NATIONAL CANCER INSTITUTE

NCI Center for Cancer Research

*Reducing the Burden of Cancer Through Exploration, Discovery and Translation*

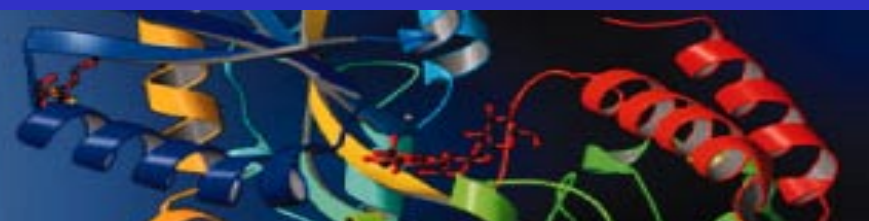
NATIONAL INSTITUTES OF HEALTH

# Computer-Assisted Mechanism-of-Action Analysis of Large Databases, Including 250,000 Open NCI Database Compounds

URL: http://cactus.nci.nih.gov/ncidb2/

National Cancer Institute at Frederick

NCI-Frederick



# Main Team Members

- **Marc C. Nicklaus**, Computer-Aided Drug Design MiniCore Facility, Laboratory of Medicinal Chemistry, Center for Cancer Research, National Cancer Institute, NIH, Frederick, MD
- **Wolf-Dietrich Ihlenfeldt** --- Computer Chemistry Center, Institute of Organic Chemistry, University of Erlangen-Nuremberg, Germany
- **Vladimir V. Poroikov, Dimitrii A. Filimonov** --- Laboratory of Structure-Function Based Drug Design, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia

# Computer-Assisted MOA Prediction for Large Databases

**GOAL:** Development of a computer program which is able to predict ***many*** kinds of biological activities for ***many*** chemical compounds on the basis of their 2D structures.

# PASS

**The computer program PASS (Prediction of Activity Spectra for Substances) is the product of ideas which originated in the framework of the National Registration System for new chemical compounds prepared in the USSR in the seventies.**

***Avidon V.V. Chem. & Pharm. J. (Rus), 1974, 8, p.22-25.***

# Applications of PASS Predictions

- Select the assays that are most relevant for a particular compound.
- Find new activities for old compounds.
- Choose compounds without undesirable side effects.
- Identify potential compounds from sample databases.
- Conduct fingerprint-type analyses.
- Use MNA descriptors for clustering analyses etc.

# Basic Elements of PASS

- Representation of biological activity.
- Description of chemical structure.
- Training set.
- Structure-activity relationships knowledge base (SAR base).
- Algorithm of the activity spectra estimation.

# Biological Activity Spectrum

- **Premise: All activities caused by the compound in biological entities are represented by the Biological Activity Spectrum of the substance.**

  ***Biological Activity Spectrum:***

  **(1) is defined as an "intrinsic“ property of the compound;**

  **(2) represents each activity qualitatively.**

  **Qualitative description of activities provides the basis to include data collected from many different sources in the training set.**

# Representation of Activities

- **Input**: Active vs. inactive; for (user-provided) additions to training set that have quantitative activity data: you have to define the threshold
- **Output**: Probabilities for activity and inactivity, respectively

# Examples of Biological Activities Included in PASS

| | |
|---|---|
| **Effects:** | **Anxiolytic** |
| | **Antineoplastic** |
| | **Vasodilator** |
| | **…** |
| **Mechanisms:** | **GABA A receptor agonist** |
| | **Topoisomerase II inhibitor** |
| | **…** |
| **Toxicity:** | **Carcinogenic** |
| | **…** |

**Currently predicted (PASS 1.51): 733 biological activities.**

# Chemical Structure Description

- Input: MOL (SD) file
- Multilevel Neighborhoods of Atoms (MNA) descriptors generated from MOL file input automatically
- MNAs constructed «de novo» if new structural feature (not yet present in training set) is found in a compound

# MNA Descriptors

Multilevel Neighborhoods of Atoms (MNA) descriptors:

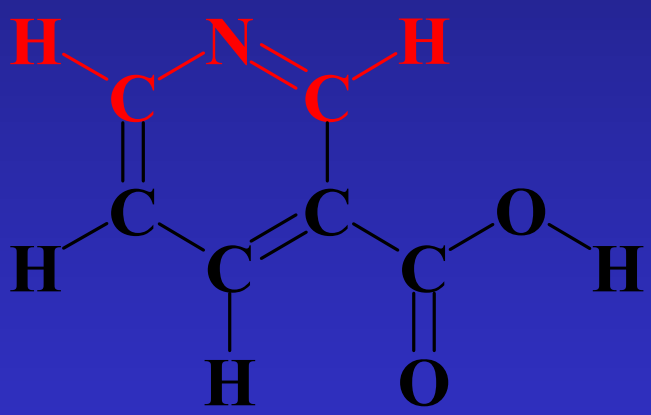
- 2D description.
- Bond types are not specified.
- All hydrogens are included according to the valencies and partial charges of atoms.

---

Filimonov D.A. Poroikov V., Borodina Yu., Gloriozova T. *J. Chem. Inf. Comput. Sci.*, 1999, 39, 666-670.

