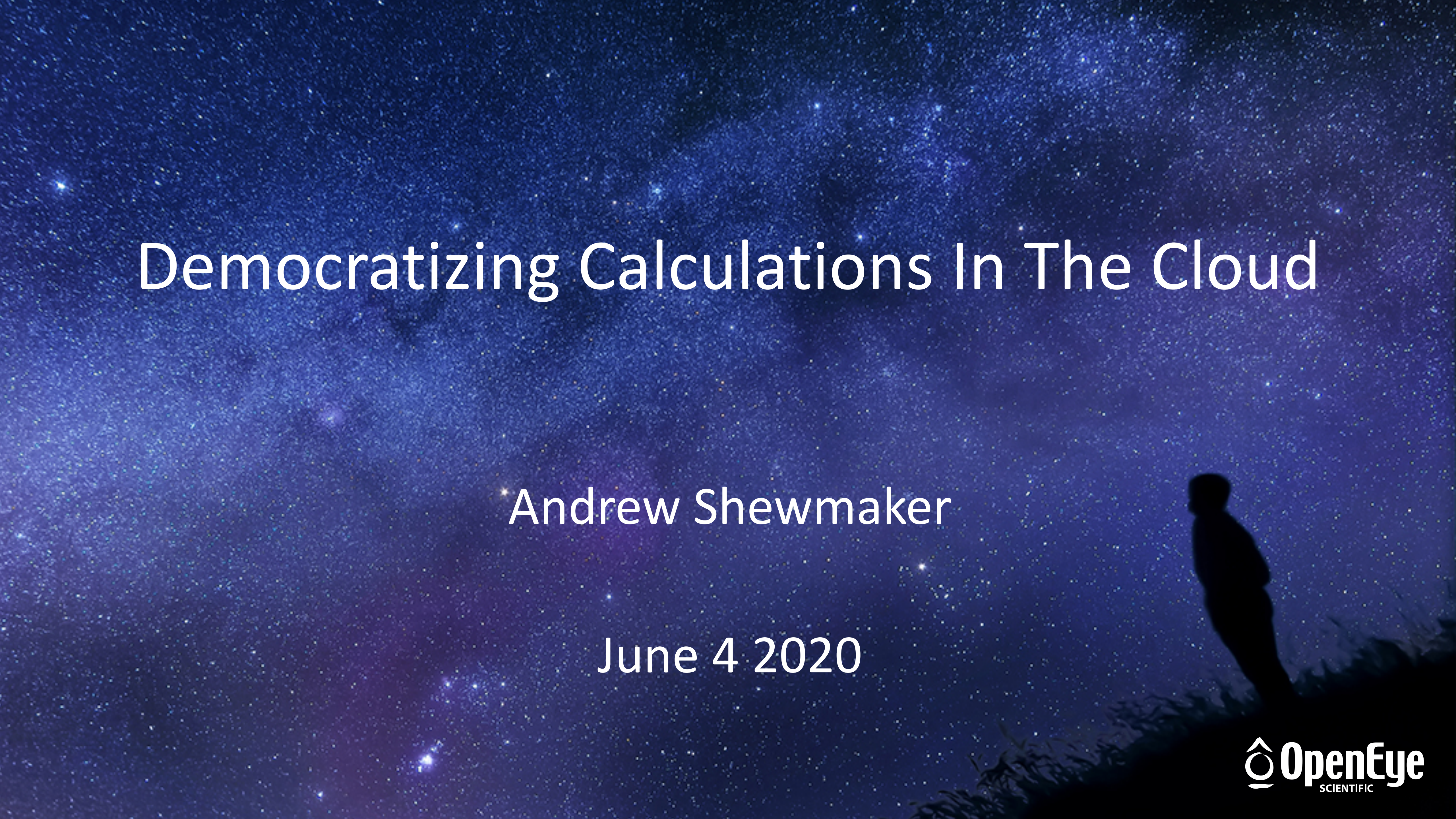


Democratizing Calculations In The Cloud

Andrew Shewmaker

June 4 2020

OpenEye SCIENTIFIC

Agenda

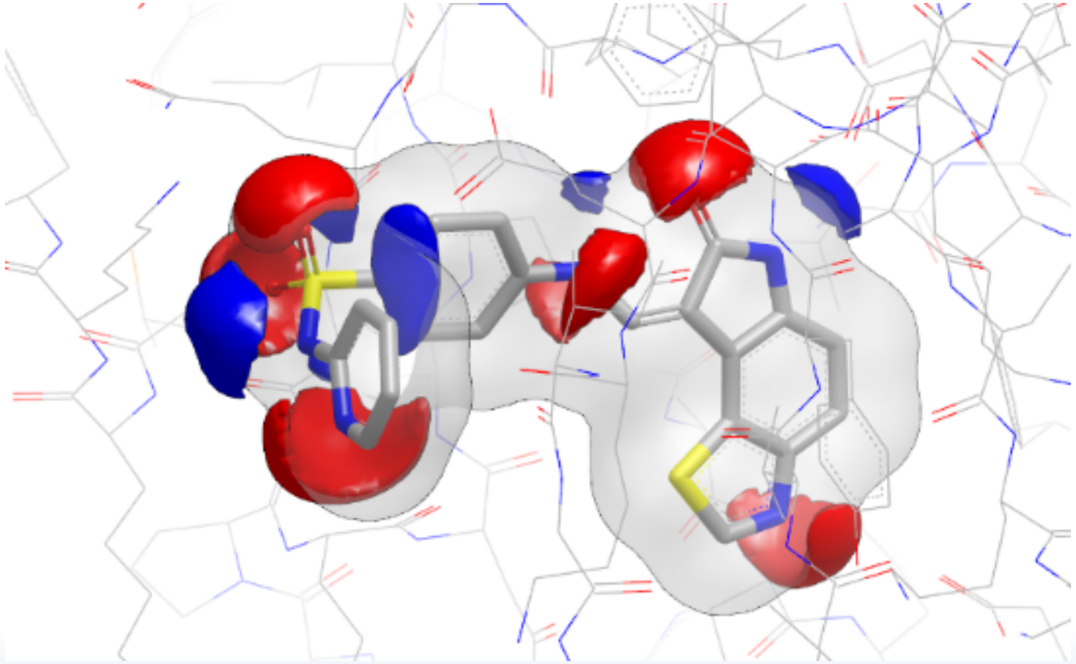
- Intro to OpenEye and Orion
- HPC Science in Orion

Stellar nursery NGC 2174

OpenEye Modeling & Cheminformatics

- Founded 1997
  - Anthony Nicholls
- Santa Fe HQ
  - Boston, Cologne, Tokyo, remote
- Organic growth – no VC
- 70 employees

Molecular Similarity (shape & electrostatics) == Biological Similarity



Looking For Potential COVID-19 Therapeutics: Docking Studies On 1.4 Billion Molecules

5TB input data, 50K CPUs,

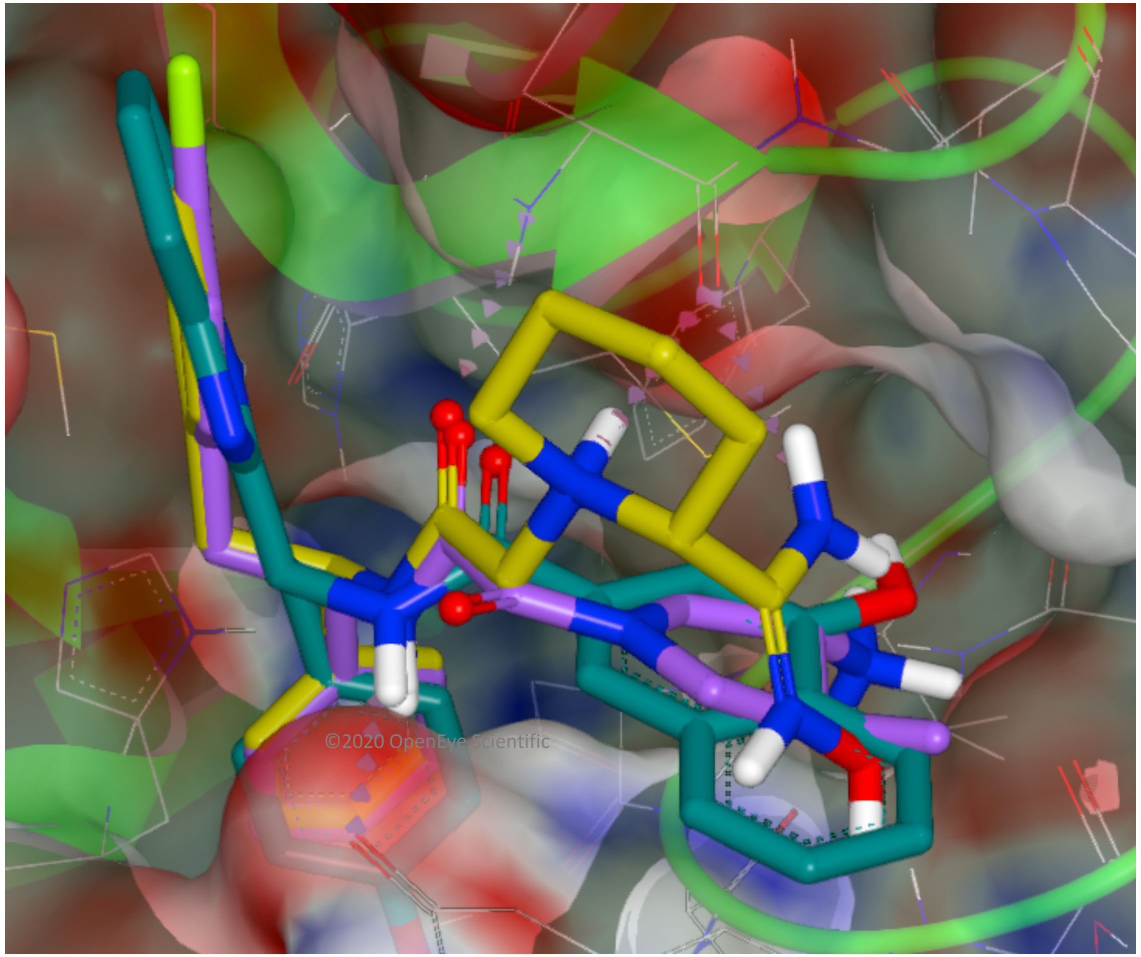
3.5 days per study

SARS-CoV-2 Mpro protease

www.eyesopen.com/blog/openeye-deploys-the-orion-molecular-design-platform-to-find-covid-19-therapeutics

Angiotensin converting enzyme 2 (ACE2)

www.eyesopen.com/blog/openeye-releases-additional-giga-scale-virtual-screening-covid-19-data-for-public-use



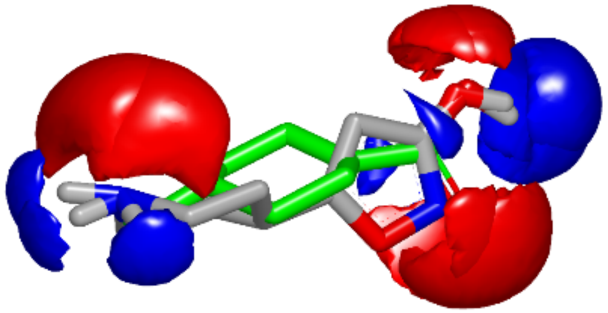
How To Enable Subject Matter Experts To Evaluate 10^{50} Molecules Using $>10^7$ CPUs?

