

# **Pushing the State of the Art in Internet Chemistry Databases: The Enhanced NCI Database Browser**

W. D. Ihlenfeldt  
Computer Chemistry Center  
Univ. Erlangen-Nuremberg  
Nägelsbachstr. 25  
D-91052 Erlangen/Germany  
wdi@ccc.chemie.uni-erlangen.de

## **Abstract**

We describe the design and implementation of an WWW-based interface to the NCI anti-tumor/anti-viral screening database. This database is the largest publicly accessible structure repository on the Internet. A variety of peculiar requirements regarding the interface to this database had to be met, among them the highest possible degree of platform independence, crosslinking to other Internet-based information sources and varied export capabilities to make this database a suitable source of structures for further examination by computational methods.

## **Introduction**

The National Cancer Institute (NCI) of the National Institute of Health (NIH) in Bethesda, U.S.A., has been collecting samples of chemical compounds within its Developmental Therapeutics Program (DTP, [1]) for more than a decade. Nearly 250,000 structures have been received and a large portion of them screened for anti-tumor activity using standardized plated tumor cell cultures. Recently, an anti-viral screening program has been added to detect potential anti-viral (especially anti-AIDS) properties of these compounds. The DTP program offers free screening of the submitted samples while the contributors retain commercial rights for interesting compounds. However, the results for inactive structures or compounds which for other reasons do not possess commercial potential are published after a time period of about two years.

Researchers who prove an authentic interest can order samples of the stored submitted compounds for further experimentation. The registration of the compounds and their associated screening results was begun with a custom in-house data repository system. Recently, the tabular numeric data has been transferred to an Oracle database. However, there was no convenient method of access to the accumulated data outside the NIH, where a simple Telnet-based alphanumeric interface was available to the researchers. We had the opportunity to develop a state-of-the art WWW interface to provide access to this data for the world-wide scientific community.

## **The Database**

At the time of writing, the database contains 246,182 structures and NSC registry numbers, about 216,000 names or WLN identifiers, 32,000 anti-viral screening table rows, 80,000 tumor cell line screening table rows and about 120,000 CAS numbers. The size of this data set is certainly smaller than the large literature-oriented databases such as CAS or Beilstein, but also notably larger than typical public data collections such as compound manufacturer catalogs. A database of this size requires a reasonable level of efficiency in the search procedures in order

to provide interactive response within a few seconds, including the time to format and transfer the dynamically generated HTML response pages.

The database is relatively static in its content. The public version is updated only about twice a year. The data are physically stored as a CACTVS system streamable scan file. [2] Because of the fast start-up time of the script interpreter which facilitates the database access (about 0.2 seconds on our servers, see below), a simple CGI script interface to this database was feasible. For other typical database access scenarios, a simple CGI implementation is often not sufficient because standard CGI starts a new access program for every retrieval request. If a database server needs several seconds or minutes to initialize, performance is unsatisfactory and more complex schemes which involve communication between a permanently running database server and a small access interface script are required. However, such a complex scenario could be avoided in our implementation.

The use of the highly portable CACTVS system notably simplified the mirroring of the database in Europe and the U.S. The European server (<http://www2.ccc.uni-erlangen.de/ncidb/>) employs a SGI Origin 200 180MHz server running IRIX6.4 and holding the data on an UW SCSI disk, while the U.S. counterpart (<http://cactvs.cit.nih.gov/ncidb/>) runs on a 400 MHz Pentium II system running Linux 2.0 and storing the data on an EIDE disk. Despite incompatible byte ordering, the same database file is usable on both sites because it employs an XDR encoding scheme with a standard byte order. The performance of both systems is about equal and certainly satisfactory for this database, guaranteeing responses within one or two seconds for typical textual, numeric or full-structure queries and less than ten seconds for 100 hits in typical substructure queries. The database contains only the minimum information necessary to allow effective queries plus all collected data that are not computable. Secondary information which can be derived from the structure data, such as 2D display coordinates or 3D atomic coordinates are only computed on demand when they are actually needed for a specific visualization task. For the computation of display coordinates an extensively modified version of the Shelley algorithm is employed [3], while 3D coordinates are generated by a loadable module which is based on the CORINA 3D coordinate generator. [4]

