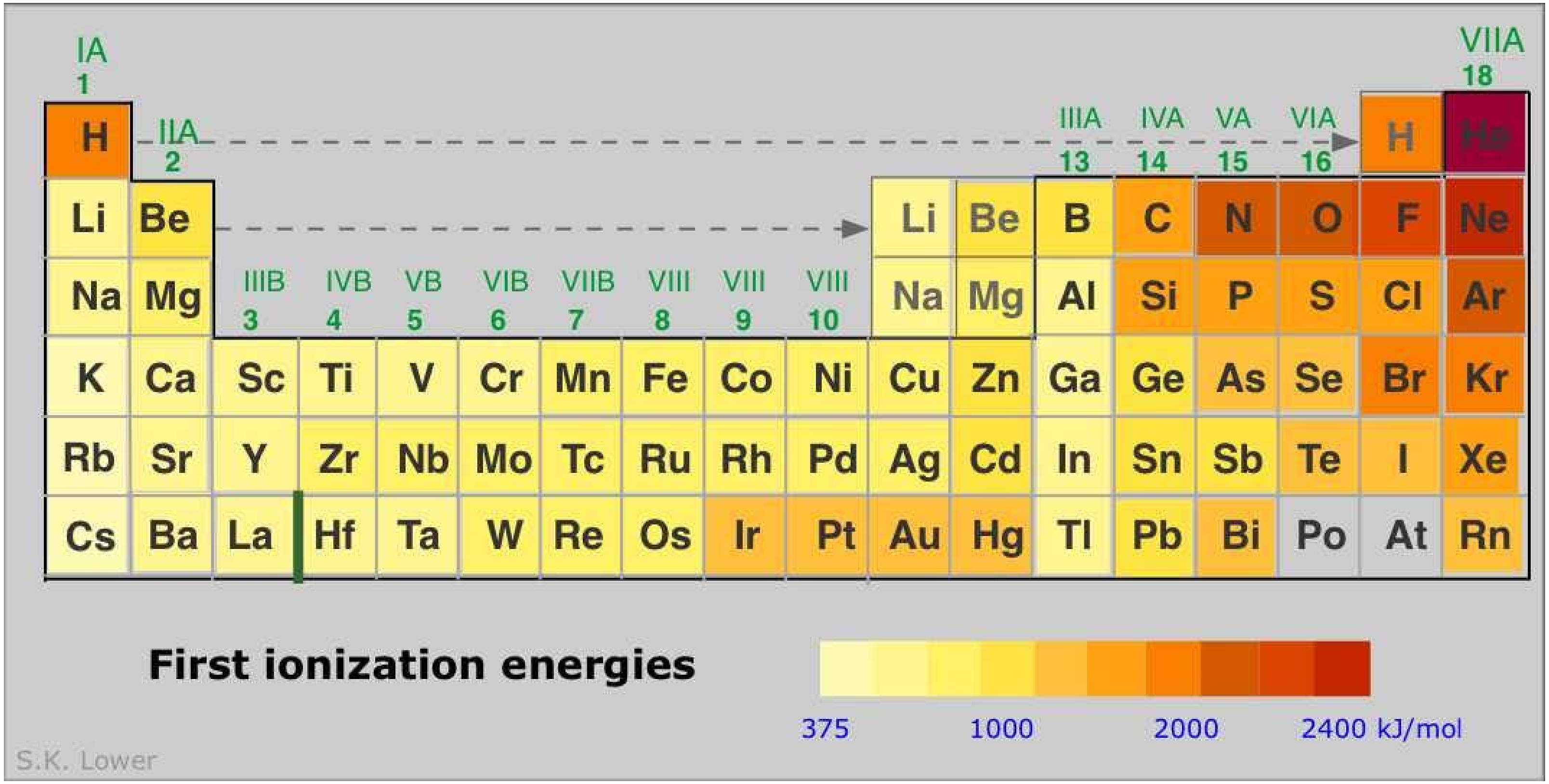
This term always refers to the formation of *positive* ions. In order to remove an electron from an atom, work must be done to overcome the electrostatic attraction between the electron and the nucleus; this work is called the *ionization energy* of the atom and corresponds to the exothermic process



