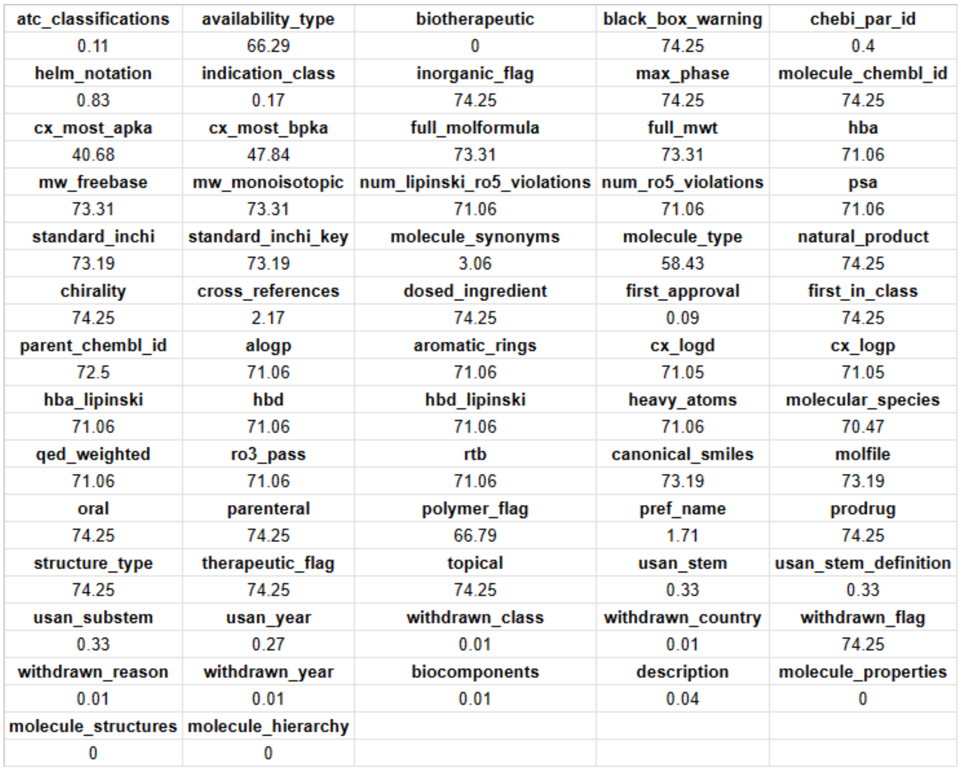
An overview of data curation activities undertaken by Adida

Data Sources

Adida is utilizing these data sources listed below to gather chemical compound records. These data sources encompass a wide range of information, and together they have yielded a total of **534,893,105** records as of **March 27, 2023**. This data can be used to investigate the properties of different compounds and to gain an understanding of the chemical makeup of materials.

Data collected as of March 27, 2023

NUMBER OF RECORDS	NOTES
ChEMBL 67 columns 2,328,900 rows 74,490,358 records	**Description**: Database of molecules with drug-like properties and biological activity **Method:** Developed a Python script using Google Colab and ChEMBL REST API. **Column Names and Data Filled Percentage:**



atc_classifications	availability_type	biotherapeutic	black_box_warning	chebi_par_id
0.11	66.29	0	74.25	0.4
helm_notation	indication_class	inorganic_flag	max_phase	molecule_chembl_id
0.83	0.17	74.25	74.25	74.25
cx_most_apka	cx_most_bpka	full_molformula	full_mwt	hba
40.68	47.84	73.31	73.31	71.06
mw_freebase	mw_monoisotopic	num_lipinski_ro5_violations	num_ro5_violations	psa
73.31	73.31	71.06	71.06	71.06
standard_inchi	standard_inchi_key	molecule_synonyms	molecule_type	natural_product
73.19	73.19	3.06	58.43	74.25
chirality	cross_references	dosed_ingredient	first_approval	first_in_class
74.25	2.17	74.25	0.09	74.25
parent_chembl_id	alogp	aromatic_rings	cx_logd	cx_logp
72.5	71.06	71.06	71.05	71.05
hba_lipinski	hbd	hbd_lipinski	heavy_atoms	molecular_species
71.06	71.06	71.06	71.06	70.47
qed_weighted	ro3_pass	rtb	canonical_smiles	molfile
71.06	71.06	71.06	73.19	73.19
oral	parenteral	polymer_flag	pref_name	prodrug
74.25	74.25	66.79	1.71	74.25
structure_type	therapeutic_flag	topical	usan_stem	usan_stem_definition
74.25	74.25	74.25	0.33	0.33
usan_substem	usan_year	withdrawn_class	withdrawn_country	withdrawn_flag
0.33	0.27	0.01	0.01	74.25
withdrawn_reason	withdrawn_year	biocomponents	description	molecule_properties
0.01	0.01	0.01	0.04	0
molecule_structures	molecule_hierarchy			
0	0			