### **General Design Principles**

The interface design to this data was guided and limited by a number of important considerations. First, since the institute was required to make this data available to a broad audience, the use of platform-dependent technology in any feature-critical access page was out of the question. Platform-dependent plug-ins such as Chemscape Chime [5] were usable only as auxiliary tools for alternative views of results, never as core components. Likewise, the system was designed to be usable with all reasonably recent major Web browsers, regardless of provenance and platform.

In spite of these limitations, we nevertheless aimed both at sophisticated chemical search functionalities, and at complete and informative result displays. Search functionality includes standard operations such as name fragment search, full-structure search, substructure search and similarity search. Because it is the most portable method, structures are displayed as GIF images generated on the fly by default. These images contain query-specific annotations such as highlighting embedded matched substructure units and therefore cannot be precomputed. Alternatively, Chime and other platform-dependent plug-ins can be utilized as display facilities if the user explicitly requests this.

In order to improve the usability of query results for further research, we provide extensive export facilities of the retrieved structures. We support about 20 2D and 3D file formats for seamless transfer into modelling programs and similar applications. In addition to single

selected structures, complete hitlists can be exported both as multi-record datasets and result set tables for later merging and manipulation of structure sets.

In order to make the use of this service as convenient as possible and to avoid technical obstacles, we designed the service to be usable through firewalls. We refrain from using auxiliary TCP ports except the standard HTTP port 80 and do not require name server resolution of the URLs of dynamically generated pages. Also, to avoid problems with the fair number of users who have configured their browser to reject HTTP Cookies [6], we implemented all features without relying on this functionality. All necessary state information is transported in hidden form fields or the parameter field of automatically generated CGI links within the response pages. An example are the registry IDs of the structures which comprise a hit list. The primary display of such a result set is a table. The structures can be recalled one by one for a more detailed view by initiating a quick database lookup using their NSC ID number which is encoded in the parameter part of a link associated with each structure. Currently, no memory of query results is kept at the server side.

### Navigation Model

The user interface of the database service centers around three separate browser windows. The first window is the query input page, where the user selects the query type and inputs parameters. Linked to this page is a Java molecule editor [7], which was kindly provided by Peter Ertl of Novartis AG. This editor delivers the user-drawn structure as a SMILES [8] or SMARTS string to the input form. Alternatively, SMILES may also be directly typed in or pasted from SMILES-capable external editors. In addition, other structure files such as a MDL Molfile [9] with query specifications may be loaded by means of a file selector box. This set of input methods (applet, cut&paste transfer, file upload) covers all portable input methods. Figure 1 shows the specification of a typical substructure query.

After submitting the query, three things may happen. First, the query may contain syntax errors (such as an illegal SMILES string, a CAS number with an invalid checksum, or an erroneous regular expression for name search), that were not detected by the client-side JavaScript check functions which perform an initial check before any data is transferred to the server. In this case, an error message is displayed in a popup window and the query input page remains the primary focus. The second and most likely result of the query is a response list which contains more than one result structure. Hitlists are output in a second result set window in tabular form. In the default setup, this is simply a list with some elementary naming and composition information plus links to detail pages for each of the result set structures. Optionally, this page can be enhanced by selecting other presentation forms, for example with GIF images for all structures or a Chime-enhanced table. The result of the first display option is shown in Figure 2 for the substructure query submitted in Figure 1. The Chime solution has the disadvantage that it is usable only on a very limited number of platforms. On the other hand, hundreds of GIF images strain the memory of the user's computer and may cause long transfer times. From the result set page, the full list can be exported in a variety of multi-record structure file formats, including MDL SD-file and SMILES.