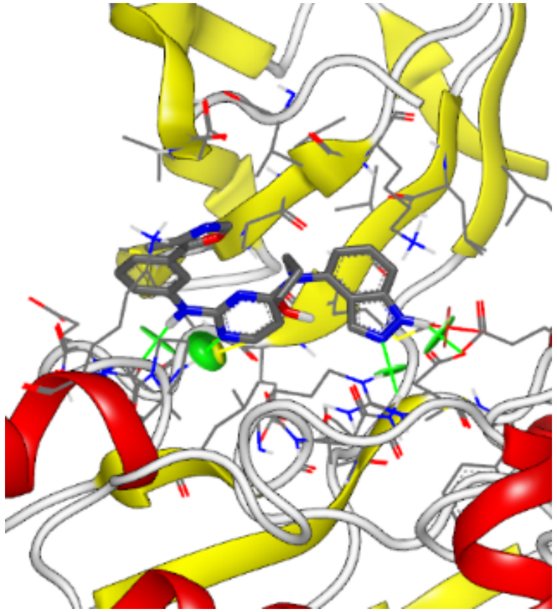
orion
OpenEye Scientific

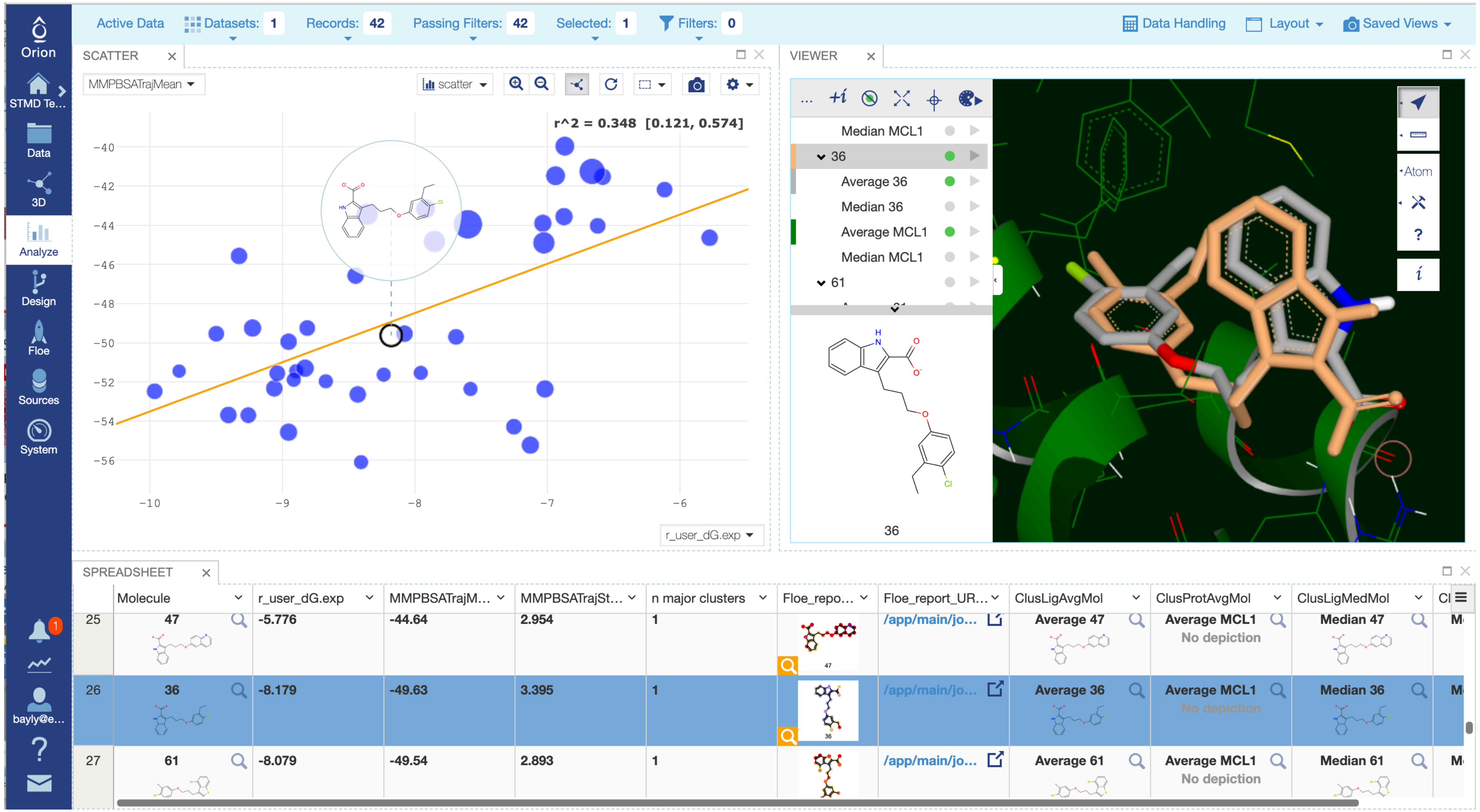
- Automatic scale-up
- Automatic parallelism
- Automatic fault tolerance
- Automatic storage and backup of results
- Automatic scale down when calculations are finished
- Automatic reproducibility of methods at a future time

amazon web services

Computer Aided Drug Design With Orion



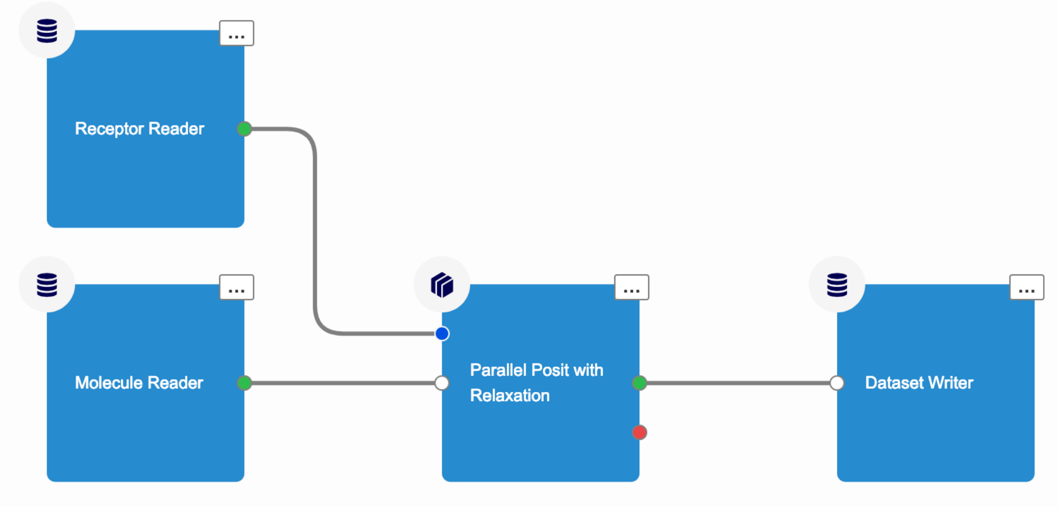


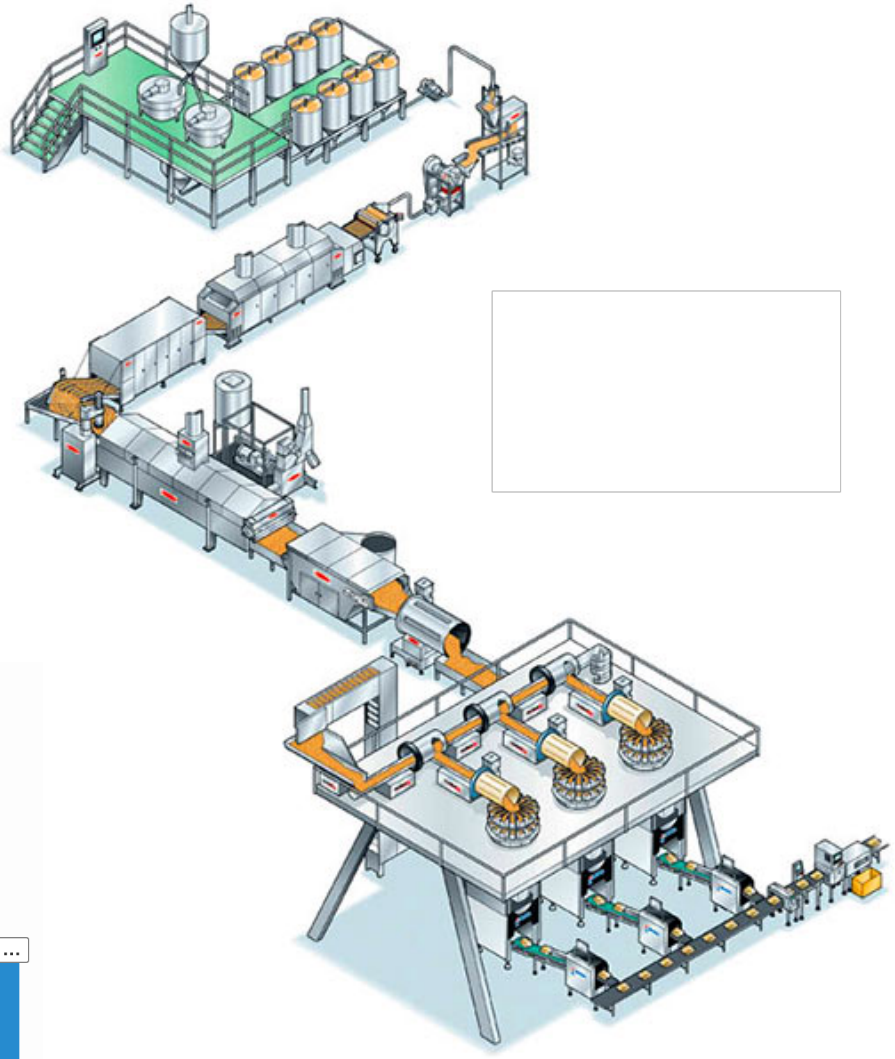


	Molecule	r_user_dG.exp	MMPBSATrajM...	MMPBSATrajSt...	n major clusters	Floe_repo...	Floe_report_UR...	ClusLigAvgMol	ClusProtAvgMol	ClusLigMedMol	Cl
25	47	-5.776	-44.64	2.954	1	47	/app/main/jo...	Average 47	Average MCL1 No depiction	Median 47	M
26	36	-8.179	-49.63	3.395	1	36	/app/main/jo...	Average 36	Average MCL1 No depiction	Median 36	M
27	61	-8.079	-49.54	2.893	1		/app/main/jo...	Average 61	Average MCL1 No depiction	Median 61	M

Flow-Based Programming

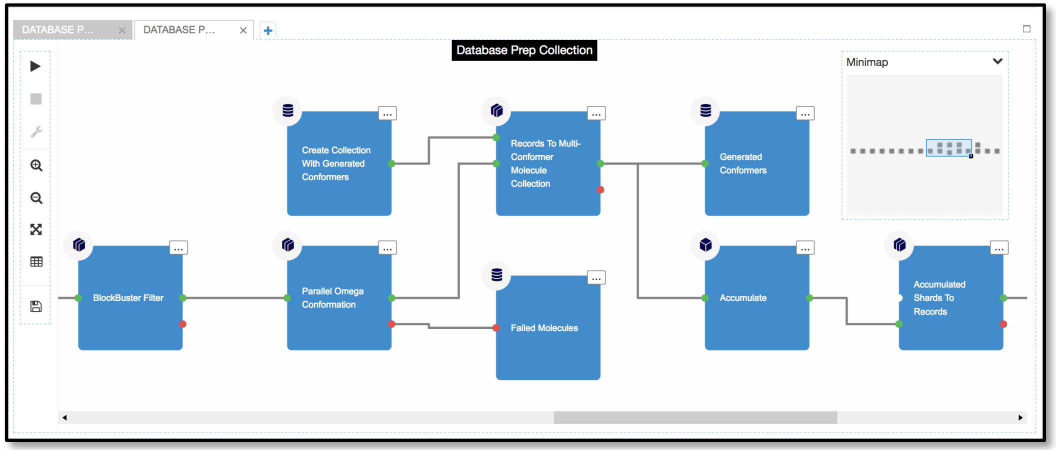
- Independent workers (Cube)
- Directed graph of workers (Floe)
- Good fit for the cloud
  - Inherent parallelism
  - Independent, reusable components
  - Easy to program

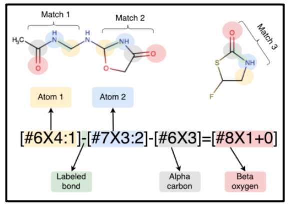




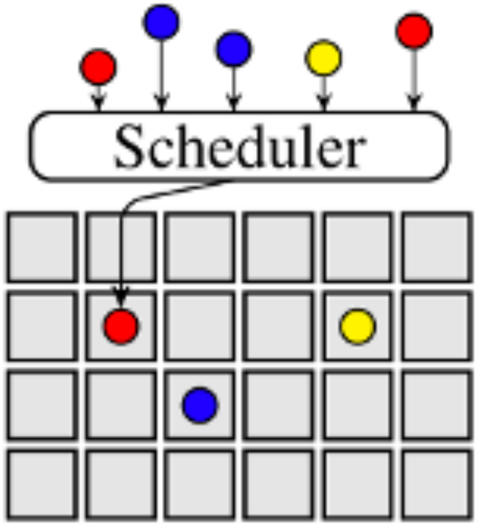
Open Programming Platform

- Cubes are Python
- Floes assembled in Python or UI
- Wrapped binaries
- Third-party code
  - PSI4, OpenMM, GROMACS, etc.
  - Scientific Python ecosystem
  - Commercial Code*

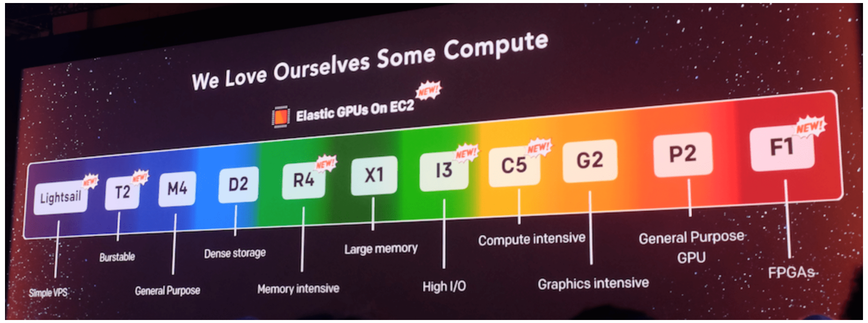

DATABASE P...
DATABASE P...
Database Prep Collection
Minimap
Create Collection With Generated Conformers
Records To Multi-Conformer Molecule Collection
Generated Conformers
BlockBuster Filter
Parallel Omega Conformation
Failed Molecules
Accumulate
Accumulated Shards To Records



Orion Scheduler



- Decides when/where everything in Orion runs
    - Manages pools of instance types
    - Fairness determined by task slowdown due to sharing
    - Dominant Resource Factor instance partitioning



We Love Ourselves Some Compute

Elastic GPUs On EC2 NEW!