# PASS Input for Nicotinic Acid:

**MOL-file** → **MNA Descriptors**




| 1st level | 2nd level … |
|---|---|
| HC | C(C(CC-H)C(CC-C)-H(C)) |
| HO | C(C(CC-H)C(CN-H)-H(C)) |
| CHCC | C(C(CC-H)C(CN-H)-C(C-O-O)) |
| CHCN | C(C(CC-H)N(CC)-H(C)) |
| CCCC | C(C(CC-C)N(CC)-H(C)) |
| CCOO | N(C(CN-H)C(CN-H)) |
| NCC | -H(C(CC-H)) |
| OHC | -H(C(CN-H)) |
| OC | -H(-O(-H-C)) |
| | -C(C(CC-C)-O(-H-C)-O(-C)) |
| | -O(-H(-O)-C(C-O-O)) |
| | -O(-C(C-O-O)) |

# Algorithm of Activity Spectra Estimation (I)

- Was chosen from more than hundred methods tested.
- Uses LOO procedure for each compound in training set during training phase.
- During the prediction phase, a training set compound found to be equivalent* to the substance under prediction is excluded from the SARbase.

*Compounds are considered as equivalent if they have :

(1) the same molecular formula;
(2) the same set of MNA descriptors.

# Algorithm of Activity Spectra Estimation (II)

- Results of prediction:
  The probability to be active: Pa
  The probability to be inactive: Pi

  **Pa (Pi) can be considered as the probability of the first (second) kind of errors for the compound under prediction or as the probability of the compound belonging to the classes of active (inactive) compounds, respectively.**

- The algorithm is described in detail in:

  ***J. Med. Chem. 2001, 44 (4), 2432-2437***

  ***J. Chem. Inf. Comput. Sci., 2000, 40 (6), 1349-1350***

  ***Bioinformatics, 2000, 16 (8), 747-748***

  ***http://www.ibmh.msk.su/PASS***

# Example: PASS Output for Nicotinic Acid

```
21  Substructure descriptors;  0  new.
Exclude Original PASS Structure 12645 with activities:
                Antihypercholesterolemic
                HDL-cholesterol increasing
                Nucleotide metabolism regulator
                Spasmolytic
 138  Possible activities at Pa > Pi.
  Pa      Pi      for Activity:
0.855   0.007   Keratolytic
0.763   0.007   HDL-cholesterol increasing
0.754   0.007   Antihypercholesterolemic
0.715   0.005   Antilipidemic
0.615   0.012   Nootropic
0.625   0.026   Antiinflammatory, intestinal
0.568   0.007   Keratoses actinic (solar) treatment
                                    ...
```

# LOO Cross-Validation

In leave-one-out cross-validation for PASS 1.51, the average accuracy through all 43,244 compounds of the training set and 733 kinds of biological activity, was:

~ 85%

# Examples of Successfully Predicted Biological Activities

## Antihypertensive, ACE inhibitor, ECE inhibitor

Lagunin A et al. (2001). Unpublished data.

## NSAID, Local anaesthetic, Antioxidant

Geronikaki A. et. al. (1999). Quant. Struct.-Activ. Relationships, 18 (1), 16-25.

## Hepatoprotective

Flekhter O.B. et al. (1999). Bioorganic Chemistry (Rus), 26 (3), 215-223.

## Anticancer

Pogrebnyak A.V. et al. (1998). Plant Resources (Rus), 34 (1), 61-64.
Islyaikin M.K. et al. (1997). Chemical & Pharmaceutical J. (Rus), 31 (8), 19-22.

## Antibacterial

Maiboroda D.A. et al. (1998). Chemical & Pharmaceutical J. (Rus), 32 (6), 24-28.

## Cognition enhancers

Filimonov D.A. et al. (1995). Abstr. SCI Conference "Design of Bioactive Compounds", Potsdam, Germany, 26.

# PASS Prediction via Internet

## http://www.ibmh.msk.su/PASS

**Lagunin A., Stepanchikova A., Filimonov D., Poroikov V.**

**PASS: prediction of activity spectra for biologically active substances.**

***Bioinformatics*, 2000, 16 (8), 747-748.**

# The NCI Database

- Approximately half a million compounds
- Collected since 1955 by the National Cancer Institute, NIH
- Tested in anti-cancer screens; since 80's also in AIDS screens
- Managed by NCI's Developmental Therapeutics Program (DTP); see http://dtp.nci.nih.gov
- Publicly available: currently 250,251 cpds. ("Open NCI Database")
- Cancer screening data (60 cell lines) available for ca. 38,000 compounds
- AIDS screening data available for ca. 43,000 compounds
- Samples available from DTP for ~60% of compounds

# Enhanced CACTVS Browser of the Open NCI Database

- Web-based interface for searching data from the Open NCI Database by numerous criteria, including 2D and 3D structural searches
- Augmented by many additional data,
  - derived: e.g. number of rotatable bonds
  - calculated/predicted: e.g. log P; biological activities
  - systematically determined: e.g. IUPAC names
  - cross-evaluated: e.g. commercial availability
- Boolean searches possible
- Requirements: Just a Web browser, several plug-ins are optional
- Based on chemical information toolkit CACTVS (Wolf-Dietrich Ihlenfeldt, see http://www2.chemie.uni-erlangen.de/software/cactvs/)

