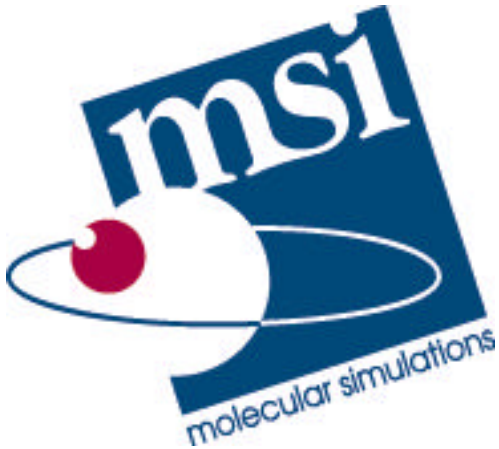


Techniques and Strategies in 3D Data Mining

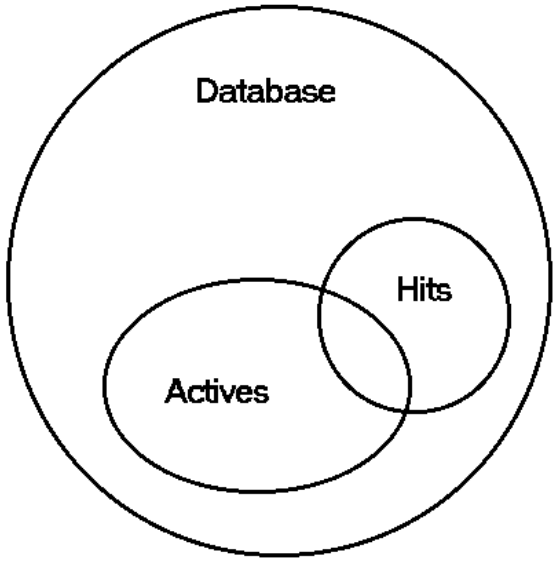
Osman F. Güner, Rémy Hofmann, and Hong Li

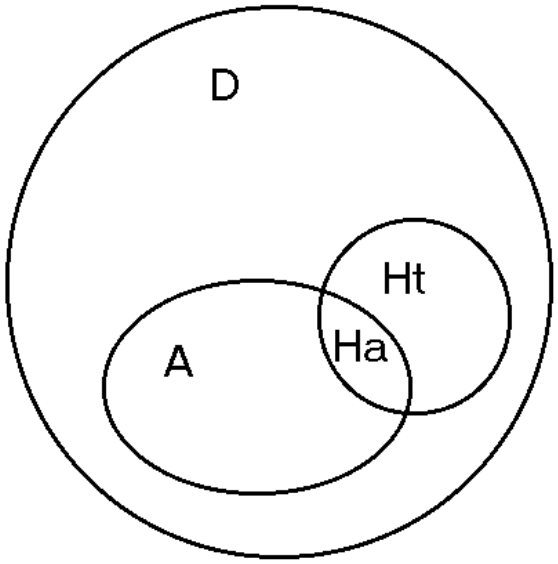
Molecular Simulations Inc, San Diego, CA

ACS Nat'l Meeting, Anaheim, March 21-25 1999

Databases, Active Compounds, and Hits

- **Schematic representation of a typical database and a hit list that contains some known active compounds**





How to Analyze Hit Lists?

- **Different metrics can be used to evaluate the quality of a hit list**
  - **Percent yield (%Y): the percentage of the known actives in the hit list**

$$\%Y = \frac{H_a}{H_t} \times 100$$

  - **Percent of actives (%A): the percentage of known active compounds retrieved from the database**

$$\%A = \frac{H_a}{A} \times 100$$

where H_t is the total number of compounds and H_a is the number of know actives in the hit list, A is the active compounds in the database.

How to Analyze Hit Lists?

- **Different metrics can be used to evaluate the quality of a hit list**
  - **Enrichment (*E*): indicates how many time more richer the hit list is than the original database with respect to the yield of actives**

$$E = \left(\frac{H_a / H_t}{A / D} \right) = \frac{H_a \times D}{H_t \times A}$$

where H_t is the total number of compounds and H_a is the number of know actives in the hit list, A is the active compounds in the database, and D is the number of compounds in the database.

How to Analyze Hit Lists?

- **Different metrics can be used to evaluate the quality of a hit list**
  - **Goodness of hit list (*GH*):[1] a weighted linear combination of *%Y* and *%A*, with correction to address database size differences**

$$GH = \left(\frac{H_a(3A + H_t)}{4H_t A} \right) \times \left(1 - \frac{H_t - H_a}{D - A} \right)$$

where H_t is the total number of compounds and H_a is the number of know actives in the hit list, *A* is the active compounds in the database, and *D* is the number of compounds in the database.

[1] Güner, O. F. and Henry, D. R. in *Pharmacophore Perception, Development, and Use in Drug Design*, **1999**, in press.

OFG

Database Domain

- **Derwent's World Drug Index (WDI) v. 96-4**
  - —**Database contains 48,405 compounds**
  - —**Number of active compounds can be determined by an text search on the fields:**
    - **Activity Keyword (PT)**
      - **38,629 compounds with data in the PT field**
      - **3,393 are listed only as TRIAL PREP. with no other activity indication**
    - **Mechanism of Action (MA)**
      - **10,318 compounds with data in the MA field**
  - —**We use the 10,318 compound subset of WDI as our search Domain (*D*) and use the "Mechanism of Action" (MA) field to identify the number of active compounds (*A*) in the database**
  - —**Unless otherwise specified, all searches are performed with the BEST algorithm**

3D Database Mining Strategies

- **Use of Query Clustering and Merging**
- **Receptor vs Ligands-based Pharmacophore Models**
- **Use of Shape vs Pharmacophore vs Merged Shape/Pharmacophore Queries**
- **The Significance of Training Set Selection: Using Similar vs. Diverse Compounds**
- **Manual vs Automated Pharmacophore Model Generation**
- **Rigid vs Flexible 3D Searching**

1. Use of Query Clustering and Merging

- **Two strategies can be applied for query merging**
  - —**to improve the ratio of known active compounds in the hit list - *increase selectivity***
  - —**to maximize the number of active compounds in the hit list - *increase coverage***
- **Use query clustering to identify similar and diverse models**
- **<u>Scenario:</u>**
- **Hypotheses generated using a topologically diverse set of 23 5-HT3 with an activity range of 0.2 to 1,400 nm. The top hypothesis has an r^2 correlation of 0.8275 with respect to predicted vs actual activities**

