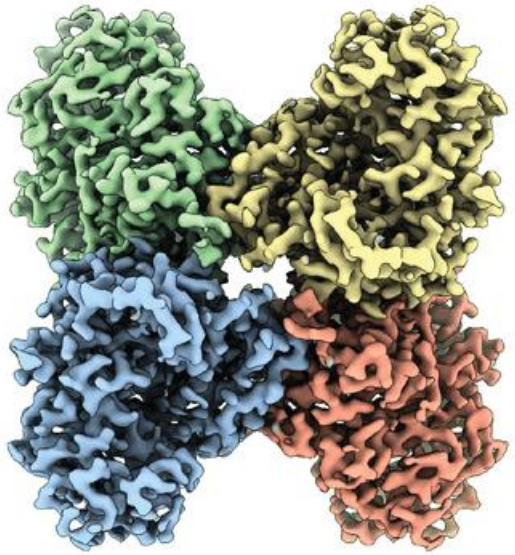
Update for Intel End2End AI Benchmarking

Martyn Winn
17th Sept 2020

Benchmark: Rabbit muscle aldolase



~150 kDa homotetrameric enzyme.

2.13Å resolution.

Involved in gluconeogenesis, the Calvin cycle and glycolysis.

Common test protein - can be bought off shelf.

Chosen because:

- Workflow well-described in https://doi.org/10.1016/j.yjsbx.2020.100020
- Original images publically available on EMPIAR
- Dataset from 200kV microscope - might be different signal / noise characteristics from the 300kV dataset.
- Several optical aberrations make image analysis more challenging.

ML relevance:

Images	Denoising, motion correction	√
Particle picking	More or smarter choice	√
Reconstruction	Low res model known	✕
Interpretation	High res so building easy	✕

