AIMSim and astartes Vlachos Group Software Workshop

Jackson Burns
Green Group @ MIT CCSE

***AIMSim*: An accessible cheminformatics platform for similarity operations on chemicals datasets**

10.1016/j.cpc.2022.108579 *and*
10.26434/chemrxiv-2022-nw6f5-v5

github.com/VlachosGroup/AIMSim

Machine Learning Validation via Rational Dataset Sampling with `astartes`

10.21105/joss.05996

github.com/JacksonBurns/astartes

AIMSim

***AIMSim*: An accessible cheminformatics platform for similarity operations on chemicals datasets**

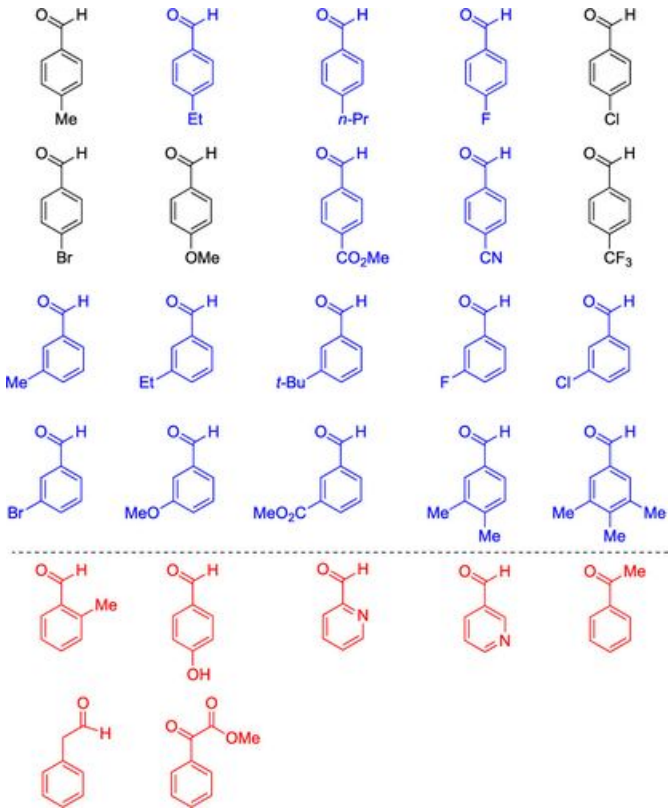
10.1016/j.cpc.2022.108579 *and*
10.26434/chemrxiv-2022-nw6f5-v5

github.com/VlachosGroup/AIMSim

- "Visualizing Diversity in your Molecular Dataset"
  - Virtual Screening
  - Lead Optimization
  - ML Dataset Preparation

- AIMSim provides a GUI and module-level set of tools for:
  - Similarity visualization
  - Clustering
  - Descriptor & Distance Selection
  - Database Comparisons
  - Robust set of molecular featurization tools

How diverse is my dataset?



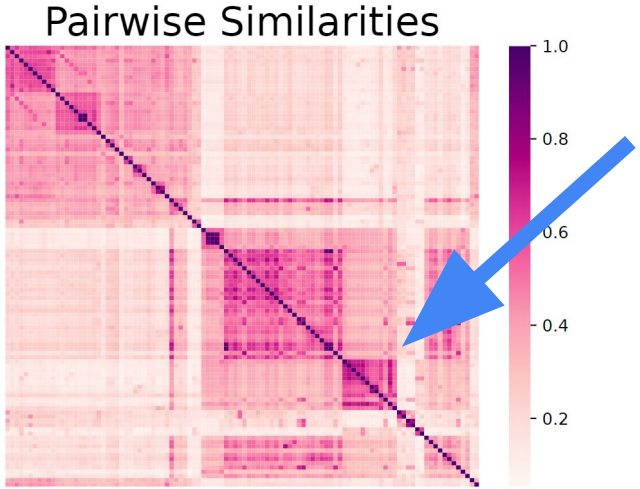
Kozlowski 10.1021/acs.orglett.2c03246

For chemists…

Validate that a method is generalizable

For ML researchers…

Ensure that a training set covers sufficient ‘chemical space’



Highly similar clusters of molecular might be redundant!

Generating this plot requires choosing…

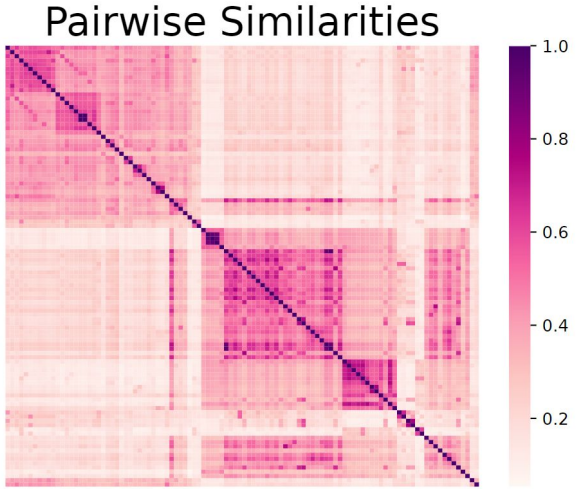
- Algorithm to decide on ‘clusters’
- Molecular Descriptor
- Distance Metric

The **AI** in AIMSim can pick this for you!

as well as the logic to handle…

- Database Comparisons
- Molecular featurization tools
- Much more!

