



# *Problems and Approaches in Computational Chemistry*



 POLITECNICO DI MILANO



## **Reasoning about molecular similarity and properties**

*Original Work by Rahul Singh*

*Dept. Comp. Science, San Francisco State University*

*Presentation by Giorgio Orsi, Davide Mazza*

*Dept. of Electronics and Information. Politecnico di Milano*



## Molecular similarity

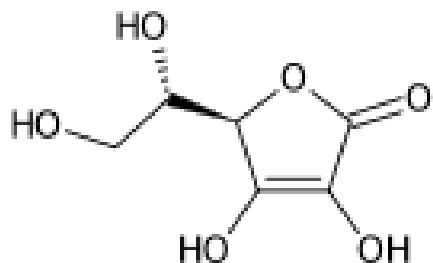
- The problem of finding similarities among molecules on the basis of a set of (chemical/physical/geometrical) properties P [5].
- Similar molecules tend to behave similarly
  - Properties represented by descriptors
  - Some properties influence other properties (e.g., geometry --> energy status)
- Similarity test is used for:
  - Exploration of molecular structural space
  - Development of structure-property models
  - Querying molecular structure databases



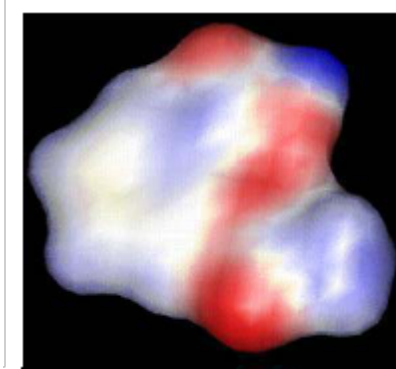
- The representation format influence similarity measures:

Example: Ascorbic Acid

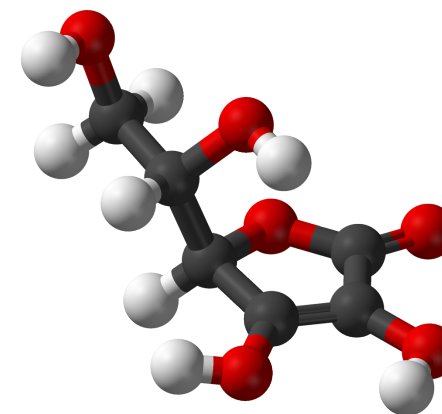
2D molecular graph



3D shape



3D structure



1D InChI notation

1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1



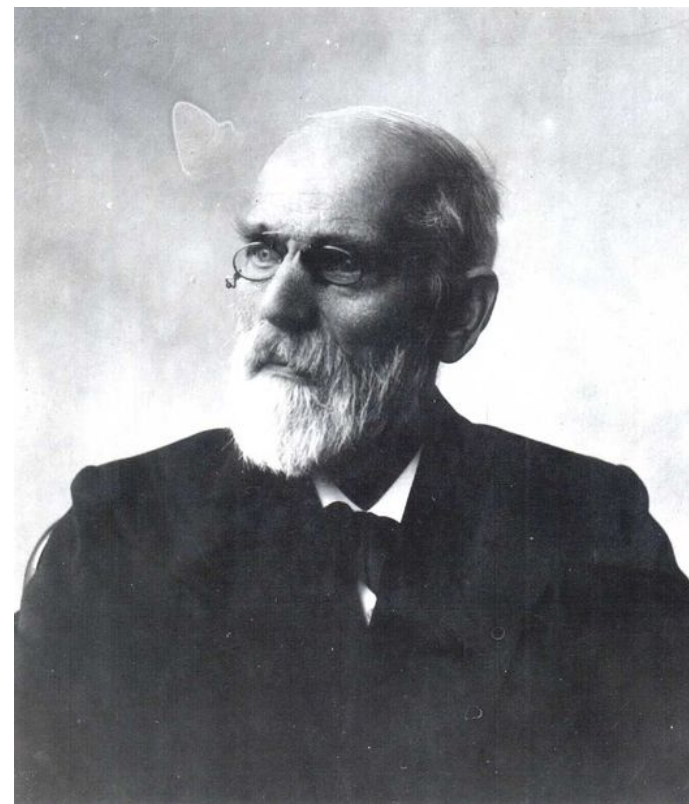
- Molecules represented as strings obtained by a depth-first or breadth-first molecular graphs traversal.
- Similarity based on ad-hoc string distance metrics.
- Similarity can be computed
  - directly on the string representation
  - among fingerprints



