Use of advanced computer science technologies for quasi-online data processing and primary analysis in the pipe-line approach – on example of experiments on EU-XFEL and CryoEM in structural biology

by V. Ilyin

Joint RSF-Helmholtz grant 18-41-06001 2018-2020

New avenues in information and data science: advanced imaging applications at the XFEL and cryo-EM frontier

NRC Kurchatov Institute (Moscow) *and* *DESY (Hamburg)*

Vyacheslav Ilyin team leader,
Alexander Vasiliev,
Anton Teslyuk, Eugene Pichkur
Sergey Bobkov, Yury Chesnokov
Ksenia Ikonnikova, Timur
Sergey Zolotarev, Baymukhametov
all NRC Kurchatov Institute

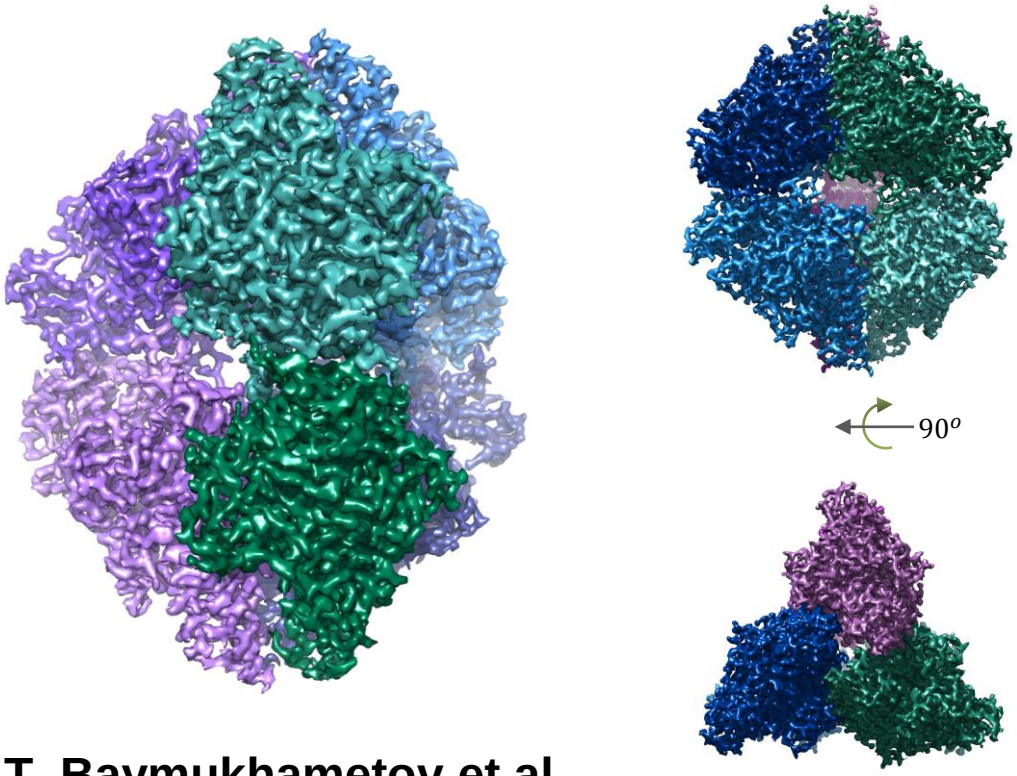
Ivan Vartanyants, team leader
DESY
Wilfried Wurth,
Max Rose, Luca Gelisio
Young Yong Kim, Ruslan Khubbutdinov
Dmitry Lapkin, Dameli Assalauova
Nastasia Mukharamova
XFEL
Adrian Mancuso, Ruslan Kurta
Giuseppe Mercurio

NEC'2019, Budva, 1st October 2019

Example of the cryo-EM analysis: use of the HPDP prototype

RSF-Helmholtz 18-41-06001

Cytochrome *C* nitrite reductase (TvNiR) from the bacterium *Thioalkalivibrio nitratireducens*



T. Baymukhametov et al.
Acta naturae. 2018, v.10, №3, p.48.

Parameters:

Molecular weight	380 kDa
Dimensions	130-150 Å
Symmetry	D3 (hexamer)

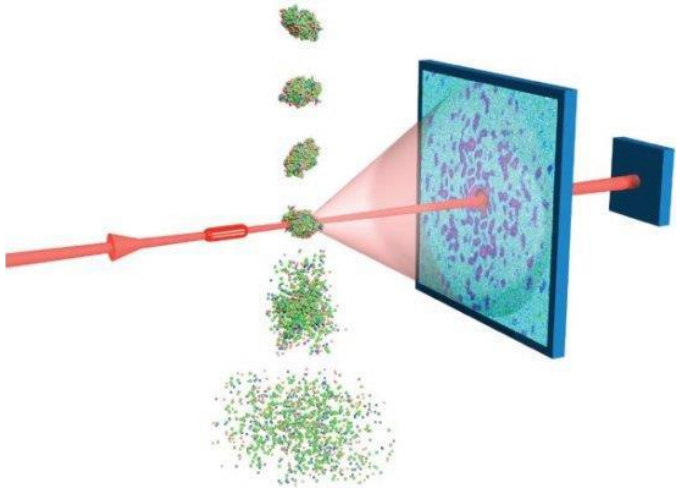
Physiological role:
involved in the catalysis of nitride reduction in bacterial cells.