Clustering the Hypotheses

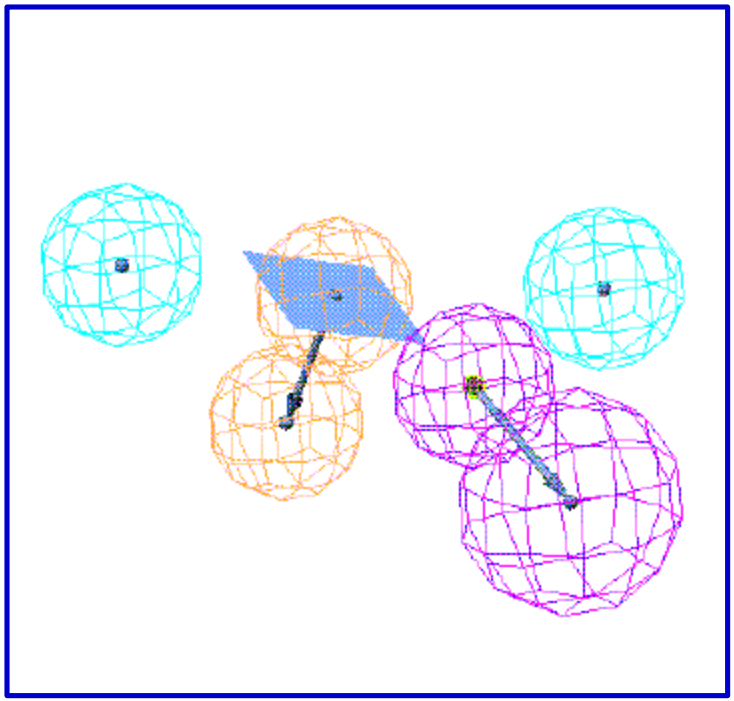
- **Cluster analysis results for the top ten 5-HT3 hypotheses**

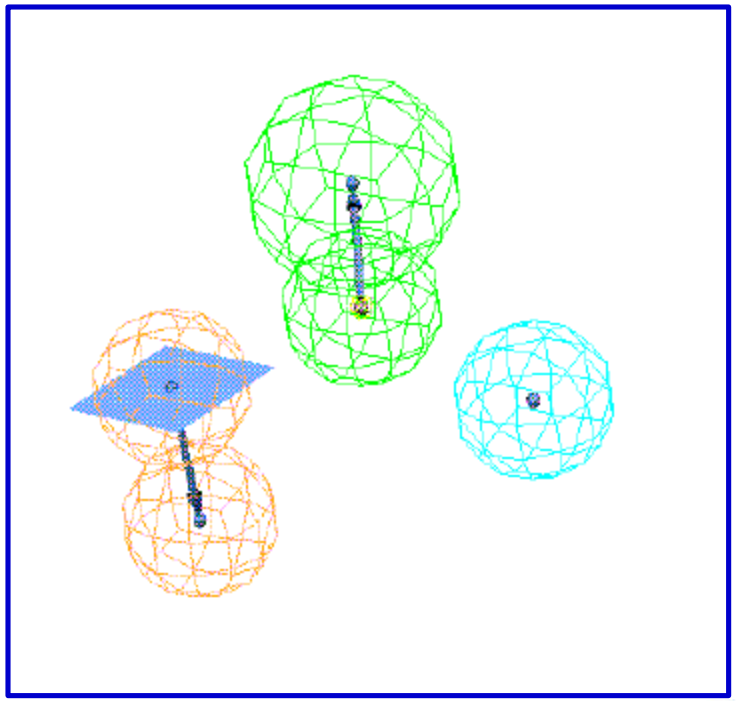
Number of Clusters	2	3	4	5	6	7	8	9
5HT3.1	**1**	1	1	1	1	1	1	1
5HT3.2	2	2	2	2	2	2	2	2
5HT3.6	2	2	2	2	2	2	2	**3**
5HT3.9	2	2	2	2	2	2	2	**3**
5HT3.5	2	2	3	3	3	3	3	4
5HT3.3	2	3	4	4	4	4	4	5
5HT3.7	2	3	4	4	4	5	5	6
5HT3.10	2	3	4	4	4	5	6	5
5HT3.4	2	3	4	5	5	6	7	8
5HT3.8	2	3	4	5	6	7	8	9

Hierarchical “average linkage” clustering

Hypotheses Considered for Merging

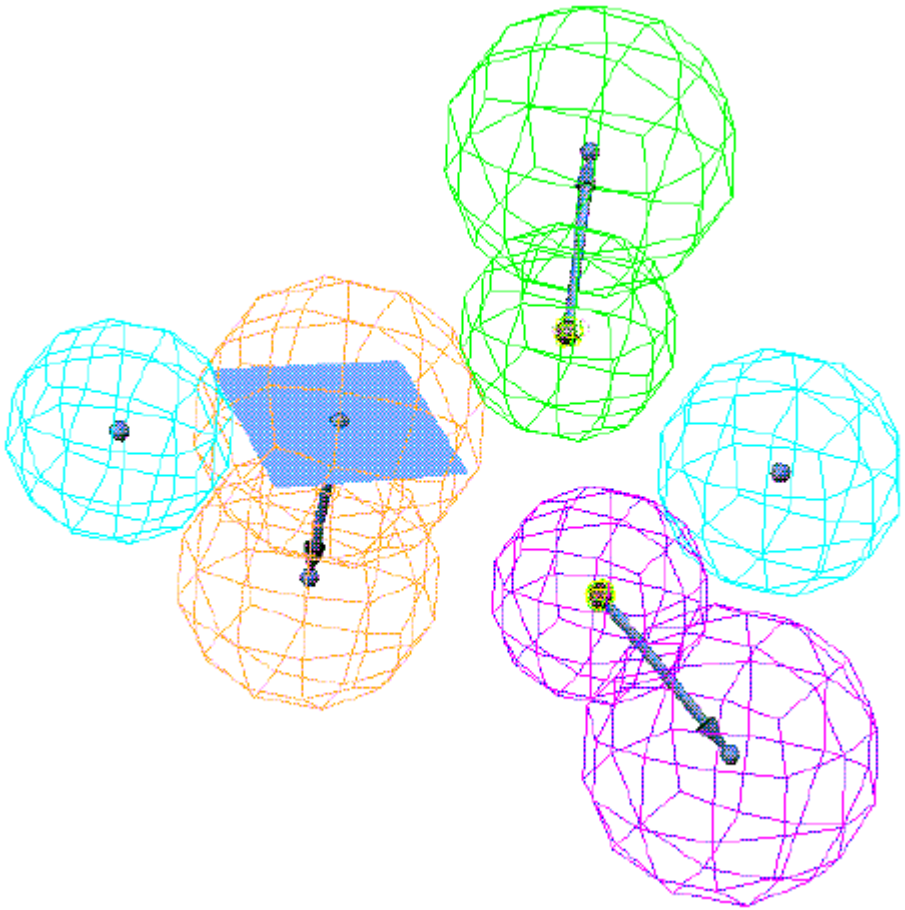
- **5HT3.1 on the left and 5HT3.5 on the right**





