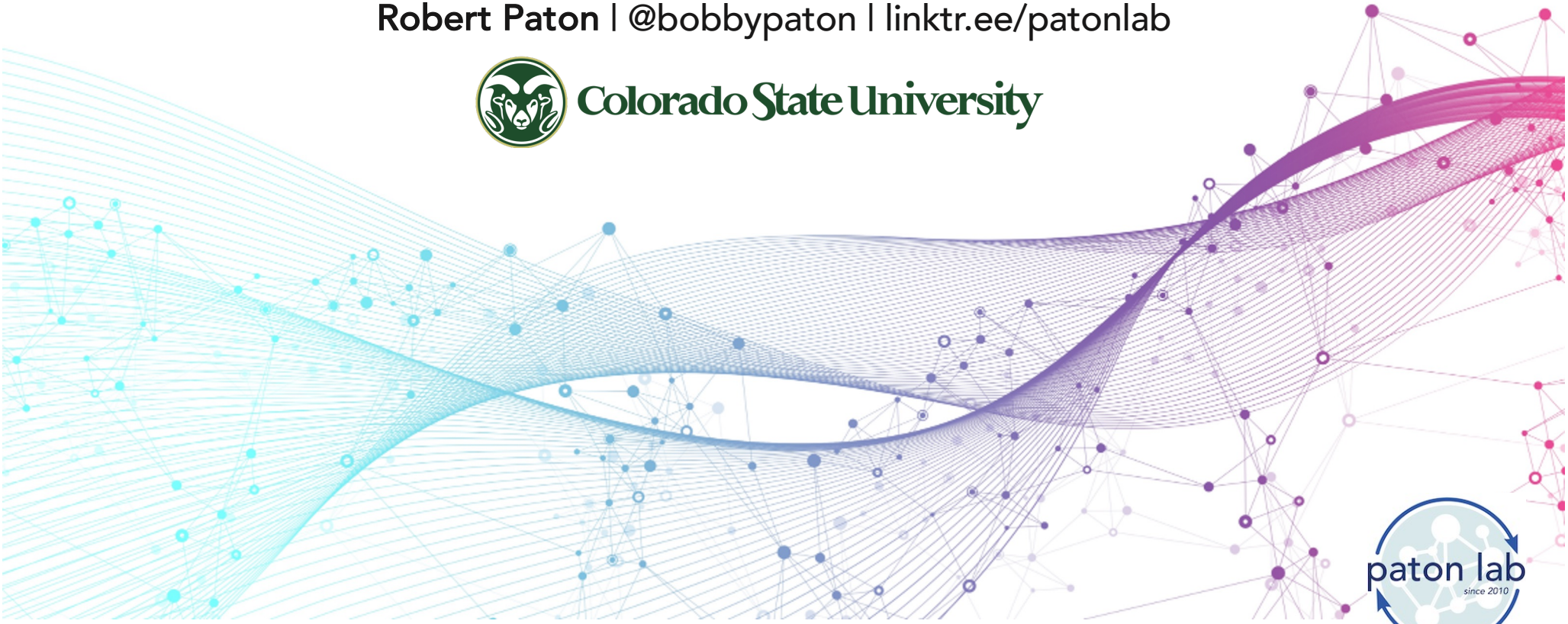
2D and 3D Representations and Molecular Descriptors

Robert Paton | @bobbypaton | linktr.ee/patonlab

Colorado State University



CAMLC 2025: Zaragoza Sept 16th 2025

2D and 3D Representations and Molecular Descriptors

Outline:

Introducing the terminology

Digitizing chemistry: How do we convert molecules into numbers?

Molecular Descriptors in Different Dimensions

Practical Application using Python

By the end of this lecture you should be able to appreciate some different strategies to *featurize* molecules, appraise what types of features are applicable to a given problem, and use molecular descriptors to perform a regression task for a target molecular property.

2D and 3D Representations and Molecular Descriptors

Useful Resources:

- Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors - WILEY-VCH, DOI: 10.1002/9783527613106*
- *Janet, J. P.; Kulik, H. J.* Machine Learning in Chemistry, 2020 DOI: 10.1021/acs.infocus.7e4001
- Gallegos, L. C. *et al.* Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021**, 54, 827–836. DOI: 10.1021/acs.accounts.0c00745
- *Wiliiams, W. L. et al.* The Evolution of Data-Driven Modeling in Organic Chemistry *ACS Cent. Sci.* **2021**, *7*, 1622–1637. DOI: 10.1021/acscentsci.1c00535
- Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comp. Mol. Sci.* **2022**, *12*, e1603. DOI: doi.org/10.1002/wcms.1603
- https://iwatobipen.wordpress.com/tag/rdkit/
- https://practicalcheminformatics.blogspot.com/
- https://greglandrum.github.io/rdkit-blog/

Useful Tools:

https://github.com/mordred-descriptor/mordred
https://github.com/digital-chemistry-laboratory/morfeus
https://github.com/patonlab/DBSTEP

House Price Predictions

Target variable | *Feature vector*

House price	Bedrooms	Sqft_living	Floors	Bathrooms
221,900	3	1180	1	1

Featurization involves the conversion of raw data into a numerical vector (***feature vector***)

We can then attempt to find a mathematical relationship between the target variable and the feature vector

In the case of Multivariate Linear Regression (MLR)

$$y = m_0x_0 + m_1x_1 + m_2x_2 + m_3x_3 + c$$

House Price Predictions

Target variable *Feature vector*

House price	Bedrooms	Sqft_living	Floors	Bathrooms	Num. rooms
221,900	3	1180	1	1	8

What attributes or properties are used to make up the feature vector?

Feature Engineering often involves the use of domain knowledge – i.e. which attributes are relevant?

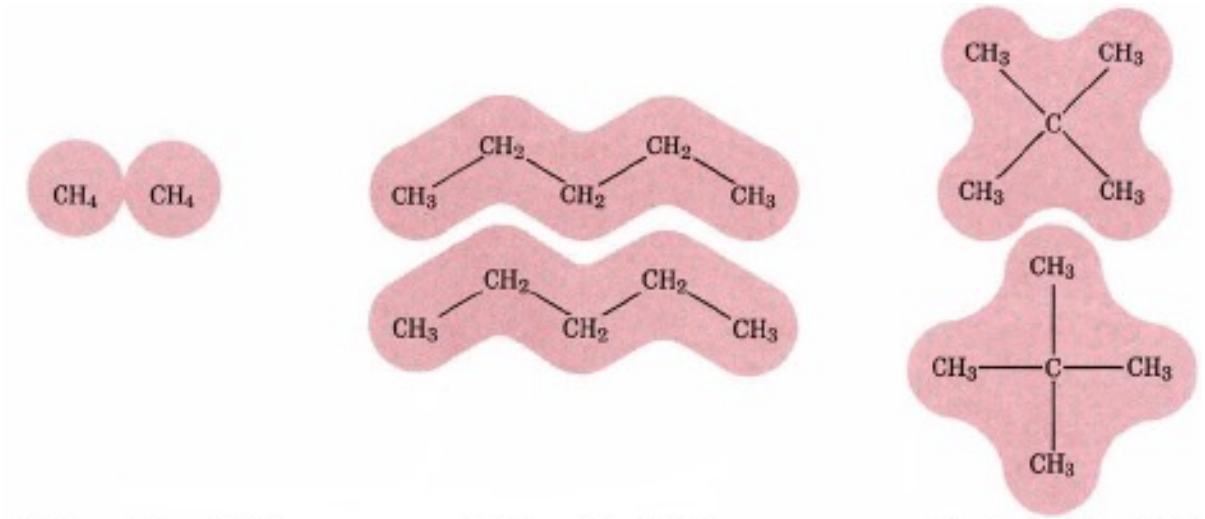
What is the best representation of the sample data to learn a solution to the problem?