Periodic Trends in ion formation

- Chemical reactions are based largely on the interactions between the most loosely bound electrons in atoms

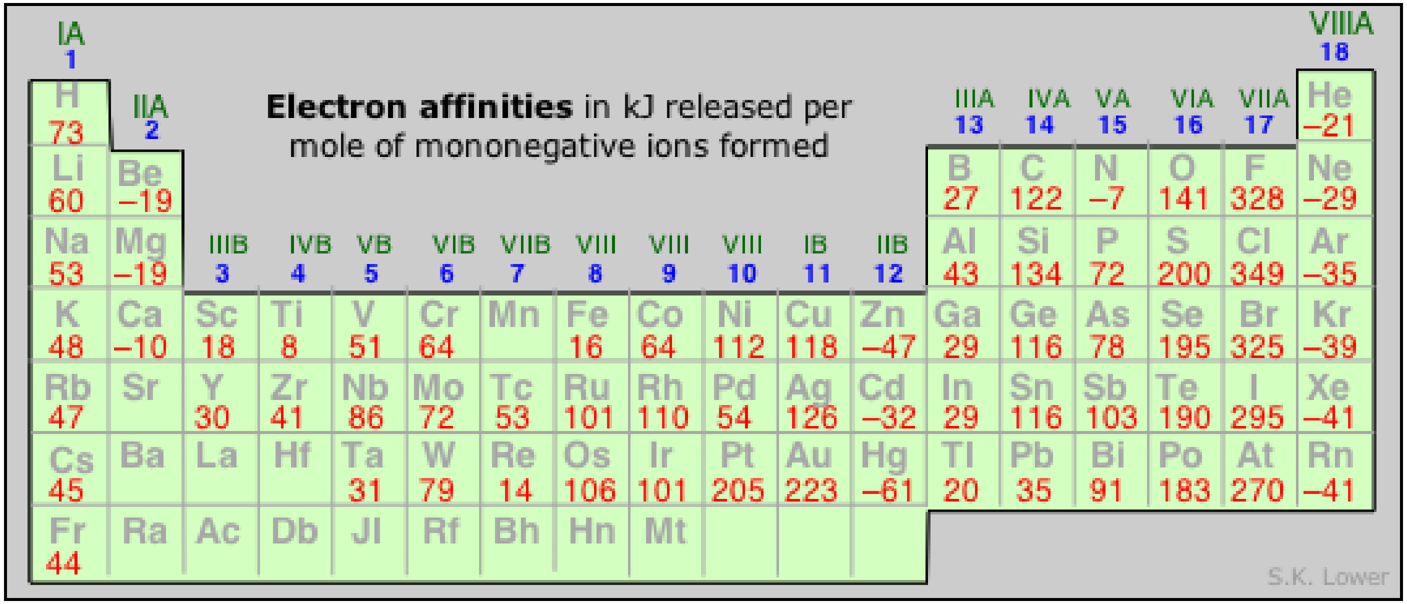
- Tendency of an atom to gain, lose or share electrons is one of its fundamental chemical properties



ML descriptor: Electron affinity

- Formation of a negative ion occurs when an electron from some external source enters the atom and become incorporated into the lowest energy orbital that possesses a vacancy (LUMO)

- Because the entering electron is attracted to the positive nucleus, the formation of negative ions is usually exothermic, i.e. the energy – **electron affinity** -- is given off



Electron affinities in kJ released per mole of mononegative ions formed

IA	IIA	IIIB	IVB	VB	VIB	VIIB	VIII	VIII	VIII	IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H 73																	He –21
Li 60	Be –19											B 27	C 122	N –7	O 141	F 328	Ne –29
Na 53	Mg –19											Al 43	Si 134	P 72	S 200	Cl 349	Ar –35
K 48	Ca –10	Sc 18	Ti 8	V 51	Cr 64	Mn	Fe 16	Co 64	Ni 112	Cu 118	Zn –47	Ga 29	Ge 116	As 78	Se 195	Br 325	Kr –39
Rb 47	Sr	Y 30	Zr 41	Nb 86	Mo 72	Tc 53	Ru 101	Rh 110	Pd 54	Ag 126	Cd –32	In 29	Sn 116	Sb 103	Te 190	I 295	Xe –41
Cs 45	Ba	La	Hf	Ta 31	W 79	Re 14	Os 106	Ir 101	Pt 205	Au 223	Hg –61	Tl 20	Pb 35	Bi 91	Po 183	At 270	Rn –41
Fr 44	Ra	Ac	Db	Jl	Rf	Bh	Hn	Mt									

