Perovskite Dataset Description

Dataset Overview

“Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management”
DOI: https://doi.org/10.1557/mrc.2019.72

Explanation of datasets

There are two datasets stored in “data” folder: “**0042.perovskitedata_RAPID.csv**” and “**0042.perovskitedata_RAPID_full.csv**”. **They contain the same experiments but only “0042.perovskitedata_RAPID.csv” is used for data analysis and machine learning**. (see explanations of files in README) In both datasets, each row corresponds to an individual experiment (a single perovskite synthesis reaction). Each column in the datasets contains the values of a specific feature for all experiments and has a column header with prefix (e.g., “_raw_”, “_rxn_”) to identify the feature type. The categorization of features is further explained in “Explanation of Header Prefixes” section below.

The only difference between the two datasets is the feature set they contain for the experiments. “0042.perovskitedata_RAPID_full.csv” includes all features: “_raw_”, “_rxn_”, and “_feat_” while “0042.perovskitedata_RAPID.csv” includes only “_rxn_” and “_feat_” features. Generally, the “_raw_” features are experiment details which are not useful in machine learning. The “_rxn_” features are the experimental conditions which are considered to affect experiment outcome. The “_feat_” features are calculated chemical descriptors of organoammonium used in each reaction. All features and descriptors will be explained in detail in the “Explanation of Features-Descriptors” section.

Explanation of Header Prefixes

raw

DO NOT LEARN ON. We strongly recommend not using these features. Raw data associated with the experiment that was performed. Combined with the other header prefixes, these columns describe the complete set of all data acquired during an experimental run.

rxn

Recommended set of data to learn on. These descriptors include experimental observables, reaction conditions, and calculated molecular features. More elaborated descriptions can be found below for individual headers.

feat or _calc_

Recommended set of data to learn on

prototype

New features that could be used with caution. These are under development and likely require additional integrity testing. Specific features may require additional process or special handling before consumption by typical ML tools. We didn’t used this feature in our ML analysis.

out

Target outputs to predict. These are a numeric representation of the manually scored crystal quality of a single reaction rated on a scale of 1 to 4. For the organohalide perovskite chemistry, a “4” indicates that large crystals were observed; this is the ideal outcome of the experiment. A “3” means small crystallines were observed; still an indicator of potentially useful material, but less desirable than a “4”. A “2” indicates that fine powder were observed. A “1” represents no observable formation of solid (clear solution).

If used for binary classification, the outcomes can be changed to booleans reasonably in two ways:

- 1s and 2s can be rated failures (0) and 3s and 4s can be rated successes (1).
- 1s, 2s, and 3s are rated as failures (0) and 4s only are successes (1) (Our ML in the paper is based on this binary classification)

Ideally, the goal would be to predict only 4s as successes, but it is acceptable to tackle the potentially easier problem of predicting both 3s and 4s as successes first.

General Notes: _raw_ to _rxn_

The non-’_raw_’ training subset was selected from the total raw data set as follows

All conditions must be satisfied for data to be pulled from the RAW set to the curated data set:

1. Supplies the model with features we would like to optimize
    - Examples of *useful* features: chemical descriptors, mmol of reactant (i.e. not solvent mmol), well temperature, volume of pure chemicals, total volumes of solutions which vary between experiments
    - Examples of *less useful* features: Operator name, run date, run id, grams of a chemical in a reagent, reagent preparation data, etc.
2. Captures variance in the current combined set of experiments
3. Proven to be easily implementable or commonly understood
4. Has been implemented successfully for new modeling campaigns without significant struggle

General Nomenclature

- **Well**
  - The location on the tray where the reaction/experiment is taking place. Some properties of wells vary throughout the tray such as temperature.
- **Experiment/Reaction**
  - A specific test which is described by the properties of the environment and the reagents/chemicals added to the “well” in which the experiment is taking place.
- **Organic**
  - In the case of the perovskite chemistry the term “organic” is referring to the ammonium salts used in the experiment.
- **Inorganic**
  - In the case of the perovskite chemistry the term “inorganic” is referring to the metal used to form the perovskite. For workflow in RAPID inorganic only refers to lead diiodide (PbI_2)
- **Reagent**
  - A chemical or combination of chemicals which create the solution added by the robot to the well.
- **Chemical**
  - A compound which can be defined by an InChIkey. Can be of various qualities and purities, but should be primarily defined by the core component molecule. This is the most granular definition of what is added to each experiment and thereby, each well.
- **Solvent**
  - The chemical used to solvate the organic and inorganic component of the reaction facilitating transfer on the robot.Namespace in the CSV:

Explanation of Features-Descriptors

The following section provides a general description of each of the features used in the current dataset.