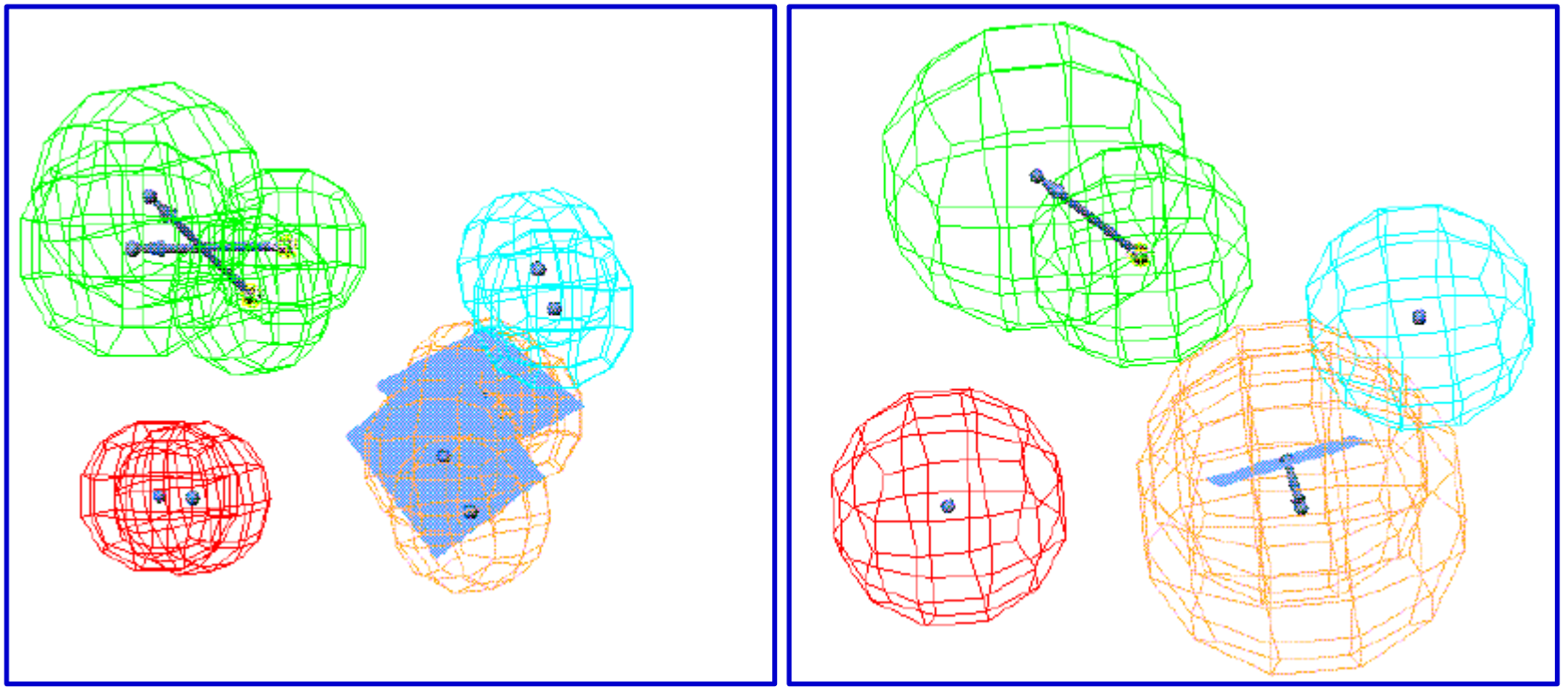
New Merged Hypothesis

- **5HT3.1 and 5HT3.5 were merged by using 1.2 Å tolerance**



Merging Similar Hypotheses

- **HT3.6 and HT3.9 aligned before merger on the left, and following merger on the right**



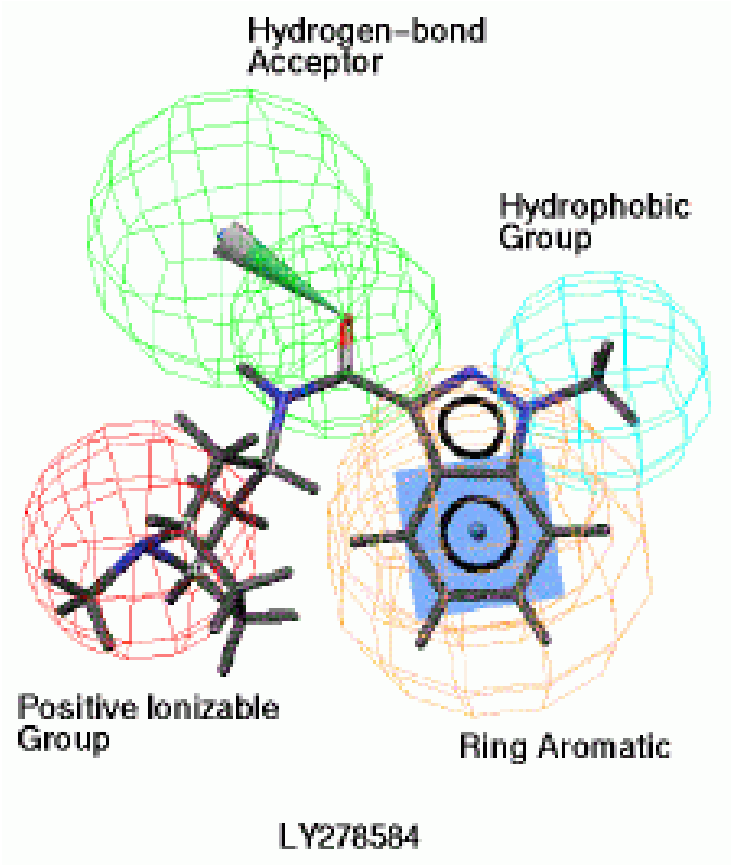
Search Results with Merged Queries

- **Comparison of the results with 5HT3.1 and the two merged hypotheses**

Query	# Actives (Ha)	# Hits (Ht)	%Y	%A	Enrichment (E)	GH score
Database	225	10,318	2.18	100.0	1.0	0
HT3.1 BEST	64	1,889	3.39	28.4	1.6	0.079
Merged (1&5)	53	1,667	3.18	23.6	1.5	0.070
Merged (6&9)	174	3,772	4.61	77.3	2.1	0.147

