Arc Science
Technical report

Arc Chemistry

Katerina S. Tatarinova, Danil D. Kotelnikov

Abstract. Rare-earth elements are separated by solvent extraction, a process that relies on differences of as little as 0.7 pm between the radii of neighbouring ions. Machine-learning models of extraction report high accuracy on random test splits. The manuscript reviews 113 sources on extraction chemistry, public data, model families, evaluation practice and uncertainty. Its central evidence is a published re-analysis of the largest public table of distribution ratios, 1,202 rows for 14 lanthanides. A random forest scores R2 = 0.94 on the original hold-out and 0.20, -0.03 and -0.24 when the test set shares no ligands, literature sources or fingerprint families with the training set. The manuscript lists nine gaps and proposes a study with six falsifiable hypotheses. The study is a proposal, and every number in it is a target set before data are seen.

In short

A review of 113 sources on machine learning for rare-earth solvent extraction, with a proposed study that tests models on ligands, laboratories and chemical families absent from training. The study is a design with six falsifiable hypotheses.

1. Data

The only machine-learning-ready extraction table found by the review is a collection of 1,202 literature values of log D for 14 trivalent lanthanides. D is the distribution ratio, the metal concentration in the organic phase divided by that in the aqueous phase. The original authors state 109 ligands, and a later analysis recovers 93 unique ligands from the released feature matrix; the manuscript keeps both counts. The matrix carries fingerprints and no SMILES strings, so the ligands cannot be encoded again with a new representation. All rows concern neutral ligands, so the acidic industrial extractants D2EHPA and PC-88A are absent.

Stability-constant sources are larger and harder to use. NIST SRD 46 holds 6,166 ligands and 112,559 data lines and is discontinued. Two recent collections report more than 30,000 log K1 values and more than 18,000 tributyl phosphate extraction points; the release status of both collections is open. The proposed study therefore starts with a row-level table. Each record carries the ligand SMILES, extractant, diluent, extractant concentration, aqueous composition and acid, ionic strength, temperature, metal, D and the DOI of origin. Rows extracted by a language-model pipeline are flagged, and a human reader audits a random sample.

2. Representations

The first lanthanide extraction model combined extended-connectivity fingerprints, 208 molecular descriptors, experimental conditions and lanthanide descriptors into 2,291 inputs. The review names three recurring problems with such inputs. A fingerprint encodes the ligand and leaves out the medium, so two rows with the same ligand and different D are indistinguishable unless the concentration, diluent and acid are also given. Metal identity enters through descriptors such as ionic radius, and adjacent lanthanides differ by 0.7 to 2.2 pm, so a descriptor that varies smoothly with atomic number may be too coarse for adjacent pairs. No study that tests this was found.

The third problem concerns the target. The separation factor is the ratio of two distribution ratios, so a value derived from two predictions accumulates the error of both, whereas a model of the log ratio can learn that shared errors cancel. In the one analysis that compares extraction strength with selectivity, the two rankings diverge, with a Spearman correlation of 0.63. The proposed study tests chemical descriptors (donor group, branching, chain length, lanthanide radius) against fingerprints alone, and it tests geometry-based features from a tight-binding method parameterised for lanthanides as additional descriptors.

3. Models

Published models span fully connected networks, classical regressors, Gaussian processes and message-passing graph networks. The first extraction model, a fully connected network, reached R2 = 0.85 and a mean absolute error of 0.34 log units on a random hold-out of 117 rows. One stability-constant study reports root-mean-square errors of 0.63 log units for a multi-metal model and about 0.76 for lanthanide-only models, and one graph network trained on more than 30,000 values reports R2 = 0.942 on an external test set. Under ligand-grouped splits an ensemble reaches R2 of 0.676 to 0.752. These are the values that the cited authors report; the review did not reproduce them.

The proposed study uses gradient-boosted trees and random forests as baselines, on fingerprints, descriptors, conditions and lanthanide descriptors. A graph network is added for comparison. Pretraining on tmQM, a set of 86,665 transition-metal complexes, is tested as transfer learning, with the known limit that the set contains no lanthanides. The attached hypothesis is that pretraining improves accuracy when fewer than 200 training rows are available. One constraint frames all of this. A Pr/Nd separation factor of 1.5 corresponds to a free-energy difference of 1.0 kJ/mol at 298 K, and a model error of 0.3 to 0.5 log units on D exceeds the quantity of interest for adjacent pairs.

4. Splits

A random split places rows of the same ligand, and often of the same paper, on both sides of the split. The model can then score well by remembering ligands that it has already seen. The review names three natural grouping variables for lanthanide extraction: the ligand, the literature source,

which carries the experimental protocol, and the fingerprint family, a cluster of structurally similar ligands. A disjoint split keeps every member of a group on one side.