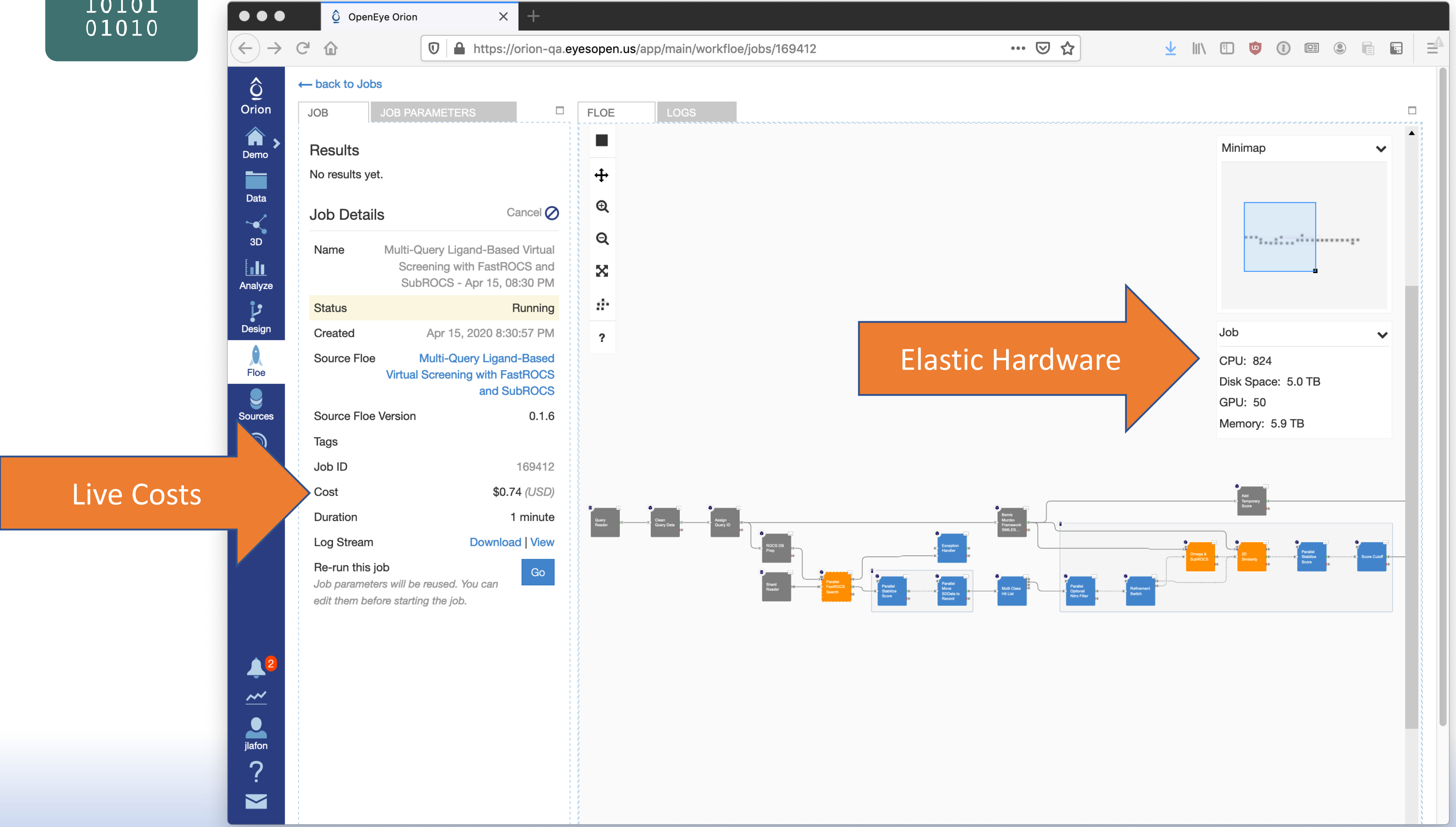
Lightsail NEW! T2 NEW! M4 D2 R4 NEW! X1 I3 NEW! C5 NEW! G2 P2 F1 NEW!

Simple VPS, Burstable, General Purpose, Dense storage, Memory intensive, Large memory, High I/O, Compute intensive, Graphics intensive, General Purpose GPU, FPGAs

- Per-Cube
    - Hardware requirements
    - Scaling parameters
    - Spot Preference

- Cyclic workflows
- Mixed software environments

Running Floes in Orion



← back to Jobs

JOB | JOB PARAMETERS

Results

No results yet.

Job Details — Cancel

Name	Multi-Query Ligand-Based Virtual Screening with FastROCS and SubROCS - Apr 15, 08:30 PM
Status	Running
Created	Apr 15, 2020 8:30:57 PM
Source Floe	Multi-Query Ligand-Based Virtual Screening with FastROCS and SubROCS
Source Floe Version	0.1.6
Tags	
Job ID	169412
Cost	$0.74 *(USD)*
Duration	1 minute
Log Stream	Download | View

Re-run this job — Go

Job parameters will be reused. You can edit them before starting the job.

Job

CPU: 824

Disk Space: 5.0 TB

GPU: 50

Memory: 5.9 TB

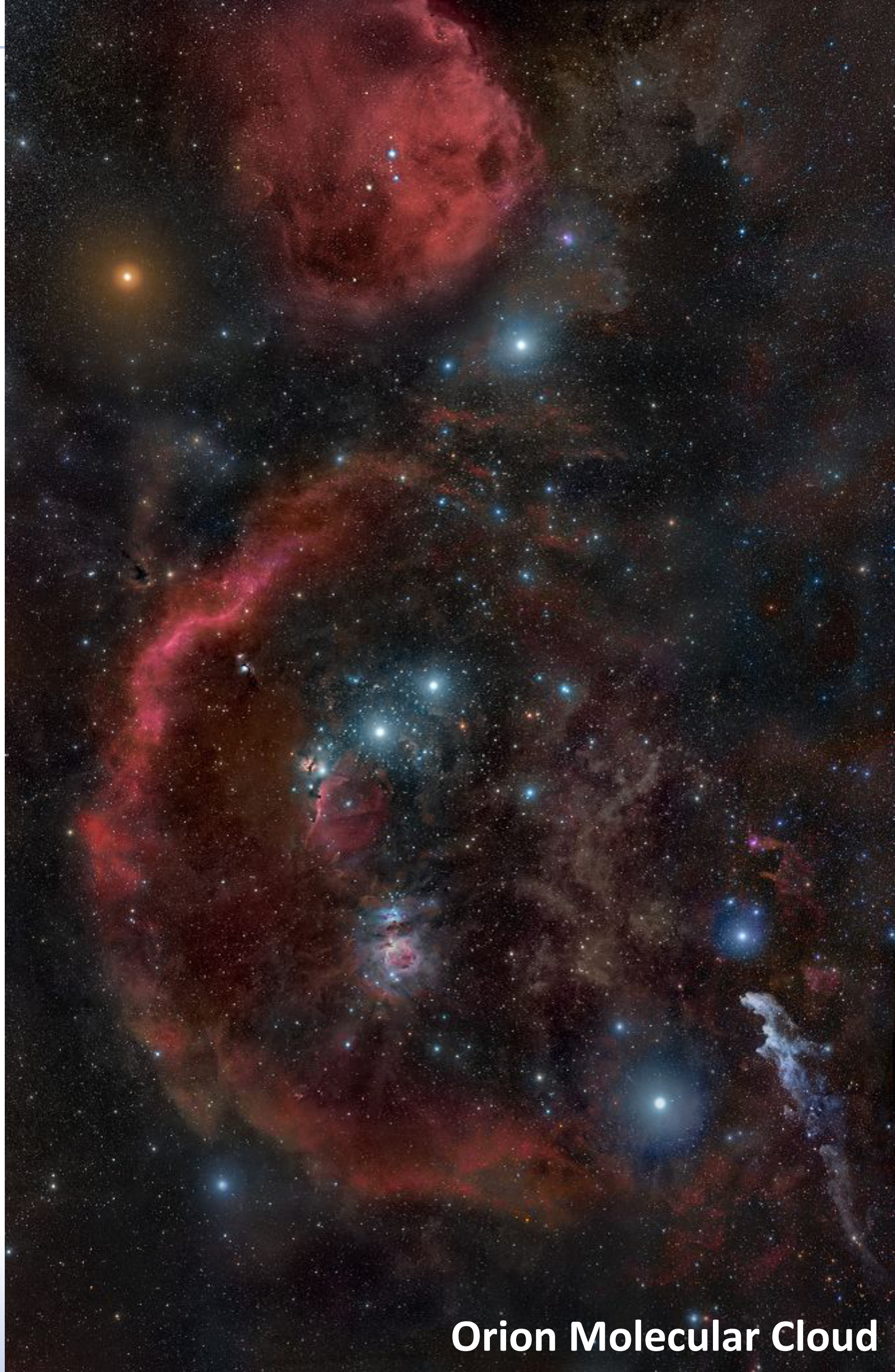
Elastic Hardware

Live Costs

Agenda

- Intro to OpenEye and Orion

- HPC Science in Orion

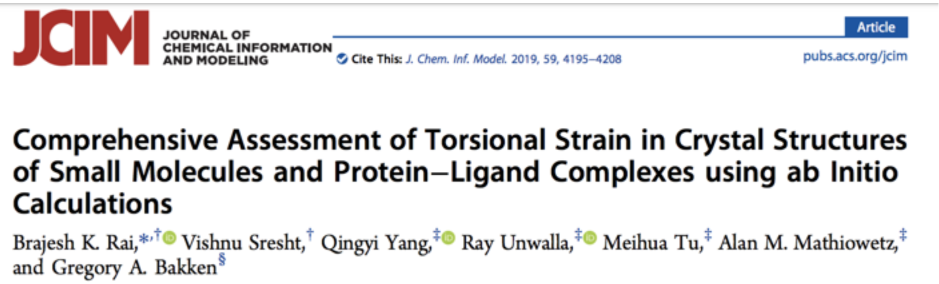


Orion Molecular Cloud

orion Launch Partner projects

Pfizer

1. QM torsion scan
2. Corporate collection
3. 184K CPUs at peak, 2 months to complete
4. ML force-field



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

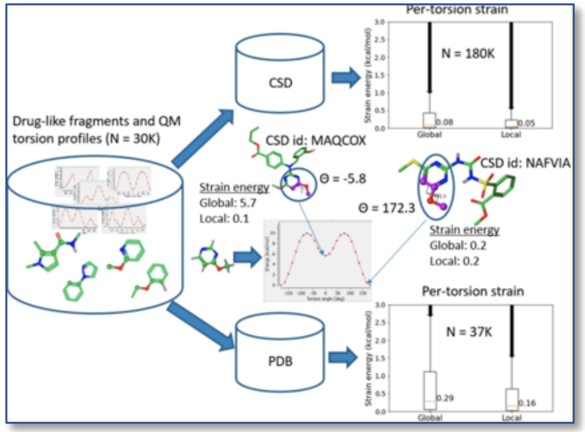
Article

Cite This: *J. Chem. Inf. Model.* 2019, 59, 4195−4208

pubs.acs.org/jcim

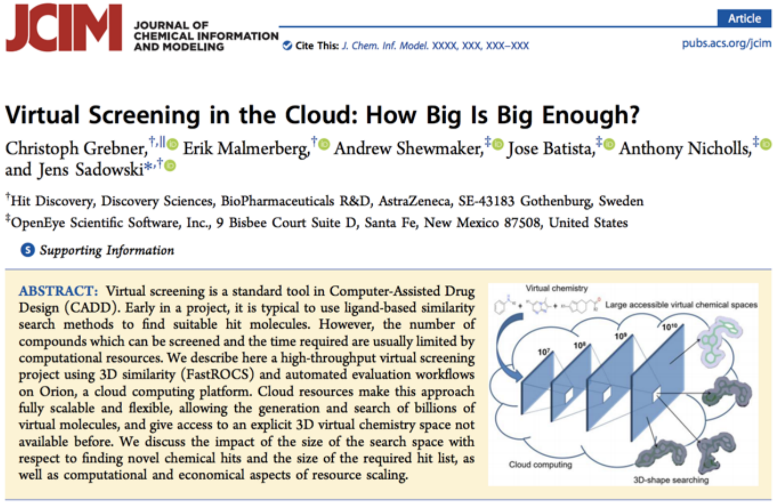
Comprehensive Assessment of Torsional Strain in Crystal Structures of Small Molecules and Protein−Ligand Complexes using ab Initio Calculations

Brajesh K. Rai,[*,†] Vishnu Sresht,[†] Qingyi Yang,[‡] Ray Unwalla,[‡] Meihua Tu,[‡] Alan M. Mathiowetz,[‡] and Gregory A. Bakken[§]



AstraZeneca

1. 12.7 Billion compounds
2. 3D FastROCS on GPUs in Orion
3. Impact on multiple projects



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

pubs.acs.org/jcim

Virtual Screening in the Cloud: How Big Is Big Enough?

