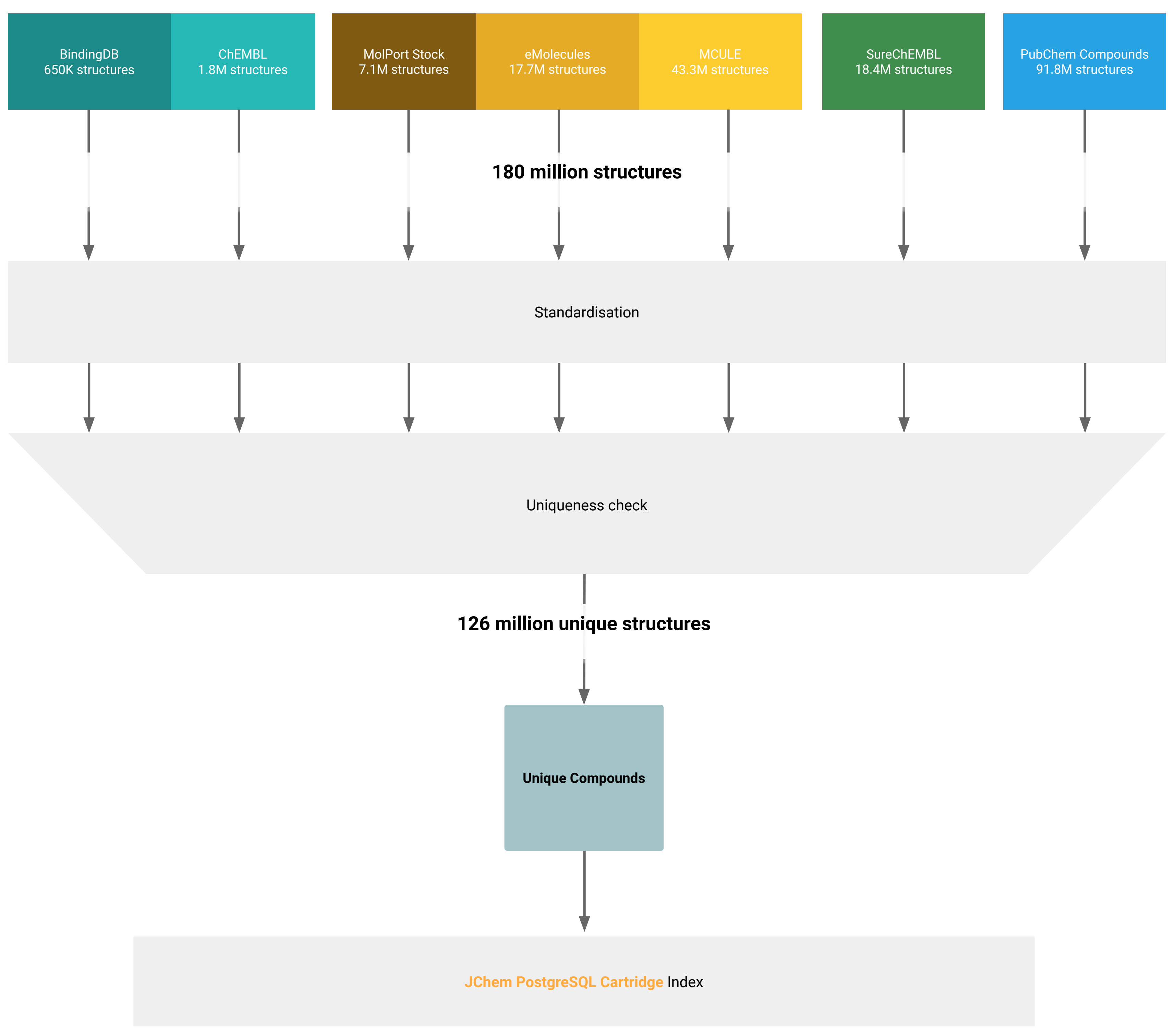
Shrinking the haystack: an overarching search in chemical databases

A. Strácz, A. Volford, N. Sas, L. Hosszú, I. Solt - CHEMAXON KFT, 1031 Zahony u. 7, Budapest, Hungary

Abstract Drug discovery is a knowledge-intensive process in which having the right information at hand can be critical in making the proper decision. With the exponentially growing amount of chemistry and biology-related data in public, commercial and corporate databases it becomes more and more challenging for chemists to find relevant information which helps them to move forward in the right direction with their research.