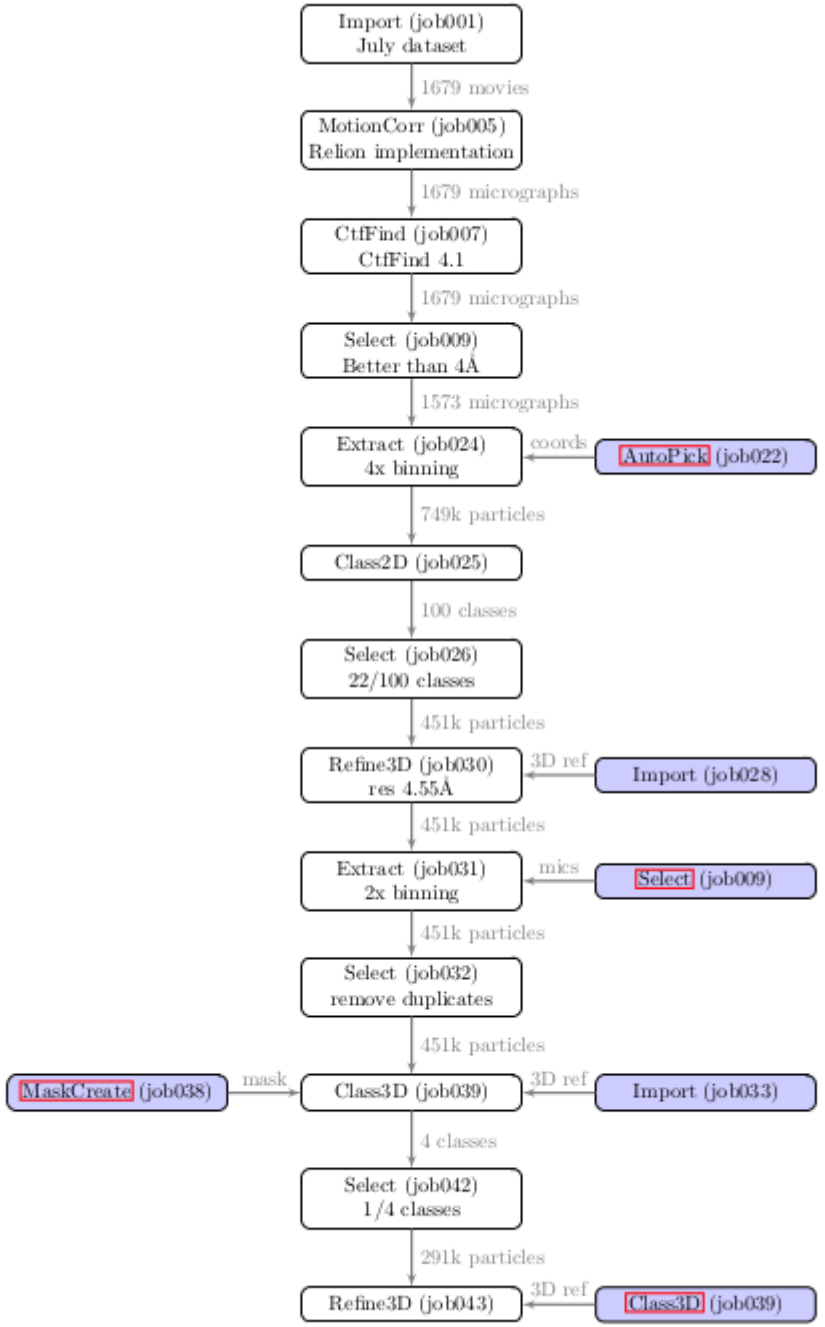
Workflow

- So far, following recommendations of paper
- See right - missing job numbers imply extra tests or failed jobs

- Job parameters are in job.star
- Command used in run_submit.script
- Deposit these in EMPIAR?

Automating:

- Relion Schedules
- New Relion/CCP-EM framework



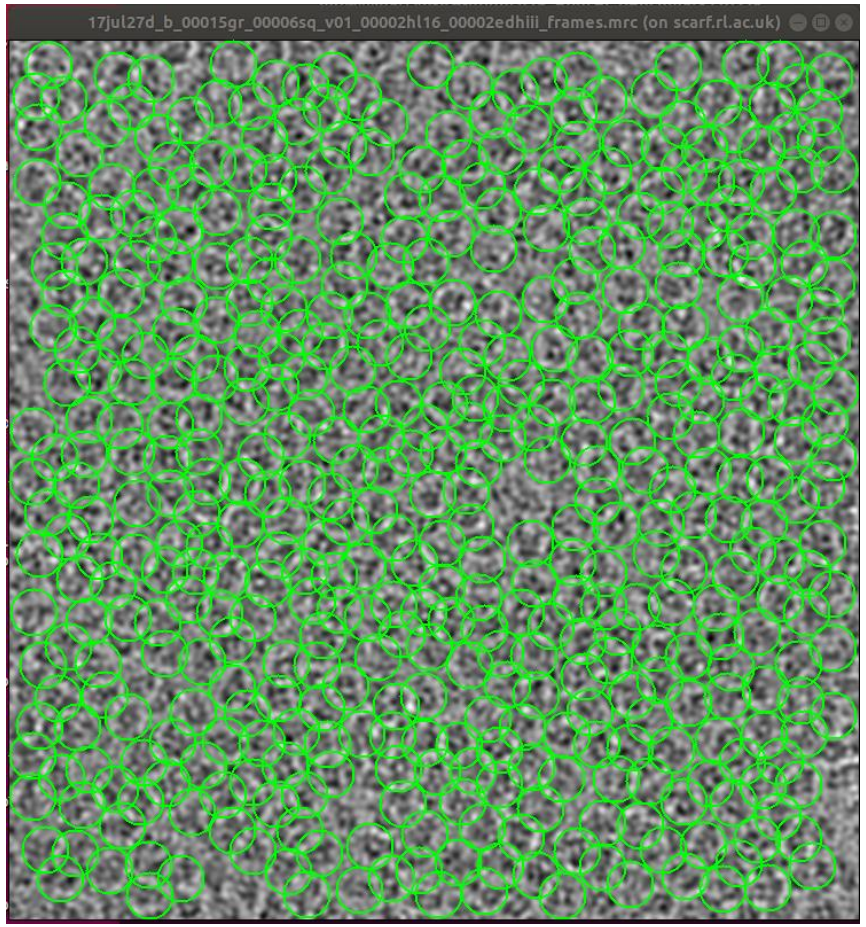
Job statistics

- Jobs from workflow + some extras included as timing examples
- Heavy jobs are Class2D, Class3D and Refine3D – can only sensibly be run on GPU
- Timings for same job are consistent (compare job034, job036, job039)
- Otherwise, timings vary with number of particles, sampling, speed of convergence, etc etc
- Original images 337 GB, summed movies 113GB, extracted particles 58 GB and 137 GB (increased sampling)

