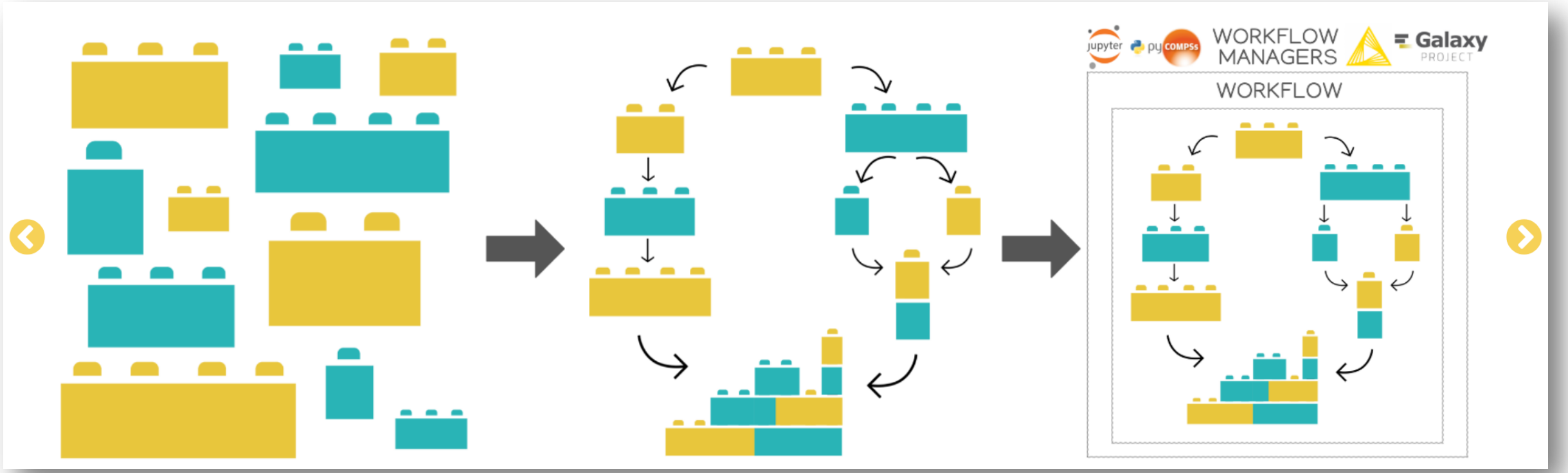
bioexcel
Centre of Excellence for Computational Biomolecular Research

Horizon 2020
European Union Funding
for Research & Innovation

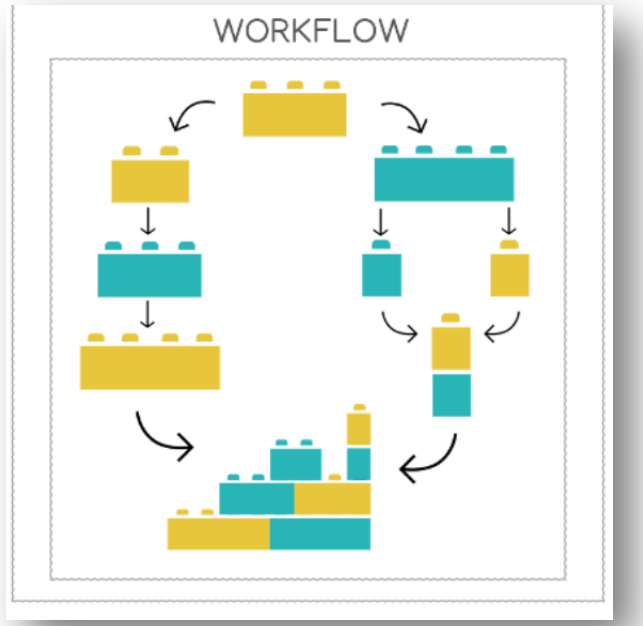
BioExcel Building Blocks (BioBB) and HPC

credit: Adam Hospital (IRB)

BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows



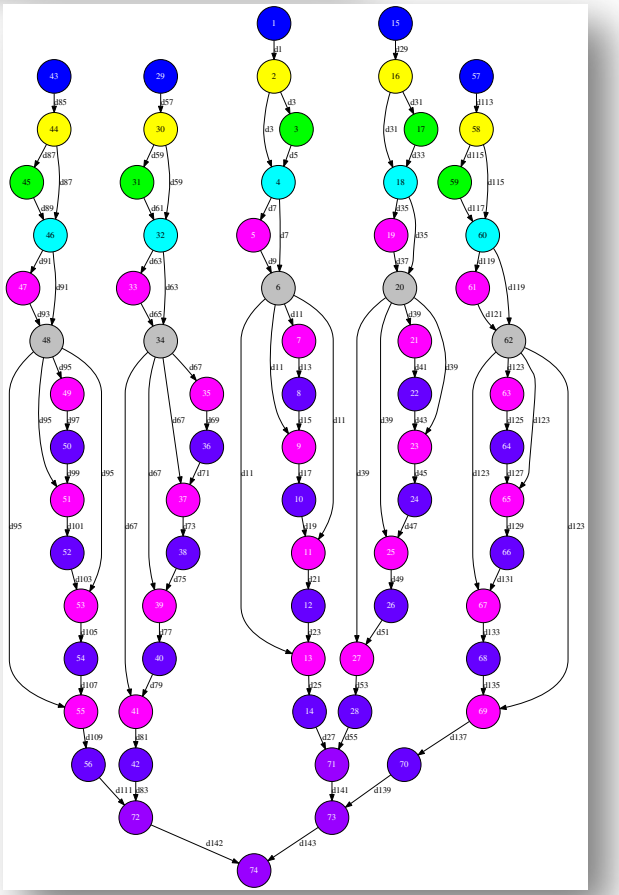
Building Blocks ***Workflows*** ***Execution***

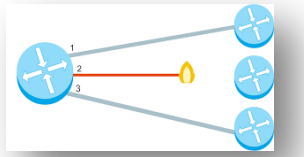


- **Biomolecular simulation workflows:** e.g. MD, free energy, Virtual Screening
- **Prepared for the pre-exascale:** efficient use of thousands of cores in one single job:
  - Highest number used: 40,000 cores
  - Common use: ~2,000 cores
- **Flexible and easy to adapt:** custom biomolecular workflows

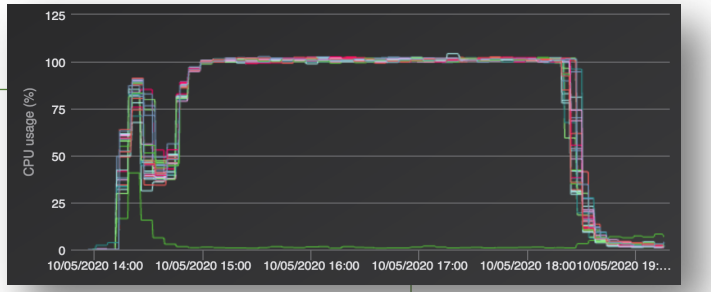
```
print 'step2: mmbuniprot -- Get mutations'
mmbuniprot = uniprot.MmbVariants(prop['pdb_code'])
mutations = mmbuniprot.fetch_variants()

for mut in mutations:
    mut_path = cdir(wd, mut)

    print 'step3: scw -- Model mutation'
    scw_path = cdir(mut_path, 'step3_scw')
    scw_pdb = opj(scw_path, prop['mutated_pdb'])
    scw = scwrl.Scwrl4(mmbpdb_pdb, scw_pdb, mut, scwrl_path=scwrl_path)
    scw_pdb2 = scw.launchPyCOMPSs()
```

"Embarrassingly parallel" executions





Fault tolerance



How it works

```python
global_log.info("step1_pdb: Download the initial Structure")
Pdb(**global_paths["step1_pdb"], properties=global_prop["step1_pdb"]).launch()

global_log.info("step2_fixsidechain: Modeling the missing heavy atoms in the structure side chains")
FixSideChain(**global_paths["step2_fixsidechain"], properties=global_prop["step2_fixsidechain"]).launch()

global_log.info("step3_pdb2gmx: Generate the topology")
Pdb2gmx(**global_paths["step3_pdb2gmx"], properties=global_prop["step3_pdb2gmx"]).launch()

global_log.info("step4_editconf: Create the solvent box")
Editconf(**global_paths["step4_editconf"], properties=global_prop["step4_editconf"]).launch()

global_log.info("step5_solvate: Fill the solvent box with water molecules")
Solvate(**global_paths["step5_solvate"], properties=global_prop["step5_solvate"]).launch()

global_log.info("step6_grompp_genion: Preprocess ion generation")
Grompp(**global_paths["step6_grompp_genion"], properties=global_prop["step6_grompp_genion"]).launch()

global_log.info("step7_genion: Ion generation")
Genion(**global_paths["step7_genion"], properties=global_prop["step7_genion"]).launch()

global_log.info("step8_grompp_min: Preprocess energy minimization")
Grompp(**global_paths["step8_grompp_min"], properties=global_prop["step8_grompp_min"]).launch()

global_log.info("step9_mdrun_min: Execute energy minimization")
Mdrun(**global_paths["step9_mdrun_min"], properties=global_prop["step9_mdrun_min"]).launch()

global_log.info("step10_energy_min: Compute potential energy during minimization")
GMXEnergy(**global_paths["step10_energy_min"], properties=global_prop["step10_energy_min"]).launch()
```