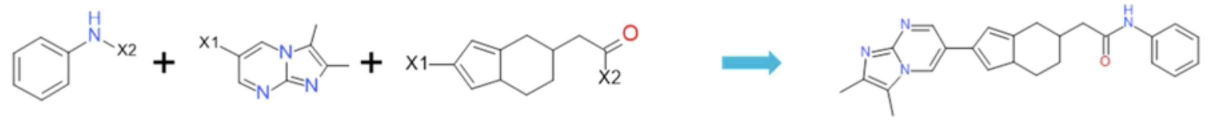
Christoph Grebner,[†,∥] Erik Malmerberg,[†] Andrew Shewmaker,[‡] Jose Batista,[‡] Anthony Nicholls,[‡] and Jens Sadowski[*,†]

[†]Hit Discovery, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, SE-43183 Gothenburg, Sweden
[‡]OpenEye Scientific Software, Inc., 9 Bisbee Court Suite D, Santa Fe, New Mexico 87508, United States

Supporting Information

ABSTRACT: Virtual screening is a standard tool in Computer-Assisted Drug Design (CADD). Early in a project, it is typical to use ligand-based similarity search methods to find suitable hit molecules. However, the number of compounds which can be screened and the time required are usually limited by computational resources. We describe here a high-throughput virtual screening project using 3D similarity (FastROCS) and automated evaluation workflows on Orion, a cloud computing platform. Cloud resources make this approach fully scalable and flexible, allowing the generation and search of billions of virtual molecules, and give access to an explicit 3D virtual chemistry space not available before. We discuss the impact of the size of the search space with respect to finding novel chemical hits and the size of the required hit list, as well as computational and economical aspects of resource scaling.

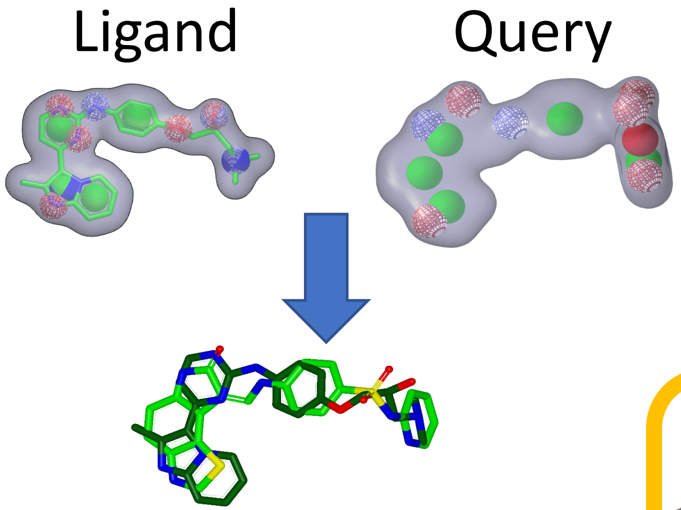
Scheme 1. Scheme for Enumerating Virtual Molecules



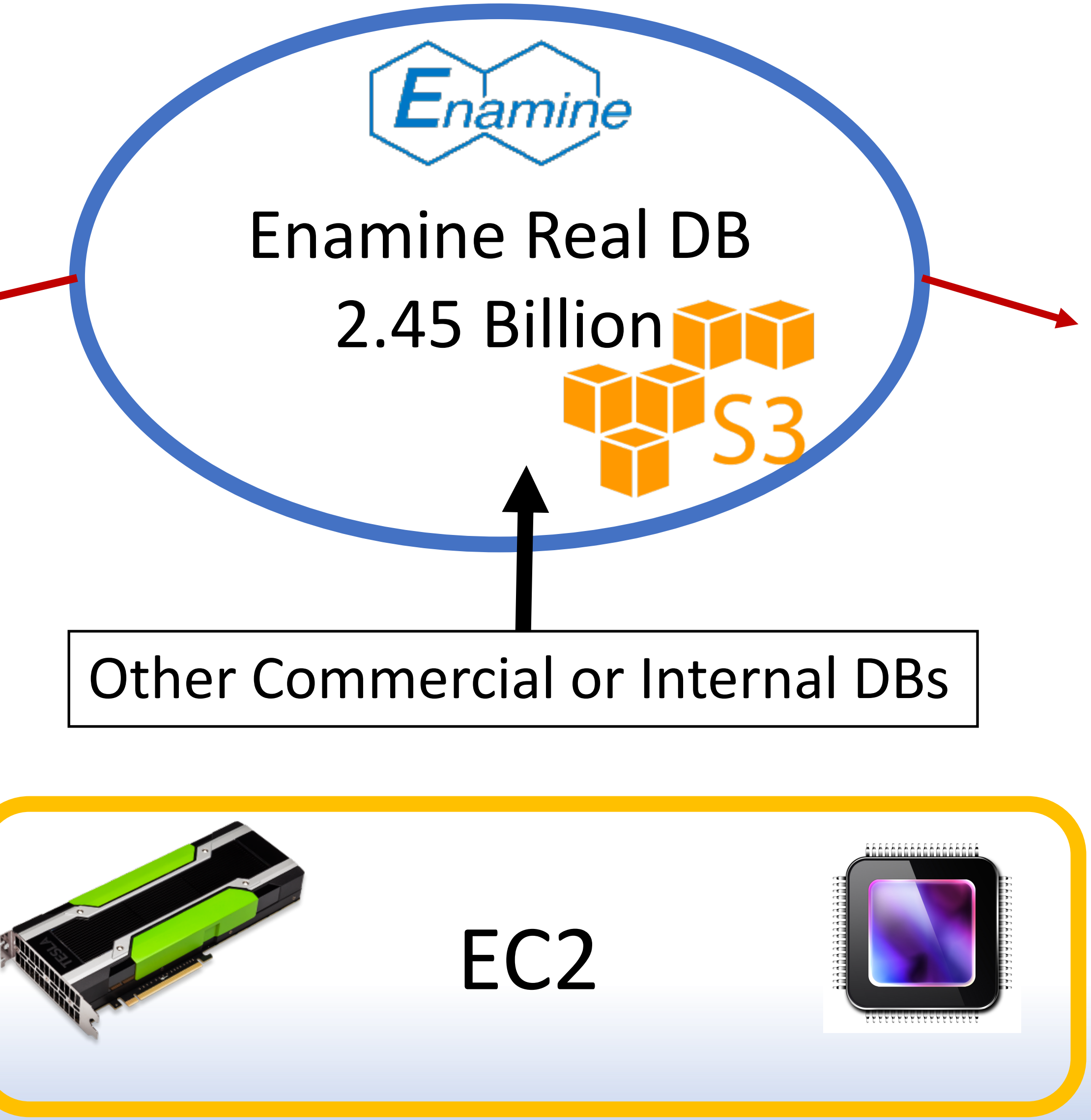
2.45 Billion Molecule Virtual Screens

<u>FastROCS</u>

- 3D Ligand Based
- Runtime ~ 30min
- Cost: **$50-$200**



3D Tanimoto Similarity



Enamine Real DB

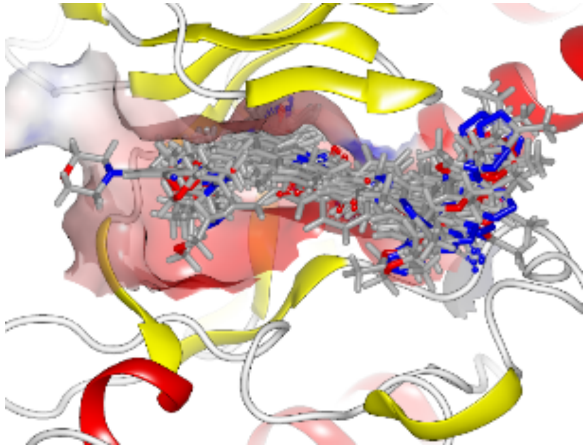
2.45 Billion

Other Commercial or Internal DBs

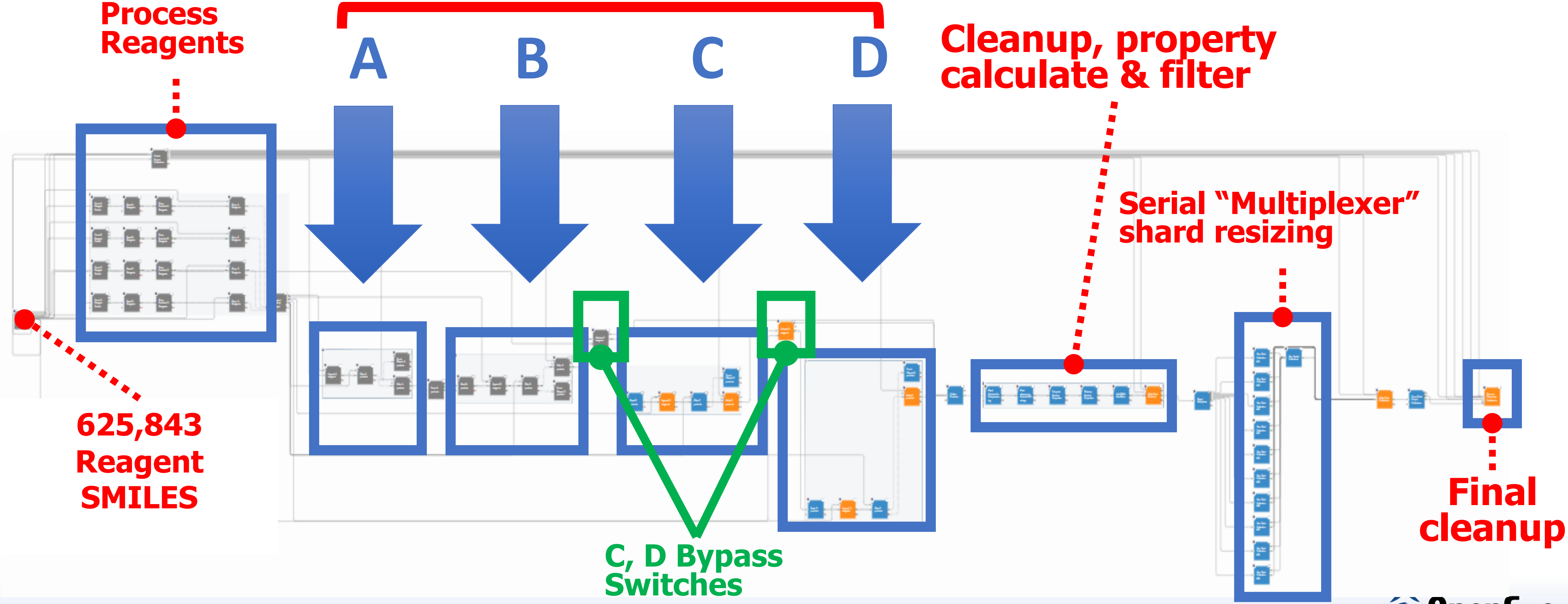
EC2

<u>Docking</u>

- Structure Based
- Runtime ~24 Hours
- Cost : **$10k – $50k**



Enamine REAL Space: 13.3 Billion SMILES



Reaction Components
(124 SMIRKS)