# How To Get There….

URLs: http://cactus.nci.nih.gov/ncidb2

(U.S. mirror)

http://www2.chemie.uni-erlangen.de/ncidb2

(European mirror)

# Other Services at http://cactus.nci.nih.gov

- **About the Erlangen/Bethesda Collaboration**

## Searching the NCI Database

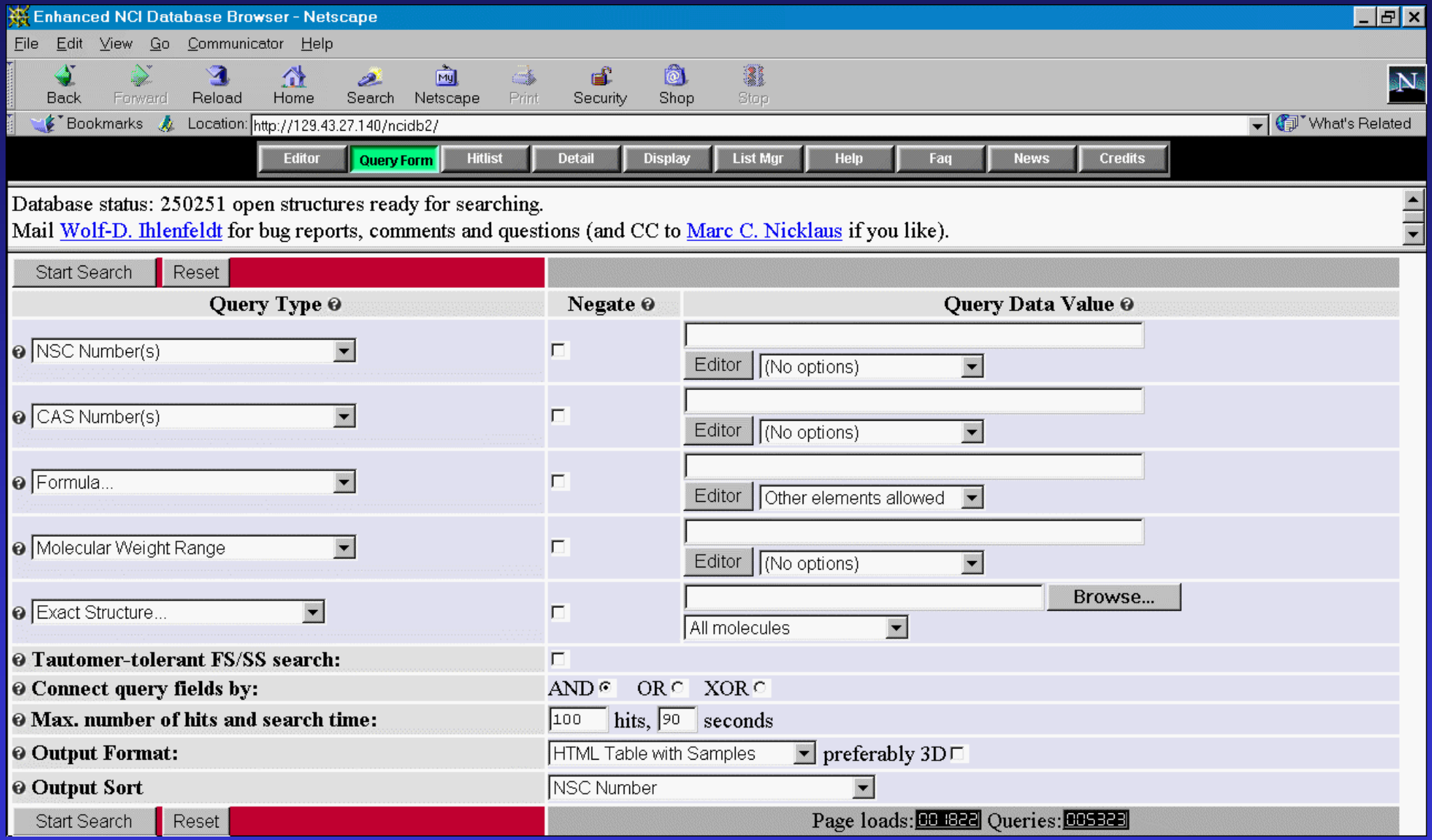
*To search and display chemical structures here, you will need Java/JavaScript to be enabled on your browser.*

- **Enhanced NCI Database Browser Release 2** (US mirror, German mirror)
  *A WWW browser to the new and enlarged collection of open NCI database compounds (>250,000 structures) with all kinds of nifty output features and links to other services for continued processing. See here for a partial list of new features. [Public beta stage -- Fully functional, full datasets, but some texts (FAQ etc.) not yet updated.]*
- **Enhanced NCI Database Browser** (old version) (US mirror, German mirror)
  *The first version of the NCI Database Browser. This link now forwards automatically to the new (U.S.) service, i.e. this machine. The old U.S. server has been turned off.*
- **Download Page**
  *Here you can download the "raw" data in bulk format that were used in building the Enhanced NCI Database Browser.*
- **Other Public Chemical Data**
  *A list of URLs that point to other public chemical (or chemistry-related) information. Currently limited to U.S. Government web sites. These sites may contain search capabilities and/or other public datasets.*