- Molecules represented as graphs
- Exact (graph-)similarity
  - Graph-isomorphism algorithms (MCS)
- Approximated similarity
  - Features Trees
    - Atoms determining certain properties are grouped in super-atoms
    - Super atoms represent **features**
    - Subtree matching algorithms (split search, match search, ...)
  - 2D fingerprints
    - modified hashing algorithms to ensure reversibility and approximated search.

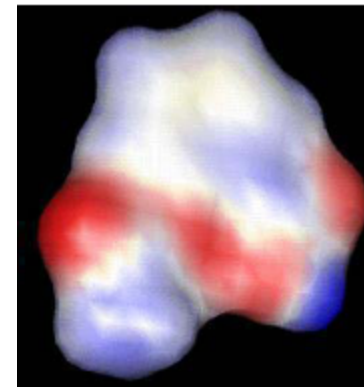


- Molecules represented as
  - 3D structures
  - 3D surfaces
- Similarities
  - Atom based
  - Pharmacophore based
  - Volume based
  - Surface based
- Techniques
  - Alignment based
  - Descriptor based
    - Van der Waals fields
    - Schrodinger wave equation



## Contextualization of the work

- Similarity test on small molecules
  - tens of atoms
  - weight ~ 1K Daltons
- Molecular representation
  - 3D surface-based descriptors
  - geometric shape, *donor* and *acceptor* fields
- Allows better query formulation with 3D surface-based descriptors
- Applications:
  - Pharmacology (drug absorption).
  - Molecular database design.
  - Molecule retrieval and mining.

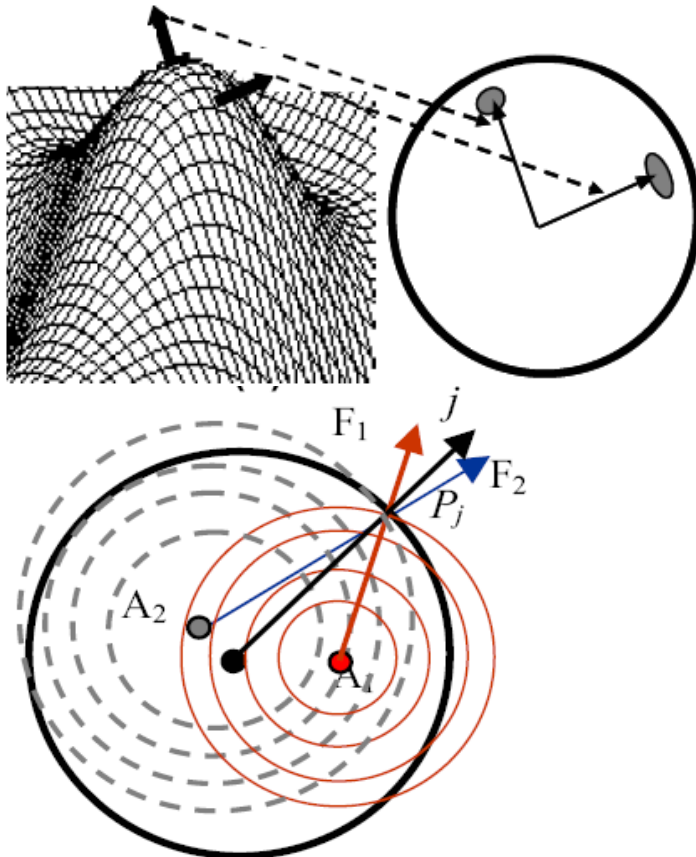




- Definition of a standard coordinate system for comparison
- Computation of property distributions (surface descriptors)
- Independency on pose and conformation
- Query efficiency (speed)
- Validation

## Standard coordinate system (1)

- Molecule representation
  - mapping between the molecule surface and a suitable sphere
  - molecule properties represented as distributions over the **unit sphere** [1].