A published re-analysis of the 1,202-row table measures the effect. A random forest reaches R2 = 0.94 on the original hold-out of 117 rows. With ligand-disjoint, source-disjoint and family-disjoint test sets the same model reaches 0.20, -0.03 and -0.24. An R2 at or below zero means that the model predicts no better than the mean of the test values. On four new ligands tested prospectively the R2 is 0.70. The manuscript attaches cautions: the re-analysis has one author, discloses generative-AI assistance, has no independent replication, and its source-disjoint result may mix extrapolation with scatter between laboratories. In the proposed study every model is evaluated under all four splits and a time split, with the random split as a reference only.

5. Evaluation design

In the proposed study accuracy is reported per split type, and each prediction carries an interval. Conformal prediction wraps any regressor in intervals whose coverage is guaranteed when calibration and test data are exchangeable. In the published re-analysis, nominal 80% intervals covered 0.72 of the test rows on the ligand-disjoint split before calibration, 0.90 after split-conformal calibration and 0.87 with Mondrian calibration. Exchangeability does not hold across disjoint splits by construction, so the manuscript treats these coverages as observations on one data set. No calibrated-uncertainty study of lanthanide stability constants was found.

The proposed study fits split-conformal and Mondrian intervals for each split type and reports coverage on source-disjoint data. The hypothesis is an empirical coverage between 0.85 and 0.95 at a nominal level fixed in advance. Acquisition rules based on model uncertainty are compared with random selection on retrospective data and, if a partner laboratory has capacity, on new extraction runs. The published benchmark found no acquisition function that won on both discovery and prediction, which is the stated reason to test each rule against random selection.

6. Study design

The design answers nine gaps that the searches behind the review did not close. Among them are an open row-level table of distribution ratios with structures for industrial extractants, a peer-reviewed model of separation factors for adjacent pairs such as Nd/Pr, independent replication of the log D benchmark, calibrated uncertainty for stability constants, and a benchmark of universal machine-learned interatomic potentials on lanthanide complexes. The review states each gap relative to the retrieval tools it used.

Six hypotheses follow, each with an outcome that would refute it. The first states that adding extraction conditions lowers error on ligand-disjoint test sets. The second states that chemical descriptors raise R2 on family-disjoint splits. The third states that conformal intervals reach the stated coverage on source-disjoint data. The fourth states that uncertainty-driven acquisition reaches a fixed error threshold with fewer new experiments than random selection. The fifth states that pretraining helps in the low-data regime. The sixth states that the top candidates ranked by

the model contain more high-selectivity ligands than candidates chosen by a panel of experts from the same pool. Every number in the design is a target set before data are seen, and a negative result is reported with the same prominence as a positive one.

7. Limits

The manuscript is a review and a proposal, and the proposed study is the work that follows. Its numbers are published values or targets set before data. It claims no separation factor for any ligand, and a claim about Nd/Pr would require experiments outside its scope. The central numbers come from a single unreplicated re-analysis whose prospective test has four ligands, and the manuscript reads them as a strong warning about the random split.

The review reads its sources at three depths: two of the 113 in full, five in part and the rest at the level of the abstract or a one-sentence summary. Claims from abstract-level reading are phrased as the authors' reported results. A statement that something was not found means that it was not found with the retrieval tools used. The manuscript declares that an AI assistant was used to retrieve and verify references, extract data, draft text and prepare figures, and that one author is preparing funding applications built on the design.

System design

The review and the public table exist; every later stage is the proposed study.

Implemented components

- Public log D table.
- Literature review.

Specified components

- Row-level table.
- Sampled audit.
- Features.
- Disjoint splits.
- tmQM pretraining.
- Tree baselines.
- Graph network.
- Conformal intervals.
- Acquisition test.
- Hypothesis tests.

Connections

- Public log D table → Literature review.
- Literature review → Row-level table.
- Row-level table → Sampled audit.
- Row-level table → Features.
- Row-level table → Disjoint splits.
- Features → Tree baselines.
- Features → Graph network.
- tmQM pretraining → Graph network.
- Disjoint splits → Tree baselines.
- Disjoint splits → Graph network.
- Tree baselines → Conformal intervals.
- Graph network → Conformal intervals.
- Conformal intervals → Acquisition test.
- Conformal intervals → Hypothesis tests.
- Tree baselines → Hypothesis tests.
- Acquisition test → Hypothesis tests.

Source

This report is the printed form of the page https://arcscience.app/arc-chemistry/. The page is the source of record for every statement in it.