https://www.ebi.ac.uk/chembl/

OQMD

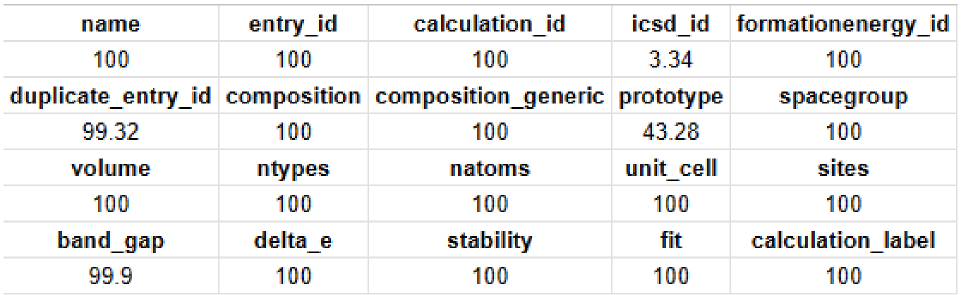
998,600
rows

20
columns

18,432,635
records

Description: The OQMD is a database of DFT calculated thermodynamic and structural properties of materials
Method: Developed a Python script using Google Colab and OQMD REST API.

Column Names and Data Filled Percentage:



name	entry_id	calculation_id	icsd_id	formationenergy_id
100	100	100	3.34	100
duplicate_entry_id	composition	composition_generic	prototype	spacegroup
99.32	100	100	43.28	100
volume	ntypes	natoms	unit_cell	sites
100	100	100	100	100
band_gap	delta_e	stability	fit	calculation_label
99.9	100	100	100	100

https://oqmd.org/

AFLOW

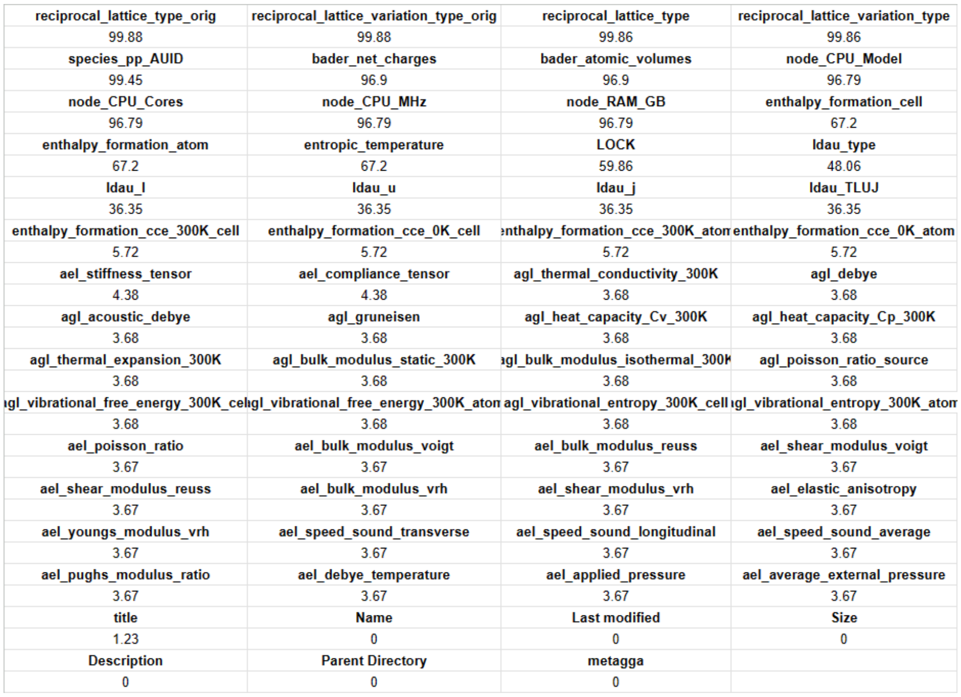
147,334
rows

194
columns

21,767,225
records

Description: Globally available database of material compounds and their calculated properties
Method: Developed a Python script using Google Colab and AFLOW API and an HTML parser.