Feature selection then automatically selects a subset that are most useful to the problem

Some may be dropped due to low statistical importance or correlation with other features.

$$y = m_0x_0 + m_1x_1 \quad + m_3x_3 + m_4x_4 + c$$

Back to Chemistry: Quantitative Structure-Property Relationships (QSPR)



BPs (°C): **-161** **36** **9.5**

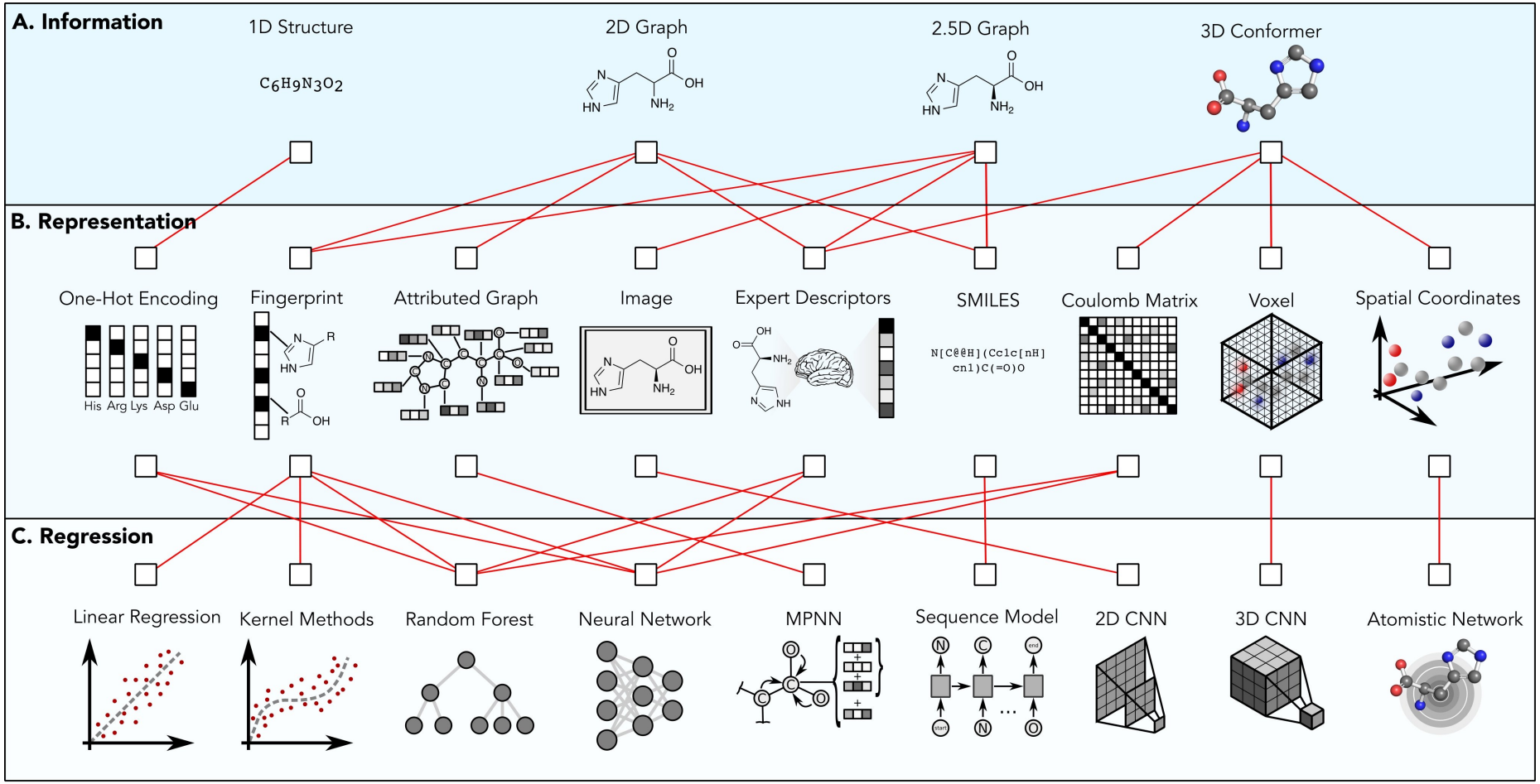
Back to Chemistry: Quantitative Structure-Property Relationships (QSPR)

Boiling Point	molecule	Mol. weight
36.1	pentane	72.15
68.8	hexane	86.18
98.5	heptane	100.20
125.6	octane	114.23

Molecular Descriptors numerically describe the shapes, electronics, energies of molecules

https://bertiewooster.github.io/2023/03/10/Revisiting-a-Classic-Cheminformatics-Paper-The-Wiener-Index.html

Representing Molecules



Pattanaik, L.; Coley, C. W. *Chem* **2020**, 6, 1204-1207

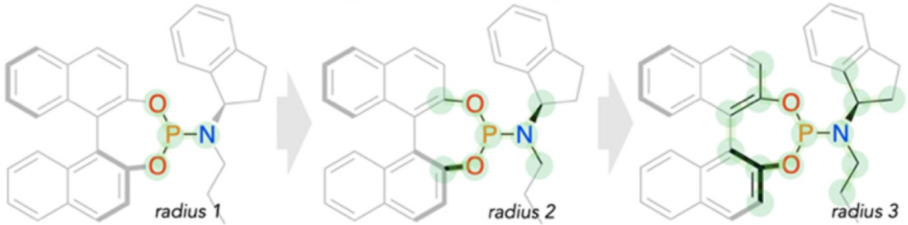
1D representations

Constitutional descriptors

- *Num. C atoms = 32 Num. Rings = 7*

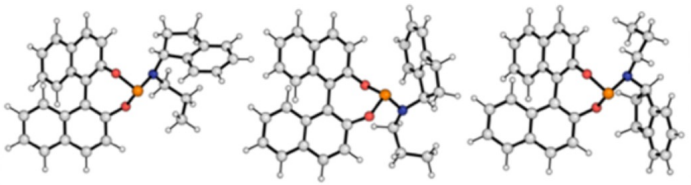
2D representations

Atomic connectivity: topological descriptors & molecular graphs

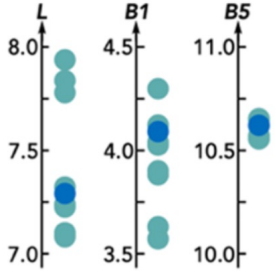


- *Morgan/circular fingerprints encode molecular fragments*

4D representations



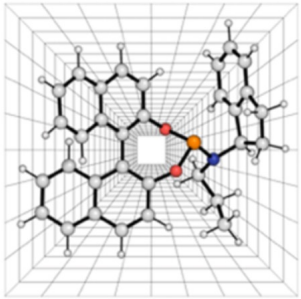
Conformational ensemble descriptors



Boltzmann average values and Min/Max values

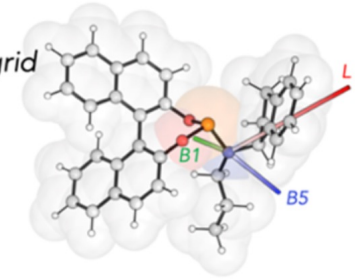
- wSterimol

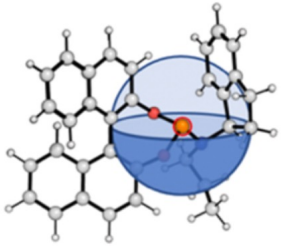
3D representations



Grid-based descriptors map properties to 3D-grid points (orientation dependent)

- *CoMFA*
- *Average Steric Occupancy (ASO)*





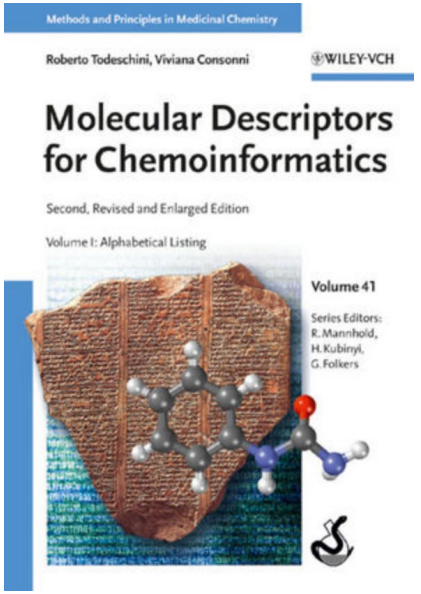
Optimized 3D-geometry described by various shape/steric descriptors:

- *Sterimol (multi-dimensional)*
- *Buried volume (scalar)*

Acc. Chem. Res. **2021**, *54*, 827−836

Molecular Descriptors

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number, or the result of some standardized experiment."



