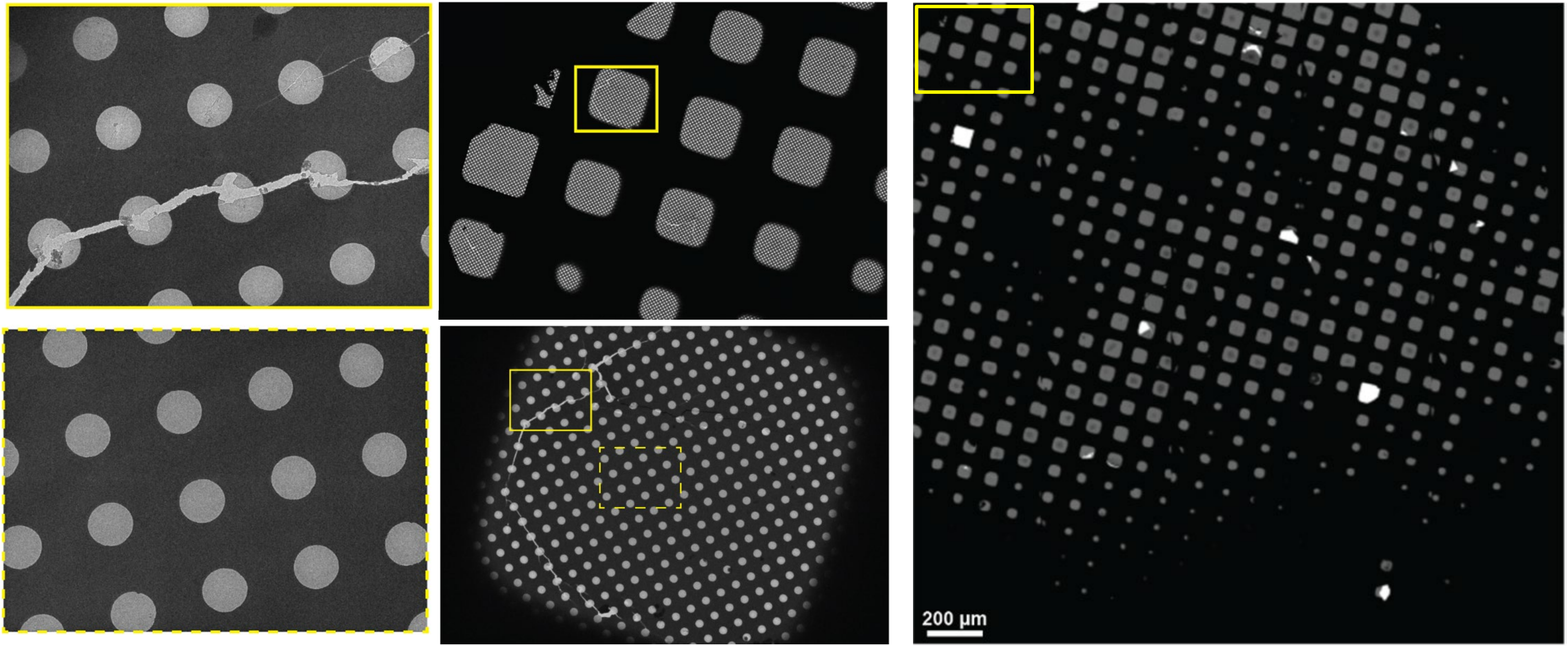
Automating cryo-EM Data Collection with Reinforcement Learning

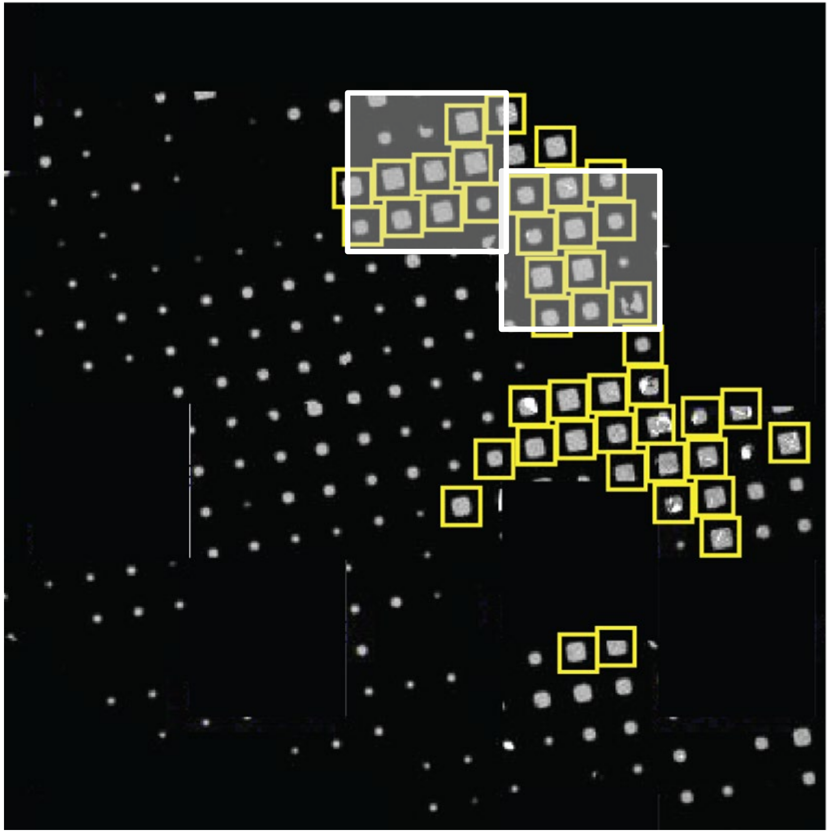
Yilai Li, Ph.D.

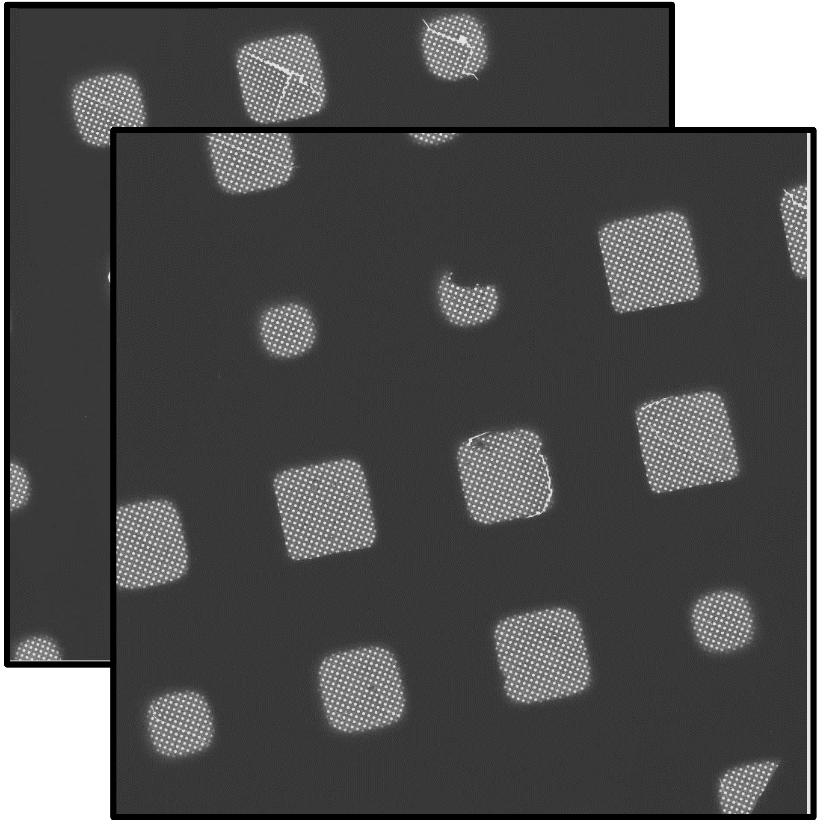
Cianfrocco Lab
Life Sciences Institute
University of Michigan