Column Names and Data Filled Percentage:

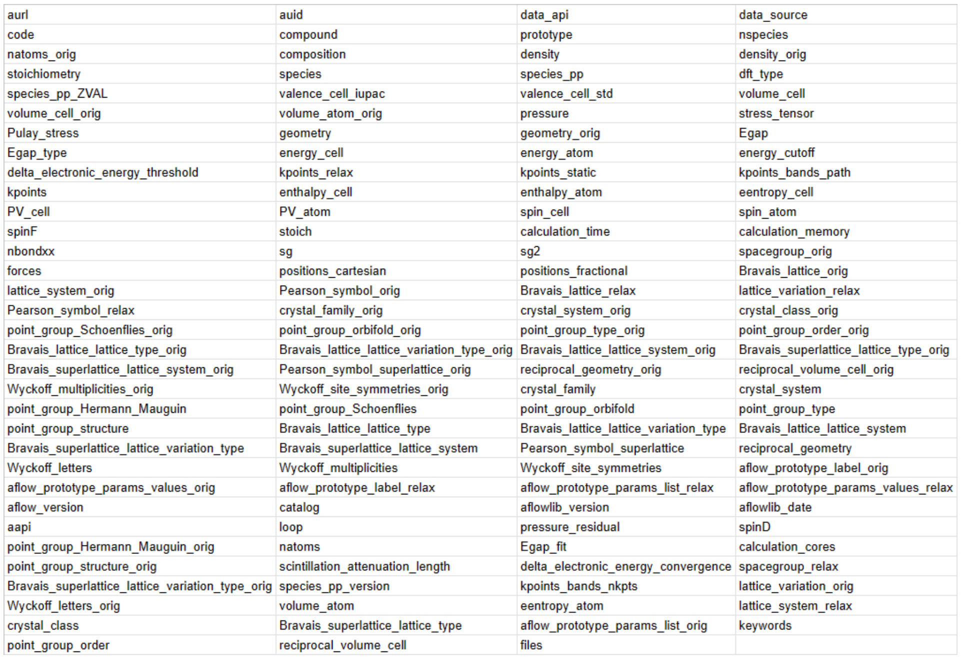


reciprocal_lattice_type_orig	reciprocal_lattice_variation_type_orig	reciprocal_lattice_type	reciprocal_lattice_variation_type
99.88	99.88	99.86	99.86
species_pp_AUID	bader_net_charges	bader_atomic_volumes	node_CPU_Model
99.45	96.9	96.9	96.79
node_CPU_Cores	node_CPU_MHz	node_RAM_GB	enthalpy_formation_cell
96.79	96.79	96.79	67.2
enthalpy_formation_atom	entropic_temperature	LOCK	ldau_type
67.2	67.2	59.86	48.06
ldau_l	ldau_u	ldau_j	ldau_TLUJ
36.35	36.35	36.35	36.35
enthalpy_formation_cce_300K_cell	enthalpy_formation_cce_0K_cell	nthalpy_formation_cce_300K_atom	enthalpy_formation_cce_0K_atom
5.72	5.72	5.72	5.72
ael_stiffness_tensor	ael_compliance_tensor	agl_thermal_conductivity_300K	agl_debye
4.38	4.38	3.68	3.68
agl_acoustic_debye	agl_gruneisen	agl_heat_capacity_Cv_300K	agl_heat_capacity_Cp_300K
3.68	3.68	3.68	3.68
agl_thermal_expansion_300K	agl_bulk_modulus_static_300K	gl_bulk_modulus_isothermal_300K	agl_poisson_ratio_source
3.68	3.68	3.68	3.68
gl_vibrational_free_energy_300K_cel	gl_vibrational_free_energy_300K_atom	agl_vibrational_entropy_300K_cell	gl_vibrational_entropy_300K_atom
3.68	3.68	3.68	3.68
ael_poisson_ratio	ael_bulk_modulus_voigt	ael_bulk_modulus_reuss	ael_shear_modulus_voigt
3.67	3.67	3.67	3.67
ael_shear_modulus_reuss	ael_bulk_modulus_vrh	ael_shear_modulus_vrh	ael_elastic_anisotropy
3.67	3.67	3.67	3.67
ael_youngs_modulus_vrh	ael_speed_sound_transverse	ael_speed_sound_longitudinal	ael_speed_sound_average
3.67	3.67	3.67	3.67
ael_pughs_modulus_ratio	ael_debye_temperature	ael_applied_pressure	ael_average_external_pressure
3.67	3.67	3.67	3.67
title	Name	Last modified	Size
1.23	0	0	0
Description	Parent Directory	metagga	
0	0	0	