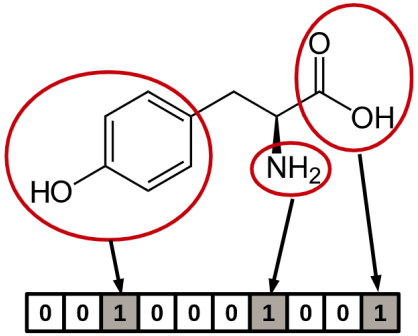
Consonni and Todeschini, Wiley 2009

Considerations for Molecular Descriptors

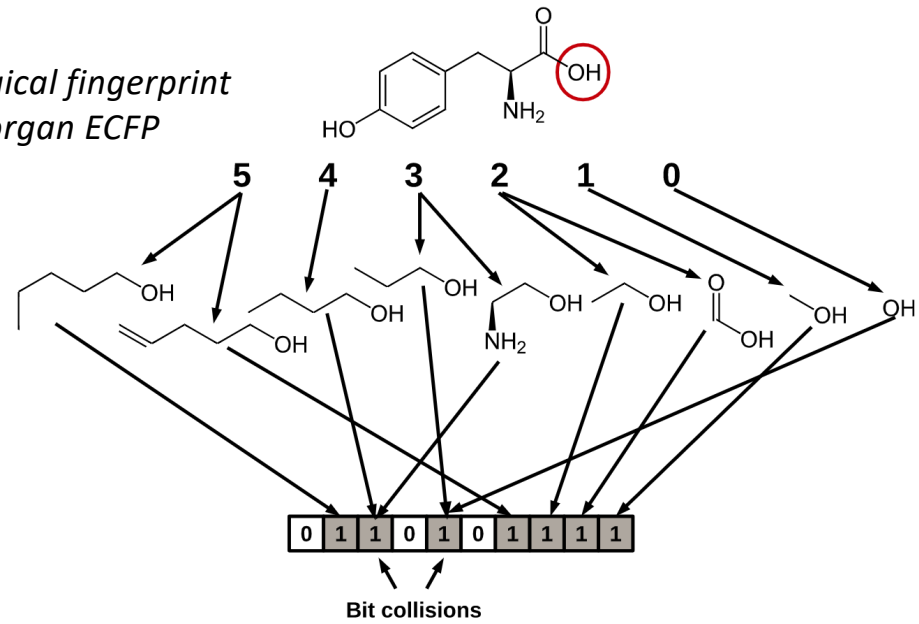
- Interpretability
- Good correlation with at least one property
- No trivial correlation with other molecular descriptors
- Gradual change in its values with gradual changes in the molecular structure
- Insensitive to how collected
- No restrictions to type of molecule
- if possible, discrimination between structural and stereo-isomers.
- Independent of other molecular descriptors

2D Molecular Descriptors: Fingerprints

substructure fingerprint
e.g. MACCS



topological fingerprint
e.g. Morgan ECFP



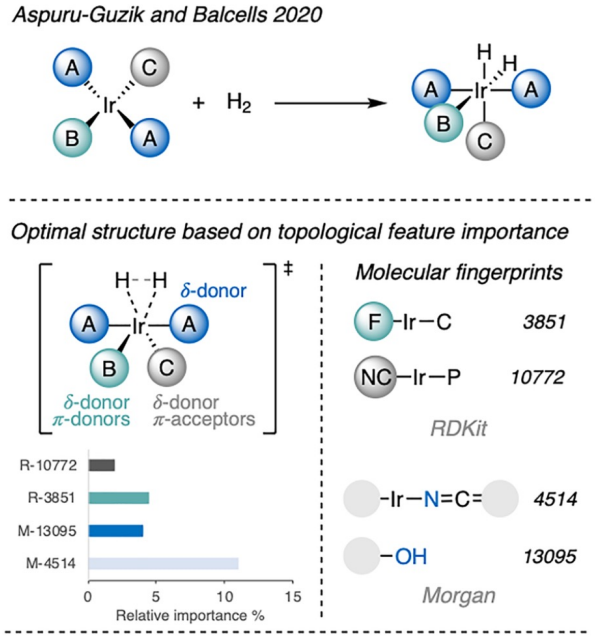
Feature vector

Hypothetical 10-bit substructure fingerprint, with three bits set because the substructures they represent are present in the molecule (circled)

Hypothetical 10-bit topological fingerprint, in this case a linear path-based fingerprint with fragments up to a length of 5.

There are two bit collisions, which are likely in fingerprints with a reduced number of bits.

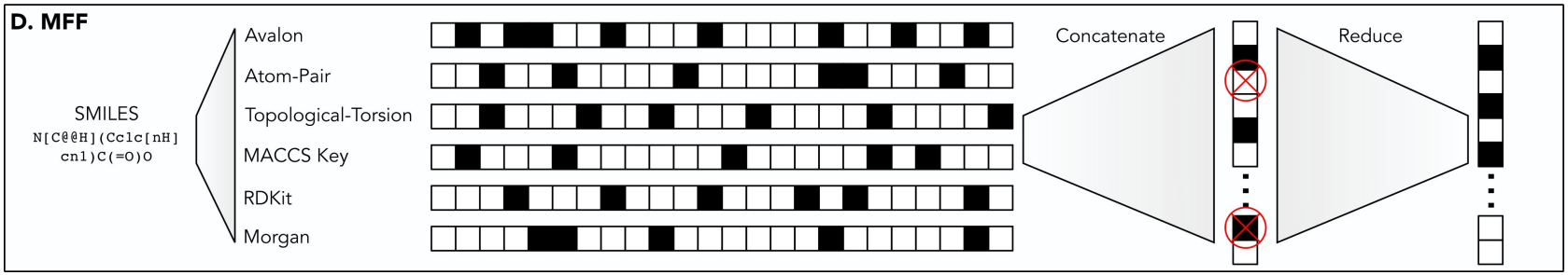
Methods **2015**, *71,* 58–63



2D Molecular Descriptors: Fingerprints

2500 barrier heights
Regression with a Gaussian process (GP) model gave an MAE = 0.6 kcal/mol, R^2= 0.95

Gomes & Aspuru-Guzik: Chem. Sci. **2020**, *11*, 4584−4601



Concatenation of multiple fingerprints used to predict selectivities and yields

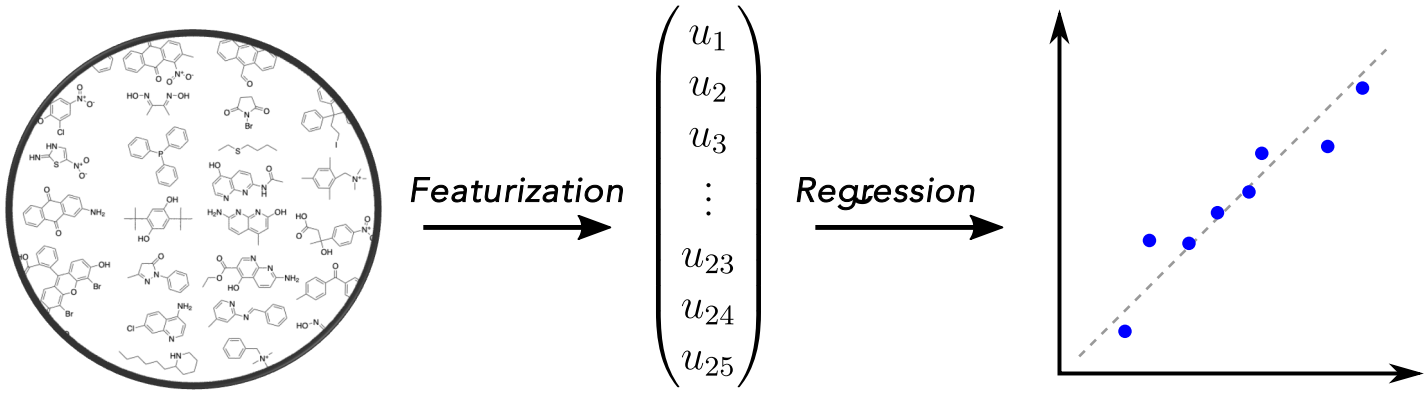
Glorius: Chem. **2020**, *6*, 1379−1390

2D Molecular Descriptors: Fingerprints

Example 1: Fingerpints https://tinyurl.com/yc8bma2b

Example 2: Topological Descriptors https://tinyurl.com/5xa7ekxm

Traditional Cheminformatics Approach

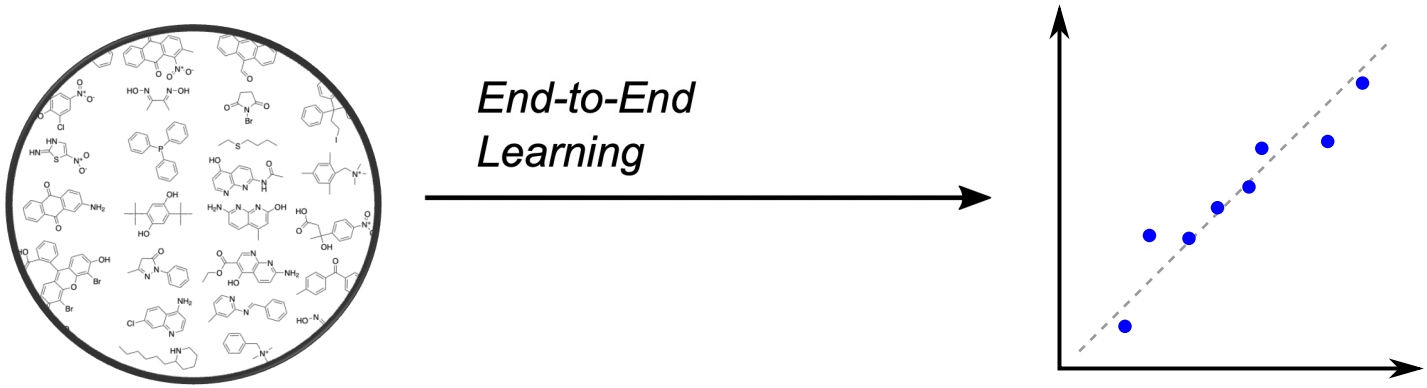


Traditional cheminformatics / machine learning

- Determine a scheme to map atoms/bonds/molecules to a continuous vector space
- Fit a function that maps the vector space to desired property

