GEMMA Time Domain Astronomy Scheduler Testing and Verification

1. Overview

This document explains how the Scheduler performance will be tested and evaluated. In general, testing will be done by both running the Scheduler for single nights and comparing to QC manual plans, and by running in validation mode over a past, recreated semester and comparing to what was actually accomplished during that semester.

In the former case, this is to ensure that plans appear sensible and resemble a typical QC plan. Scores for both manual and Scheduler plans will be produced and analyzed to understand any differences. These single night plans will also be used to test the Real-time Scheduler by including any changes in weather conditions or incoming ToOs during the night. The time needed for the Scheduler to generate a new plan will be compared to the time needed for observers to switch observations using the manual plans.

For generating a semester's worth of observations, this will be done using complete observation sets from past semesters including information on nightly weather conditions during that period, times of incoming rToOs, and daily instrument and configuration availability to reproduce the semester as realistically as possible. The resulting end of semester metric scores (based on the numbers of B1 and B2 programs that would be completed assuming the plans were observed as scheduled) will be compared to actual results from those semesters. The Scheduler is expected to perform better than the current manual QC planning process in both reaching a higher fraction of completed programs (to at least 80% completion) and in minimizing time lost at night when conditions change.

For each Scheduler release, specific tests will be extracted from those listed below in Section 2.3 for the relevant requirements included in each release. The requirements and associated tests will be included together in a testing spreadsheet.

2. Testing plan

2.1 Initial Considerations

1. First testing will likely occur before the new database structure. There won't be an easy way to correctly relate all calibrations or relative timings of science observations.

- Extract all available information from the current database.
- May be able to use existing scheduling groups to relate observations. Assume rules for calibration frequency.
- Otherwise, may have to make assumptions about length of typical calibrations and include these space holders with science observations as per standard rules per instruments.
- Observations with specific relative timings and frequencies may require the use of groups and additional specific timing windows.
- GPP will provide minimum block sizes (atoms). This information is not available in the current database. May need to make assumptions about the number of steps constituting an atom for different instrument/modes or manually create atoms for a set of observations.

2. Testing of full past semesters (or longer periods than single nights) - To compare to what actually happened, we will need to fully recreate the semester with the weather conditions through the night -IQ/CC/(WV for GN), the GMOS configurations, instruments available, systems/components (un)available due to faults or Instrument checkout dates/times, when different programs actually came online (near start of semester or throughout for FT/DD), etc.

- May need assumptions about when new (FT,DD) programs came online for a rough comparison with actual semesters. Queue programs can be assumed to be available from the start of the semester.

- Can also simulate *realistic* past semesters in order to compare simulations for the full semesters with each other, for different scheduling algorithms and metrics/scores.

3. Each testing simulation should have a final metric/score. Repeated runs with any changes should be stored with the final scores and schedules.

- Manual QC plans, including historical QC plans can have scores calculated in order to compare to the Scheduler results.

- A saved plan should contain the full input information, at a minimum the input resources and observing database dates or snapshot used. This is to keep a record of the input factors and to allow re-creation of the test/runs.

- Each generated plan should have a breakdown of the factors in the score to determine what changes in the scoring may be needed. This is also to compare the factors in the scoring to the “manual” factors used by the QC.

- Simulations run over multiple nights should have each nightly plan saved along with the associated scores.

2.2 Semester Re-creation

We have re-created several semesters: 18B and 19A for both Gemini North and South and 18A for Gemini North. These include both typical semesters for each telescope and semesters with above average time losses (from weather, faults, or other issues). The recreation includes the instrument and telescope schedule, the GMOS daily configuration changes, nightly weather conditions, and fault time losses. The telescope/instrument schedules are taken from existing calendars. The GMOS daily configurations are re-created using the historical record of component change requests in the Instrument Control Tracking Database (ICTD). The weather information for each night is taken from guider probe seeing values, Weather sensors in GEA, and QAP (Quality Assessment Pipeline) values. Fault losses are taken from the fault reporting system.

We do not have information for when different programs came online. We therefore assume all queue programs are ready from the start of the semester. For FT, we can use the 3 month cycle period during which they were active. For DD programs, we can assume the program was activated within a week of when the first observations were originally obtained.

For rToOs, we can use the default 24h timing windows when they were triggered. For sToOs, we may be able to parse the trigger emails for the date of trigger when timing windows were not used.

We also take into account time loss and component unavailability for significant faults (those with time loss > 1h or instrument/mode became unavailable for a period of time) on the dates where these were disruptive to the queue. For less significant faults, we spread the actual time loss over each month randomly. Short faults (<0.1h) are ignored as it is very rare that these types of faults stop an observation. For on-sky checkouts, we can take the actual times from past nightlogs or assume a set amount of time per instrument in the first CC70/IQ70 or better conditions after an instrument has been installed.

For each of these semesters, we have extracted from the database the following information for the programs/observations including: Program ID, affiliate, Band, rollover flag, thesis flag, program mode, ToO type, program time allocation, time award per participant, program notes, [ordered list of observations in a scheduling group], observation ID, phase2 status, execution status, internal priority, setup time and acquisition overhead, constraints (CC, IQ,SB, WV, AM, elevation, timing windows), target and guide star coordinates, execution and overhead times for all steps in sequences. We have also grabbed information about target and condition requirement changes that occurred during the course of the semesters, whenever this information was noted in the program. While program change is unlikely something that could be included in the simulations, we at least have the programs flagged so that we are aware of which programs were affected by changes. This amounts to less than 5% of programs each semester.

The original observations for these semesters are first copied over to a test database, where they are manually cleaned up in order to recreate as close as possible the original number and length of observing sequences in each program before actual observation attempts caused significant changes to the programs.

For the number of reacquisitions needed for a long sequence, we assume the expected frequency (and duration) for each instrument. The scheduler will be able to schedule partial observations, with rules for minimum lengths to schedule. The standard assumed time for acquisitions for each instrument/mode will be included each time an observation is scheduled.

The Scheduler must also know the minimum length of an observation that can be scheduled, which we call ‘atoms’. These will be defined in GPP, but are not currently programmatically defined. We will assign a default set of sequence steps for each instrument and mode to be used for the atoms, but will manually define these for special case observations that require non-standard sequence portions to be observed such as, for example, the full sequence to be taken in one go.

2.3 Testing and verification

2.3.1 Successful generation of nightly plan:

Run the Scheduler for a single night.

