U.S. FOOD & DRUG ADMINISTRATION

Scalable Filtering of Chemical Substances on HPC Clusters

40 DAYS TO < 10 MINUTES

Mike Mikailov, Computer Scientist
DIDSR/OSEL
CDRH

Mike Mikailov[1], Yulia Borodina[1], Fu-Jyh Luo[1], Kenny Cha[1]

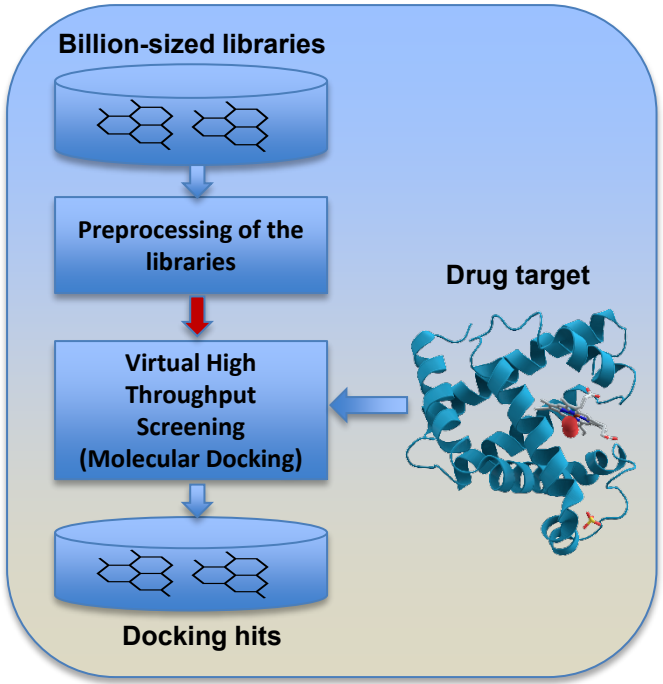
[1]U.S. Food and Drug Administration, Silver Spring, 20993, MD, USA

Disclaimer

- The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.
- The information in this presentation represents the opinions of the speaker and does not necessarily represent FDA's position or policy.

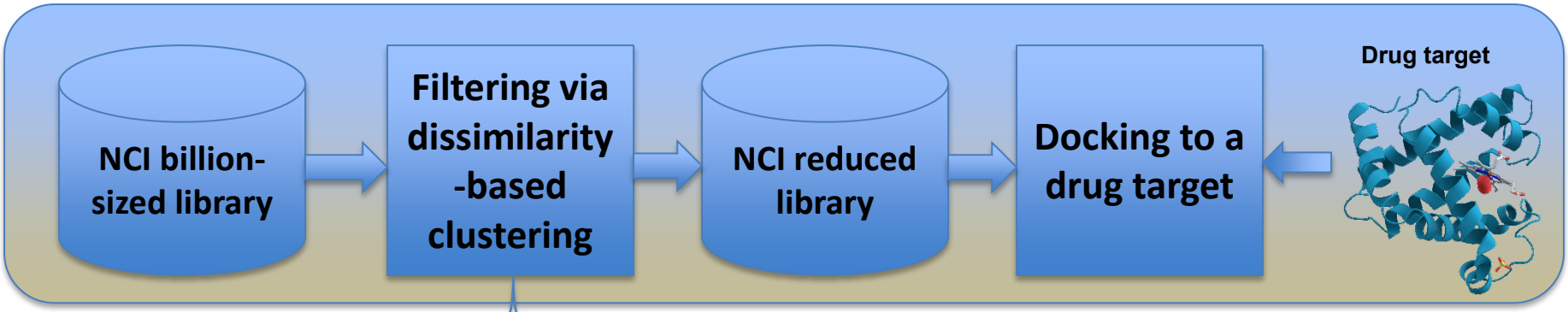
Introduction/Hypothesis

- The recent explosion of chemical libraries of National Cancer Institute (NCI) beyond a **billion molecules** led to large-scale simulations for **Virtual Screening (VS)**.
- **VS** is a simulation technique used in drug discovery to search libraries of molecules to identify structures, likely to bind to a drug target. It is estimated that over 950 years are needed for processing the billion-sized libraries: 30 sec. for docking per molecule leads to ~950 years per one billion molecules.
- To make **VS** more efficient, it is **desired that a very large database of chemicals was pre-filtered and only a subset of molecules was used for docking.**
- The FDA CDRH High-Performance Computing (HPC) team is working with NCI to apply innovative scaling techniques to make this mission critical task feasible.



