Elissa's notes for Zinc22 on Wynton as off 2/11/2022

This is probably constantly under development. Just ask me/us as you run the screen questions. I will update the tutorial/scripts/steps as things change. Currently my workflow is:

- Setup in fragment, leads1, leads2, etc subsets
- Run with 8 hour run time (can also do 30 mins in short queue, but this might have issues later on)
- Check for failed jobs, corrupt mol2 files
- Submit these jobs with multiple hour (5-6) runtime to avoid failure/corruption
- Double check all jobs have run/data is there

If running on gimel, super script below, but rest of tutorial not adapted for gimel
/mnt/nfs/exa/work/ak87/UCSF/NSP14/DOCKING/TYR368_0.2-ALA353_0.2-GLY333_0.2/ZINC611-XTAL/LSD/ALL-SPH/WUXI-MCULE/SUBS1/super

JK's Color matching: http://wiki.docking.org/index.php/Coloring_and_Subcluster_Matching
This is not used in this tutorial, so ask him if you want to use it.

Setting up docking

http://wiki.docking.org/index.php/How_to_dock_in_DOCK3.8

Setup directory:

1. Copy dockfiles and INDOCK from gimel to wynton (no symlinks!)
    a. You can no longer copy from ex6 or ex4, you must use scp
        i. On the wynton side, always log into dt1 or dt2
            1. ssh user@[dt1/dt2].wynton.ucsf.edu
        ii. On the gimel side, use files 2 or files 3
            1. ssh user@169.230.75.3 0r .33
        iii. Example
            1. ssh user@169.230.75.3 and in a separate window ssh user@dt1.wynton.ucsf.edu
            2. From files2 window,
                a. scp -r [folder_to_copy] user@dt1.wynton.ucsf.edu:[path/to/work/directory]/.

2. **ONLY launch LSDs from /wynton/group/bks/work/username not on home dir**
    a. ssh usea
    b. cd /wynton/group/bks/work/username/
3. ~~cp /wynton/group/bks/work/elisfink/D2R/super .~~ Deprecated as of 2023-11-17 due to dock score asterisks bug; use new method (DOCK3R) as follows:

a. **Note, as of 2/22/22** the super script is to be placed in same dir as dockfiles/ so that it can set $DOCKFILES correctly – you do not need to change the path to dockfiles anymore using this script
b. Copy super script that uses a DOCK3R container instead of a DOCK3 executable:
    cp /wynton/group/bks/soft/docking_super_scripts/super_dock3r.sh super
c. Obtain the correct subdock.bash and rundock.bash scripts:
    cp -r /wynton/group/bks/soft/SUBDOCK/ .
d. Check DOCK3R related environmental variables in super script:
    i. DOCK3R_IMAGE: This parameter specifies the DOCK3R image to be used for the container. The DOCK3R image contains the DOCK 3 program and all the dependencies required for it to run.
    ii. USE_APPTAINER_DOCK3R: Whether to use Apptainer (formerly Singularity) instead of Docker as the containerization method. Wynton uses Apptainer, in which case this should be set to true.
    iii. MOLECULES_DIR_TO_BIND: This is a configuration parameter for the DOCK3R container. It requires the absolute path to a directory on the host system that contains all the molecules to be docked. E.g., /wynton/group/bks/ . When set, this directory is bind-mounted into the Docker container, allowing it to access and process these files at runtime. This setup ensures the necessary molecular data is available to the container for its tasks.

4. Add lines at bottom of INDOCK before the last line for mid-docking strain filtering: /wynton/group/bks/work/elisfink/strain_INDOCK_lines
    a. remove the # beside check... Total.. max…
5. Make sure to set mol2 max low to avoid writing out too much data, extra low is safer, redocking is fast if you need to increase the cutoff
6. On line 1 of INDOCK, change to 3.8
7. Copy/move INDOCK to dockfiles
8. Change super script to reflect your dockfiles directory path
    a. This is not necessary if you use Elissa's new version of super script
9. **DO NOT UPDATE /scratch**

Get SDI:

10. Get zinc22 charge subsets for all HACs you want to dock
    http://wiki.docking.org/index.php/Selecting_tranches_in_ZINC22
    Charges: J=-4; K=-3; L=-2; M=-1; N neutral; O=+1; P=+2; Q=+3; R=+4
    a. Check in this path which date/current you want to use, it may change: /wynton/group/bks/zinc-22/sets, I named it as <pick> below
    b. For all anions: grep "\-M\-" /wynton/group/bks/zinc-22/sets/<pick>/*.wyn > precalc-sdi-anions
    c. For neutrals: grep "\-N\-" /wynton/group/bks/zinc-22/sets/<pick>/*.wyn > precalc-sdi-neutrals

d. For cations: grep "\-O\-" /wynton/group/bks/zinc-22/sets/<pick>/*.wyn > precalc-sdi-cations
e. For +2: grep "\-P\-" /wynton/group/bks/zinc-22/sets/<pick>/*.wyn > precalc-sdi-dications
f. For -2: grep "\-L\-" /wynton/group/bks/zinc-22/sets/<pick>/*.wyn > precalc-sdi-dianions

11. Other ZINC22 command examples:
    a. For (cations, H17 to H25, logP to 3.4999): grep "\-O\-" /wynton/group/bks/zinc-22/sets/<pick>/*.lead-like.wyn > leadlikecationsonly
    b. For all lead-like ls /wynton/group/bks/zinc-22/sets/<pick>/*lead-like.wyn > precalc-sdi-leadlike
    c. For everything ls /wynton/group/bks/zinc-22/sets/<pick>/*wyn > precalc-sdi-all

In-Stock sets:
Gimel: /nfs/exd/zinc-22g
Wynton: /wynton/group/bks/zinc-22g

Prepare SDI/directories for the screen:

12. Separate by HAC to multiple screens
    a. If you have multiple charge group sdi files, cat all charge groups you are interested in tol precalc-sdi.final
    b. Use this script but it might not accomplish everything you want it to do
        i. It will separate into multiple SDI files: HAC up to 16 (fragments), HAC 17-19 (leads1), HAC 20-23 (leads2), HAC 24-25 (large), HAC 26 and up (larger)
        ii. This also splits the respective *in files into chunks of 10,000 automatically
        iii. Use this if you have a ":" between a *wyn file and *db2.tgz file
            1. cp /wynton/group/bks/work/elisfink/5HT2AR/prep/test/make_sdi_files_version2.sh
            2. sh make_sdi_files_version2.sh precalc-sdi.final fragments leads1 leads2 large larger
        iv. Use this if you got your molecules by *lead-like.wyn set and don't have a semicolon
            1. cp /wynton/group/bks/work/elisfink/5HT2AR/prep/make_sdi_files_version1.sh
            2. sh make_sdi_files_version1.sh precalc-sdi.final fragments leads1 leads2 large larger
                a. qsub /wynton/group/bks/work/ttummino/GALR_screens/scripts/qsub_make_sdi_files_version1.sh zinc22u_g.sdi
        v. For either script: ***in files are the sdi files**
        vi. Skip, automatically done in the scripts above, but if you need JK's script…

1. **NOTE** if any of your *in files has more than 10,000 lines, split into multiple files, ie. if leads2.in has 12,000 makes leads2a.in with 10,000 lines and leads2b.in with 2,000 lines
2. cp /wynton/group/bks/work/jklyu/LSD/colored_subcluster_docking/zinc22_3d_build_20210504/script/split_sdi.csh
3. csh /wynton/group/bks/work/jklyu/LSD/colored_subcluster_docking/zinc22_3d_build_20210504/script/split_sdi.csh

13. Make multiple directories for each .final you want to dock
    a. for example ~/VGLUT2_zinc22_lsd/fragment_subset_lsd/ , ~/VGLUT2_zinc22_lsd/leads1_subset_lsd/ , etc.
14. In each of these directories of a subset:
    a. Copy dockfiles/ with the INDOCK inside
    b. mkdir sdi
    c. mv fragments*.in ~/VGLUT2_zinc22_lsd/fragment_subset_lsd/sdi
    d. Copy super script into each one, update dockfiles path if needed

Submit and run LSD

15. To submit docking, go into each lsd_subset
    a. sh super
        i. Make sure to update super script to reflect names of sdi files (*.in, *.in.*, *.sdi
    b. This submits one job with each line of the sdi/*.in file being a subjob
    c. Start with submitting one of the subsets to make sure things run okay, then launch rest of the subsets
16. If you want jobs to only run 30 mins, change line 9 to 28 mins and 30 mins. This gets you the short job queue but you could run into problems getting docking output later.
17. Otherwise we recommend changing to a longer run time of 8:28 and 8:30 on line 9. This will mostly let your jobs run until they finish. This helps the docking output stay intact when you go to get poses later.
18. All jobs might not be done after the run time you specified, so to restart docking
    a. sh super
19. At some point you will be resubmitting a small number of jobs over and over again, so change to longer run time! 10 hours or so on...

