CITRINE
INFORMATICS

CITRINATION FEATURES DOCUMENTATION

CHICAGO, IL
PITTSBURGH, PA
REDWOOD CITY, CA

CITRINE.IO

CITRINATION FEATURES DOCUMENTATION

TABLE OF CONTENTS

CITRINATION FEATURES

OVERVIEW

Summary

There are over 100 features that are used as inputs to machine learning models on the Citrination Platform. The purpose of this document is to provide descriptions of these features as well as some background on how they are generated.

The default features generated by the platform are labelled here as the **Standard Set** of features. If needed, more features are available in the **Extended Set**. These features are generally more expensive to compute and of less value to our models.

The **Standard Set** contains a total of 63 features and consists of three categories:

Elemental Properties – 24 features

Molecule Features – 7 features

Analytic Features – 32 features

The **Extended Set** contains a total of 71 features that can be added:

Elemental Properties – 30 features

Molecule Features – 41 features

A brief description of each type of feature can be found at the beginning of each section.

Feature Generation

Not all of the features will be generated for every material or chemical formula. If the input formula is organic, then only the Molecule Features will be generated based on the formula. If the formula is inorganic the platform will generate both the Elemental Properties and Analytic Features.

STANDARD SET

DEFAULT ELEMENTAL PROPERTIES

24 FEATURES

These are properties of pure elements. Each element has its own value for each property. To find the value for a given formula, the component elements' values are used. For example, the elemental properties of Ag, N, and O would be considered for the chemical formula $AgNO_3$. These features are generated by default for an inorganic chemical formula.

Taken from software package Magpie maintained by Wolverton Research Group.
bitbucket.org/wolverton/magpie

FORMAT:

Citrine Name – Magpie Name

Description

Elemental atomic volume - ICSDVolume

Volume per atom of ICSD phase at STP[1]

Elemental bulk modulus - BulkModulus

Bulk modulus in elasticity theory; measures resistance to compression

Elemental crystal structure (space group) - SpaceGroupNumber

Space group of T=0K ground state structure; a number that defines the crystal structure of the element in its ground state[2]

Elemental density - Density

Density of element at STP

[1] ICSD: www2.fiz-karlsruhe.de/icsd_home.html
[2] Chapter 1.3 of International Tables for Crystallography, Volume A, 6th edition

Elemental electron density - n_ws^third

Electron density at the surface of a Wigner-Sietz cell. A Wigner-Seitz cell around a lattice point is the locus of points that are closer to that lattice point than any other[3]

Elemental magnetic moment - GSmagmom

DFT magnetic moment of T=0K ground state; measures tendency to align with a magnetic field[4]

Elemental polarizability – Polarizability

Static average electric dipole polarizability; measures ability to form instantaneous electric dipoles[5]

Elemental melting temperature - MeltingT

Melting temperature of element

Elemental work function - phi

Adjusted work function; energy required to remove a free electron from the surface of the material

Mendeleev number - MendeleevNumber

Mendeleev Number. Numbers elements by column on the periodic table starting with Li (H placed at the top of column 17)[6]

Number of s valence electrons - NsValence

Number of filled s valence orbital

Number of p valence electrons - NpValence

Number of filled p valence orbitals

Number of d valence electrons - NdValence

Number of filled d valence orbitals

[3] E. Wigner and F. Seitz, Phys. Rev. 43(804), 1933
[4] OQMD: oqmd.org
[5] CRC Handbook of Chemistry and Physics
[6] Villars et al., J. Alloys Comp., 2004, 367(1-2): 167-175

Number of f valence electrons - NfValence

Number of filled f valence orbitals

Number of unfilled s valence electrons - NsUnfilled

Number of unfilled s valence orbitals

Number of unfilled p valence electrons - NpUnfilled

Number of unfilled p valence orbitals

Number of unfilled f valence electrons - NfUnfilled

Number of unfilled f valence orbitals

Pauling Electronegativity – Electronegativity

Pauling electronegativity; measures the tendency of an atom to attract electrons[7]