Gallegos, Luchini, St. John, Kim & Paton *Acc. Chem. Res.* **2021,** *54*, 827; Sigman, Doyle, Anslyn, *ACS Cent. Sci.* **2021**, 7, 1622

Deep Learning Approach



Traditional cheminformatics / machine learning

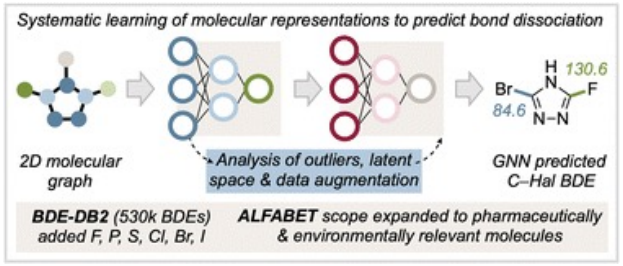
- Determine a scheme to map atoms/bonds/molecules to a continuous vector space
- Fit a function that maps the vector space to desired property

- Instead of manually determining a method to convert molecules into a finite-dimensional numerical representation, can we "learn" an optimal embedding?

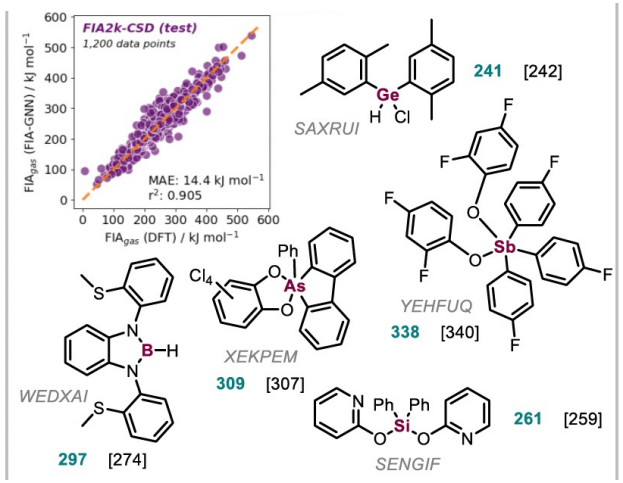
Gallegos, Luchini, St. John, Kim & Paton *Acc. Chem. Res.* **2021,** *54*, 827

Organic Parameters of Interest

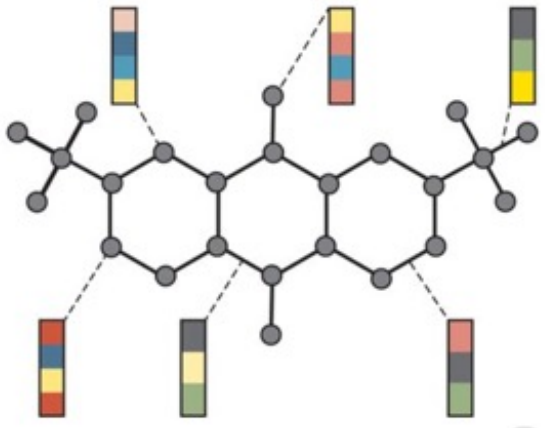
Reaction Thermochemistry



BDFE/BDE prediction: *Nat. Commun.* **2020,** *Digit Discov* **2023**

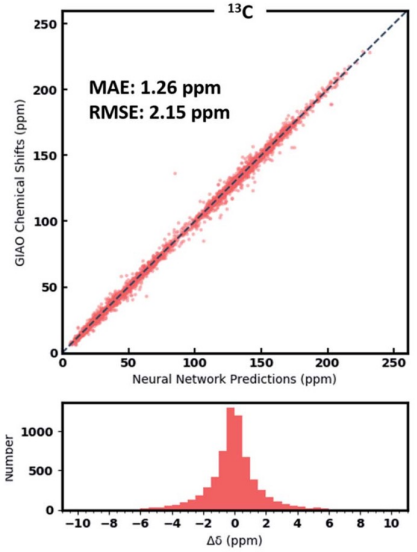


Lewis Acidity prediction: *ACIE* **2024**
(with Greb, Heidelberg)



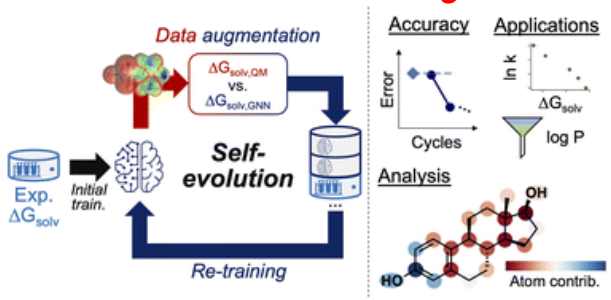
Graph Neural Network-based predictions

Spectroscopy



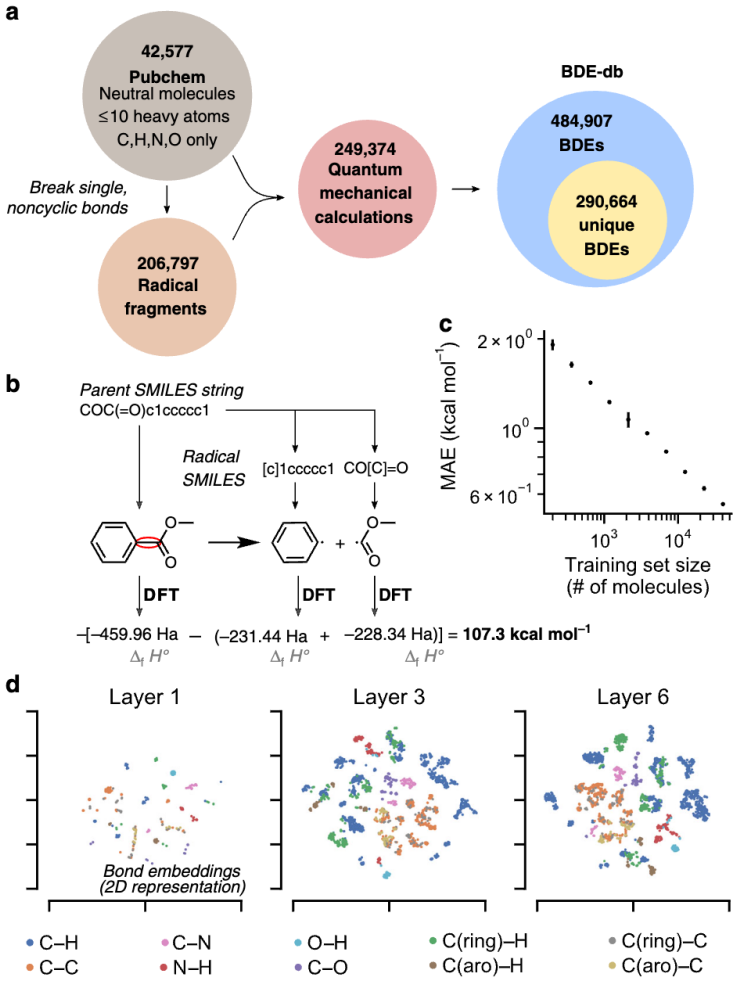
^{13}C/^{1}H NMR : *Chem. Sci.* **2021**