## Other Online Services

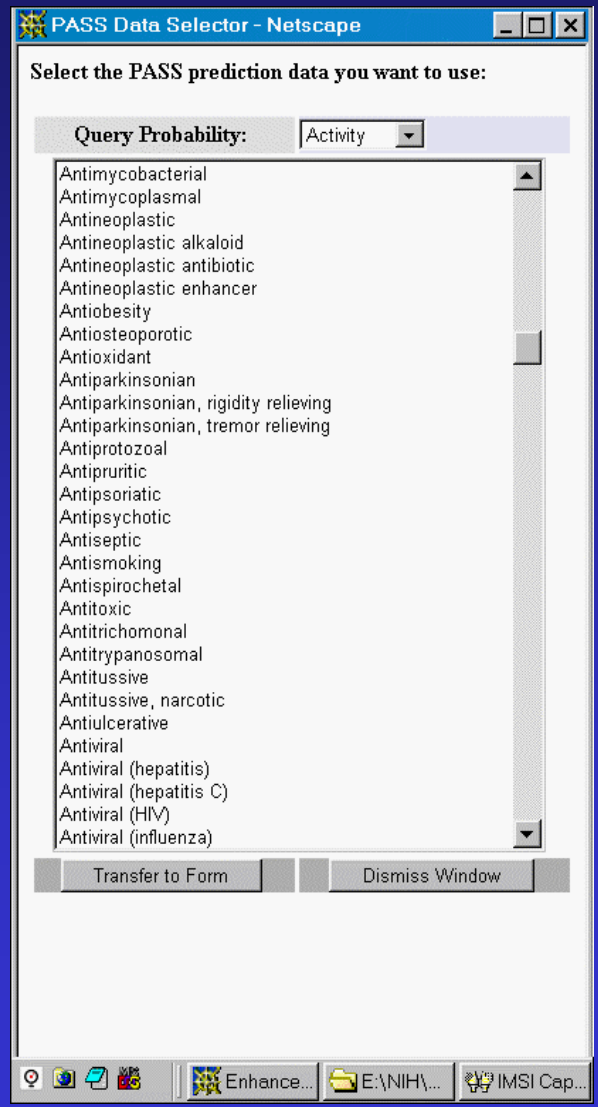
- **SDF Toolkit**
  *Toolkit (programmed in Perl 5) providing functions to read and parse structure files in MDL's SDF format, such as filter and add/remove properties, select individual records out of large SD files etc. A Win32 version is also available.*
- **VRML Creator for Chemical Structures** (US mirror, German mirror)
  *Generation of VRML scenes from your 2D or 3D data files. Automatic generation of 3D coordinates if not contained in the input structure. Many display options to control the visual appearance.*
- **GIF Creator for Chemical Structures** (US mirror, German mirror)
  *Computer-generated GIF and PNG images of chemical structures for WWW pages etc. from your 2D or 3D input files. Forms-based interface, automatic generation of 2D display coordinates for structures without them.*
- **Online SMILES Translator**
  *Web-based SMILES Translation Service. Converting of SMILES strings in USMILES, PDB and MOL file formats.*
- **Online Pseudorotation Tool**
  *Web-based pseudorotation parameter calculation tool. Calculates, and displays in tabular format, the pseudorotation parameters (P, chi, nu_max) for nucleosides, nucleotides and analogs thereof, as well as for DNA and RNA single and double strands.*
- **Self-Organized Map (SOM) of Compounds Tested in the NCI anti-HIV Screen**
  *A rectangular self-organizing map of 65 x 50 clusters, with about 42,000 AIDS-screened compounds clustered by structure similarity (CACTVS fingerprint descriptor). You can map onto this SOM compounds from the 42k AIDS set, sets of such compounds, predefined datasets (DTP's list of anti-viral screen compounds categorized by chemical class; compounds tagged in the World Drug Index with the HIV keyword, categorized by MOA classification), or even your own compound(s) (to be submitted, e.g., as SMILES string or an SDFile). Display of, and coloring by, fragments present in the clusters is possible, as well as searches, in the whole database, starting from seed compounds obtained from the SOM. Stigmata coloring will enable fragment analysis. [Pre-BETA version: Some bugs still present, no help text available.]*