Workflow script

Python code
Building blocks
Loops / conditionals
Global log
Output folders hierarchy

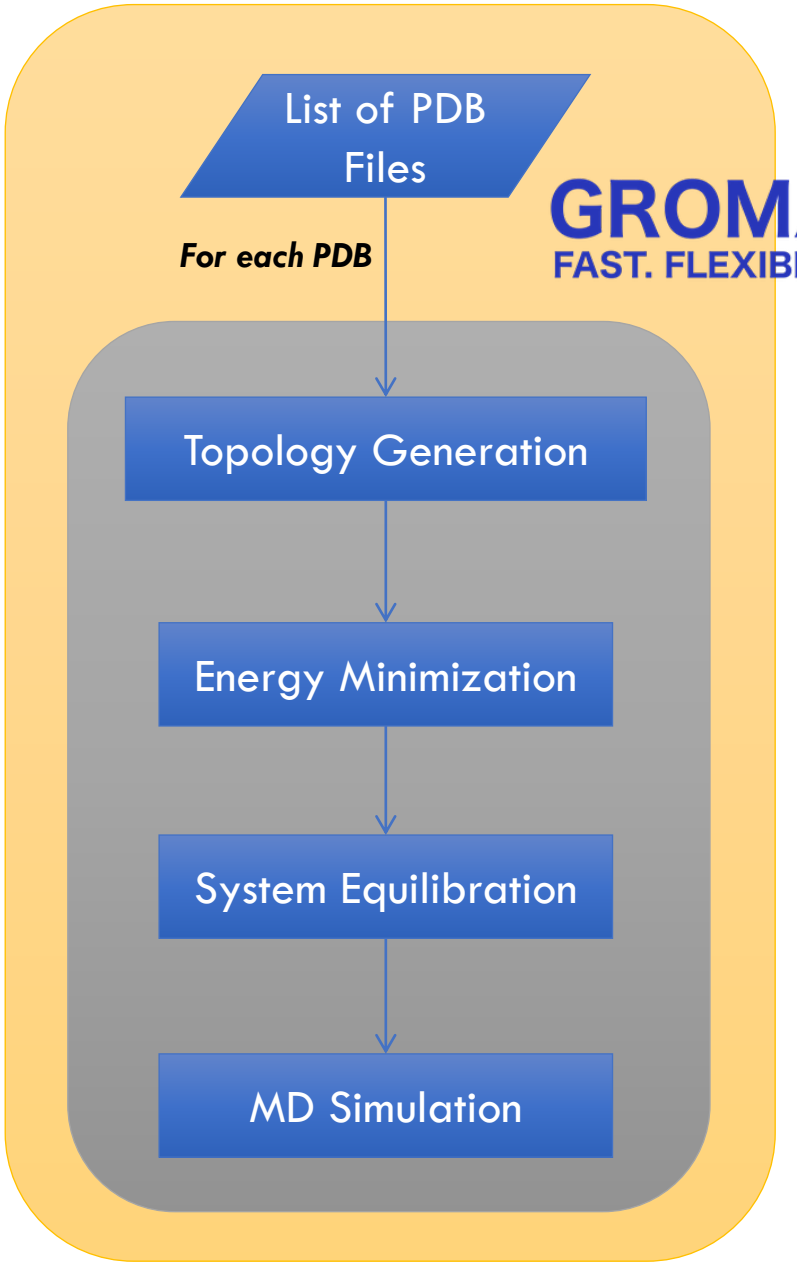

```yaml
# Example of a YAML configuration file for a BioExcel building blocks workflow

working_dir_path: md_tutorial     # Folder to write i/o files of the workflow steps
can_write_console_log: False      # Verbose writing of log information
restart: False                     # Skip steps already performed
remove_tmp: True

step1_pdb:
  paths:
    output_pdb_path: structure.pdb
  properties:
    pdb_code: 1aki

step2_fixsidechain:
  paths:
    input_pdb_path: dependency/step1_pdb/output_pdb_path
    output_pdb_path: fixsidechain.pdb

step3_pdb2gmx:
  paths:
    input_pdb_path: dependency/step2_fixsidechain/output_pdb_path
    output_gro_path: pdb2gmx.gro
    output_top_zip_path: pdb2gmx_top.zip

step4_editconf:
  paths:
    input_gro_path: dependency/step3_pdb2gmx/output_gro_path
    output_gro_path: editconf.gro

step5_solvate:
  paths:
    input_solute_gro_path: dependency/step4_editconf/output_gro_path
    output_gro_path: solvate.gro
    input_top_zip_path: dependency/step3_pdb2gmx/output_top_zip_path
    output_top_zip_path: solvate_top.zip
```