In this poster we present an ongoing development aiming at providing chemists and other scientists in the pharmaceutical and biotech industries with relevant hits from vendor catalogs, virtual libraries, corporate inventories, in house and publicly available bioassay, metabolic and toxicology databases, as well as patent collections matching the chemical series standing in the focus of their research. Our novel chemical search technologies utilized in this development allow for an instant feedback from very extensive data collections, opening new perspectives in data driven molecule design.

Data processing



BindingDB 650K structures | ChEMBL 1.8M structures | MolPort Stock 7.1M structures | eMolecules 17.7M structures | MCULE 43.3M structures | SureChEMBL 18.4M structures | PubChem Compounds 91.8M structures

180 million structures

Standardisation

Uniqueness check

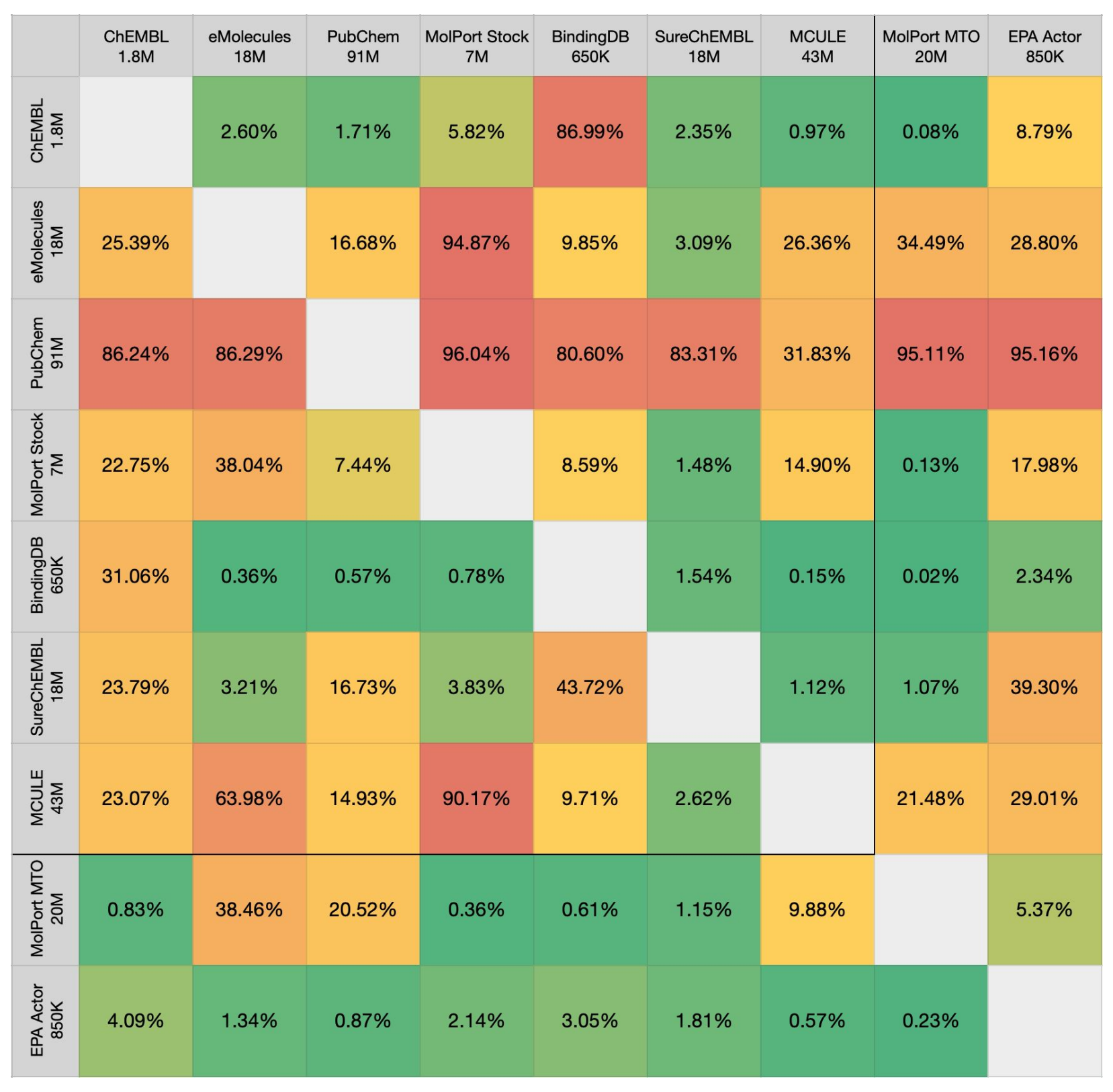
126 million unique structures

Unique Compounds

JChem PostgreSQL Cartridge Index

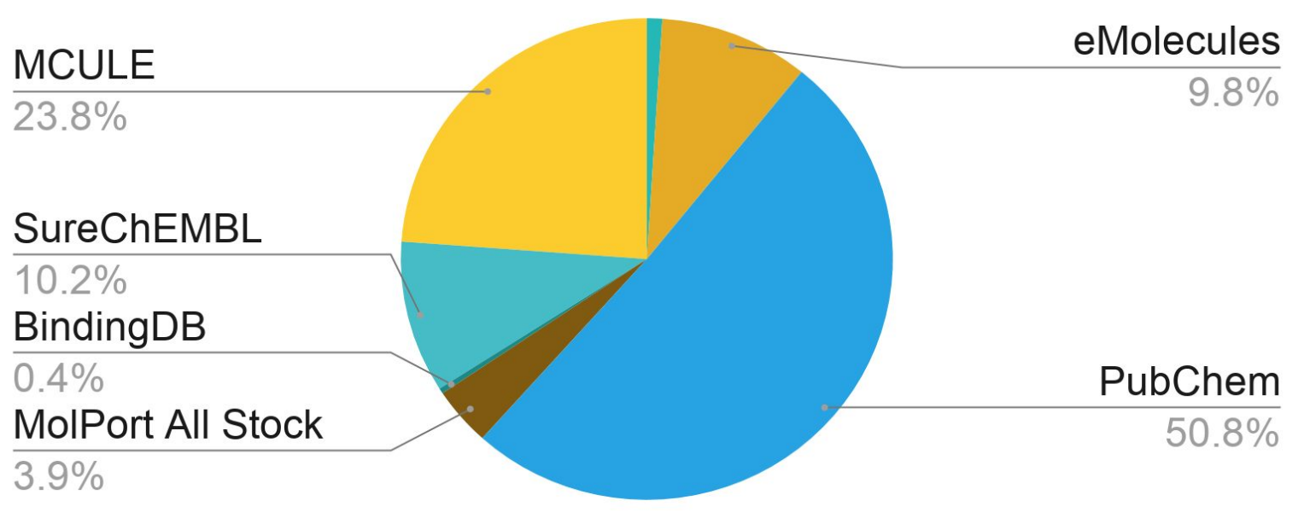
Overlap and diversity of databases

In order to understand the value and the amount of novelty compared to the rest of the data of a newly added database, we have investigated three main aspects: i) the amount of compounds which are only present in a given database, ii) the diversity of the database to be added and iii) the distribution of molecules in a *"Sweet Spot"* [11] chart.