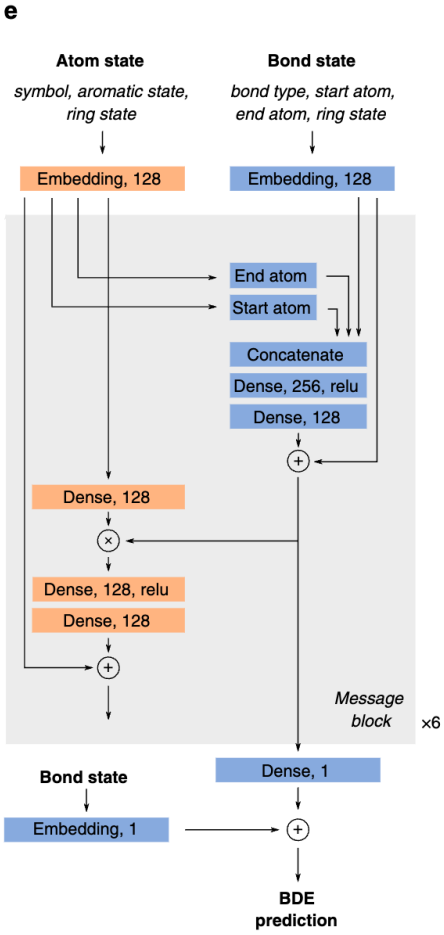
Solvation Free Energies

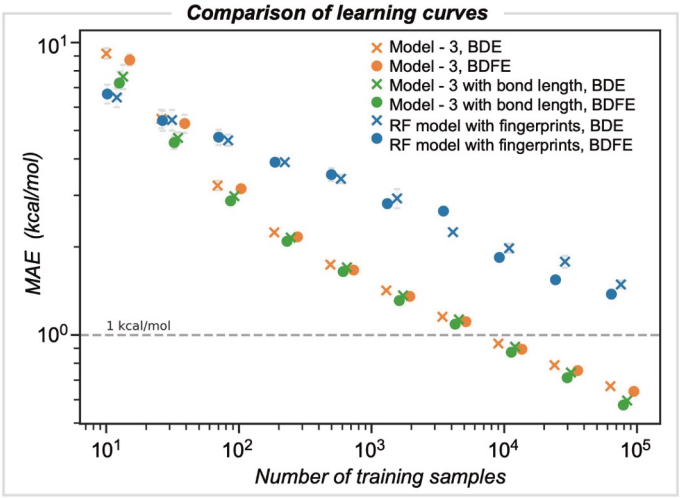


ΔG_{solv}: *Chem. Sci.* **2024**
(with Kim, CSU)

Deep Learning Approach

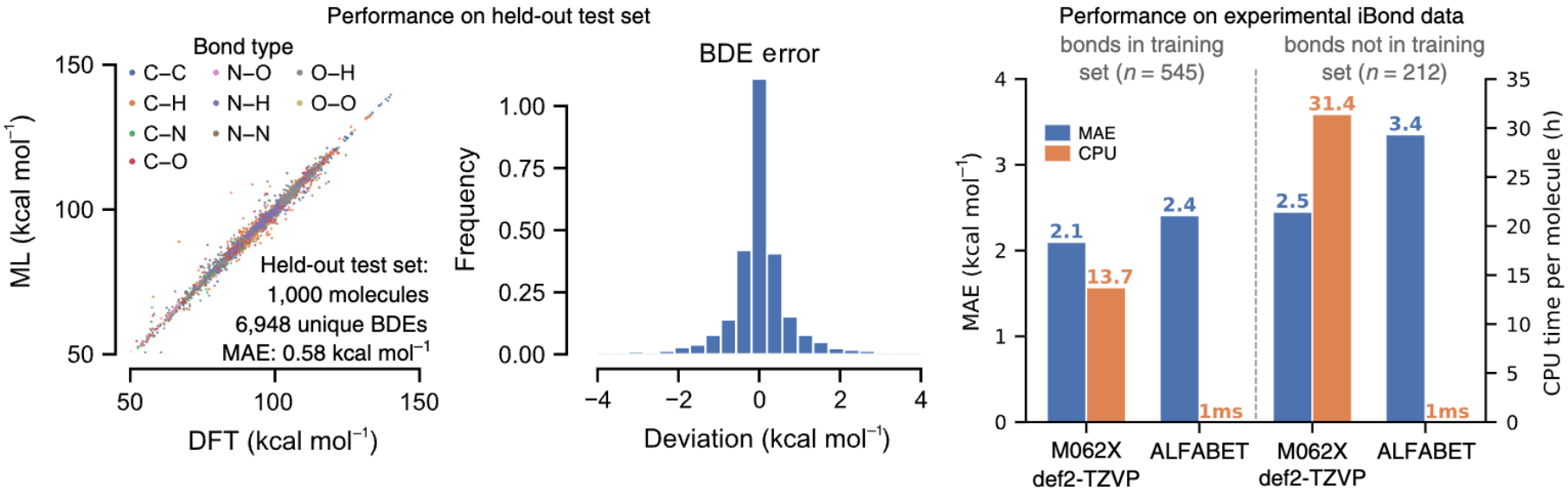


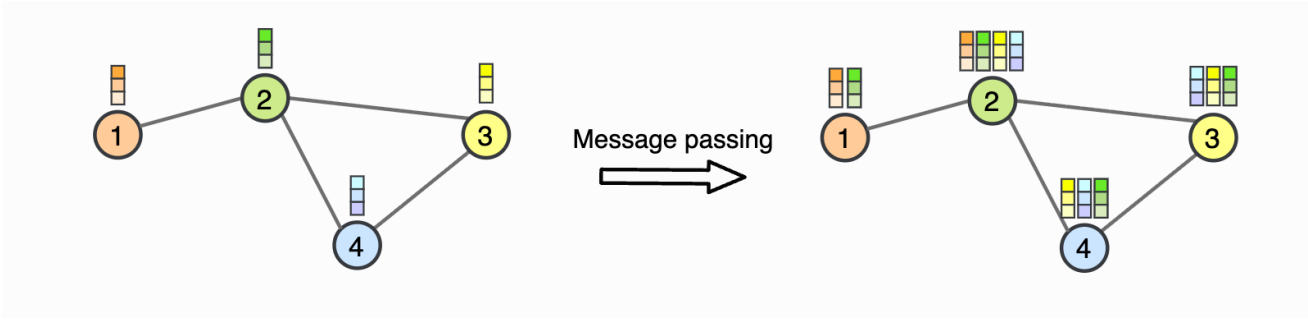




St. John, Guan, Kim, Kim & Paton *Nat. Commun.* **2020**, *11*, 2328; Sowndarya, St. John, Kim, Kim & Paton *Digit. Discov.* **2023**, *2*, 1900

An ML derived, Fast, Accurate Bond Dissociation Enthalpy Tool (ALFABET)