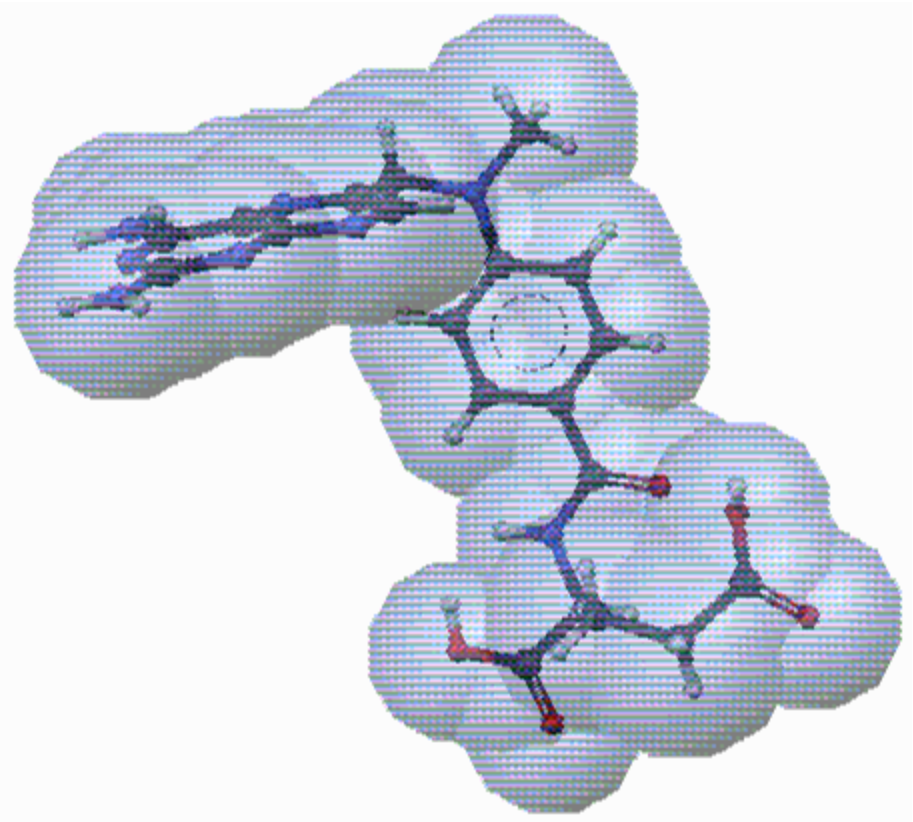
Merged Queries Conclusions

- 5HT3.1 represents the activities of the training set well (r^2=0.8275 vs for the merged[6&9] query, r^2=0.6088), but not necessarily accommodate the diverse sets of active compounds in the entire database
- The merged[6&9] query has a much better coverage but also surprisingly retrieved a list with improved selectivity as well
- On the right, a known 5-HT3 antagonists patented by Eli Lilly is displayed. Note how well the features of the compounds maps on the the merged[6&9] query.

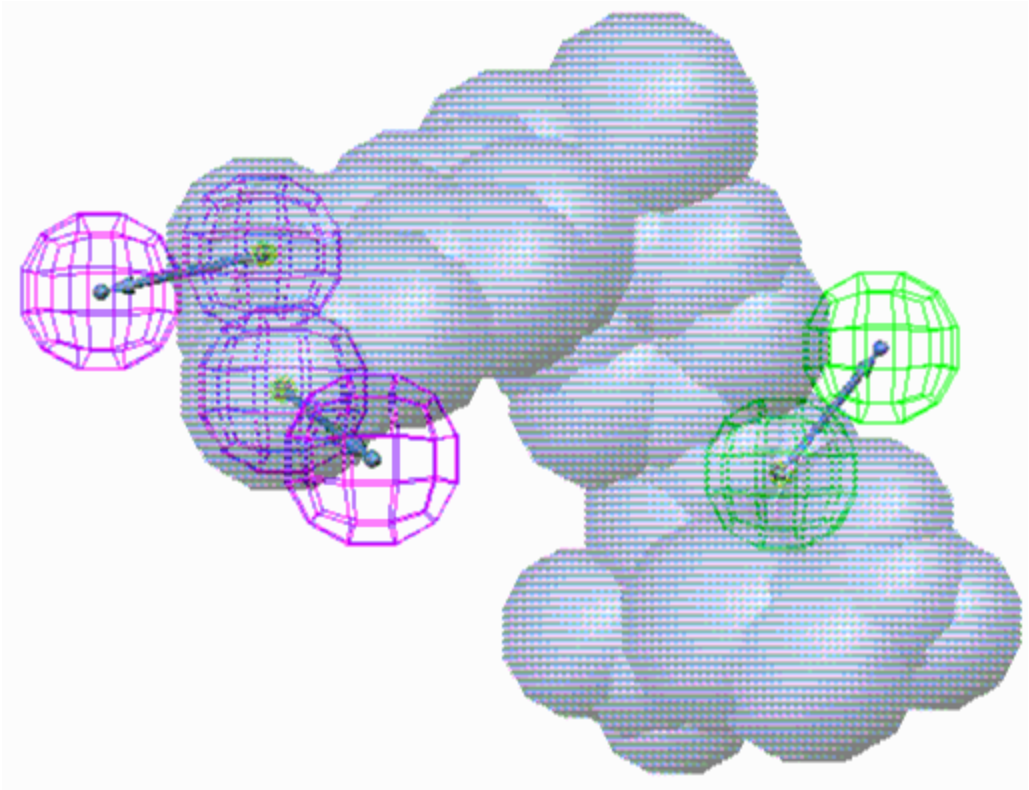


2. Shape vs Pharmacophore vs Combination

- **DHFR and methotrexate example is used to generate three queries:**
  - **—Shape query, based on the bound conformation of methotrexate**
  - **—Pharmacophore query with features representing the H-bonds, and**
  - **—Merged shape and pharmacophore query**
- **On the right is the shape query with methotrexate mapped on to it.**