Gauss Map:

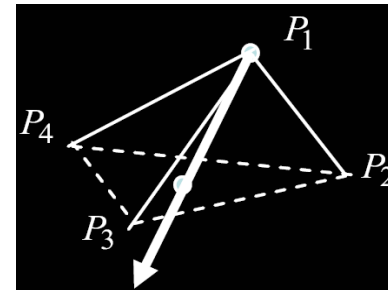
- $G \rightarrow \mathbb{R}^3$  the molecule surface
- $S$  the unit sphere

$M: G \rightarrow S$   
is a gauss map iff

- For each point  $p$  of  $G$ , the  $p$  normal is translated to the origin of  $S$
- The normal endpoint lies on the surface of  $S$

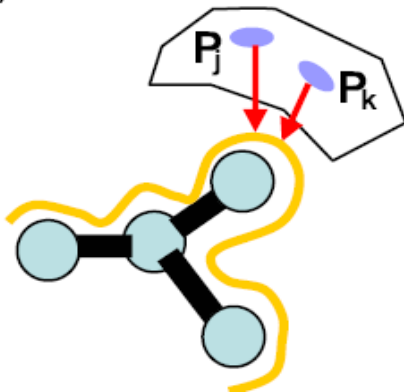
## Standard coordinate system (2)

- Mapping all points in unpractical
  - EGI (Extended Gaussian Image)
  - sampled gaussian map
  - each surface normal belonging to the same sample is associated with a single point on the sphere.
- EGI properties
  - two convex objects with the same EGI are provably congruent (Minkowski theorem)
  - if the surface rotates, the EGI rotates in the same way
  - EGI “mass” is proportional to the inverse of the surface curvature
  - the “center of mass” of the EGI coincides with the sphere's center.



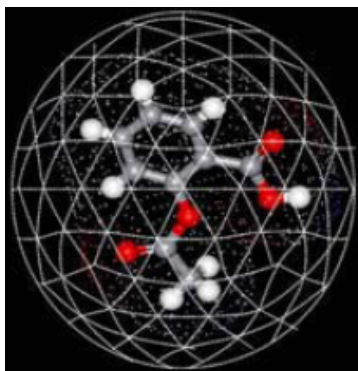
## Standard coordinate system (3)

- Problem: In non-convex shapes, multiple points can be mapped to the same point on the sphere.



*In computer vision the problem can be approached with SAI (Spherical Attribute Image).*

- SAI is computationally expensive:
  - tessellate the sphere
  - compute the needed properties on surface samples



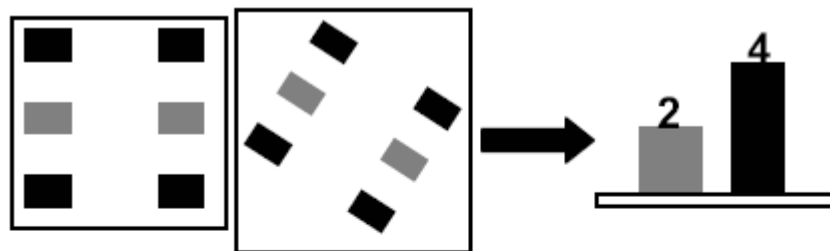
*Computed properties:*

- *geometric shape*
- *donor field*
- *acceptor field*

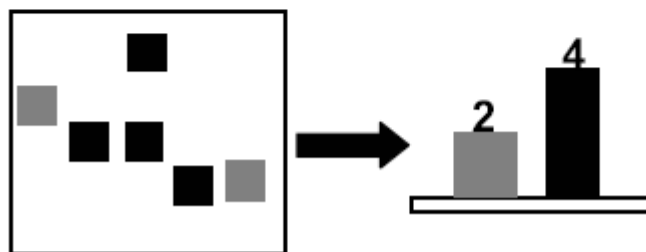
$$\left. \begin{array}{l} \text{donor field} \\ \text{acceptor field} \end{array} \right\} f(P_j, X_i) = \left( \frac{a^2}{2\pi r_i^2} \right)^{\frac{3}{2}} \exp\left( \frac{-a^2}{2r_i^2} |X_i - P_j|^2 \right)$$

