Learning Molecular Properties: Structural Representation of Molecular Data

Chloé-Agathe Azencott

Institute for Genomics and Bioinformatics
University of California, Irvine

Advancement to Candidacy

Some Questions

- Is the compound active / inactive against the disease? (*classification*)
- Is the compound toxic? carcinogenic? mutagenic? (*classification*)
- What is the aqueous solubility of the compound? Its partition coefficient? (*regression*)
- Which compounds in the library are active against the disease? (*early recognition*)
- ...

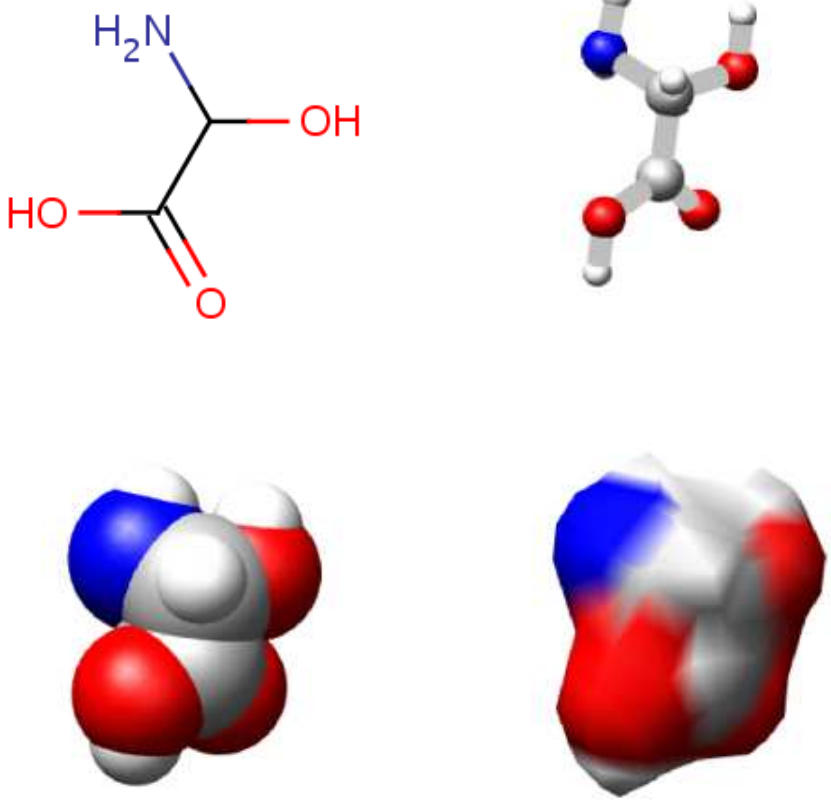
Goal

Define a framework in which computers can learn this kind of information *in silico*

Representing Molecular Data: An Issue I

- Finding molecular descriptors that encapsulate the necessary information is hard
- Expert knowledge may be incomplete
- We need a representation that is *"complete"* enough.

Representing Molecular Data: An Issue II



Outline

Using Structural Information

Using Structural Information

- Based on substructures of the molecule
- At different levels (1D to 4D)
- Allow to define *spectral kernels*

Spectral Kernels

- Define *feature vectors* that record the presence/absence (or number of occurrences) of particular substructures in a given structure
- $$\phi(A) = (\phi_s(A))_{s \text{ substructure}}$$

  where
  $$\phi_s(A) = \begin{cases} 1 & \text{if } s \text{ occurs in } A \\ 0 & \text{otherwise} \end{cases}$$
- Extension of traditional chemical fingerprints
- Then, define a *similarity* between these feature vectors

Classic Similarities

$A = (a_i)_{i=1...N}$ and $B = (b_i)_{i=1...N}$ two feature vectors

- Dot product

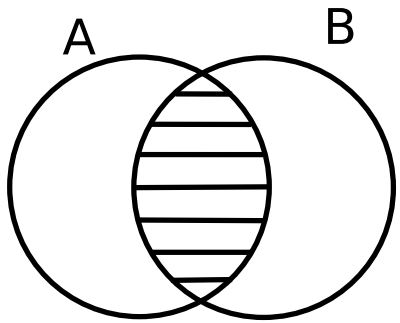
$$k(A, B) = \sum_{i=1}^{N} (a_i \times b_i)$$