# Enhanced NCI Database Browser: Query Form

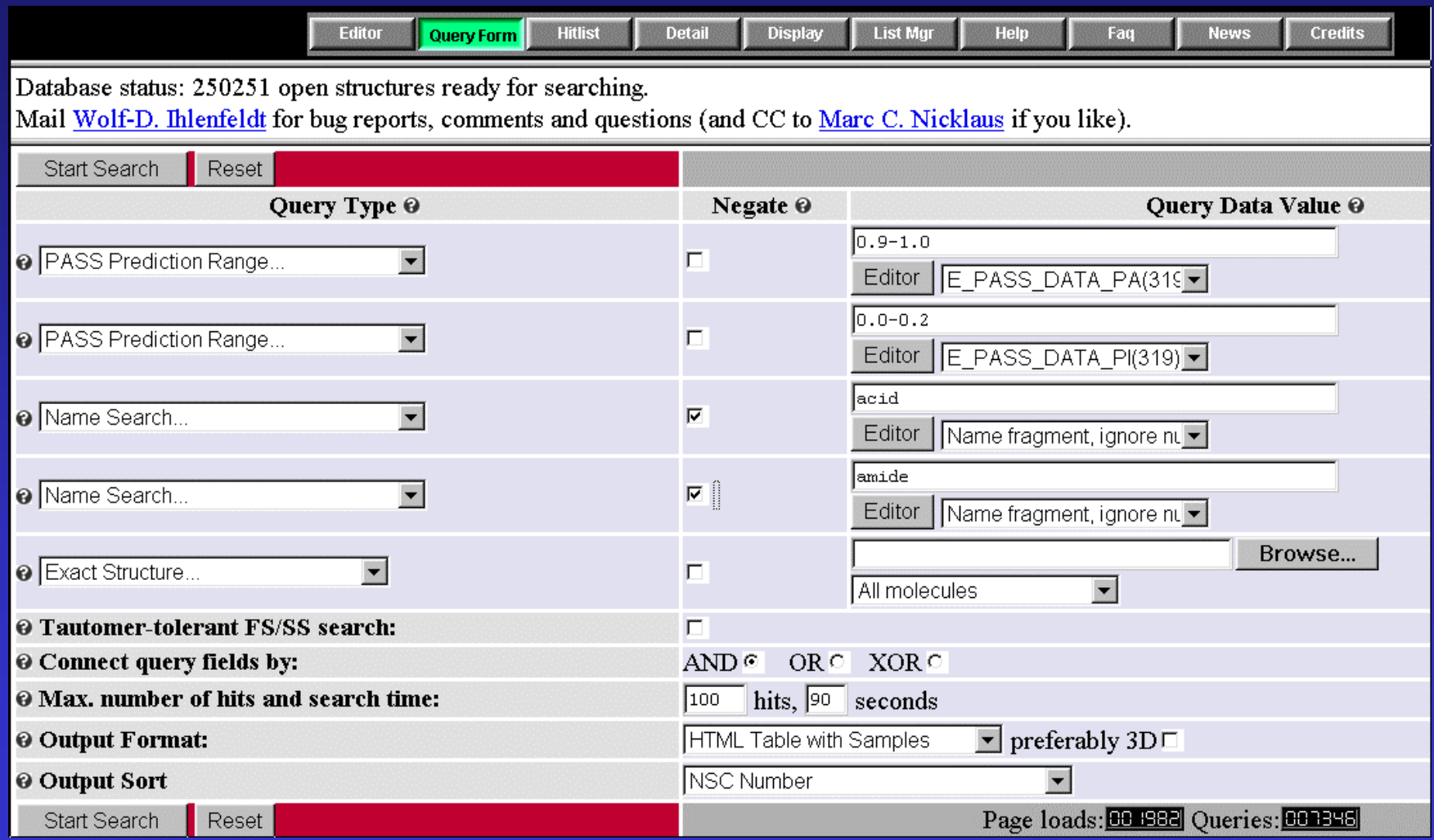


# PASS Predictions Searchable in NCI DB Browser

- More than 64 million PASS predictions included
- More than 500 activities available
- Predictions separately searchable by probabilities of activity and inactivity
- Both types combinable by logical AND
- Predictions searchable by probability ranges (in subintervals of 0.0 – 1.0)
- PASS searches combinable with any other search criteria



# Combined Search: PASS Antiangiogenesis Prediction & Name (Fragment) Exclusion



# Search Result: Hitlist

Editor | Query Form | Hitlist | Detail | Display | List Mgr | Help | Faq | News | Credits

Database status: 250251 open structures ready for searching.
Mail Wolf-D. Ihlenfeldt for bug reports, comments and questions (and CC to Marc C. Nicklaus if you like).

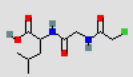
**Operations with this Dataset of 83 Structures:**

**Data Retrieval:** Format: SDFile 3D ☐ Fields: NSC Number, Molecular Weight, Name (ACD) — Retrieve

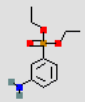
**Visualization:** GIF Image Gallery — Display

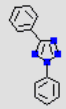
**Miscellaneous:** Restart Query (at first record) — Execute

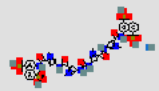
**Sample Structures**

89667 

96667 

141177 

141869 

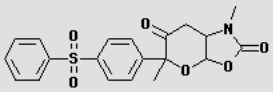
645795 

| | NSC Number | Formula | CAS | #Names | Sample Name |
|---|---|---|---|---|---|
| ☑ | 7965 | $C_3H_4ClN_5$ | 3397-62-4 | 8 | 6-chloro-1,3,5-triazine-2,4-diamine |
| ☑ | 9665 | $C_{16}H_{26}O_4$ | *(None)* | 7 | 5-methoxy-4-(2-methyl-3-(3-methyl-2-butenyl)-2-oxiranyl)-1-oxaspiro[2.5]octan-6-ol |
| ☑ | 10374 | $C_5H_8ClNO_3$ | 691-80-5 | 1 | N-(chloroacetyl)alanine |
| ☑ | 13914 | $C_5H_8ClN_5$ | 32998-04-2 | 1 | 6-chloro-$N^2$,$N^2$-dimethyl-1,3,5-triazine-2,4-diamine |
| ☑ | 32859 | $C_5H_8ClNO_3$ | 6092-47-3 | 1 | ethyl chloroacetylcarbamate |
| ☑ | 32864 | $C_6H_{11}ClN_2O_2$ | 7248-86-4 | 1 | N-(chloroacetyl)-N'-isopropylurea |
| ☑ | 33713 | $C_{10}H_{18}O_3$ | *(None)* | 1 | 2,2,5,5-tetramethyltetrahydro-3-furanyl acetate |
| ☑ | 51808 | $C_{12}H_8F_3N$ | 401-17-2 | 2 | 2,5-difluoro-N-(4-fluorophenyl)aniline |