## Similarity measure (1)

- Similarity of two molecules is defined in terms of the similarity of their property distributions.
- Property histograms' intersection is a rapid way to compute it, due to its high efficiency and invariance to pose (translations and rotations).
- But property histograms by themselves are insufficient to disambiguate distributions.

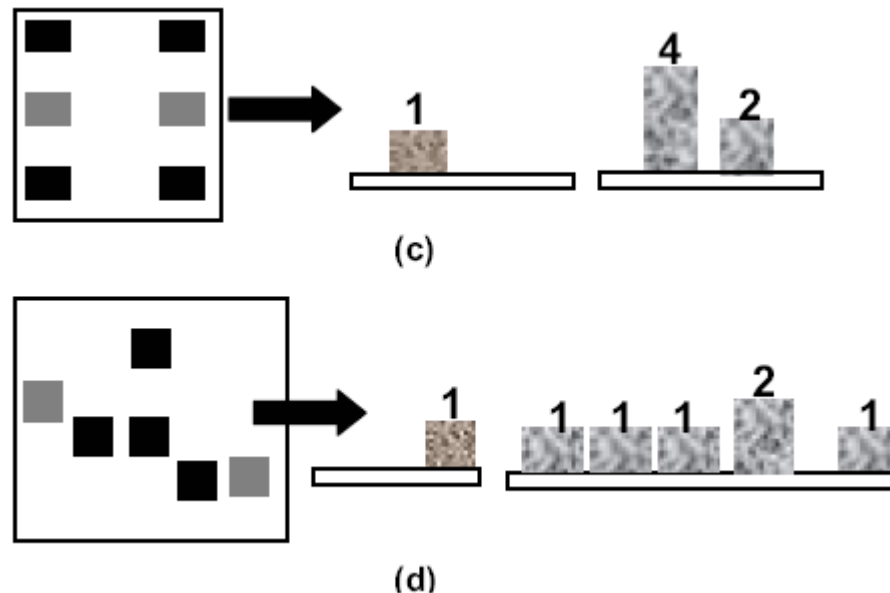


(a)



## Similarity measure (2)

- Other information needed: *spatial* distributions.
- Compute the pair-wise distances between elements with same properties (geometry, donor, acceptor → same color in figure, same class). As many histograms as many classes of elements.
- Quantize the distances in  $k$  bins from  $[0, 1[$ ,  $[1, 2[$ , ...  $[C/2-1, C/2[$  where  $C$  is the circumference of the sphere.





## Similarity measure (3)

- Similarity measure is obtained by the intersections of the same-property histograms of the two molecules to compare.
- Some figures to measure intersections are computed in sequence, to give a final similarity measure.

$$H(D_I, D_M) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{M_j}}; H(D_M, D_I) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{I_j}} \quad H_{TC}(M_1, M_2) = \frac{H(M_1, M_2) \cdot \gamma + H(M_2, M_1) \cdot \gamma}{2}$$

$$H(M_a, M_b) \gamma = \frac{\sum_{j=1}^K \min(M_{aj}, M_{bj}) \times \gamma_j}{\sum_{j=1}^K M_{aj}}$$

- $H_{full}$  is the average over all property distributions of the corresponding TC histogram intersection values.