Data acquisition:
Titan Krios with Falcon II DED (NRC KI, Moscow)

Data processing:
By use of HPDP platform installed on supercomputer resources at NRC KI

Resolution: 2.56 Å

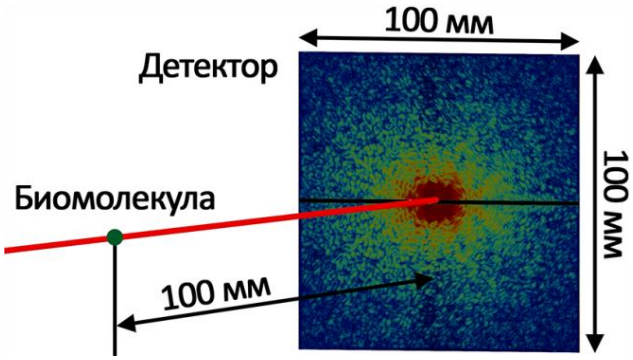
EY-XFEL single particle experiment



EU-XFEL parameters:

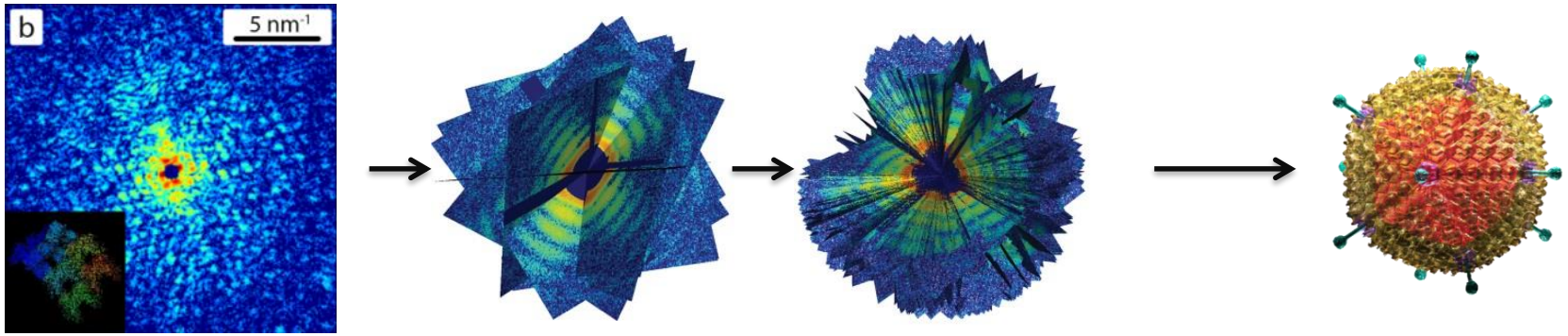
X-Ray λ	0.5 – 47 Å
Frequency	27000 Hz
Start	Dec. 2017

Up to 2 millions images per hour (2020?)



Resolution 224x224 px (3 Å)

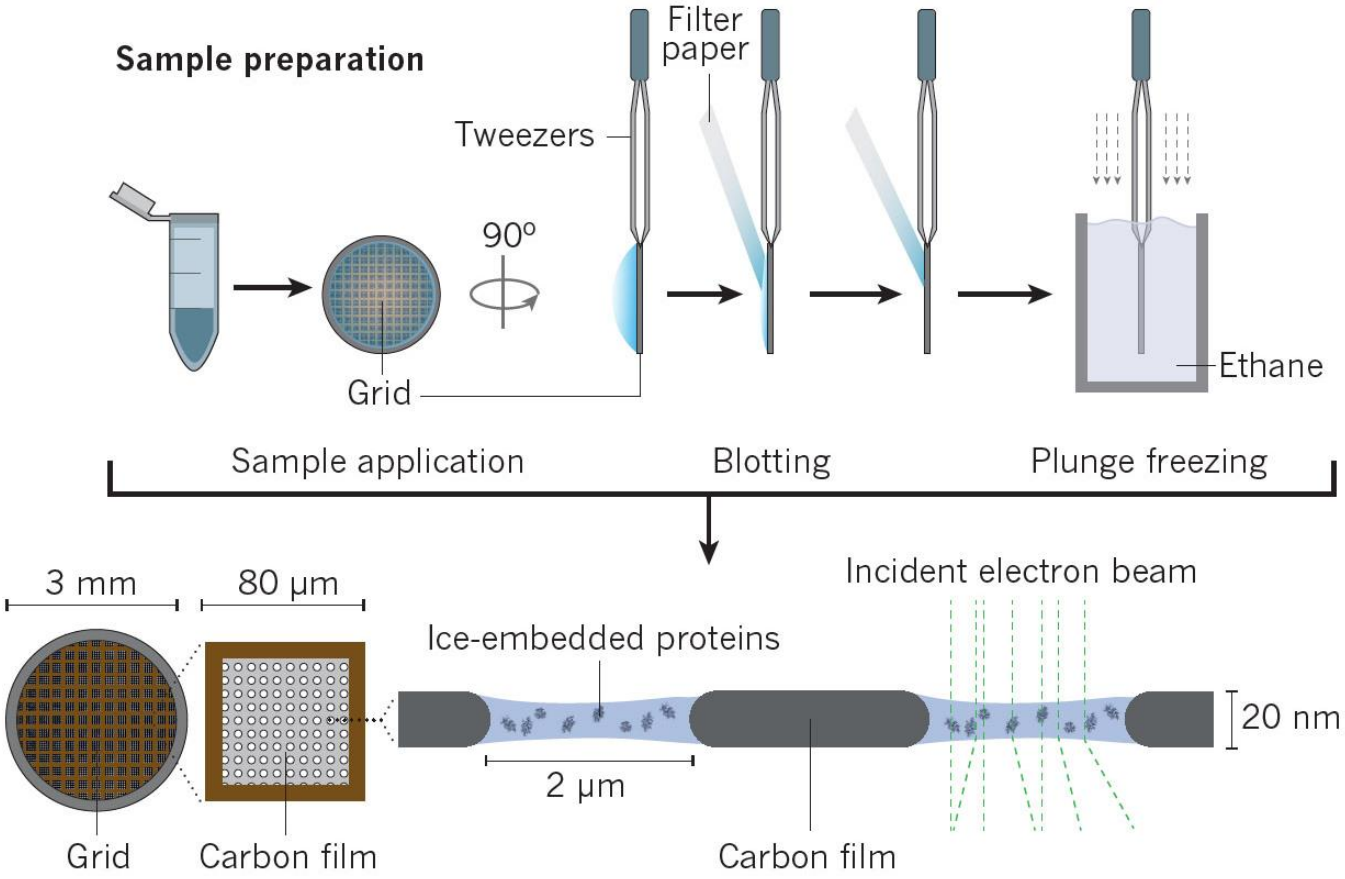
Image in inverse space *To collect images with different orientations* *To get 3D structure*



Just a bit on Cryogenic Transmission Electron Microscopy (CryoTEM)

RSF-Helmholtz 18-41-06001

Cryo-EM allows to restore the 3D structure of bio macromolecules - viruses in (almost) native state (due to instant freezing) with near atomic resolution.