https://aflowlib.org/

AFLOW

The following columns are 99.97% filled



aurl	auid	data_api	data_source
code	compound	prototype	nspecies
natoms_orig	composition	density	density_orig
stoichiometry	species	species_pp	dft_type
species_pp_ZVAL	valence_cell_iupac	valence_cell_std	volume_cell
volume_cell_orig	volume_atom_orig	pressure	stress_tensor
Pulay_stress	geometry	geometry_orig	Egap
Egap_type	energy_cell	energy_atom	energy_cutoff
delta_electronic_energy_threshold	kpoints_relax	kpoints_static	kpoints_bands_path
kpoints	enthalpy_cell	enthalpy_atom	eentropy_cell
PV_cell	PV_atom	spin_cell	spin_atom
spinF	stoich	calculation_time	calculation_memory
nbondxx	sg	sg2	spacegroup_orig
forces	positions_cartesian	positions_fractional	Bravais_lattice_orig
lattice_system_orig	Pearson_symbol_orig	Bravais_lattice_relax	lattice_variation_relax
Pearson_symbol_relax	crystal_family_orig	crystal_system_orig	crystal_class_orig
point_group_Schoenflies_orig	point_group_orbifold_orig	point_group_type_orig	point_group_order_orig
Bravais_lattice_lattice_type_orig	Bravais_lattice_lattice_variation_type_orig	Bravais_lattice_lattice_system_orig	Bravais_superlattice_lattice_type_orig
Bravais_superlattice_lattice_system_orig	Pearson_symbol_superlattice_orig	reciprocal_geometry_orig	reciprocal_volume_cell_orig
Wyckoff_multiplicities_orig	Wyckoff_site_symmetries_orig	crystal_family	crystal_system
point_group_Hermann_Mauguin	point_group_Schoenflies	point_group_orbifold	point_group_type
point_group_structure	Bravais_lattice_lattice_type	Bravais_lattice_lattice_variation_type	Bravais_lattice_lattice_system
Bravais_superlattice_lattice_variation_type	Bravais_superlattice_lattice_system	Pearson_symbol_superlattice	reciprocal_geometry
Wyckoff_letters	Wyckoff_multiplicities	Wyckoff_site_symmetries	aflow_prototype_label_orig
aflow_prototype_params_values_orig	aflow_prototype_label_relax	aflow_prototype_params_list_relax	aflow_prototype_params_values_relax
aflow_version	catalog	aflowlib_version	aflowlib_date
aapi	loop	pressure_residual	spinD
point_group_Hermann_Mauguin_orig	natoms	Egap_fit	calculation_cores
point_group_structure_orig	scintillation_attenuation_length	delta_electronic_energy_convergence	spacegroup_relax
Bravais_superlattice_lattice_variation_type_orig	species_pp_version	kpoints_bands_nkpts	lattice_variation_orig
Wyckoff_letters_orig	volume_atom	eentropy_atom	lattice_system_relax
crystal_class	Bravais_superlattice_lattice_type	aflow_prototype_params_list_orig	keywords
point_group_order	reciprocal_volume_cell	files	

Materials Project

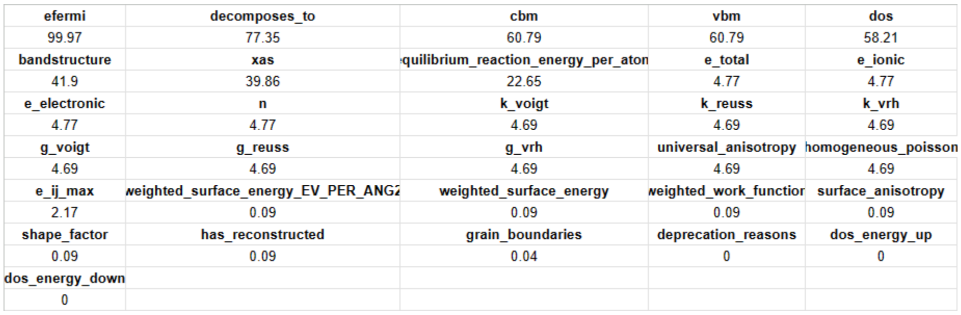
151,247
rows

74
columns

7,291,433
records

Description: The database is free to use and includes information on thousands of materials, including their chemical composition, physical properties, and other relevant data.
Method: Developed a Python script using Google Colab and MP REST API Python library. Needs API key to work.

Column Names and Data Filled Percentage:



efermi	**decomposes_to**	**cbm**	**vbm**	**dos**
99.97	77.35	60.79	60.79	58.21
bandstructure	**xas**	**quilibrium_reaction_energy_per_aton**	**e_total**	**e_ionic**
41.9	39.86	22.65	4.77	4.77
e_electronic	**n**	**k_voigt**	**k_reuss**	**k_vrh**
4.77	4.77	4.69	4.69	4.69
g_voigt	**g_reuss**	**g_vrh**	**universal_anisotropy**	**homogeneous_poisson**
4.69	4.69	4.69	4.69	4.69
e_ij_max	**weighted_surface_energy_EV_PER_ANG2**	**weighted_surface_energy**	**weighted_work_functior**	**surface_anisotropy**
2.17	0.09	0.09	0.09	0.09
shape_factor	**has_reconstructed**	**grain_boundaries**	**deprecation_reasons**	**dos_energy_up**
0.09	0.09	0.04	0	0
dos_energy_down				
0				