From the result set page, or directly from the query page - if only a single database record is the result of a query - , a third window with detail information is opened. This window contains all available information about the selected compound, including the full table of screening results. Figure 3 demonstrates this for a structure from the result set in Figure 2. Starting from this page about two dozen auxiliary display and export formats can be requested. The structure can be stored on disk in a broad range of single-record structure exchange file formats (such as XYZ, PDB or Sybyl Molfile). Alternatively, a number of auxiliary visualization options can be selected. Most important among these are a generator for the 3D WWW geometry language

VRML [10]. The molecule selected in Figure 3 is shown as a 3D VRML model in Figure 4. In contrast to the more chemistry-specific display functions via a plug-in, these 3D depictions can be examined with standard viewers that are included in recent Web browser distributions. As a final alternative, a Java 3D viewer applet from the ChemSymphony suite [11], that uses an XYZ file as input and provides platform-independent wireframe visualization, can be selected. With respect to performance, this option is the slowest display alternative, and additionally the applet code needs some time to be transferred to the user's computer when it is invoked for the first time, but this is the most portable way of displaying 3D structure information.

Figure 5 shows the complete result set from the query in Figure 2 exported as SD-file for further processing.

### **Crosslinking**

For providing the maximum benefit from the data collection and the available search methodology, the database structures are crosslinked to a number of other Internet-based free information and computational services. Depending on the nature of these links, three main approaches can be identified.

For the direct submission of structure data from the database to input forms of typical WWW-based public computational services, we have developed a general-purpose Web page rewrite engine. This application runs on our server and downloads a form page via HTTP. The form page is allowed to contain arbitrarily nested subframes. The program performs textual substitution operations on the HTML tags of the raw HTML code of the form page. In the simplest case, this causes insertion of e.g. a SMILES string or a Molfile dump of the current database structure into a specified form field. The modified page is presented to the user so that he or she can supply additional parameters or enable further options of the particular service. Corrective <BASE> tags [12] are inserted into the source text of the form pages to keep the submission button operational. A more complicated rewrite procedure is required in case of nested frames. On these pages, the links which load the subcomponents of complex layouts need to be carefully modified so that the rewrite script is called recursively for each subcomponent. When the user presses the form submission button, it acts as if the user had manually transferred the structure information. Examples where this kind of data transfer is used in our NCI database implementation include a VRML generator service for 3D molecular models [13], a chemical structure GIF image generator service [14] and the TeleSpek spectroscopy simulation service. [15]

Sometimes the submission form has a simpler structure without user-adjustable additional parameters. Had he visited the page manually, the user would not do much more than simply inserting the data and immediately afterwards press the submission button. The complexities and overhead of the form rewrite mechanism are not needed in this case. Instead, a link or form with hidden preset elements is generated which directly contacts the service if clicked at. An example for this kind of link is the connection to the ChemFinder WWW structure database [16]. For ChemFinder queries, we supply either a CAS number, if available for the selected database record, or a SMILES full-structure query string otherwise. ChemFinder automatically identifies the type of query. We do not have access to the full content of the ChemFinder database, so the success of such a query cannot be predicted. If ChemFinder does not contain data about the selected compound, an error page will be generated. However, this behavior is unavoidable given the lack of a structure index of the ChemFinder database.

For the LiqCryst database [17], we were able to obtain such an index thanks to the courtesy of its maintainers. This database is a web-accessible information source about liquid crystal properties of many compounds. Now, every NCI database record contains information whether

a corresponding LiqCryst entry does exist, and if so, what its LiqCryst identifier is. At the presentation level, the user will simply find a clickable link which leads directly to the corresponding LiqCryst entry, if and only if a LiqCryst entry exists.