Radius of s orbitals - ZungerPP-r_s

Pseudopotential radius of s orbital[8]

Radius of p orbitals - ZungerPP-r_p

Pseudopotential radius of p orbital[8]

Radius of d orbitals - ZungerPP-r_d

Pseudopotential radius of d orbital[8]

Row in periodic table - Row

Row on periodic table

Total number of unfilled valence electrons - NUnfilled

Number of unfilled valence orbitals

Total number of valence electrons - NValance

Number of valence electrons

[7] L. Pauling. Journal of the American Chemical Society, 54(9): 3570-358, 1932
[8] OQMD: oqmd.org

STANDARD SET

DEFAULT MOLECULE FEATURES

7 FEATURES

These are features specific to organic molecules. The bulk of the features come from QSAR modeling (Quantitative structure-activity relationship). These features are generated by default for an organic chemical formula.

Taken from open source software package, Chemistry Development Kit: cdk.github.io

FORMAT:

Feature Name

Description

AcidGroupCount

Number of acidic groups in the molecule

AtomCount

Total number of atoms in the molecule

AtomicPolarizability

Sum of the atomic polarizabilities (including implicit hydrogens)[9]

BondCount

Total number of bonds, ignores bonds to H atoms

HBondAcceptorCount

Number of hydrogen bond acceptors in the molecule. An acceptor in a hydrogen bond is the atom to which the H is NOT covalently bonded

[9] See Elemental Polarizability in **Default Elemental Properties**

MassAutocorr

Broto-Moreau autocorrelation descriptor weighted by scaled atomic mass. Autocorrelation descriptors capture information about molecular structure and physical/chemical properties of the individual atoms[10]

PolarizabilityAutocorr

Broto-Moreau autocorrelation descriptor weighted by polarizability. Autocorrelation descriptors capture information about molecular structure and physical/chemical properties of the individual atoms[10]

[10] Moreau G. and Broto P., Nouveau Journal de Chimie, 1980, 4:359-360

STANDARD SET

ANALYTIC FEATURES

32 FEATURES

These are features calculated by Citrine that have been found to be useful in our machine learning models. Some of the Analytic Features are "elementwise" meaning each pure element has its own value. The others are specific to a formula. Many of these features are calculated using elemental properties in the Standard or Extended Set.

FORMAT:

Feature Name

Description

Elementwise Features

These features are similar to the elemental properties in that a distinct value is computed for every element in the material.

BCC Efficiency

Ratio of estimated FCC lattice parameter to estimated BCC lattice parameter

Conduction ionization energy

Difference between first ionization energy and DFT band gap energy

DFT energy density

Energy of DFT ground state divided by atomic weight. Yields energy per unit mass

DFT volume ratio

Energy of DFT ground state divided by atomic volume. Yields energy per volume of atom

Elastic Poisson Ratio

Poisson Ratio calculated using bulk and shear moduli from elasticity theory

Interatomic distance

Distance between two atoms at STP

Ionization Affinity Ratio

Ratio of electron affinity to first ionization energy

Liquid range

Difference between boiling temperature and melting temperature. Defines the range of temperatures where liquid is the stable phase

Liquid ratio

Ratio of boiling temperature to melting temperature

Miracle Ratio

Ratio of Miracle radius to covalent radius[11]

Modulii sum

Sum of bulk modulus and shear modulus

Mulliken electronegativity

Electronegativity as described by Mulliken; also called "absolute electronegativity"[12]

Non-dimensional band gap

DFT band gap energy normalized by ionization energy

Non-dimensional heat of fusion

Ratio of electron affinity to heat of fusion

Non-dimensional liquid range

Liquid range normalized by the melting temperature

Non-dimensional work function

Work function normalized by ionization energy

---

[11] See MiracleRadius in **Extended Elemental Properties**

[12] R.S. Mulliken, Journal of Chemical Physics. 2(11): 782-793, 1934

Packing density

Fraction of unit cell occupied by atoms; also called Atomic Packing Factor

Ratio of Electron Affinity to Electronegativity

Ratio of electron affinity to Pauling electronegativity

Shear Modulus Melting Temp Product

Product of shear modulus and melting temperature

Trouton's Ratio

Trouton's Ratio of heat of fusion to melting temperature

Valence electron density

Number of valence electrons divided by atomic volume

Zunger Pseudopotential radius ratio

Ratio of radius of p orbitals to radius of s orbitals

Formula Features

These features are computed based on the elemental properties of the elements in the formula. To clarify, "maximum" in these descriptions refers to the maximum property value of all of the elements in the formula.