A. Initially run for each site independently. Compare the generated plan with a manual QC plan for that night. This could be historic using a test database or for the present night using the production database. Do for single sets of conditions (over a night), instrument configurations, no ToOs or faults and same database observation set (no changes during the night). May need to use assumptions about required calibrations/time lengths.

B. For each site independently, have a few (>2) QCs make the plans with the same dataset, then compare all the plans, manual + the Scheduler. Do for GN and GS. Initial tests such as these have been called Ghost QC experiments, as the idea was to compare how similar manual QC plans actually are and the typical expected variation. See the Appendix for some representative results.

C. When implemented, run the Scheduler for both sites together.

Specific Tests:

- Check that a plan is generated for various specified conditions for the night. [Req 2.20, 2.21]

- Check that the Scheduler plan does not violate any of the implemented rules [Req. 1.7,3.1,5.1-5.2, 6.1c,f,l,r, 5.5, 5.4]
    - Observations are scheduled in the correct conditions,
    - At allowable times,
    - With available instruments, modes, and components
    - Scheduled non-sidereal observations have ephemerides,
    - Scheduled observations have available guide stars (for non-sidereal observations or due to conditions),
    - Scheduled observations have been set to Prepared (or equivalent state in GPP) or Ongoing
    - Scheduled observations are from programs which have not exceeded their allocated time (plus some buffer)

    - Scheduled observations are from programs which have not expired (outside of their active window)

- Check that prioritization is as expected: Evaluate the different QC plans (see e.g. the Ghost QC experiments), modify Scheduler metric/algorithm as needed in the event it is clearly making poor choices compared to the agreement within the QC plans. Check that it does not differ by more than the QC plans differ from each other, or if it does, evaluate why and whether any modifications are needed. Check that e.g. long programs or observations with long on-sky calibrations have not been penalized. [Req 6.1, 3.9, 7.9]

- Check that all available program types are considered for the plans (eg Q/LP/FT/DD/SV/C/CAL/ENG) and are scheduled when appropriate [Req 3.2 - 3.4]

- Check that Scheduler observations are scheduled on average at AM <= that for QC plans for the same observation sets and conditions variants. [Req 6.8]

- Check that there are no large gaps in the Scheduler plan, compare amount of unscheduled time in the QC and Scheduler plans. Ensure the Scheduler time losses are <= the mean of the QC plans. On average these should be less than 2% of the night based on the requirements. [Req 1.3]

- Check that the Scheduler plan includes Band 4 and specphot standards for very poor conditions [Req 1.8]

- Check that the Scheduler schedules timing window observations appropriately and can schedule around them well. [Req 1.1, 6.6]

- Check that non-sidereal observations are scheduled correctly - when they are accessible at reasonable AM and when guide stars are available. [Req. 5.4,5.6]

- Check that high priority observations being lost are not missed and that target remaining visibility is adequately accounted for. [Req 1.9, 1.10]

- Check that the Scheduler does not schedule low priority observations (e.g. B3, B4) if there are higher priority observations available for the conditions. Compare to manual QC plans. Check that it also does not schedule too many observations in better than requested conditions (also in comparison with the manual QC plans). [Req. 6.1, 1.1.1]

- Check that unstarted B2 and B3 programs are not started if it is no longer possible to complete to the minimum (80% or B3 minimum) level. [Req 6.16]

- Check that the Scheduler does not overly split observations and does not schedule partial observations that can no longer be completed to at least 80%. [Req 6.2, 6.5]

- For partial observations, check that sequences are split at appropriate steps, and check that the Scheduler includes complete atoms with all required calibrations. [Req 1.5, 6.3]

- If the scheduler can suggest GMOS component changes, allow changes by QCs and Scheduler for the given dataset and a conditions ‘forecast’, compare with QC choices and analyze any differences. [Req. 1.9]

- When calibrations can be associated with science observations in the database, check that they are correctly scheduled. [Req 3.6 - 3.8]

- Check that all necessary calibrations have been included in the night plan. [Req 3.6 - 3.9]

- Check that extra unnecessary calibrations are not included in the night plan when a single calibration can be shared between observations. [Req 6.4]

- Check plans assigned as PV nights. Ensure observations have atom sizes < 45min. [Req. 3.5]

- Check that the schedule includes allowed observations outside of the nautical twilight period (NIR observations or calibrations) following instrument and scheduling rules. [Req 1.2]

- Check that observations with relative timing windows are scheduled correctly. [Req 6.7]

- Check that AND and OR groups are correctly scheduled [Req 6.13 - 6.14]

- Check that all instrument specific scheduling rules are adhered to, including for all the current visitor instruments. [Req 4.1, 4.3]

- Check that intra-program priorities for similar RA targets are correctly handled [Req 6.1o]

- For LGS observations, check that it schedules around closure windows well, and where unavoidable, factors in the extra time needed to wait through any closure windows [Req 6.10 - 6.11]

- If users can specify the time period within a night for which the Scheduler should generate a plan, check that this time is scheduled correctly. [Req 1.11]

- Check that the QC interface has all capabilities (and on all supported O/S and browsers): the QC can check the plan and how the plan changes after adjusting conditions and potential instrument/component availability, view (and save) the plan for each set of conditions), examine the scores for each observation (both for those included in the plan and those not), view plots of Elevation-Time and Alt-Az, and manually insert observations into the plan. [Req 1.14, 1.15, 2.11, 2.20-2.26]

- Check that the QC interface provides additional information including lists of unschedulable observations, programs which cannot be completed in the remaining time, and programs which have recently or will soon expire. [Req 2.27-2.29]

- When running both sites together, do checks for each telescope plan as above, plus check: [Req 1.4]
    - All observations are scheduled at an allowed site
    - No observations are scheduled at both sites
    - Observations are scheduled at the more appropriate site
    - A previously started observation at one site is only continued at that same site. [Req 6.18]
    - Associated calibrations are scheduled at the same site/s as the science. If observations from a program are scheduled at both sites, check that calibrations are scheduled at both sites. [Req 6.17]
    - Whether any higher priority observations are missed at a site which could have been avoided by scheduling an allowed observation at the other site. [Req 1.1]
    - Both sites maintain high scores for the nightly plan; there are no systematic trends in scores over multiple tests.

2.3.2 Test of Real-time scheduling mode:

Can use either the test or production database.

A. Start with a snapshot of the database. Run for each telescope independently. Either run in the background, reading in real conditions during the night or, for a past night, provide the changing conditions and fault time losses at the given times along with the starting DB observations and any additional rToOs triggered during the night. Assume

observations scheduled over the night are observed. Will require to keep a log of all scheduler updates.