S.K. Lower

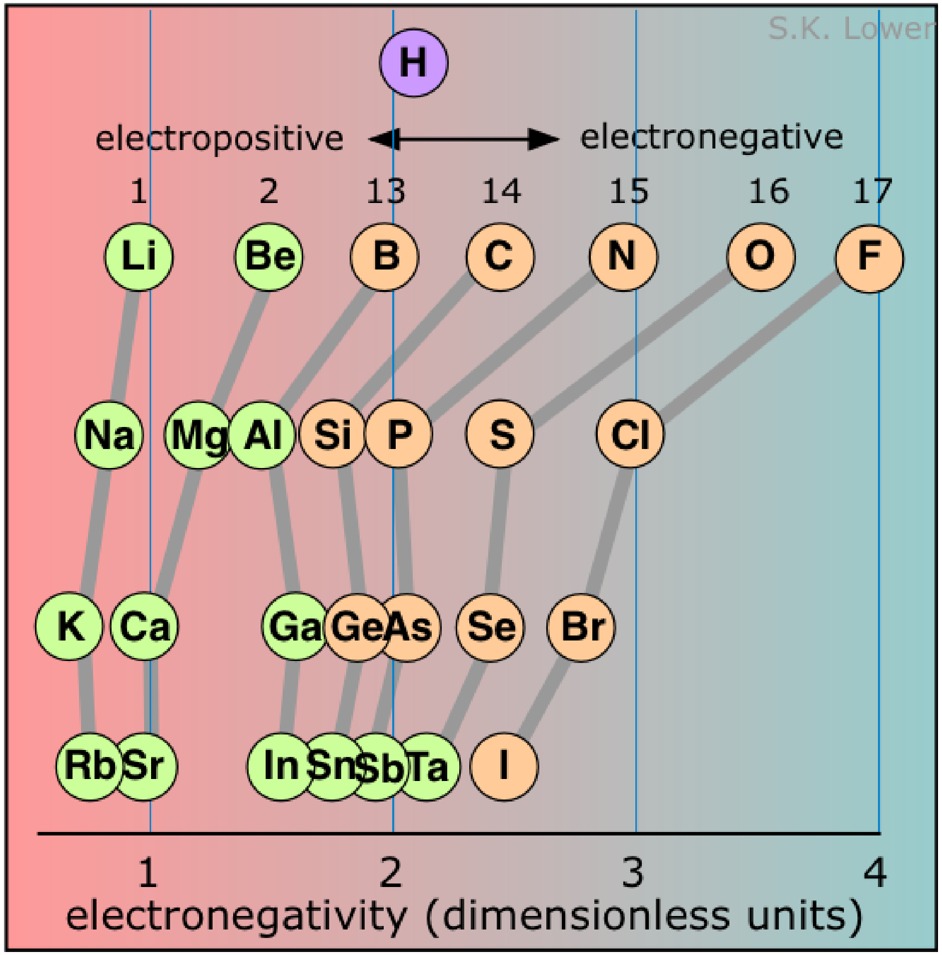
ML descriptor: Electronegativity

When two elements are joined in a chemical bond, the element that attracts the shared electrons more strongly is more electronegative

Elements with low electronegativities (the metallic elements) are said to be electropositive.

