

advancing structural science

What's Up CCDC Customer Update

23 January 2020





Today's presenters



Carmen Nitsche

General Manager, CCDC Inc



Juliette Pradon, PhD Research & Applications Scientist Ultra-large docking: scaling GOLD



Jonathan Betts, PhD MBA Director of Business Development CCDC update: 2020 outlook



Ioana Sovago, PhD Research & Applications Scientist

H-bond quick view

Overview

A regular update on what's happening with CCDC and an opportunity for your news/feedback and suggestions:

- CCDC update: 2020 outlook
- Ultra-large docking: scaling GOLD
- H-bond quick view
- The floor is yours



CCDC Update: 2020 Outlook



Jonathan Betts

Business Development Director



5

2019 highlights

- Welcomed 20 new customer organisations into the CCDC community
- Organisational
 - Discovery and Materials Science teams
 - Product management
 - Fully implemented Agile
 - Built customer success team
- New Scientific Advisory Board
- Four significant product releases
- CSD surpassed 1M entries
- Great engagement through UGMs and meetings
 globally

6

Powering into the 2020s

We want to continue to build on the success of 2019 in pursuit of our purpose to:

- Bring together the world's knowledge on molecular structure to promote new science and help create life changing products
- We do this by continuously driving knowledge creation by:
 - Delivering scientific software that uniquely combines enriched data with advanced search and analysis
 - Promoting collaboration amongst the world's leading scientists
 - Encouraging a new generation of leaders in structural science



CCDC Portfolio Roadmap 2020

Data:		CSP Landsca	pe Visualisatio	n		
Insights:	CSD D	eep Insights				
Community:	P	orous Materials			- * (<i>1</i>	
Core:	CSD Sketcher C	SD-Search				-
Discovery:	D	iscovery Search		Discovery Store		
Materials:	Crystallisability Prediction		Co-Crystal De	esign		
Particle:	Impro	ved surface analys	sis			
Integration:	CSD KNIME Nodes	Pipelir	ne Pilot Compo	onents		
Foundation:	Database Evolution	User Exp	perience			
						Č

GOLD: How to run ultra-large jobs on cloud resources





Juliette Pradon

Research & Applications Scientist



Ultra-large docking

- Virtual screening of *hundreds of millions to billions* of compounds
- Context:
 - Availability of massive libraries of real compounds, e.g. Enamine REAL database, ~700 million pharma-oriented molecules
 - Availability & reduced cost of large/distributed computing resources clusters, cloud
- Challenges:
 - Preparation of compound libraries
 - Deployment of screening job onto computing resources
 - Speed of screening library
 - Size of output from virtual screen
 - Analysis & visualisation of screening results

CCDC

Ultra-large GOLD – vision

- Input/output: to address the disk usage issues due to the required unzipping of ligand files & the large number and size of files being output
- Cloud: to improve deployment and execution on common highperformance computing platforms
- Fast: to screen ultra-large virtual libraries within an acceptable timeframe for lead generation, using appropriate distributed computing resources



Ultra-large GOLD, input/output – achieved

- Input/output: to address the disk usage issues due to the required unzipping of ligand files & the large number and size of files being output
 - No need to unzip: the script provided for job submission DOES allow for ligand files and/or directories to be zipped
 - e.g. as *ligands.sdf.gz* files downloaded from ZINC website
 - No extraneous output files: with the MIN_OUT write option, output and retain only best_ranking.lst & gold.log files
 - e.g. VEGFR-2 DUD set with 105 actives & 3204 decoys: full output 601.7 MB, min_out output 1.5 MB

Ultra-large GOLD, cloud – achieved

- Cloud: to improve deployment and execution on common highperformance computing platforms
 - CCDC provides a GOLD Docker image for easy deployment onto a Kubernetes cluster, which can be run on many common cloud platforms e.g. Azure, AWS
 - CCDC provides two Python scripts (submit_tasks.py & collect_results.py) as examples of how to batch up the GOLD jobs & how to collect results from the queues
 - User documentation: <u>https://www.ccdc.cam.ac.uk/support-and-</u> resources/ccdcresources/GOLD_Cloud_Computing_User_Guide.pdf
 - Licensing: GOLD requires access to a floating licence server (bare metal install, not a VM), for all unlimited CSD-Discovery & CSD-Enterprise users



Ultra-large GOLD, cloud – further details

- Recommendation: IT personnel sets up and manages the Kubernetes cluster, and the scientist provides protein + ligand + GOLD settings relevant for the ULD job to be run
- Requires two local machines: one for GOLD job submission & results collection, and one for the licence server
- Requires one Kubernetes cluster: made up of one RabbitMQ pod and many GOLD worker pods (tested with over 1,300 GOLD pods)
- Requires several sets of credentials (secrets): for RabbitMQ (admin/user/cookie), for setting up workers, & licence server URL

Ultra-large GOLD, cloud – further details

- GOLD jobs are generated locally and pushed to a message queue running on a RabbitMQ pod in the K8s cluster
 - RabbitMQ comes standard from DockerHub
 - See documentation for how to set up RabbitMQ
- When setting up K8s cluster, Docker image for GOLD worker is copied onto pod from CCDC's container image registry once, then is cached & cached copy used for all subsequent GOLD worker.
- GOLD worker pods will pick up jobs off the job queue and return results to the RabbitMQ results queue
- License server will need to communicate with GOLD worker pods
 - Requires external IP address, must be accessible, think about firewall (valid for all local/cloud connections!)



