

Dataset Overlap Density Analysis



Andreas H. Göller

Bayer Healthcare AG, Global Drug Discovery, Computational Chemistry, Wuppertal, Germany

Bayer HealthCare

Introduction

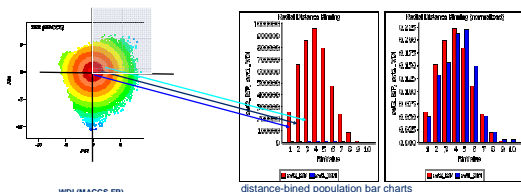
The need to compare compound datasets arises from various scenarios, like mergers and acquisitions, gap analysis campaigns, library extension programs or combinatorial library design.

Whereas it is "relatively easy" to find identical compounds in two datasets, the quantification of the overlap of two datasets is not straightforward. There are various approaches described in the literature as are

- pairwise nearest neighbor comparisons [1]
- Statistics based on clustering [2,3]
- Property space distribution bar charts of e.g. the "rule-of-five" or the "oral PhysChem Score" [3]
- BCUT binned N-dimensional spaces [4]
- ChemGPS classification by PCA projection onto a drug-like and satellite molecule reference space [5]

But how does one quantify the overlap of two datasets in one single interpretable number?

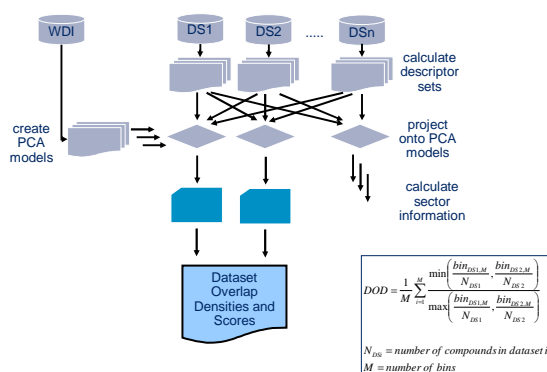
PCA map projections of dataset A onto dataset B in principle inherently contain this information on overlaps and gaps, but the visual inspection is hampered by the **crowdedness of the maps** of large sets.



Dataset Overlap Density

Dataset Overlap Density (DOD) is introduced as a score for the overlap of two datasets. A 2D descriptor PCA map of the World Drug Index (WDI) as drug-like reference space is created and other datasets are projected onto this map. PCA population areas are defined via distance to center and radial sector placement from the 2D projection for each point. Absolute and population number-normalized binned populations of each such PCA area segment for each dataset are calculated and visually compared via distance-binned bar charts.

By summing up the overlap ratios of all distance bins of the two datasets one finally comes up with a single number for the dataset overlap density in a particular descriptor space. By doing so in a set of descriptor spaces one finally creates a signature vector for dataset overlap. The approach can be adjusted for N-dimensional mappings or cube-binning schemes, and can be as fine-grained as needed for a particular application. It allows to quantify local gaps or overlaps. Proprietary datasets can be compared just by exchange of data files of the first N principal components from the original descriptor sets.



Datasets

The datasets and their identical match overlaps are given in the Table. Compounds were cleaned up by Pipeline Pilot [6]. The World Drug Index (WDI) [7] with 48129 compounds is used as a reference chemistry space for the creation of the 2D PCA maps. It is compared to the iResearch (IRS) [8] and ZINC [9] libraries as available chemical spaces and the ChEMBL [10] and Pubchem [11] databases as currently employed chemical spaces.

	WDI	IRS	ZINC	CHEMBL	PUBCHEM
WDI	48129	3	75492	11802	3588
IRS		6058461	1373472	13095	199212
ZINC			8665863	88492	244829
CHEMBL				621956	8007
PUBCHEM					404268

Descriptors

Four exemplary different descriptors sets implemented in Pipeline Pilot were applied:

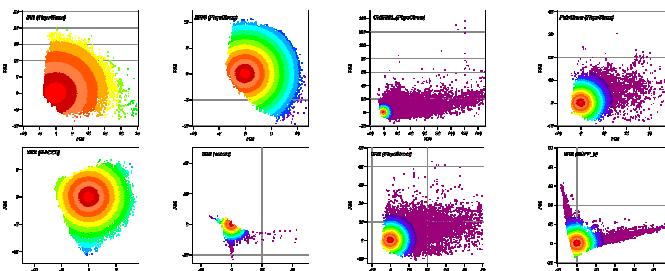
- MDL MACCS keys (166)
- Estate keys (organics)
- ECFP-6
- Physicochemical descriptors (Lipinski parameters: ALogP, PSA, MW, Number of H-Acceptors, Number of H-Donors)

Conclusions

- The overlap of two large chemical spaces in any descriptor space can be expressed as a single number
- The approach is generic and can be used with any numerical descriptor or descriptor combination
- Specific areas (or N-dimensional volumes) with significant gaps in a one dataset can be identified
- Absolute or normalized values allow for the identification of absolute or relative gaps and over-representations
- Dataset similarity is always highest in MACCS metrics followed by PhysChem and Estate
- Sector-based normalized scores indicate that WDI and ChEMBL are most similar spaces, and IRS, ZINC and PUBCHEM form a set distinct from WDI and ChEMBL

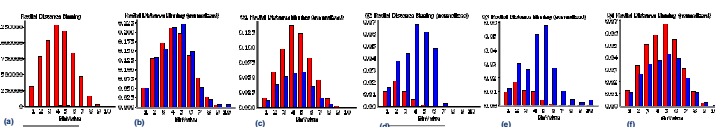
PCA maps

The PCA maps of four different datasets projected onto the WDI space (top). All have the same overall shape in the area near the center but differ largely in the eastern tail. WDI PCA plots with different descriptors (bottom) have different overall shapes. The population densities for the plot area elements even near the center vary significantly but this can not be determined for any area element due to the crowdedness of the plots.