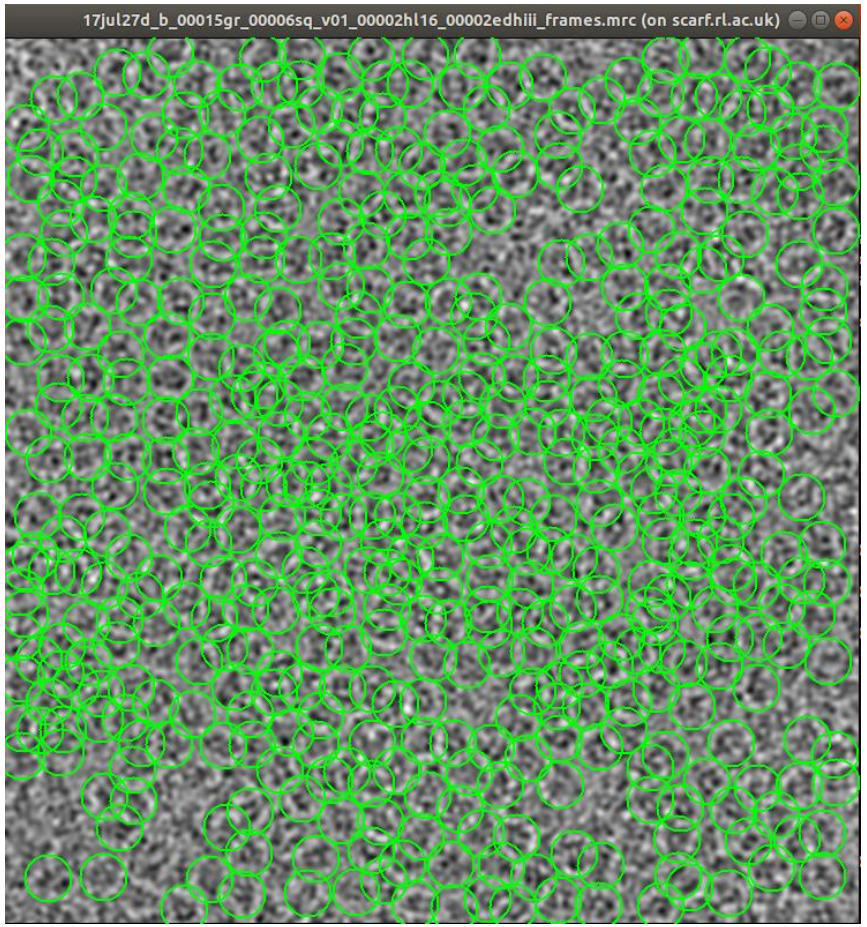
Job ID	Description	Where run	MPI/th/GPU	Runtime	Data produced	Notes
job001	Import	interactive	n/a	negligible	1 MB	Links to 337 GB in Movies
job005	Motion correction	cn488 (scarf15)	1 / 12 / 0	Not recorded	113 GB	No GPU for Relion implementation
job007	CtfFind	cn880 (scarf18)	12 / 0 / 0	2m:03s	3 GB	Ctffind not GPU
job009	Select	interactive	n/a	negligible	2 MB	
job013	Reference import	interactive	n/a	negligible	32 MB	
job022	Autopick	cn793	16 / 0 / 0	2h:59m:50s	219 MB	
job024	Extract 4x	4xcn778 () + 9xcn833 () + 3xcn884 ()	16 / 0 / 0	20m:01s	58 GB	749k particles extracted
job025	Class2D	1xcn630 + 4xcn657 + 4xcn658	9 / 6 / 0	27h:16m:56s	12 GB	100 2D classes
job026	Select	interactive	n/a	negligible	262 MB	
job028	Reference import	interactive	n/a	negligible	11 MB	
job029	Refine3D	3xcn659 + 3xcn663 + 3xcn664	9 / 6 / 0	2 days (timeout)	4 GB	12/25 cycles completed
job030	Refine3D	cn2g16	5 / 1 / 4	16h:16m:11s	7 GB	Repeat 29 with GPU
job031	Extract 2x	12xcn540 + 4xcn648	16 / 0 / 0	18m:45s	137 GB	451k particles extracted
job032	Select	interactive	n / a	negligible	269 MB	
job033	Reference import	interactive	n /a	negligible	80 MB	
job034	Class3D	cn2g19	5 / 1 / 4	7h:29m:03s	16 GB	Poor result
job036	Class3D	cn2g19	5 / 1 / 4	7h:28m:54s	16 GB	Poor result
job037	Refine3D	cn2g19	5 / 1 / 4	22h:55m:54s	10 GB	Not used
job038	MaskCreate	interactive	n / a	negligible	80 MB	
job039	Class3D	cn2g19	5 / 1 / 4	7h:57m:26s	16 GB	4 3D classes
job042	Select	interactive	n / a	negligible	131 MB	

CryOLO

Tested on one example image. Spot The Difference (there are many).



Relion template-based Autopick



CryOLO using general model (imported into Relion)