B. As in A, but run over the course of the night for both telescopes concurrently.

Specific Tests:

- Compare the initial plan with real QC plans for the conditions at the start of the night.
- Redo the checks for Test 2.3.1, but for real-time scheduling. Check the new plans for the remainder of the night after changes are made. [Req. 1.6.6]
- Check that the scheduler can access Resource (systems, instrument, and component availability) and weather information. [Req 1.10, 5.1, 5.2, 5.7]
- Check that the Scheduler updates the plan after any events (any updates to the database/fault/weather/ToO/resources) and check how quickly the Scheduler updates the plan after these changes. This should take < 1min. [Req 1.6, 1.6.2, 1.3.1]
- Check that plans are updated when observations are taking more/less time than expected when running in the background on the actual night [Req 1.6.1]
- Check that when the plan is updated for changes in the weather or database, the new plan includes completion of at least the current observation atom including associated calibrations, whenever possible. [Req 6.3]
- Check that any triggered rToOs are included in the updated plan, but do not replace the same or higher priority observations. [Req. 1.6.4, 7.1 - 7.4]
- Check that the Scheduler suggests stop or abort of the current observation, whichever is more appropriate, when receiving an interrupting rToO and places this as the next (current) observation. [Req 7.3]
- Check that multiple incoming ToOs can be handled well.
- Check that when ongoing observations can no longer be completed to minimum length within a timing window (due to some delay such as a fault or weather changes), that the Scheduler suggests to stop or abort the observation and updates the plan. [Req 2.8]

- Check that non-interrupting rToOs do not cause a stop or abort of the current observation. Check that these are included in the updated plan/s at an appropriate time (based on priorities) within the 24h window and after at least the current ongoing atom observation is completed. [Req 7.2]

- Check that the Scheduler updates when changes occur to the LGS clearance windows. [Req 1.6.5]

- Check that the Scheduler communicates with LTCS and does not schedule LGS or non-LGS targets when it would violate 'first on target' policies. [Req 6.12]

- Check that the Scheduler does not schedule instruments/components/systems that are not available, until becoming available again (eg fault resolved, mode checked out). [Req 5.1 - 5.3]

- Check that rToOs are not sent to a telescope that is not currently accepting rToOs, and check that this information is also available to users. [Req 7.8]

- Check that schedulable high priority observations are not missed (time critical, rToO, losing B1, etc). [Req 6.1]

- For tests during visitor instrument blocks, check that the real-time scheduler can handle observations that are not using the seqexec. [Req 4.4]

- Compare what the observer actually did during the night with what the Scheduler requested over the course of the night.
    - Did the Scheduler find higher priority targets than the observer?
    - For cases of a change in the weather conditions (or fault or rToO trigger), compare the amount of time it took the Scheduler to produce a new plan with how much time was spent by the observer trying to determine what to do next. [Req 1.3.1]

- When 2-site planning is implemented, check that any triggered rToOs which can be observed at either site are scheduled at the most appropriate site, and if missed at that site, is rescheduled for the next available site for as long as the rToO remains active. [Req 7.5, 7.6, 7.8]

- For 2-site planning, ensure that realtime updates take into account both telescopes. [Req 1.4]

- Check for full functionality on all supported O/S and browsers. [Req 2.13]

2.3.3 Longer term planning (validation or simulation mode):

Do with a starting database snapshot. Run using historical data and a snapshot of the test database from the starting date (validation mode), or with a starting snapshot of the production database and simulated conditions over some future period (simulation mode).

- Run over the course of several nights to a week or longer using actual (or simulated) conditions over each night.
- For forward look tests, include the long term forecast in the Visibility calculations.
- Include any ToOs (actual or simulated) and other changes to the DB.
- Include fault time losses and any loss of system/component availability.
- Allow component change suggestions from the Scheduler if/when this is implemented. Otherwise use what was actually installed or provide a component schedule.

If not running in real time, the scheduler should assume the set of conditions at the start of the night for the full night. Then for each event (conditions change, rToO, fault, etc) throughout the night, it should make the appropriate changes to the subsequent schedule. It should not make full nightly plans with prior knowledge of any changes through the night. When running with historical data, assume all observations are specific to the particular site. When running over some forward look/simulated period, test also the planning of both telescopes concurrently with observation sets that can be observed at either site.

Specific Tests:

- Test the use and functionality of simulation mode and on different supported browsers - [Req. 1.16, 1.17, 2.30-2.39]

- Check the progress of B1-3 programs over the period. For validation mode, compare what was actually observed with what the Scheduler planned. Also compare individual plans with the QC plans during that period. [Req 1.1, 6.1]

- Check GMOS component change suggestions if this is implemented. [Req 1.9, 2.38]

- Check that any high priority and time critical observations during that period were scheduled. [Req 6.1]

- Check for and compare any observations that were lost during this period, by either the actual queue scheduling, or by the simulation. [Req 6e]

- Try running over periods of regular queue and also during schedule blocks (e.g. visitor instrument, LGS). Check that blocked instruments are prioritized similarly in the scheduler and actual queue plans. Compare simulations with actual block outcomes. [Req 4.2]

- Check that observations which can never be scheduled (due to visibility or problematic conditions requirements, lack of guide stars, etc) in the remaining time are flagged. [Req. 2.27]

2.3.4 Success over a full semester (validation and simulation modes):

Do for a full simulated (realistic or re-created) semester starting with a test database. For the simulation case, use frequency of conditions, add time for faults affecting systems, number of rToOs and frequency, add in additional programs each month (for DD/FT), etc. Use the simulation case to compare different metrics/algorithms using the same dataset, and also to test the scheduling of both telescopes concurrently with observation sets that could be observed at either site. For re-created semesters (see Semester Re-creation above), over an actual past semester (validation case), use actual weather conditions through each night, the instrument schedule and configurations, timing of significant faults affecting systems, rToO triggers on the correct dates, and FT/DD programs coming online in the correct months.

When the Scheduler can choose instrument configurations, this should be tested rather than simply using the historical GMOS configurations.

Specific Tests:

- Check that simulation and validation modes work as expected [Req. 1.17, 2.1, 2.30 - 2.37]

- Calculate final metric score for each telescope for the semester (where the metric score is based on the fraction of B1/B2 completed programs). Is this similar to or higher than what QCs have accomplished for the particular semester and on average? [Req 1.1.1]
    - If a re-created semester, compare to what was actually accomplished - number of B1 and B2 programs reaching at least 80% completion.