$$\text{Similarity}(M_i, M_j) = \arg \max_{C_i, C_j} [H_{full}(C_i, C_j)]$$

$$C_i = \{C_i^1, C_i^2, \dots, C_i^r\}, C_j = \{C_j^1, C_j^2, \dots, C_j^r\}$$



## Accuracy in query-retrieval settings

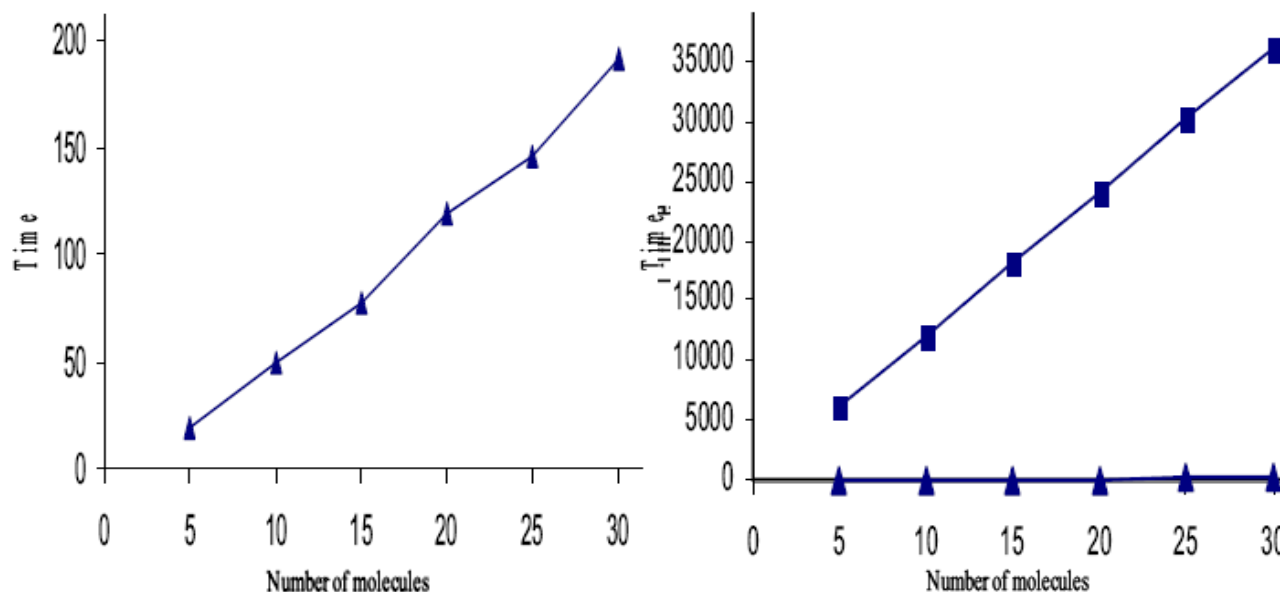
- Test set: 5000 molecules randomly selected from MDDR database.
- Query and target molecules represented by 20 conformers each
  - 400 distinct molecular structures
- Comparison made against ISIS (2D database)

<b>Method</b>	<b>Data Size</b>	<b>Number of Conformations</b>	<b>Accuracy</b>
ISIS	5000	none	100%
Proposed	5000	20/20	100%
Proposed	5000	20/20*	98.2%



## Evaluation of performance (speed)

- Test set: 30 molecules selected as maximally-diverse
- Query and target molecules represented by 20 conformers for the model and one for the query
  - 20 distinct structures
- Comparison with other surface-based algorithms like COMPASS [3] and Molecular Hashkeys [2]





