Better Data Splits for Machine Learning with `astartes`

Jackson W. Burns and William H. Green (advisor)

Massachusetts Institute of Technology

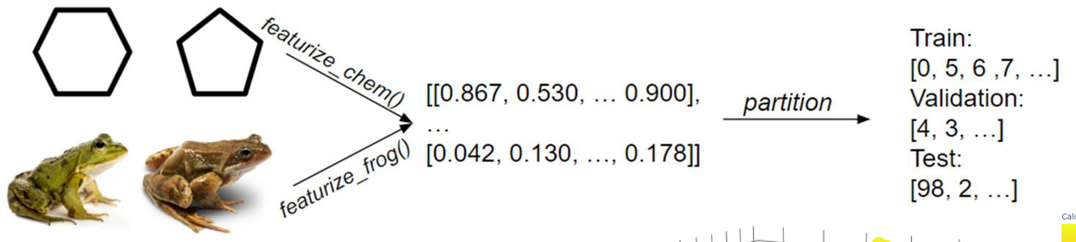
Motivation

- Implementing Machine Learning is highly accessible
- **Data leakage** and **extrapolation** often not addressed
- Rigorous model **validation** lacks a unified tool
- `astartes` makes validation as easy as modeling

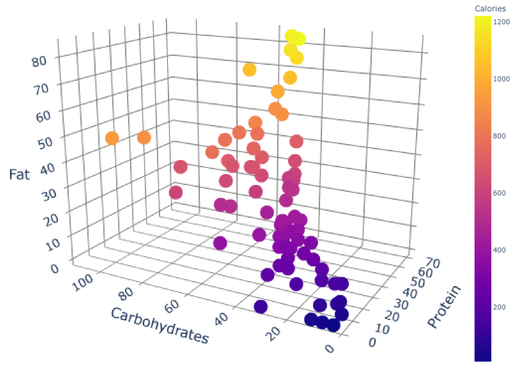
`astartes` algorithmic samplers: test, validation, train, partitions, molecules, arrays

Background

- ML requires domain knowledge to properly featurize data
- `astartes` provides featurizations and operates on arbitrary arrays

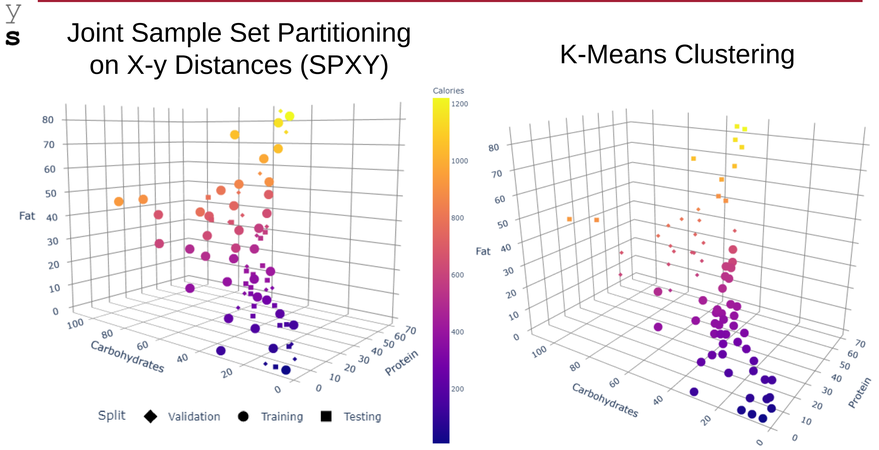


- Trained models are used in two ways:
- **Interpolation** – inference
  - Random splits commonly used
  - Kennard-Stone and SPXY enforce similar composition
- **Extrapolation** – discovery
  - Difficult but often more realistic
  - Clustering-based approaches



Representative Problem – Burger King Menu in three-dimensional space

Interpolation & Extrapolation Comparison



Representative Problem – partitioning by different sampling algorithms

Scan to interact with these figures and see how they were generated with `astartes`

- Interpolation exposes the model to the most representative subset during training
  - Random sampling *likely* to enforce
  - SPXY *guaranteed* to enforce
- Extrapolation withholds feature space until the validation and testing phase
  - K-Means, Sphere Exclusion
- Partition into training:validation:testing for critical evaluation

Discovery with ML requires **rigorous evaluation** of the model's capacity to **extrapolate**

Performance Impact at HPC Scale

- Common feature space is high-dimensional and non-interpretable
- **Compare** different data splits by **tracking impact on** D-MPNN **performance**
  - Directed-Message Passing Neural Network train with pytorch via chemprop
  - Yang et al. *in* J. Chem. Inf. Model (10.1021/acs.jcim.9b00237)
- Case study with two established cheminformatics prediction tasks (10^5 samples with 10^3 features)

QM9 Multi-Objective

Split	MAE	RMSE
Random	2.02 ± 0.06	3.63 ± 0.21
Scaffold	2.20 ± 0.27	3.46 ± 0.49
K-means	2.48 ± 0.33	4.47 ± 0.81

Ramakrishnan et. al *in* Scientific Data *(10.1038/sdata.2014.22)*

RDB7 Barrier Prediction

Split	MAE	RMSE
Random	3.87 ± 0.05	6.81 ± 0.28
Scaffold	6.28 ± 0.43	9.49 ± 0.50
K-means	5.47 ± 1.14	8.77 ± 1.85

Spiekermann et. al *in* Scientific Data *(10.1038/s41597-022-01529-6)*

- Real world datasets see similar or worse performance for extrapolation
- Extrapolation yields larger variance due to different composition between train and test sets
- Better equipped to use these models or *improve* them

Reproducibility & Accessibility

Scientific Software is a Science

- `astartes` produces identical results for *every* user *every* time
- These figures and the paper are reproduced before every change to ensure backward compatibility

Run Tests passing | Reproduce Paper passing

GitHub Repository

- Maintainability and ease of use are a top priority
  - Lower barrier of entry for users
  - Easily added to existing workflows

```
pip install astartes
conda install astartes
```

Future Work

- More and better **featurization schemes** for chemicals and generic data
  - Automated learned encodings *directly* without user input
  - Community contributors from multiple fields
- Find and develop new interpolative and extrapolative algorithms

Contact