Before you process results

Make sure you run 'sh super' until you get 0 jobs launching, before proceeding to the next section

Check for failed jobs

1. Make dirlist
    a. cp /wynton/group/bks/work/fliu/zzz.script/make_dirlist_0819.sh .

i. Use
1. /wynton/home/shoichetlab/ttummino/work/new_scripts/make_dirlist.sh
2. If you are using subdock.bash for docking now
ii. Edit to reflect ~results/ or ~output/ as written in your super script
b. sh make_dirlist.sh

2. Find finished/not finished jobs, remove output from failed jobs, compile errors
a. Run this to remove failed jobs output (OUTDOCK, test.mol2.gz, etc)
i. cp /wynton/group/bks/work/elisfink/VGLUT2/lsd_05082021/fragments/check_finished_notfinished.sh
b. If you don't want to delete output or resubmit jobs due to broken db2.gz files:
i. cp /wynton/group/bks/work/fliu/zzz.script/check_finished_notfinished_0819.sh .
c. Run script:
i. sh check_finished_notfinished.sh
ii. sh check_finished_notfinished_0819.sh
iii. Or qsub /wynton/group/bks/work/ttummino/GALR_screens/scripts/qsub_get_finished_not_finished.sh
d. Errors located in LIST_OF_ERRORS if needed, NOT-FINISHED directories are now ready for resubmission

Check for corrupted test.mol2.gz files

1. Find corrupted mol2 files
a. cp /wynton/group/bks/work/elisfink/VGLUT2/lsd_05082021/leads1/find_corrupt_mol2.sh
b. Make sure you have a dirlist
c. sh find_corrupt_mol2.sh dirlist out.corrupt
d. awk -F"." '{print $2}' out.corrupt | awk -F"/test" '{print $1}' | sort | uniq > resub.from.corrput
e. sed -i 's/\///' resub.from.corrput
2. Remove output from jobs where there are corrupted mol2 files
a. cp /wynton/group/bks/work/elisfink/VGLUT2/lsd_05082021/leads1/remove_failed_contents.sh
b. sh remove_failed_contents.sh resub.from.corrupt

Resubmit jobs with errors

3. Resubmit these failed/corrupted jobs that you now have empty dirs for:
a. Change runtime to 5-6 hours!!! For example 6:00 and 6:02 in your super file
b. sh super

c. Once this run has finished, run through the entire “before you process results” just in case things still need to run again
d. If jobs didn’t finish, just follow “before you process results” and increase run time to 8:00 and 8:02 for example
e. This may take a few rounds

To process docking results

Running mostly like old docking screen for now (eventually migrate to Ben’s new script)
/wynton/group/bks/soft/DOCK = $DOCKBASE

20. To make dirlist
    a. cp /wynton/group/bks/work/elisfink/VGLUT2/lsd_05082021/fragments/make_dirlist.sh
    b. sh make_dirlist.sh
21. Run this to get to no .# on end of test.mol2.gz files (it cat s all .# poses together)
    a. cp /wynton/group/bks/work/elisfink/VGLUT2/lsd_fragments_zinc22_04212021/cat_getposes.sh
    b. sh cat_getposes.sh dirlist
22. Extract all
    a. python /wynton/group/bks/work/jklyu/LSD/DOCK_version/DOCK/analysis/extract_all_blazing_fast.py dirlist extract_all.txt -40
        i. qsub /wynton/group/bks/work/ttummino/GALR_screens/scripts/qsub_extract_all.py [min energy to be extracted]
    b. OR update with JK’s extract_all parallel scripts that needs updated...
       cp /wynton/group/bks/work/jklyu/LSD/5HT2A/zinc22_3d_build_20210824/script/extract_all_blazing_fast_parallel.csh
       Change energy_cap and dir_in variables in extract_all_blazing_fast_parallel.csh
       Also pay attention to foreach subdir (`ls -d output/H*/`), just in case you have a different way to name the subdir
       csh extract_all_blazing_fast_parallel.csh
       Once all the parallel jobs are done, cat *extract_all.sort.uniq.txt > extract_all.txt
       ssh dev1(dev2 and dev3 are two alternatives) and use csh
       setenv TMPDIR /wynton/group/bks/
       sort -nk 22 extract_all.txt > extract_all.sort.txt
       touch dirlist
       python /wynton/group/bks/work/jklyu/LSD/sigma2/scripts/make_uniq.py dirlist extract_all.txt 1000
    c. OR get the paths for top X molecules to redock with lower energies