Process
Reagents

A

B

C

D

Cleanup, property
calculate & filter

Serial "Multiplexer"
shard resizing

625,843
Reagent
SMILES

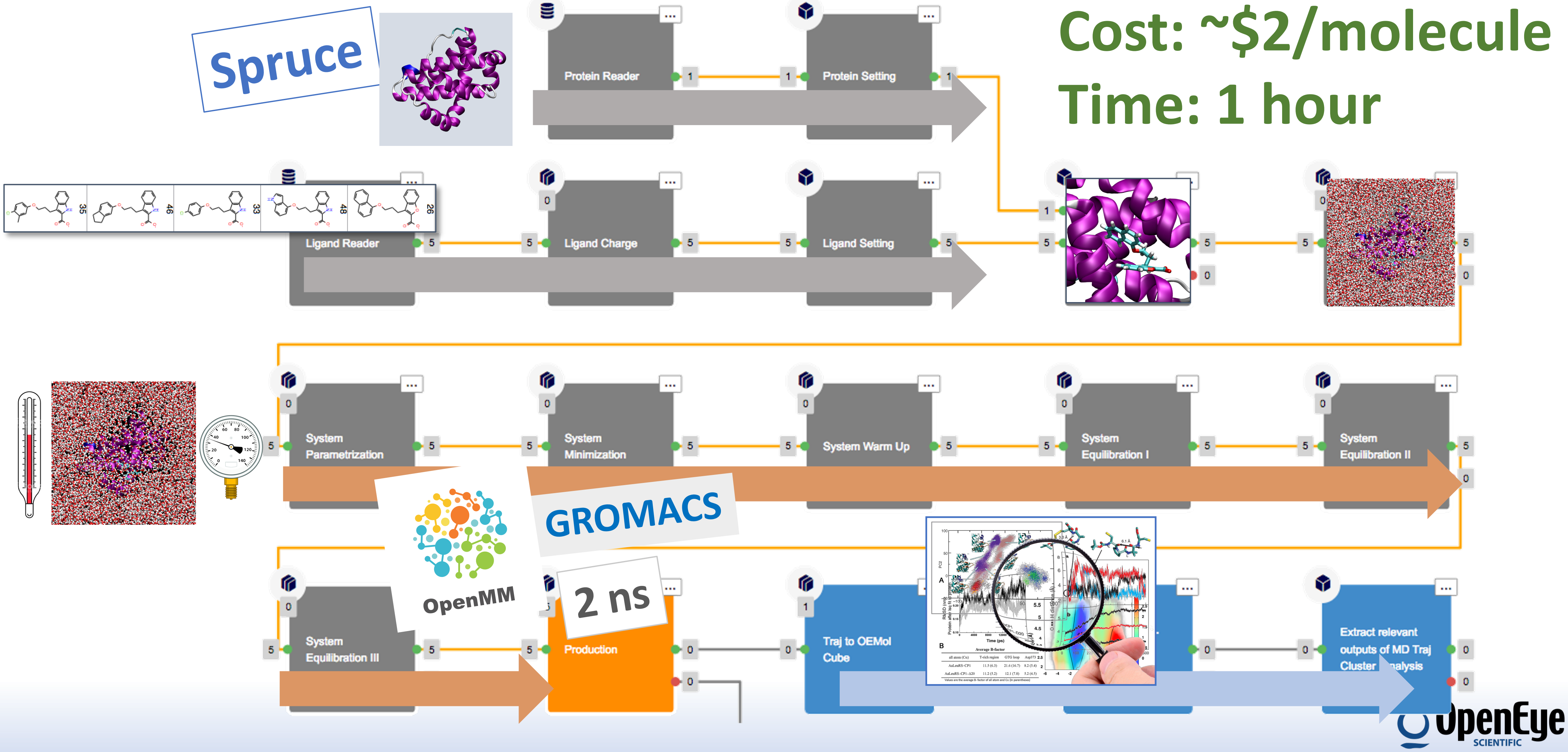
C, D Bypass
Switches

Final
cleanup

Short Trajectory Molecular Dynamics Floe

Cost: ~$2/molecule

Time: 1 hour



Spruce

Protein Reader

Protein Setting

Ligand Reader

Ligand Charge

Ligand Setting

System Parametrization

System Minimization

System Warm Up

System Equilibration I

System Equilibration II

GROMACS

OpenMM

2 ns

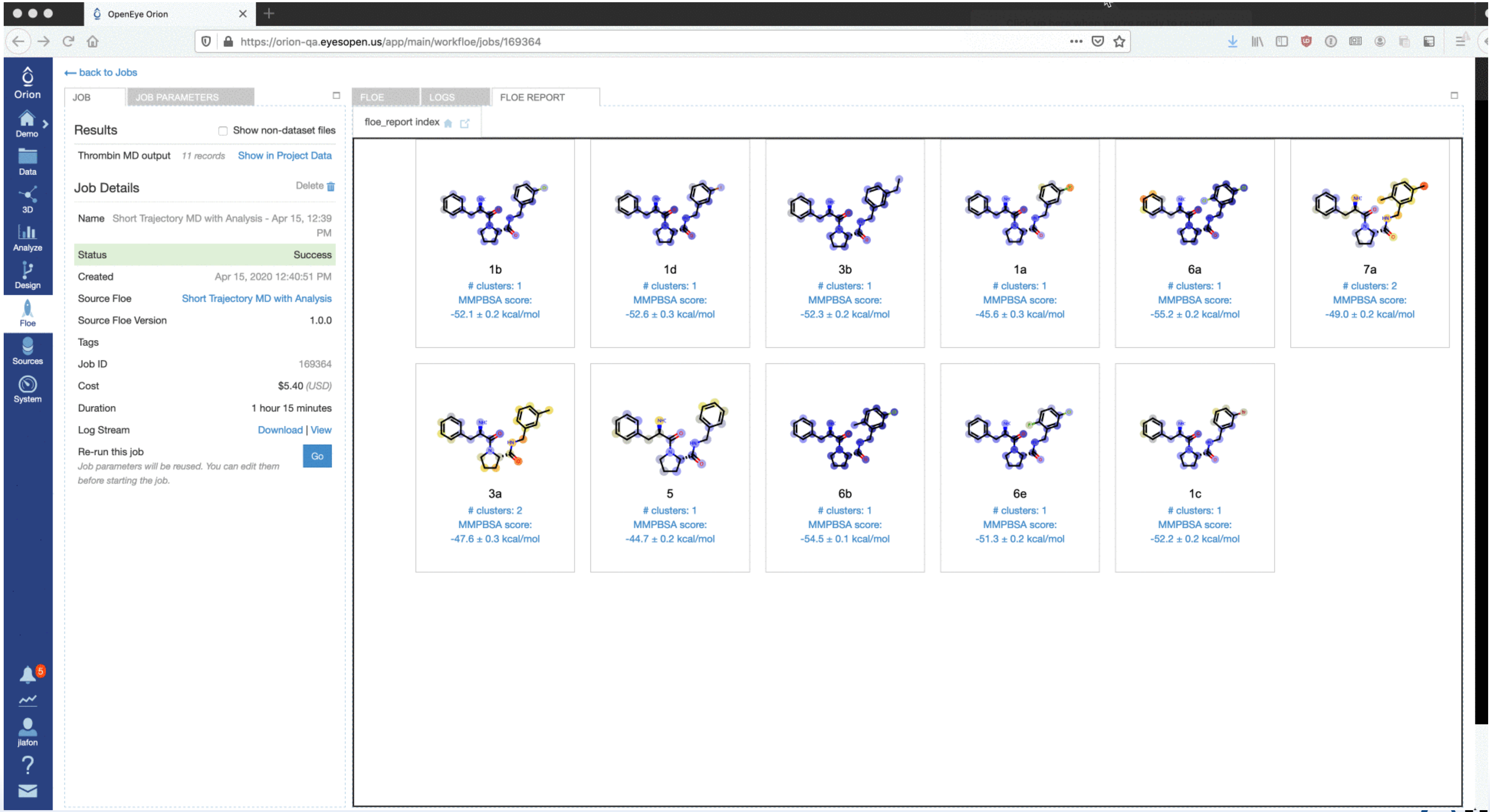
System Equilibration III

Production

Traj to OEMol Cube

Extract relevant outputs of MD Traj Cluster analysis

Short Trajectory Molecular Dynamics Report



Results □ Show non-dataset files

Thrombin MD output *11 records* Show in Project Data

Job Details Delete

Name	Short Trajectory MD with Analysis - Apr 15, 12:39 PM
Status	Success
Created	Apr 15, 2020 12:40:51 PM
Source Floe	Short Trajectory MD with Analysis
Source Floe Version	1.0.0
Tags	
Job ID	169364
Cost	$5.40 *(USD)*
Duration	1 hour 15 minutes
Log Stream	Download | View
Re-run this job	Go

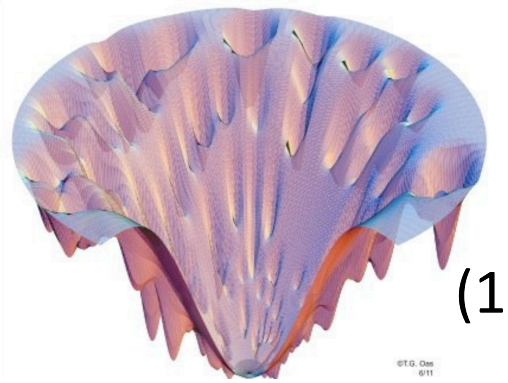
Job parameters will be reused. You can edit them before starting the job.

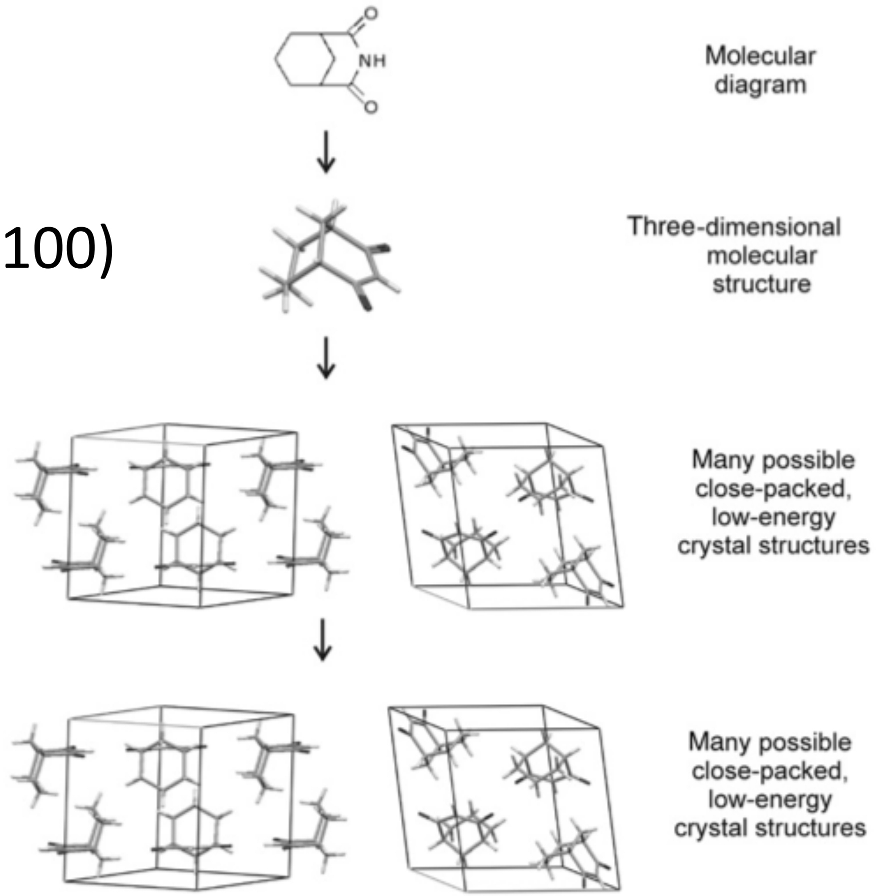
Ligand	# clusters	MMPBSA score
1b	1	-52.1 ± 0.2 kcal/mol
1d	1	-52.6 ± 0.3 kcal/mol
3b	1	-52.3 ± 0.2 kcal/mol
1a	1	-45.6 ± 0.3 kcal/mol
6a	1	-55.2 ± 0.2 kcal/mol
7a	2	-49.0 ± 0.2 kcal/mol
3a	2	-47.6 ± 0.3 kcal/mol
5	1	-44.7 ± 0.2 kcal/mol
6b	1	-54.5 ± 0.1 kcal/mol
6e	1	-51.3 ± 0.2 kcal/mol
1c	1	-52.2 ± 0.2 kcal/mol

Small-molecule Crystal Structure Prediction

- Experimental crystal structures are obtained at the formulation stage
- Much effort and money would have been spent on optimizing the compound
- Not all polymorphs are realized in experiments
- Discovery of a new polymorph after formulation and marketing can result in recall and significant losses

Typical physics-based modeling workflow of CSP




NH
O
O
Molecular diagram
Three-dimensional molecular structure
Many possible close-packed, low-energy crystal structures
Many possible close-packed, low-energy crystal structures

Start with 2D structure

(10-100)

Conformer generation

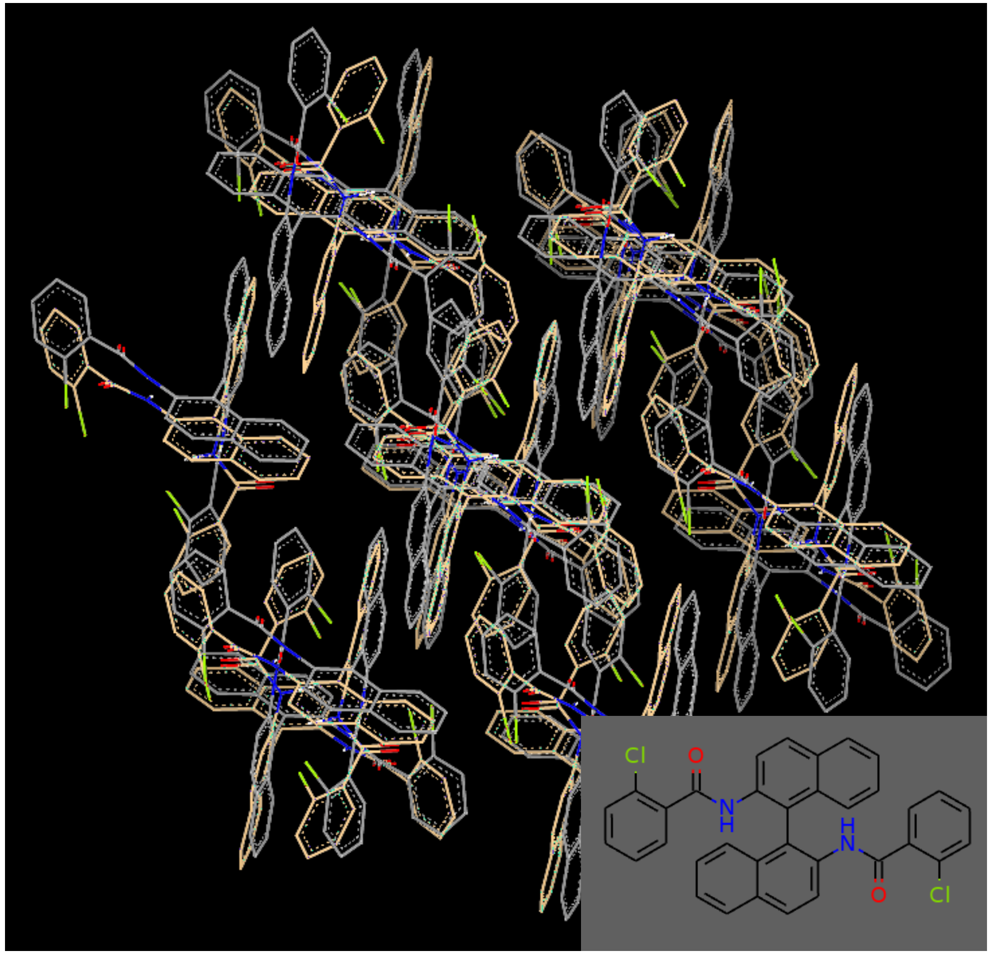
(10^5-10^6)

Rigid packing -> FF minimization -> FF ranking

(100)