AIMSim



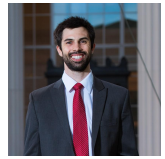
github.com/VlachosGroup/AIMSim

- `astartes` makes rigorous ML as simple as sklearn/pytorch
  - Non-random dataset splitting to enforce inter/extrapolation
  - Three-way splitting for validation

- Explicit focus on interoperability with existing code and dependencies

- Packaging and software quality are the *same importance* as scientific accuracy

algorithmic **s**amplers **t**est **a**lidation **r**ain **t**ition **e**cules **s**

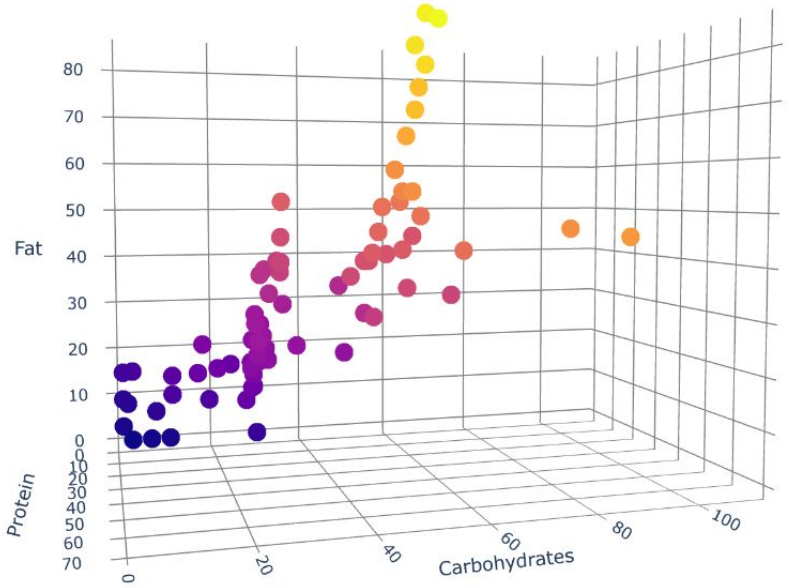


Machine Learning Validation via Rational Dataset Sampling with `astartes`

10.21105/joss.05996

github.com/JacksonBurns/astartes

My dataset covers all the feature space I want, now how do I train on it?

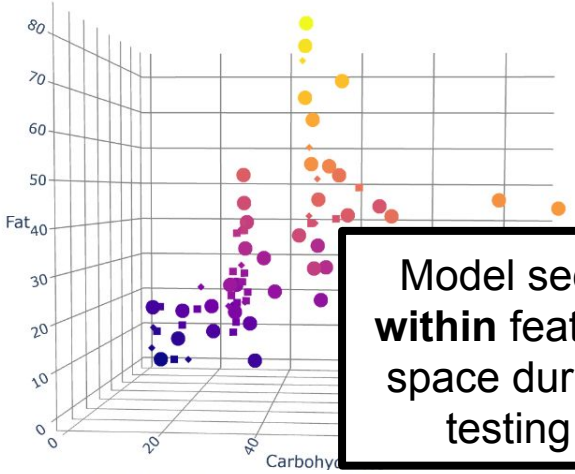


Burger King Menu in 3D Space

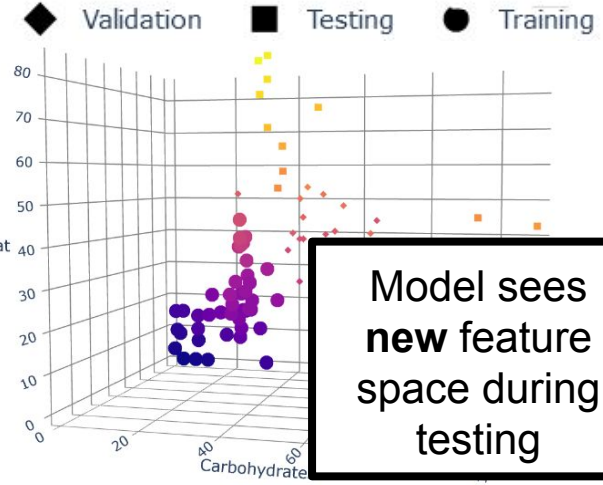
Inference
(Interpolation)

astartes

Discovery
(Extrapolation)



Model sees **within** feature space during testing



Model sees **new** feature space during testing

Demos! (and a usage note)

- Remainder of this presentation will be a mix of interactive demos hosted online
  - If you are experienced Python user, you can also follow our installation instructions and run these locally
- You can access everything on our GitHub pages
  - Jupyter notebooks (runnable in Google Colab) that demonstrate theory and application

- General usage note: AIMSim and astartes have very thorough documentation that should answer many questions, but we are also happy to add to it!

Additional Note: please star the GitHub repo to let us know you are using the software and file an issue or reach out (jwburns@mit.edu) for issues/collaborations!