- Gaussian RBF

$$k(A, B) = \exp(-\frac{\sum_{i=1}^{N} (a_i + b_i)^2}{2\sigma^2})$$

- Both are kernels

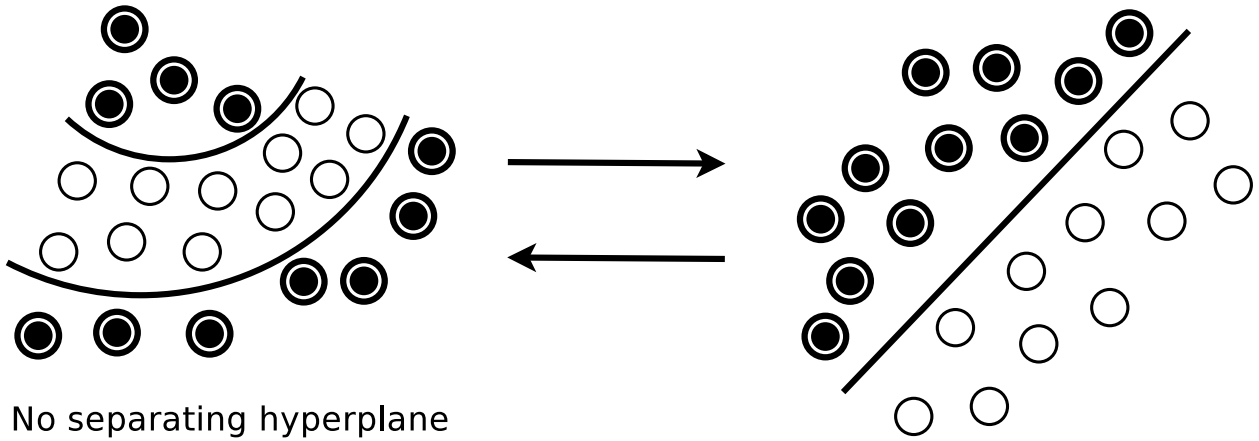
Tanimoto and MinMax



- Tanimoto (binary setting): $k(A,B) = \frac{A \cap B}{A \cup B}$
- MinMax (counts setting): $\frac{\sum_{i=1}^{N} \min(A_i, B_i)}{\sum_{i=1}^{N} \max(A_i, B_i)}$

Support Vector Machines

Linear separation in a feature space



No separating hyperplane
in the input space

Separating hyperplane
in the feature space

$$\exists \phi, k(A,B) = \langle \phi(A), \phi(B) \rangle$$

$$f(A) = \sum_{i=1}^{M} \alpha_i k(A, A_i) + b$$

SVM-Related Issues

- Hyperparameters optimization and model selection
  - Exhaustive grid-search
  - Cross-validation
- Unbalanced data
  - Different C parameter (error/margin trade-off) for different classes
  - Data re-sampling
- Dataset (and fingerprints) size
  favor *online* implementations

Representation of Molecules by 1D-4D Structural Information

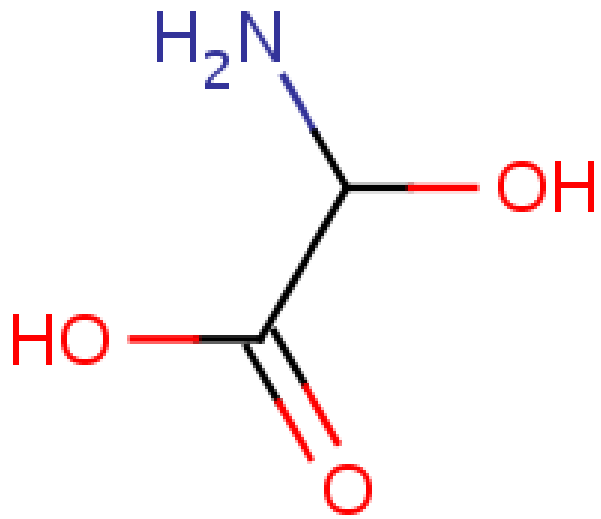
Example

Melting Point Prediction - Karthikeyan Dataset

- 4,173 molecules with melting point value
- (Previous) best published results: squared correlation coefficient of 0.42 and average absolute error of 41.3°C
- Hard problem!