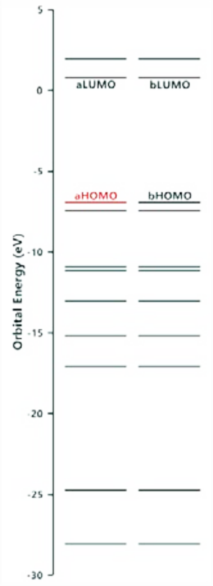
Electronegativities are properties of atoms that are chemically bound to each other, not of an isolated atom

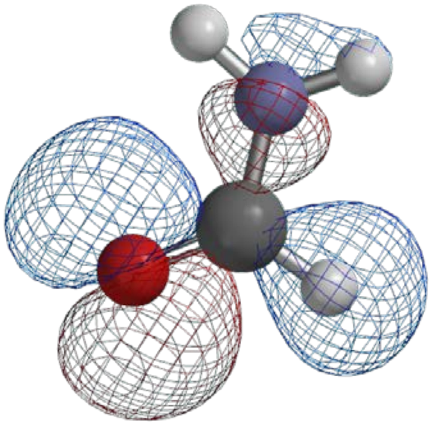
Try the Pauling scale (others are available in ‘Mendeleev’)



- Computed molecular properties
- HOMO (~ionization energy)
- LUMO (~ electron affinity)
- Dipole moment (~ stacking)
- Polarizability (response to electric field)

Absorption and emission of electromagnetic radiation (light) corresponds to the difference between the energy levels





formamide

ML descriptors or design targets:

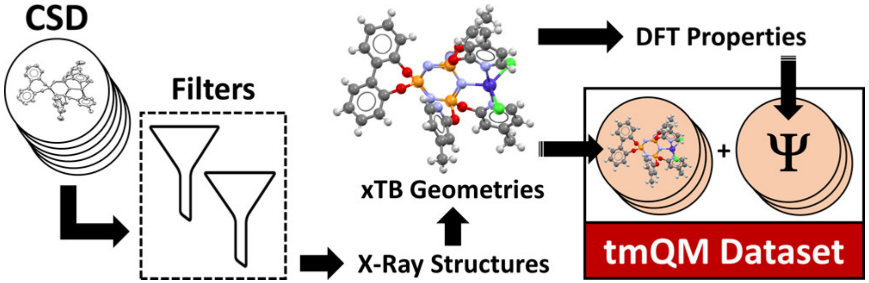
E_{HOMO} , E_{LUMO}

Energy gap = $E_{HOMO} – E_{LUMO}$

Fermi level or chemical hardness

$\eta = (E_{HOMO} + E_{LUMO})/2$

tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes



Computational protocol used to generate the tmQM data set. xTB = extended tight-binding; DFT = density functional theory; μ = dipole moment; α = polarizability; *q* = charge

- tmQM contains 86,665 complexes (we will use first 6000)
- large diversity of the TM–organic chemical space
- variety of organic ligands bound to **thirty** *3d*, *4d,* and *5d* TMs from groups 3 to 12

- Cartesian coordinates optimized at the GFN2-xTB level
- Quantum properties computed at the DFT(TPSSh-D3BJ/def2-SVP) level