## Conclusion

We have demonstrated that it is possible to implement a sophisticated interface to a large structure database using standard WWW technology. In contrast to typical commercial solutions, we were able to provide a completely platform-independent access to the structure data and thus are able to reach a maximum audience. Another highlight of our system is the extensive crosslinking to other Internet-based information sources. Our database contains a large number of structures and offers superior query capabilities. Therefore it is a suitable starting point for the collection of data on compounds from public information repositories. Finally, because all the structures and their associated data can be exported in a variety of exchange formats, the database is a valuable resource for the building of test and reference data sets for computational chemistry studies.

## References:

- [1] The NCI screening projects are described in:  
Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., Hose, C., Langley, J., Cronise, P., Vaigro-Wolff, A., Gray-Goodrich, M., Campbell, H., Mayo, J., and Boyd, Michael. *J. Nat. Cancer Inst* **83**, 757-766 (1991)  
Weislow, O.S., R. Kiser, D.L. Fine, J.P. Bader, R.H. Shoemaker, and M.R. Boyd. *J. Nat. Cancer Inst.* **81**, 577-586 (1989)  
see also the DTP homepage at <http://dtp.nci.nih.gov>
- [2] W. D. Ihlenfeldt, *J. Chem. Inf. Comput. Sci.*, submitted (1999)
- [3] Shelley, C. A., *J. Chem. Inf. Comput. Sci.* **23**, 61-65 (1983)
- [4] J. Sadowski, J. Gasteiger, G. Klebe, *J. Chem. Inf. Comput. Sci.*, **34**, 1000-1008 (1994)
- [5] <http://www.mdli.com/support/chime/>
- [6] <http://www.cookiecentral.com/faq/>
- [7] P. Ertl, *Chimia* **52**, 673-677 (1998)
- [8] D. Weininger, *J. Chem. Inf. Comput. Sci.* **28**, 31-36 (1988)  
see also <http://www.daylight.com/dayhtml/smiles/smiles-intro.html>
- [9] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer, *J. Chem. Inf. Comput. Sci.* **32**, 244-255 (1992)
- [10] J. Hartmann, J. Wernecke, *The VRML 2.0 Handbook*, Addison-Wesley, Reading, Massachusetts, 1996  
see also <http://www.vrml.org>
- [11] A. Krassavine, *Chimia* **52**, 668-672 (1998)  
see also <http://www.chemsymphony.com>

- [12] J. Steffen, *Referenzhandbuch HTML 3.2*, Sybex, Düsseldorf, 1996  
see also <http://www.w3c.org/TR/REC-html40/>
- [13] <http://www2.ccc.uni-erlangen.de/services/vrmlcreator/>
- [14] <http://www2.ccc.uni-erlangen.de/services/gifcreator/>
- [15] P. Selzer, *Chimia* **52**, 678-682 (1998)  
see also <http://www2.ccc.uni-erlangen.de/research/ir/>
- [16] J. S. Brecher. *Chimia* **52**, 658-663 (1998)  
see also <http://chemfinder.camsoft.com>
- [17] V. Vill, *Adv. Mater.* **6**, 527 (1994)  
see also <http://liqcryst.chemie.uni-hamburg.de>

## Figures:

[1] Specification of a substructure query

The screenshot displays the 'Enhanced NCI Database Browser' interface within a Netscape browser window. The browser's address bar shows the URL: <http://www2.ccc.uni-erlangen.de/ncidb/fraxe.html>. The main content area features a search form with the following elements:

- Start Search** and **Reset** buttons.
- Query Type**: A dropdown menu set to 'Substructure Search'.
- Negate**: A checkbox that is currently unchecked.
- Query Data Value**: A text input field containing the SMILES string 'O=C1OC=CC=CC=C12'.
- Start Editor** and **Import Structure** buttons.
- CAS Number(s)**: A text input field.
- Formula/Other Elements ok**: A text input field.
- Pull Structure/All**: A text input field.
- Highlight matched SS in structures:**
- Allow multi-fragment SS overlap:**
- Suppress match of arobornds on plain single/double bonds:**
- Enforce embedding:**
- Tautom PS/SS search:**
- Connect query fields by:** AND (selected), OR, XOR.
- Max. number of hits and search time:** 100 hits, 30 sec.
- Output Format:** Text Table (selected).