Merged Shape/Pharmacophore Query

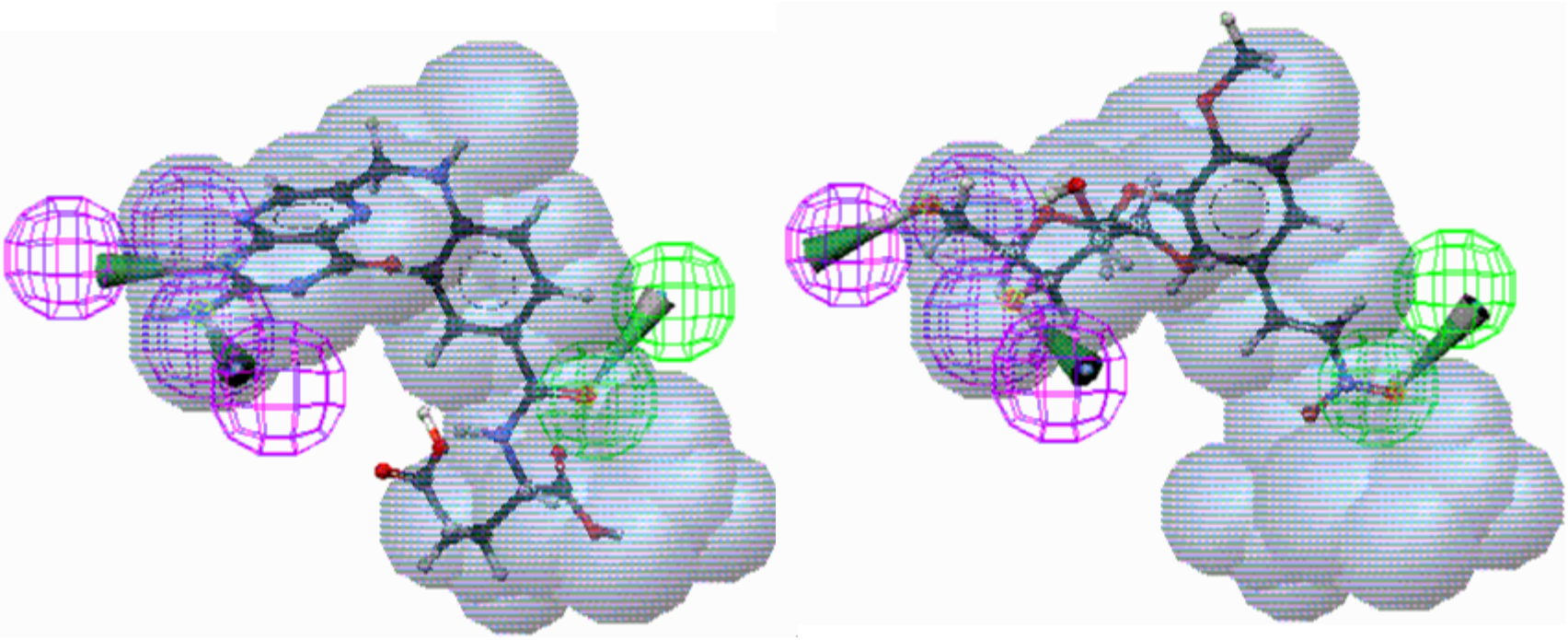


Shape vs Pharmacophore vs Merged Query Results

- **Comparison of results with shape, Pharmacophore, and merged query**

Query	# Actives (Ha)	# Hits (Ht)	%Y	%A	Enrichment (E)	GH score
Database	80	10,318	0.78	100.0	1.0	0
Shape	13	2,244	0.58	16.3	0.8	0.035
Pharmacophore	23	1,144	2.01	**28.8**	2.6	0.077
Merged	4	20	**20.00**	5.0	**25.8**	**0.163**

Hits and Leads

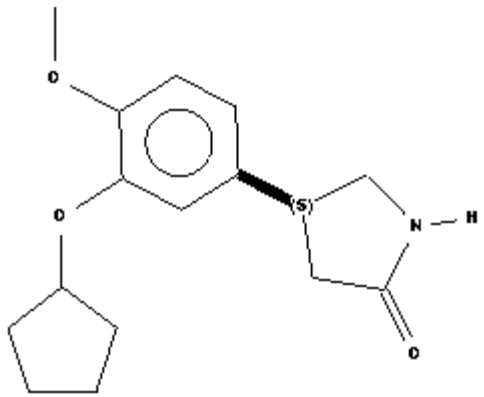


On the left, folate retrieved as a false positive since it was not listed as folate antagonists

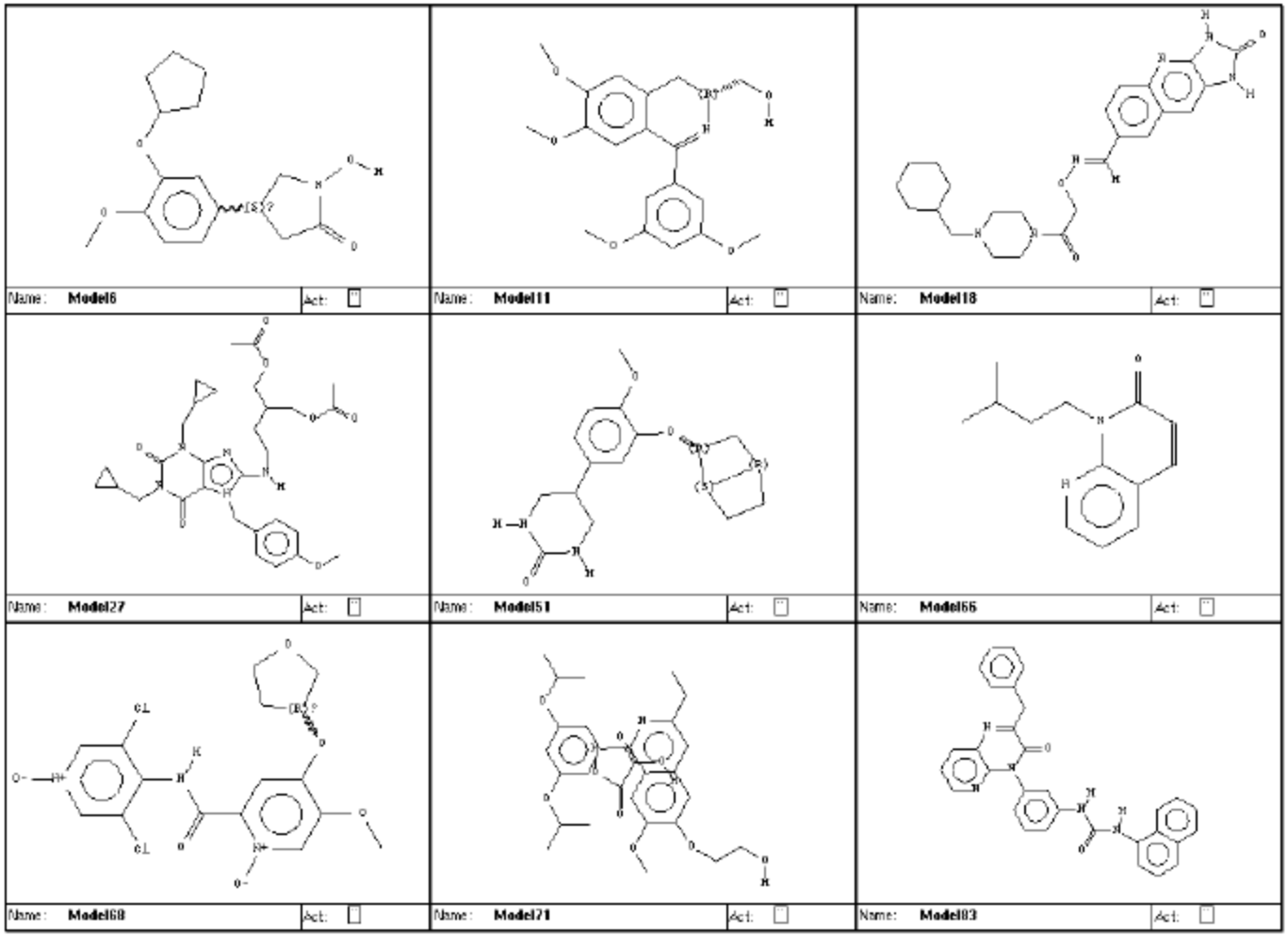
On the right is a commercially available chemical retrieved from ACD