  (i) the electronic and dispersion energies

  (ii) metal center natural charge

  (iii) HOMO/LUMO energies and gap

  (iv) dipole moment

  (v) polarizabilities

CSD = Cambridge Structural Database

1.25M accurate 3D structures with data from X-ray and neutron diffraction

CCS = chemical compound space

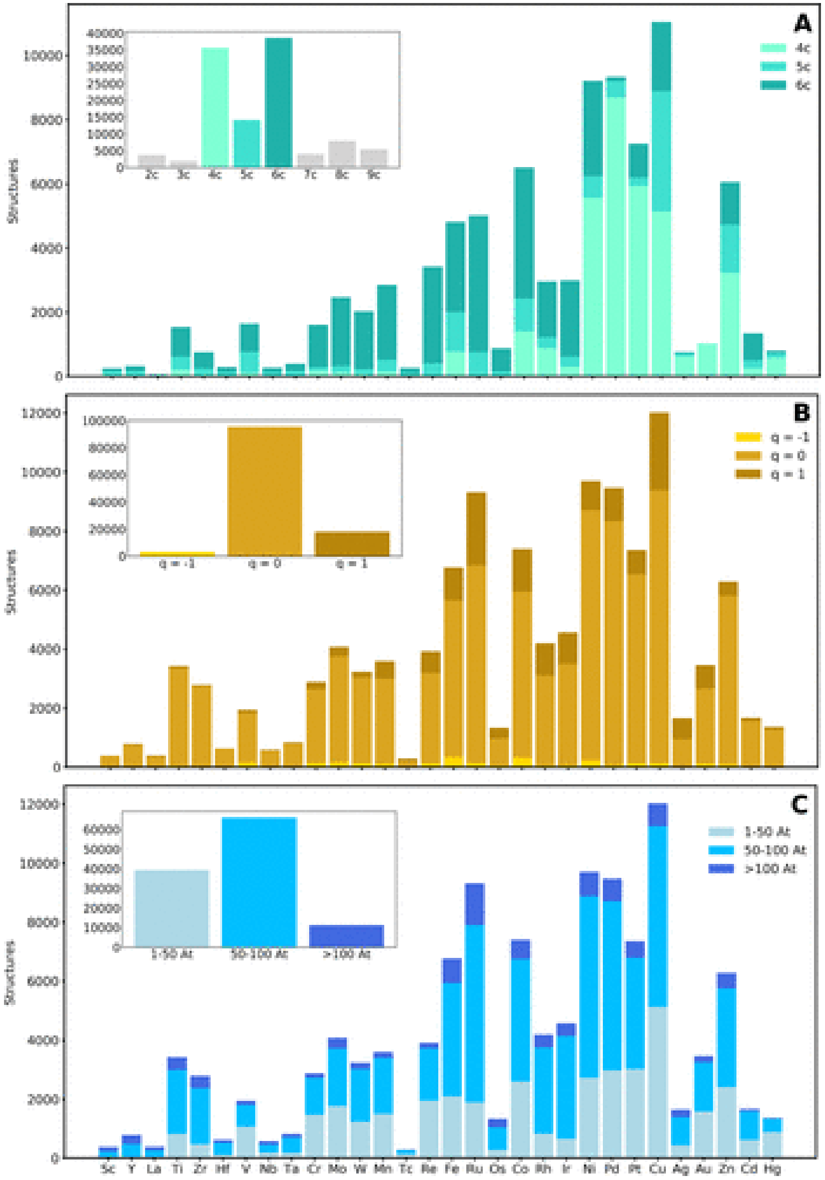
"The pairwise representations of the properties revealed unusual regions within the CCS, for example, TM complexes with large polarizabilities and wide HOMO/LUMO gaps"

Goal: What predictive models, new descriptors and correlations can you find? Can we interpret the results and learn new chemistry trends?

Data set construction: Apply filters to CSD

1. Composition filter (metal elements): Excluded all structures except those containing a single transition metal (TM)
2. Composition filter (non-metal elements): Excluded all structures except those containing a minimum of one C and one H atoms. The otherallowed elements: B, Si, N, P, As, O, S, Se, F, Cl, Br and I.
3. Components filter: Excluded the structure of all molecular components, except that of the metal complex (remove solvent and counterions)
4. Polymers filter: Excluded all polymeric structures.
5. Spatial coordinates filter: Excluded all structures without 3D-coordinates.
6. Disorder filter: Excluded all structures with disordered atoms.
7. Charge filter: Excluded all structures with charge higher than 1 and lower than -1

Result 116,332 structures



Distributions over the 3–5d TM series by

(A) metal node degree (MND), number of bonds to the metal

(B) molecular charge *q*

(C) size in number of atoms

More distributions in the paper

Using experimental structures has benefits:

chemists know how to make them!

Using quantum chemistry codes optimize geometry in gas phase: apply filters to QM data

1. Calculation did not converge
2. Computed geometry is too different from CSD (7%)
3. Exclude species odd number of electrons