i. qsub /wynton/group/bks/work/ttummino/GALR_screens/scripts/qsub_get_paths_to_top_hits.py [# of mols you want to redock]
ii. Takes about 1.5 hrs to run 300000

23. Getting poses
    a. python /wynton/group/bks/work/jklyu/LSD/DOCK_version/DOCK/analysis/getposes_blazing_faster.py
    b. Or parallel:
        i. cp /wynton/group/bks/work/elisfink/code/process_lsd/getposes_blazing_fast_parallel.csh
        ii. Csh getposes_blazing_fast_parallel.csh extract_all.sort.uniq.txt 300000 20000
        iii. Arguments: 1st extract_all file, 2nd topX you want, 3rd chunk size
        iv. **Note**: path for /bks/soft/dock/versions/dock37/DOCK-3.7-trunk/analysis/getposes_blazing_faster.py is gone on line 68 of the cshell script; can modify to path /wynton/group/bks/work/jklyu/LSD/DOCK_version/DOCK/analysis/getposes_blazing_faster.py for temporary solution
    c. If you want to submit array jobs and need longer running time for getting poses:
        i. cp /wynton/group/bks/work/fliu/zzz.script/getposes_blazing_fast_parallel_btingle.csh
        ii. cp /wynton/group/bks/work/fliu/zzz.script/par_getposes_btingle.csh
        iii. Csh getposes_blazing_fast_parallel_btingle.csh extract_all.sort.uniq.txt 300000 20000
        iv. Arguments: 1st extract_all file, 2nd topX you want, 3rd chunk size

24. Rsync poses/extract all to gimel for strain/cluster/etc analysis
    a. rsync -avhc username@log1.wynton.ucsf.edu:/path .

25. To get strain from poses run with ZINC22 db2 files… (update)

26. Temp process for getting smiles from poses with obabel
    a. cp /mnt/nfs/ex4/work/elisfink/VGLUT2/lsd_process_05132021/picks_06032021/bfc/run_bfc_zinc22.csh
    b. Arguments: 1) list of zincIDs to run BFC on 2) path to a single old mol2 file, could be those you used for strain 3) path to extract all file 4) topX to run BFC on 5) TC for clustering
    c. Example command: csh run_bfc_zinc22.csh zincid.list ../strain/all_poses.mol2 ../extract_all.sort.uniq.txt 40000 0.5
    d. This scripts uses obabel to get canonical smiles from mol2 file of poses, then does BFC automatically

i. Here are individual commands to get smiles from mol2, but they are in the script above
ii. source /mnt/nfs/soft/openbabel/current/env.csh
iii. obabel -imol2 for_bfc.mol2 -osmi -O out.smi
iv. obabel -i smi out.smi -o can -xi > out.can.smi

If you want to check for failed molecules

a. Login to gimel5
b. Make smiles tab separated
c. `java -jar /nfs/soft/molinsp/mib.2020 -singlepart -normalizeCharges -f combined.protonated.smi > combined.tab.smi`
d. Do this with those for BFC (it may also fix some broken ones??) or on all top X-K poses, IFP should catch broken ones for now

27. Get docking stats:
    a. cp /wynton/group/bks/work/elisfink/5HT2AR/leads1/docking_stat_parallel.csh
    b. Change DOCKBASE path at the top ONLY
        i. /wynton/group/bks/soft/DOCK
    c. csh docking_stat_parallel.csh /current_path (USE FULL PATH!)
       If you have e.g. OUTDOCK.1, OUTDOCK.2 AND OUTDOCK.all, compounds will be counted multiple times. Adjust your scripts!
       E.g.:
       /wynton/group/bks/work/sgahbauer/scripts/docking_stat_parallel.csh
    d. This submits jobs for the number of dirs in /output
    e. Manual version:
        i. grep ZINC output/fragments.aa/*/OUTDOCK.all | grep -v scratch | awk '{print $3}' | sort -u | wc -l
        ii. qsub /wynton/group/bks/work/ttummino/GALR_screens/scripts/manual_count.csh
        iii.