SMILES Strings

- unique representation of a molecule as a string
- widely used for characterization (especially in databases)
- E.g. NC(O)C(=O)O



One-Dimensional Substructures

- 1D substructures:
  substrings of length l of the SMILES string
- $l = \infty$ is possible
- NC(O)C(=O)O

NC	1
C(	2
(O	1
O)	2
)C	1
(=	1
=O	1
)O	1

Results on Karthikeyan

Method	r^2	RMSE	AAE
1D	**0.52**	**44.88**	**34.30**
2D	0.56	42.71	32.58
2.5D Delaunay	0.49	46.07	35.37
3D Delaunay	0.50	45.62	35.01
3D Histogram	0.27	55.01	43.38
3.5D Delaunay	0.44	48.35	37.44
4D Delaunay	0.35	55.36	43.43
4D Histogram	0.40	50.40	39.85
Previous Best	**0.42**	**52.0**	**41.3**

Molecular Graph

- Vertices ≡ atoms
  Labeling:
  - Element: by symbol (e.g. C, O, N)
  - Element-Connectivity: by symbol + number of neighbors (e.g. C3, N1)
  - Element-Hybridization: by symbol + hybridization state
  - Sybyl: Tripos atom typing system (e.g. N.ar, N.am, N.pl3)
- Edges ≡ bonds
  Labeling:
  - Bond-type: by bond type (e.g. s, d, ar)
  - All bond equivalents, . . .

Two-Dimensional Substructures I

- Labeled sub-paths (walks)

