



# Excel as a Cheminformatics Interface

Bin Zhou, Shumei Jiang, Yingyao Zhou

ChemAxon 2011 US UGM

*Sep. 27, 2011*



Genomics Institute of the  
Novartis Research  
Foundation

# Excel - Another Programming Interface

---

- **Cheminformatics Application Platforms**

  - Web-based (Chemicalize.org, PubChem)

  - Standalone (Instant Jchem, SEURAT@Synaptic)

  - Excel (Helium@GSK, CeuticalSoft)

- **Excel as a development platform**

  - Pros:**

    - Large user base
    - Easy to “use”
    - Flexible (edit, format, calculate)

  - Cons:**

    - lack of out-of-box chemical intelligence, no fuzzy logic
    - creates data silos
    - performance issue on large datasets

# Considering Excel

---

- Excel is still the place where most data are captured initially.
- Excel is still the tool that most users to analyze the data.
- Excel is still the format that most users to share data with.
- Excel is not going away soon, we need to help increase user productivity within Excel. Therefore, GNF started seriously considering Excel as another application platform, but we do not want to do Excel programming ...

# GNF LDDDB Excel Add-In

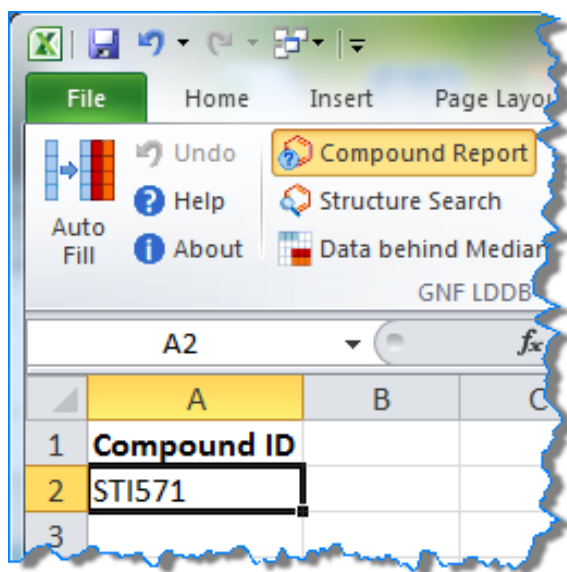
---

- Excel Add-In, when designed right, takes care of the most tedious part of the client-side programming, while enabling developers to focus on and reuse server-side features.
- **Some Applications of the GNF Excel Add-In**
  - Web application integration
  - Compound ID conversion
  - *In silico* Calculations
  - Structure Searches
  - Bioinformatics Tools

# Web Application Integration

Data Lookup: highlight compound IDs, click “Compound Report” to view all in-house data.

Idea: selected IDs are used to construct the target URL string.

