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 Header Prefixes

#### **\_raw\_**

**DO NOT LEARN ON. We strongly recommend not using these** unless you know what the header means and how the column of data relates to experimental outcomes. 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 should 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.

#### **\_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.

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)

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 \*good\* 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 \*bad\* 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 (PbI2)
* **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.

## \_rxn\_ → Primary Experimental Descriptions

* \_rxn\_M\_inorganic
  + molarity of the inorganic component added to the experiment well.
* \_rxn\_M\_acid
  + molarity of the inorganic component added to the experiment well.
* \_rxn\_M\_organic
  + molarity of the organic component added to the reaction well. Descriptions of the variance in the organic components are found in the \_feat\_ section of the dataset. Additional descriptions are under development. The identity of the organic cation is described in \_rxn\_organic-inchikey
* \_rxn\_mixingtime1S
  + Duration of the first mixing time after the addition of solvent, organic, inorganic and the first addition of formic acid. (See workflow section above for more information)
* \_rxn\_mixingtime2S
  + Duration of mixing time (in seconds) after the addition of the second portion of formic acid. See workflow section above for more information)
* \_rxn\_reactiontimeS
  + Duration of time the reaction was allow to sit and for crystals to form after the initial chemical mixing.
* \_rxn\_stirrateRPM
  + The rate at which the tray was shaken during the mixing time. **Note:** This number is not consistent for all runs in the dataset, but is potentially hazardous to learn on. The runs performed using different shake rates also have other characteristics which distinguish them. Care should be used when drawing conclusions from models which show this as an important feature.
* \_rxn\_temperatureC
  + The temperature of the reaction in C.

## \_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

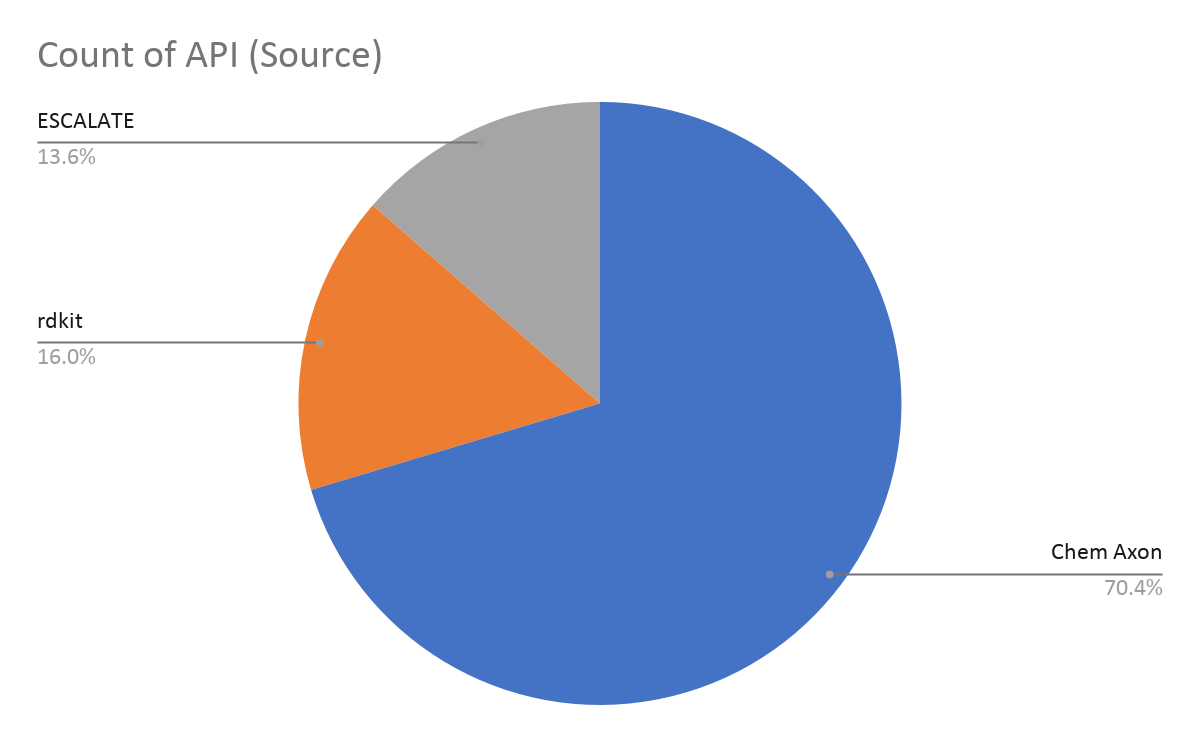
* \_raw\_reagent\_<###>\_v1-conc\_<inchikey>
  + current version: of concentration calculation delineated by reagent ### and inchikey
* \_raw\_reagent\_<###>\_conc\_<inchikey>
  + deprecated: concentration calculations delineated by reagent ### and inchikey
* \_raw\_reagent\_<###>\_volume
  + Volume 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 molecule (more details on web) | Chem Axon: cxcalc | 19.24.0 | acceptorsitecount |
| \_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)](https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP#ExtendedConnectivityFingerprintECFP-Introduction) | 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 |