---

4. Result 88,699 molecules

Do more accurate(expensive) property QM calcs

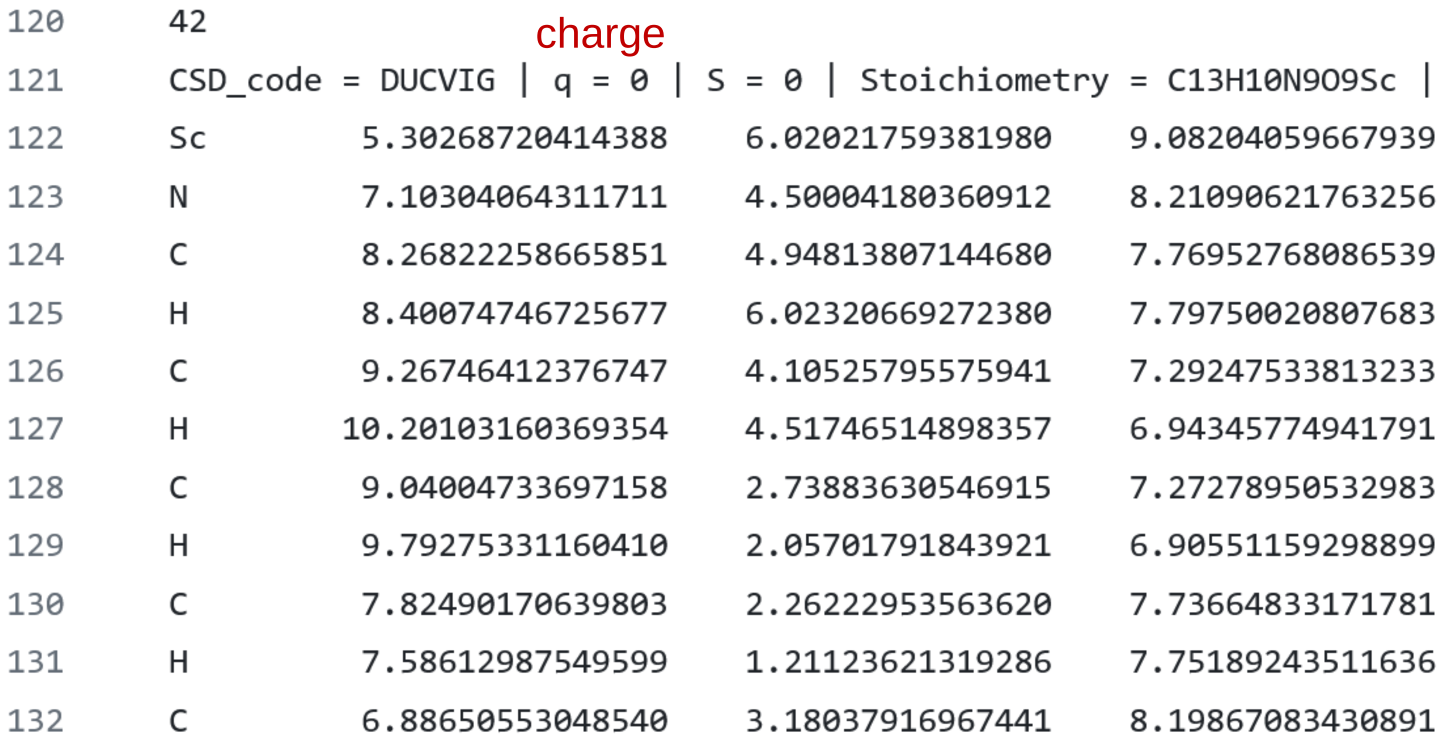
1. Electronic and dispersion energies
2. HOMO and LUMO energies
3. HOMO-LUMO gap
4. Dipole moment
5. Metal center charge
6. Polarizability