28. Getting supplier codes

a. Save ZINCIDs in file
b. `curl https://cartblanche22.docking.org/substances.txt -F zinc_id-in=@test.txt -F output_fields='smiles,zinc_id,supplier_code'`
c. Elissa note: I've seen some ZINCIDs change so just make sure you check structures. Also some may not have matches (not in output file) so input those in SW/Arthor private in 22/40B and you can find the s/m enamine codes. Some codes may also be WuXi codes.

d. **Another note:** check w/ Ben/Khanh/John when generating these to make sure databases are loaded correctly. Double check smiles that come out against smiles you got in docking. Once you send the quote to Illia, check the smiles for compounds you get back, on the event enamine codes are wrong Illia may switch molecules without you realizing it. A good way to do this is make a csv file with smiles you sent Illia in one column and smiles you got back from illia, open csv in DataWarrior and it'll populate structures of both for easy side by side comparison: https://openmolecules.org/datawarrior/download.html
e. Can also use the GUI: https://cartblanche22.docking.org to look up by zincid/smiles/vendor code
f. **ANOTHER NOTE:** When ordering w/ WuXi, emphasize you want to order based on smiles primarily. JK has spotted discrepancies between smiles from docking and WuXi vendor codes (11/02/21)

SKIP for not not updated: Running with Ben's new faster extract/get poses scripts and CartBlanche for smiles (will update)
http://wiki.docking.org/index.php/Docking_Analysis_in_DOCK3.8

29. Load python3.8 for analysis on wynton
    a. Follow directions here: /wynton/group/bks/work/sgahbauer/DOCK-dev/analysis/top_poses/README
    b. Don't run this until it's updated, it won't get you extract all/single pose for top ranked, will update once it's fixed
30. To get smiles for zinc22: https://cartblanche22.docking.org/search/zincid ask Chinzo any questions, ask about cap of #/request
    a. curl https://cartblanche22.docking.org/substances.txt -F zinc_id-in=@/your/file/pathl/pass_zincid.txt -F output_fields="zinc_id,sub_id,smiles" -F chunk=5000 -F timeout=15

Building molecules with the ZINC22 pipeline

Look at the wiki: https://wiki.docking.org/index.php/Building_The_3D_Pipeline_ZINC22
Use the preinstalled pipeline scripts described in the wiki page. The crossed-out information below is out of date!

To repack for docking:
https://wiki.docking.org/index.php?title=Repackaging_DB2_DOCK38

Consider editing the script to write out to your local directory and not tmp (especially if you have a lot of db2s!)

~~Make sure to git clone the repo for all the most up to date scripts!!!! **Every time, I mean it!** This also includes recopying the scripts for ~/soft if you hard copied rather than linked~~

~~Required files:~~

1. ~~DOCK.3.8.X.X.3d.tar.gz in /wynton/home/shoichetlab/USERNAME/soft~~
    a. ~~Make sure to use an up-to-date version! This contains the building pipeline code~~
    b. ~~cp /wynton/home/shoichetlab/sgahbauer/soft/DOCK3.8.4.0.3d.tar.gz ~/soft/.~~
2. ~~Openeye and jchem licenses and corina~~
    a. ~~cp /wynton/home/shoichetlab/sgahbauer/soft/.jchem-license.cxl ~/soft/.~~
    b. ~~cp /wynton/home/shoichetlab/sgahbauer/soft/.oe-license.txt ~/soft/~~
        i. ~~On gimel it needs to be in home (~/.)~~
        ii. ~~Make sure to echo $OE_BASE and $OE_DIR and that they are empty, if they are not empty, comment out sourcing of dock3.7 environment in your .bashrc or .cshrc, log out, and then log back in and make sure these variables are empty!~~
    c. ~~cp /wynton/home/shoichetlab/sgahbauer/soft/corina.tar.gz ~/soft/.~~
        i. ~~On gimel it needs to be in home (~/.)~~
3. ~~ZINC22 building and submission scripts (e.g. from here https://github.com/btingle/zinc-3d-build-3)~~

4. Go to your directory where you want to build molecules which contains the smiles file (***YourSMILESfile.smi***).
5. For BKS, ssh gimel5 and use slurm. For Wynton, ssh dev3 (login nodes should work too) and use SGE.