_RunID_vial

- This is the first column in the dataset. It records unique experiment ID each reaction. For example, in “2019-09-30T17_16_09.419113+00_00_LBL_A1”, “2019-09-30T17_16_09” records the year-month-date and time when the experiment is generated by ESCALATE. “419113+00_00_LBL” is a unique string for each microplate of experiment. “A1” is the specific location of the reaction in the microplate of experiment.

__out_crystalscore

It records reaction outcomes which are scored into four classes:

- **1**: clear solution without any solid.
- **2**: fine powder.
- **3**: small crystallites (average crystal dimension < 0.1 mm).
- **4**: large (> 0.1 mm) crystals suitable for structure determination by single crystal X-ray diffraction.

_rxn_organic_inchikey

This column specifies the INCHI-key the identity of the organoammonium iodide. This key can be used to look up the chemical formula and chemical name in the inventory.csv file located in the same folder as this document.

rxn → Primary Experimental Descriptions

- _rxn_M_inorganic
  - This column specifies the molarity of the inorganic component (PbI_2) in the reaction solution (unit: molar/liter). (See selection of compounds in Figure S4)
- _rxn_M_acid
  - This column specifies the molarity of the formic acid in the reaction solution (unit: molar/liter). (See selection of compounds in Figure S4)
- _rxn_M_organic
  - This column specifies the molarity of the organic component (organoammonium iodide) in the reaction solution (unit: molar/liter). (See selection of compounds in Figure S4) Descriptions of the variance in the organoammonium iodide are found

in the _feat_ section of the dataset. The identity of the organic cation is described in _rxn_organic_inchikey.

- _rxn_mixingtime1S
    - This column specifies the duration of the first mixing time (in seconds) after the addition of solvent, organic component, inorganic component, and the first addition of formic acid. (See Robotic Workflow section in the SI for more information)
- _rxn_mixingtime2S
    - This column specifies the duration of the second mixing time (in seconds) after the addition of the second portion of formic acid. (See Robotic Workflow section in the SI for more information)
- _rxn_reactiontimeS
    - This column specifies the duration of time (in seconds) that the reaction was are heated undisturbed for to allow for crystal growth. (See Robotic Workflow section in the SI for more information)
- _rxn_stirrateRPM
    - This column specifies the rate at which the reaction microplate was shaken during the two mixing time (mixingtime1S, mixingtime2S). (See Robotic Workflow section in the SI for more information)
- _rxn_temperatureC
    - This column specifies the temperature of the reaction solution at which the crystals growth occurred. (See Robotic Workflow section in the SI for more information)

raw → Experimental Descriptions

- _raw_v0-M_acid -
    - deprecated acid concentration values
- _raw_v0-M_organic
    - deprecated organic concentration values - organic identity indicated by _rxn_organic-inchikey
- _raw_v0-M_inorganic
    - deprecated PbI2 concentration values
- _raw_M_<inchikey>_final
    - deprecated: concentration of a given chemical in the experiment delineated by inchikey
- _raw_v1-M_<inchikey>_final
    - current version: concentration of a given chemical in the experiment delineated by inchikey
- _raw_mmol_<inchikey>_final
    - deprecated mmol value of the indicated inchikey in the experiment. Many values are 0. For more information on InChIKeys, please see: https://en.wikipedia.org/wiki/International_Chemical_Identifier
- _raw_v1-mmol_