Virtual Screening

Virtual screening via clustering filtering



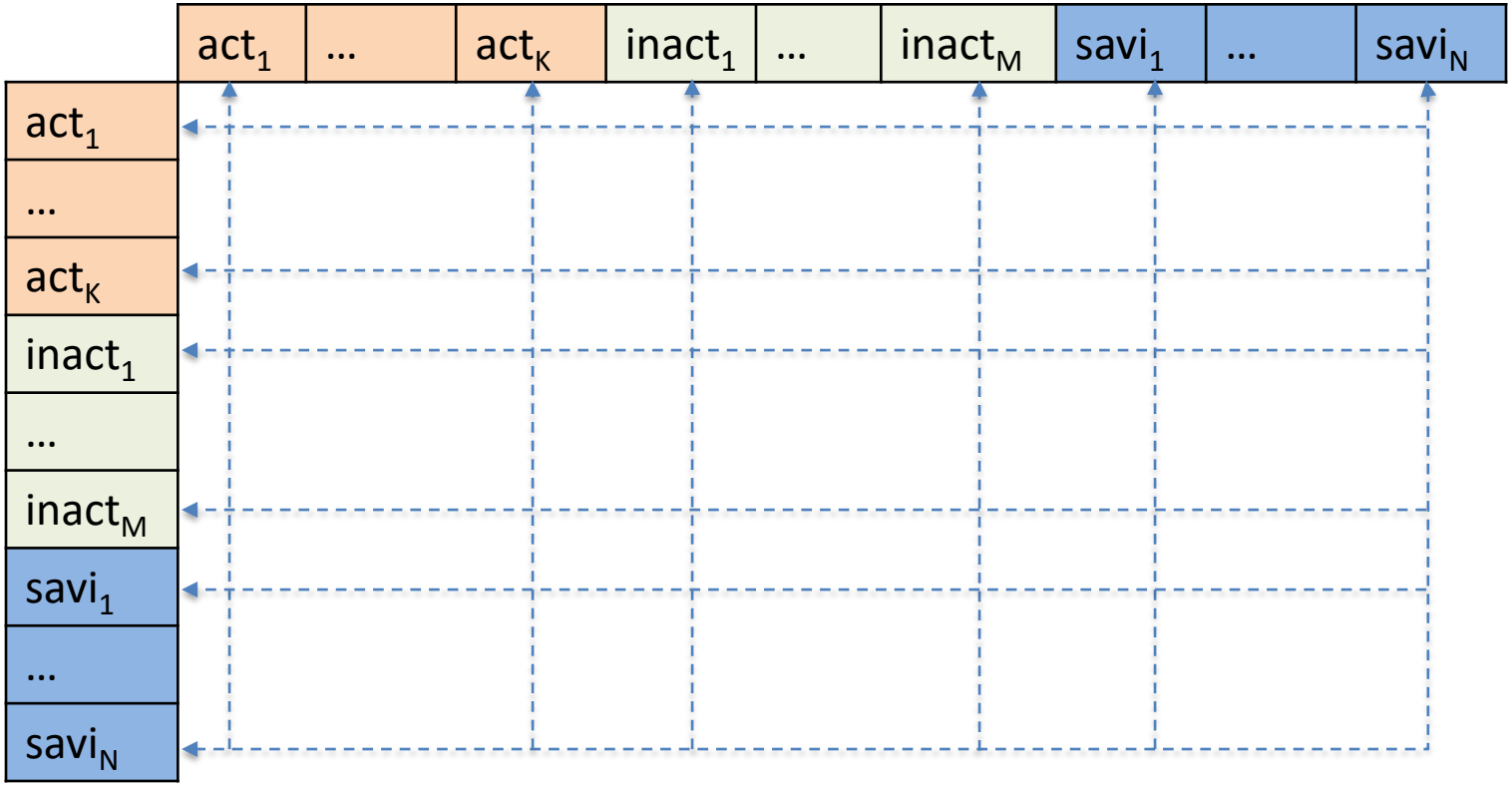
The pre-filtering not only decreases the dimensionality of the dataset but also enriches it with more structurally diverse and/or more drug-like substances.