Cryo-EM grids are complicated data landscapes

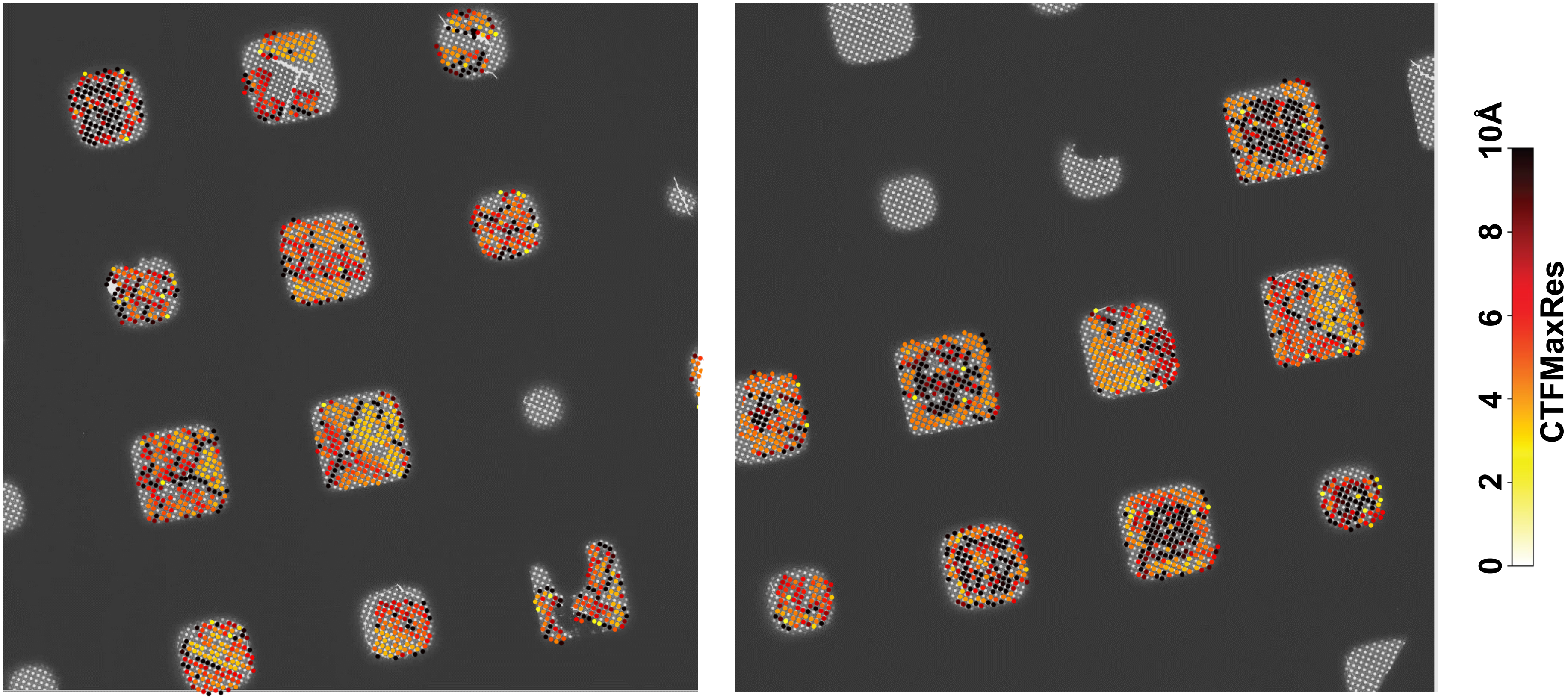


Cryo-EM grids are complicated data landscapes

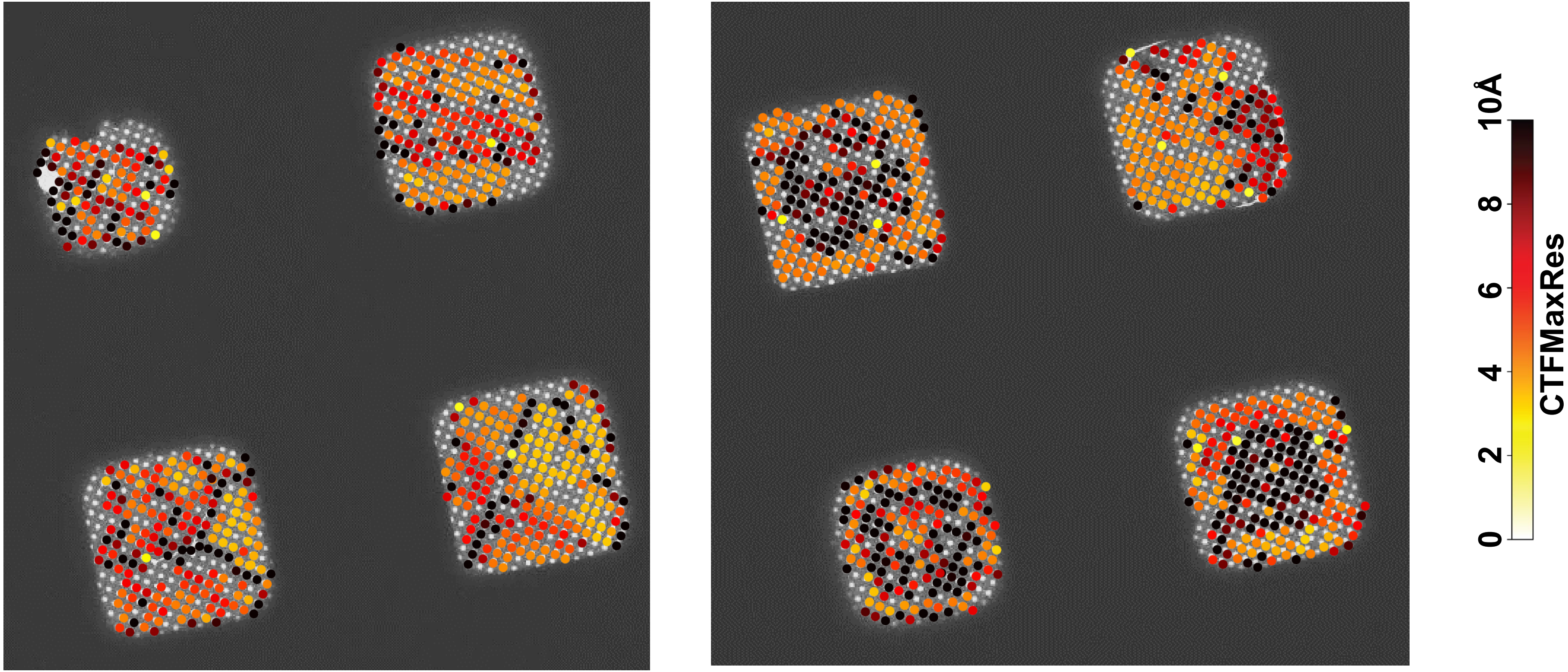




Cryo-EM grids are complicated data landscapes

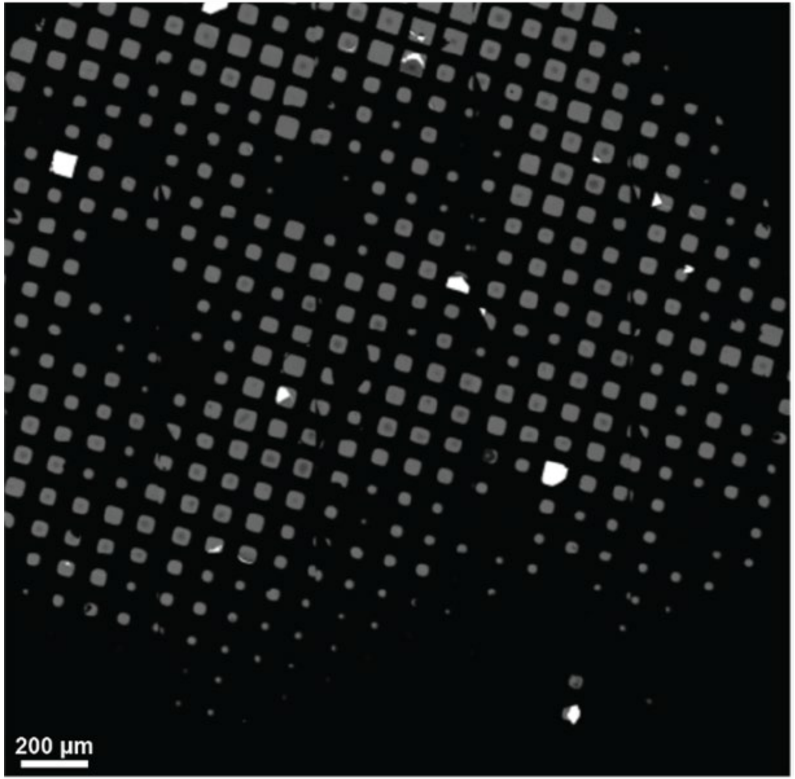


Cryo-EM grids are complicated data landscapes



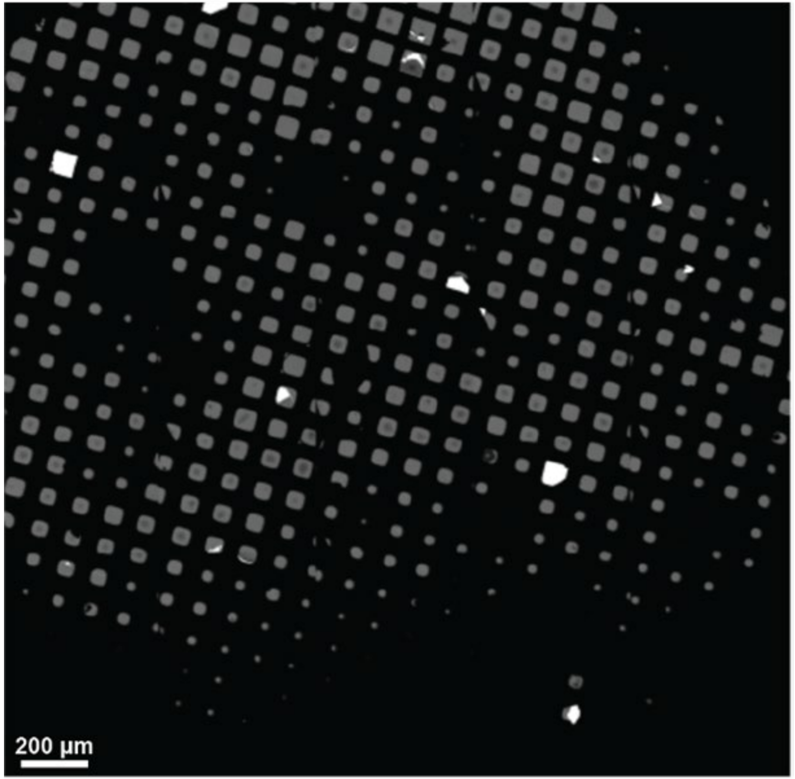
Cryo-EM grids are complicated data landscapes

- ➔ Square and hole variability
- ➔ Need to decide how to navigate through the grid and to find the ‘best’ way through
- ➔ Image <5% of grid



How is our data quality now?

- ➔ From all micrographs collected from Jan 2019 to May 2021 in our lab, about 50% does not contain high resolution information at all
- ➔ The efficiency of data collection greatly depends on the expertise