Formula weight

Sum of all atomic weights, weighted by stoichiometric values

Maximum weight fraction

Maximum weight fraction (ratio of element's weight to the formula weight)

Minimum weight fraction

Minimum weight fraction (ratio of element's weight to the formula weight)

Maximum atomic fraction

Maximum stoichiometric value in the formula (normalized values)

Minimum atomic fraction

Minimum stoichiometric value in the formula (normalized values)

Maximum electronegativity difference

The difference between the maximum Pauling electronegativity and the minimum Pauling electronegativity

Maximum radius difference

Difference between maximum and minimum covalent radii

Maximum radius ratio

Ratio of maximum covalent radius to minimum

Min atomic radius plus max electronegativity difference

Minimum covalent radius plus the maximum electronegativity difference (see above)

Number of elements

Number of distinct elements in the formula

EXTENDED SET

EXTENDED ELEMENTAL PROPERTIES

30 FEATURES

These are properties of pure elements. Each element has its own value for each property. To find the value for a given material, the component elements' values are used. For example, the elemental properties of Ag, N, and O would be considered for the chemical formula $AgNO_3$. These features are **NOT** generated by default for an inorganic chemical formula.

Taken from software package Magpie maintained by Wolverton Research Group: bitbucket.org/wolverton/magpie

FORMAT:

Feature Name

Description

AtomicVolume

Volume of a single atom derived from atomic weight and density

AtomicWeight

Atomic mass/weight; mass of a single atom

BoilingT

Boiling temperature

Column

Column on periodic table

CovalentRadius

Covalent radius; half the bond length in an A-A bond where A is the element

ElectronAffinity

Electron affinity; amount of energy released when an electron is added to a neutral atom

FirstIonizationEnergy

Energy required to remove the first electron from an element

GSbandgap

DFT bandgap energy of T=0K ground state

GSenergy_pa

DFT energy per atom (raw VASP value) of T=0K ground state

GSestBCClatcnt

Estimated BCC lattice parameter based on the DFT volume of the ground state

GSestFCClatcnt

Estimated FCC lattice parameter based on the DFT volume of the ground state

GSvolume_pa

DFT volume per atom of the T=0K ground state

HeatCapacityMass

Heat capacity per unit mass at STP; Specific Heat Capacity; amount of energy needed to raise the temperature of a unit mass of the element by 1K

HeatCapacityMolar

Heat capacity per mole at STP; Molar Heat Capacity; amount of energy needed to raise the temperature of a mole of the element by 1K

HHIp

Herfindahl–Hirschman Index (HHI) production values. A measure of market concentration based on elemental production[13]

HHIr

Herfindahl–Hirschman Index (HHI) reserves values. A measure of market concentration based on known elemental reserves[13]

HeatFusion

Enthalpy of fusion for elements at melting temperatures per mole of atoms

[13] Gaultois et al., Chem. Mater., 2013, 25 (15), 2911–2920

IsAlkali

Whether an element is an alkali or alkali earth metal

IsDBlock

Whether an element is a d-block metal

IsFBlock

Whether an element is an f-block metal

IsMetal

Whether an element is a metal

IsMetalloid

Whether an element is a metalloid

IsNonmetal

Whether an element is a nonmetal

MiracleRadius

Assessed radii of elements in metallic glass structures[14]

