Supplementary Information: Quantifying the performance of machine learning models in materials discovery

Christopher K. H. Borg[1], Eric S. Muckley[1], Clara Nyby[1], James E. Saal[1], Logan Ward[2], Apurva Mehta[3], and Bryce Meredig[1]

[1]Citrine Informatics, Redwood City, CA, USA
[2]Argonne National Laboratory, Lemont, IL, USA
[3]SLAC National Accelerator Laboratory, Menlo Park, CA, USA
[*]corresponding author(s): James E. Saal (jsaal@citrine.io)

October 24, 2022

1 Starrydata ingestion pipeline

As illustrated in Figure 1, the Starrydata ingestion pipeline consists of four processes. First, data from the Starrydata2 database is queried via the API. Specifically, a generalized query for any sample containing an element from the periodic table was constructed, and queried samples were cached in a local directory (note, samples not containing any elements were not returned). Queried samples were then converted to individual Pandas dataframes (raw, sample, paper, property) as shown in the Starrydata documentation [1]. Second, the raw data was combined with data from sample, paper, and property dataframes. This step generates one easily viewable table with all important linkages (process-composition-property). Third, since many properties of interest are recorded as a function of temperature, these properties are interpolated at defined temperatures (T = 300K). This allows for direct comparison of samples and their properties at defined temperatures. Fourth, thermoelectric figures of merit (e.g., σ_{E0} [2]) are calculated and subsequently filtered to identify erroneous data points. To denote a compound as being "111-type", we first ensure a composition has 3 or more elements and that all elements are either a metal or metalloid. Then, if the sum of the stoichiometry is between 2.9 and 3.1 and no single element has greater than 0.4 atomic fraction, the compound was denoted as being "111-type".

After querying for Starrydata records, it was noted that many records are sparse, only having a few properties reported for a given composition. This is likely due to the nature of the way experimental properties for TEs are reported and how they are extracted by the Starrydata team. When applicable, TE properties are calculated from extracted (interpolated) values using empirical relationships. For example, if Power Factor (PF) is not reported for a sample but Seebeck and electrical conductivity are reported for that sample, PF is calculated via: $PF = S^2 * \sigma$. These calculated values are then combined with values extracted (interpolated) from the Starrydata2 database. In instances where a sample has both an extracted and calculated value for a property, the extracted value was used. The number of properties extracted and calculated for samples at room-temperature are shown in Table 2. All property calculations are listed below:

1. $\sigma = \rho^{-1}$ (σ = electrical conductivity, ρ = electrical resistivity)
2. $PF = S^2 * \sigma$ (PF = Power Factor, S = Seebeck coefficient)
3. $ZT = PF * T * \kappa_{total}^{-1}$ (T = Temperature, κ_{total} = Total thermal conductivity)
4. $\sigma_{E0} \rightarrow e(2m_e\kappa_B T)^{3/2}/2\pi^2\hbar^3 * \mu_0(m^*/m_e)^{3/2}$. ($\sigma_{E0}$ = transport model prefactor, $\mu_0(m^*/m_e)^{3/2}$ = weighted mobility)

Table 2 highlights the number of records at distinct point in the data processing pipeline. Raw data refers to the number of data points for a particular property obtained from the initial query of the database. The initial query returns records with dense property curves (e.g. Seebeck as a function of temperature). To standardize the temperature at which properties are compared, property-temperature curves were interpolated at 300K; after filtering