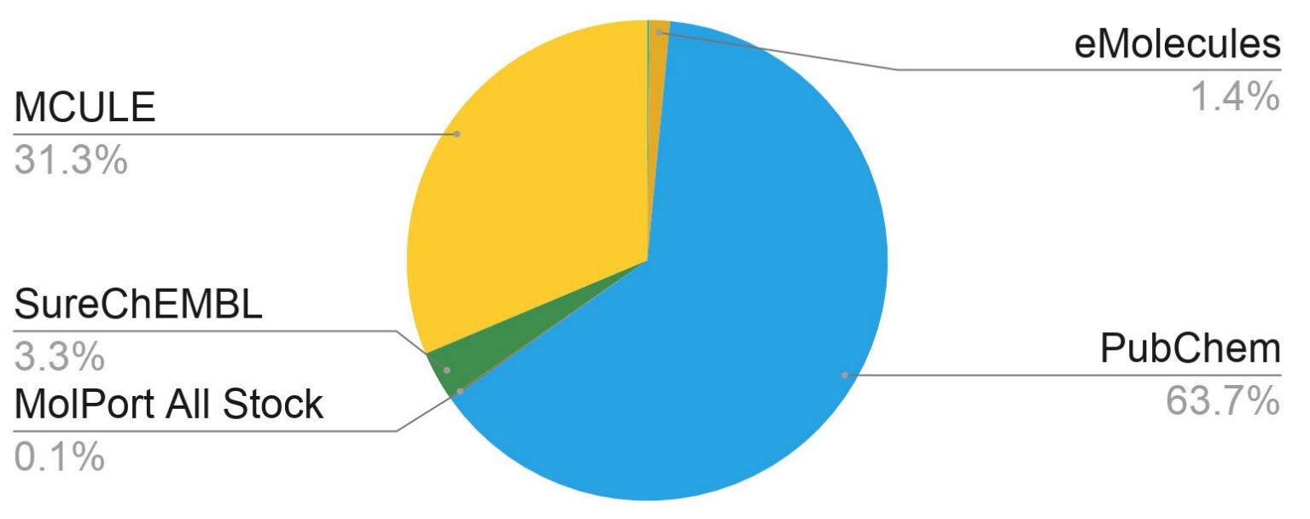


	ChEMBL 1.8M	eMolecules 18M	PubChem 91M	MolPort Stock 7M	BindingDB 650K	SureChEMBL 18M	MCULE 43M	MolPort MTO 20M	EPA Actor 850K
ChEMBL 1.8M		2.60%	1.71%	5.82%	86.99%	2.35%	0.97%	0.08%	8.79%
eMolecules 18M	25.39%		16.68%	94.87%	9.85%	3.09%	26.36%	34.49%	28.80%
PubChem 91M	86.24%	86.29%		96.04%	80.60%	83.31%	31.83%	95.11%	95.16%
MolPort Stock 7M	22.75%	38.04%	7.44%		8.59%	1.48%	14.90%	0.13%	17.98%
BindingDB 650K	31.06%	0.36%	0.57%	0.78%		1.54%	0.15%	0.02%	2.34%
SureChEMBL 18M	23.79%	3.21%	16.73%	3.83%	43.72%		1.12%	1.07%	39.30%
MCULE 43M	23.07%	63.98%	14.93%	90.17%	9.71%	2.62%		21.48%	29.01%
MolPort MTO 20M	0.83%	38.46%	20.52%	0.36%	0.61%	1.15%	9.88%		5.37%
EPA Actor 850K	4.09%	1.34%	0.87%	2.14%	3.05%	1.81%	0.57%	0.23%	