- ➔ **How do we collect as many "good" micrographs in a limited time frame?**

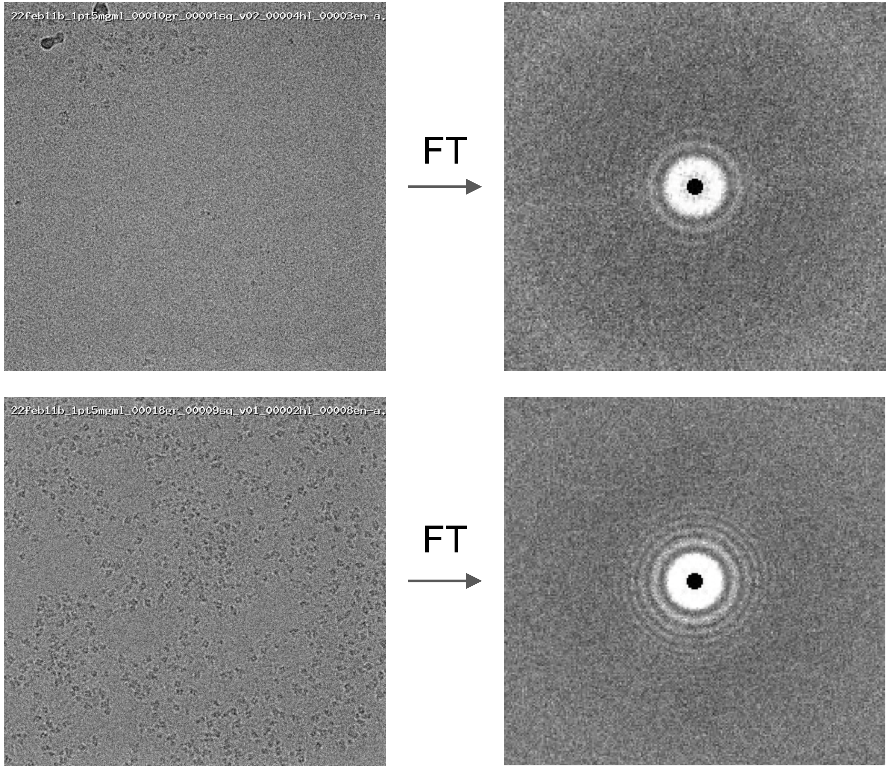


Path planning across a cryo-EM grid is challenging

How do we collect as many “good” micrographs in a limited time frame?

- ➔ What is a “good” micrograph?
- ➔ How do we assess data quality at low and medium magnified images?
- ➔ How do we balance trade offs in the time for switching patches, squares, and regions of atlas with data quality

How do we defined good/bad exposures? (**CTF Max Resolution**)

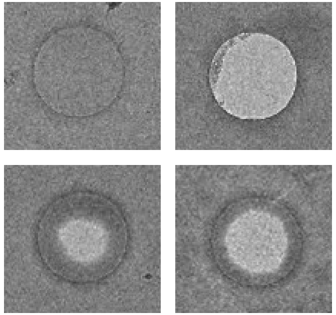


Pros:
-unbiased
-generally correlated with data quality
-quick to calculate

Cons:
-unrelated to particle quality
-ice thickness dependent

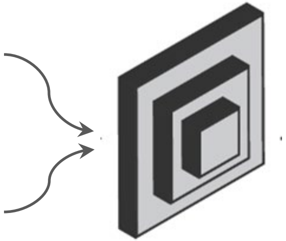
Specifically for our approach: CTFFIND4

How do we assess hole quality? (“**Deep regressor**”)



