Authors: Mohammad S.A

Affiliation: Autodidacts et.al

هاكاثون الابتكار الصحي
Health Innovation Hackathon

DNA:Target fishing in bigdata

Introduction

Big biomedical data e.g omics play an essential rule for the development of AI&ML applications in computational medicinal chemistry furthermore Chemogenomics data which generally refers to the activity data of chemical compounds on an array of protein targets represents an important source of information for building in silico target prediction models. The increasing volume of chemogenomics data offers exciting opportunities to build models based on Big Data. Alongside with an ever expanding libraries of small molecules and the novelty of alpha-fold2 in target structure prediction there's a new challenges of manipulating these data e.g ELT and crud operations. Herein we introduce DeNovoAutomata a platform with new breed of tools that efficiently store &analyze it to give meaningful insights which reflects a paradigm shift in biomedical research

The platform integrate automation tools and expose API endpoints for the Researcher in the filed

Methods

in Polypharmacology an important aspect of designing drugs is activity cliffs a minor structural modification which changes the target biological activity significantly hence it affect ADME-DMPK profiles . analysis tools for Pharmacokinetic and Pharmacodynamic are develop to address this namely SAR -QSAR however due to it's nature Qsar models prediction relays on data quality and availability with the diverse databases available Online and lack of clean dataset particularly for QSAR analysis data fusion had to be performed for curation of such database "ExCAPE-DB" have 18GB This dataset comprises over 70 million SAR data points we download the entire data set then convert it into hadoop5 opening with Vaex was very fast (less than a second) As a bonus, Vaex can automatically generate a Machine Learning pipeline ,then we had to recreate subset of RDKIT on daskDF to generate similarity mapping plots(1)which use less memory usage at any given moment when dealing with large dataset , another way way of analysis is matched molecular pair analysis (MMPA) a quite useful technique to create, compile, store, retrieve, and use MMP rules.in mmpdb DB a semi automated pipeline can be created to do ELT .which emphasizing the new analysis schemes therefore to visualize these data we integrate CIME a webapp hat can help explore these data (figure5)

In docking-based IVS, (2) given small molecule is docked to the binding site of each protein in a target database through a docking engine "BLinDPyPr" is such an example of an automated workflow (3) moreover since cavity detection is crucial in blind docking we integrate PocketPipe (figure4) which can do automatic ELT for protein pocket space in parallel. thus accelerate the docking process .we investigate integrating a consensus docking for validations

a new approach of storing protein is structure in binary format (mmtf) has been developed hence new set of tools has emerged namely MMTF PySpark and Lemons framework which has been used for parsing and retrieving structural data for the targeted protein

Results

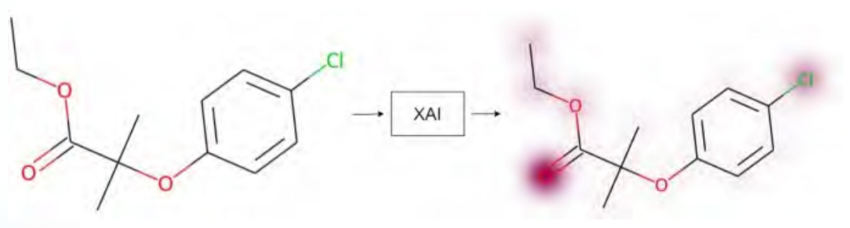
BLinDPyPr pipeline reported to achieve 45.9% success rates on the PDBbind benchmarking set, while a classic DOCK6 blind docking run yields 20.4% and classic DOCK6 site-specific docking achieves 49.7%.

-The entire PDB archive (~121,000 structures) can be stored in less than 7GB. Meaning it can fit into RAM on a Desktop machine performing a structural query for a protein reported to take 10 minutes on an 8 core machine.