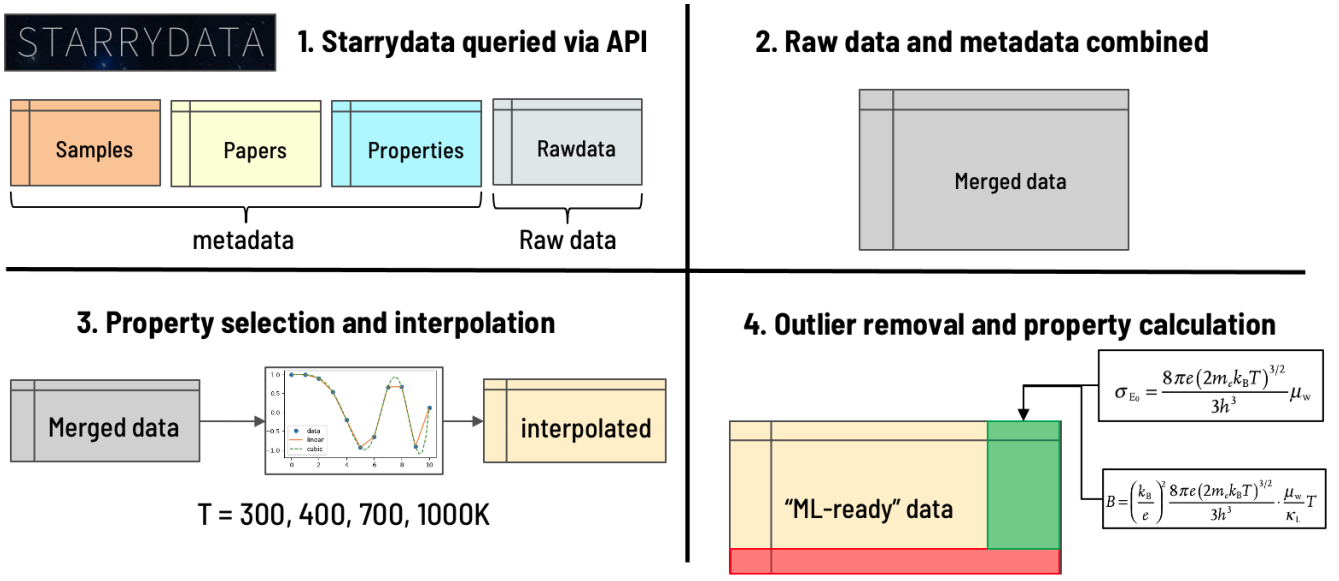

Figure 1: Starrydata ingestion pipeline. Raw data is queried via the starrydata api and merged into a single table. Then, Data points are interpolated at standard temperatures, thermoelectric figures of merit (e.g., σ_{E0}) are calculated, and physically-intuitive filters are applied to remove erroneous data points.

Table 1: Physically-relevant ranges for properties of interest. Property values outside of these ranges were filtered out by the SL pipeline.

Property	units	minimum value	maximum value
Seebeck coefficient (S)	V/K	-0.005	0.005
Electrical conductivity (σ)	S/m	0	10000000
Thermal conductivity (κ_{total})	$W/(mK)$	0	100
Power factor ($S^2\sigma$)	$W/(mK^2)$	0	10
Figure of merit (ZT)	-	0	3
Temperature (K)	K	200	1200
Transport coefficient (σ_{E0})	S/m	0	10000000

Table 2: Thermoelectric properties extracted from the Starrydata2 database. Raw data, interpolated, calculated, and final refer to distinct caches at points in the data ingestion pipeline 111-type refers to records that are labeled as 111-type by our composition classifier.

Property ID	Property	Units	raw data	interpolated	calculated	final	111-type
2	Seebeck coefficient	V/K	498527	21508	0	17315	986
3	Electrical conductivity	S/m	184924	9931	10366	16438	970
4	Thermal conductivity	$W/(mK)$	276597	15508	0	12785	789
5	Electrical resistivity	Ωm	324818	10399	0	8495	462
6	Power factor	$W/(mK^2)$	184900	10570	8529	15437	913
8	ZT	-	221091	13730	1713	12794	808
sigma_E_0	Transport coefficient (σ_{E0})	S/m	0	0	18181	14742	889

for curves with a physically-relevant temperature range (shown in table 1). This reduced the amount of data per record to the number of data points shown in the "interpolated" column. As previously described, the "calculated" column shows the number of data points for a particular property that were calculated from interpolated values using empirical relationships. The "final" dataset shows a filtered view of both calculated and interpolated values based on the physically-relevant properties filters shown in Table 1. The "111-type" column shows the same data points from "final" for only compounds that agree with our 111-type composition classifier (described in previous paragraph). The "111-type" dataset was used as an input to the SL pipeline resulting in the record counts shown in Section 2.4 (after indexing on chemical formula and filtering $ZT_{max} = 2$).

2 Acquisition function notation

When making predictions using random forests with uncertainty estimates, a prediction can be represented as a predicted distribution (rather than a single value) where the width of the distribution is correlated with the uncertainty of a prediction.

We note that previous work on sequential learning[3] used alternative definitions for acquisition functions:

1. Maximum Expected Improvement (MEI) = The candidate with the highest (or lowest, for minimization) target value.
2. Maximum Likelihood of Improvement (MLI) = The candidate with the highest (or lowest, for minimization) target value when including prediction uncertainty.
3. Maximum Uncertainty (MU) = The candidate with the greatest prediction uncertainty, entirely independent of its expected value.
4. Random search (RS) = A candidate is selected at random from the pool.

Herein, we use definitions that are more consistent with other literature:

1. Expected Value (EV) = The candidate with a predicted distribution mean value closest to the target value. EV strictly favors values closer to the target without considering uncertainty.
2. Expected Improvement (EI) = The candidate with a predicted distribution curve that has the highest probability of intersecting with the target value.
3. Maximum Uncertainty (MU) = The candidate with the greatest prediction uncertainty, entirely independent of its expected value.
4. Random search (RS) = A candidate is selected at random from the pool.

Note: Expected Value (EV) is consistent with the previously defined Maximum Expected Improvement (MEI) and Expected Improvement (EI) is consistent with the previously defined Maximum Likelihood of Improvement (MLI).

References

[1] Use datafiles in python. www.starrydata.wordpress.com/2018/09/14/use-datafiles-in-python/.

[2] Stephen Dongmin Kang and G Jeffrey Snyder. Charge-transport model for conducting polymers. *Nature materials*, 16(2):252–257, 2017.

[3] Julia Ling, Maxwell Hutchinson, Erin Antono, Sean Paradiso, and Bryce Meredig. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation*, 6(3):207–217, 2017.