- If a realistic simulated semester, compare program completion (metric score) with the average obtained in previous semesters. Can also use these to compare multiple simulations using different scoring functions, e.g. try identical situations with different scoring functions to find the cases that optimize the final metric score (most B1/2 programs completed).

- Specifically also check (and compare with actual): [Req 1.1.1]

  - The fraction of B1 that reach 100%, 80%. Fraction B1 < 80% (trying to minimize).
  - The fraction of B2 that reach 80%, Fraction at < 80% vs Fraction that gets to 100%. (Was time wasted getting some to 100% that could have been used to get others to 80%?)
    (For GN for previous 10 semesters, on average 62% B2 reached 100%, 21% reached < 80% (excluding LGS and instrument blocked time)
  - The fraction of B3 that were started but ended < 80% (or less than B3 min)
    (For GN for previous 10 semesters, on average 18% of B3 programs were started but ended with "insufficient" data - left at 20-79% complete.)

- Check fraction of thesis programs completed compared to non-thesis. Compare to actual completion fractions. [Req 6.1q]

- Check fraction of DD/FT programs completed (do both retain completion rates as high as actual without added bumps?).[Req 1.1.1, 3.2]

- Check for systematic trends for programs not completed (All long/ all short programs? All have short/few timing windows or no timing constraints? All optical or NIR? All a blocked visitor instrument?). If any trends found, were they also present in the actual semester outcome? [Req. 1.1.2]

- Check that individual observations (< 1%) are not left partially completed. [Req 6.15]

- Check that instrument configuration choices follow frequency rules (how many and how often, and on which days - if Scheduler can make configuration choices). [Req 4.5]

- Check total time lost on sky over a semester when updating plans (< 1%) or for gaps in plans (< 2%) is < 3%. [Req 1.3, 1.3.1]

- Check that all scheduled ('observed') observations/programs also had a full set of calibrations scheduled by the end of the semester. [Req 3.6 - 3.7]

- For observations that could be scheduled at either site, check that for either site the science was scheduled ("observed") at, the full set of calibrations was also scheduled. [Req 6.17]

- Check the individual nightly plans to ensure other scheduling rules were not violated (all calibrations scheduled and correctly scheduled, relational groups correctly scheduled, timing frequency correctly adhered to, etc). [As in Test 2.3.1]

2.3.5 Testing in Real Time:

Test in simulation mode over a period of time (once resource/component/weather servers are running), before release for general use, running in background each night with a test database and assuming observations are taken as scheduled by the Scheduler. The Scheduler should respond to changing conditions, faults, ToOs. Start with the data in the production database, and add FT/DD programs to the test database as they are set up. Use actual conditions, faults, ToOs. Use actual GMOS components or as suggested by Scheduler (preferably the latter if/when possible).

Specific Tests:

- Check manual vs Scheduler plans daily ahead of the night for any issues.

- Check what was accomplished during night vs what the Real-time Scheduler suggested.

- Compare fractions of B1/2/3 completed (actual vs simulated) as go through the period. [Req 1.1.1]

- Compare numbers of targets lost and even whole programs that cannot be completed to 80/100% as the test continues. [Req 1.1.1]

- Check final fraction of B3 programs that were started but may not get to useful level (actual and simulated). [Req 1.1.1, 6.16]

- Compare time lost on sky due to decision making (observer vs scheduler) or gaps in either plan. [Req 1.3, 1.3.1]

- Check that the scheduler does not suggest changes to observations requiring better conditions before conditions have stabilized. Check expected time losses for cases of

the scheduler suggesting such changes when conditions have not actually improved. [Req 1.3.2, 1.6.3]

- Check that the scheduler does suggest changes in poorer conditions after the appropriate time delay to be sure that conditions actually have degraded. Compare with how long it takes the observer to change plans. [Req. 1.6.3]

- Check for any systematic trends (by either the manual or Scheduler queues) - (in completion vs. program length, GMOS components, timing windows, program type - FT/DD, number/length of calibrations, etc) [e.g. Req 1.1.2, 3.2, 3.9]

- When running in real time, test the use of a manual plan using the system with only reception of rToOs. Check that it is clear to the observer that they are in manual mode. Check that any rToOs are received as expected. [Req 1.13, 2.11]

- Check that observations which can be scheduled at either site are appropriately scheduled: rToOs at the first available site if interrupting or at the most appropriate site if non-interrupting (taking into account the current observation priority, current site conditions, and, if relevant, time loss caused by the trigger), time critical at the most appropriate available site (per conditions and other observation priorities), non-time critical observations per observation priorities, conditions, schedules, AM and visibility. [Req. 1.4, 1.4.1, 1.4.2, 7.5]

- Check that (un)available resources are correctly accounted for in the plans [Req 5.1-5.3]. Check that updates by the Observer to the available resources are accounted for and plan updated as needed. [Req 2.12]

- Check that the information from the conditions server and model predictions are correctly used to schedule observations (at appropriate AM, sky location, and times within correct seeing, sky brightness, CC, and WV constraints). [Req 5.8 - 5.10, 5.11 - 5.13]

2.3.6 Using real-time scheduling:

Use Real-time scheduler at night with manual plans only as back-up. (The above test (2.3.5) is still in simulation mode since observations are not necessarily actually observed. This test has the observer running the real-time generated plan).

- Check that Schedule/Explore/Observe communicate and run as expected [Req 2.1, 1.6.2]

- Check that all calibrations are correctly scheduled by the system, including daytime calibrations following a night of observing [Req 3.7]

- Check that interrupting ToOs are charged correctly for any time loss in the current observation when atoms cannot be completed. [Req 7.7]

- Check that Observers can still load and use a static manual plan. [Req 1.12]

- Check ease of use by Observers, get Observer feedback.
    - Check that it is always clear to observers what to observe [Req 2.3]
    - Check that the plots (elevation vs time, Alt-Az with wind) are easy to view and aid the observer [Req 2.2]
    - Check that it is clear what to do when a rapid ToO arrives [Req 2.9.1, 7.2 - 7.4]
    - Check that eavesdropping requests are clearly visible and a warning/reminder provided to the observer ahead of the scheduled slew time. [Req 6.9]
    - Check that all information observers require about an observation is provided and/or easy to find. [Req 2.5]
    - Check that the remaining time in an atom and scheduled sequence are clearly indicated. [Req 2.4]
    - Check that when updates are made to the plan due to changes in the weather, it is clear to the observer what to do [Req 2.7, 2.9.3]. Check also that observers are comfortable with the timing of the proposed changes [per Req. 1.6.3] or can request changes as needed. [Req, 2.10]
    - Check that when other non-imminent changes to the plan occur that the observer has been made sufficiently aware of the changes. [2.9.2]
    - Check that < 1.5% of the time on average is lost to failed attempts at better conditions [Req 1.3.2]
    - Check that observers can easily skip an observation when needed and the plan is correctly updated [Req 2.6]