## Validation through application (1)

- The proposed molecular similarity test has been applied to a real case: the building of a *structure-activity* model for modeling and predicting human intestinal drug absorption.
- Model construction (mapping between molecules) is built by employing a single-hidden-layer neural network using 20 training molecules.
- The descriptor design for the structure-activity model involves computing the similarity of the participating molecules with a predefined set of *characteristic molecules*.
- The similarity test is validated by using:

$$r^2 = 1 - \frac{\sum_i (V_i - P_i)^2}{\sum_i (V_i - \bar{V})^2}$$

cross-validation

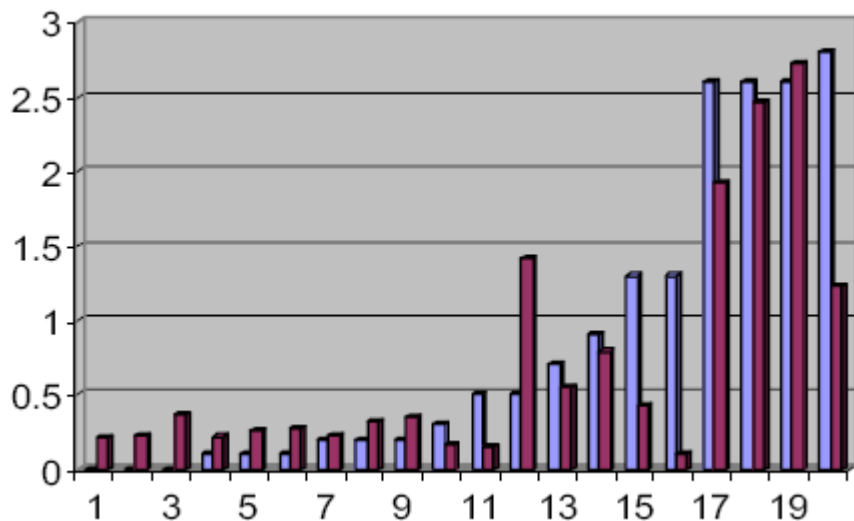
$$\tau = \frac{\text{correct\_ordering} - \text{incorrect\_ordering}}{n(n-1)/2}$$

Kendall's tau ordering measure

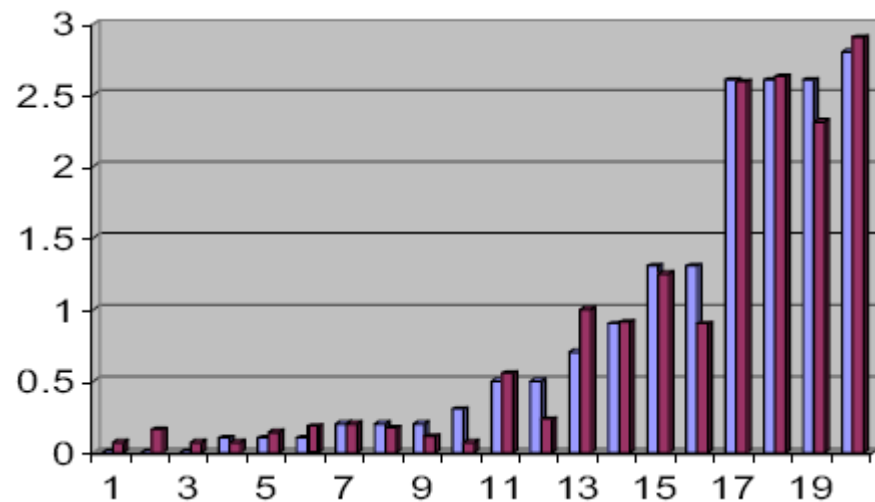


## Validation through application (2)

- Ordinal measure is important because the ordering of the molecules is typically more robust to experimental variability than pure error-based measures (enforce test reliability).
- Leave-one-out prediction to test performances of the models. Results are shown below



Molecular Hashkeys algorithm



Proposed method



## Basic improvements

- Comparison of molecule retrieval systems should be based on classical performance measures.
- Given a certain set of properties  $S_p$

Recall:

$$\frac{|S_p \text{ Molecules} \cap \text{Returned Molecules}|}{|S_p \text{ Molecules}|}$$