Workflow parameters

Steps Inputs / Outputs
Steps Dependencies
Steps Properties
Workflow inputs
Workflow parameters

Example:

MD setup and run for a list of structures

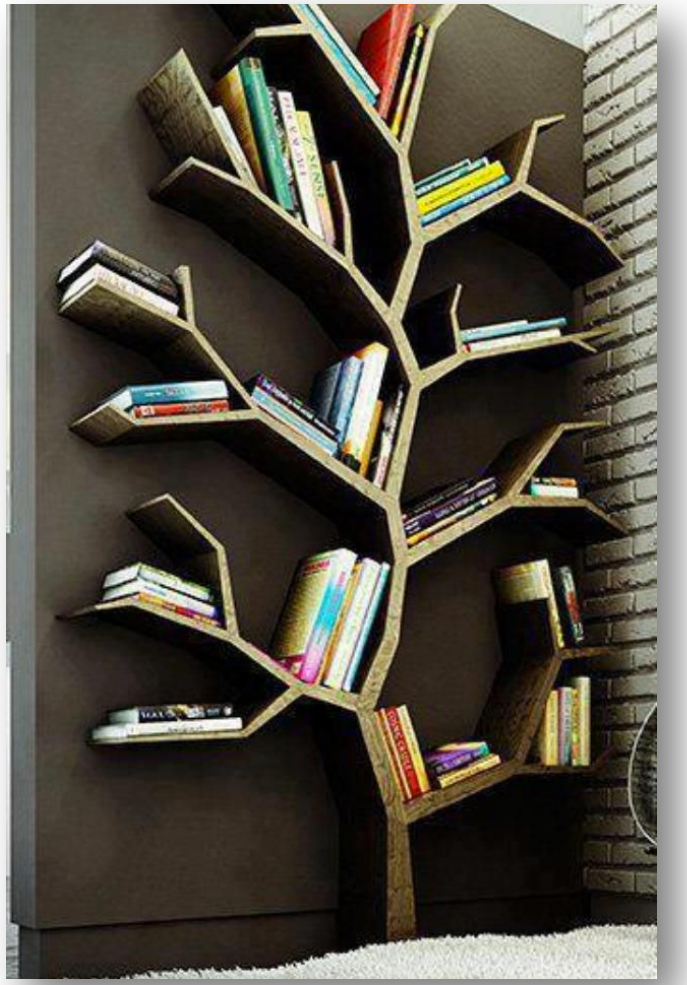


GROMACS
FAST. FLEXIBLE. FREE.

```python
for structure in conf.properties['input_structures'].split(','):

    prefix_str = os.path.basename(structure)
    prefix_str = prefix_str.replace('.','_')

    mut_prop = conf.get_prop_dic(prefix=prefix_str)
    mut_paths = conf.get_paths_dic(prefix=prefix_str)

    global_log.info("Starting setup process for PDB: " + prefix_str)
    mut_paths['step1_pdb2gmx']['input_pdb_path'] = structure

    global_log.info("step1_pdb2gmx: Generate the topology")
    pdb2gmx_pc(**mut_paths["step1_pdb2gmx"], properties=mut_prop["step1_pdb2gmx"])

    global_log.info("step2_editconf: Create the solvent box")
    editconf_pc(**mut_paths["step2_editconf"], properties=mut_prop["step2_editconf"])

    global_log.info("step3_solvate: Fill the solvent box with water molecules")
    solvate_pc(**mut_paths["step3_solvate"], properties=mut_prop["step3_solvate"])
```

```yaml
# List of structures to be simulated
input_structures: ['/home/bscxxxx/str.pdb','/home/bscxxxx/str2.pdb',
                   '/home/bscxxxx/str3.pdb','/home/bscxxxx/str4.pdb']

step1_pdb2gmx:
  paths:
    output_gro_path: mut_gmx.gro
    output_top_zip_path: mut_gmx.top.zip
  properties:
    force_field : amber99sb-ildn

step2_editconf:
  paths:
    input_gro_path: dependency/step1_pdb2gmx/output_gro_path
    output_gro_path: editconf.gro
  properties:
    box_type: octahedron
    distance_to_molecule: 1.2

step3_solvate:
  paths:
    input_solute_gro_path: dependency/step2_editconf/output_gro_path
    output_gro_path: solvate.gro