3. Significance of Training Set Selection

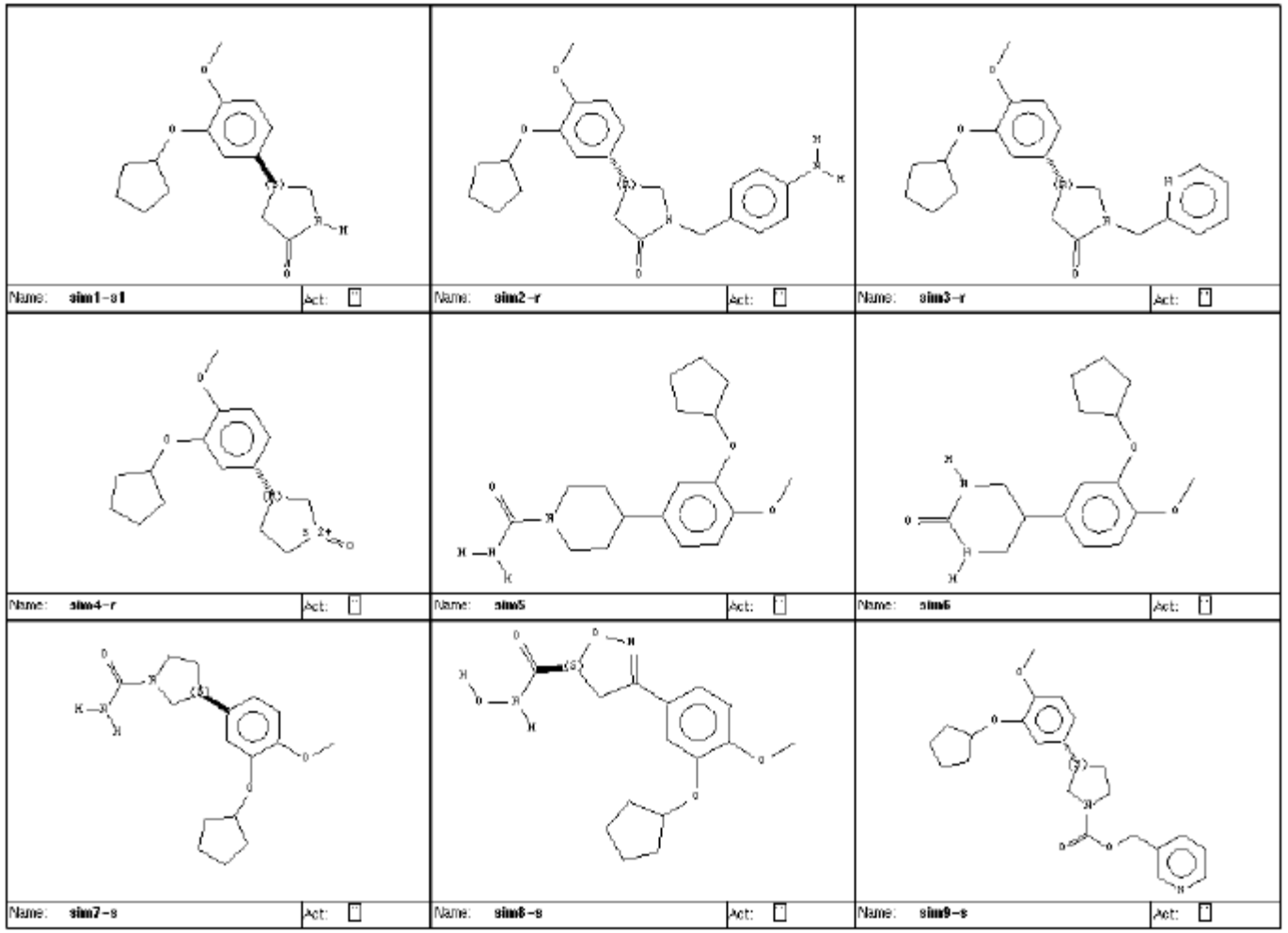
- **Diverse set of patented phosphodiesterase IV inhibitors selected based on cluster analysis of topological descriptors**
- **Similar set is selected based on the compounds most similar to rolipram**