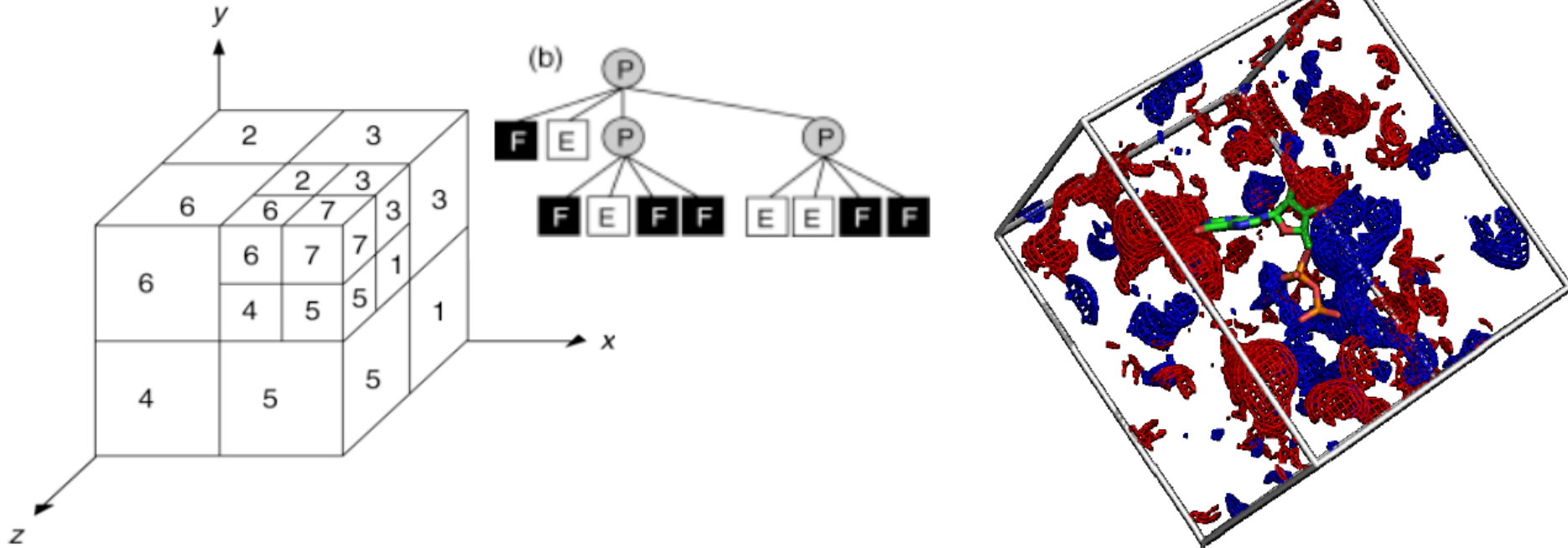
Precision:

$$\frac{|S_p \text{ Molecules} \cap \text{Returned Molecules}|}{|\text{Returned Molecules}|}$$