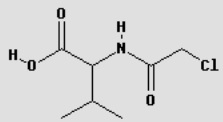
# Search Result: Image Gallery

Editor | Query Form | Hitlist | Detail | Display | List Mgr | Help | Faq | News | Credits

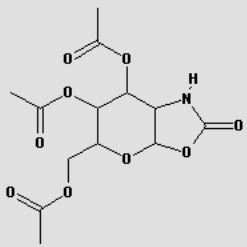
Database status: 250251 open structures ready for searching.
Mail Wolf-D. Ihlenfeldt for bug reports, comments and questions (and CC to Marc C. Nicklaus if you like).



| | | |
|---|---|---|
| **NSC:** | 373541 | Transfer to Java Editor |
| **Formula:** | $C_{20}H_{19}NO_6S$ | |
| **CAS No:** | 81830-87-7 | |
| **Anti-HIV Screening:** | Confirmed Inactive: $IC_{50}$ data; $EC_{50}$ data | |
| **Cancer Cell Screening:** | Yeast screen data | |
| **#Names:** | 1 | |
| **Sample Name:** | 1,5-dimethyl-5-(4-(phenylsulfonyl)phenyl)dihydro-3aH-pyrano[3,2-d][1,3]oxazole-2,6(1H,5H)-dione | |



| | | |
|---|---|---|
| **NSC:** | 401063 | Transfer to Java Editor |
| **Formula:** | $C_7H_{12}ClNO_3$ | |
| **CAS No:** | 4090-17-9 | |
| **Anti-HIV Screening:** | *(no data)* | |
| **Cancer Cell Screening:** | Yeast screen data | |
| **#Names:** | 1 | |
| **Sample Name:** | N-(chloroacetyl)valine | |



| | | |
|---|---|---|
| **NSC:** | 407103 | Transfer to Java Editor |
| **Formula:** | $C_{13}H_{17}NO_9$ | |
| **CAS No:** | 10163-09-4 | |
| **Anti-HIV Screening:** | *(no data)* | |
| **Cancer Cell Screening:** | Yeast screen data | |
| **#Names:** | 1 | |
| **Sample Name:** | 6-(acetyloxy)-5-((acetyloxy)methyl)-2-oxohexahydro-3aH-pyrano[3,2-d][1,3]oxazol-7-yl acetate | |

# Search Result: Detail View (top)

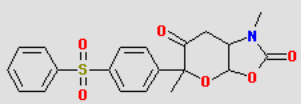
Editor | Query Form | Hitlist | Detail | Display | List Mgr | Help | Faq | News | Credits

Database status: 250251 open structures ready for searching.
Mail Wolf-D. Ihlenfeldt for bug reports, comments and questions (and CC to Marc C. Nicklaus if you like).

**Operations with this Structure (NSC 373541):**

| | | | |
|---|---|---|---|
| **Structure Retrieval:** | Format: MDL Molfile<br>3D ☐ File Name: *NSC* ○ *Name* ○ | Fields: NSC Number, Molecular Weight, Name (ACD) | Retrieve |
| **Cell Screens:** | Yeast Screen, Format: HTML | | Retrieve |
| **Visualization:** | Format: 3D Java Viewer | | Display |
| **External Services:** | Format: Cambridge Soft ChemFinder Search | | Contact |

**Structure Data:**



| | | | |
|---|---|---|---|
| **NSC Number:** | 373541 | **Date:** | 2001-06-13 14:00 |
| **File Record:** | 187645 | **CAS Number:** | 81830-87-7 |
| **Formula:** | $C_{20}H_{19}NO_6S$ | **Weight:** | 401.4332 gr/mol |
| **Complexity:** | 709.8 | **Anti-HIV Screening:** | Confimed inactive |
| **Druglikeness(std):**<br>**Druglikeness(neg):**<br>**WDI Record:** | **Is drug**<br>**Is drug**<br>No | **logP(KOW):**<br>**logP(exp):**<br>**logP(ACD):** | 1.82<br>*No data*<br>*No data* |
| **H-Bond Acceptors:**<br>**H-Bond Donors:** | 6<br>0 | **Available on DTP Plates:** | **Yes** |
| **# Rotatable Bonds: (CACTVS)** | 3 | **WLN:** | *No data* |
| **Stereochemistry Potential**<br>***R/S atoms and E/Z bonds*** | **Yes** | **Yeast Screen Level** | 1 |
| **# Catalyst Conformers:**<br>***(0 if Catalyst could not handle structure)*** | 25 | **Matched Conformer:** | *None* |
| **Composition:** | C 59.84% H 4.77% N 3.49% O 23.91% S 7.99% | | |