Nine Most Diverse PDE IV Inhibitors

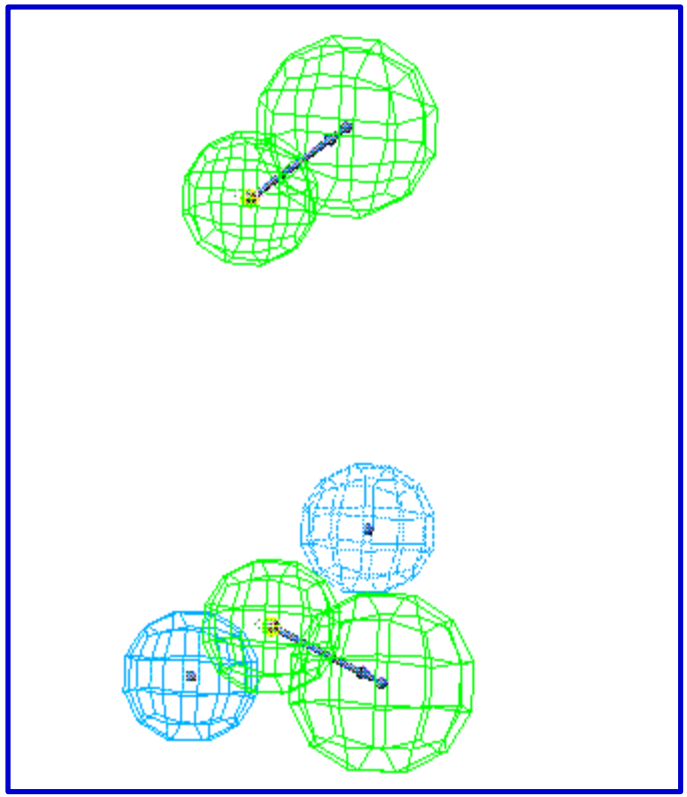


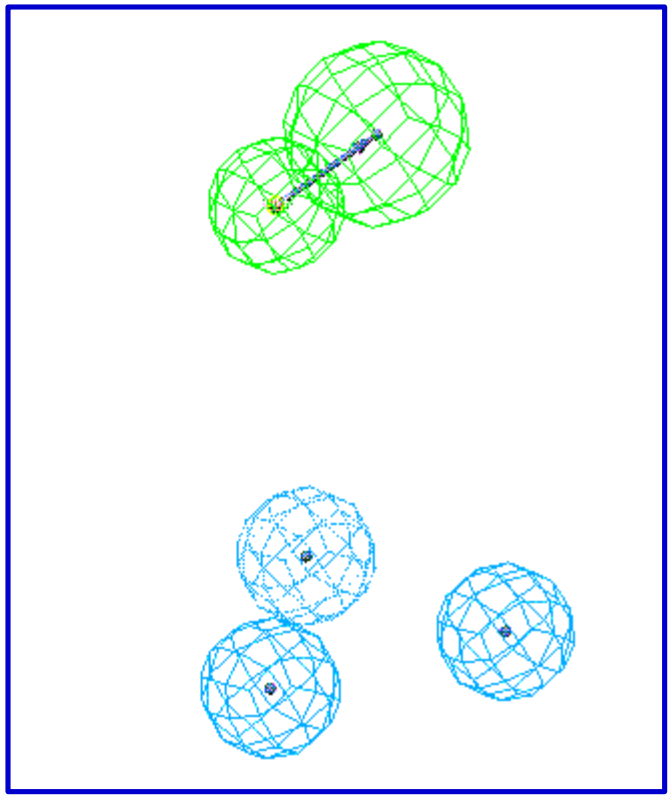
Nine Most Similar PDE IV Inhibitors



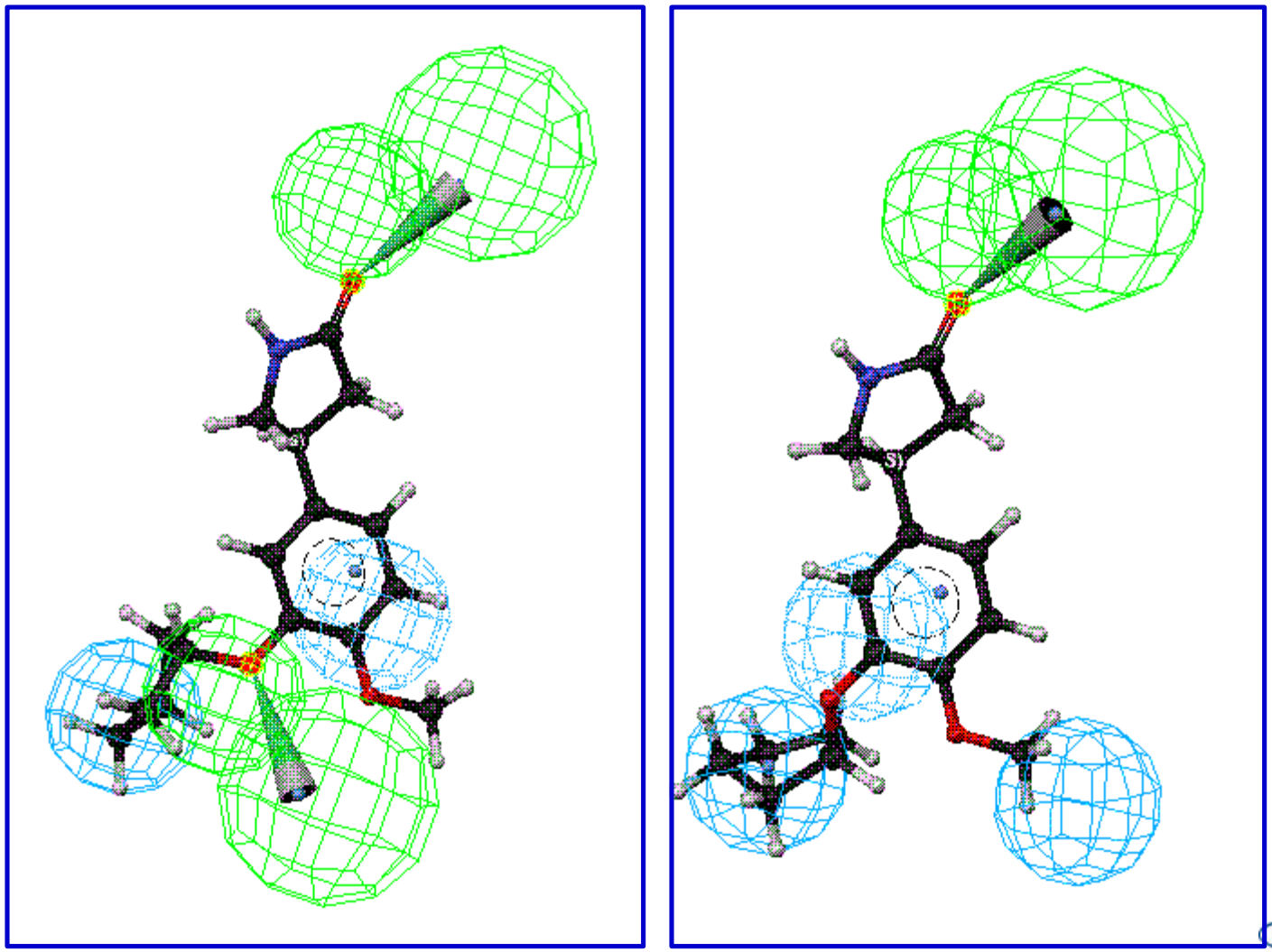
Top Scoring Hypotheses

- **On the left is the top hypothesis obtained from the diverse training set; on the right is the one from the similar set**