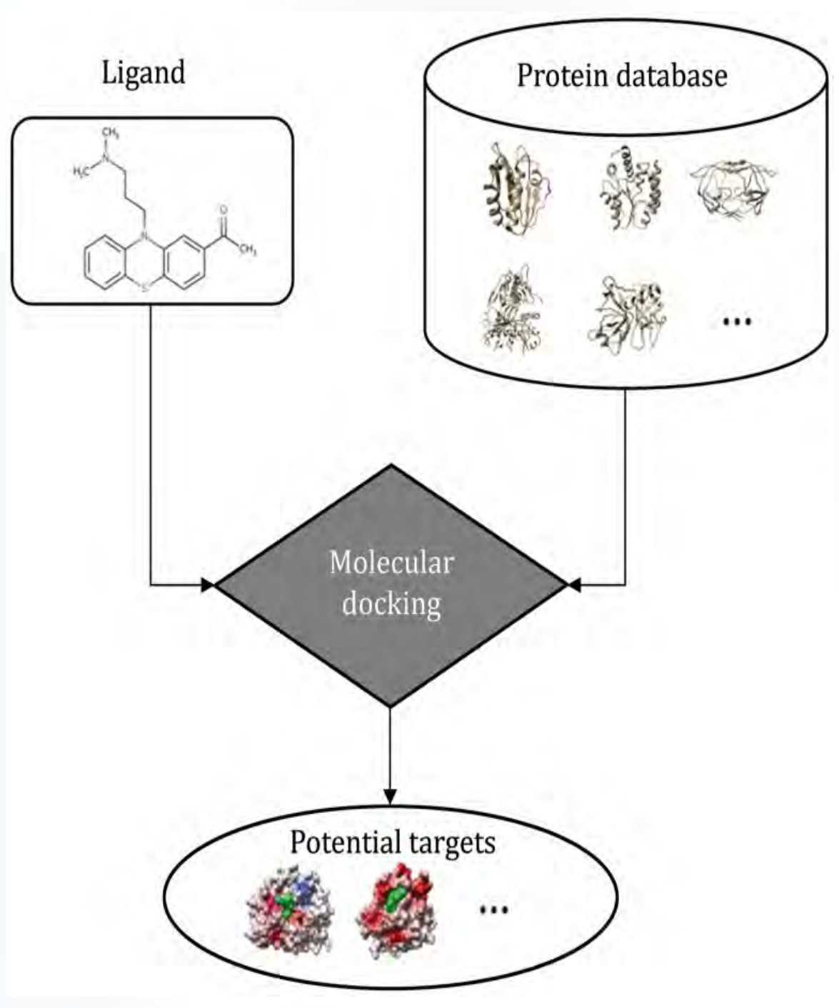
visualizing 'The structure of PDB_ID 4V60 protein vault at 3.5 angstrom resolution' Loaded and parsed 241956 atoms in 2.0 seconds. Prepared rendering in 0.082 seconds using the visualizing tool

a new set of tools had to be develop / reimplemented for usage in Bigdata and distributed computing in which interpretable QSAR can answer questions like -If the compound is predicted to be active against some target what are the driving factors of its activity? Or if it is predicted to be inactive, how its activity can be modulated? hence it can be used in hit to lead optimizations or re-purpose drug candidates or even using this knowledge for finding novel scaffold variants in target-focused pharmacophore modeling