    - current version of the mmol value of the indicated inchikey in the experiment.
- _raw_model_predicted
    - ML model predictions from ongoing live campaigns
- _raw_ChallengeProblem
    - CP participation value (not useful except for audit trail / record keeping)
- _raw_ExpVer
    - Workflow version of the experiment
- _raw_GenVer
    - Version of the ESCALATE_Capture code used to generate the experiment
- _raw_datecompleted
    - Date the final values of an experiment were recorded
- _raw_datecreated
    - Date that the initial experiment was created (when the run was staged for execution in the laboratory)
- _raw_jobserial
    - Run UID same as name without the (vial ID)
- _raw_lab
    - identity of the lab where the experiment was performed
- _raw_labwareID
    - Identity of the equipment that the experiment was performed on

raw → Reagent Descriptions

General: where ### can be any number 0-9 and #### can be any number 0-3). These values describe the exact preparation of each reagent. Each reagent can have up to 4 different chemicals. There are up to 5 unique reagents in WorkFLow 1 experiments

- _raw_reagent_<###>_v1-conc_<inchikey>
    - current version: of concentration calculation delineated by reagent ### and inchikey
- _raw_reagent_<###>_conc_<inchikey>
    - deprecated: concentration calculations delineated by reagent ### and inchikey
- _raw_reagent_<###>_volume
    - Volume dispensed of a reagent into the experiment
- _raw_reagent_<###>_chemicals_<#####>_InChIKey
    - the identity of reagent ### chemical #### defined by the inchikey
- _raw_reagent_<###>_chemicals_<#####>_amount
    - the amount of reagent ### chemical #### used in the defined reagent associated with a given experiment. These are defined by the inchikey above.
- _raw_reagent_<###>_chemicals_<#####>_units
    - Units which describe the amount above (each value is paired)
- _raw_reagent_0_chemicals_0_nominal_amount

    - the amount of reagent ### chemical #### targeted for used by the defined reagent associated with a given experiment. These are defined by the inchikey above.
- _raw_reagent_0_chemicals_0_nominal_amount_units
    - units for the above entry (each value is paired)
- _raw_reagent_0_date
    - preparation date of the reagent
- _raw_reagent_0_id
    - uid of the reagent
- _raw_reagent_0_instructions...
    - instructions for the preparation of reagents (see workflow for variable parameters)

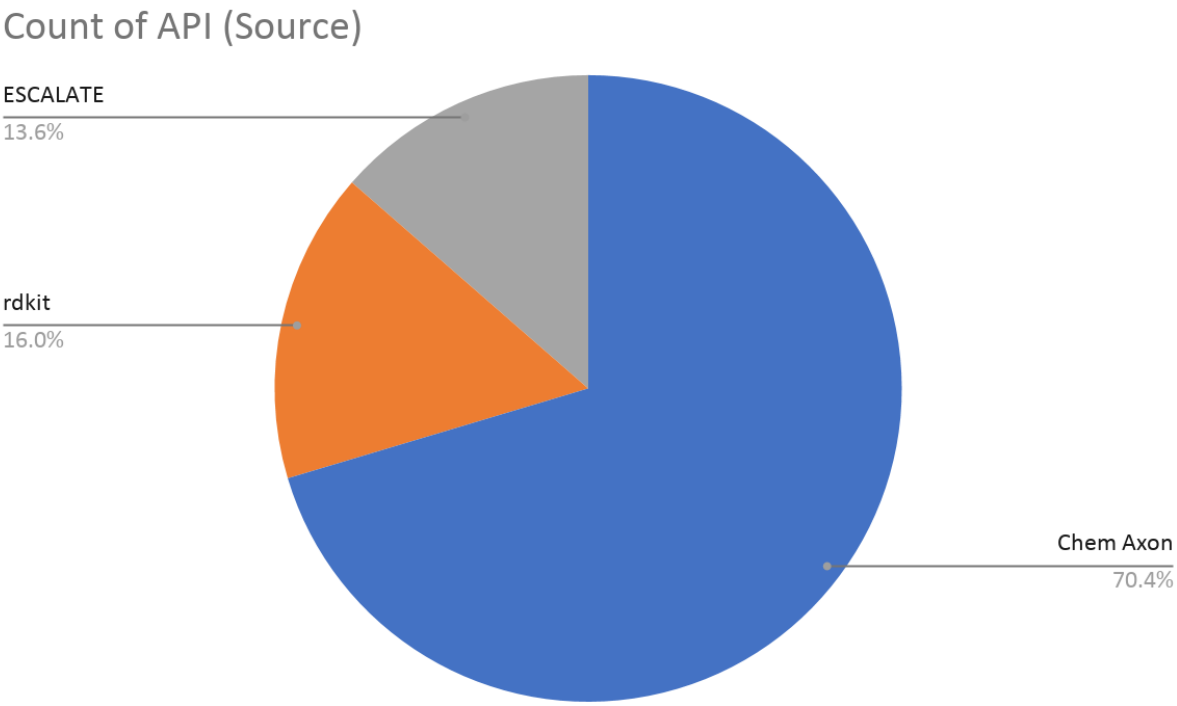
feats → Physicochemical Descriptors Overview