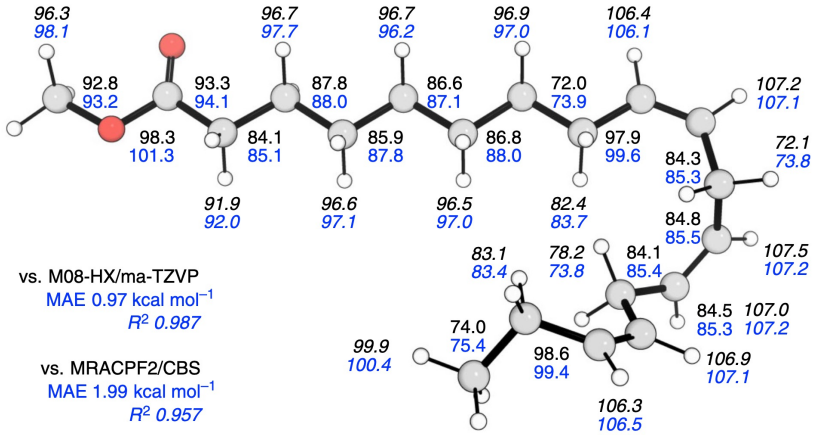


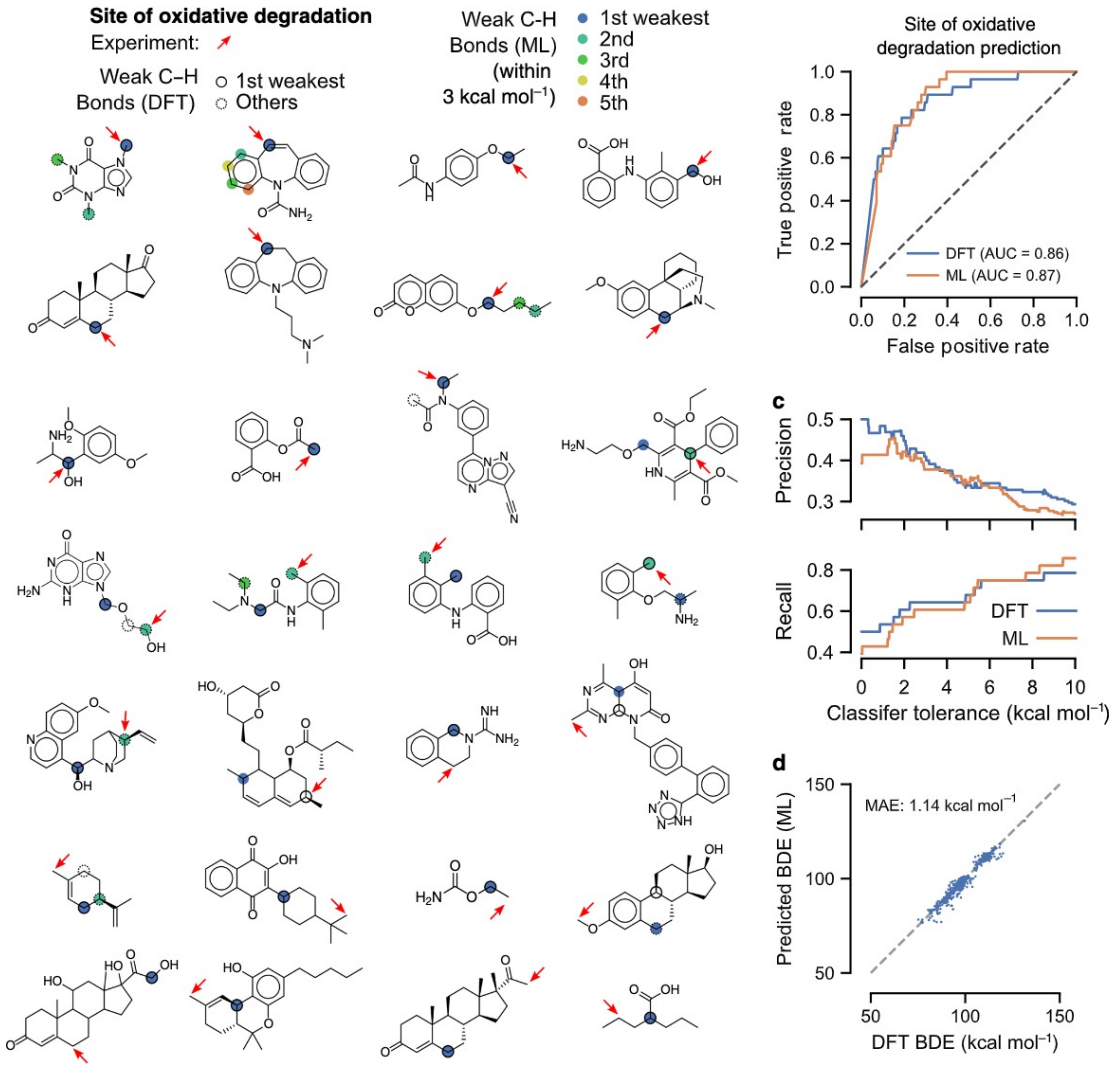


St. John, Guan, Kim, Kim & Paton *Nat. Commun.* **2020**, *11*, 2328

Code: https://github.com/NREL/nfp

Performance of ALFABET





St. John, Guan, Kim, Kim & Paton *Nat. Commun.* **2020**, *11*, 2328

Applications of ALFABET

ALFABET

A machine-**L**earning derived, **F**ast,
Accurate **B**ond dissociation **E**nthalpy **T**ool

Enter a SMILES string, e.g. 'CC1=CC(=CC(=C1)O)C'

DRAW SUBMIT

Machine learning predictions of bond dissociation energies (BDEs)

This tool predicts BDEs for single, noncyclic bonds in neutral organic molecules consisting of C, H, O, and N atoms. Mean absolute errors are typically less than 1 kcal/mol for most compounds. To use, enter a SMILES string above (or use the drawing tool) and press submit. Reference DFT-calculated BDEs used as training can be displayed for any predicted bond using the neighbors link.

For more info, see our publications:

St. John, P.C., Guan, Y., Kim, Y., Kim, S., and Paton, R.S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat Commun* **11**, 2328 (2020). https://doi.org/10.1038/s41467-020-16201-z

St. John, P.C., Guan, Y., Kim, Y., Etz, B.D., Kim S., and Paton, R.S. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci Data* **7**, 244 (2020). https://doi.org/10.1038/s41597-020-00588-x

The National Renewable Energy Laboratory is a national laboratory of the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, operated by the Alliance for Sustainable Energy, LLC.

Security & Privacy Apply for a Job

HOME

RESEARCH

Advanced Manufacturing
Bioenergy
Buildings
Chemistry and Nanoscience
Computational Science
Energy Analysis
Energy Systems Integration Facility
Geothermal
Grid Modernization
Hydrogen and Fuel Cells
Integrated Energy Solutions
Materials Science
Solar
Transportation
Water Power
Wind Energy

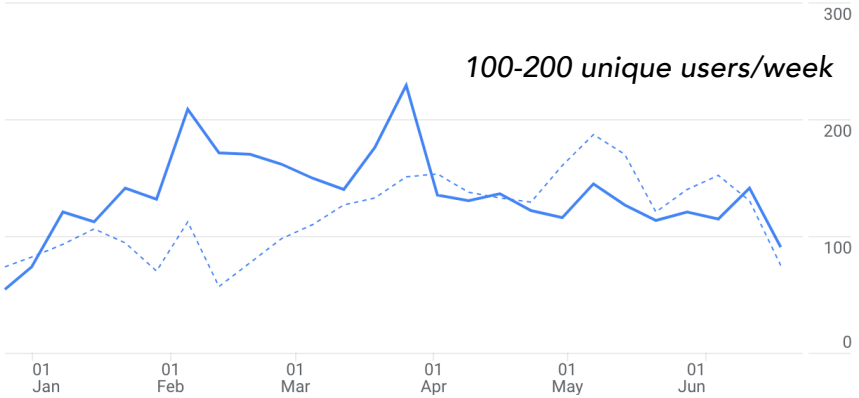
FOLLOW NREL

St. John, Guan, Kim, Kim & Paton *Nat. Commun.* **2020**, *11*, 2328

pip install alfabet

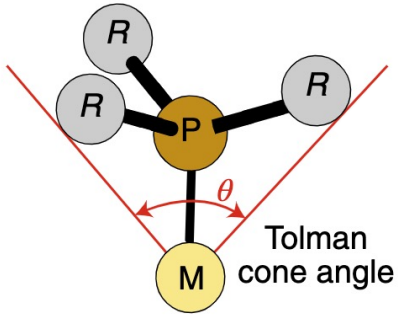
```
>>> from alfabet import model
>>> model.predict(['CC', 'NCCO'])
```

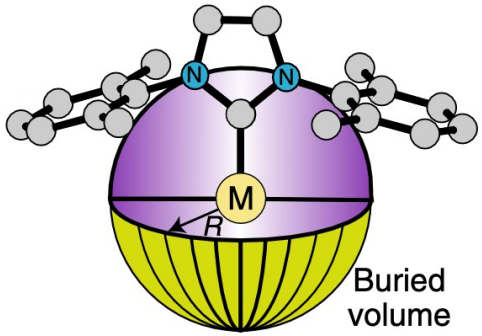
	molecule	bond_index	bond_type	fragment1	fragment2	...	bde_pred
0	CC	0	C–C	[CH3]	[CH3]	...	90.278282
1	CC	1	C–H	[H]	[CH2]C	...	99.346184
2	NCCO	0	C–N	[CH2]CO	[NH2]	...	89.988495
3	NCCO	1	C–C	[CH2]O	[CH2]N	...	82.122429
4	NCCO	2	C–O	[CH2]CN	[OH]	...	98.250961
5	NCCO	3	H–N	[H]	[NH]CCO	...	99.134750
6	NCCO	5	C–H	[H]	N[CH]CO	...	92.216087
7	NCCO	7	C–H	[H]	NC[CH]O	...	92.562988
8	NCCO	9	H–O	[H]	NCC[O]	...	105.120598