3. Acceptance criteria

---

1. The Scheduler simulation for an entire semester completes at least as many B1 and B2 programs as manual QC has accomplished on average in past semesters.
2. The minimum set of requirements for each release met, with full requirements defined in the “GEMMA TDA Scheduler Requirements” document Section 6 “Adaptive Queue Scheduler Requirements”
3. The scheduler correctly follows all applied scheduling rules.
4. Ease of use by observers verified, observers sign off on the Real-time Scheduler.
5. Core QCs sign off on the real-time and simulation modes of the Scheduler.

Appendix - GhostQC Analysis

Several “ghostQC" trials have been run to compare the variance in manually-generated plans and to compare QC plans with the prototype scheduler. Between 13-18 October 2020 Gemini South queue coordinators (QCs) took advantage of the Gemini South COVID19 shutdown to make plans for Gemini North, which was operating. On April 7, 2021 all seven GN QCs created plans for Gemini North.

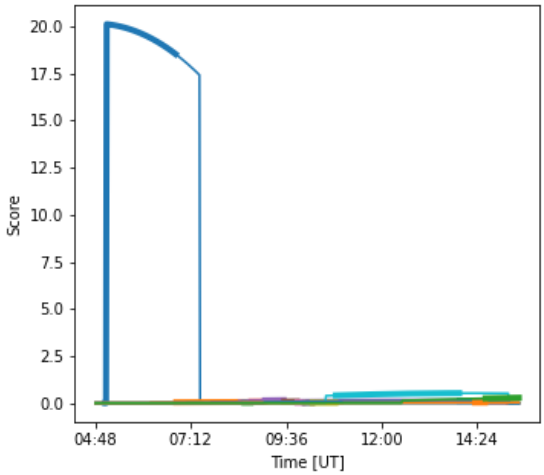
The plans from the ghost QCs and the scheduler were analyzed with the same procedure, calculating both the total score, just the metric component of the score (using the Parabolic 2 option from the Queue Scoring and Metric document), and the fraction of the night scheduled. For this analysis the score is the product of the metric, visibility fraction, and hour angle weight.

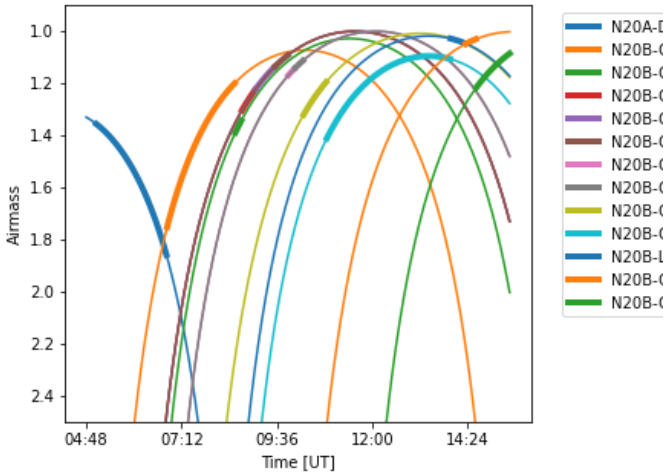
The initial analysis is done for the four most common variants: IQ70CC50, IQ70CC70, IQ85CC50, and IQ85CC70. For each plan, the metric and score are calculated for each observation and summed. The mean, standard deviation, fractional variation (standard deviation/mean), min, and max for the different QC plans are then calculated. The fraction of the night scheduled is also determined as another measure of schedule quality. These statistics are presented in Table 1 and shown graphically in Figure 2.

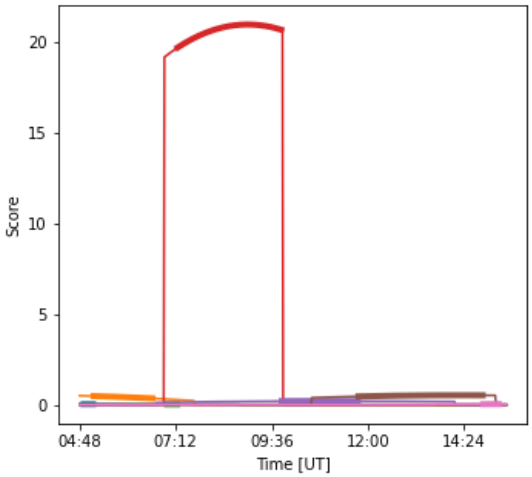
Inspection of Figure 2 and Table 1 shows many similarities between the QC plans but occasional large variations in the score and metrics. The differences are typically between 10% and 50%. These variations are easily seen in the plots of score and airmass versus time in Figure 1. Each row in the figure is a plan from a different QC or the automatic scheduler (see below) for the IQ70CC50 variant from 2020 Oct 13. The fraction of the night scheduled tends to be high, above about 90%. This is one cause for some of the large variations in the score and metric sums.

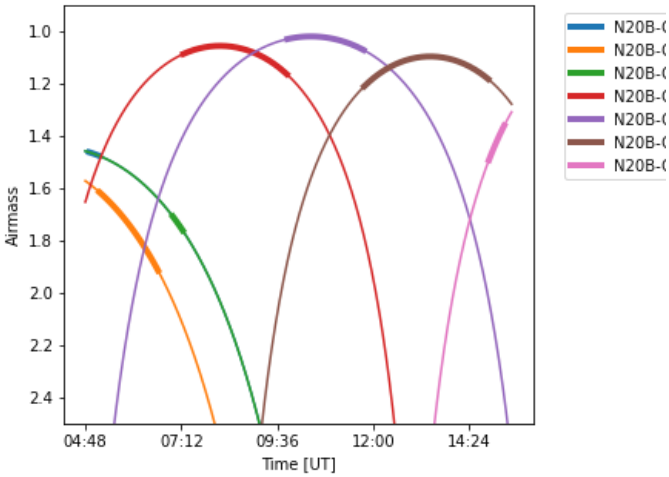
Finally, we run the current version of the GreedyMax scheduling algorithm to see how its results compare with the human plans. The algorithm attempts to maximize the summed score for a night. It evaluates the night in its entirety by scheduling the observation with the highest score and then iteratively filling in around it by selecting the highest-scoring observation in a given time interval. The pool of observations is taken from the ODB XML backup written in the morning after the night in question so that it includes all the observations that were available to the QC.