6. ~~Copy the building and submission script from the repo you just cloned~~
    a. ~~cp zinc-3d-build-3-master/build-3d.bash .~~
    b. ~~cp zinc-3d-build-3-master/submit-all-jobs-slurm.bash .~~
7. Make a run file (run.sh) BKS version

```
#!/bin/bash
source /nfs/soft/dock/versions/dock38/pipeline_3D_ligands/activate
export INPUT_FILE=YourSMILESfile.smi
export OUTPUT_DEST=YourOutputDirectory_full_path
export SBATCH_ARGS="--time 2:00:00" # limits time to 2 hours
submit-all-jobs-slurm.bash
```

8. Make a run file Wynton version

```
#!/bin/bash
source /wynton/group/bks/soft/pipeline_3D_ligands/activate
```

```
export INPUT_FILE=YourSMILESfile.smi
export OUTPUT_DEST=YourOutputDirectory_full_path
export QSUB_ARGS="-l s_rt=01:28:30 -l h_rt=01:30:00 -r y" # limits hard limit to 1.5 hours, with a soft limit 1.5 minutes before hard limit to ensure graceful exit
submit-all-jobs-sge.bash
```

9. Run the building scripts
    a. Chmod +x run.sh
    b. ./run.sh
10. When done, molecules will be stored in ***YourSMILESfile.smi.batch-3d.d***
    a. Built molecules are /out/
    b. You can check the ligand tar files
       tar -tf 1.tar.gz
11. db2 files for every protomer need to be extracted from the tar.gz files and repacked for docking. Copy script into your directory where you want to build.
    a. cp /wynton/group/bks/work/sgahbauer/scripts/20220124_get-db2-tgz.sh .
    b. ./20220124_get-db2-tgz ***YourSMILESfile.smi.batch-3d.d***
    c. This script might flag some errors if there aren't molecules from each charge group. Should be benign
12. This will generate a directory containing db2.tgz files (tarballs) and a directory with sdi files ready for docking
    a. tarballs_***YourSMILESfile.smi.batch-3d.d***
    b. sdi_***YourSMILESfile.smi.batch-3d.d***

Status of ZINC22 building
https://docs.google.com/spreadsheets/d/14QpCiB78feWPQKT5LkpfKDNR6VCAMbEs7Md-QMX0glc/edit?ts=608b31dc#gid=485268055

/wynton/group/bks/soft/DOCK

Error message for dockdev team

What version of 3.8 on Wynton can we source? Instead of pulling everytime it's needed from github… (to go in super script)
What is the current super script? Potentially reflecting changes for job queuing?
Updated list of OUTDOCK errors that we have:

Error 1:

open the file:
/scratch/elisfink/9857570_2153/input/H16P090-M-paa.db2.tgz/H16/H16P090/BP/Gc/ZINCgf000000BPGc.0.M.db2
Error. seenint > db2lig%total_confs
3 2
Error. seenint > db2lig%total_confs
3 2
close the file:
/scratch/elisfink/9857570_2153/input/H16P090-M-paa.db2.tgz/H16/H16P090/BP/Gc/ZINCgf000000BPGc.0.M.db2

Error 2:

open the file: /scratch/sgahbauer/3154694_10/input/55.db2.tgz/ZINC000000033953.0.N.db2.gz
5 7 bonds with error
Error. newlist is not big enough
5 7 bonds with error
Error. newlist is not big enough
close the file: /scratch/sgahbauer/3154694_10/input/55.db2.tgz/ZINC000000033953.0.N.db2.gz

Temp notes

```
#!/bin/bash
tmp_dir=$(mktemp -d -t ci-XXXXXXXXXX --tmpdir=/home/gabor/tmp)

echo $tmp_dir
```

Isha changed for extract_all_blazing_fast on wynton