The following column names are 100% filled
Builder_meta, nsites, elements, nelements, composition, composition_reduced, formula_pretty, formula_anonymous, chemsys, volume, density, density_atomic, symmetry, property_name, material_id, deprecated, last_updated, origins, warnings, structure, task_ids, uncorrected_energy_per_atom, energy_per_atom, formation_energy_per_atom, energy_above_hull, is_stable, band_gap, is_gap_direct, is_metal, es_source_calc_id, is_magnetic, ordering, total_magnetization, total_magnetization_normalized_vol, total_magnetization_normalized_formula_units, num_magnetic_sites, num_unique_magnetic_sites, types_of_magnetic_species, possible_species, has_props, theoretical, database_IDs, fields_not_requested

https://materialsproject.org/

PubChem

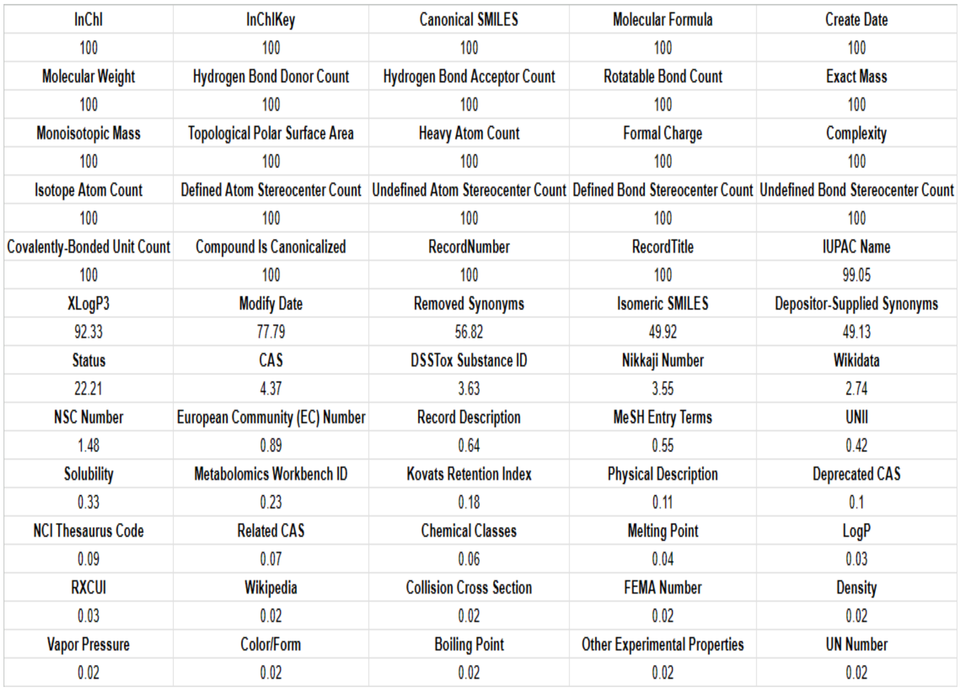
14,346,971
rows

94
columns

411,344,654
records

Description: PubChem is a database of chemical structure, physical properties, biological activities, and reactions
Method: Developed a Python script running in a Virtual Private Server using PubChem REST API to gather records.

Column Names and Data Filled Percentage:



InChI	InChIKey	Canonical SMILES	Molecular Formula	Create Date
100	100	100	100	100
Molecular Weight	Hydrogen Bond Donor Count	Hydrogen Bond Acceptor Count	Rotatable Bond Count	Exact Mass
100	100	100	100	100
Monoisotopic Mass	Topological Polar Surface Area	Heavy Atom Count	Formal Charge	Complexity
100	100	100	100	100
Isotope Atom Count	Defined Atom Stereocenter Count	Undefined Atom Stereocenter Count	Defined Bond Stereocenter Count	Undefined Bond Stereocenter Count
100	100	100	100	100
Covalently-Bonded Unit Count	Compound Is Canonicalized	RecordNumber	RecordTitle	IUPAC Name
100	100	100	100	99.05
XLogP3	Modify Date	Removed Synonyms	Isomeric SMILES	Depositor-Supplied Synonyms
92.33	77.79	56.82	49.92	49.13
Status	CAS	DSSTox Substance ID	Nikkaji Number	Wikidata
22.21	4.37	3.63	3.55	2.74
NSC Number	European Community (EC) Number	Record Description	MeSH Entry Terms	UNII
1.48	0.89	0.64	0.55	0.42
Solubility	Metabolomics Workbench ID	Kovats Retention Index	Physical Description	Deprecated CAS
0.33	0.23	0.18	0.11	0.1
NCI Thesaurus Code	Related CAS	Chemical Classes	Melting Point	LogP
0.09	0.07	0.06	0.04	0.03
RXCUI	Wikipedia	Collision Cross Section	FEMA Number	Density
0.03	0.02	0.02	0.02	0.02
Vapor Pressure	Color/Form	Boiling Point	Other Experimental Properties	UN Number
0.02	0.02	0.02	0.02	0.02