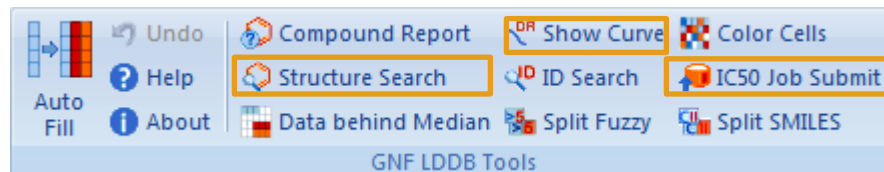


<http://LDDB.gnf.org/CpdRpt?CpdID=STI571>

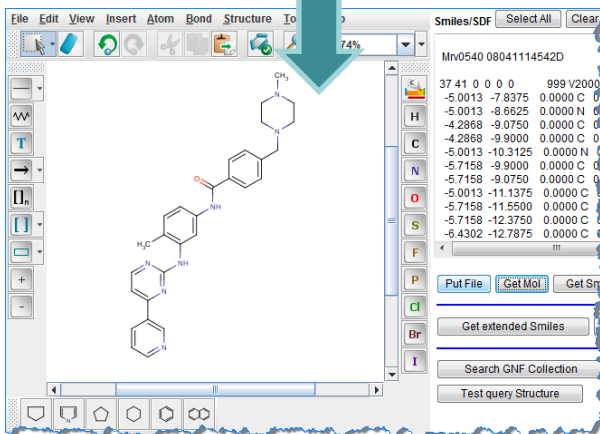
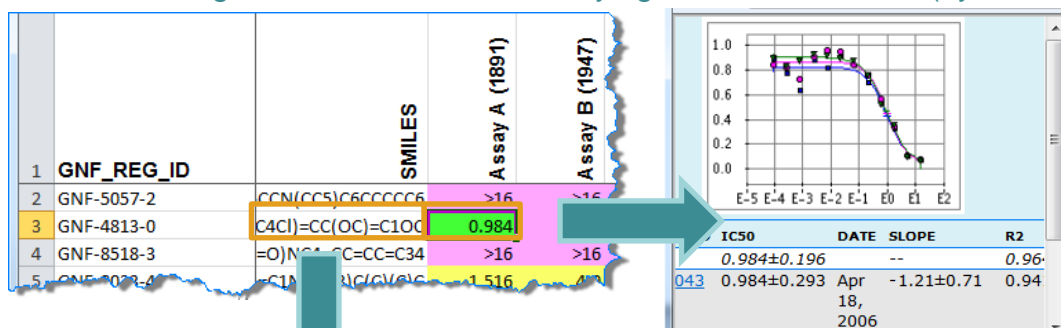
The screenshot shows the Compound Report page for STI571. The page displays the chemical structure of STI571, its molecular weight (493.603), ClogP (4.529), H-dnr (2), H-apt (8), # rot (7), PSA (86.28), and LogS(mol/l,PH7.4) (0.00012). Below the structure, there is a description: "★★★ [Imatinib(in Wikipedia),Sti-571][CAS:152459-95-5,220127-57-1][MOA:Antineoplastic Agents; Protein Kinase Inhibitors][Indication:For the treatment of newly diagnosed adult patients with Philadelphia chromosome positive chronic myeloid leukemia (CM ...)]". Below the description, there is a table of batch numbers, salt purity, project, chemist, and vendor. The table has columns for Batch Number, Salt Purity(%), Project, Chemist, and Vendor. The rows are: UV214:100, UV220:100, and UV254:100. The vendor is Novartis. To the right of the main content, there is a table of activity data for various assays. The table has columns for Assay and IC50(uM). The rows are: Non-MedChem Compounds, 0.136 %Eff=107.2, >10 %Eff=22.433, 13.95 %Eff=99.61, 0.271 FC=0.963, 11.352 FC=0.971, >10 FC=0.897, 9.284 FC=0.969, 6.124 FC=0.992, 11.391 FC=1, 17.292 FC=1, 13.702 FC=0.993, 8.735 FC=1, 10.066 FC=0.991, >10 FC=0.898, 8.477 FC=0.997, 2.657 FC=1, 10.006 FC=0.942, 11.938 FC=1, 11.581 FC=1, 8.414 FC=0.997, 10.514 FC=0.989, >10 FC=0.892, and 0.0328 FC=0.895.

# Web Application Integration

Other ideas including visualizing dose-response curves behind an average IC50 value, or automate a data upload process.



Link an average IC50 value to the underlying curves on the web (by URL construction).



Link a SMILES string to the corporate database search interface.

# Compound ID Conversion

“Auto Fill” takes a column of IDs and automatically creates additional attribute columns, e.g., compound ID to SMILES conversion.

	A	B
1	<b>GNF ID</b>	<b>IC50</b>
2	GNF-3730-1	0.093
3	GNF-7710-4	0.021
4	GNF-4992-0	
5	GNF-4183-1	0.027
6	GNF-4183-1	0.098

Highlight compound IDs

Auto Fill

Undo Compound Report Show Curve Color Cells  
Help Structure Search ID Search IC50 Job Submit  
About Data behind Median Split Fuzzy Split SMILES

GNF LDDB Tools

Favorites: Load Save Delete

Cheminformatics Bioinformatics General Tools

By ID In Database In Silico

The selected column in the excel is GNF ID

The output columns are as the follows,

Structure ID  
Novartis ID  
Vendor  
Legacy ID  
Stripped Novartis ID  
MedChem State  
Transfer Permission

SMILES  
Amount

Specify rules to convert GNF IDs into SMILES strings and amount available.

Add-In automatically add additional columns of data into the same spreadsheet.