Ultra-large GOLD, cloud – further details

- Results are retrieved from the RabbitMQ results queue locally by running the provided collect_results.py script
 - Collect script runs all the time, connects to the RabbitMQ results queue, and if there are any results, it writes these to disk in local collect machine & then deletes them from RabbitMQ queue
 - Output gold_results directory contains x directories with a tmp name (i.e. one directory per batch of 2,000 compounds, each with a bestranking.lst file listing for each compound its name & fitness score)
 - Output gold_results directory also contains a results.lst file with the overall top-scoring 1,000 results



Ultra-large GOLD, fast – achieved

- Fast: to screen ultra-large virtual libraries within an acceptable timeframe for lead generation, using appropriate distributed computing resources
 - Screened 130 million ZINC compounds (as 50,229 zipped batches of 2,000 ligands) against COMT DUD target in 40 hrs 17 mins (1.979 secs/ligand)
 - Deployed onto a Kubernetes cluster with 99 nodes of type F16s_v2 (i.e. 16-core virtual machines), equating to 1372 GOLD Worker pods, with a ~90% CPU load
 - GOLD ULD settings: MIN_OUT option; Library Screening GA settings (10% search efficiency)
 - Results: collection script resulted in 50,229 directories written to local machine (50,229 *bestranking.1st* files total) and top scoring 1,000 compounds written in overarching *results.1st* file
 - Total cost on Azure: £2,200

Note: upcoming white paper describing this virtual screen in greater detail

CCDC

CSD-Materials Updates

H-Bond Coordination Quick-View



Ioana Sovago Applications Scientist Materials Science Team



18

2020.0 CSD Release

• CSD-Materials

- H-Bond Coordination Quick-View
- Colored Hydrogen Bond Propensity (HBP) chart

😵 H-bond Coordination Quick-view							×
				Up	date res	ults H-	bond criterion
INTRO	^		Atom (D/A)	= 0	= 1	= 2	= 3
The table shows calculated likelihoods for allowed coordination numbers for each donor and acceptor observed		1	N2 of acyclic_a	0.081	0.889	0.030	0.000
in the current structure, computed using CSD derived models.		2	N1 of acyclic_nh	0.547	0.453	0.000	0.000
ighlighted table cells indicate the likelihood for the bserved coordination number for the atom of that row. reen highlighting indicates a maximum likelihood is bserved. a bightighting indicate there is a mare likely alternative		3	O1 of acyclic_a	0.319	0.660	0.018	0.002
		4	O2 of nitro (a)	0.888	0.105	0.006	0.000
coordination number for that atom.		5	O3 of nitro (a)	0.888	0.105	0.007	0.000
		6	S1 of cyclic_thio	0.974	0.026	0.000	0.000
SUMMARY		7	S2 of cyclic_thio	0.974	0.026	0.000	0.000
H-Bond analysis	~						
Total # models: 596							
Mean highlighted co-ordination likelihood: 0.769264							Save

Hydrogen bond networks



)C

H-Bond Coordination Quick-View

- Quick assessment of the likelihood of H-bond behaviour based on coordination numbers in the observed structure
- Green highlighting indicates that the observed outcome is optimal based on CSD-derived likelihoods
- Red highlighting indicates there is a more optimal coordination outcome for that donor or acceptor based on CSD data
- Released to CSD-Materials users in Mercury in the 2020.0 CSD Release

H-bond Coordination Quick-view							×
					Update res	ults H-bon	d criterion
INTRO	^		Atom (D/A)	= 0	= 1	= 2	= 3
The table shows calculated likelihoods for allowed coordination numbers for each doors and accenter observed in the current structure, computed using		1	N2 of acyclic_amide (d)	0.081	0.889	0.030	0.000
CSD derived models.		2	N1 of acyclic_nhn (a)	0.547	0.453	0.000	0.000
Highlighted table cells indicate the likelihood for the observed coordination number for the atom of that row.		3	O1 of acyclic_amide (a)	0.319	0.660	0.018	0.002
Green highlighting indicates a maximum likelihood is observed. Red highlighting indicates there is a more likely alternative coordination number for that atom		4	O2 of nitro (a)	0.888	0.105	0.006	0.000
		5	O3 of nitro (a)	0.888	0.105	0.007	0.000
SUMMARY		6	S1 of cyclic_thioether (a)	0.974	0.026	0.000	0.000
H Pand applying		7	S2 of cyclic_thioether (a)	0.974	0.026	0.000	0.000
n-bonu analysis							
Contacts must be closer than 0 + sum of vdw radii Angstroms D-H A angle must be greater than 120.000000 degrees type: INTRA and INTER, min bond path: 4, max bond path: 999, line-of- sight filter: 0							
donors: NH1(atom 0 of acyclic_amide)	~	<					>
otal # models: 596							
Mean highlighted co-ordination likelihood: 0.769264							Save