The remaining columns are less than 0.02% filled
GlyTouCan Accession,JECFA Number, Refractive Index,Dissociation Constants,ICSC Number, RTECS Number, Odor,Stability/Shelf Life,Decomposition,Henry's Law Constant, DEA Code Number,Corrosivity, pH, Optical Rotation, Flash Point,Autoignition Temperature, Vapor Density, Taste, Ionization Efficiency, Acid Value, Viscosity, Heat of Combustion, Surface Tension, Odor Threshold, Polymerization, Hydrophobicity, Isoelectric Point, Caco2 Permeability, Heat of Vaporization, Relative Evaporation Rate, LogS, Enthalpy of Sublimation, Ionization Potential, Dielectric Constant

https://pubchem.ncbi.nlm.nih.gov/

PubChem database consists of over 100 million compounds. The collection of data is still in process.

Open Citrination Datasets

NUMBER OF RECORDS

NOTES

Dataset 150146

50000 rows

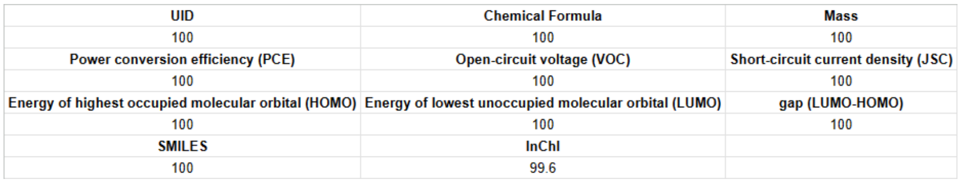
11 columns

549,798 records

HARVARD CLEAN ENERGY PROJECT
DATASET NO. 150146

Description: Properties in this dataset include principle energy levels, photovoltaic performance parameters, mass and stoichiometric formulas. Top properties: Power conversion efficiency energy of the highest occupied molecular orbital energy of the lowest occupied molecular orbital
Method: Developed a Python script using Google Colab and Citrination Python client and an API key.

Column Names and Data Filled Percentage:



UID	Chemical Formula	Mass
100	100	100
Power conversion efficiency (PCE)	Open-circuit voltage (VOC)	Short-circuit current density (JSC)
100	100	100
Energy of highest occupied molecular orbital (HOMO)	Energy of lowest unoccupied molecular orbital (LUMO)	gap (LUMO-HOMO)
100	100	100
SMILES	InChI	
100	99.6	

https://citrination.com/datasets/150146/

Dataset 114201

11,644 rows

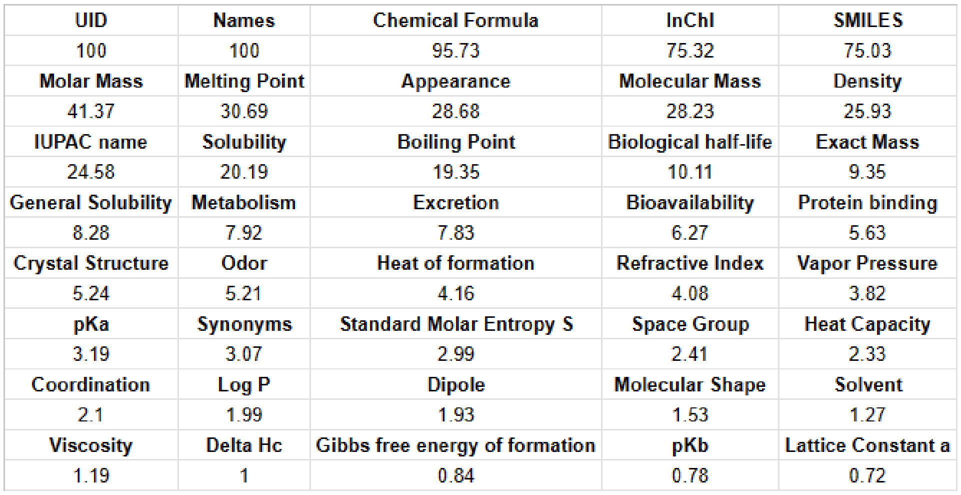
294 columns

90,811 records

WIKIPEDIA
DATASET NO. 114201

Description: Chemical and pharmacology data from English Wikipedia. Top properties: Molecular mass Density Melting point
Method: Developed a Python script using Google Colab and Citrination Python client and an API key