    input_top_zip_path: dependency/step1_pdb2gmx/output_top_zip_path
    output_top_zip_path: solvate_top.zip
```

Workflow Templates: Library



MD setup + run for a list of structures

python YAML

Model + MD setup + run for a set of protein mutations

python YAML

Free energy calculations with fast-growth TI

python YAML

Virtual Screening

python YAML

And many more...

Workflow Launcher

 # BioExcel Building Blocks
module load ANACONDA/2019.10
module load biobb/covid_dev

 # [Py]COMPSs release
module load COMPSs/2.6.4

 # Singularity
module load singularity

 # GROMACS 2019
module load gromacs/2019.1

 # Specific Run Parameters

Environment Modules

Welcome to the Environment Modules open source project. The Environment Modules package provides for the dynamic modification of a user's environment via modulefiles.

pmx: generate hybrid protein structure and topology
Computational Biomolecular Dynamics Group

GROMACS
FAST. FLEXIBLE. FREE.

```
enqueue_compss
--job_name=mds_pmx_covid
--num_nodes=8
--exec_time=2800
--worker_in_master_cpus=48
--base_log_dir=$PWD
--qos=bsc_ls
wf.py  --config wf.yaml
```

Workflow Launcher

 # BioExcel Building Blocks
module load BioBB/2.6

 # [Py]COMPSs release
module load COMPSs/2.6

 # PCoCC
module load PCoCC

 # GROMACS 2019
module load gromacs/2019.4

 # Specific Run Parameters

Environment Modules

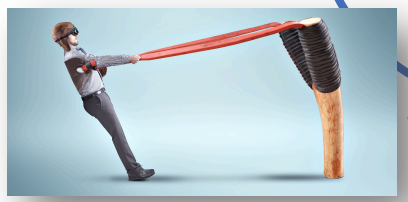
Welcome to the Environment Modules open source project. The Environment Modules package provides for the dynamic modification of a user's environment via modulefiles.

HPC HIGH PERFORMANCE COMPUTING

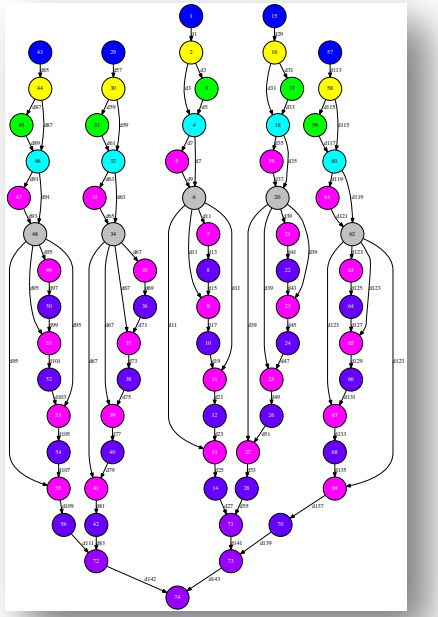
pmx: generate hybrid protein structure and topology
Computational Biomolecular Dynamics Group

GROMACS
FAST. FLEXIBLE. FREE.