Dissimilarity-based clustering on one CPU

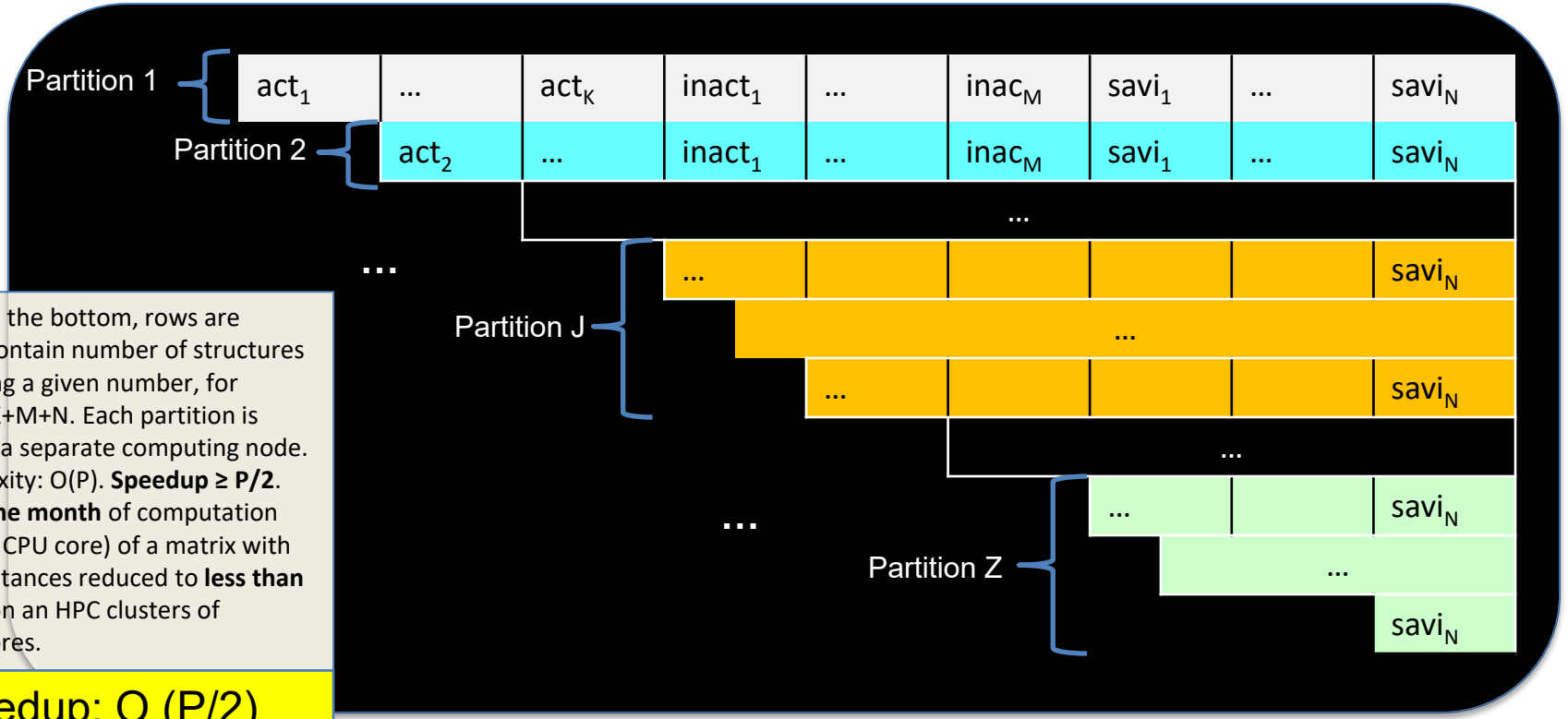
Dissimilarity matrix for one Synthetically Accessible Virtual Inventory (SAVI) file after merging with active and inactive structures. Computing dissimilarity matrix for the entire dataset (**all-to-all**) is prohibitively expensive on **one CPU core** : more than a month is needed for computing a matrix of only 100,000 substances,