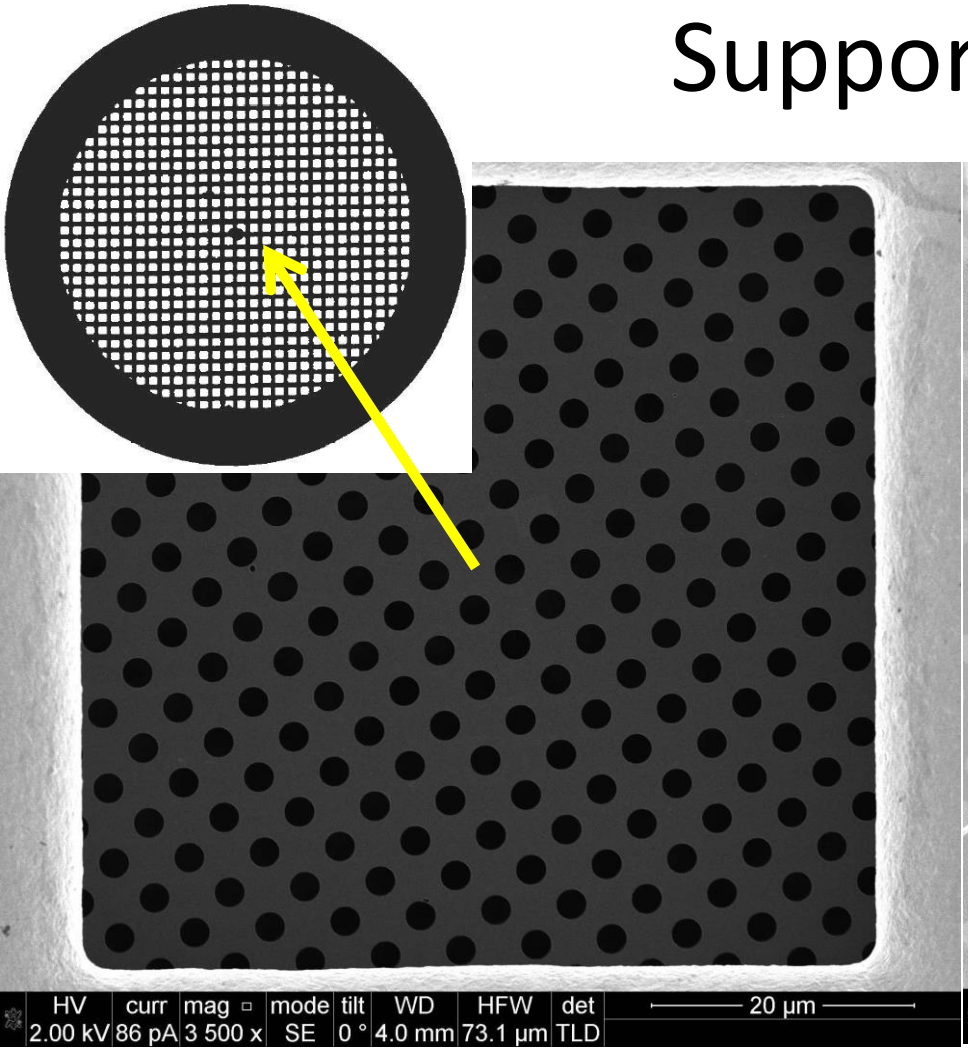


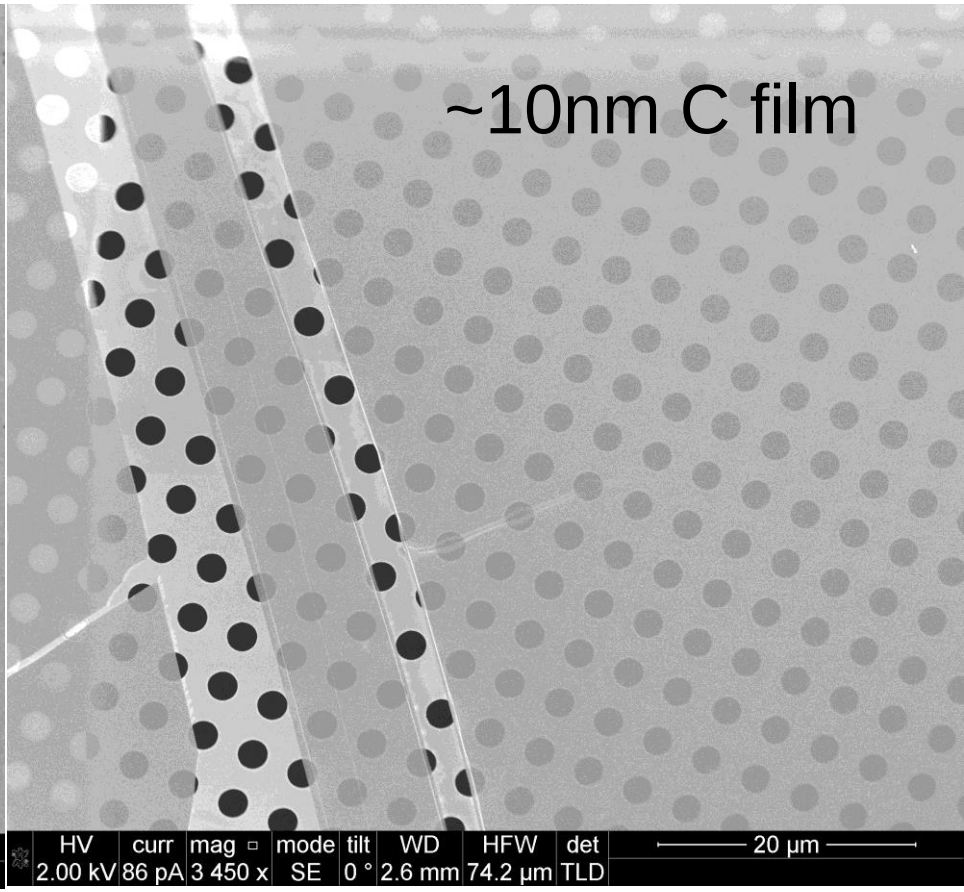
NEC'2019, Budva, 1st October 2019

Just a bit on Cryogenic Transmission Electron Microscopy (CryoTEM)