H-Bond Coordination Quick-View

• Feature available in Mercury under CSD-Materials

	AABHTZ (P-1)	- Mercury											V
	File Edit Sele	ection Display Calculat	e CSD-Community CSD-Syst	em CS	D-Materials	CSD-Discovery	CSD	Python API Hel	p				2
	Picking Mode: Pick	Atoms	▼ Clear Measurements	1	Search		•	🕆 with	Atom Label 💎				
	Style: Ball and Stick 🔻 Colour: by Element 💌 Manage Styles P		P	P Calculations •			Select by SMARTS: [c]						
	Animate	Default view: b 🔻	a b c a* b* c* x-	x+	Polymorp	h Assessment	•	Hydrogen B	Bond Propensities	zoom+			
Mercury					Co-Crysta	l Design	۰	H-bond Co	ordination Quick-view				
					Full Intera	ction Maps			145				
					Hydrate A	nalyser			H-bond Coordination Quick	-view			×
					Aromatics Analysei				Update results				
			Conforme	r Generation			INTRO	^	Atom (D/A)	= 0 = 1	= 2 = 3		
					Crystal Str	ucture Prediction			The table shows calculated likelih coordination numbers for each do	oods for allowed nor and acceptor observed	1 N2 of acyclic_a	0.081 0.889 (0.030 0.000
					Launch D	ASH			in the current structure, compute models.	d using CSD derived	2 N1 of acyclic_nh	0.547 0.453	0.000 0.000
					ADDoPT		•		Highlighted table cells indicate the	e likelihood for the the atom of that row.	3 O1 of acyclic_a	0.319 0.660 v	0.018 0.002
				_				-	Green highlighting indicates a ma observed.	ximum likelihood is	4 O2 of nitro (a)	0.888 0.105 /	0.006 0.000
									Red highlighting indicates there is coordination number for that ator	a more likely alternative n.	5 O3 of nitro (a)	0.888 0.105 /	0.007 0.000
											6 S1 of cyclic_thio	0.974 0.026	0.000 0.000
									SUMMARY		7 S2 of cyclic_thio	0.974 0.026	0.000 0.000
									H-Bond analysis				
									Total # models: 596				
									Mean highlighted co-ordination lik	elihood: 0.769264			Save

CCDC

21

H-Bond Coordination Quick-View Polymorphs applicability



N'-(1,3-dithiolan-2-ylidene)-4-nitrobenzohydrazide (refcode **DEDMUX**). Form III (purple) has a different geometry

CCDC

22

Quick assessment of H-Bond network

- Hydrogen bond networks varies in polymorphic materials
- Often several options are available due to multiple donors and acceptors present
- H-Bond Coordination Quick-View will quickly assess the most optimal network based on CSD driven information



CCDC

Examples of H-bond network in polymorphic material

Form I and II



24 Form III

CCDC

Coordination assessment in polymorphic material

Atom (D/A)	= 0	= 1	= 2	= 3
N2 of acyclic_amide (d)	0.105	0.866	0.029	0.000
N1 of acyclic_nhn (a)	0.800	0.200	0.000	0.000
O1 of acyclic_amide (a)	0.329	0.652	0.017	0.002
O2 of nitro (a)	0.888	0.105	0.006	0.000
O3 of nitro (a)	0.888	0.105	0.006	0.000
S1 of cyclic_thioether (a)	0.974	0.026	0.000	0.000
S2 of cyclic_thioether (a)	0.974	0.026	0.000	0.000

Form I



Co-ordination scores

(To refresh table: left-click chart point)

	Atom (D/A)	= 0	= 1	= 2	= 3
1	N2 of acyclic_amide (d)	0.081	0.889	0.030	0.000
2	N1 of acyclic_nhn (a)	0.547	0.453	0.000	0.000
3	O1 of acyclic_amide (a)	0.319	0.660	0.018	0.002
4	O2 of nitro (a)	0.888	0.105	0.006	0.000
5	O3 of nitro (a)	0.888	0.105	0.007	0.000
6	S1 of cyclic_thioether (a)	0.974	0.026	0.000	0.000
7	S2 of cyclic_thioether (a)	0.974	0.026	0.000	0.000



Hydrogen Bond Propensity colored chart













2020 customer events

- US UGM East, Boston (April 24)
- CFC Meeting, Cambridge (April 29-30)
- Europe UGM, Cambridge (June 2-3)
- US UGM West, San Diego (August 13)





- ACS CSD 1M Symposium report <u>https://info.ccdc.cam.ac.uk/acs-report</u>
- Watch out for "Matwalls" white paper







Next What's Up Webinar

- Next webinar: March 19th
- Send us your ideas and news <u>hello@ccdc.cam.ac.uk</u>





Thank you

hello@ccdc.cam.ac.uk



The Cambridge Crystallographic Data Centre 12 Union Road, Cambridge CB2 1EZ, United Kingdom Registered Charity No. 800579