Column Names and Data Filled Percentage:



UID	Names	Chemical Formula	InChI	SMILES
100	100	95.73	75.32	75.03
Molar Mass	Melting Point	Appearance	Molecular Mass	Density
41.37	30.69	28.68	28.23	25.93
IUPAC name	Solubility	Boiling Point	Biological half-life	Exact Mass
24.58	20.19	19.35	10.11	9.35
General Solubility	Metabolism	Excretion	Bioavailability	Protein binding
8.28	7.92	7.83	6.27	5.63
Crystal Structure	Odor	Heat of formation	Refractive Index	Vapor Pressure
5.24	5.21	4.16	4.08	3.82
pKa	Synonyms	Standard Molar Entropy S	Space Group	Heat Capacity
3.19	3.07	2.99	2.41	2.33
Coordination	Log P	Dipole	Molecular Shape	Solvent
2.1	1.99	1.93	1.53	1.27
Viscosity	Delta Hc	Gibbs free energy of formation	pKb	Lattice Constant a
1.19	1	0.84	0.78	0.72

https://citrination.com/datasets/114201/

Dataset 114201	**The remaining columns are less than or equal to 0.6% filled** Henry Constant, Molecular Shape, Band Gap, log P, Electron Mobility, Coordination, Lattice Constant a, Thermal Conductivity, Lattice Constant alpha, Lattice Constant beta, Lattice Constant gamma, Odour, Dipole, Point Group, Specific Surface Area, Pore Volume, Average Pore Size, Boiling Point 2, Solubility in dimethylformamide, Solubility in glycerol, Solubility Product, Melting Point F, Solubility in methanol, Solubility in ethyl acetate, Lattice Constant b, Lattice Constant c, Crystal system, Formula, Monoisotopic Mass, Tg, Young's Modulus, Atmospheric OHRate Constant, center, released, pmid, versioning, curation, Melting notes, Pearson symbol, Onset of actions, Solubility in ethanol, Solubility in diethyl ether, pKb, Solubility in Dimethylformamide, Solubility in Dimethyl sulfoxide, Solubility in Sulfolane, Solubility in Methylpyrrolidone, Solvent 1, Critical Temperature, Critical Pressure, download, frequency, appearance, Melting poit, boxwidth, heat, scoville, drug name, type, component 1, class 1, component 2, class 2, tradename, Optical activity, Solubility in glycerin, Solubility in propylene glycol, month, Structure, Crit Temp, p a, Solubility in amyl acetate, "Solubility in 1, 4-Dioxane", 6.351, Surface Tension, Solubility in glycerine, Solubility in ethanediol, Magnetic Susceptibility, Enthalpy of Formation, Solubility in hexane, Solubility in carbon tetrachloride, LD50, p sp, Solubility in ethylene glycol, Solubility in ammonium hydroxide, Orbital Hybridisation, 69, Solubility in Water, Solubility in water, chapter, Solvent 2, Solubility in isopropanol, -gaseous:-(21481 pm 18), "-gaseous:-280, 4", "Solubility in ethanol, acetone, diethyl ether, ", Heat of fusion, Heat of vaporization, Absorbance, Refractive index, Solubility in acetic acid, Ph, Solubility in chloroform, Solubility in slightly soluble in alcohol, Solubility in formamide, Solubility in hydroxylamine, Isoelectric Point, routes of administration, Solubility in water, Original research, Hydrogen Bond Donor, Hydrogen Bond Acceptor, Physiological charge, work, Dipole Moment, "Solubility in benzene, THF", Melting Point V, "Solubility in alcohol, ether, benzene, acetic acid", Solubility in butyric acid, Boiling Point F, Molar Mass notes, Solubility in selenium(IV) oxychloride, Solubility in dichloromethane, Critical Relative Humidity, "Solubility in acetone, toluene, octane", Spec Rotation, Solubility in sulfuric acid, Solubility in hydrochloric acid, Solubility in Xylene, Solubility in Acetone, Solubility in Ethyl acetate, Solubility in 1-Octanol, C, H, N, O, Boilin notes, Sublimation Conditions, Boiling notes, Vapor Density, Solubility in formic acid, Solubility in dimethyl sulfoxide, Solubility in tetrahydrofuran, Solubility in ester, Solubility in arene, Flash Point, Solubility in chlorine, Lambda-max, Solubility in sulfur dioxide, Solubility in liquid ammonia, -align, 32.4g/100mL at 0, 60.4g/100mL at 100, Soluble 2, Point group, 15-6-3-12(4-7-15), Solubility Constant, "General Solubility-luene, benzene, dichloromethane, ", Average mass, Solubility in alcohol, specific rotation, Solubility in Ethanol, location, Molar Volume, "Solubility in alcohol, acetone, ammonia", "Solubility in benzene, ether, ethyl ether, sulfuric acid", 0.7904-0.7928, Solubility in chlorocarbons, "Solubility in acetone, methanol, ethanol, glycerol, ethyl acetate", Band gap, Pub Chem, Chem Spider ID, Solubility in methyl acetate, Viscocity, Boiling PCt, pmc, Color, Formula Weight, Henry's Law Constant, SrI2c6H2O(hexahydrate), Solubility in all organic solvents, flash Point, logp, surface tension, enthalpy of vaporization, Solubility Product As, Exact MAss, Acidity (p H), C27H31O16+, Solubility in Ammonia, Isoelectric point, Triple point, Triple point kPa, critical point, pka 2, Dipole moment, Optical Rotary Power, Fluorescence, Protein bound, Solubility in olive oil, Solubility in ammonia, chemical description, Solubility in silver nitrate, Ionicity, Solubility in acetonitrile, Solubility in octanol, Solubility 7, Charge, Boiling Point c, Diplole Moment, Critical Point, Solubility in pyridine, Flashpt, Crystal Struc, Hole Mobility, "Solubility in ether, benzene", Solubility in acids, Enthalpy of vaporization, Solubility in bases, Boiling comment, Solubility in oils, sp, Melting Point H, sp (hemihydrate), Solubility in benzoyl chloride, Solubility in p-Cymene, Solubility in diiodomethane, Solubility in diethylether, "Solubility in isopropanol, acetone, xylene", T c, P c, V c, G f, H f, H v, Ground electronic state bond length, Ground electronic state dissociation energy, Solubility in Diethyl ether, Solubility in Chloroform, Melting Point L, "Solubility in eth, ace, CH3CN", PEL, IDLH, REL, Main Hazards, "Solubility in ethanol, diethyl ether, ", Solubility in dilute acids, Melting PC, r(Si-H), IUPACName