RSF-Helmholtz 18-41-06001

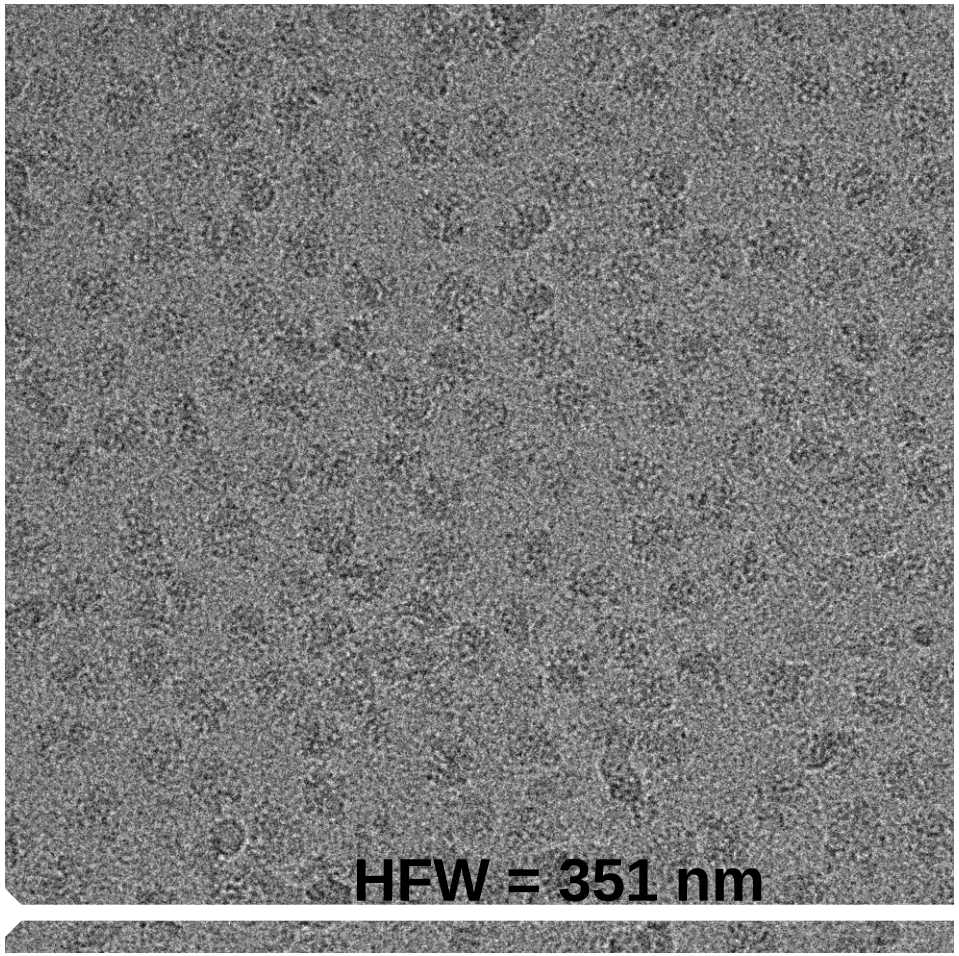
Support grids





Just a bit on Cryogenic Transmission Electron Microscopy (CryoTEM)