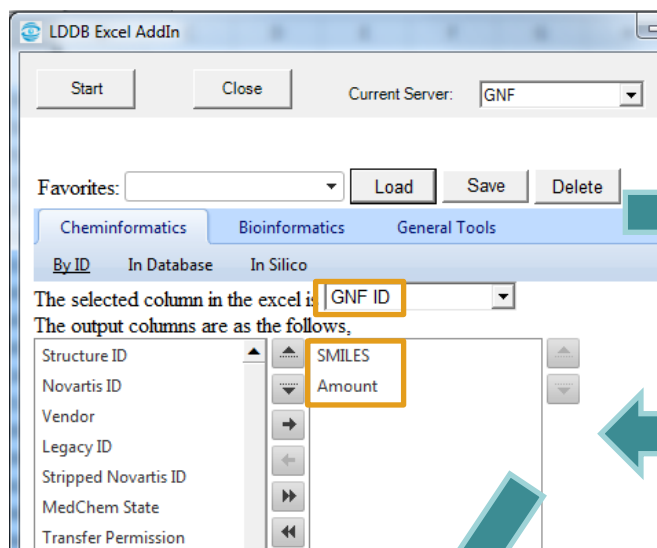
	A	B	C	D
1	<b>GNF ID</b>	<b>SMILES</b>	<b>Amount</b>	<b>IC50</b>
2	GNF-3730-1	CNC1=CC(=NC=N1)N2		0.093
3	GNF-7710-4	CNC1=CC(=NC=N1)N2		0.021
4	GNF-4992-0	CNC1=CC(=NC=N1)N2	6.08	0.039
5	GNF-4183-1	COC1=CC=C(C=C1)N2C		0.027
6	GNF-4183-1	COC1=CC=C(C=C1)N2C		0.098

# Architecture of Auto Fill

“Auto Fill” fetches data from an Excel sheet, based on the conversion rules specified, it invokes the corresponding web service to obtain results in a C# DataTable format, then modifies the original sheet with the additional attribute data. The web services are not aware of any Excel API.

Initial Sheet

	A	B
1	<b>GNF ID</b>	<b>IC50</b>
2	GNF-3730-1	0.093
3	GNF-7710-4	0.021
4	GNF-4992-0	
5	GNF-4183-1	0.027
6	GNF-4183-1	0.098



Resultant Sheet

	A	B	C	
1	<b>GNF ID</b>	<b>SMILES</b>	<b>Amount</b>	<b>IC50</b>
2	GNF-3730-1	<chem>CNC1=CC(=NC=N1)N2</chem>		0.093
3	GNF-7710-4	<chem>CNC1=CC(=NC=N1)N2</chem>		0.021
4	GNF-4992-0	<chem>CNC1=CC(=NC=N1)N2</chem>	6.08	0.039
5	GNF-4183-1	<chem>COC1=CC=C(C=C1)N2C</chem>		0.027
6	GNF-4183-1	<chem>COC1=CC=C(C=C1)N2C</chem>		0.098