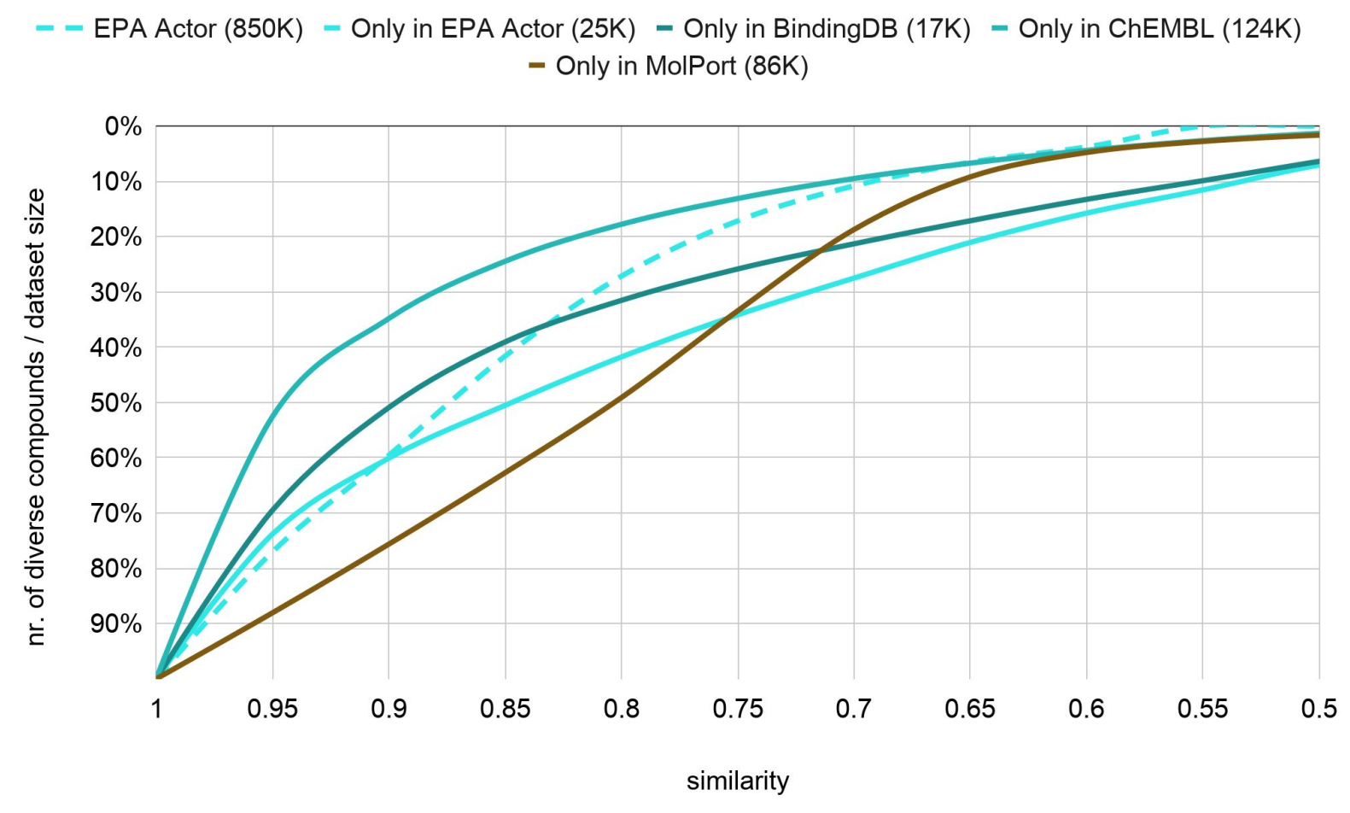
▲ **HEATMAP.** Pairwise overlap ratio of databases. Example: 2.80% of eMolecules is found within ChEMBL, and 25.39% of ChEMBL is found in eMolecules. The total amount of compounds are displayed in the header cells



▲**CHART.** The total contribution of each dataset to the duplicate filtered dataset of 126M molecules.