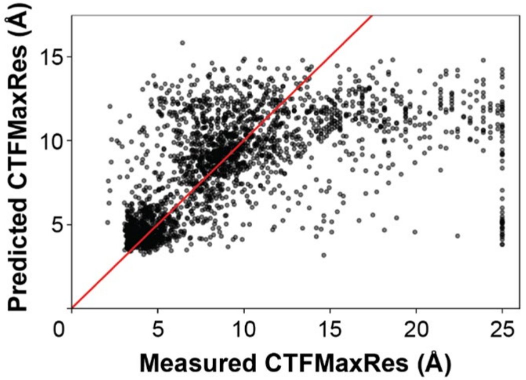
+

CTFMaxRes
per hole

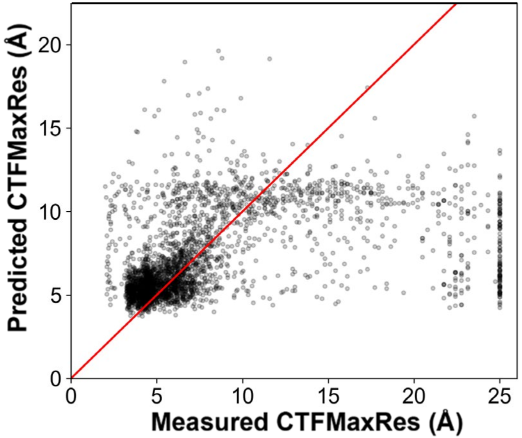


ResNet50

Deep regressor



Trained & tested on same grid

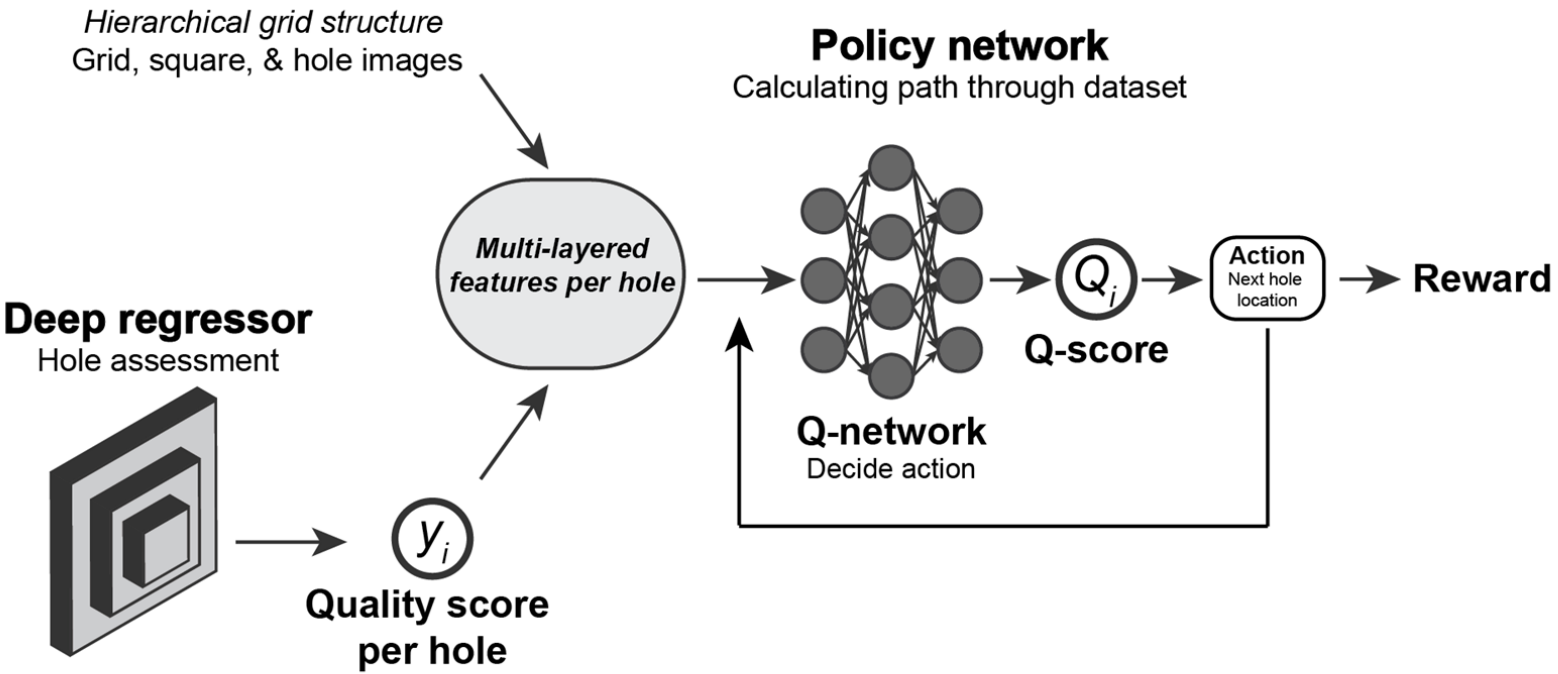


<u>General</u> deep regressor tested on single grid

Grid-specific regressor: 1,325 holes

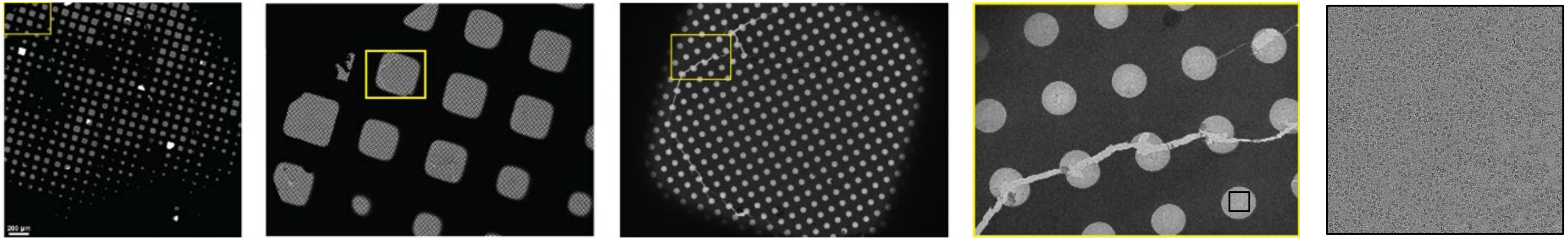
General regressor: 100,578 holes

How do we plan a path across a grid? (**Deep Q-network**)



How do we plan a path across a grid? (**Deep Q-network**)

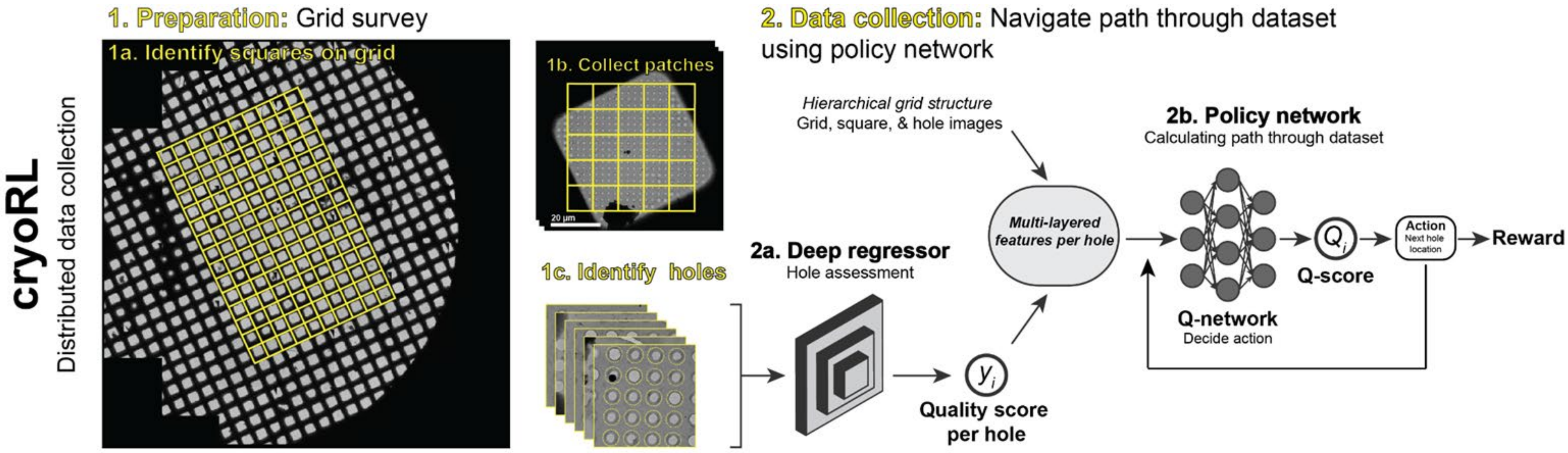
Designing rewards for training



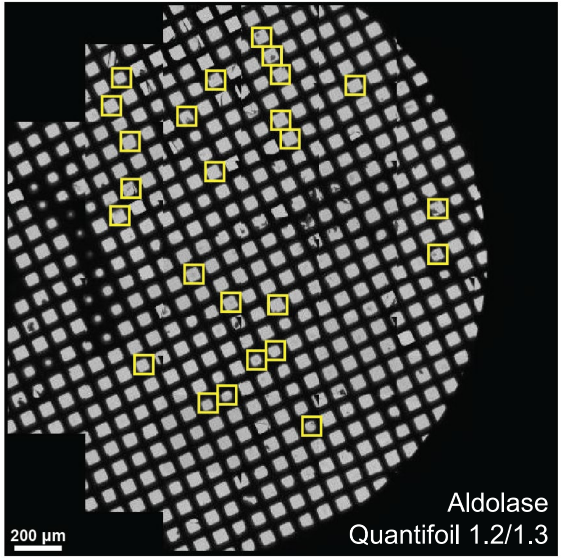
Atlas **Grid** **Square** **Patch** **Micrograph**