Transfer to Java Editor

# Search Result: Detail View (continued)

| | |
|---|---|
| **SMILES:** | CN1C2CC(=O)C(C)(OC2OC1=O)C3=CC=C(C=C3)S(=O)(=O)C4=CC=CC=C4 |
| **Name:** | 1,5-dimethyl-5-(4-(phenylsulfonyl)phenyl)dihydro-3aH-pyrano[3,2-d][1,3]oxazole-2,6(1H,5H)-dione (ACD/Name 4.0) |
| **Commercial Availability:** | *No* |
| **Commercial Database Keys:** | *None* |
| **Available Screening Data:** | Yeast, $EC_{50}$, $IC_{50}$ |
| **Anti-HIV Screening:** | **Conclusion:** Confirmed Inactive<br><br>$IC_{50}$(0.0001m) > 0.0001<br>$EC_{50}$(0.0001m) > 0.0001 |
| **Cancer Screening Summary:** | *No data ($GI_{50}$/TGI/$LC_{50}$) available.* |

**PASS Predictions:**

| Predicted Activity | p(active) | p(inactive) |
|---|---|---|
| Acetylcholine muscarinic antagonist | 0.159 | 0.141 |
| Acetylcholine nicotinic agonist | 0.218 | 0.096 |
| Acetylcholine nicotinic antagonist | 0.135 | 0.132 |
| Acetylcholine release stimulant | 0.317 | 0.191 |
| Adenosine A3 receptor antagonist | 0.267 | 0.261 |
| ADP ribosyl transferase inhibitor | 0.267 | 0.136 |
| Adrenergic transmitter uptake inhibitor | 0.261 | 0.205 |
| Alcohol dehydrogenase inhibitor | 0.248 | 0.248 |
| Aldosterone antagonist | 0.185 | 0.113 |
| Alpha 1 adrenoreceptor agonist | 0.247 | 0.161 |
| Alzheimer's disease treatment | 0.408 | 0.096 |
| Aminopeptidase microsomal inhibitor | 0.287 | 0.219 |
| Analeptic | 0.229 | 0.162 |
| Anesthetic inhalation | 0.231 | 0.089 |
| **Angiogenesis inhibitor** | 0.923 | **0.005** |
| Antiamebic | 0.259 | 0.147 |
| Antiarthritic | 0.820 | 0.007 |
| Antiasthmatic | 0.298 | 0.221 |
| Antibacterial | 0.222 | 0.069 |
| Anticoccidial | 0.283 | 0.029 |
| Antidyskinetic | 0.378 | 0.241 |
| Antiglaucomic | 0.174 | 0.171 |
| Antihypercholesterolemic | 0.310 | 0.099 |
| Antiinfective (HIV) | 0.388 | 0.086 |

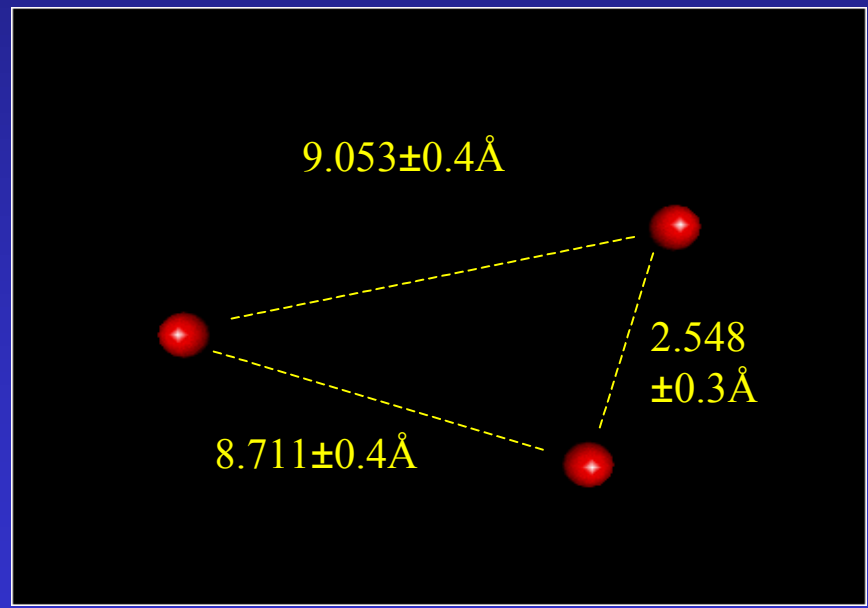
# 3D Pharmacophore Searches

- Up to 25 conformers pre-calculated by the program Catalyst (MSI) are stored for each compound.
- Searches are possible by distance constraints and other query features. Most ISIS features are implemented, such as exclusion spheres, centroids, points on lines….
- There are two ways to define a query: Query file is prepared in an external program, such as Catalyst, ISIS/Draw etc., and submitted in .mol format; or use JME Editor available within the service.