▲**CHART.** The proportion of exclusive contributions (*i.e.* the unique molecules which are present in only one dataset). Actual values in spreadsheet. Example: PubChem [7] and MCULE have an order of magnitude larger exclusive compound collection than others. **TABLE ▼**



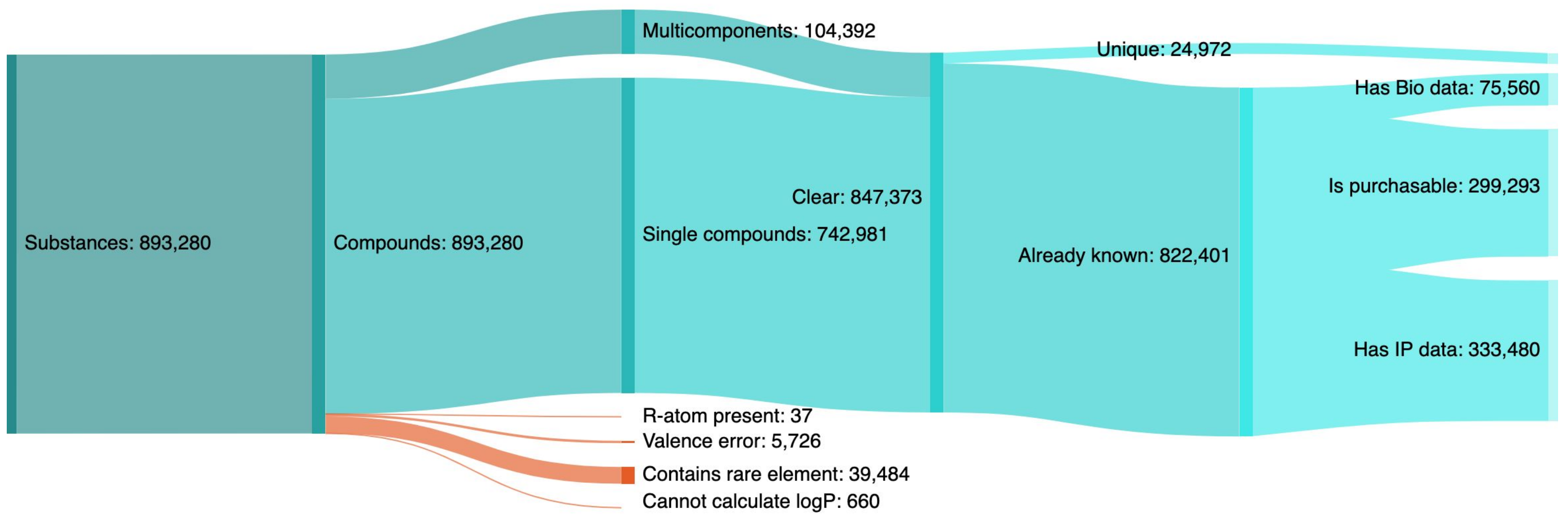
▲**CHART.** Number of diverse picks required - normalized to each database's size - to cover the database at different similarity thresholds. Example: at 0.95 ChEMBL[2] shows low diversity, while Molport is highly diverse. Some databases have been omitted from this chart.

Database	Nr. of exclusives
ChEMBL	122 913
eMolecules	1 021 050
PubChem	44 932 492
MolPort Stock	85 039
BindingDB	16 919
SureChEMBL	2 979 143
MCULE	28 221 313
EPA Actor	24 890
MolPort MTO	357 001