NdUnfilled

Number of unfilled d valence orbitals

Number

Atomic number

OxidationStates

Observed oxidation states for each element

ShearModulus

Shear modulus in elasticity theory

[14] Miracle et al., International Materials Reviews, 2013, 55:4, 218-256

ZungerPP-r_sigma

Sum of the radii of s and p orbitals[15]

ZungerPP-r_pi

Absolute value of the different between the radii of s and p orbitals[15]

[15] OQMD: oqmd.org

EXTENDED SET

EXTENDED MOLECULE FEATURES

41 FEATURES

These are features specific to organic molecules. The bulk of the features come from QSAR modeling (Quantitative structure-activity relationship). These features are **NOT** generated by default for an organic chemical formula.

Taken from open source software package, Chemistry Development Kit: cdk.github.io

FORMAT:

Feature Name

Description

ALOGP

A combination of two values expressed as a vector: Ghose-Crippen LogKow which measures the tendency of a chemical to partition itself between organic and aqueous phases, and Ghose-Crippen molar refractivity which measures the polarizability of a mole of the molecule[16,17]

AromaticAtomCount

Total number of aromatic atoms in the molecule

AromaticBondCount

Total number of aromatic bonds in the molecule

AutocorrelationDescriptorCharge

Broto-Moreau autocorrelation descriptor weighted by electronegativity. Autocorrelation descriptors capture information about molecular structure and physical/chemical properties of the individual atoms[18]

[16] Ghose, A.K. and Crippen, G.M., Journal of Computational Chemistry, 1986, 7:565-577
[17] Ghose, A.K. and Crippen, G.M., Journal of Chemical Information and Computer Science, 1987, 27:21-35
[18] Moreau G. and Broto P., Nouveau Journal de Chimie, 1980, 4:359-360

BCUT

Eigenvalue based descriptor based on a weighted version of the Burden matrix which takes into account molecular connectivity and atomic properties. The descriptor contains values related to atomic charge, polarizability and H bond abilities[19,20]

BPol

Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens)[21]

BasicGroupCount

Number of basic groups in the molecule

CPSA

Descriptors that capture information about the molecular features responsible for polar intermolecular interactions [22]

CarbonTypes

Topological descriptor characterizing the carbon connectivity

ChiChain

Counts of simple and valence chains of orders 3–7 present in the molecule

ChiCluster

Counts of simple and valence clusters of orders 3–6 present in the molecule

ChiPathCluster

Counts of simple and valence path clusters of orders 4–6 present in the molecule

ChiPath

Counts of simple and valence paths of orders 0–7 present in the molecule

---

[19] Burden, F.R., J. Chem. Inf. Comput. Sci., 1989, 29:225-227
[20] Burden, F.R., Quant. Struct.-Act. Relat., 1997, 16:309-314
[21] See Elemental Polarizability in **Default Elemental Properties**
[22] Stanton, D.T. and Jurs, P.C., Analytical Chemistry, 1990, 62:2323-2329

EccentricConnectivityIndex

Topological descriptor combining distance and adjacency information of the molecular graph that correlates well with many physical properties[23]

FMF

Characterizes complexity of a molecule by taking the ratio of the number of heavy atoms in the Murcko framework of the molecule to the total number of heavy atoms in the molecule[24]

FragmentComplexity

Complexity of a system (See source for definition)[25]

GravitationalIndex

Characterizes the mass distribution of a molecule[26]

HBondDonorCount

Number of hydrogen bond donors in the molecule. A donor in a hydrogen bond is the atom to which the H is covalently bonded

HybridizationRatioDescriptor

Ratio of the number of sp3 carbon atoms to the total number of both sp3 and sp2 carbon atoms

KappaShapeIndices

Three Kier and Hall kappa molecular shape indices capturing information about size, degree of cyclicity, and the degree of centralization/separation in branching[27]

KierHallSmarts

Counts of electrotopological state fragments[28]

---

[23] Sharma et al., Journal of Chemical Information and Computer Sciences, 1997, 37:273-282
[24] Yang et al., J. Med. Chem., 2010, 53(21):7709-14
[25] Nilakantan et al., Journal of chemical information and modeling, 2006, 46:1069-1077
[26] Katritzky et al., J. Phys. Chem., 1996, 100:10400-10407