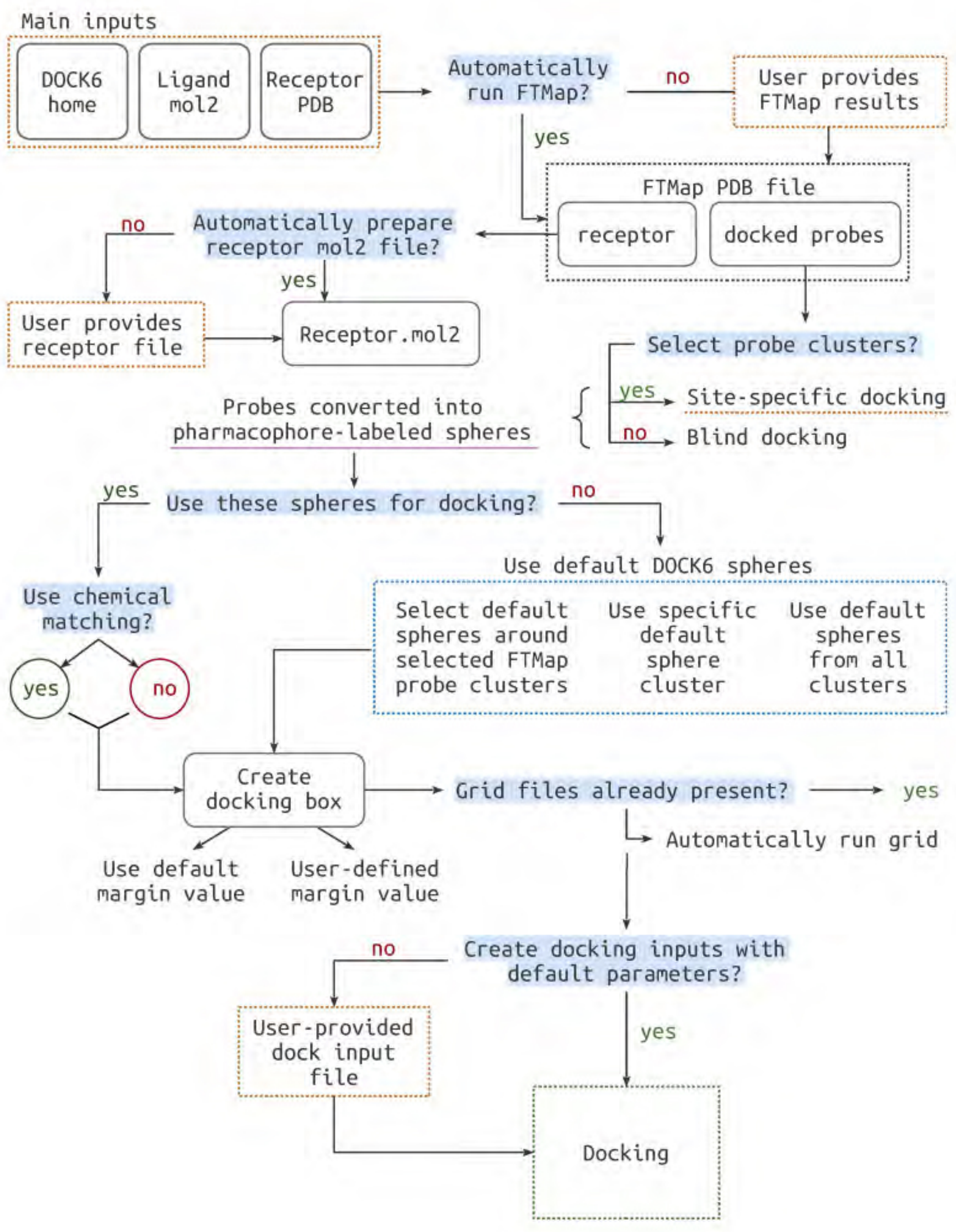
Graphs and Tables



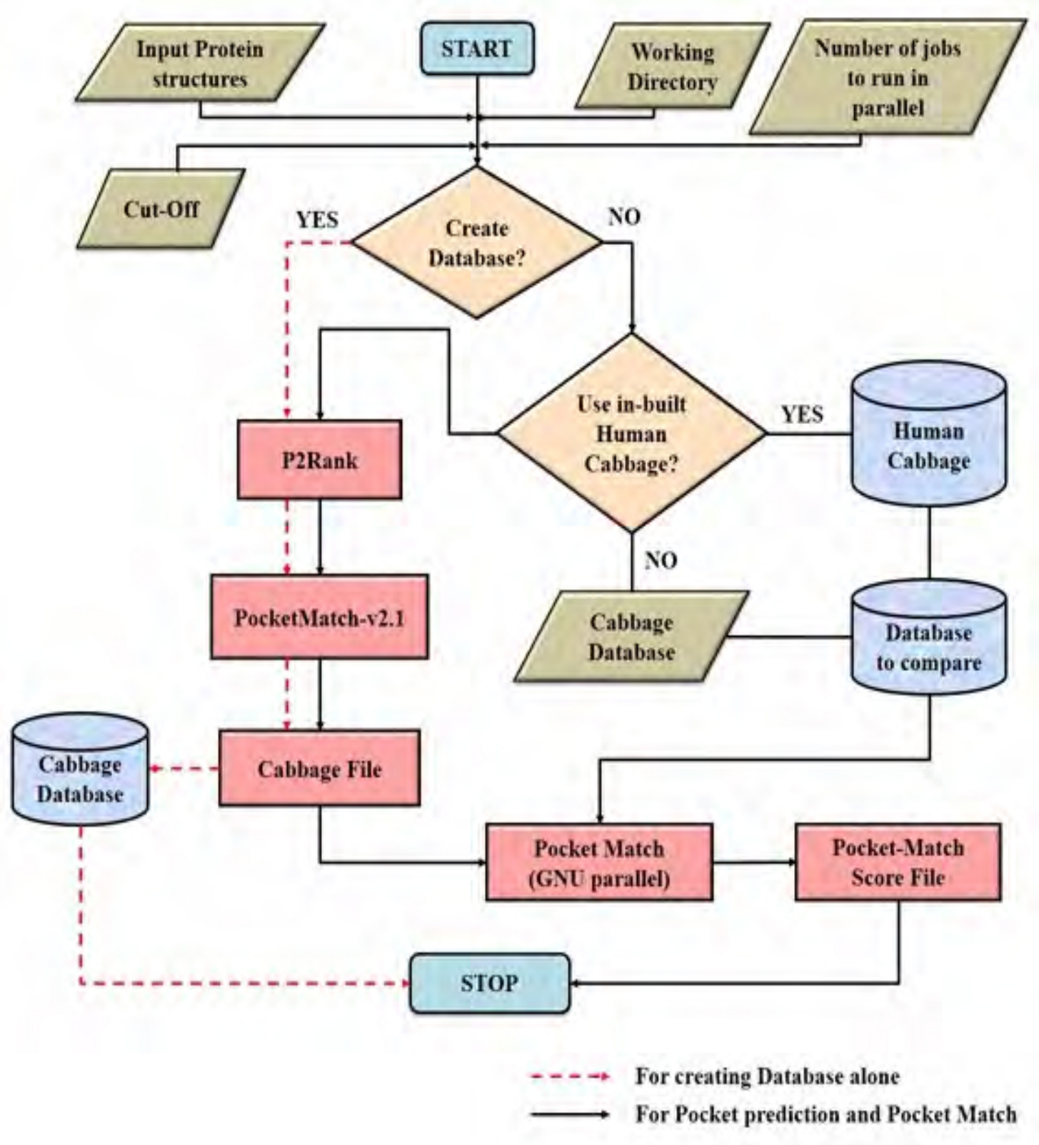
1
similarity mapping



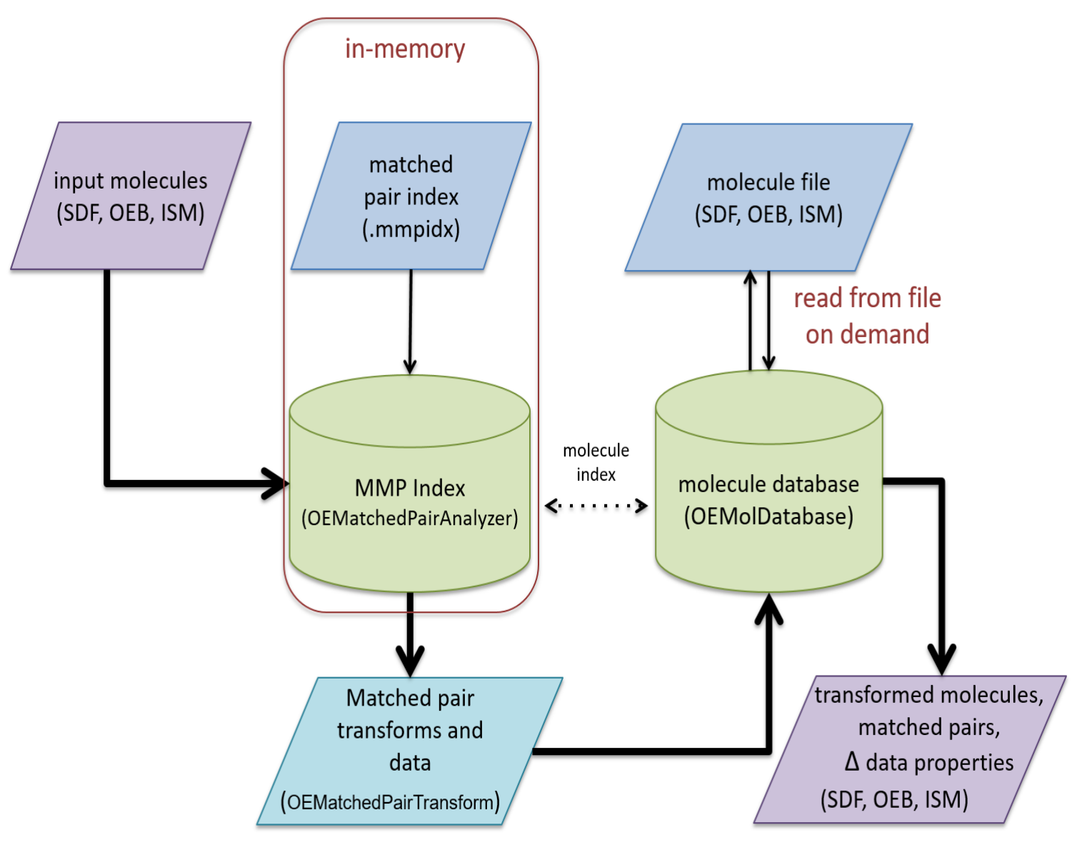
2
A flowchart of the docking-based inverse virtual screening


Main inputs
DOCK6 home
Ligand mol2
Receptor PDB
Automatically run FTMap?
no
User provides FTMap results
yes
FTMap PDB file
receptor
docked probes
Automatically prepare receptor mol2 file?
no
yes
User provides receptor file
Receptor.mol2
Select probe clusters?
yes
Site-specific docking
no
Blind docking
Probes converted into pharmacophore-labeled spheres
yes
Use these spheres for docking?
no
Use default DOCK6 spheres
Select default spheres around selected FTMap probe clusters
Use specific default sphere cluster
Use default spheres from all clusters
Use chemical matching?
yes
no
Create docking box
Use default margin value
User-defined margin value
Grid files already present?
yes
Automatically run grid
Create docking inputs with default parameters?
no
yes
User-provided dock input file
Docking