Cost and benefit of adding a new dataset

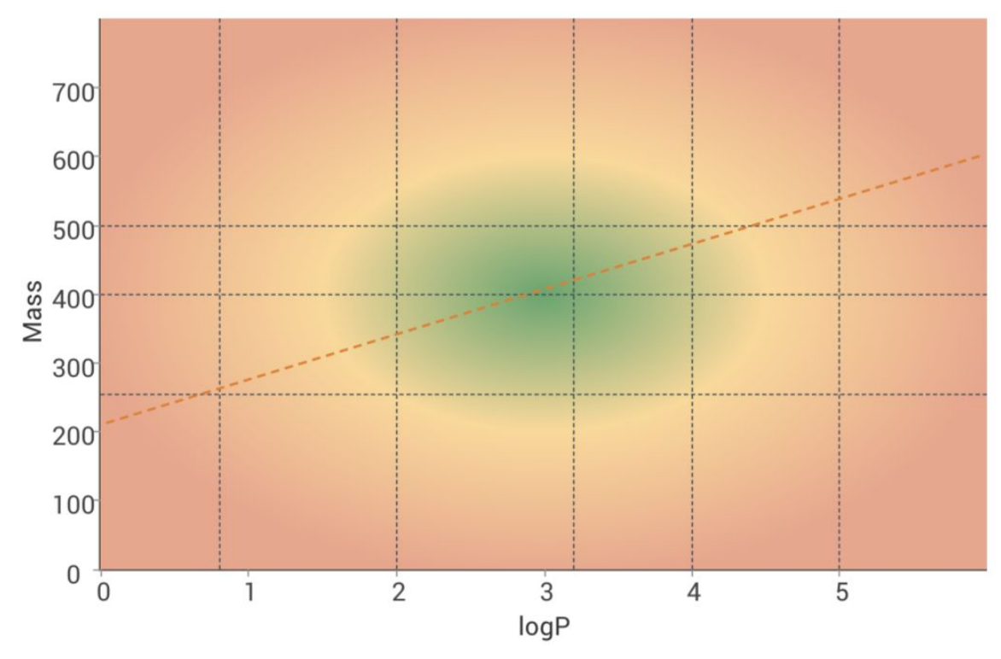
EPA Actor database compiles toxicology data from thousands of public sources [9]. Over 44M assay results of 506 534 assays are made available for 893 280 structures. Integrating it into this content database could provide useful new data for compounds previously known, or surface purchasability and IP information for the compounds. The process below is used to analyze the quality of the content and the size of the most valuable portions. Quality checks were performed with ChemAxon Structure Checker.