Time complexity: **$O(P^2)$**

where P=K+M+N



Distributed clustering via matrix partitioning

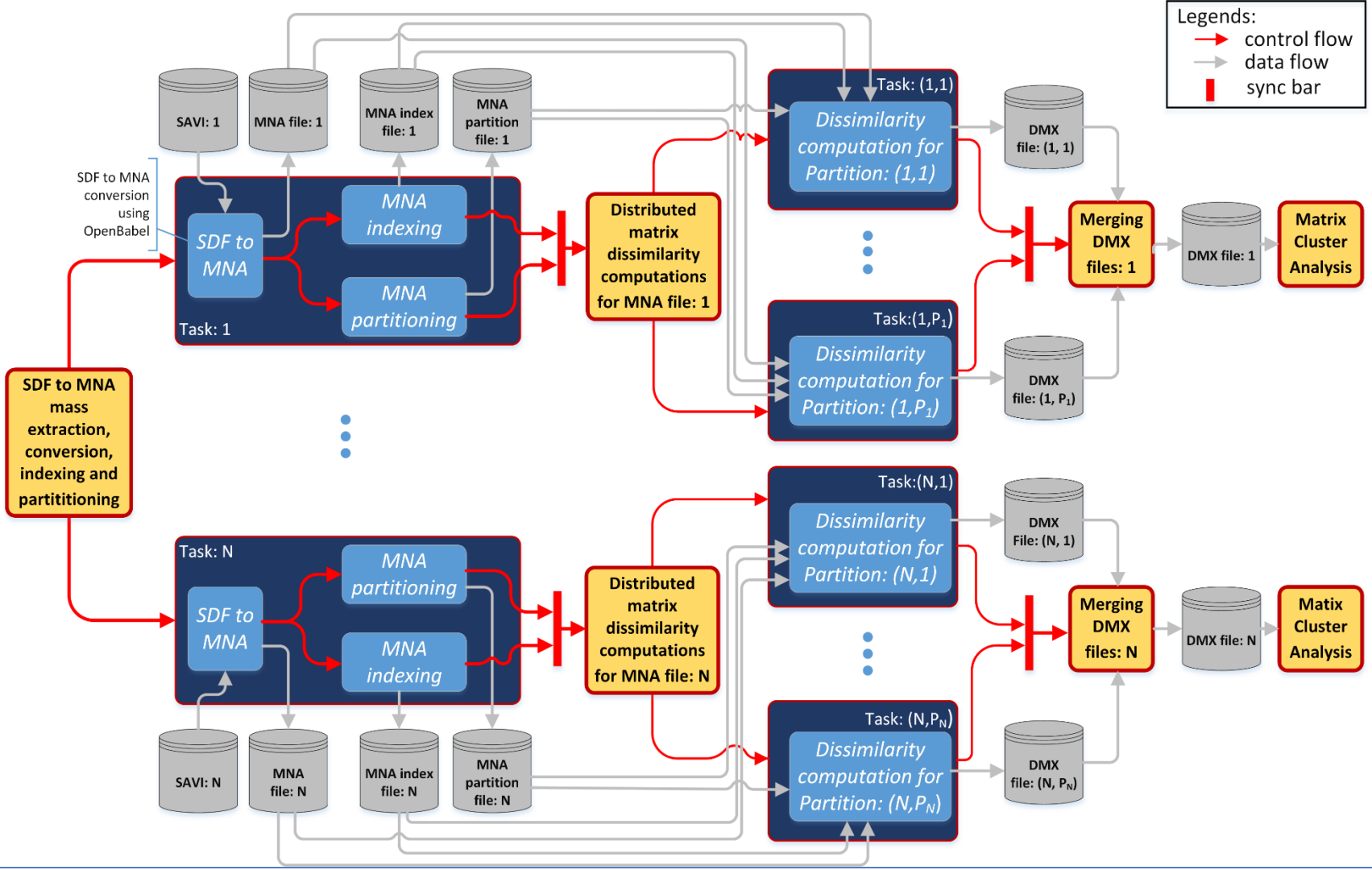


Starting from the bottom, rows are grouped to contain number of structures not exceeding a given number, for instance, P=K+M+N. Each partition is delegated to a separate computing node. Time complexity: O(P). **Speedup ≥ P/2**. More than **one month** of computation time (on one CPU core) of a matrix with 100,000 substances reduced to **less than 10 minutes** on an HPC clusters of 5,000 CPU cores.

Speedup: O (P/2)

Performance on the HPC cluster of 5,000 CPUs

Number of SAVI files	Number of structures or partition size in a SAVI file, P	Number of structures in the top triangle, T=(P+1)*P/2	Number of partitions or tasks, Z=T/P or Z = ⌈ (P+1)/2⌉	Number of batches, B= Z/5,000	Before scaling	After scaling
1	100,000	5,000,050,000	50,001	10	~40 days	< 10 minutes



Scalable Workflow Diagram for Filtering Chemical Substances on HPC Clusters

U.S. FOOD & DRUG ADMINISTRATION

Thank you!

Disclaimer: *The information in this presentation represents the opinions of the speaker and does not necessarily represent FDA's position or policy.*


FDA
U.S. FOOD & DRUG
ADMINISTRATION