Re-rank top FF packings using Quantum Mechanics

Example: CSP6 XXVI

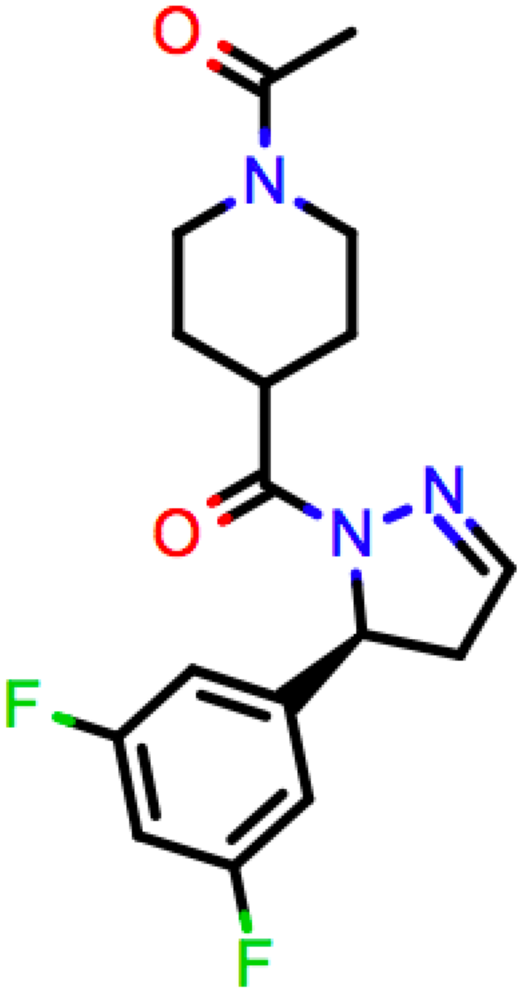


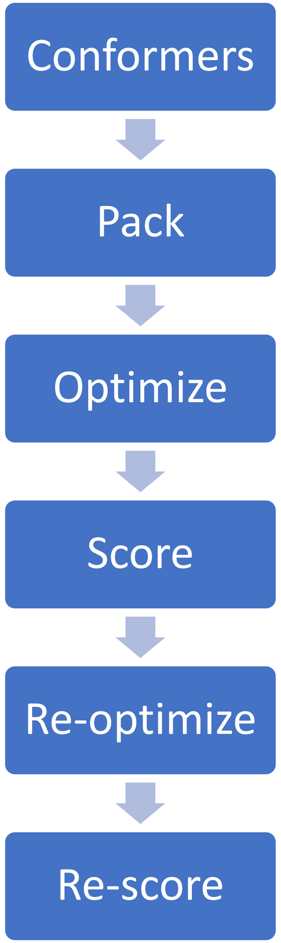
- Plane-wave DFT took ~1 month
- Dimer-expansion takes ~14 hours on Orion
  - 100 single point QM Lattice Energy calculations
  - 4838 dimer calculations (80 heavy atoms per dimer)

Performance Summary

Wall Time, hours	Total CPU hours	Peak CPUs	Peak RAM
14	~6500	4000	27.5 TB

OE-Collaborator Blind Prediction Challenge





Conformers

Pack

Optimize

Score

Re-optimize

Re-score

200

800,000

IEFF optimization

Rank = 21, RMS_20 = 1.1 A

Top 100 IEFF packings
IEFF@LR + DFT@SR

Rank = 1, RMS_20 = 0.18 A

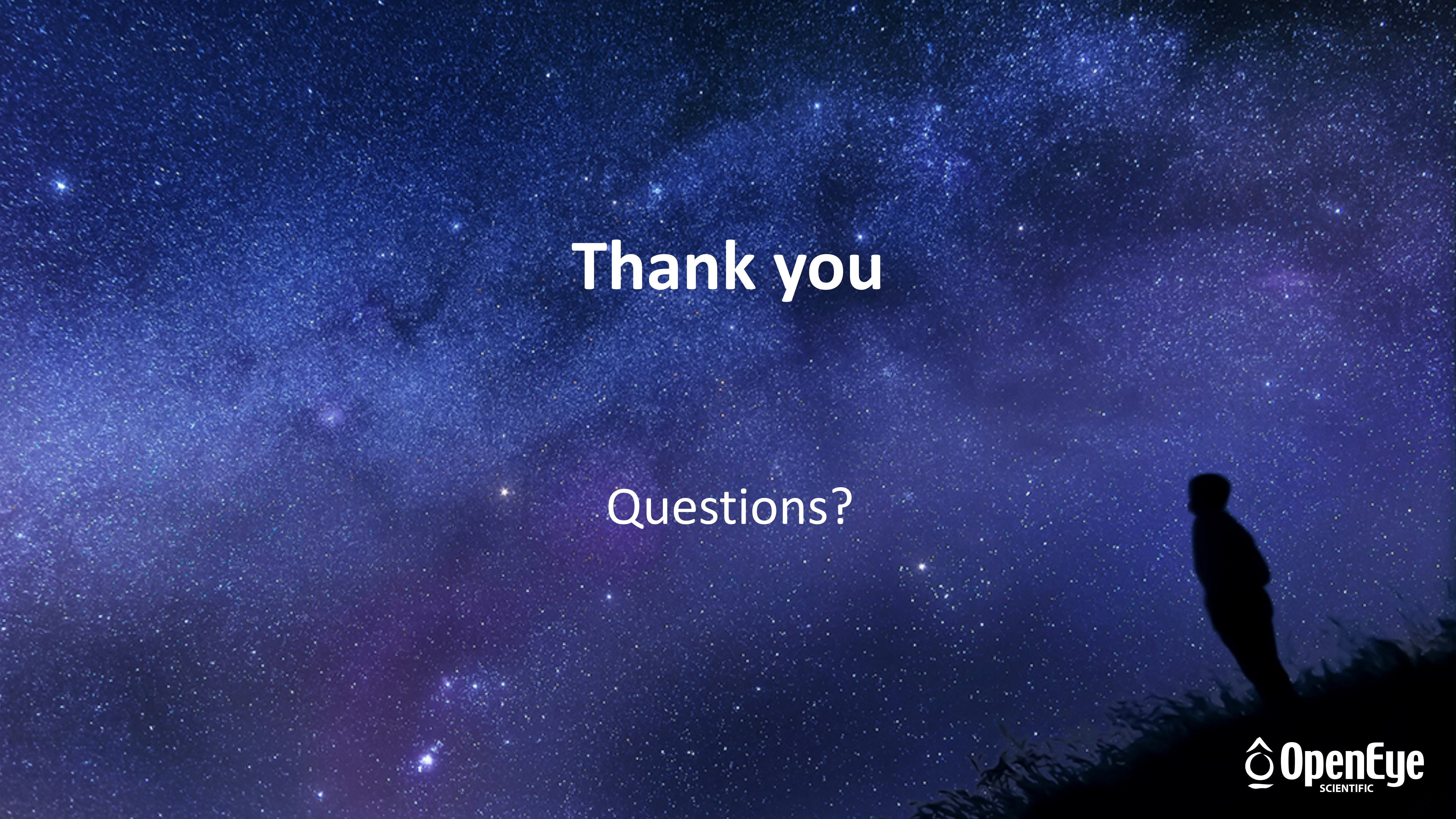
290K CPU hours
24 hours wall time
~12000x speed-up

Cost: $5k-$50k

Time: 1 day

Making Large Calculations In The Cloud Accessible Is Difficult, But Worth It

- Orion is an ambitious project to create a complete cloud-native platform for Computer Aided Drug Design
- Orion is driving new science at OpenEye and our partners
- Scalable calculations are stable, automated, and reproducible
  - Large-scale calculations become routine
- Open development environment
  - Opportunities to integrate your code, 3rd party code, on-prem services



Thank you

Questions?

OpenEye SCIENTIFIC

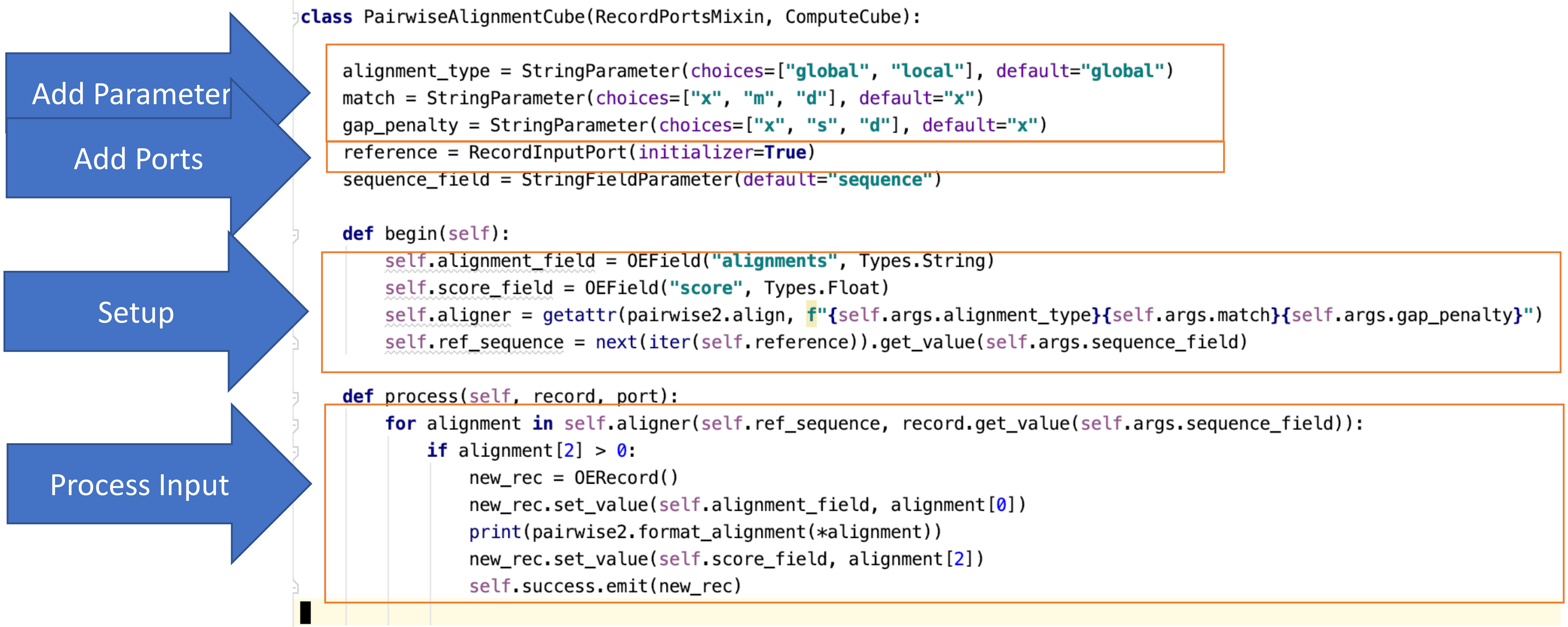
For more information, please contact:

info@eyesopen.com

www.eyesopen.com

+1-505-473-7385

Cube Example: Sequence alignment (BioPython)



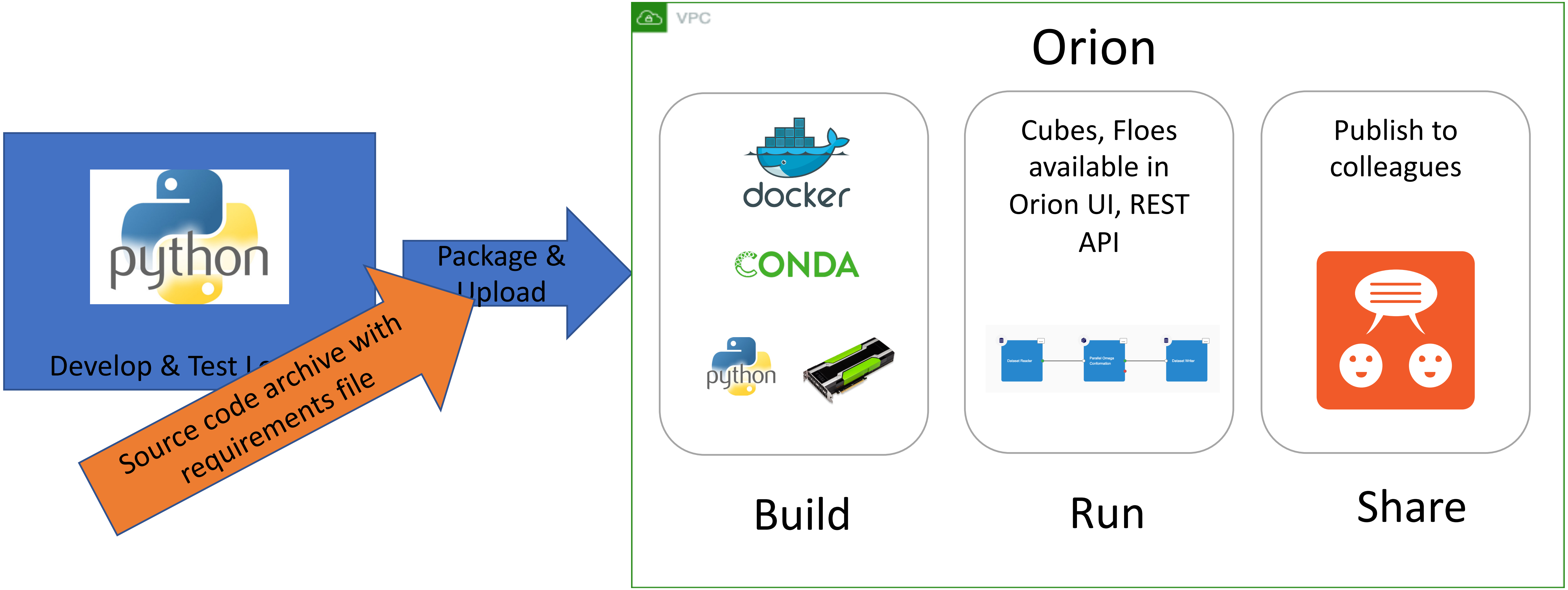
```python
class PairwiseAlignmentCube(RecordPortsMixin, ComputeCube):

    alignment_type = StringParameter(choices=["global", "local"], default="global")
    match = StringParameter(choices=["x", "m", "d"], default="x")
    gap_penalty = StringParameter(choices=["x", "s", "d"], default="x")
    reference = RecordInputPort(initializer=True)
    sequence_field = StringFieldParameter(default="sequence")

    def begin(self):
        self.alignment_field = OEField("alignments", Types.String)
        self.score_field = OEField("score", Types.Float)
        self.aligner = getattr(pairwise2.align, f"{self.args.alignment_type}{self.args.match}{self.args.gap_penalty}")
        self.ref_sequence = next(iter(self.reference)).get_value(self.args.sequence_field)

    def process(self, record, port):
        for alignment in self.aligner(self.ref_sequence, record.get_value(self.args.sequence_field)):
            if alignment[2] > 0:
                new_rec = OERecord()
                new_rec.set_value(self.alignment_field, alignment[0])
                print(pairwise2.format_alignment(*alignment))
                new_rec.set_value(self.score_field, alignment[2])
                self.success.emit(new_rec)
```

