Within only a couple of years the number of patent-extracted structures in PubChem has gone from 2 up to 14.5 million. The major sources, in order of contribution, are Thomson Pharma, IBM, SCRIPDB and SureChemOpen. The latter deposited 8.3 million in December 2012, of which 4.6 million had unique compound identifiers (CIDs). This new public patent chemistry ‘feast’ has a range of implications and utilities, some of which will be touched upon in this poster. Estimating a potential total for useful structures extractable from the entire patent corpus is a subject of conjecture. However, those related to medicinal chemistry (i.e. International Patent Classification C07D) are only in the order of 25% of the patent document families. Thus, at least the major proportion of example structures specified in drug discovery patents, including those with activity results, can now be found in PubChem. While statistical updates will be presented, currently ~30% of the 47 million CIDs now have at least one patent extraction (although some of these will be common chemistry) and, significantly, ~15% of these are from patent-only sources. A valuable consequence is that compounds with structure-activity (SAR) data extracted from medicinal chemistry papers in ChEMBL now have a high probability of either identifying or similarity intersects with structures from patents. Two sources, IBM and SCRIPDB, have patent numbers indexed in the PubChem CID records but for SureChem the substance entries link out to the free SureChemOpen application. This provides not only the location of each structure, typically as an IUPAC name, in the full-text and/or images but also a view of the entire chemistry extracted from that document. Examples will be shown where SAR tables in patents are of an order of magnitude larger than those in the eventual paper. This information expansion that patents bring to PubChem (and via the use of SureChemOpen extrinsically) is transformative considering not only that ~70% more data is thereby ‘unlocked’ but also the increased connectivity between papers, abstracts, patents and protein targets (e.g. from PubChem BioAssay) via their chemical content becomes easier to navigate and exploit. A comparative analysis of Mw distributions between the major patent sources and ChEMBL as the standard for bioactive content will also be presented. The differences give an insight into complementary for PubChem content between manual expert extraction and the automated name-to-structure conversion pipelines.

Despite estimates that ~70% of all published medicinal chemistry data are patent-only these have hitherto been a ‘Cinderella’ data source for bioactive chemical structures. However, these recent trends have revolutionised public access, 1) availability of full-text via the major patent authorities, 2) automated Chemical Named Entity Recognition, or CNER technology used here to also encompass image conversions and Complex Work Unit processing and 3) the submission of structures to PubChem by major patent-extraction sources. An example is SureChem, owned by DigitalScience. This uses a proprietary pipeline that incorporates four commercial 3rd party CNER tools to automatically recognise chemistry. Currently, 14 million unique structures have been extracted from 22 million full-text patents, made publically accessible and searchable in SureChemOpen.

The coverage (by these 14.5 million) of pharmacologically significant sources and property cuts is shown below. The 7.7 million patent structures passing the Lipinski rule-of-five plus an expanded molecular weight (Mw) filtered encompass lead-like examples, most of which are associated with bioactivity data in filings. The similarity coverage of vendor and bioactivity space of ChEMBL and BioAssay will potentially extend well beyond the exact matches intersects. Note also that coverage of drug-like space, in the last three rows, is between 65% and 95%.

**SOURCE OF PATENT**  
**PATENT COVERAGE**  
ROF <250-800 Mw (31,480,051) 7,647,104  
ChEMBL (804093) 391,081  
Chemical Vendors 722,382  
BioAssays, Active 271,188  
DrugBank (6720) 4,542  
MedSH Pharmacology (11,528) 8,067  
FDA Drugs (1216) 1,181

As an example of ‘opening up’ access to patent chemistry we can take the new antimalarial candidate MMV390048. Via the image in the press release, the lead structure can be mapped to CID 53311393. PubChem links this to WO2011086531 and we can find example structures, in the cases below converted from images of compounds 10 and 11. The SureChem conversions can be used to populate a table of structures aligned to the SAR data.

**Conclusion**

1. There has been a ‘big bang’ in our ability to interrogate and exploit the patent corpus via open sources
2. Positively impact drug discovery, pharmacology and chemical biology
3. The similarity envelope for the 14.5 million patent CIDs is pre-computed with 6.9 million of these also connect directly to non-patent sources
4. PubChem continues to enhance linkages to patents in parallel with those to targets and publications