At the bottom of the search form are **Start Search** and **Reset** buttons. A 'Netscape: Structure Editor' window is overlaid on the main interface, showing a chemical structure of a benzimidazole derivative (SMILES: O=C1OC=CC=CC=C12). The structure editor includes a toolbar with various drawing tools and buttons for **Transfer**, **Clear**, **Close**, and **Help**.

[2] Result set of substructure query

Netscape: NCI Database Query Image Gallery

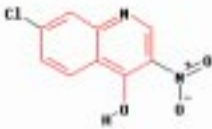
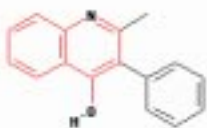
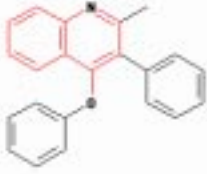
File Edit View Go Communicator Help

Back Forward Reload Home Search Print Security Stop Find

Location: <http://131.188.128.230/cgi-bin/nci.tc1>

Projects Lookup

### NCI Database Query Image Gallery

Image	Basic Data	
	NSC:	<a href="#">35</a> <input type="button" value="Transfer to Java Editor"/>
	Formula:	C <sub>9</sub> H <sub>5</sub> ClN <sub>2</sub> O <sub>3</sub>
	CAS No:	5350-50-5
	AIDS Screening:	(no data)
	Tumor Cell Screening:	(no data)
	#Names:	0
	Sample Name:	No Name
	NSC:	<a href="#">47</a> <input type="button" value="Transfer to Java Editor"/>
	Formula:	C <sub>16</sub> H <sub>13</sub> NO
	CAS No:	5350-61-8
	AIDS Screening:	(no data)
	Tumor Cell Screening:	(no data)
	#Names:	0
	Sample Name:	No Name
	NSC:	<a href="#">51</a> <input type="button" value="Transfer to Java Editor"/>
	Formula:	C <sub>22</sub> H <sub>17</sub> NO
	CAS No:	5350-65-2
	AIDS Screening:	(no data)
	Tumor Cell Screening:	(no data)
	#Names:	0
	Sample Name:	No Name

[3] Detailed view of a selected result structure

Netscape: NCI Database Query Result Table

File Edit View Go Communicator Help

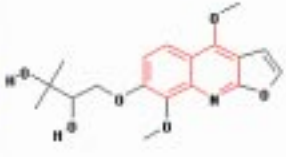

Back Forward Reload Home Search Print Security Stop Find

Bookmarks Location: <http://131.188.128.230/cgi-bin/nci.tc1?opt=nsc&data1=94653&dbfile=/usr/local/w...>