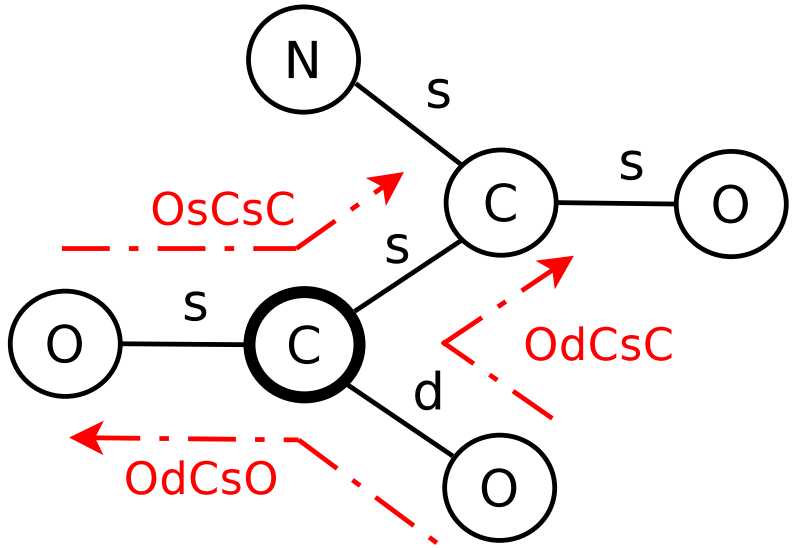


Figure: Some sub-paths of depth 2

- Labeled sub-trees - Extended-Connectivity (or Circular) features

Two-Dimensional Substructures II

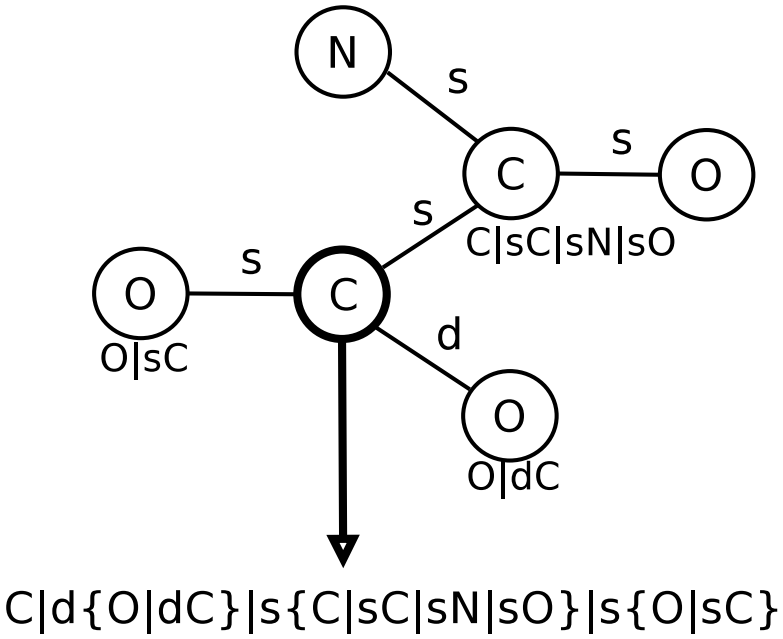


Figure: Example of a circular substructure of depth 2

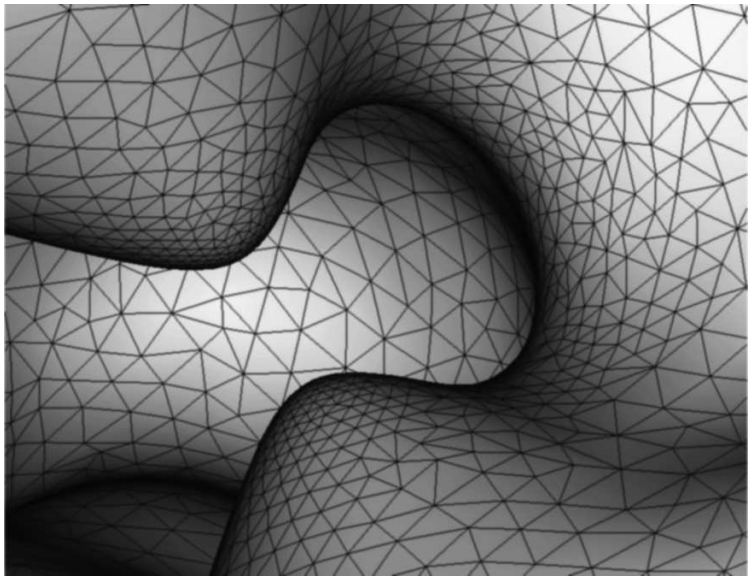
Results on Karthikeyan

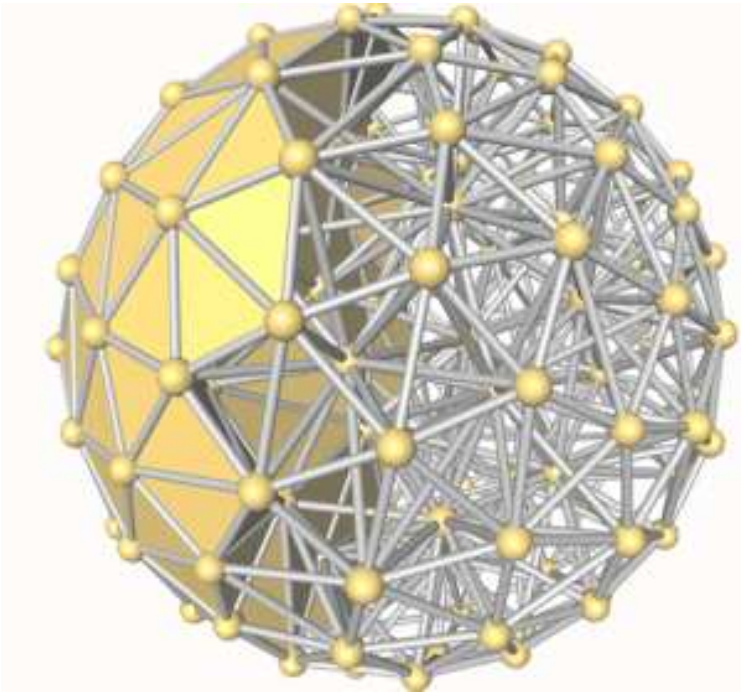
Optimal parameters: paths of depth up to 10, labeling, counts

Method	r^2	RMSE	AAE
1D	0.52	44.88	34.30
2D	**0.56**	**42.71**	**32.58**
2.5D Delaunay	0.49	46.07	35.37
3D Delaunay	0.50	45.62	35.01
3D Histogram	0.27	55.01	43.38
3.5D Delaunay	0.44	48.35	37.44
4D Delaunay	0.35	55.36	43.43
4D Histogram	0.40	50.40	39.85
Previous Best	**0.42**	**52.0**	**41.3**