If information regarding a particular feature cannot be found in the list below, please refer to the linked documentation.

Overview

Expert features have been curated by domain experts as a possible alternative to existing physicochemical descriptors from rdkit, chemaxon, openbabel, etc. These features are hand-curated and could possibly incorporate unintentional errors. For now they are included under the _prototype_ namespace.

This table last updated = Thursday, November 21, 2019



ChemAxon

The complete description of the ChemAxon functions can be found here: https://docs.chemaxon.com/display/docs/cxcalc+calculator+functions

Feature UID	Description	API (Source)	Source Version	Source Function
_feat_acceptorcount	Hydrogen bond acceptor atom count in molecule	Chem Axon: cxcalc	19.24.0	acceptorcount
_feat_Accsitecount	Hydrogen bond acceptor multiplicity in	Chem Axon:	19.24.0	acceptorsitecount

	molecule (more details on web)	cxcalc		
_feat_Aliphatic AtomCount	Counts the number of aliphatic atoms in the molecule	Chem Axon: cxcalc	19.24.0	aliphaticatomcount
_feat_AliphaticRingCount	Aliphatic ring count	Chem Axon: cxcalc	19.24.0	aliphaticringcount
_feat_AromaticAtomCount	Aromatic atom count	Chem Axon: cxcalc	19.24.0	aromaticatomcount
_feat_AromaticRingCount	Aromatic ring count	Chem Axon: cxcalc	19.24.0	aromaticringcount
_feat_ASA	solvent accessible surface area calculated using the radius of the solvent (1.4 Å for the water molecule)	Chem Axon: cxcalc	19.24.0	wateraccessiblesurfacearea
_feat_ASA_H	solvent accessible surface area of all hydrophobic (|qi|<0.125) atoms (|qi| is the absolute value of the partial charge of the atom)	Chem Axon: cxcalc	19.24.0	wateraccessiblesurfacearea
_feat_ASA_P	solvent accessible surface area of all polar (|qi|>0.125) atoms (|qi| is the absolute value of the partial charge of the atom)	Chem Axon: cxcalc	19.24.0	wateraccessiblesurfacearea
_feat_ASA-	solvent accessible surface area of all atoms with negative partial charge (strictly less than 0)	Chem Axon: cxcalc	19.24.0	wateraccessiblesurfacearea
_feat_ASA+	solvent accessible surface area of all atoms with positive partial charge (strictly greater than 0)	Chem Axon: cxcalc	19.24.0	wateraccessiblesurfaceareag
_feat_AtomCount_C	Number of Carbon atoms in the molecule	Chem Axon: cxcalc	19.24.0	atomcount -z 6
_feat_AtomCount_N	Number of Nitrogen atoms in teh molecule	Chem Axon: cxcalc	19.24.0	atomcount -z 7
_feat_AvgPol	Average molecular polarizability calculation	Chem Axon: cxcalc	19.24.0	avgpol
_feat_BalabanIndex	The Balaban index	Chem Axon: cxcalc	19.24.0	balabanindex
_feat_BondCount	Bond count	Chem Axon: cxcalc	19.24.0	bondcount
_feat_CarboaliphaticRingCount	Carboaliphatic ring count	Chem Axon: cxcalc	19.24.0	carboaliphaticringcount
_feat_CarboaromaticRingCount	Carboaromatic ring count	Chem Axon: cxcalc	19.24.0	carboaromaticringcount
_feat_CarboRingCount	Number of rings containing only carbon atoms	Chem Axon: cxcalc	19.24.0	carboringcount
_feat_ChainAtomCount	Number of atoms in aliphatic chains	Chem Axon: cxcalc	19.24.0	chainatomcount
_feat_ChiralCenterCount	The number of tetrahedral stereogenic center atoms	Chem Axon: cxcalc	19.24.0	chiralcentercount
_feat_CyclomaticNumber	The cyclomatic number (complexity of molecule metric)	Chem Axon: cxcalc	19.24.0	cyclomaticnumber