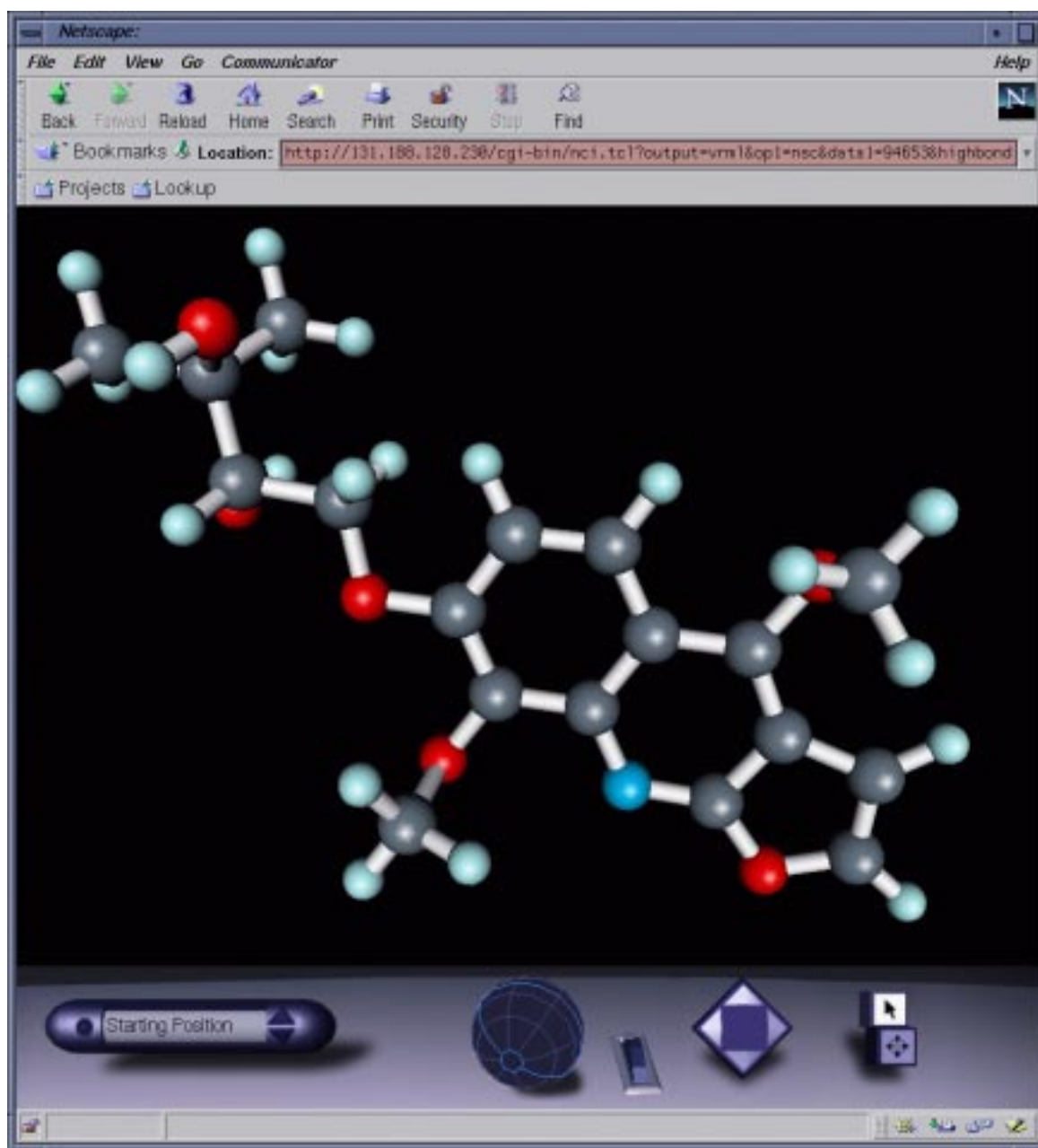
Projects Lockup

### NCI Database Query Result Table

Database Retrieval: Format:  in 3D if supported

<b>NSC Number:</b>	94653	<b>CAS Number:</b>	522-11-2	<b>Date:</b>	1999-03-30 22:08																																																																																																																																												
<b>Formula:</b>	C <sub>18</sub> H <sub>21</sub> NO <sub>4</sub>	<b>Weight:</b>	347.367 gr/mol	Record 7412																																																																																																																																													
<b>Complexity:</b>	447.3	<b>Operations with this structure:</b>																																																																																																																																															
<b>Plot:</b> <a href="#">Customize</a>		<input type="button" value="Transfer to Java Editor"/>																																																																																																																																															
		 <input type="button" value="Search ChemFinder"/>																																																																																																																																															
		Email: <input type="text" value="nciuser@somewhere.on.earth"/> <input type="button" value="Simulate IR Spectrum (~ 2 mins)"/>																																																																																																																																															
<b>Names:</b>	Evoxine EVOXINE																																																																																																																																																
<b>AIDS Screening:</b>	Confirmed inactive.																																																																																																																																																
<b>GI50 Screening Results:</b>	<table border="1"> <thead> <tr> <th>Conc</th> <th>Unit</th> <th>LoConc</th> <th>Panel</th> <th>Cell</th> <th>Panel#</th> <th>Cell#</th> <th>-logGI50</th> <th>#Tests/Line</th> <th>Max#Tests/Cp</th> </tr> </thead> <tbody> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>NCI-H23</td><td>1</td><td>1</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>NCI-H522</td><td>1</td><td>3</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>A549/ATCC</td><td>1</td><td>4</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>BRVM</td><td>1</td><td>8</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>NCI-H226</td><td>1</td><td>13</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>NCI-H322M</td><td>1</td><td>17</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>NCI-H460</td><td>1</td><td>21</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>LNB</td><td>HOP-92</td><td>1</td><td>29</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>COL</td><td>HT29</td><td>4</td><td>1</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>COL</td><td>HCC-2998</td><td>4</td><td>2</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>COL</td><td>HCT-116</td><td>4</td><td>3</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>COL</td><td>SW-620</td><td>4</td><td>9</td><td>4.000</td><td>1</td><td>1</td></tr> <tr><td>M</td><td></td><td>-4.0</td><td>COL</td><td>COLO 