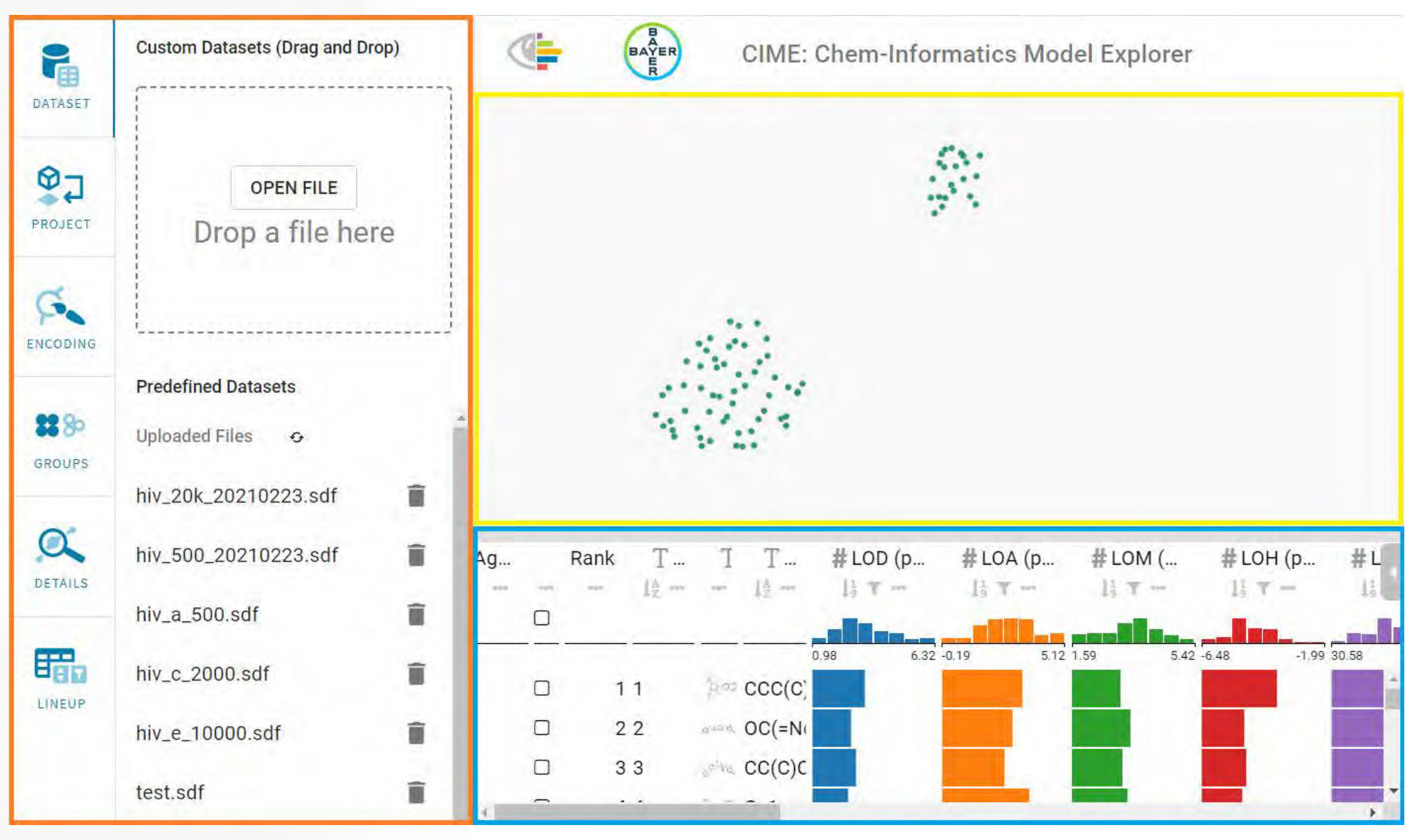
3
semi automatic blind docking approach



4
automated ELT parallel tool for pockets comparison


in-memory
input molecules (SDF, OEB, ISM)
matched pair index (.mmpidx)
molecule file (SDF, OEB, ISM)
read from file on demand
MMP Index (OEMatchedPairAnalyzer)
molecule index
molecule database (OEMolDatabase)
Matched pair transforms and data (OEMatchedPairTransform)
transformed molecules, matched pairs, Δ data properties (SDF, OEB, ISM)

5
Semi automated ELT for chemical molecules which could be visualized using web UI



Custom Datasets (Drag and Drop)

Drop a file here

Predefined Datasets

Uploaded Files

hiv_20k_20210223.sdf

hiv_500_20210223.sdf

hiv_a_500.sdf

hiv_c_2000.sdf

hiv_e_10000.sdf

test.sdf

Conclusion

We found that integration of automation pipelines and big data with parallel computing has great advantage over traditional workflows DeNovoAutomata platform cloud be summarized as building blocks of modules (a set of tools and API) that deal with specific domain problem i.e. ligand space then incorporate tools for processing efficiently

we suggest that more tools needs to be developed towards unified api that utlize automation and and harness the full potential and power of parallel distributed computing

Jobs can be developed in average computer laptop then replicated using BioExcel workflow to scale from workstations to high end HPCS

With the rapid increasing volume and diversity of data concerning drug related targets and their ligands, the simple ligand-based target fishing approach would play an important role in assisting future drug design and discovery

SO_KSAUHS