_feat_donorcount	Hydrogen bond donor atom count in molecule	Chem Axon: cxcalc	19.24.0	donorcount
_feat_donsitecount	Hydrogen bond donor multiplicity in molecule (more details on website)	Chem Axon: cxcalc	19.24.0	donorsitecount
_feat_Hacceptorcount	Hydrogen bond acceptor multiplicity in molecule (at pH 3.0)	Chem Axon: cxcalc	19.24.0	acceptorcount -H 3.0
_feat_Hdonorcount	Hydrogen bond donor atom count in molecule (at pH 3.0)	Chem Axon: cxcalc	19.24.0	donorcount -H 3.0
_feat_HeteroaliphaticRingCount	number of heteroaliphatic rings in molecule	Chem Axon: cxcalc	19.24.0	heteroaliphaticringcount
_feat_HeteroaromaticRing Count	number of heteroaromatic rings in molecule	Chem Axon: cxcalc	19.24.0	heteroaromaticringcount
_feat_HyperWienerIndex	Hyper Wiener index	Chem Axon: cxcalc	19.24.0	hyperwienerindex
_feat_LargestRingSize	Number of atoms in largest ring	Chem Axon: cxcalc	19.24.0	largestringsize
_feat_LengthPerpendicularToTheMaxArea	Calculates the size of the molecule perpendicular to the maximal projection area surface	Chem Axon: cxcalc	19.24.0	maximalprojectionsize
_feat_LengthPerpendicularToTheMinArea	Calculates the size of the molecule perpendicular to the minimal projection area surface	Chem Axon: cxcalc	19.24.0	minimalprojectionsize
_feat_MaximalProjectionArea	Calculates the maximal projection area	Chem Axon: cxcalc	19.24.0	maximalprojectionarea
_feat_MaximalProjectionRadius	Calculates the maximal projection radius	Chem Axon: cxcalc	19.24.0	maximalprojectionradius
_feat_maximalprojectionsize	Calculates the size of the molecule perpendicular to the maximal projection area surface	Chem Axon: cxcalc	19.24.0	maximalprojectionsize
_feat_MinimalProjectionArea	Calculates the minimal projection area	Chem Axon: cxcalc	19.24.0	minimalprojectionarea
_feat_MinimalProjectionRadius	Calculates the minimal projection radius	Chem Axon: cxcalc	19.24.0	minimalprojectionradius
_feat_minimalprojectionsize	Calculates the size of the molecule perpendicular to the minimal projection area surface	Chem Axon: cxcalc	19.24.0	minimalprojectionsize
_feat_MolPol	Molecular polarizability calculation	Chem Axon: cxcalc	19.24.0	molpol
_feat_molsurfaceareaASAp	solvent accessible surface area of all atoms with positive partial charge (at pH 3.0)	Chem Axon: cxcalc	19.24.0	molecularsurfacearea -t ASA+ -H 3.0
_feat_molsurfaceareaVDWp	calculates the van der Waals surface of the molecule (at pH 3.0)	Chem Axon: cxcalc	19.24.0	molecularsurfacearea -t vanderwaals -H 3.0
_feat_msareaASAp	Molecular Surface Area calculation of atoms with positive partial charge (at pH 3.0)	Chem Axon: cxcalc	19.24.0	msa -t ASA+ -H 3.0