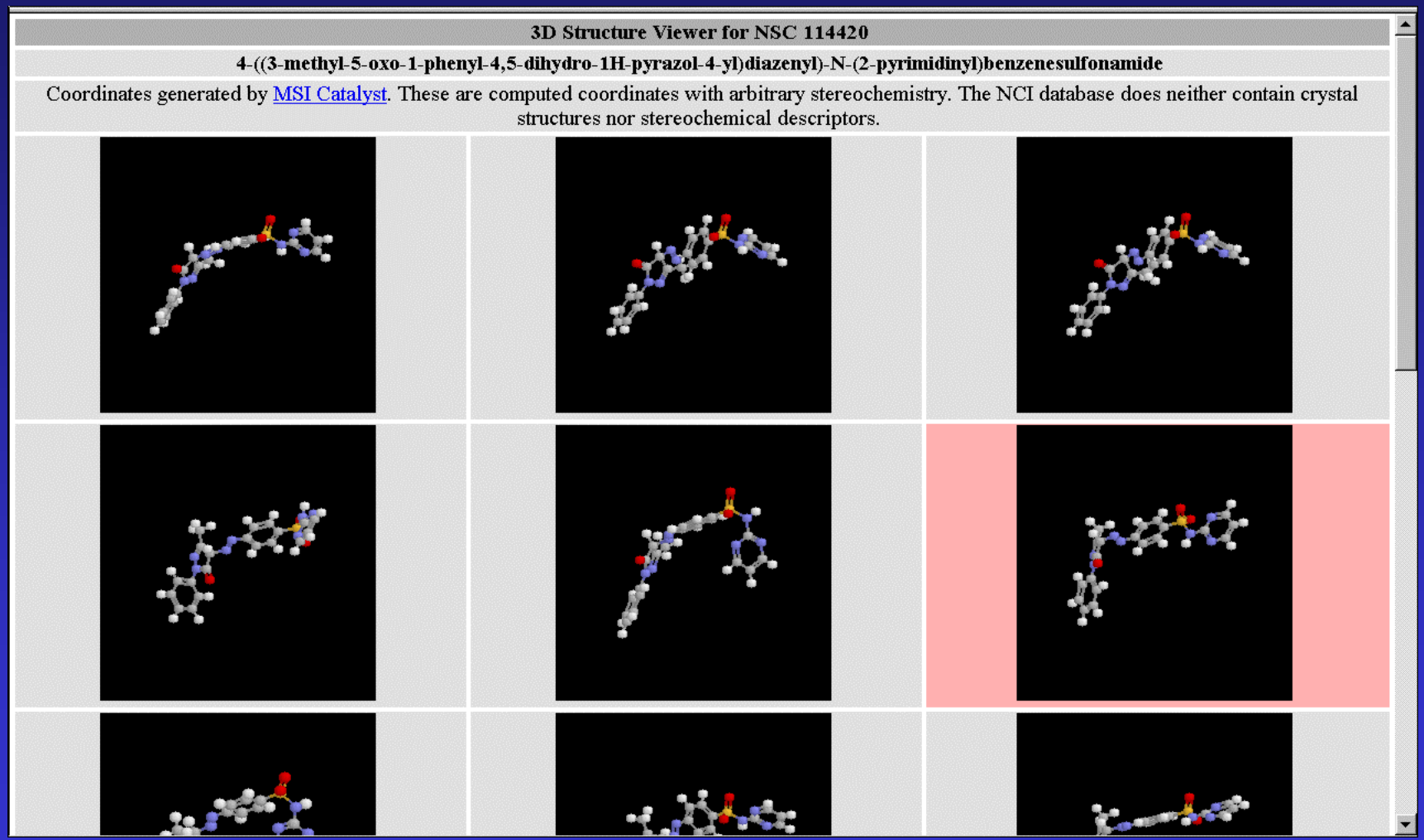
## Example:

**3-Point pharmacophore used in previous study on HIV-1 integrase inhibitor discovery.**

*J.Med.Chem.* 1997, *40*(6), 920-929.



# 3D Pharmacophore Search -- Result



# First Applications of PASS Prediction for NCI Database Compounds

- Examples so far:
  - Angiogenesis inhibitors
  - Topoisomerase I inhibitors
  - HIV-1 integrase inhibitors
- Assaying in progress
- Possibly others, conducted by users, which we don’t know about

# Acknowledgments


- **Alexey A. Lagunin, Yulia V. Borodina, Tatiana A. Gloriozova, Alla V. Stepanchikova** --- Laboratory of Structure-Function Based Drug Design, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia
- **Dan Zaharevitz** --- Developmental Therapeutics Program (DTP), NCI, NIH, Rockville, MD, U.S.
- **Frank Oellien** --- Computer Chemistry Center (CCC), Institute of Organic Chemistry, University of Erlangen-Nuremberg, Germany
- **Markus Wagener** --- NV Organon, Oss, The Netherlands
- **Peter Ertl** --- Novartis Crop Protection AG, Basel, Switzerland
- **Bruno Bienfait** --- formerly Laboratory of Medicinal Chemistry, NCI, NIH, Frederick, MD Currently ChemCodes, Inc., CCC, Erlangen, Germany
- **Johannes Voigt** --- formerly Laboratory of Medicinal Chemistry, NCI, NIH, Frederick, MD, now Schering-Plough Corp..
- **Gary Mallard** --- National Institute of Standards and Technology, Gaithersburg, U.S.
- **Valeri Kulkov** --- ACDLabs, Toronto, Canada
- **Yun Tang** --- Laboratory of Medicinal Chemistry, CCR, NCI, NIH, Frederick, MD
- **CRDF** – Financial support for the PASS Project