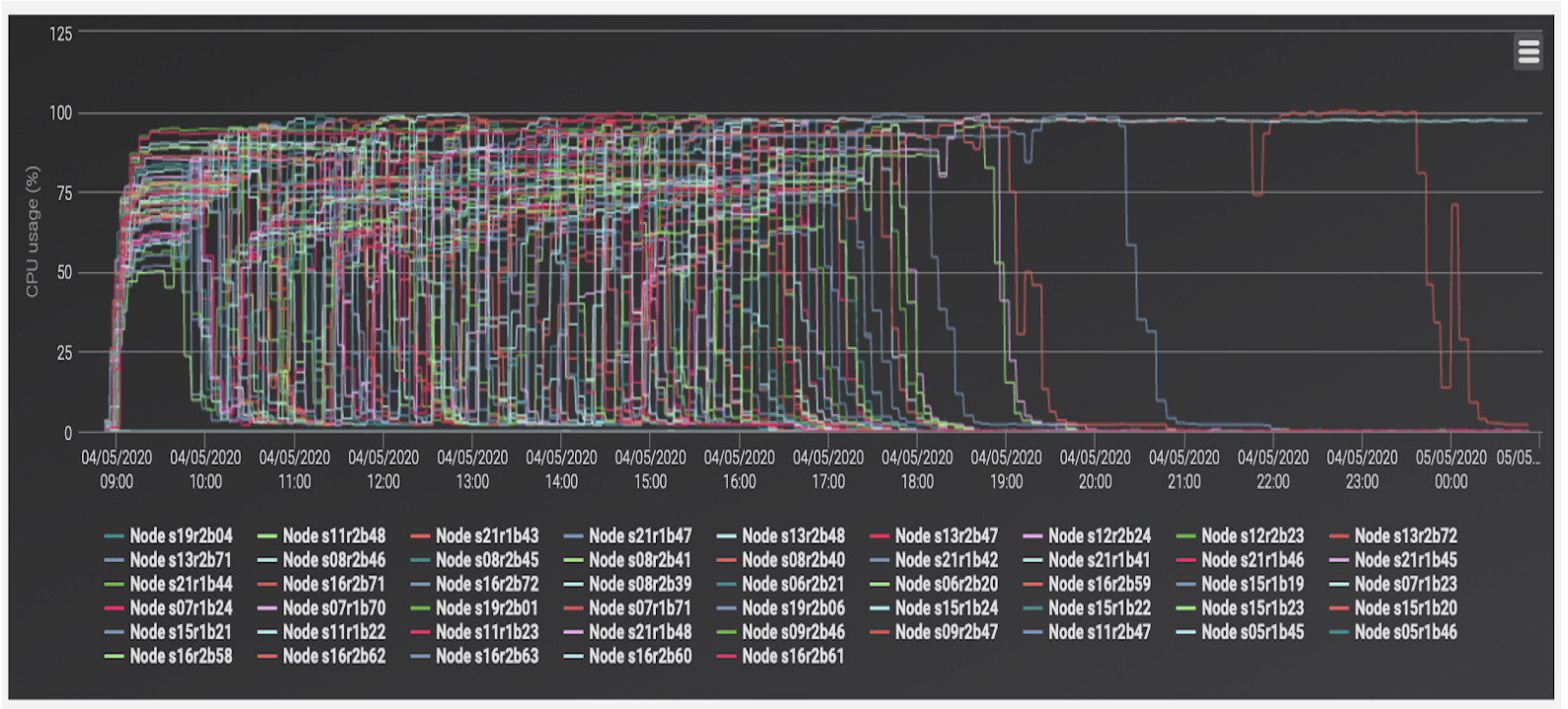
```
enqueue_compss
--job_name=mds_pmx_covid
--num_nodes=8
--exec_time=2800
--worker_in_master_cpus=128
--base_log_dir=$PWD
--qos=bsc_ls
wf.py  --config wf.yaml
```



Examples



Efficient usage of resources



n° nodes < n° executions

Protein-protein benchmark
200 HADDOCK executions

BSC MareNostrum 50 nodes (2400 cores)
1 execution x node

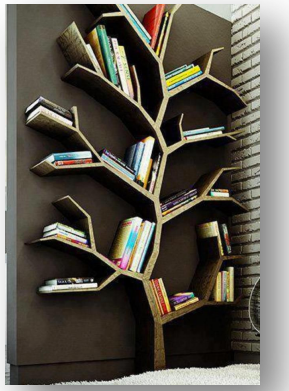
BioExcel Building Blocks (BioBB) High Performance Computing (HPC) Workflow repository

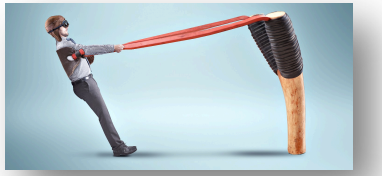
GitHub

https://github.com/bioexcel/biobb_hpc_workflows

We are working on some new workflows that will be added to the current list of workflows and launchers:

- Workflows for MD:
  - md_list: Performs a system setup and runs a molecular dynamics simulation on each one of the structures listed in the YAML properties file.
  - md_muts_sets: Performs a system setup and runs a molecular dynamics simulation for each one of the listed mutations in a given structure.
  - md_add_muts_wt: Applies a list of mutations over the initial structure obtaining a set of structures (initial structure + one mutation, initial structure + two mutations, initial structure + three mutations, ...). Finally performs a system setup and runs a molecular dynamics simulation for each of the structures in the set.
- Workflows for Free Energy Calculation (PMX):
  - pmx_cv_cufix_term: Performs a fast-growth mutation free energy calculation from two equilibrium trajectories.
- Launchers:
  - md_launch: Launcher for the md_list workflow.
  - mdmut_launch: Launcher for the md_muts_sets and md_add_muts_wt workflows.
  - pmx_launch: Launcher for the pmx_cv_cufix_term.





```
bscxxxxx@login1:~>/apps/BIOBB/workflows/MN4/md_launch.py  -h
usage: md_launch.py [-h] -i INPUT_STRUCTURES [-s SYSTEM] [-f FORCE_FIELD]
                    [-bs BOX_SIZE] [-bt BOX_TYPE] [-c CONCENTRATION]
                    [-q QUEUE] [-t TIME] [-nn NUM_NODES] [-cv COMPSS_VERSION]
                    [-d] [-l MD_LENGTH] [-mpi MPI_NODES] [--base_dir BASE_DIR]
                    [-o OUTPUT_DIR] [-jn JOB_NAME] [-gl GROMACS_LIB]
```

Workflow to setup and run MD simulations for a set of **PDB structures.**

```
/apps/BIOBB/workflows/md_launch.py
        -i structures.list
        -l 100
        -nn 16
        -mpi 4
        -o MDs
        -q bsc_ls
        -t 2800
        -jn biobb_mds
```

Structures.list file:

/home/bscxxxx/PDBs/str1.pdb
/home/bscxxxx/PDBs/str2.pdb
/home/bscxxxx/PDBs/str3.pdb
/home/bscxxxx/PDBs/str4.pdb

Table S1. DNA sequences in the miniABC library.

Seq. number	Watson strand (5'-3' direction)
1	GCAACGTGCTATGGAAGC
2	GCAATAAGTACCAGGAGC
3	GCAGAAACAGCTCTGCGC
4	GCAGGCGCAAGACTGAGC
5	GCATTGGGGACACTACGC
6	GCGAACTCAAAGGTTGGC
7	GCGACCGAATGTAATTGC
8	GCGGAGGGCCGGGTGGGC
9	GCGTTAGATTAAAATTGC
10	GCTACGCGGATCGAGAGC
11	GCTGATATACGATGCAGC
12	GCTGGCATGAAGCGACGC
13	GCTTGTGACGGCTAGGGC

https://mmb.irbbarcelona.org/miniABC/

The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules

Pablo D Dans ✉, Alexandra Balaceanu, Marco Pasi, Alessandro S Patelli, Daiva Petkevičiūtė, Jürgen Walther, Adam Hospital, Genís Bayarri, Richard Lavery, John H Maddocks ... Show more

Author Notes

Nucleic Acids Research, Volume 47, Issue 21, 02 December 2019, Pages 11090–11102, https://doi.org/10.1093/nar/gkz905

/apps/BIOBB/workflows/MN4/md_launch.py

1. **13 MDs with Na ions –** 4,992 cores (8 MareNostrum nodes x each MD)
2. **13 MDs with K ions –** 4,992 cores (8 MareNostrum nodes x each MD)
3. **13 MDs with Na + K ions –** 4,992 cores (8 MareNostrum nodes x each MD)

parmBSC0* / *parmBSC1

```
/apps/BIOBB/workflows/MN4/md_launch.py
-i structures.list
-q class_a
-nn 104 –mpi 8
-jn miniABC –t 3800 –l 1000
-o ~/miniABC/Na_MDs
-f amber99bsc1
```

MENU SCIENTIFIC DATA

Article | Open Access | Published: 10 September 2019

BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows

Pau Andrio, Adam Hospital, Javier Conejero, Luis Jordá, Marc Del Pino, Laia Codo, Stian Soiland-Reyes, Carole Goble, Daniele Lezzi, Rosa M. Badia, Modesto Orozco & Josep Ll. Gelpi