Collect micrograph…	Time (min.)	Reward (if CTFMaxRes < T)

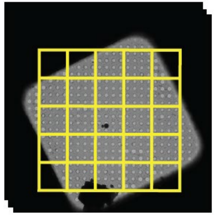
CryoRL: Reinforcement learning-guided data collection



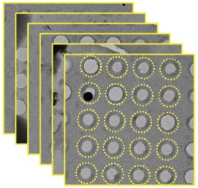
How do we evaluate cryoRL? Systematic data collection



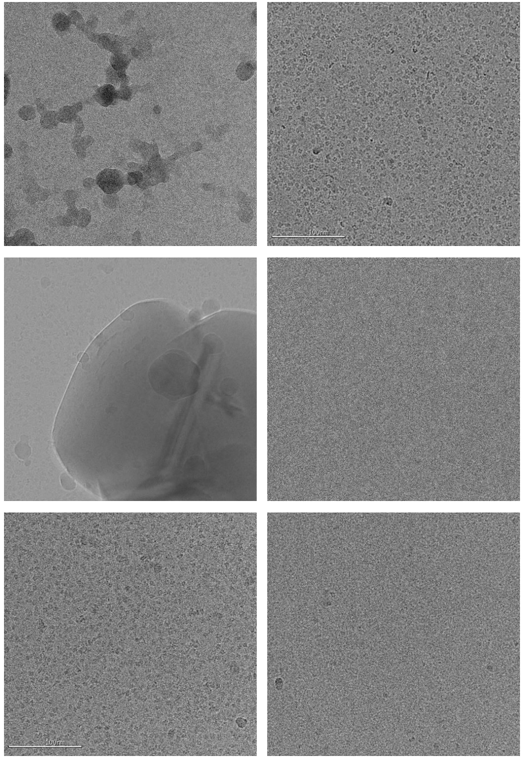
25 squares



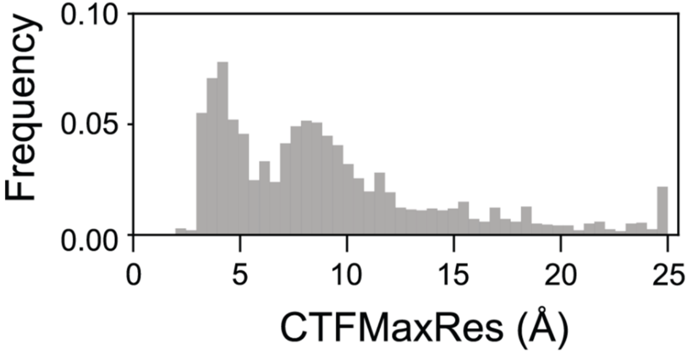
3,538 holes



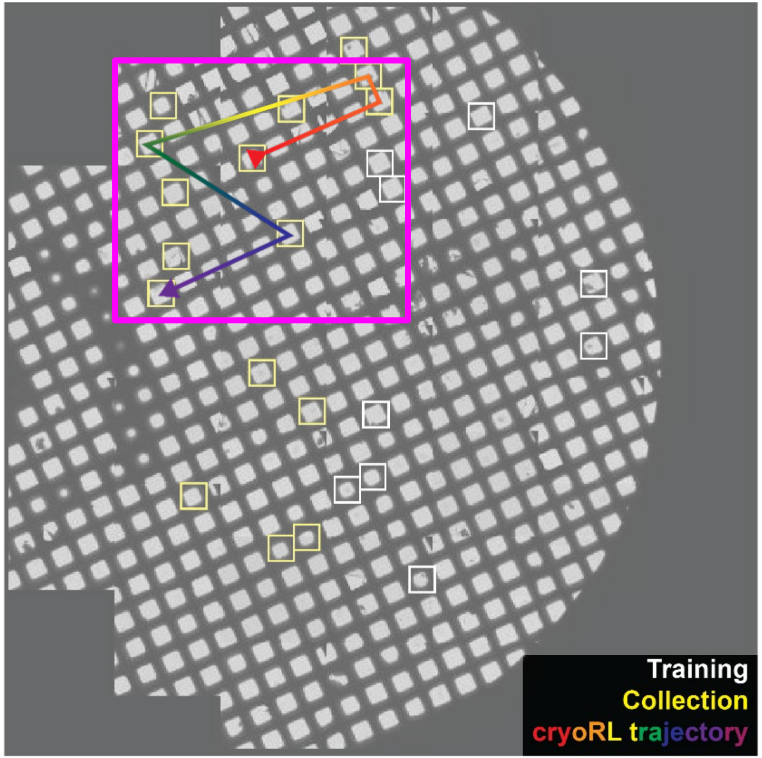
Example micrographs



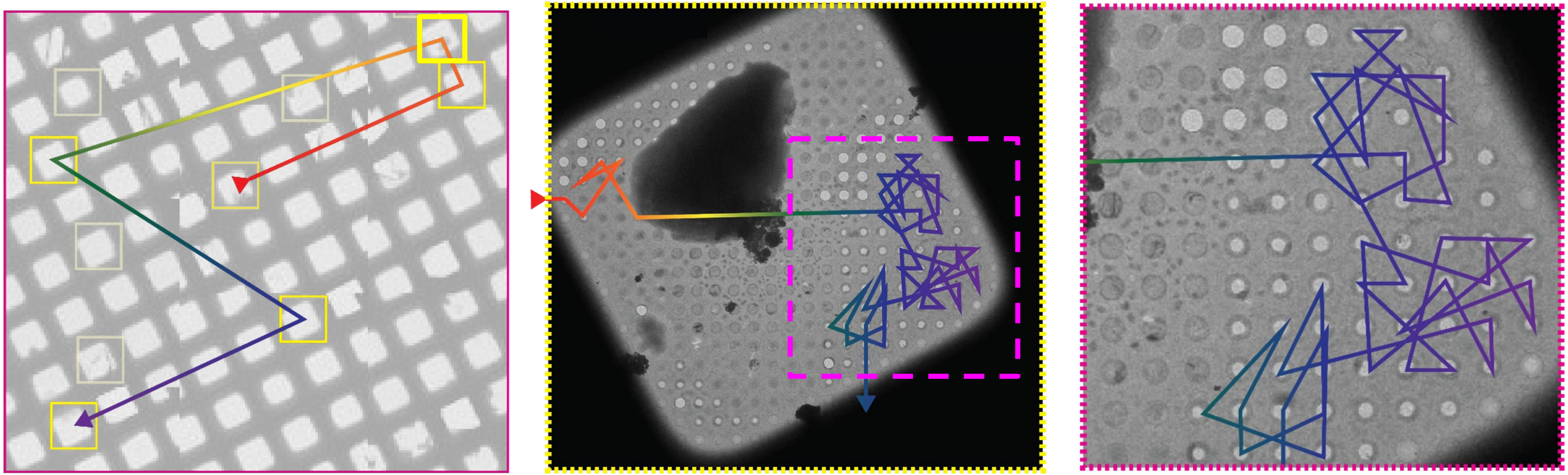
CTFMaxRes for all micrographs



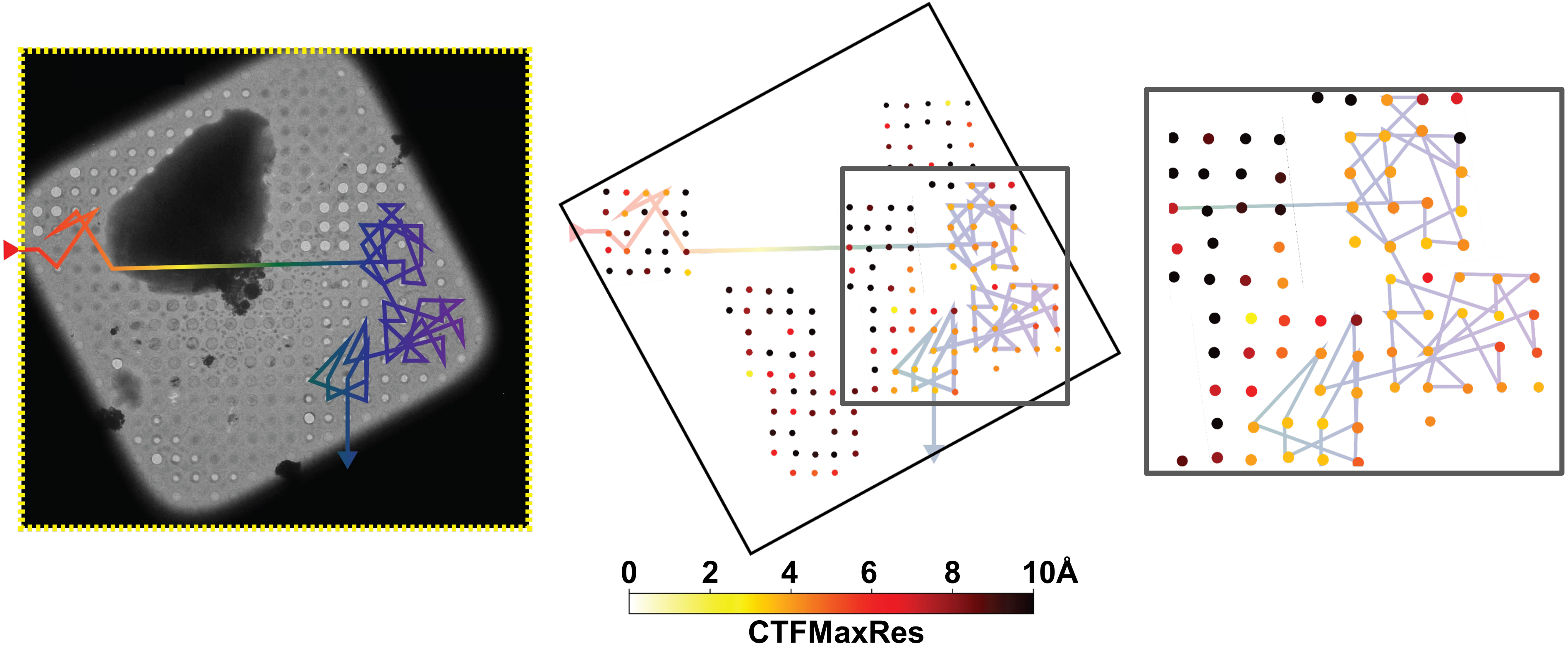
