PubChem (PC) is a publicly available collection of over 10 million unique small molecules and their biological activities. PubChem is a component of the NIH Molecular Libraries Roadmap Initiative and is implemented in the NCBI Entrez system as a set of three databases:

PC Substance, PC Compound, and PC BioAssay

**PC Substance** — contains original data from depositors. Substances may or may not represent well-characterized chemicals.

A Substance is assigned a SID and a link to the CID of the Compound that represents it (when one exists).

Each Compound is linked to the BioActivity data for all of the identical Substances that it represents.

Compounds may contain links to other data, including Protein Sequences, Toxicology information and protein Structures.

Each Compound is linked to the set of identical Substances that it represents and to pre-computed sets of structurally Related Compounds.

**PC Compound** — contains standardized versions of only those Substance records that represent well-defined chemicals.

A Compound is assigned a CID.

**PC BioAssay** — contains links to active, inactive and inconclusive Compounds and Substances, along with links to the target Protein and Gene (where applicable).

A BioAssay is assigned an AID and is linked to the original depositor.

Each Compound is linked to the BioActivity summary data.
PubChem Structure Search and Clustering

Compounds may be retrieved based on identity, similarity (by Tanimoto score), substructure, superstructure, or molecular formula. The search results can be clustered on a tree or uploaded to Entrez for further analysis.

PubChem Structure Activity Analysis

The PubChem SAR tool displays bioactivity data in a heatmap where each row represents a CID and each column an AID. The Compounds can be clustered either by structural or activity similarity, and the BioAssays by activity or protein target similarity.

PubChem BioActivity Summary

Shows all BioAssays in which the input set of Substances/Compounds have been tested

<table>
<thead>
<tr>
<th>#</th>
<th>AID</th>
<th>Active</th>
<th>Inactive</th>
<th>Total Tested</th>
<th>Data</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>372</td>
<td>9</td>
<td></td>
<td>9</td>
<td>Data</td>
<td>HIV-1 RNase H Inhibition</td>
</tr>
<tr>
<td>2</td>
<td>366</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>Data</td>
<td>Human RNase H Inhibition</td>
</tr>
<tr>
<td>3</td>
<td>367</td>
<td>7</td>
<td>2</td>
<td>9</td>
<td>Data</td>
<td>HIV-2 RNase H Inhibition</td>
</tr>
<tr>
<td>4</td>
<td>365</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>Data</td>
<td>E. coli RNase H Inhibition</td>
</tr>
</tbody>
</table>