## Possible improvement: Octrees

- Used to recursively encode 3D spaces <sup>[41]</sup>

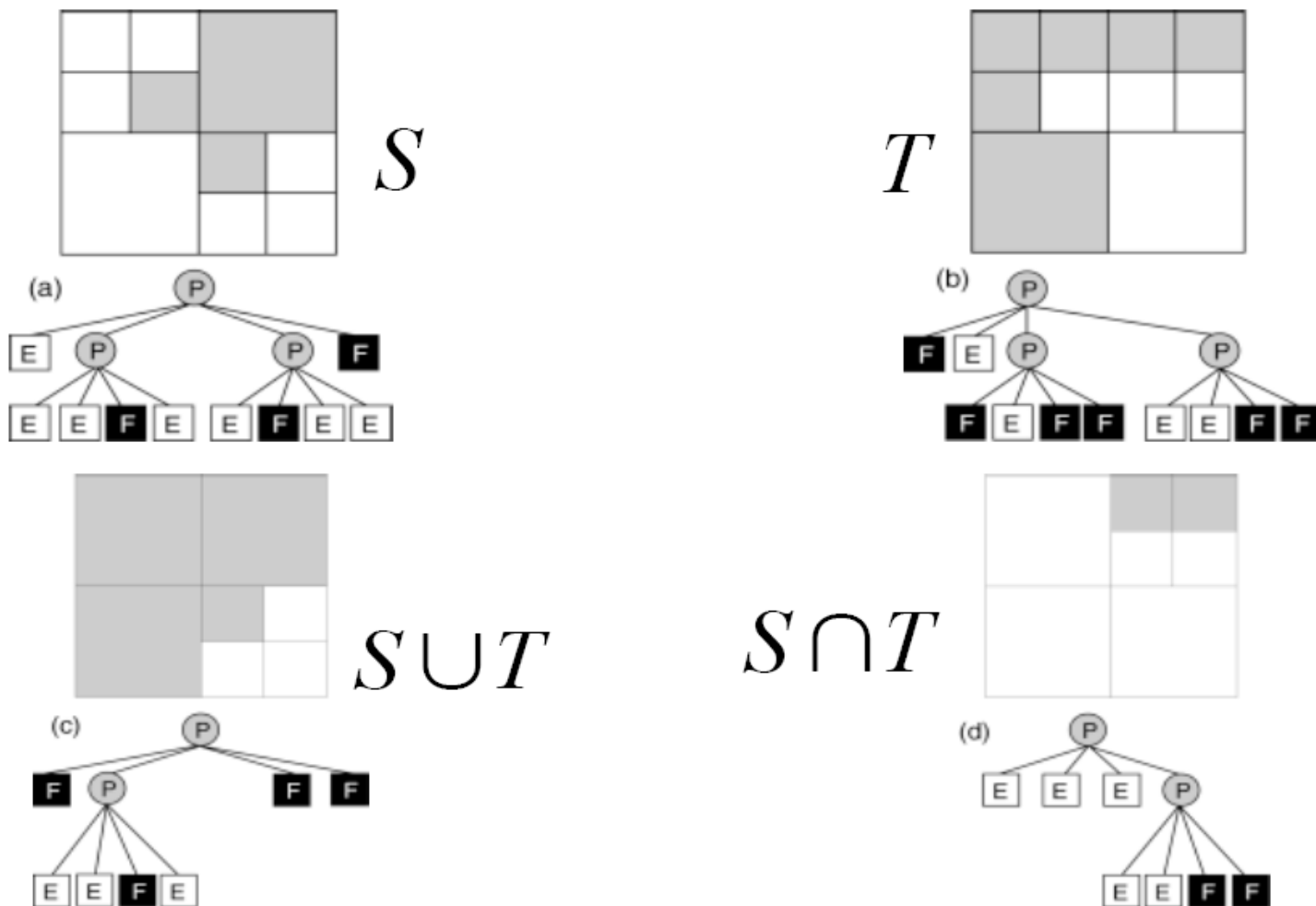


- Colored octrees: might have multiple attributes (descriptors) per node
- Object shape becomes irrelevant
- Pose and conformation independence through canonization
- Looking for a certain property in a (canonic) tree is easier than graph isomorphism (matching on canonic forms).



# Boolean Operations on octrees

- Useful to compute aggregate properties





## References

- [1] K. F. Gauss, *General investigation of curved surfaces*, Raven Press, New York, 1965
- [2] A. Ghuloum, C. Sage, A. Jain, “Molecular Hashkeys: A Novel Method for Molecular Characterization and its Application for Predicting Important Pharmaceutical Properties of Molecules”, *J. Med. Chem.*, 42, 10, pp 1739-1748, 1999
- [3] A. Jain, K. Koile, and D. Chapman, “Compass: Predicting Biological Activity from Molecular Surface Properties. Performance Comparison on a Steroid Benchmark”, *J. Med. Chem.*, 37, pp 2315-2327, 1994
- [4] H. Tsai-Hong, M. Shneier, “Rotation and translation of objects represented by octrees”. *Intl Conf. on Robotics and Automation*, 1987
- [5] D. Baum, “A point-based algorithm for multiple 3D-surface alignment of drug-sized molecules” - Chap. 3, PhD Dissertation, 2007.