Orion Development Lifecycle



VPC

Orion

Develop & Test L

Source code archive with requirements file

Package & Upload

docker

CONDA

python

Cubes, Floes available in Orion UI, REST API

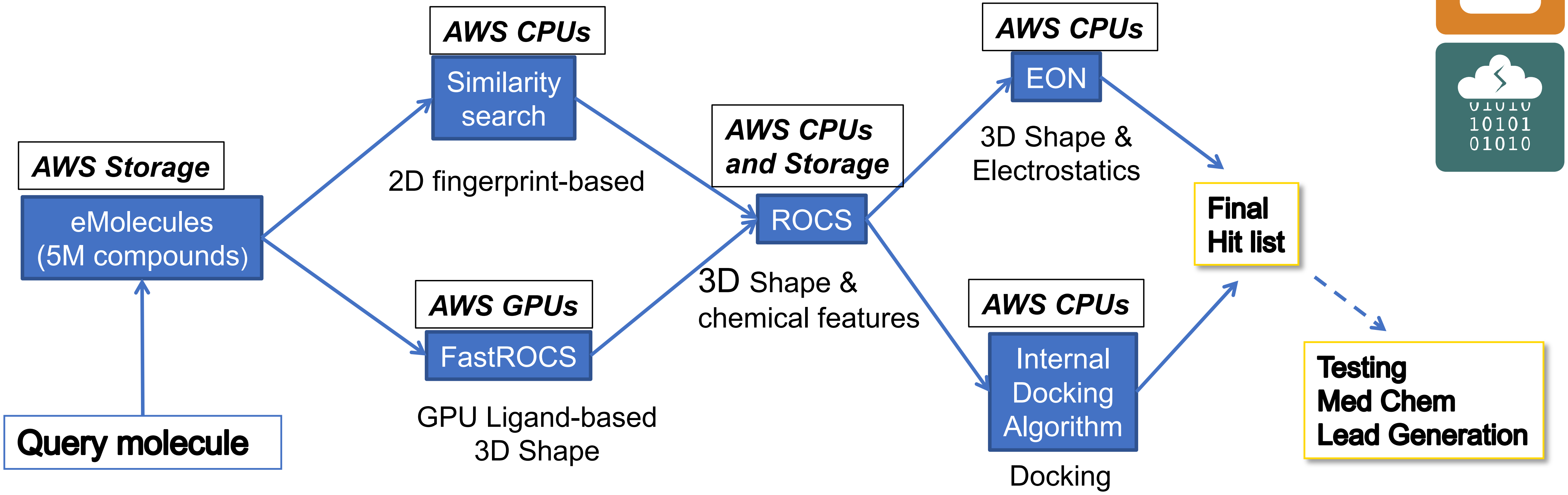
Publish to colleagues

Build

Run

Share

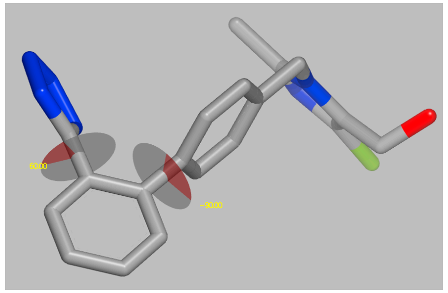
Conventional virtual screening deployed in Orion

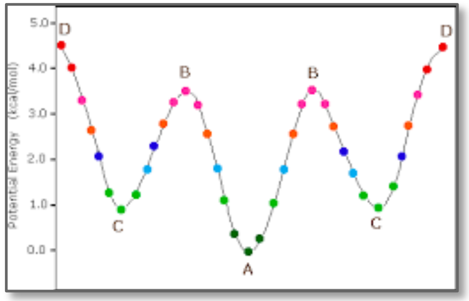

AWS Storage
eMolecules (5M compounds)
Query molecule
AWS CPUs
Similarity search
2D fingerprint-based
AWS GPUs
FastROCS
GPU Ligand-based 3D Shape
AWS CPUs and Storage
ROCS
3D Shape & chemical features
AWS CPUs
EON
3D Shape & Electrostatics
AWS CPUs
Internal Docking Algorithm
Docking
Final Hit list
Testing Med Chem Lead Generation

- **Before Orion: 5-7 days**
  - Very manual and time consuming (both setup and calculation)

- **After Orion: *15 minutes- 1 hour***

Improving torsion potentials of corporate collection



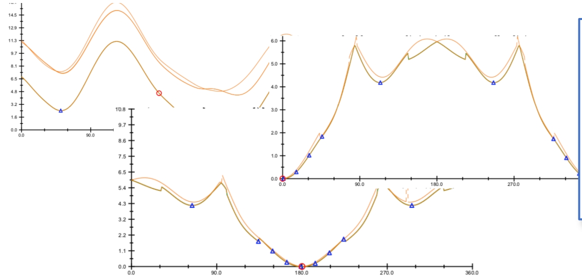


1.7M QM optimizations

Torsion Energy Profiles

5.5M Torsion Library

DFT energies



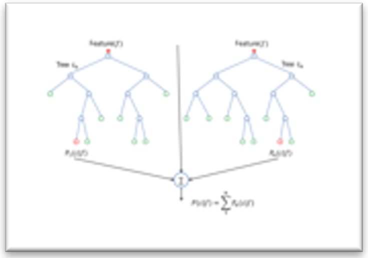
100k Torsions

PSI4
OPEN-SOURCE QUANTUM CHEMISTRY

Corporate Library

Generate fragments

Machine learning
Torsion energy lookup



QM-level
Molecular mechanics FF

aws **Peak = 184,000 CPUs (spot market rates)**

OpenEye SCIENTIFIC

Pfizer

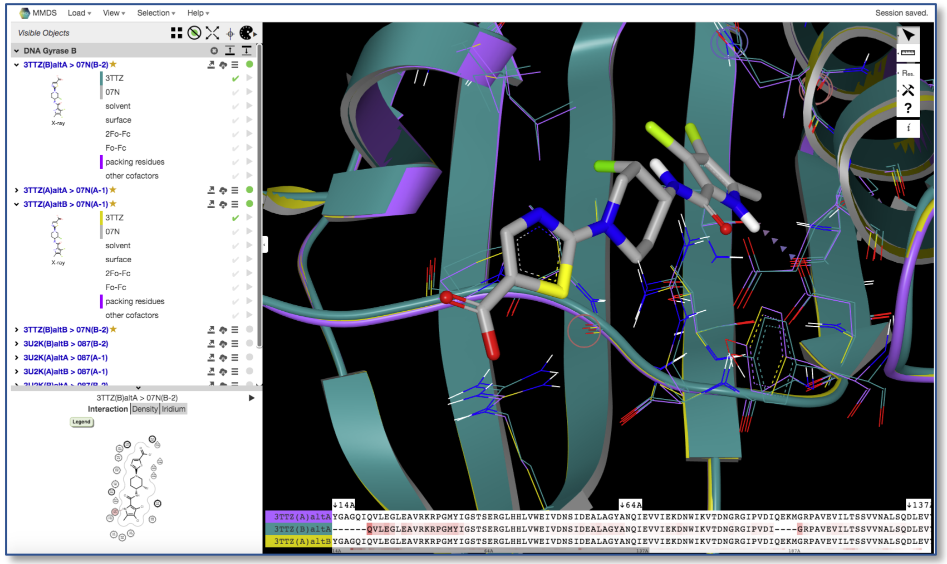
Pfizer take-home messages

- ML >75% of molecules MAE close to DFT
  - MMFF & OPLS < 50%
- CSD structures have very low median strain
  - 0.05 kCal/M/torsion → 0.25 kCal/M (4-6 rotatable bonds)
- PDB structures median strain
  - 0.3 kCal/M/torsion → 1.5 kCal/M
  - Better resolution structures have lower strain

Rai et al, CUP 2018

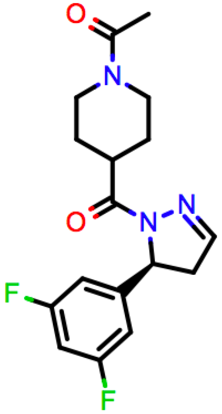
orion Partner projects 2

- Merck - MMDS
  - Protein-ligand structure database
  - Project-based
  - Evaluation of structure quality
  - Modeling-ready

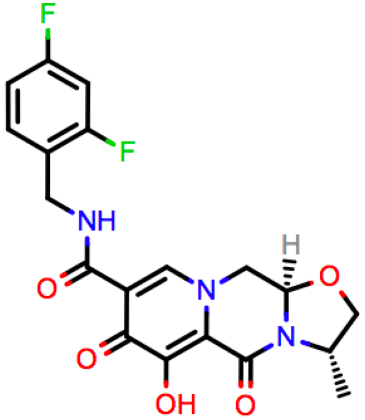


- GSK – CSP
  - Crystal from 2D
  - Crystal properties

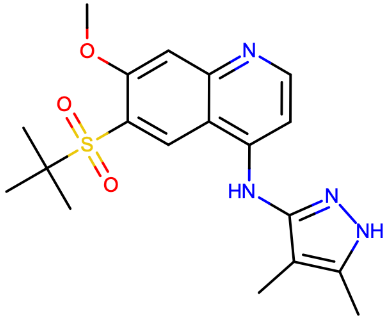
290K CPU hours
24 hours wall time
~12000x speed-up



Rank = 1, RMS_20 = 0.18 A



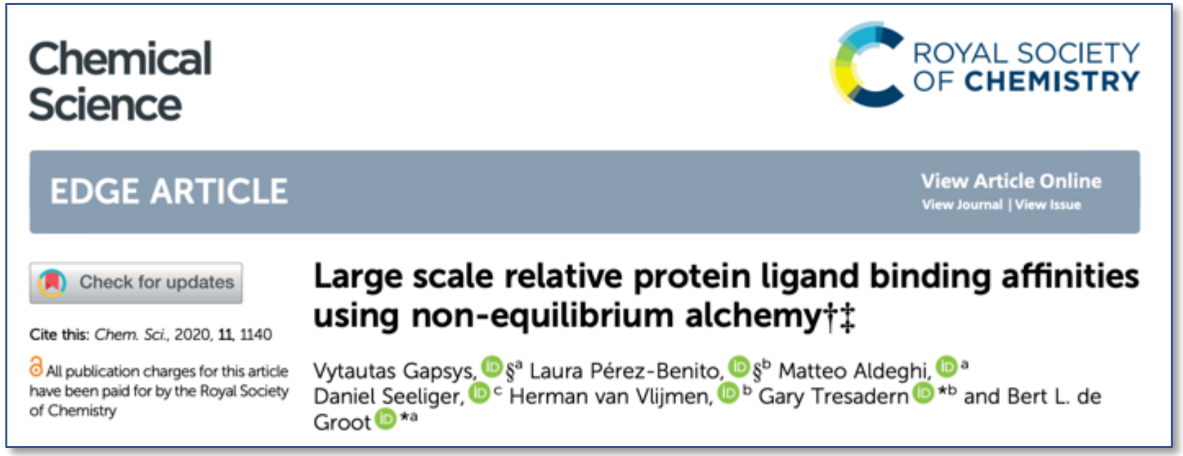
Rank = 1, RMS_20 = 0.18 A
RMS_50 = 0.26 A



Rank = 1, RMS_20 = 0.16 A

Partner projects 3

- D. Mobley, J. Chodera, B. de Groot
  1. Relative FEP
     - Non-equilibrium switching (NES)
  2. Molecular Dynamics
     - Basic dynamics with Analysis



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2020, 11, 1140

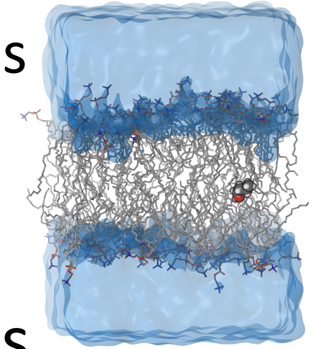
All publication charges for this article have been paid for by the Royal Society of Chemistry