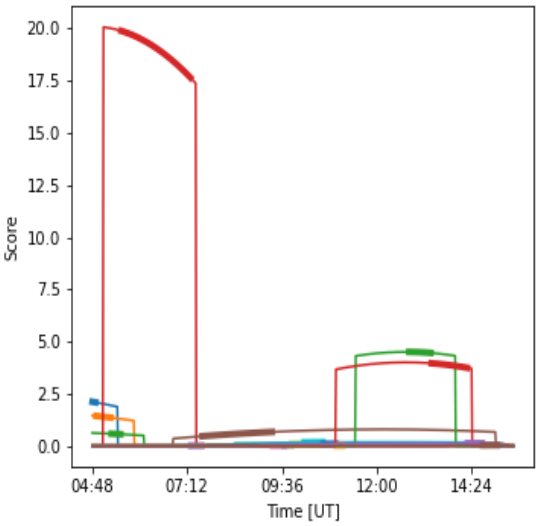
The scheduler results are also included in Figures 1 and 2 and Table 1. In general GreedyMax makes plans that are competitive with human plans as judged by the current criteria. The summed scores tend to be higher than those from the best QC plans and the summed metrics well within the range of the QC results. This suggests that the QCs are using different criteria with more emphasis on the straight metric for selecting observations and that the algorithm needs some adjustments. These comparisons will be repeated as new terms and other changes, e.g. using an instrument calendar in the visibility calculations, are made to the scoring algorithm (see the Queue Scoring and Metric document).

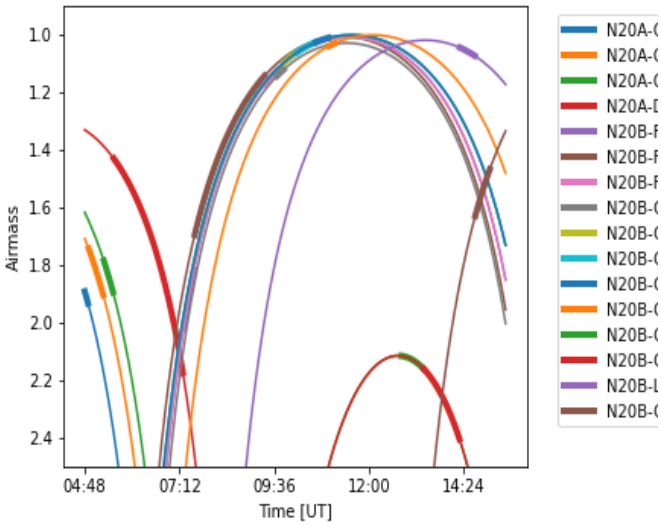

Score
Time [UT]
04:48
07:12
09:36
12:00
14:24

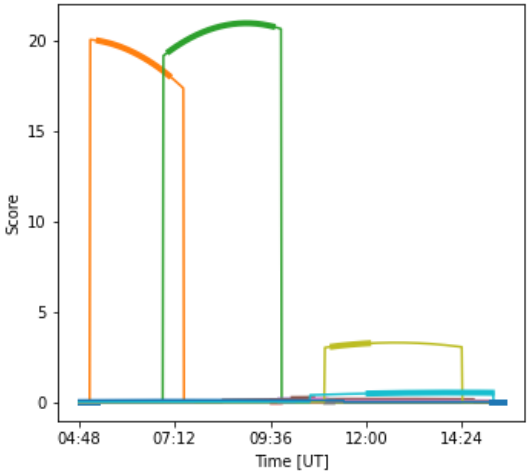

Airmass
Time [UT]
04:48
07:12
09:36
12:00
14:24
N20A-D
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-L
N20B-(
N20B-(

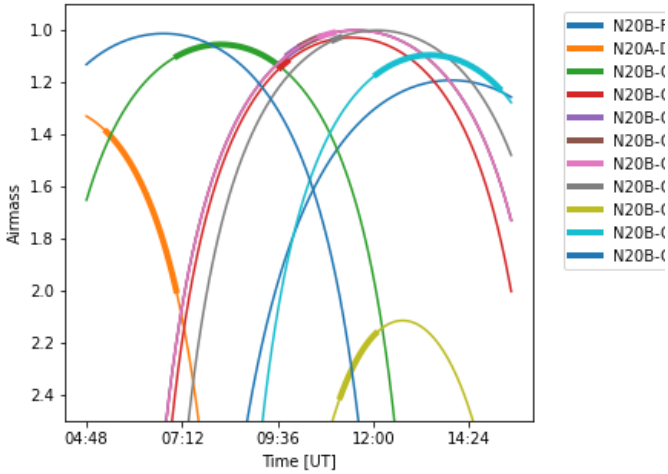

Score
Time [UT]
04:48
07:12
09:36
12:00
14:24


Airmass
Time [UT]
04:48
07:12
09:36
12:00
14:24
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(


Score
Time [UT]
04:48
07:12
09:36
12:00
14:24


Airmass
Time [UT]
04:48
07:12
09:36
12:00
14:24
N20A-(
N20A-(
N20A-(
N20A-D
N20B-F
N20B-F
N20B-F
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-L
N20B-(


Score
Time [UT]
04:48
07:12
09:36
12:00
14:24


Airmass
Time [UT]
04:48
07:12
09:36
12:00
14:24
N20B-F
N20A-D
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(
N20B-(