FLOW CHART ▼

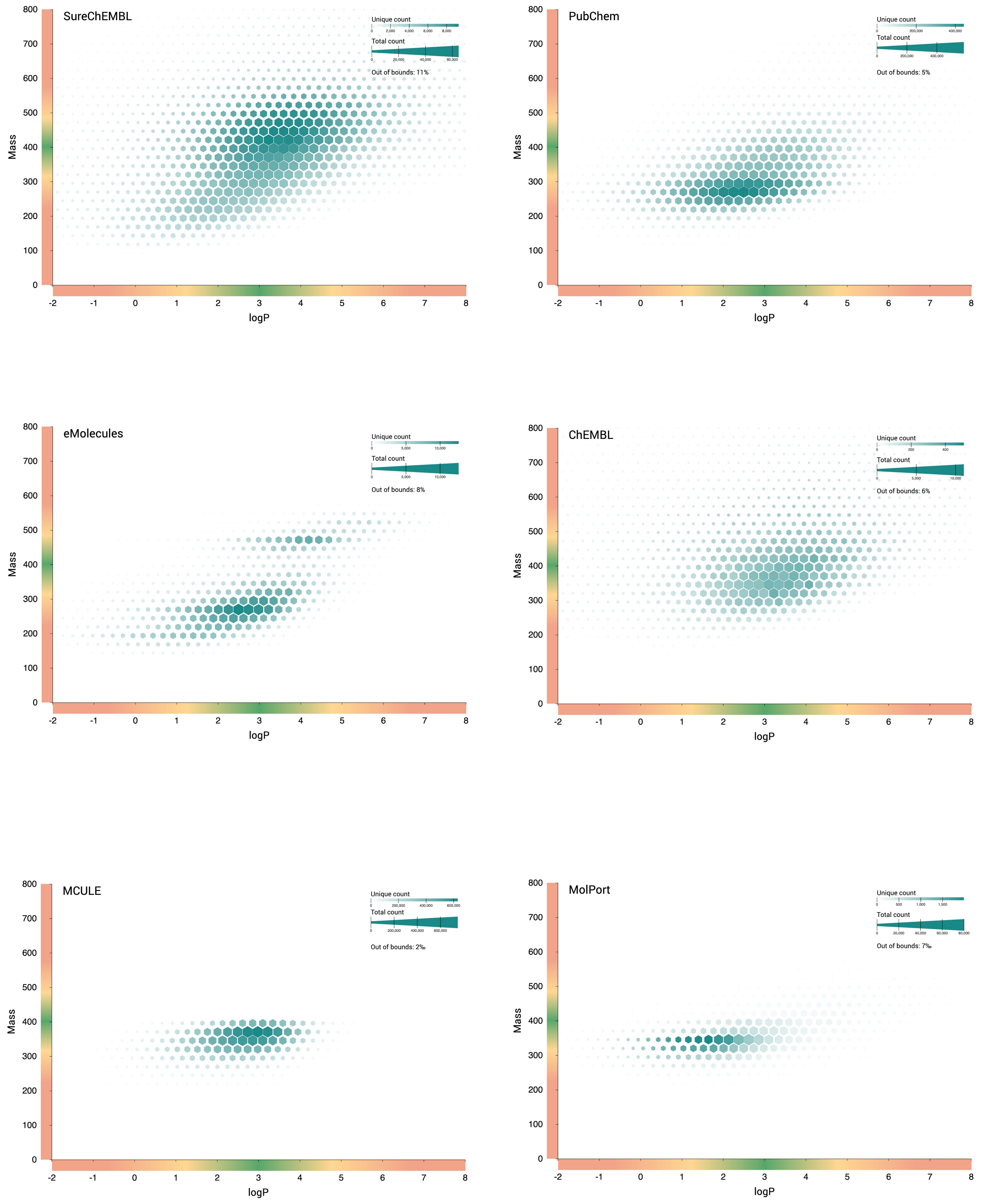


Density plot of databases

While measuring uniqueness and different data types may provide much needed insight into the value of a database, understanding the drug-likeness of chemical space it covers is crucial. Furthermore, as the cleanliness of these databases vary, understanding where the single source compounds spread out is critical. Traditional scatter plots would misrepresent the density of overlapping points in sets of millions of points, so we selected hexbin plots, and used the coloring suggestions of *Hann, et al.*'s "Sweet Spot" visualization [11] to highlight where in mass and logP chemists might want to work. logP values were predicted using ChemAxon Partitioning Plugin. **HEXBIN PLOTS ▼**



▲ **SWEET SPOT CHART**



Conclusions

A new compound database has been created with a collection rivaling the size of leading content services. Each investigated database appears to contribute novel compounds, data. A robust workflow has been developed for adding more databases where the limitation will be comprehension of analysis results and the availability of tools and algorithms that deal with hundreds of millions of records. Operating the service requires modest hardware but great performance is already achievable.

Work should continue on the understanding and exploitation of collected biological/IP/purchasability data.

Citations

[1] BindingDB www.bindingdb.org/ Accessed: 2019-04-03
[2] ChEMBL v24, ftp.ebi.ac.uk/pub/databases/chembl/ Accessed: 2019-04-08
[3] MolPort https://www.molport.com/shop/database-download Accessed: 2018-09-11
[4] eMolecules https://downloads.emolecules.com/ Accessed: 2018-07-26
[5] MCULE https://mcule.com/database/ Accessed: 2018-08-18
[6] SureChEMBL ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL Accessed: 2018-09-10
[7] PubChem ftp.ncbi.nlm.nih.gov/pubchem/Compound/ Accessed: 2018-10-16
[8] JChem PostgreSQL Cartridge https://chemaxon.com/products/jchem-engines v18.23.0
[9] EPA Actor https://actor.epa.gov/actor/ Accessed: 2019-10-01
[10] Marvin Live, https://chemaxon.com/products/marvin-live v19.17.1
[11] M. Hann, G. Keserű, Nat. Rev. Drug Discov., 2012, 11, 355–365
[12] JKlustor https://chemaxon.com/products/jklustor v19.22.0

{ astracz, avolford, isolt, nsas, lhosszu }@chemaxon.com

ChemAxon