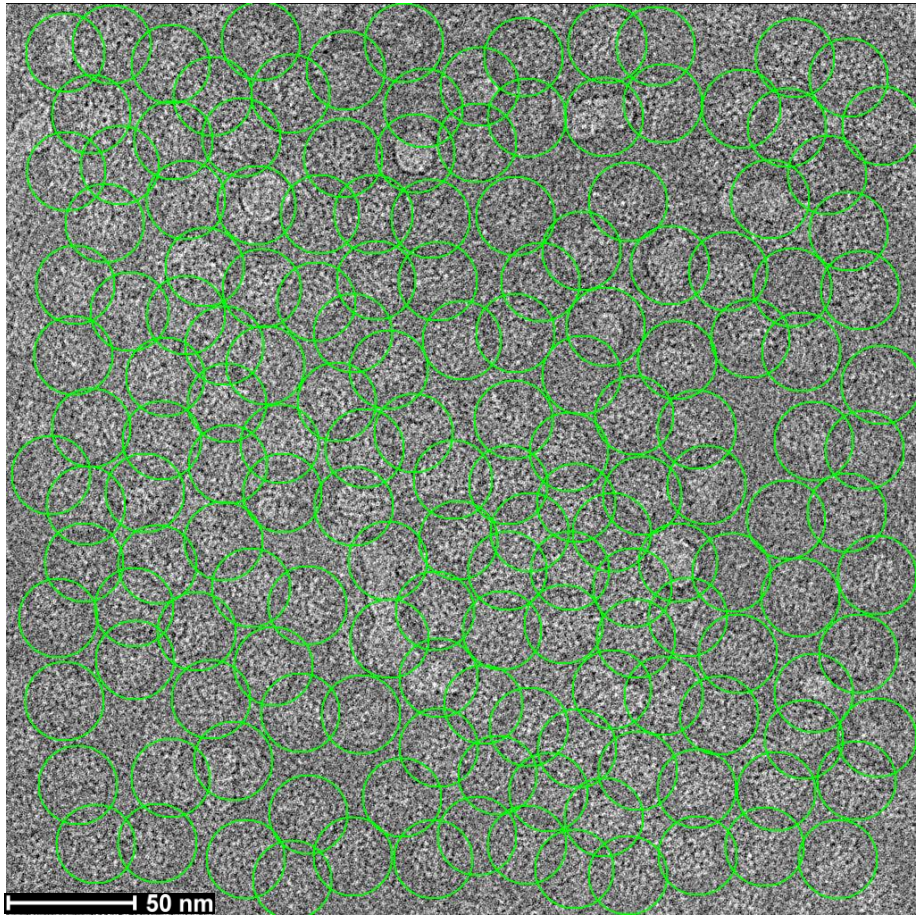
RSF-Helmholtz 18-41-06001



An example of cryo-EM image
4k x 4k, pix size: 0.859Å

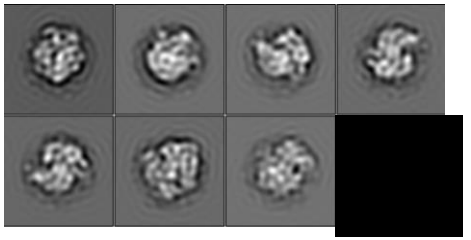
RSF-Helmholtz 18-41-06001

Preprocessing / Auto picking



Gautomatch

references

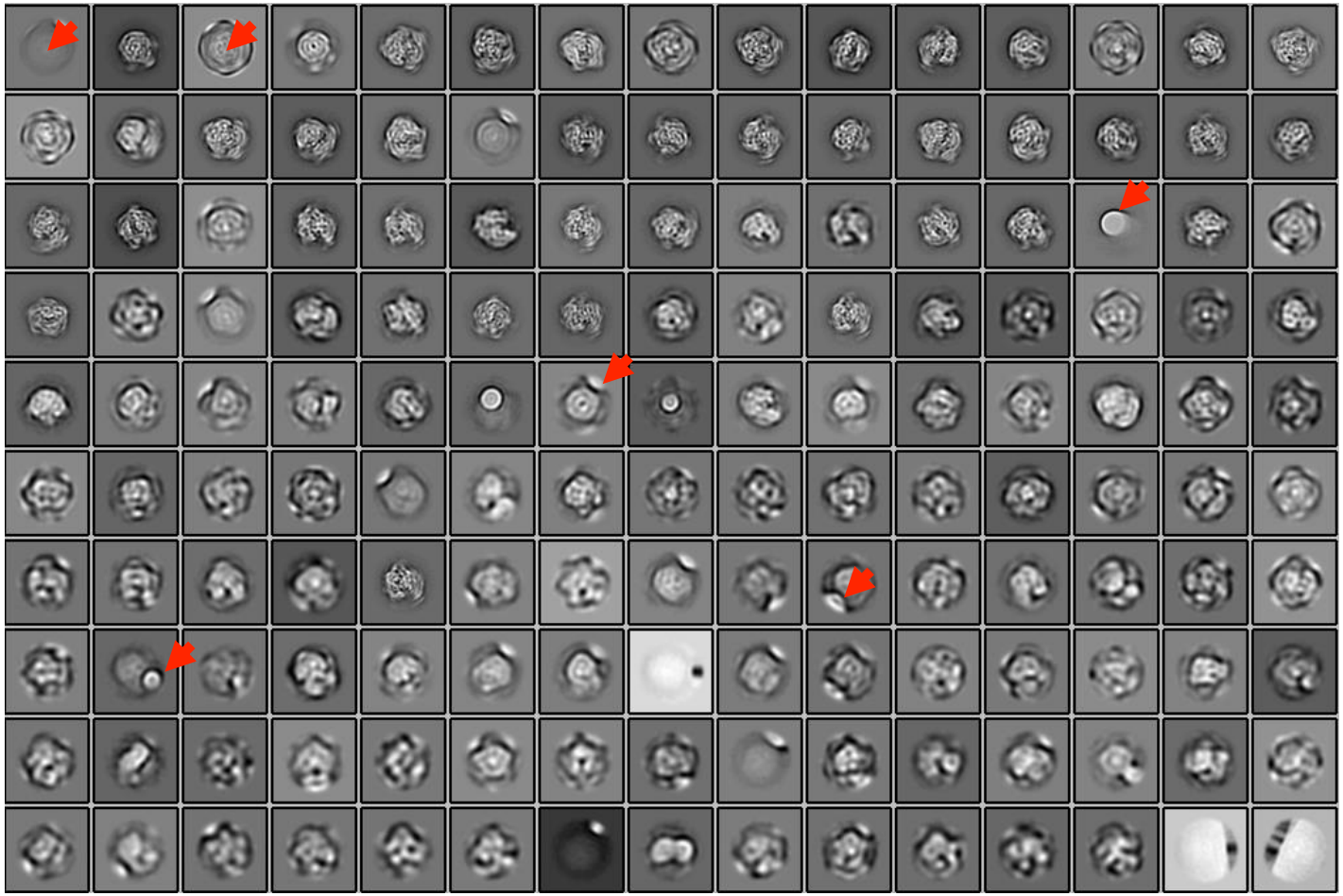


>500k particles

Just a bit on Cryogenic Transmission Electron Microscopy (CryoTEM)

RSF-Helmholtz 18-41-06001

Relion / 2D classification: run1 (C8)