_feat_msareaVDWp	van der Waals surface calculation of atoms with positive partial charge (at pH 3.0)	Chem Axon: cxcalc	19.24.0	msa -t vanderwaals -H 3.0
_feat_PolarSurfaceArea	Topological Polar Surface Area calculation (2D)	Chem Axon: cxcalc	19.24.0	polarsurfacearea
_feat_ProtPolarSurfaceArea	Topological Polar Surface Area calculation (2D) (at pH 3.0)	Chem Axon: cxcalc	19.24.0	polarsurfacearea -H 3.0
_feat_Protpsa	Topological Polar Surface Area calculation (at pH 3.0)	Chem Axon: cxcalc	19.24.0	psa -H 3.0
_feat_Refractivity	Molecular refractivity calculation (derived from polarizability)	Chem Axon: cxcalc	19.24.0	refractivity
_feat_RingAtomCount	Number of atoms in molecular rings	Chem Axon: cxcalc	19.24.0	ringatomcount
_feat_RotatableBondCount	Number of rotatable atomic bonds in the molecule(s)	Chem Axon: cxcalc	19.24.0	rotatablebondcount
_feat_SmallestRingSize	Number of atoms in smallest ring	Chem Axon: cxcalc	19.24.0	smallestringsize
_feat_VanderWaalsSurfaceArea	Van der Waals Surface Area calculation	Chem Axon: cxcalc	19.24.0	vdwsa
_feat_VanderWaalsVolume	Calculates the van der Waals volume of the molecule	Chem Axon: cxcalc	19.24.0	volume
_feat_WienerIndex	Wiener index	Chem Axon: cxcalc	19.24.0	wienerindex
_feat_WienerPolarity	Wiener polarity	Chem Axon: cxcalc	19.24.0	wienerpolarity
_raw_molweight	molecular weight based on given SMILES representation	Chem Axon: cxcalc	19.24.0	mass
_raw_standard_molweight	molecular weight of the standardized smiles string	Chem Axon: cxcalc	19.24.0	mass

RDkit and ESCALATE

The complete description of the rdkit functions can be found here:
http://www.rdkit.org/Python_Docs/rdkit.Chem.Fragments-module.html

For ESCALATE API related, link a git merge request with the API. Source version should be the version of ESCALATE which the feature was FIRST included.

Feature UID	Description	API (Source)	Source Version	Source Function
_prototype_ECFP4_hexstring	circular topological fingerprints designed for molecular characterization (advanced - please reach out for help)	Chem Axon	19.24.0	eneratemd c input.smiles -k ECFP -c ecfp_config.xml -2
_raw_smiles_standard	curates a standardized smiles string from input smiles	Chem Axon	19.24.0	standardize
_raw_inchikey	the inchikey of the organoammonium species in the reaction (for ESCALATE_report <0.8.1)	ESCALATE	0.8.1	ExpertCurated
_raw_smiles	the smiles string of a given species	ESCALATE	0.8.1	User Input
_feat_fr_amidine	Number of amidine groups	rdkit.Chem.Fragments	2019.03.4	fr_NH2
_feat_fr_Ar_NH	Number of aromatic amines	rdkit.Chem.Fragments	2019.03.4	fr_NH1
_feat_fr_ArN	Number of N functional groups attached to aromatics	rdkit.Chem.Fragments	2019.03.4	fr_NH0
_feat_fr_dihydropyridine	Number of dihydropyridines	rdkit.Chem.Fragments	2019.03.4	fr_quatN
_feat_fr_guanido	Number of guanidine groups	rdkit.Chem.Fragments	2019.03.4	fr_ArN
_feat_fr_Imine	Number of Imines	rdkit.Chem.Fragments	2019.03.4	fr_Ar_NH
_feat_fr_NH0	Number of Tertiary amines	rdkit.Chem.Fragments	2019.03.4	fr_Imine
_feat_fr_NH1	Number of Secondary amines	rdkit.Chem.Fragments	2019.03.4	fr_amidine
_feat_fr_NH2	Number of Primary amines	rdkit.Chem.Fragments	2019.03.4	fr_dihydropyridine
_feat_fr_piperdine	Number of piperdine rings	rdkit.Chem.Fragments	2019.03.4	fr_guanido
_feat_fr_piperzine	Number of piperzine rings	rdkit.Chem.Fragments	2019.03.4	fr_piperdine
_feat_fr_pyridine	Number of pyridine rings	rdkit.Chem.Fragments	2019.03.4	fr_piperzine
_feat_fr_quatN	Number of quarternary nitrogens	rdkit.Chem.Fragments	2019.03.4	fr_pyridine