[27] H. Hall & L. Kier. Rev Comput Chem. 2(9): 367 – 422, 2007
[28] Butina, D., Molecules, 2004, 9:1004-1009

LargestChain

Number of atoms in the largest chain in the molecule

LargestPiSystem

Number of atoms in the largest pi system in the molecule

LengthOverBreadth

Maximum length to breadth ratio and the length to breadth ratio of the rotation (about the z-axis) of the molecule with the minimum area

LongestAliphaticChain

Number of atoms in the longest aliphatic chain

MDE

Molecular Distance Edge descriptors that contain information about distance between certain classes of atoms[29]

MannholdLogP

Prediction of logP based on the number of carbon and hetero atoms[30]

MomentOfInertia

Moment of inertia and radius of gyration values. Characterizes mass distribution of a molecule

PetitjeanNumber

Petitjean Number describing molecular shape as defined by Petitjean[31]

PetitjeanShapeIndex

Topological and geometric shape indices as defined by Petitjean and Bath et al[28&32]

[29] Liu, S. and Cao, C. and Li, Z., Journal of Chemical Information and Computer Sciences, 1998, 38:387-394
[30] Mannhold et al., J.Pharm.Sci., 2009, 98:861--893
[31] M. Petitjean, J. Chem. Inf. Comput. Sci., 1992, 32: 331-337
[32] Bath et al., Journal of Chemical Information and Computer Science, 1995, 35:714-716

RotatableBondsCount

Number of rotatable bonds. A rotatable bond is a single bond that is not in a ring and has a single nonterminal heavy atom[33]

RuleOfFive

Number of violations of the rule of five (Lipinski's Rule of Five) that determines if a compound is likely orally active drug in humans

TopologicalSurfaceArea

Calculation of topological polar surface area based on fragment contributions. Surface area of all polar atoms (and attached H's)

VertexAdjacencyMagnitude

Vertex adjacency information of a molecule. Found by taking 1 + log(m) where m is the number of heavy-heavy bonds

Volume

Van der Waals volume (see source for calculation method)[34]

WHIM

Three-dimensional molecular indices capturing holistic information and based on a variety of weighting schemes[35]

Weight

Molecular Weight

WeightedPath

Five weighted path descriptors that characterize molecular branching. Calculated based on implementation in the ADAPT software package[36]

[33] F. Veber et al., J.Med.Chem. 2002, 45, 2615-2623
[34] Zhao et al.,The Journal of Organic Chemistry, 2003, 68:7368-7373
[35] Todeschini, R. and Gramatica, P., 3D QSAR in Drug Design, 1998, 2:355-380
[36] Randic, Journal of Chemical Information and Computer Science, 1984, 24:164-175

WienerNumbers

Wiener Path Number and Wiener Polarity Number calculated from the distance matrix. Classic topological indices that capture distance information in the molecule[37]

XLogP

Prediction of logP based on the atom-type method called XLogP[38]

ZagrebIndex

Zagreb Index found by taking the sum of the squared atom degrees of all heavy atoms

[37] Wiener, Harry, Journal of the American Chemical Society, 1947, 69:17-20

[38] Wang et al., Journal of Chemical Information and Computer Sciences, 1997, 37:615-621 and Wang et al., Perspectives in Drug Discovery and Design, 2000, 19:47-66

CITRINATION FEATURES
NOMENCLATURE

ATS - Autocorrelation of a Topological Structure

BCC - Body Centered Cubic (crystal structure)

DFT - Density Functional Theory

FCC - Face Centered Cubic (crystal structure)

ICSD - Inorganic Crystal Structure Database

OQMD - Open Quantum Materials Database

QSAR - Quantitative Structure-Activity Relationship

STP - Standard Temperature and Pressure (0°C and 100 kPa)

VASP - Vienna Ab initio Simulation Package