150 classes
20 it

↓

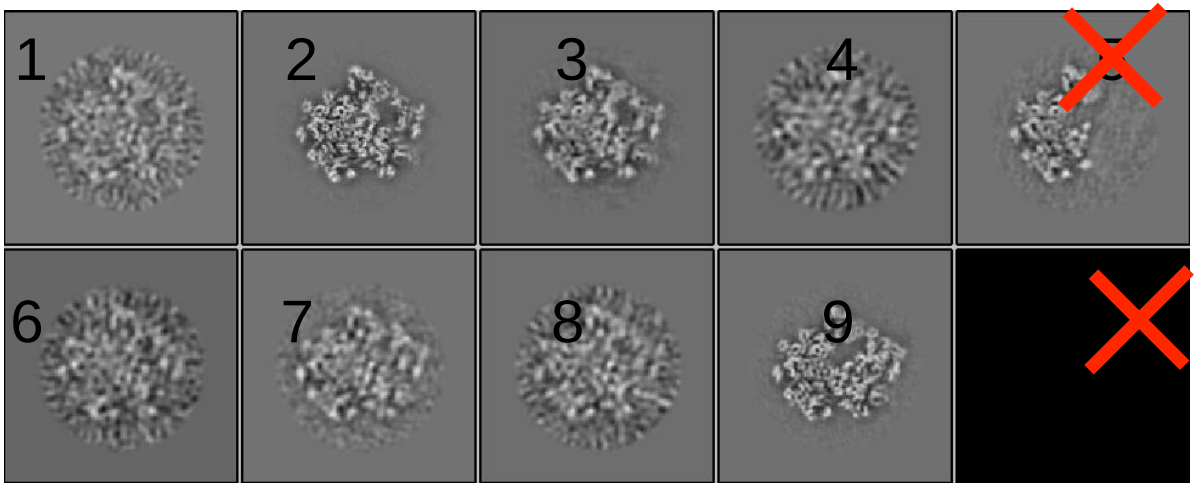
ice/empty classes

↓

443k particles

RSF-Helmholtz 18-41-06001

Relion / 3D classification (C4)



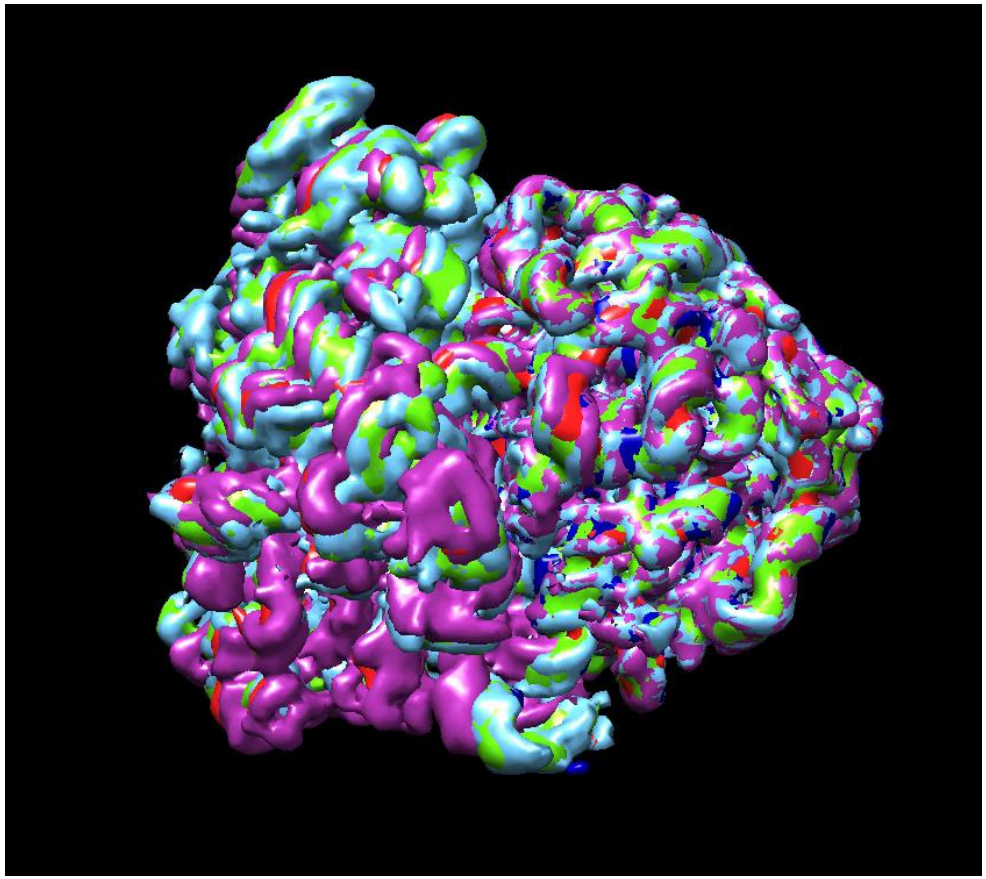
10 classes
20 it

↓

Class 1 - #1 (36k)
Class 2 - #2,3 (161k)
Class 3 - #4,6,8 (85k)
Class 4 - #7 (1.8k)
Class 5 - #9(41k)

RSF-Helmholtz 18-41-06001

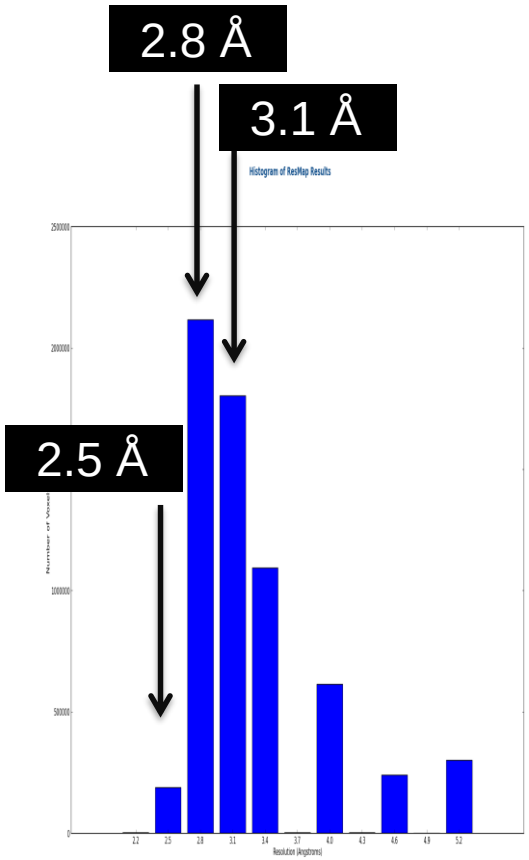
Relion / 3D classification (C4)