Result 88,665 molecules

Geometry tmQM_X_tiny.xyz

Number of atoms

charge

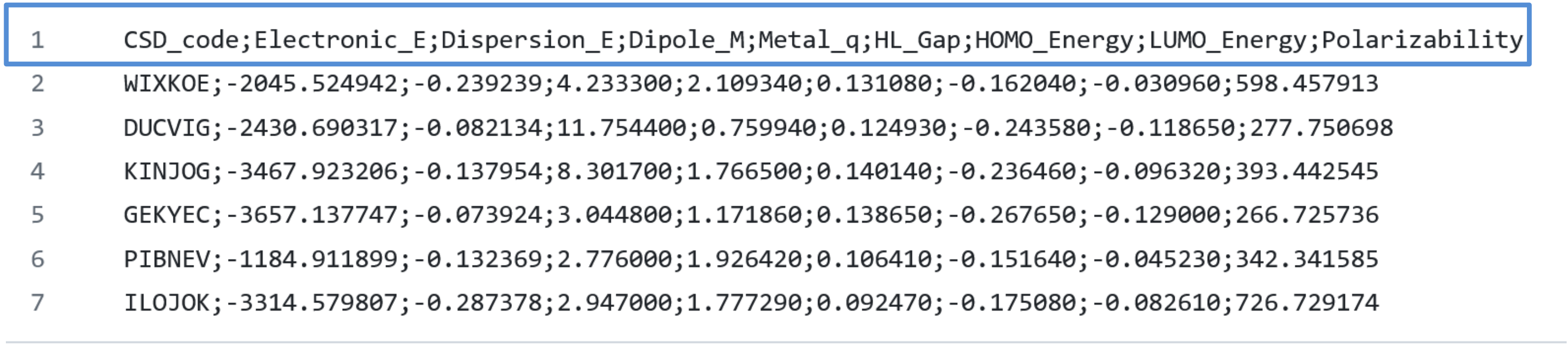


```
42
CSD_code = DUCVIG | q = 0 | S = 0 | Stoichiometry = C13H10N9O9Sc | MND = 8
Sc        5.30268720414388    6.02021759381980    9.08204059667939
N         7.10304064311711    4.50004180360912    8.21090621763256
C         8.26822258665851    4.94813807144680    7.76952768086539
H         8.40074746725677    6.02320669272380    7.79750020807683
C         9.26746412376747    4.10525795575941    7.29247533813233
H        10.20103160369354    4.51746514898357    6.94345774941791
C         9.04004733697158    2.73883630546915    7.27278950532983
H         9.79275331160410    2.05701791843921    6.90551159298899
C         7.82490170639803    2.26222953563620    7.73664833171781
H         7.58612987549599    1.21123621319286    7.75189243511636
C         6.88650553048540    3.18037916967441    8.19867083430891
```

metal node degree (MND), number of bonds to the metal

Element x y z [Angstrom]

Computed properties tmQM_y_tiny.csv

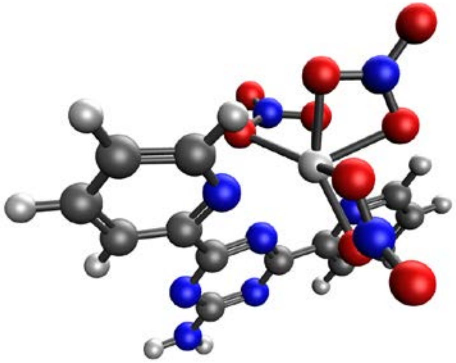


```
CSD_code;Electronic_E;Dispersion_E;Dipole_M;Metal_q;HL_Gap;HOMO_Energy;LUMO_Energy;Polarizability
WIXKOE;-2045.524942;-0.239239;4.233300;2.109340;0.131080;-0.162040;-0.030960;598.457913
DUCVIG;-2430.690317;-0.082134;11.754400;0.759940;0.124930;-0.243580;-0.118650;277.750698
KINJOG;-3467.923206;-0.137954;8.301700;1.766500;0.140140;-0.236460;-0.096320;393.442545
GEKYEC;-3657.137747;-0.073924;3.044800;1.171860;0.138650;-0.267650;-0.129000;266.725736
PIBNEV;-1184.911899;-0.132369;2.776000;1.926420;0.106410;-0.151640;-0.045230;342.341585
ILOJOK;-3314.579807;-0.287378;2.947000;1.777290;0.092470;-0.175080;-0.082610;726.729174
```

Add your own ML descriptors (analyze .xyz, use ‘mendeleev’)

Bonds to the metal:

TM size (we’ve seen several radii)
bond lengths (short or long? Scale by TM radius? you decide)
number of bonds
nearest neighbor atom type



Whole structure:

size, shape, symmetry

Metal identity:

number, group, period
Ionization energy, electron affinity, electronegativity

Everything can be converted to numbers!
Use *tiny* files to test your codes

Chemical hardness

$\eta =$ (Ehomo+Elumo)/2

can be easily added to .csv

Try η as your prediction target