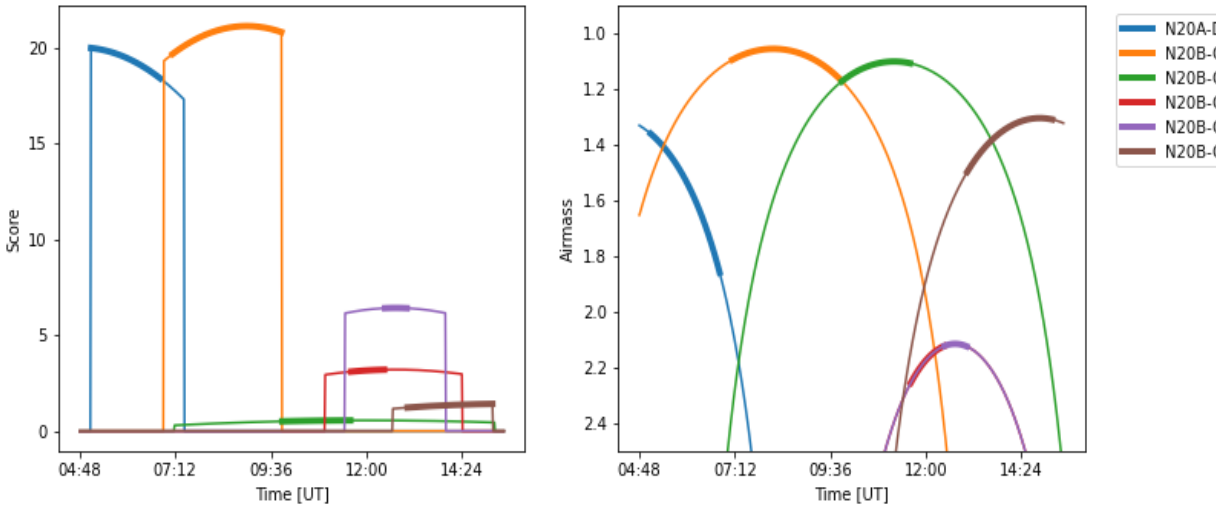


Figure 1. Score and airmass versus time for IQ70CC50 plans on 2020 October 13. Each row is the plan from a different QC or the greedy-max scheduler. Many of the plans have similarities but large differences are obvious. The greedy-max plan is shown in the bottom row, the fact that it is hard to distinguish from the human-generated plans indicates that it is generating qualitatively reasonable results.

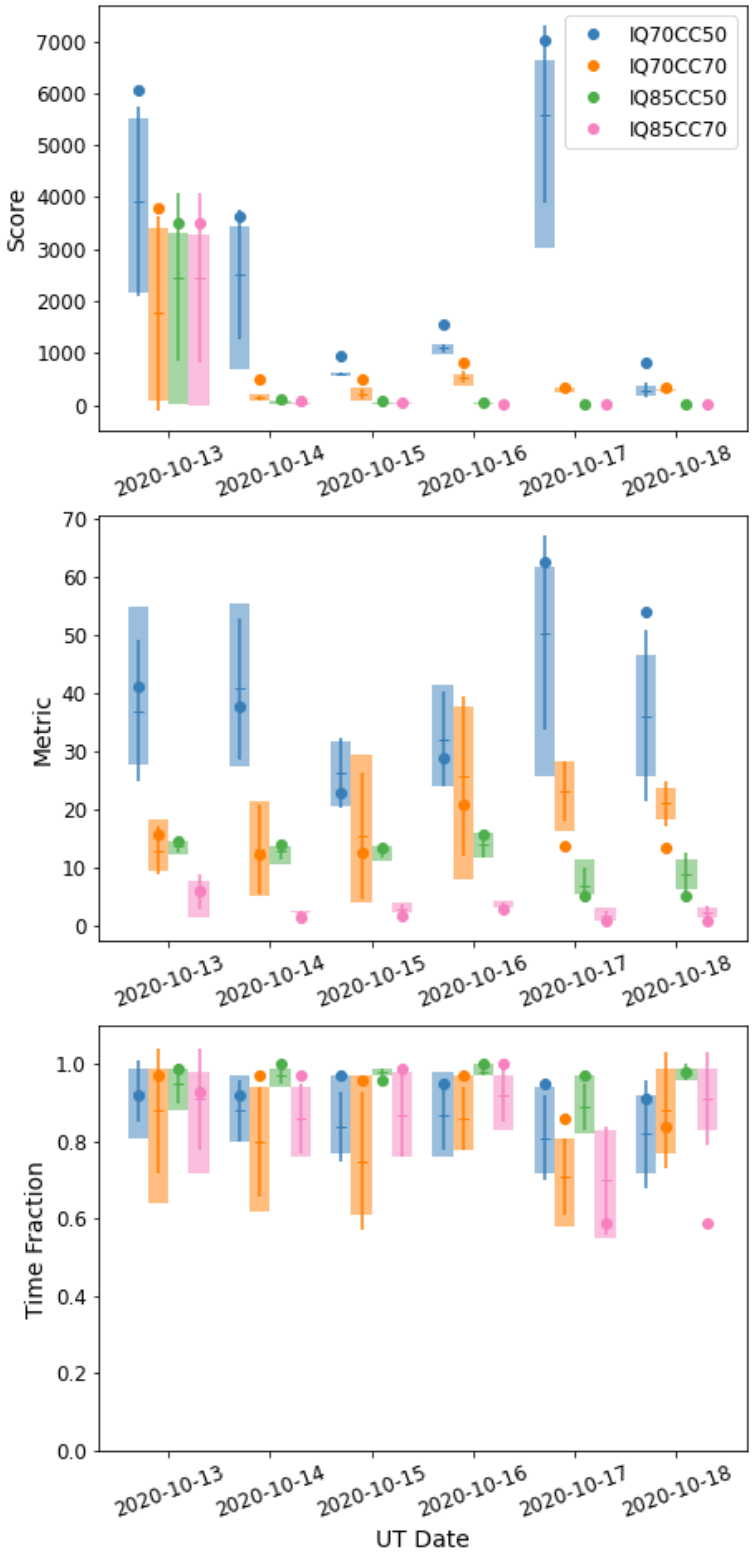

IQ70CC50
IQ70CC70
IQ85CC50
IQ85CC70
Score
Metric
Time Fraction
UT Date
2020-10-13
2020-10-14
2020-10-15
2020-10-16
2020-10-17
2020-10-18

Figure 2. Graphical representation of the data in Table 1. Rectangular bars show the min/max ranges for the QC results. The horizontal bar and vertical error bars indicate the mean and standard deviation. The circular symbols show the results of the greedy-max scheduler.

Table 1. Statistics from the ghost-QC analysis.