Delaunay Tetrahedrization I

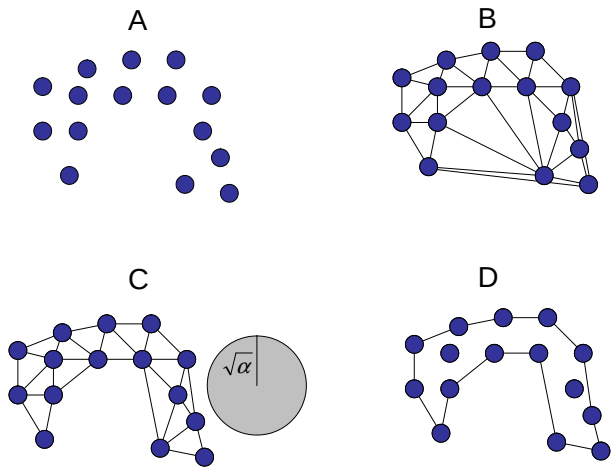
- Approximate the convex hull of a set of points (ie. atoms) by a set of tetrahedra (Delaunay)

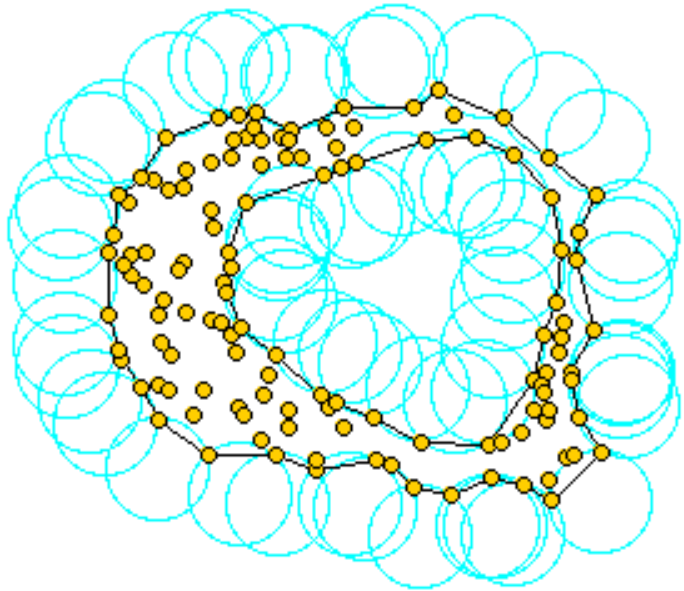




- Alpha-shape (≡ solvent-accessible surface)

Delaunay Tetrahedrization II





- Keep the interior edges (3D) or not (2.5D)
- => Labeled graph

Surface-Based Substructures

- Substructures:
  Same as 2D, on the Delaunay tetrahedrization graph

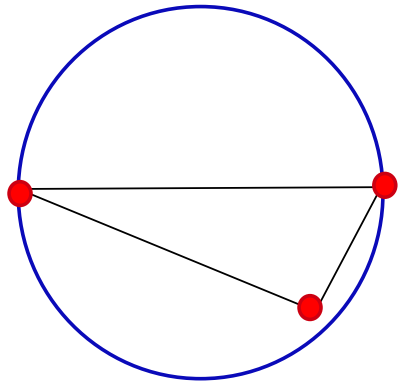
Results on Karthikeyan

Optimal parameters: depth up to 3, Sybyl labeling, counts

Method	r^2	RMSE	AAE
1D	0.52	44.88	34.30
2D	0.56	42.71	32.58
2.5D Delaunay	**0.49**	**46.07**	**35.37**
3D Delaunay	**0.50**	**45.62**	**35.01**
3D Histogram	0.27	55.01	43.38
3.5D Delaunay	0.44	48.35	37.44
4D Delaunay	0.35	55.36	43.43
4D Histogram	0.40	50.40	39.85
Previous Best	**0.42**	**52.0**	**41.3**