205</td><td>4</td><td>10</td><td>4.000</td><td>1</td><td>1</td></tr> </tbody> </table>					Conc	Unit	LoConc	Panel	Cell	Panel#	Cell#	-logGI50	#Tests/Line	Max#Tests/Cp	M		-4.0	LNB	NCI-H23	1	1	4.000	1	1	M		-4.0	LNB	NCI-H522	1	3	4.000	1	1	M		-4.0	LNB	A549/ATCC	1	4	4.000	1	1	M		-4.0	LNB	BRVM	1	8	4.000	1	1	M		-4.0	LNB	NCI-H226	1	13	4.000	1	1	M		-4.0	LNB	NCI-H322M	1	17	4.000	1	1	M		-4.0	LNB	NCI-H460	1	21	4.000	1	1	M		-4.0	LNB	HOP-92	1	29	4.000	1	1	M		-4.0	COL	HT29	4	1	4.000	1	1	M		-4.0	COL	HCC-2998	4	2	4.000	1	1	M		-4.0	COL	HCT-116	4	3	4.000	1	1	M		-4.0	COL	SW-620	4	9	4.000	1	1	M		-4.0	COL	COLO 205	4	10	4.000	1	1
Conc	Unit	LoConc	Panel	Cell	Panel#	Cell#	-logGI50	#Tests/Line	Max#Tests/Cp																																																																																																																																								
M		-4.0	LNB	NCI-H23	1	1	4.000	1	1																																																																																																																																								
M		-4.0	LNB	NCI-H522	1	3	4.000	1	1																																																																																																																																								
M		-4.0	LNB	A549/ATCC	1	4	4.000	1	1																																																																																																																																								
M		-4.0	LNB	BRVM	1	8	4.000	1	1																																																																																																																																								
M		-4.0	LNB	NCI-H226	1	13	4.000	1	1																																																																																																																																								
M		-4.0	LNB	NCI-H322M	1	17	4.000	1	1																																																																																																																																								
M		-4.0	LNB	NCI-H460	1	21	4.000	1	1																																																																																																																																								
M		-4.0	LNB	HOP-92	1	29	4.000	1	1																																																																																																																																								
M		-4.0	COL	HT29	4	1	4.000	1	1																																																																																																																																								
M		-4.0	COL	HCC-2998	4	2	4.000	1	1																																																																																																																																								
M		-4.0	COL	HCT-116	4	3	4.000	1	1																																																																																																																																								
M		-4.0	COL	SW-620	4	9	4.000	1	1																																																																																																																																								
M		-4.0	COL	COLO 205	4	10	4.000	1	1																																																																																																																																								

[4] Studying a 3D model of the molecule in Fig. 3 with a VRML viewer



[5] Export of a full result set as SD-file for further examination