Rolipram Mapped to the Hypotheses

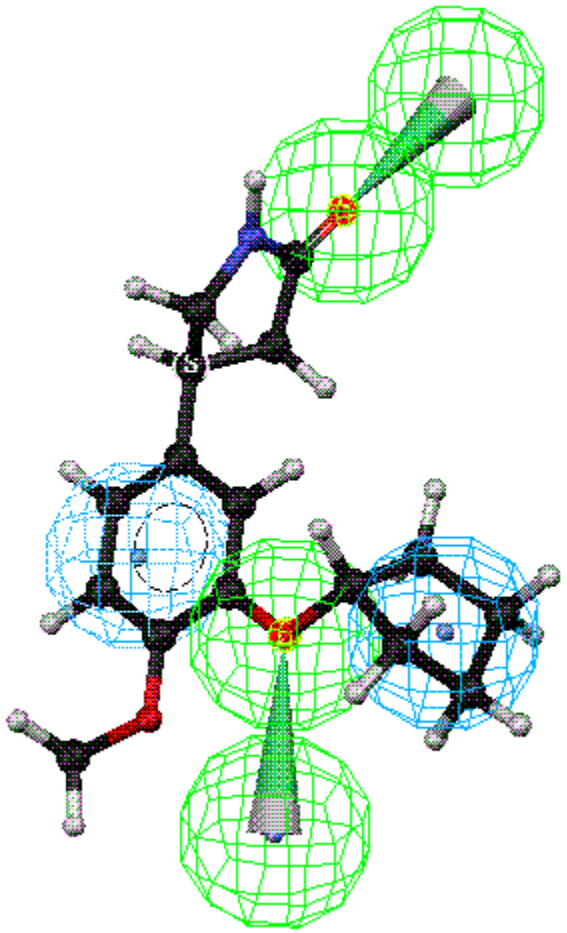


Diverse vs Similar Training Set Search Results

- **Comparison of the results obtained from searching with the top scoring hypotheses from diverse and similar training sets**

Query	# Actives (Ha)	# Hits (Ht)	%Y	%A	Enrichment (E)	GH score
Database	207	10,318	2.01	100.0	1.0	0
Diverse	73	1,589	4.59	**35.3**	2.3	**0.105**
Similar	51	986	**5.17**	24.6	**2.6**	0.091

4. Manual vs Automated Pharmacophore Model Generation



- When detailed information about the receptor-ligand interaction is available, and especially, when the bound conformation of the ligand is available, manual derivation of a pharmacophore model can be quite powerful.
- On the right is the manually generated model for PDE inhibition based on the lowest energy conformation of rolipram.

Manual vs Automated Results

- **Manually developed model is compared to the hypothesis obtained from the diverse and similar training set**

Query	# Actives (Ha)	# Hits (Ht)	%Y	%A	Enrichment (E)	GH score
Database	207	10,318	2.01	100.0	1.0	0
Diverse	73	1,589	4.59	**35.3**	2.3	**0.105**
Similar	51	986	**5.17**	24.6	**2.6**	0.091
Manual	66	1,595	4.14	31.9	2.1	0.094

5. Flexible vs Rigid Searching

- **Three places where the flexibility can be addressed:**
  - **—Flexibility in the database: (multiple conformations)**
    - Murrall, N. W.; Davies, E. K. "Conformational Freedom in 3-D Databases," *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 312-316.
  - **—Flexibility in the queries: (flexible queries)**
    - Güner, O. F.; Henry, D. R.; Pearlman, R. S. "Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures," *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 101-109.
  - **—Flexibility in the search: (on-the-fly flexible fitting)**
    - Moock, T. E.; Henry, D. R.; Ozkabak, A. G.; Alamgir, M. "Conformational Searching in ISIS/3D Databases," *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 184-189.
    - Hurst, T. "Flexible 3D Searching: the Directed Tweak Technique," *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 190-196.

Conformational Coverage in 3D Databases

- A. Smellie, S.L. Teig, and P. Towbin, "Poling: Promoting Conformational Coverage", *J. Comp. Chem.*, **1995**, *16*, 171-187.
- A. Smellie, S.D. Kahn, and S. Teig, "An Analysis of Conformational Coverage 1. Validation and Estimation of Coverage", *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 285-294.
- A. Smellie, S.D. Kahn, and S. Teig, "An Analysis of Conformational Coverage 2. Applications of Conformational Models" , *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 295-304.

How does conformational coverage approach affect the flexibility in 3D search?

Flexible vs Rigid Search Results

Fast (rigid) vs Best (flexible) search results [D=10,318, A=225 for ht3, A=80 for mtx, and A=207 for pde]

Query	# Actives (Ha)	# Hits (Ht)	%Y	%A	Enrichment (E)	GH score
Merged (6&9)-[ht3] FAST	173	3,724	4.65	76.9	2.13	0.227
Merged (6&9) [ht3] BEST	174	3,772	4.61	77.3	2.12	0.228
P HR [mtx]-FAST	21	476	4.11	26.25	5.69	0.099
PHR [mtx]-BEST	24	870	2.76	30.00	3.56	0.096
RCP1 [mtx] FAST	25	1,335	1.87	31.3	2.42	0.092
RCP1 [mtx] BEST	26	1,454	1.79	32.5	2.31	0.095
Similar [pde] FAST	48	841	5.71	23.2	2.84	0.101
Similar [pde] BEST	51	986	5.17	24.6	2.58	0.100
Manual [pde] FAST	61	1,205	5.06	29.5	2.52	0.112
Manual [pde] BEST	66	1,595	4.14	31.9	2.06	0.111

Conclusions

- **Different 3D database mining strategies are available for different problems**
  - **—Cluster and merge hypotheses, as needed, to improve selectivity or coverage, or both.**
  - **—Use the information available from receptor in enhancing your queries:**
    - **Bound conformations of ligands, if available, provides you the opportunity to develop powerful manual pharmacophore models, receptor-based, and shape-based queries.**
  - **—Diversity is an important contributor to training sets; it provides rich information that allows for development of "better" pharmacophore models**
  - **—Save time by performing FAST searches on a multi-conformation database, and reserve the more expensive BEST searches for the final part of the project.**

Acknowledgements

- **Marvin Waldman**
- **Jong-Hoon Kim**

- **Daniel McDonald**
- **Jon Sutter**
- **Bernard Chang**