cryoRL successfully navigates aldolase cryo-EM grid



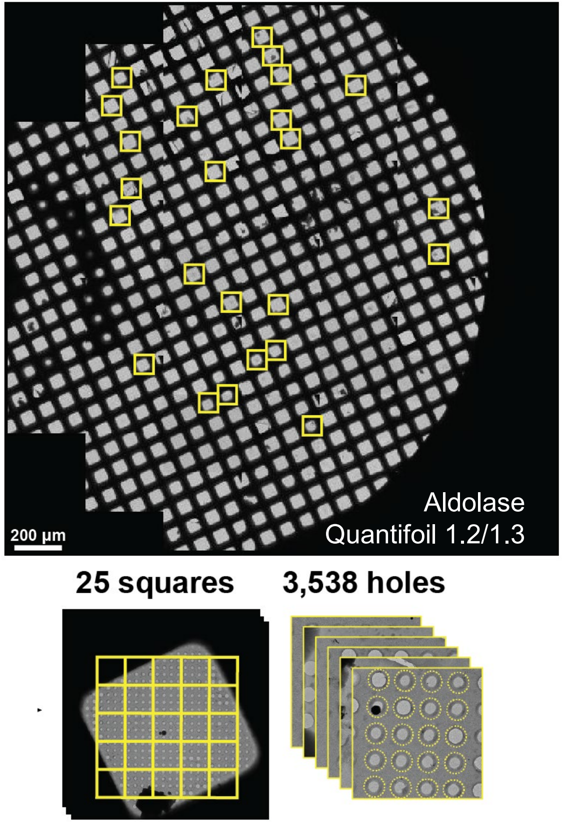
cryoRL successfully navigates aldolase cryo-EM grid



cryoRL successful navigates aldolase cryo-EM grid



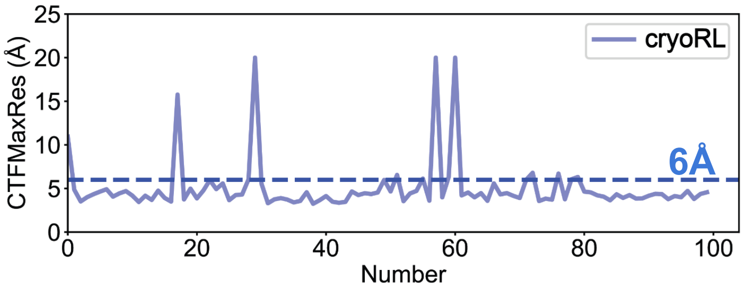
How do we evaluate the result from cryoRL? Naïve baseline

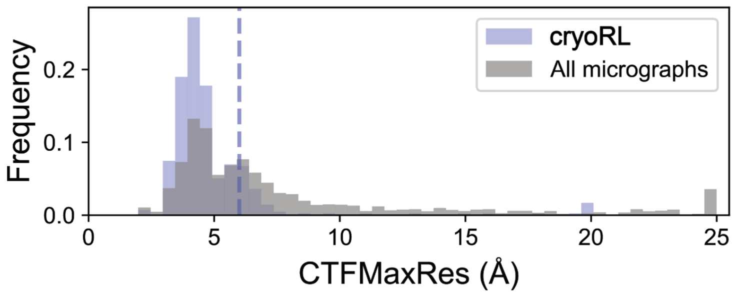


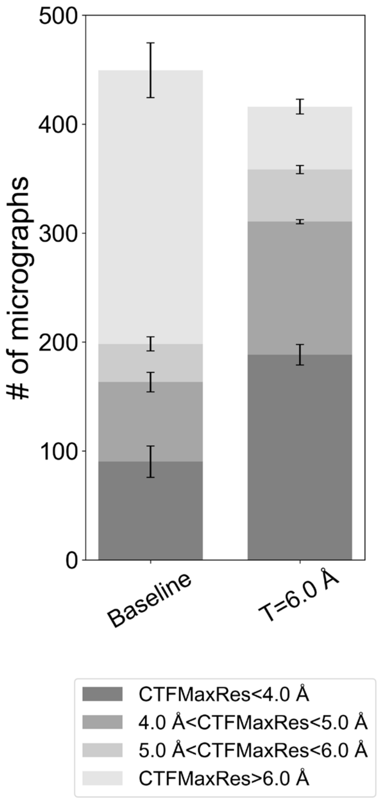
➔ Start from a random position

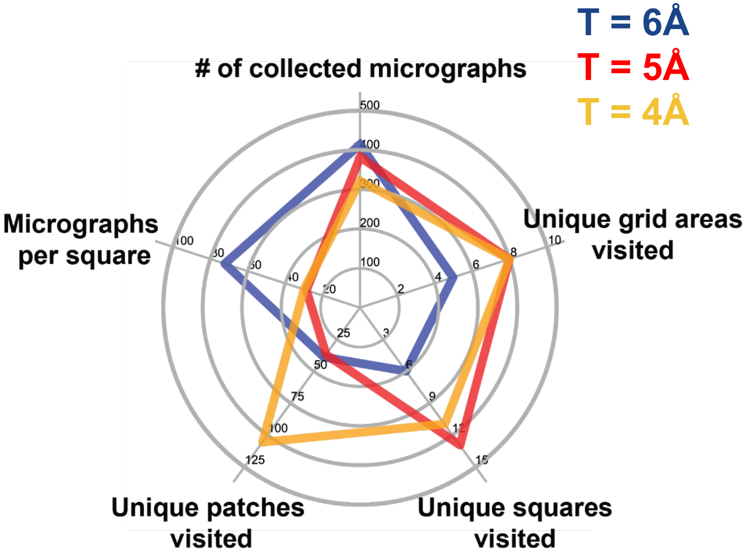
➔ Collect as many micrographs as possible in the given time limit