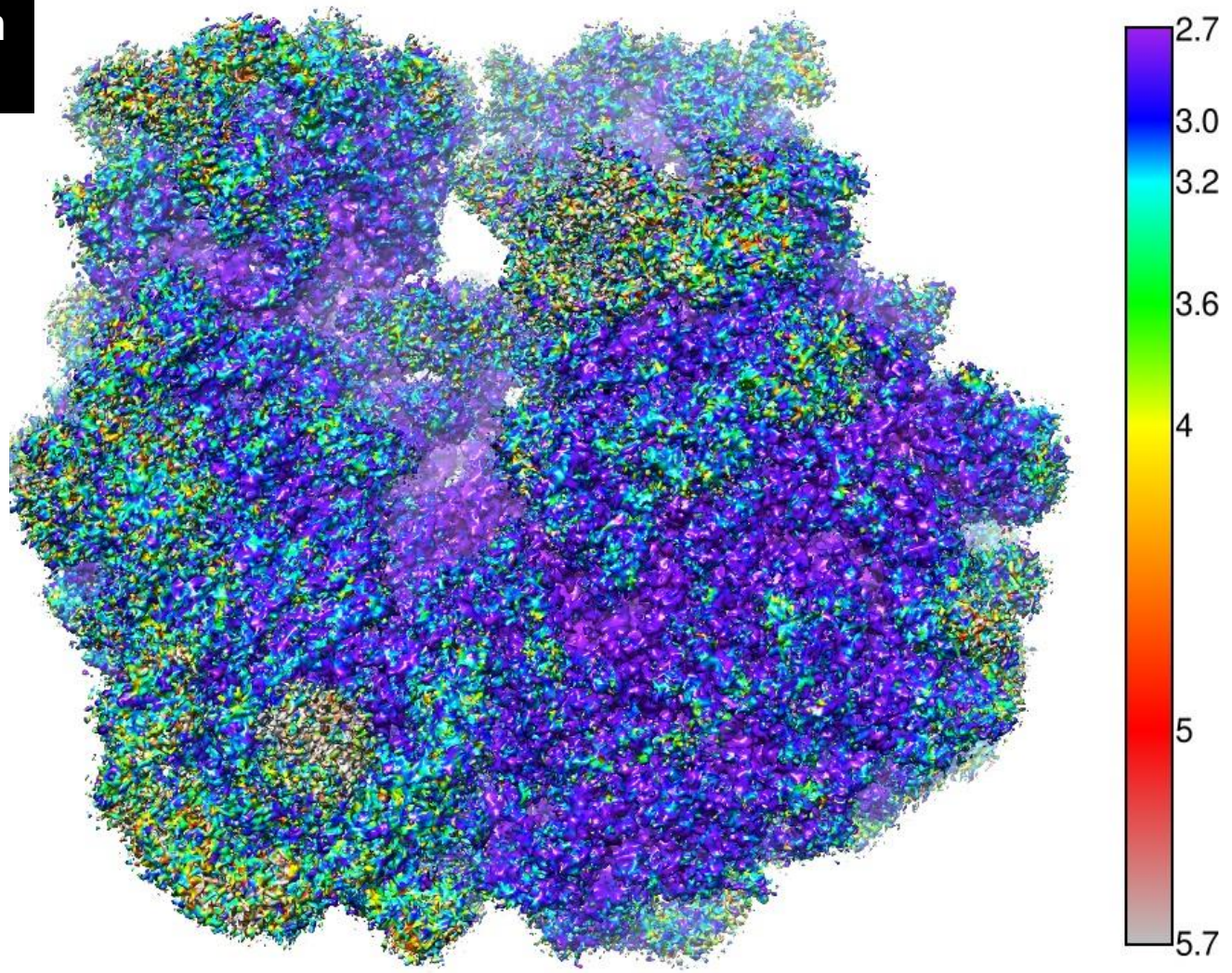


Just a bit on Cryogenic Transmission Electron Microscopy (CryoTEM)

RSF-Helmholtz 18-41-06001

Local resolution estimation (*ResMap*)