Dataset 163382

35215 rows

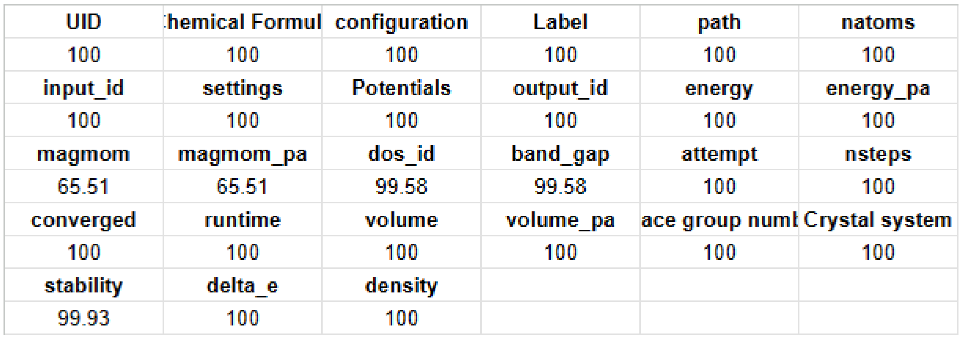
27 columns

926,191 records

OQMD
DATASET NO. 163382

Description: This database includes DFT calculations on a large number of metallic compounds, energy materials, and also mechanical properties of many materials.
Method: Developed a Python script using Google Colab and Citrination Python client and an API key.

Column Names and Data Filled Percentage:



UID	:hemical Formul	configuration	Label	path	natoms
100	100	100	100	100	100
input_id	settings	Potentials	output_id	energy	energy_pa
100	100	100	100	100	100
magmom	magmom_pa	dos_id	band_gap	attempt	nsteps
65.51	65.51	99.58	99.58	100	100
converged	runtime	volume	volume_pa	ace group numl	Crystal system
100	100	100	100	100	100
stability	delta_e	density			
99.93	100	100			

https://citrination.com/datasets/163382/

Each Open Citrination datasets listed above has its own script to capture records but uses the same API key to access the dataset.

As of this date, the Open Citrination was decommissioned. Some of the datasets are not accessible anymore, but some are available to download in JSON format. Due to that event, the Citrination Python client will also stop working. Development of a new script to parse the JSON is proceeding.

All collected data are saved in a compressed pickle or compressed CSV format. We have written a script to read and convert those data into a DataFrame that can be exported to a MySQL database.

534,893,105
Total Collected Records

Collection of records from various open databases of chemical compounds is still ongoing. Tools to help gather these records are always in constant development to ensure the quality of data.