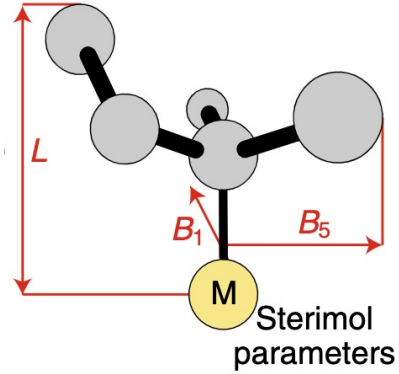


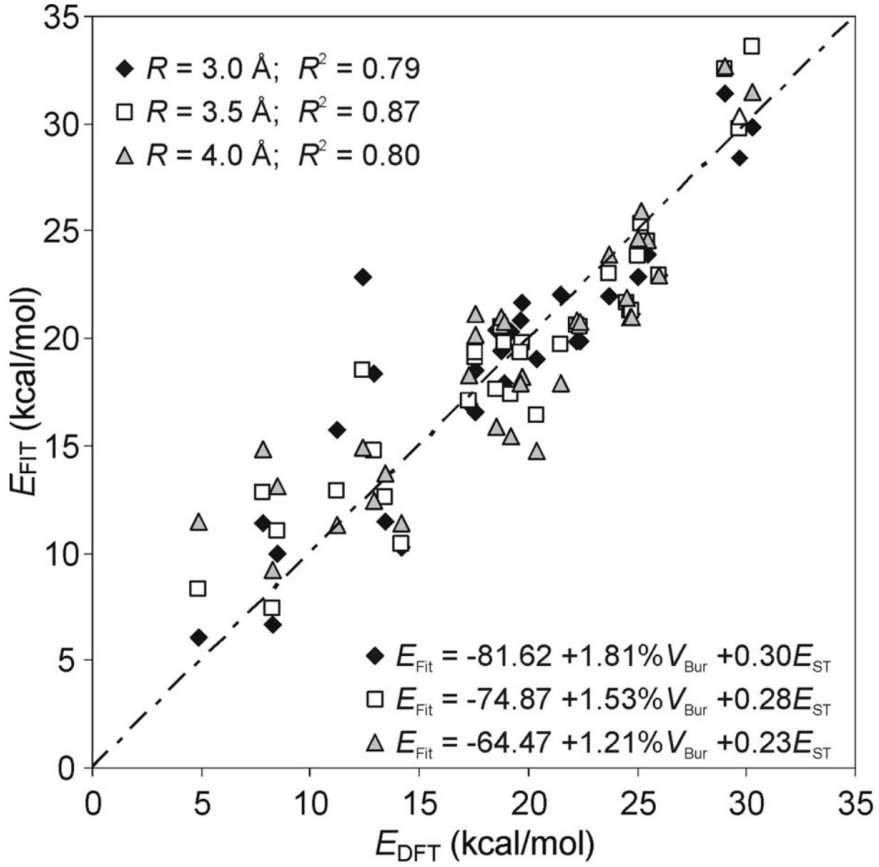
Webserver: bde.ml.nrel.gov

3D Molecular Descriptors: Buried Volume





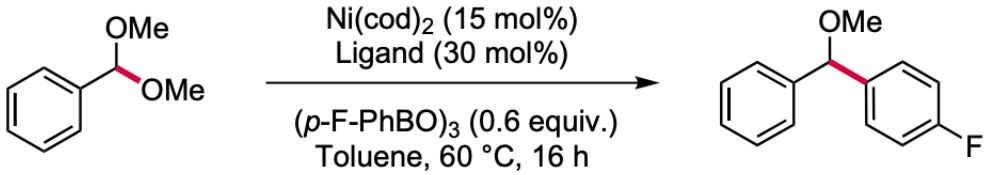


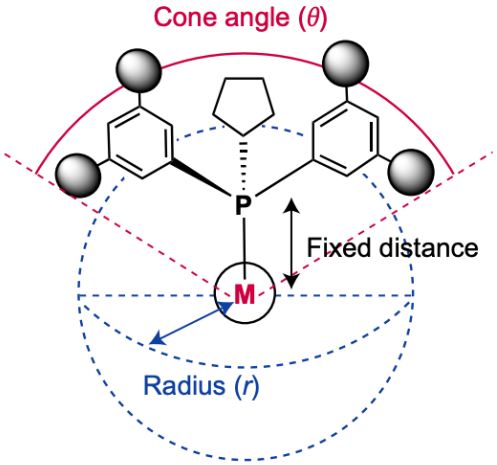


https://www.molnac.unisa.it/OMtools/sambvca2.1/index.html

Eur. J. Inorg. Chem. **2009**, 1759–1766; Nat. Chem. 2019, 11, 872

3D Molecular Descriptors: Buried Volume

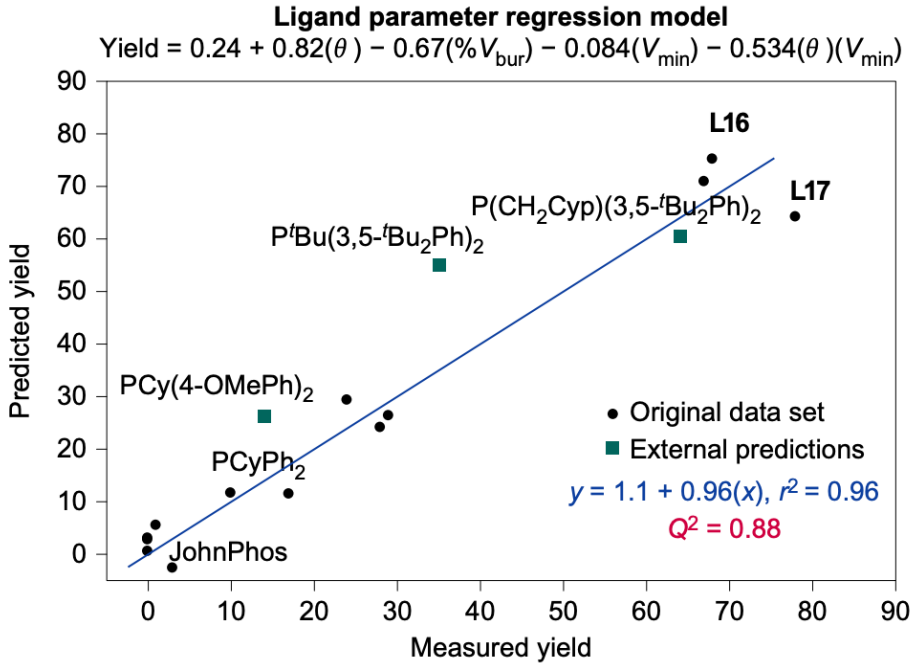




$\%V_{bur}$ = % volume of sphere of radius r occupied by ligand; any group beyond r is not captured

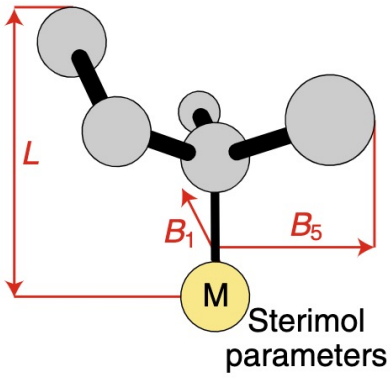
θ = angle swept by cone that encloses all ligand groups

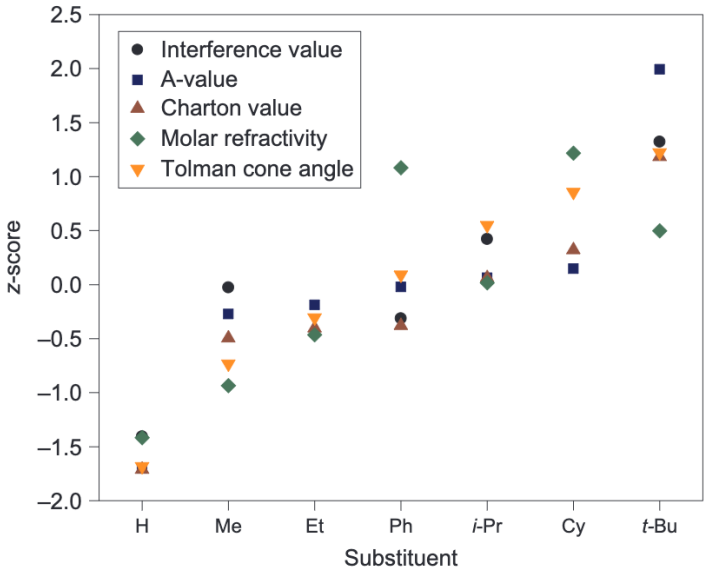
Remote steric hindrance = high θ and low $\%V_{bur}$