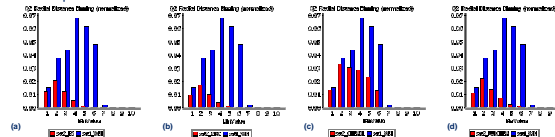


Dataset Overlap: pair barcharts

Pairwise bar charts of the absolute (a) and normalized (b) distance-from-center binned radial occupations of the PCA maps (MACCS fingerprint) of WDI and IRS datasets as well as the normalized per-sector bar charts indicate areas of differing densities between the datasets and gaps in chemical space. Normalized values are useful to identify hot spots and absolute values will be useful to decide if there is need to fill such a gap.

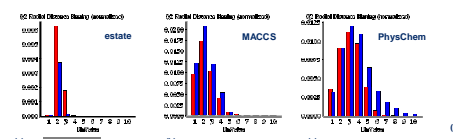


Normalized sector-2 bar charts (MACCS) for the four datasets each compared to WDI show different distance profiles and occupations.



The different shapes and relative populations of the PCA maps for three different descriptor sets by comparing ZINC and IRS are due to different molecule information encoded by the descriptors. Normalized sector-2 bar charts are shown:

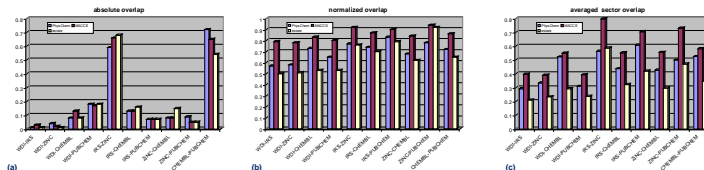
* ECFP-6 data were not finished at the time of printing.



In summary, the bar charts describe the overlap in certain areas of the 2-dimensional PCA projection between two datasets and allow to identify areas of under-population of one dataset with respect to the other one. The procedure is generally applicable to any descriptor and can be extended to N-dimensional analyses.

Dataset Overlap: Scores

The area population densities (as visualized in the bar charts) can be transformed into a dataset density overlap score. This is shown in the following for all dataset pairs and three different score calculation schemes.



Since the absolute overlap scores of the datasets derived from the un-normalized overlap densities (a) are to a large extent determined by the relative sizes of the datasets, normalized scores are preferable (b). As expected, the scores from the sector-based analysis (c) better reflecting the compound neighborhood though still coarse-grained are lower than the radial ones. MACCS keys always yield highest and Estate keys mostly lowest scores.

1. Turner, D. B.; Tyrrell, S. M. & Willett, P. **Rapid Quantification of Molecular Diversity for Selective Database Acquisition.** *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18-22.
2. Engels, M.F.M.; Gibbs, A.C.; Jaeger, E.P.; Verbinen, D.; Lobanov, V. S. & Agrafiotis, D. K. **A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition.** *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 2651-2660.
3. Schamberger, J.; Grimm, M.; Steinmeyer, A.; Hillisch, A. **Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering.** *Drug Discov. Today*, **2011**, article in Press.
4. Passlman, R. & Smith, K. M. **Novel software tools for chemical diversity.** *Perspectives in Drug Discovery and Design*, **1998**, *9/10/11*, 339-353.
5. Oprea, T.I. & Gottfrides, J. **Chemography: The Art of Navigating in Chemical Space.** *J. Comb. Chem.* **2001**, *3*, 157-166.
6. Pipeline Pilot 8.0.1, 2010 Accelrys Software Inc; components **Standardize Molecules** (Standardize Stereo/Charges and Keep Largest Fragment) and **Canonical Smiles**
7. World Drug Index, extracted Dec. 2009, 48129 compounds after clean-up.
8. ChemNavigator.com Inc.: the subset of sourceable compounds was extracted in Dec. 2009 and Bayer in-house compounds were subtracted.
9. ZINC: <http://zinc.docking.org/>; the Zinc 8 set 3 of "drug-like" compounds was extracted in Dec. 2009.
10. ChEMBL db: <https://www.ebi.ac.uk/chembl/index.php>; all chemical structures available May 2011.
11. Pubchem: <http://pubchem.ncbi.nlm.nih.gov/>; compounds from 30 most recent projects were collected to represent current screening chemical space.

Next Steps

- Extend the 2D approach to more and smaller sectors
- Extend to spherical 3D and "cubic" N-dimensional analyses
- How many dimensions are needed in different fingerprints for saturation of DOD scores?
- Compare areas/volumes with uneven populations and find the gaps to fill

Acknowledgements

Stefan Mundt for help in the creation of the PCA plots. Jens Schamberger, Kristin Beyer and Alexander Hillisch for fruitful discussions. Michael Grimm and Hans-Peter Wrona-Metzinger for tips and tricks with Pipeline Pilot.