cryoRL successful navigates aldolase cryo-EM grid

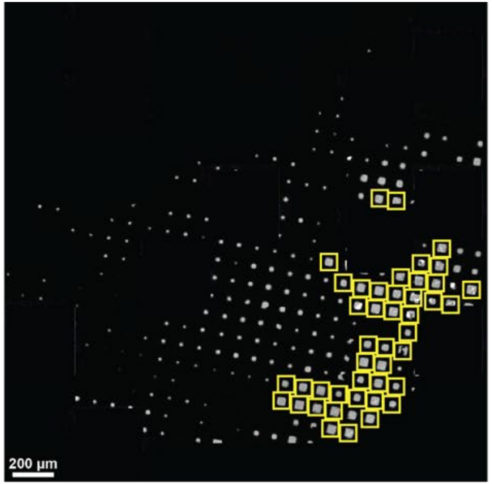




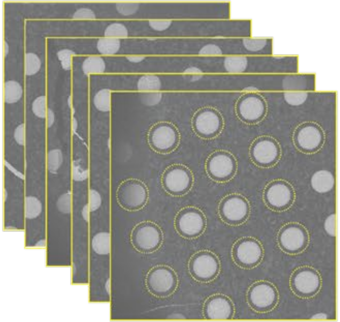




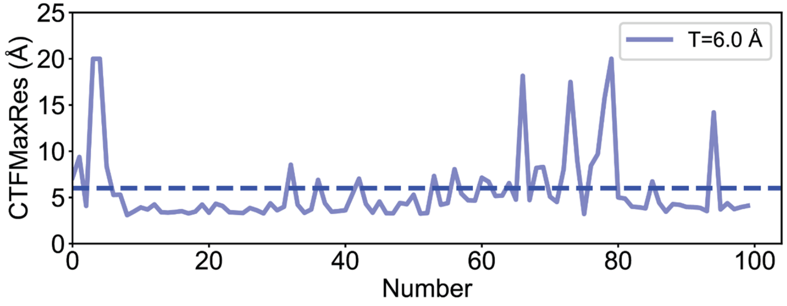
Transferability - can we can train cryoRL offline and use on a different sample?

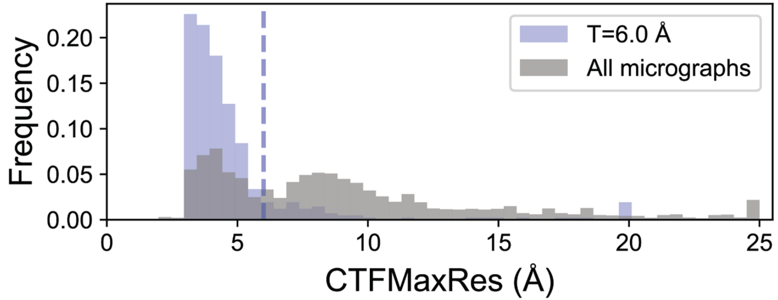


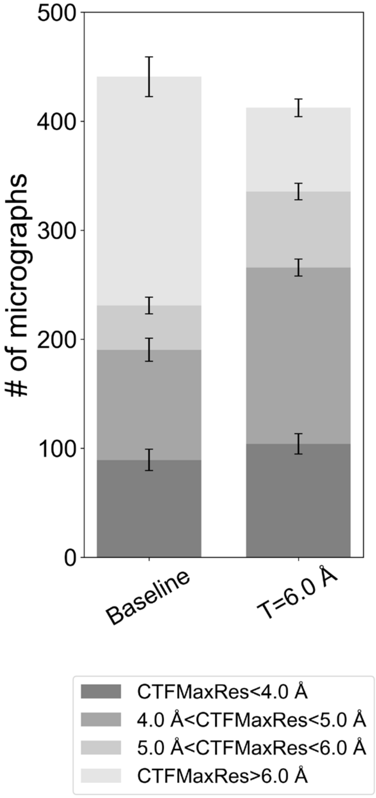
Sample: Apoferritin
Grid: UltrAuFoil 1.2/1.3
Imaging: Talos Arctica + K2