Scientific Data **6**, Article number: 169 (2019) | Cite this article

biobb http://mmb.irbbarcelona.org/biobb/

Home | Availability | Workflows | Build your own BioBB | About

BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows

Home

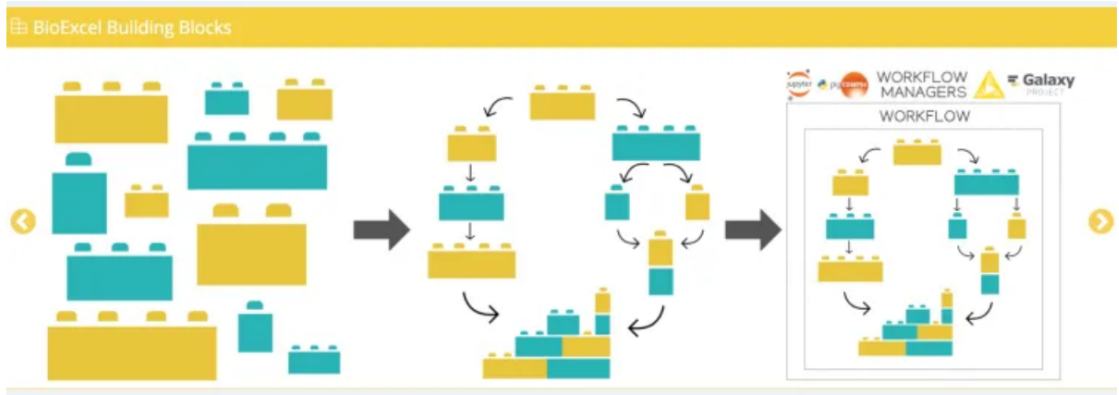
BioExcel Building Blocks

Enabling real science

Building biomolecular simulation workflows to solve scientific problems

Webinar: Computational biomolecular simulation workflows with BioExcel building blocks

August 7, 2020 Arno Proeme



BioExcel's webinar series continue with a presentation by Adam Hospital

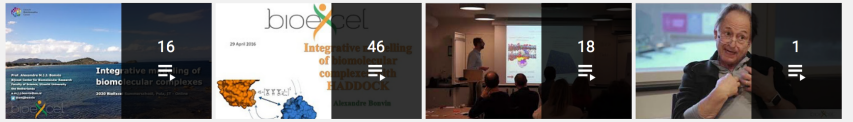
YouTube ES Search

BioExcel Center of Excellence for Computational Biomolecular Research

830 subscribers SUBSCRIBE

HOME VIDEOS PLAYLISTS CHANNELS DISCUSSION ABOUT

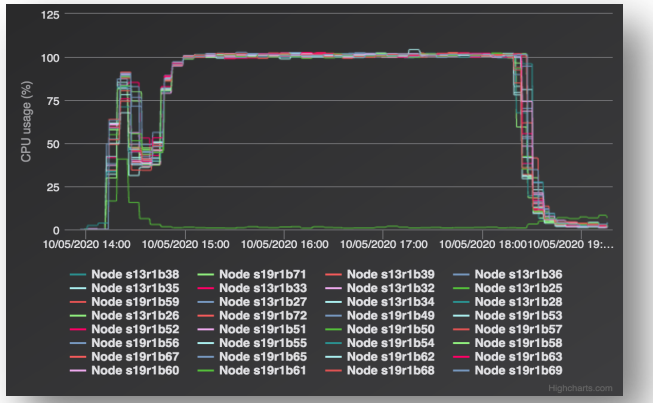
Created playlists SORT



2020 Remote BioExcel Summer School on Biomolecular... | Webinars VIEW FULL PLAYLIST | SDMS18 VIEW FULL PLAYLIST | Interviews VIEW FULL PLAYLIST

Try it!

- Openly available on BSC MareNostrum 4 supercomputer
  1. module load python/3.6.1
  2. /apps/BIOBB/workflows/md_launch.py
  3. /apps/BIOBB/workflows/mdmut_launch.py
  4. /apps/BIOBB/workflows/pmx_launch.py



- European supercomputers:
  - MareNostrum 4 supercomputer (BSC – Spain)
  - Irene Rome (CEA – France)
  - Hawk (HLRS – Stuttgart)

- If interested to try, contact
  - adam.hospital@irbbarcelona.org