Large scale relative protein ligand binding affinities using non-equilibrium alchemy†‡

Vytautas Gapsys,§[a] Laura Pérez-Benito,§[b] Matteo Aldeghi,[a] Daniel Seeliger,[c] Herman van Vlijmen,[b] Gary Tresadern*[b] and Bert L. de Groot*[a]

- **Weighted-Ensemble Methods**
  - WESTPA, Lillian Chong, U. Pitt.
  - Path sampling approach

- Multiple trajectory walkers in parallel
  - Replicate walkers with forward progress
- Yields *unbiased* pathways
- Rigorous weighting of trajectories enables calculations of rate constants



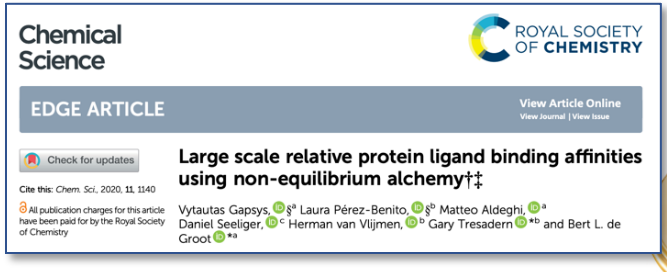
WE review, Zuckerman & Chong, *Ann. Rev. Biophys.* 2017

Binding Free-Energy

orion

- Competitive method
  - Relative vs Absolute
  - Free-energy perturbation, Thermodynamic Integration

- Non-equilibrium switching
  - Fundamentally new BFE method
  - Many small calculations



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2020, **11**, 1140

All publication charges for this article have been paid for by the Royal Society of Chemistry

Large scale relative protein ligand binding affinities using non-equilibrium alchemy†‡

Vytautas Gapsys, §[a] Laura Pérez-Benito, §[b] Matteo Aldeghi, [a] Daniel Seeliger, [c] Herman van Vlijmen, [b] Gary Tresadern *[b] and Bert L. de Groot *[a]

www.eyesopen.com/blog/relative-binding-free-energy-calculations-with-non-equilibrium-switching-in-orion

OpenEye SCIENTIFIC

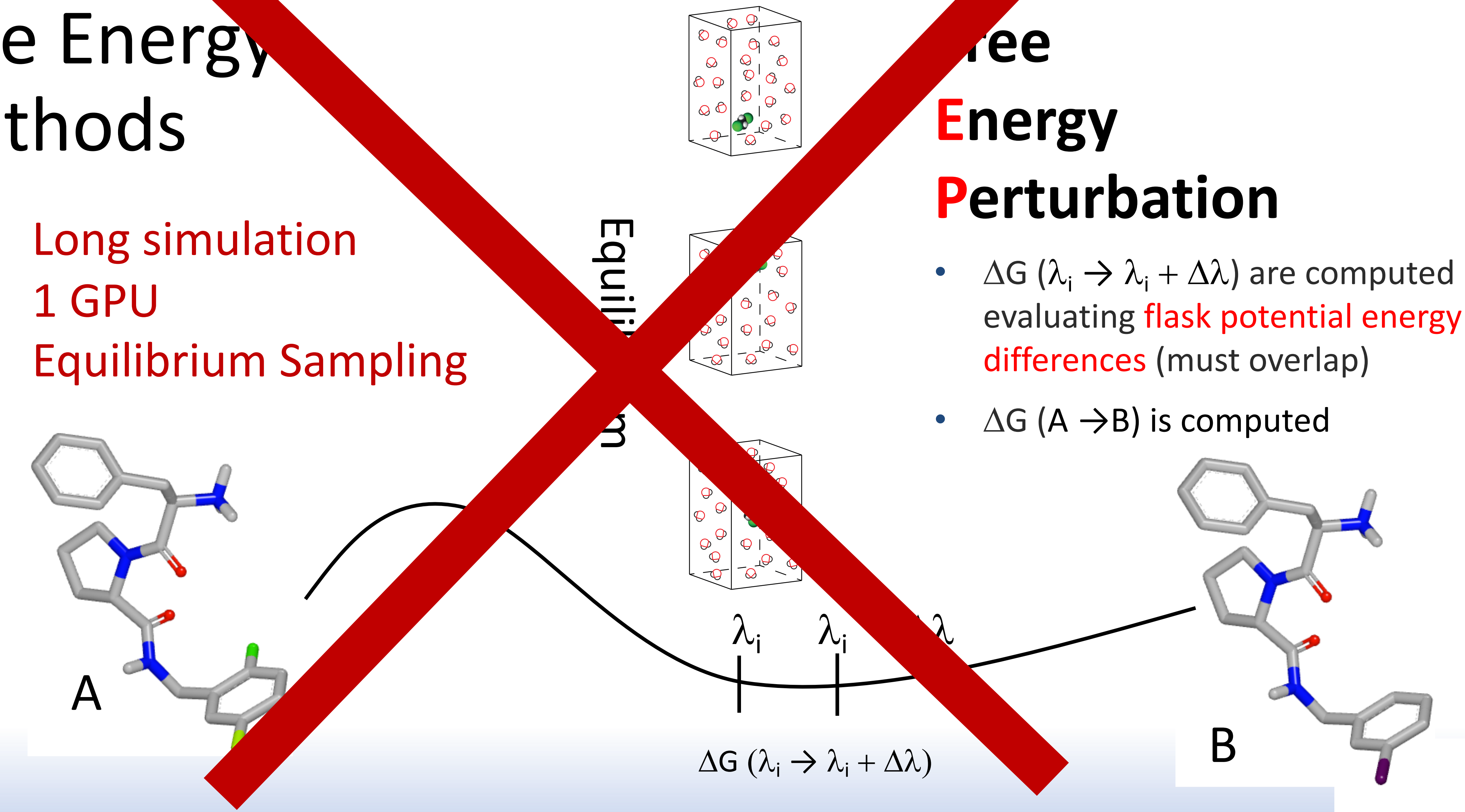
Relative Binding Free Energy Methods

Long simulation
1 GPU
Equilibrium Sampling

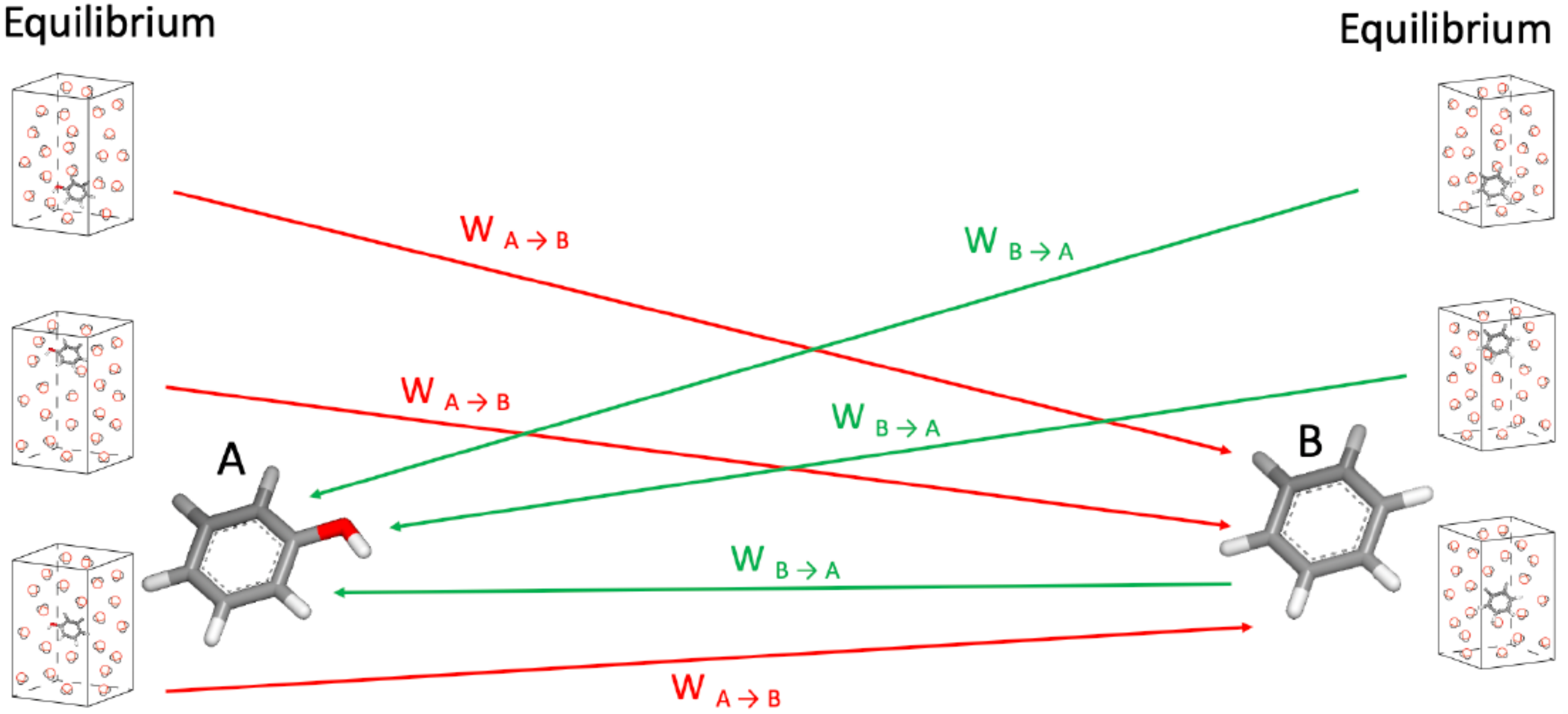
FEP

Free
Energy
Perturbation

- ΔG ($\lambda_i \rightarrow \lambda_i + \Delta\lambda$) are computed evaluating flask potential energy differences (must overlap)
- ΔG (A → B) is computed


Equilibrium
A
λ_i
λ_i
λ
ΔG ($\lambda_i \rightarrow \lambda_i + \Delta\lambda$)
B

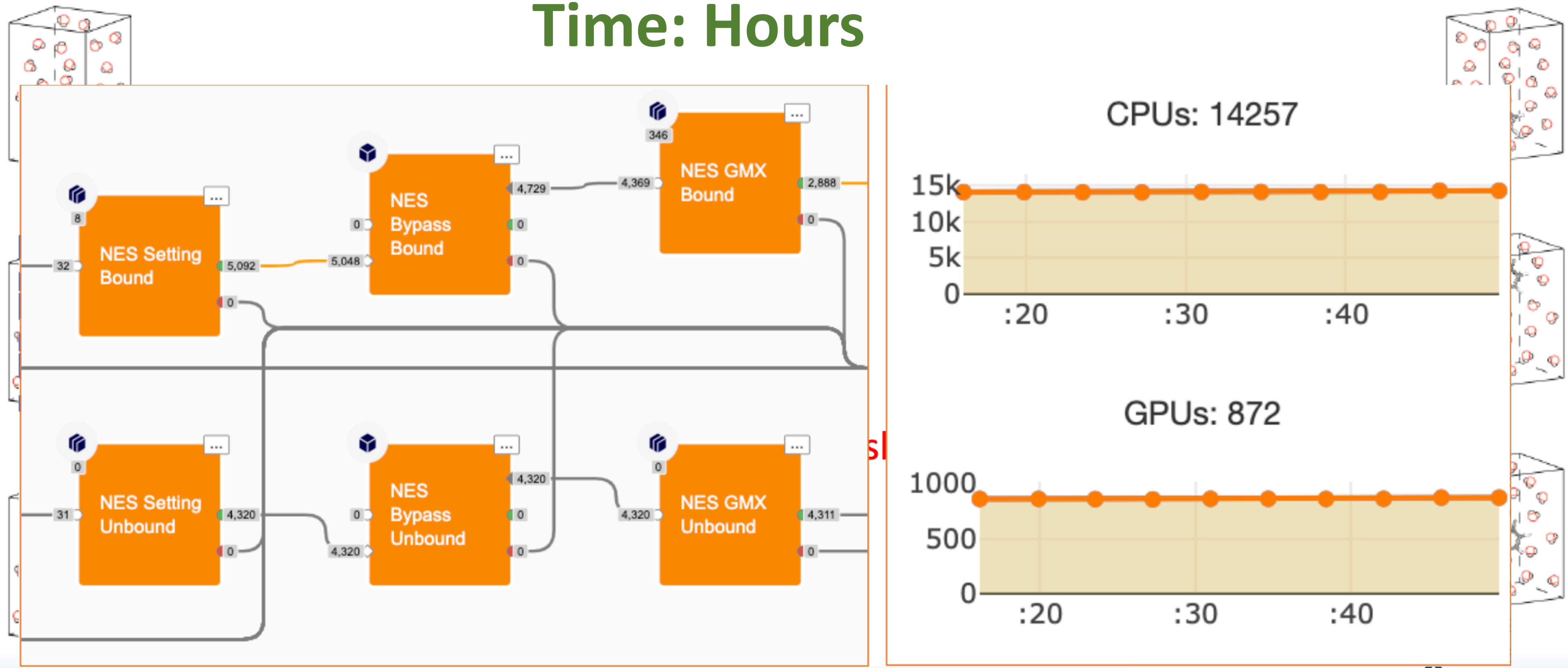
NES: Widely Parallelizable



NES: Widely Parallelizable

Cost: ~$200/comparison

Time: Hours



11 ligands, 16 edges