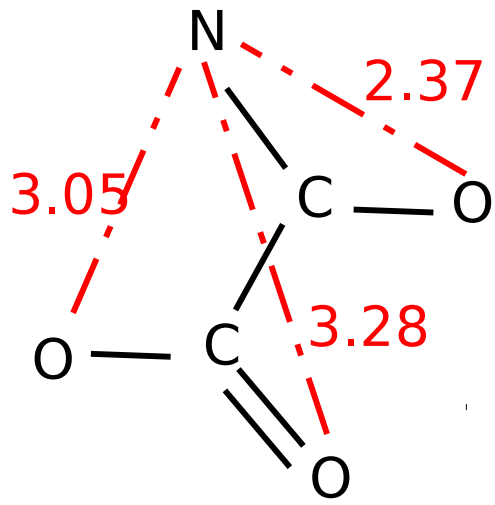
Atom Coordinates / Pharmacophores I

- Groups of k atoms
- Associated size:
  - Pairwise distances ($k = 2$)
  - diameter of the smallest sphere that contains all k atoms



- One histogram per class of k-tuple (e.g. C-C-C, C-C-O)

Atom Coordinates / Pharmacophores II



Histogram for (N-O) pairs:

dist	0.0-0.5	0.5-1.0	1.0-1.5	1.5-2.0	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0	...
#	0	0	0	0	1	0	2	0	

Three-Dimensional Substructures

- 3D substructure: k-tuple of atoms contained in a sphere of diameter d

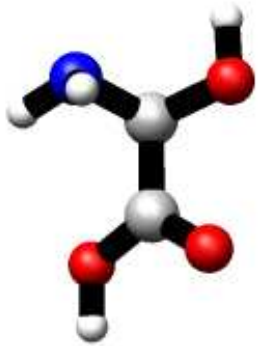
Results on Karthikeyan

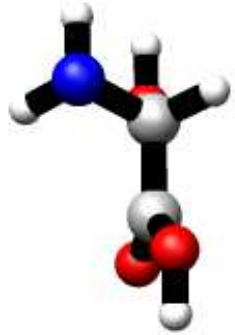
Optimal parameters: k=2, bin=0.1

Method	r^2	RMSE	AAE
1D	0.52	44.88	34.30
2D	0.56	42.71	32.58
2.5D Delaunay	0.49	46.07	35.37
3D Delaunay	0.50	45.62	35.01
3D Histogram	**0.27**	**55.01**	**43.38**
3.5D Delaunay	0.44	48.35	37.44
4D Delaunay	0.35	55.36	43.43
4D Histogram	0.40	50.40	39.85
Previous Best	**0.42**	**52.0**	**41.3**

Multiple Conformations

- Rotatable bonds => Conformers
- Stereocenters => Isomers





Kernel on Multiple Conformations

- Generate up to 15 conformations (A^i) for each molecule (A)
  ie. up to 255 pairs
- $$k_{multiple}(A,B) = \frac{\sum_{i \in \text{conf}(A)} \sum_{j \in \text{conf}(B)} k(A^i, B^j)}{|\text{conf}(A)|.|\text{conf}(B)|}$$

Results on Karthikeyan

Optimal parameters

- 3.5D: depth up to 3, Element-Hybridization labeling, counts
- 4D: k=2, bin=0.1

Method	r^2	RMSE	AAE
1D	0.52	44.88	34.30
2D	0.56	42.71	32.58
2.5D Delaunay	0.49	46.07	35.37
3D Delaunay	0.50	45.62	35.01
3D Histogram	0.27	55.01	43.38
3.5D Delaunay	**0.44**	**48.35**	**37.44**
4D Delaunay	**0.35**	**55.36**	**43.43**
4D Histogram	**0.40**	**50.40**	**39.85**
Previous Best	**0.42**	**52.0**	**41.3**

Conclusion on Chemical Data Representation

- 2D representation yields the best results
- 1D loses branching information
- 3D information is predicted by CORINA (not known) averaging over multiple conformations improves the quality of the prediction

Some More Results

Classification

Classification accuracy on small inhibitors datasets

dataset (train, valid)	best performance	previous best
BZR (181, 125)	**79.8 %**	76.4 %
COX2 (178, 125)	70.1 %	**73.6 %**
DHFR (233, 160)	**83.0 %**	81.9 %
ER (266, 180)	**82.1 %**	79.8 %