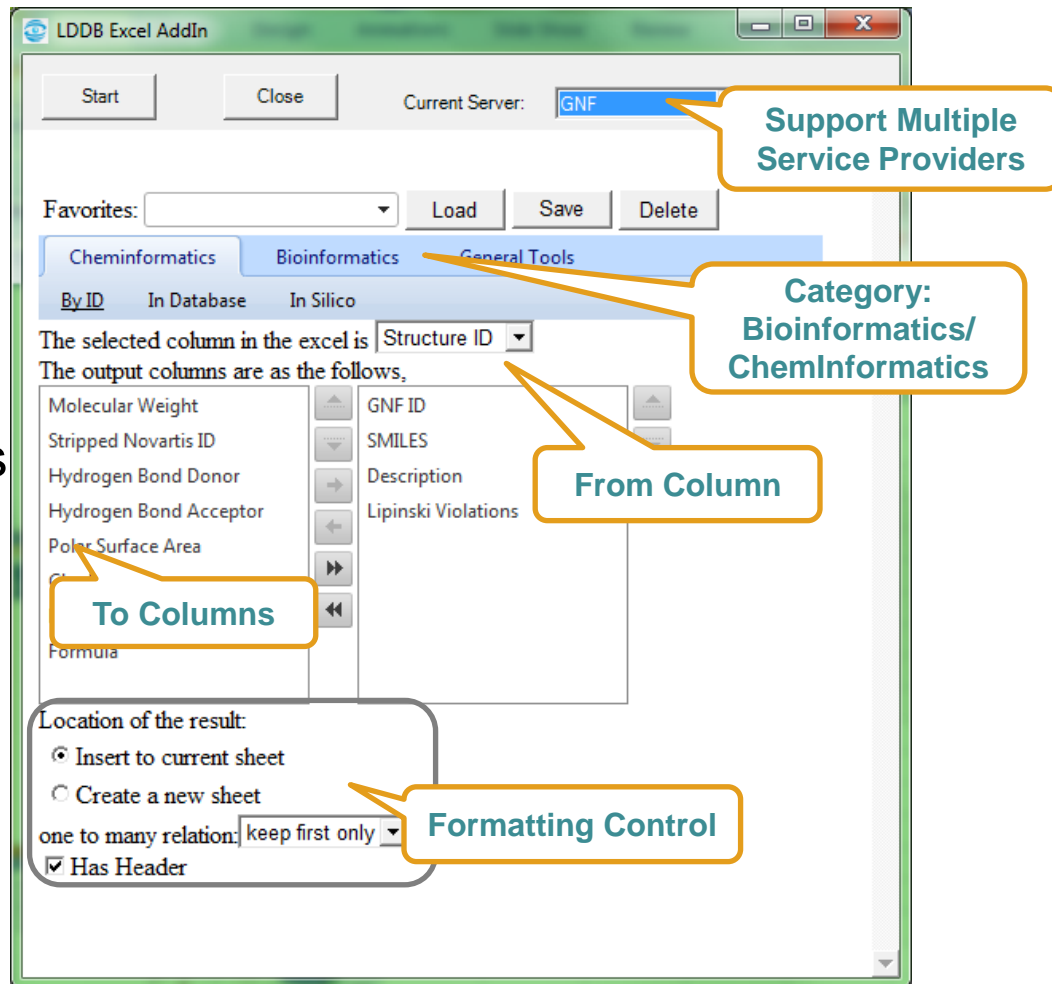


Various Web Services

# Auto Fill Web Interface

“Auto Fill” provides a generic user interface, where developers provide “plug-ins” intelligence to convert attributes from one into others. New conversions can be added without redeploying the Excel Add-In program.

- (1) Provide new web service
- (2) Provide mapping between new interface options and backend web services



# In silico Calculations

As an alternative to JChem4Excel, Add-In enables users to carry out various calculations without learning Excel formulas. In addition, one could apply the same idea to support batch vendor catalog search, compound clustering, table join, etc.

The screenshot illustrates the workflow for performing in silico calculations. On the left, an Excel spreadsheet shows a table with columns A and D. Column A contains 'GNF\_REG\_ID' and column D contains SMILES strings. The SMILES column is highlighted in yellow, with a blue arrow pointing to the 'In Silico' menu. The 'In Silico' menu is open, showing a list of calculations. The 'SMILES' option is selected, and a sub-menu is displayed with 'Molecular Weight', 'ClogP', and 'Chemical Name' highlighted in yellow. A blue arrow points from the 'ClogP' option to a second table on the right. This second table shows the results of the calculation, with columns for SMILES, Molecular Weight, ClogP, and Chemical Name. The SMILES column is highlighted in blue, and the ClogP column is highlighted in yellow.

GNF_REG_ID	SMILES	Molecular Weight	ClogP	Chemical Name
GNF-5057-2	<chem>CCN(CC5)C6CCCCC6</chem>	649.655	4.129	-4-carboxylic acid
GNF-4813-0	<chem>C4Cl)=CC(OC)=C1OC</chem>	437.879	4.007	irimidin-2-amine
GNF-8518-3	<chem>=O)NC4=CC=CC=C34</chem>	332.3129	3.266	o-1H-indol-2-one
GNF-8033-4	<chem>=C1N=CN3)C(C)(C)C</chem>	403.502	4.535	anesulfonamide

# Demo with Bioinformatics Examples

---

## Examples:

Convert gene symbols/RefSeq/ProbeSet IDs into Entrez Gene IDs

Based on Entrez Gene ID, fetch description, GO annotation, etc.

Fetch mouse orthologs for human genes

Convert data within NCBI and Ensembl systems and in-between.

Perform siRNA RSA hit picking analysis