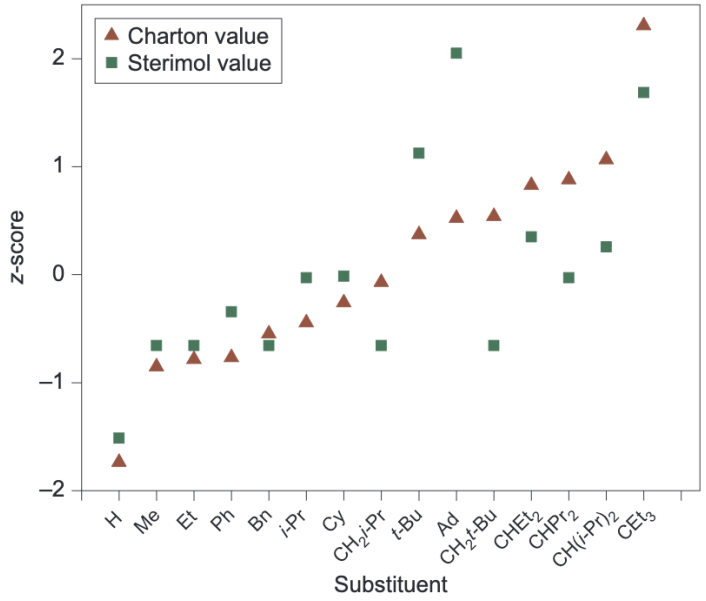


Nat. Chem. **2017**, 9, 779

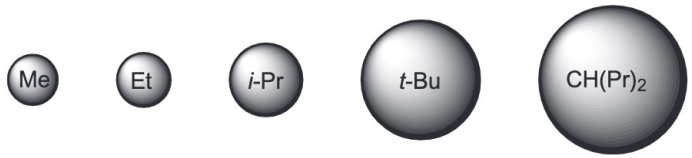
3D Molecular Descriptors: Sterimol





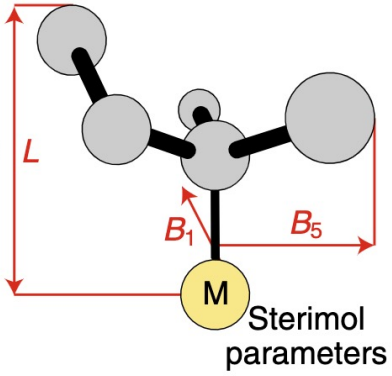


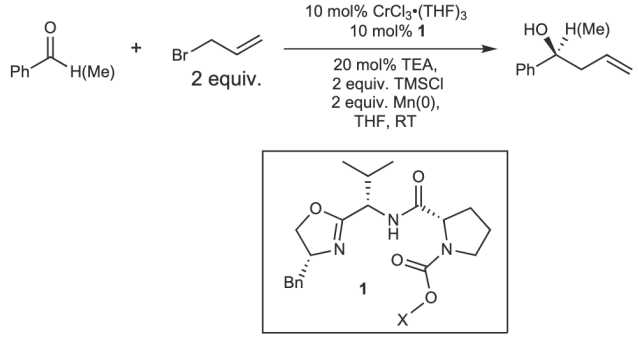
♦ Scaled sphere approximations

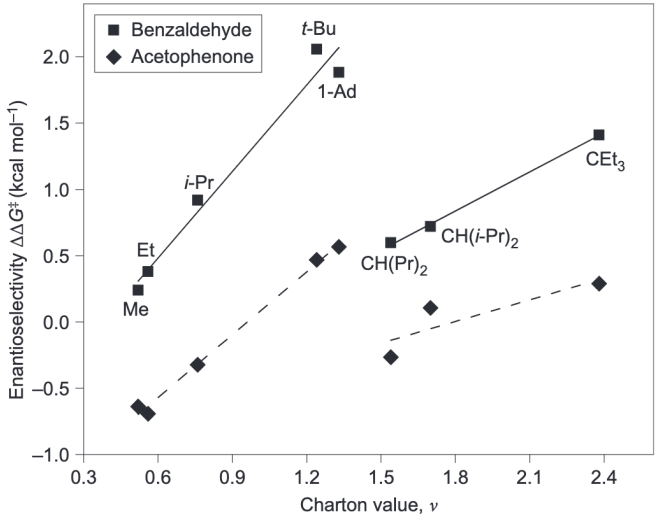


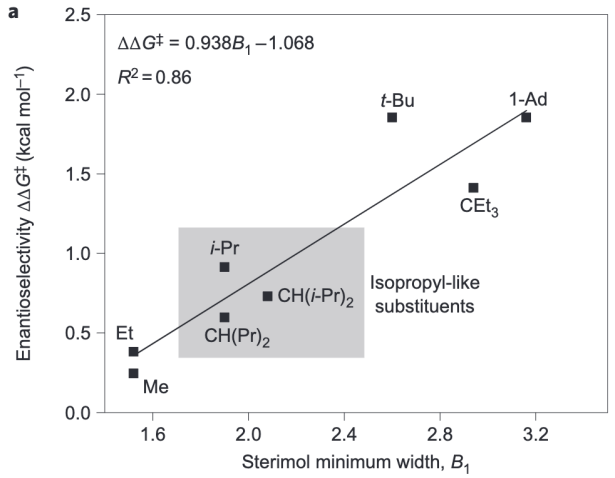
Nat. Chem. **2012**, 4, 366

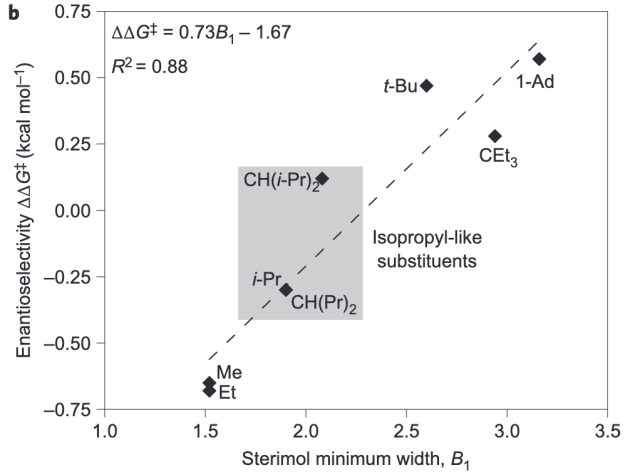
3D Molecular Descriptors: Sterimol





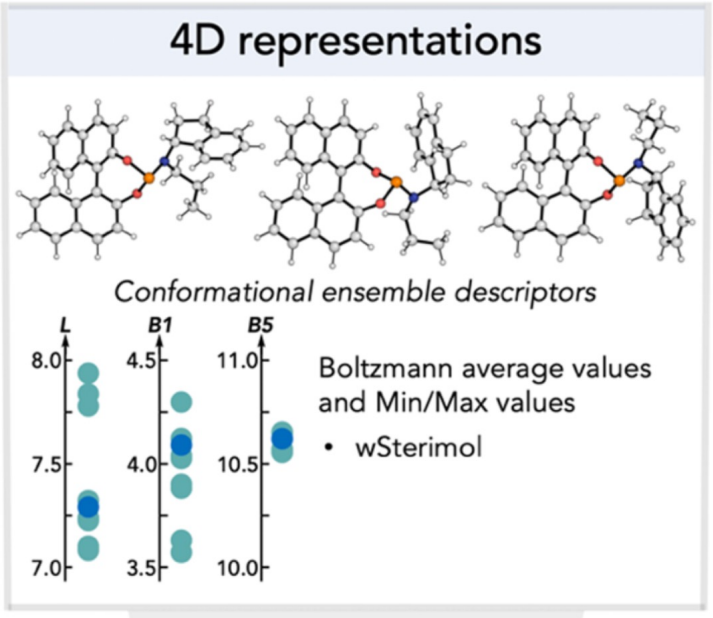




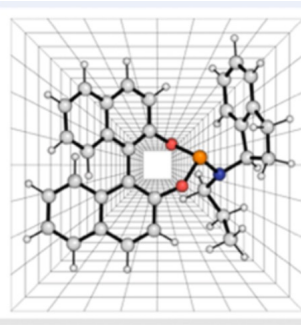


Nat. Chem. **2012**, 4, 366

4D Molecular Descriptors?



- Boltzmann average
- Most stable conformer
- Minimum/Maximum values



COMFA / Grid QSAR
- Alignment problem

ACS Catalysis **2019**, 9, 2313