Regression: Prediction of Aqueous Solubility

- Huuskonen dataset: 1,026 compounds, 10-fold cross-validation

method	r^2
2D	0.90
4D (histograms)	**0.91**
previous best	0.90

- Delaney dataset: 1,144 compounds, 10-fold cross-validation

method	AAE
2D	**0.44**
4D (histograms)	0.45
previous best	0.75

Regression: Prediction of Octanol/Water Partition Coefficient

XLOGP dataset: 1,991 compounds,

- 10-fold cross-validation

method	r^2
2D	**0.94**
4D (histograms)	0.92

- Training (1.853) and validation (138) sets

method	r^2	RMSE
2D	**0.946**	**0.338**
previous best	0.944	0.348

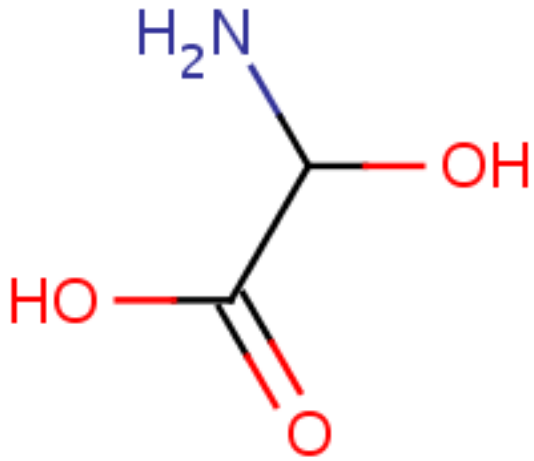
High-Throughput Screening

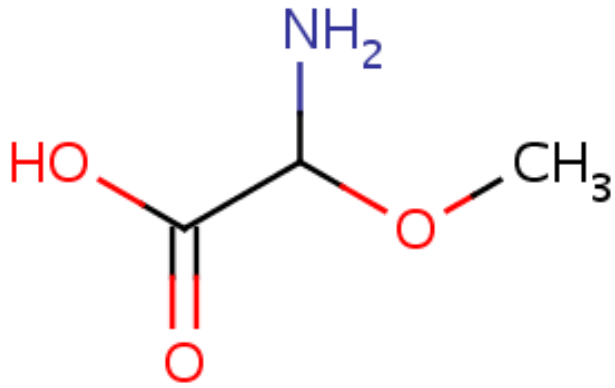
- HIVA dataset from the Agnostic Learning vs. Prior Knowledge Challenge (IJCNN07)
- 42,678 chemicals with activity against HIV
- Training on 4,229 compounds, testing on 38,449 remaining compounds
- Goal: best BER on the testing set
- 2D (circular): BER = **0.2693** **winning entry**

Further Improvement Of The Predictions

Improvement of the Predictions I

- Get more data
- Improve on the molecular representation
  - tan("O=C=Nc1ccccc1", "O=C=Nc1ccc**(C)**cc1") = 0.35





  - Improvement on the 3D
  - Improvement on the multiple conformations
  - ...
- Improve on the machine learning

Improvement of the Predictions II

- Model selection / hyperparameters optimization
- Neural Networks
- ...

Conclusion

- Capturing the necessary information regarding molecules is a **hard problem**
- Representations based on molecular **structural information** are a possible solution
  - Alleviate the need for relying on (potentially incomplete) expert knowledge
  - Produces a **fixed-size** representation for data of varying size
- **2D** representations based on molecular graphs appear (generally) as the most powerful representation

Acknowledgments I

- Pierre Baldi
- Joshua Swamidass
- Ryan Benz, Jonathan Chen, Kenny Daily, Ramzi Nasr

Publications I

C.-A. Azencott, A. Ksikes, S. J. Swamidass, J. H. Chen, L. Ralaivola, and P. Baldi.
One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties.
J. Chem. Inf. Model, 2007.

C.-A. Azencott and P. Baldi.
Virtual High-Throughput Screening with Two-Dimensional Kernels, Hands-On Pattern Recognition. Challenges in Data Representation, Model Selection, and Performance Prediction,
I. Guyon and G. Cawley and G. Dror and A. Saffari editors
Lulu press, *in press*.

Publications II

 C.-A. Azencott, S. J. Swamidass, and P. Baldi. Learning High-Throughput Screening Data, *draft*.

Thank you!