2020 Oct 13

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	3916.64	1828.77	0.47	2171.28	5527.10	6071.09
Metric	36.91	12.17	0.33	27.56	54.81	41.17
Time Fraction	0.93	0.08	0.09	0.81	0.99	0.92
IQ70 CC70						
Scores	1762.15	1884.41	1.07	74.97	3405.98	3793.50
Metric	12.92	4.20	0.33	9.55	18.24	15.77
Time Fraction	0.88	0.16	0.18	0.64	0.99	0.97
IQ85 CC50						
Scores	2457.66	1620.26	0.66	27.85	3303.93	3504.69
Metric	13.72	1.05	0.08	12.19	14.59	14.61
Time Fraction	0.95	0.05	0.05	0.88	0.99	0.99
IQ85 CC70						
Scores	2452.71	1633.25	0.67	3.13	3288.62	3492.23
Metric	5.85	3.00	0.51	1.38	7.79	6.11
Time Fraction	0.91	0.13	0.14	0.72	0.98	0.93

2020 Oct 14

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	2508.42	1242.38	0.50	678.76	3447.42	3619.43
Metric	40.76	12.10	0.30	27.52	55.53	37.67
Time Fraction	0.88	0.08	0.09	0.80	0.97	0.92
IQ70 CC70						
Scores	137.99	56.91	0.41	78.34	212.43	491.24
Metric	13.16	7.62	0.58	5.23	21.55	12.38
Time Fraction	0.80	0.14	0.17	0.62	0.94	0.97
IQ85 CC50						
Scores	49.19	17.18	0.35	33.39	69.64	105.43
Metric	12.79	1.50	0.12	10.56	13.77	13.97
Time Fraction	0.97	0.02	0.02	0.94	0.99	1.00

IQ85 CC70

Scores	29.74	29.16	0.98	4.50	59.53	76.93
Metric	2.36	0.18	0.08	2.23	2.62	1.54
Time Fraction	0.86	0.09	0.11	0.76	0.94	0.97

2020 Oct 15

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	592.93	31.58	0.05	556.67	632.03	941.69
Metric	26.40	6.00	0.23	20.55	31.69	22.77
Time Fraction	0.84	0.09	0.11	0.77	0.97	0.97
IQ70 CC70						
Scores	206.13	105.85	0.51	86.95	338.93	484.00
Metric	15.33	10.86	0.71	3.96	29.37	12.67
Time Fraction	0.75	0.18	0.23	0.61	0.97	0.96
IQ85 CC50						
Scores	36.73	11.28	0.31	24.03	51.30	74.79
Metric	12.92	1.14	0.09	11.24	13.77	13.39
Time Fraction	0.98	0.01	0.01	0.97	0.99	0.96
IQ85 CC70						
Scores	14.10	15.95	1.13	4.89	37.90	45.80
Metric	2.87	0.75	0.26	2.22	3.95	1.84
Time Fraction	0.87	0.11	0.12	0.76	0.98	0.99

2020 Oct 16

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	1092.99	91.13	0.08	969.90	1169.70	1555.12
Metric	32.07	8.17	0.25	23.97	41.46	28.82
Time Fraction	0.87	0.09	0.11	0.76	0.98	0.95
IQ70 CC70						
Scores	537.04	107.52	0.20	377.13	610.16	823.89
Metric	25.75	13.66	0.53	7.96	37.56	20.75
Time Fraction	0.86	0.08	0.09	0.78	0.97	0.97
IQ85 CC50						
Scores	38.26	8.39	0.22	30.15	45.59	63.74

Metric	14.03	2.23	0.16	11.57	15.92	15.65
Time Fraction	0.98	0.01	0.01	0.97	1.00	1.00
IQ85 CC70						
Scores	7.76	1.61	0.21	6.31	9.69	12.11
Metric	3.56	0.50	0.14	3.13	4.14	2.72
Time Fraction	0.92	0.07	0.07	0.83	0.97	1.00

2020 Oct 17

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	5593.16	1719.12	0.31	3022.23	6628.62	7024.90
Metric	50.37	16.78	0.33	25.63	61.79	62.54
Time Fraction	0.81	0.11	0.13	0.72	0.94	0.95
IQ70 CC70						
Scores	295.27	51.14	0.17	250.74	342.91	343.03
Metric	23.10	5.13	0.22	16.26	28.20	13.69
Time Fraction	0.71	0.10	0.14	0.58	0.81	0.86
IQ85 CC50						
Scores	19.06	8.44	0.44	11.04	30.75	30.85
Metric	6.99	3.03	0.43	5.34	11.53	5.16
Time Fraction	0.89	0.06	0.07	0.82	0.97	0.97
IQ85 CC70						
Scores	6.20	1.76	0.28	4.42	8.61	9.18
Metric	1.74	0.96	0.55	0.96	3.14	0.74
Time Fraction	0.70	0.14	0.20	0.55	0.83	0.59

2020 Oct 18

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	283.50	136.43	0.48	187.03	379.97	817.95
Metric	36.07	14.67	0.41	25.70	46.44	54.11
Time Fraction	0.82	0.14	0.17	0.72	0.92	0.91
IQ70 CC70						
Scores	285.33	19.71	0.07	271.40	299.27	329.54
Metric	21.11	3.87	0.18	18.37	23.84	13.28
Time Fraction	0.88	0.15	0.17	0.77	0.99	0.84

IQ85 CC50						
Scores	24.26	10.40	0.43	16.91	31.61	33.95
Metric	8.90	3.76	0.42	6.24	11.56	5.19
Time Fraction	0.98	0.02	0.02	0.96	0.99	0.98
IQ85 CC70						
Scores	7.86	1.29	0.16	6.95	8.78	11.43
Metric	2.30	1.19	0.51	1.46	3.14	0.74
Time Fraction	0.91	0.12	0.13	0.83	0.99	0.59

2021 April 7

Stat	Mean	Std	Std/Mean	Min	Max	Greedy-Max
IQ70 CC50						
Scores	1460.82	261.67	0.18	1167.96	1736.17	2570.84
Metric	61.70	14.30	0.23	47.18	89.36	59.86
Time Fraction	1.02	0.02	0.02	1.01	1.05	0.94
IQ70 CC70						
Scores	1098.48	245.31	0.22	895.70	1428.21	2111.90
Metric	29.00	4.21	0.15	25.64	34.55	30.90
Time Fraction	1.01	0.04	0.04	0.95	1.05	0.97
IQ85 CC50						
Scores	132.62	49.02	0.37	90.25	227.92	269.69
Metric	31.17	4.25	0.14	23.28	35.90	30.54
Time Fraction	0.97	0.05	0.05	0.92	1.02	1.00
IQ85 CC70						
Scores	58.93	39.67	0.67	38.44	139.52	104.27
Metric	4.60	0.54	0.12	4.00	5.23	4.73
Time Fraction	0.95	0.06	0.06	0.87	1.00	0.99