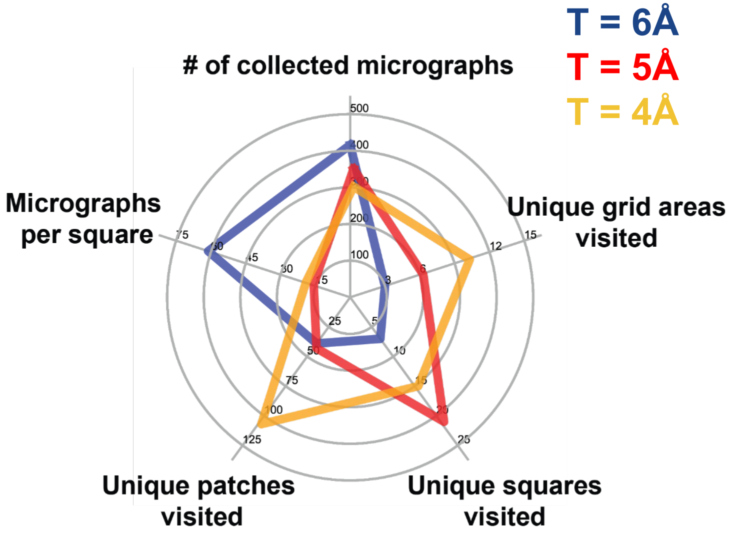


Transferred models from aldolase allows effective data collection on apoferritin

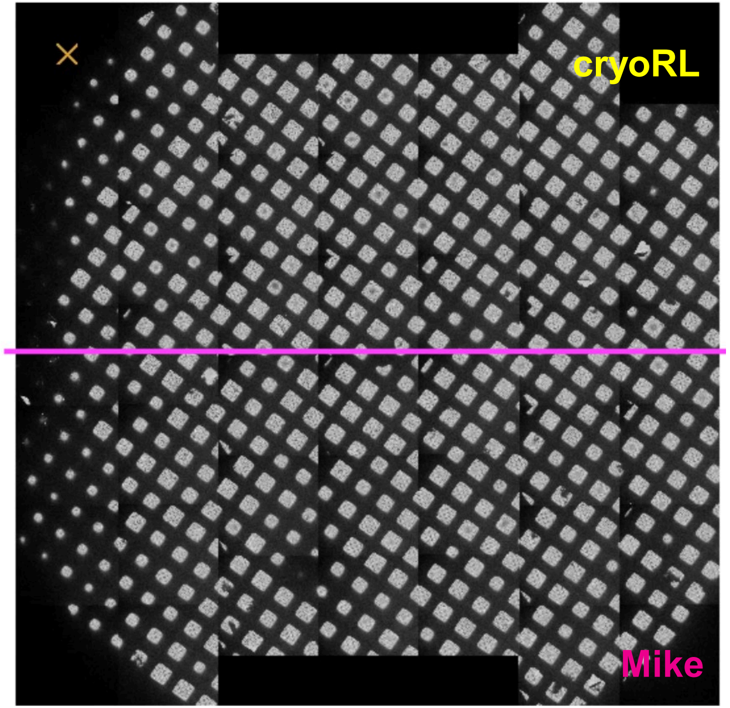


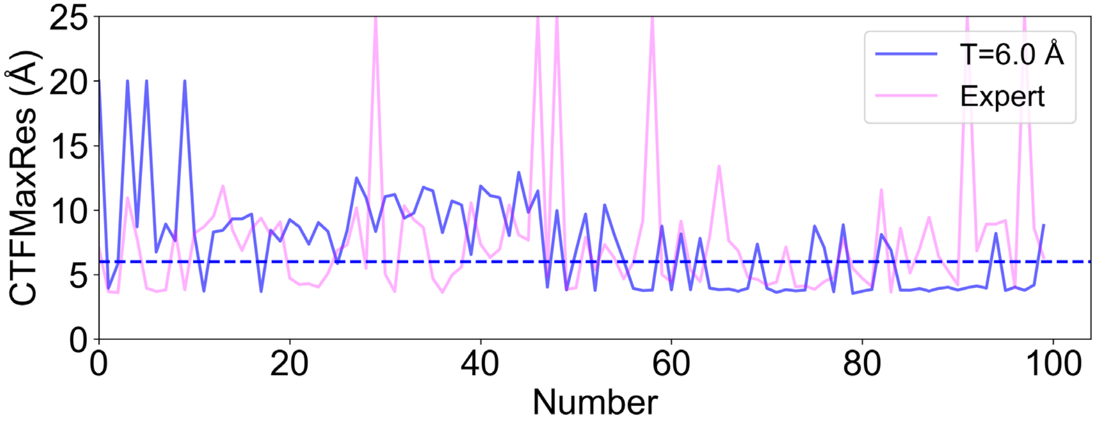




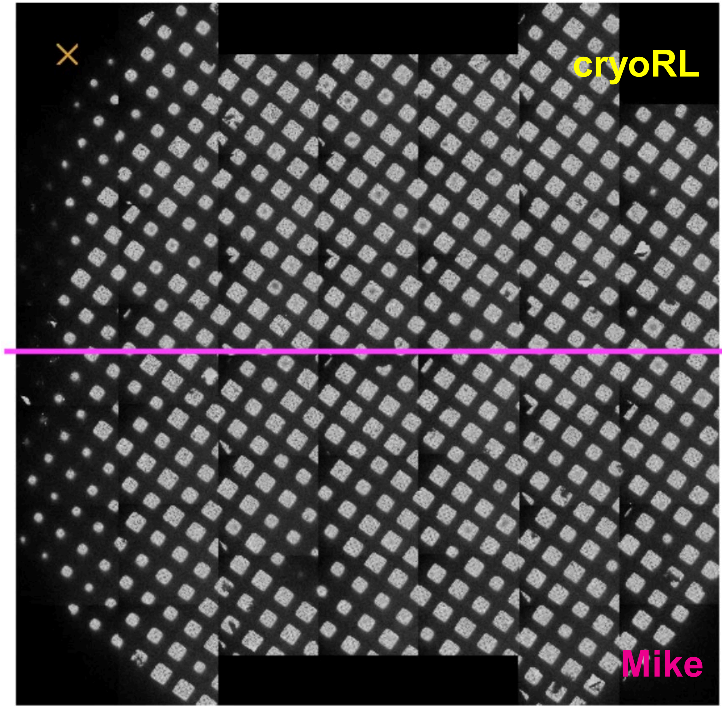


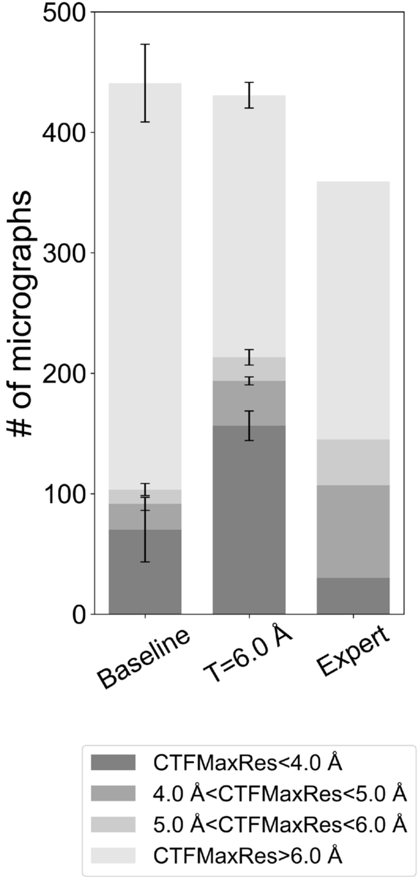
cryoRL collected data with better quality than an expert

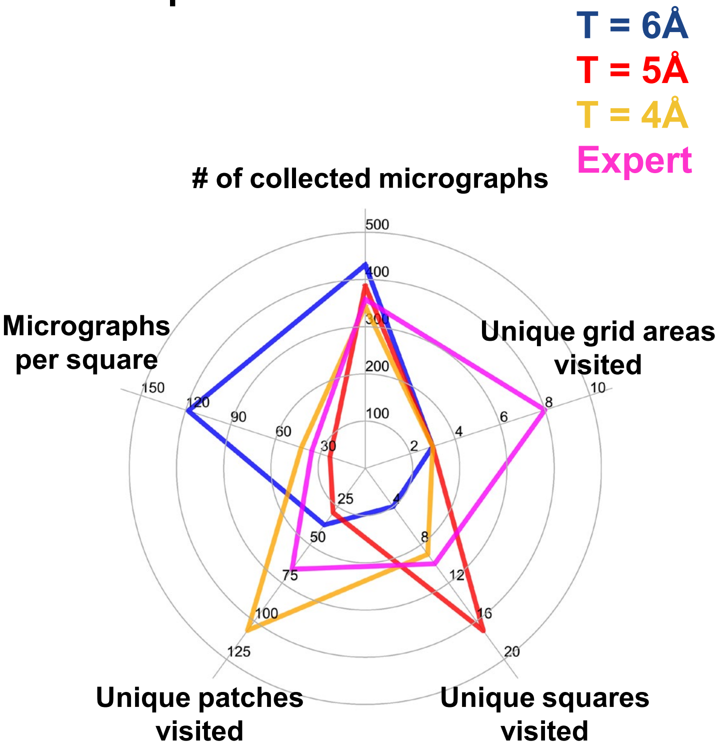




cryoRL collected data with better quality than an expert





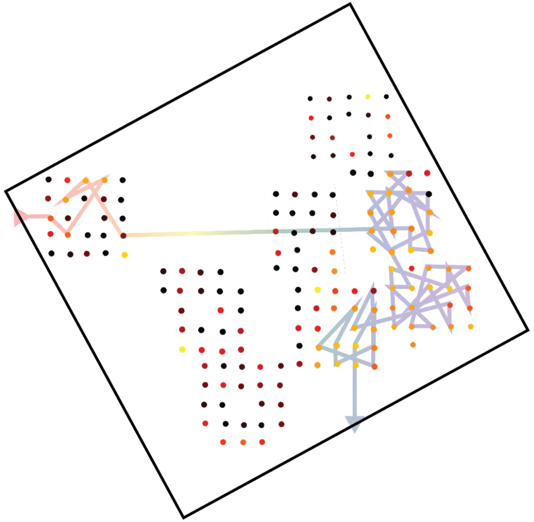


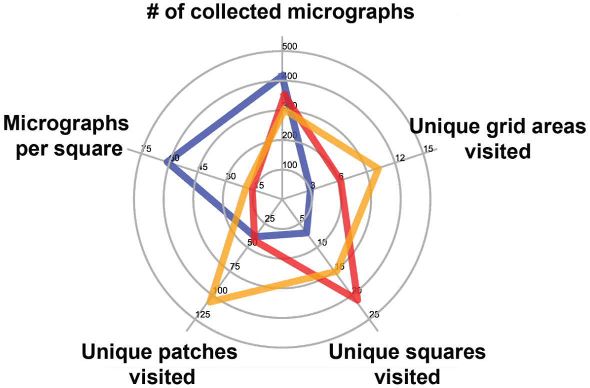
Conclusions

➔ Reinforcement learning combined with hole regressor allows successful 'data collection'

➔ cryoRL learns policy for collecting images that maximizes data quality given limited time

➔ Parameter setup allows for relaxed vs. stringent data collection





Future directions - next steps with cryoRL

- ➔ More cryoRL vs expert comparison
- ➔ Update the regressor during data collection
- ➔ What is a ‘good’ micrograph?
- ➔ How do we know when to stop data collection?
- ➔ Incorporate into software (SerialEM, Leginon/Magellon, EPU)

Acknowledgements

cryoRL

Seychelle Vos, Ph.D. (MIT)
Michael Cianfrocco, Ph.D. (U-M)
Quanfu Fan, Ph.D. (IBM)*
John Cohn, Ph.D. (IBM)

Cianfrocco lab

Somaye Badieyan, Ph.D.
Nick Vangos
Emily Eberhardt
Lily Hahn
Zhenyu Tan

Data collection simulator

Ja Young Lee (IBM)
Veronique Demers (IBM)
Lucy Yip (IBM)

U-M Cryo-EM Facility

Ashleigh Razcowoski
Min Su, Ph.D.

Funding

NIH R01 GM143805 (Magellon - Scott Stagg & Gabe Lander)
NSF DBI ABI 1759826