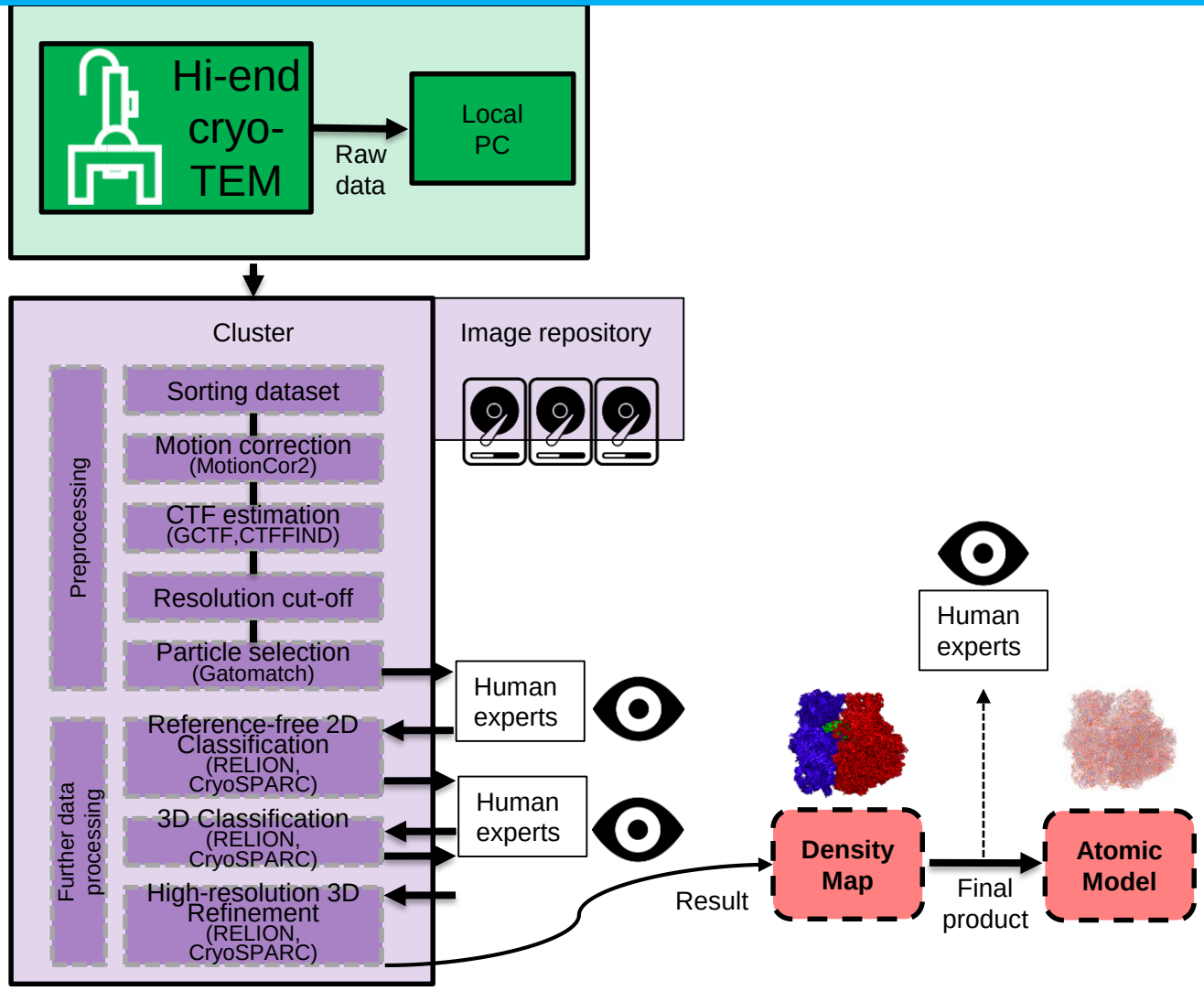
NEC'2019, Budva, 1st October 2019

The project target

- to provide processing and analysis of the **XFEL/cryo-EM** data flow to reconstruct 3D-structure in a **quasi real-time**
- corresponding **high performance data processing (HPDP) platform** is under development in the frame of this project
- the main experimental facilities are the **European XFEL** and the **cryogenic electron microscopy (cryo-EM)**
- the main advantage of this platform will be complete processing **pipeline from the experiment to reconstructed 3D structure of the biomolecule**

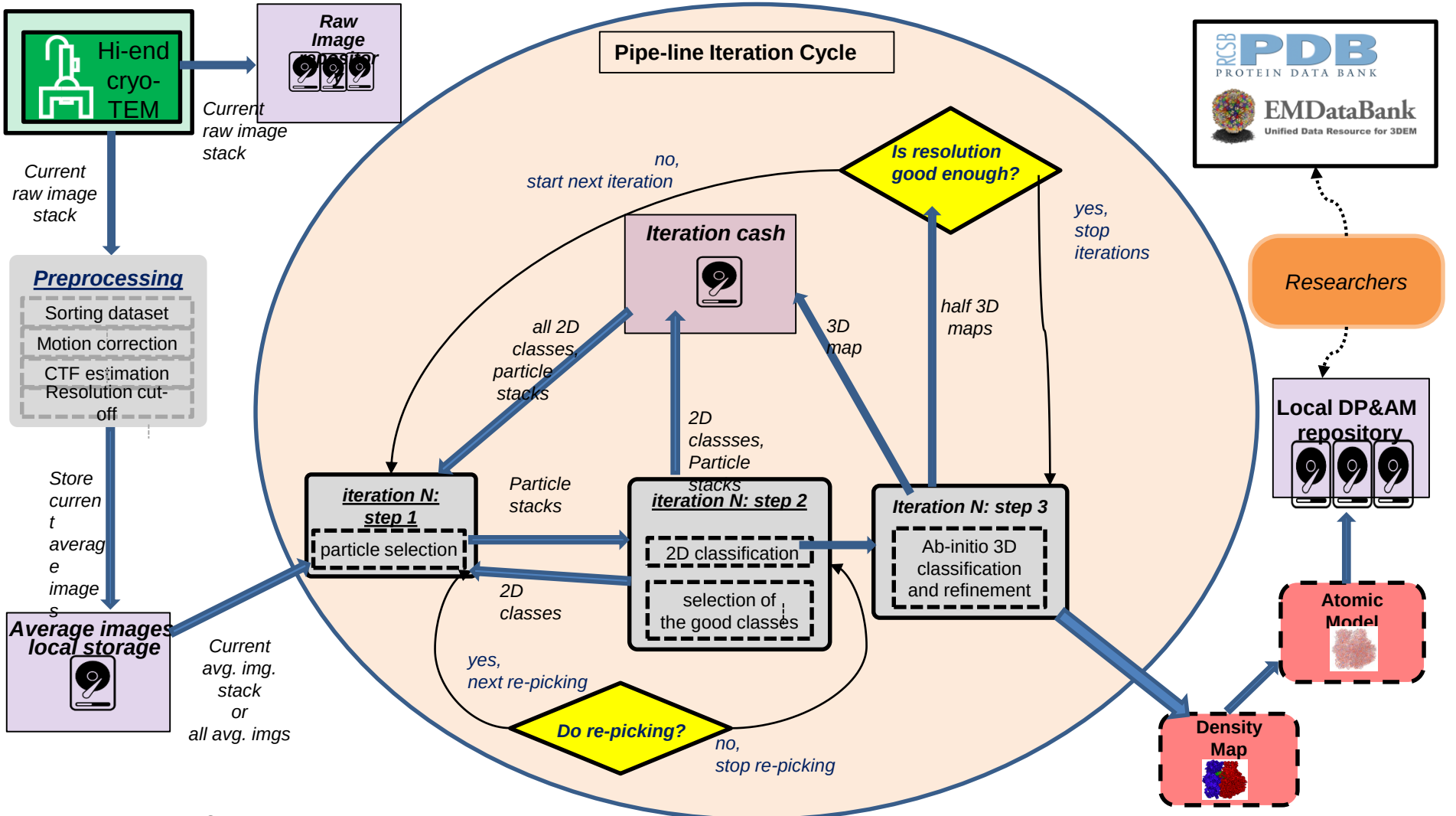
CryoEM current workflow (commonly used)

RSF-Helmholtz 18-41-06001



CryoEM pipe-line workflow *under realization*

RSF-Helmholtz 18-41-06001

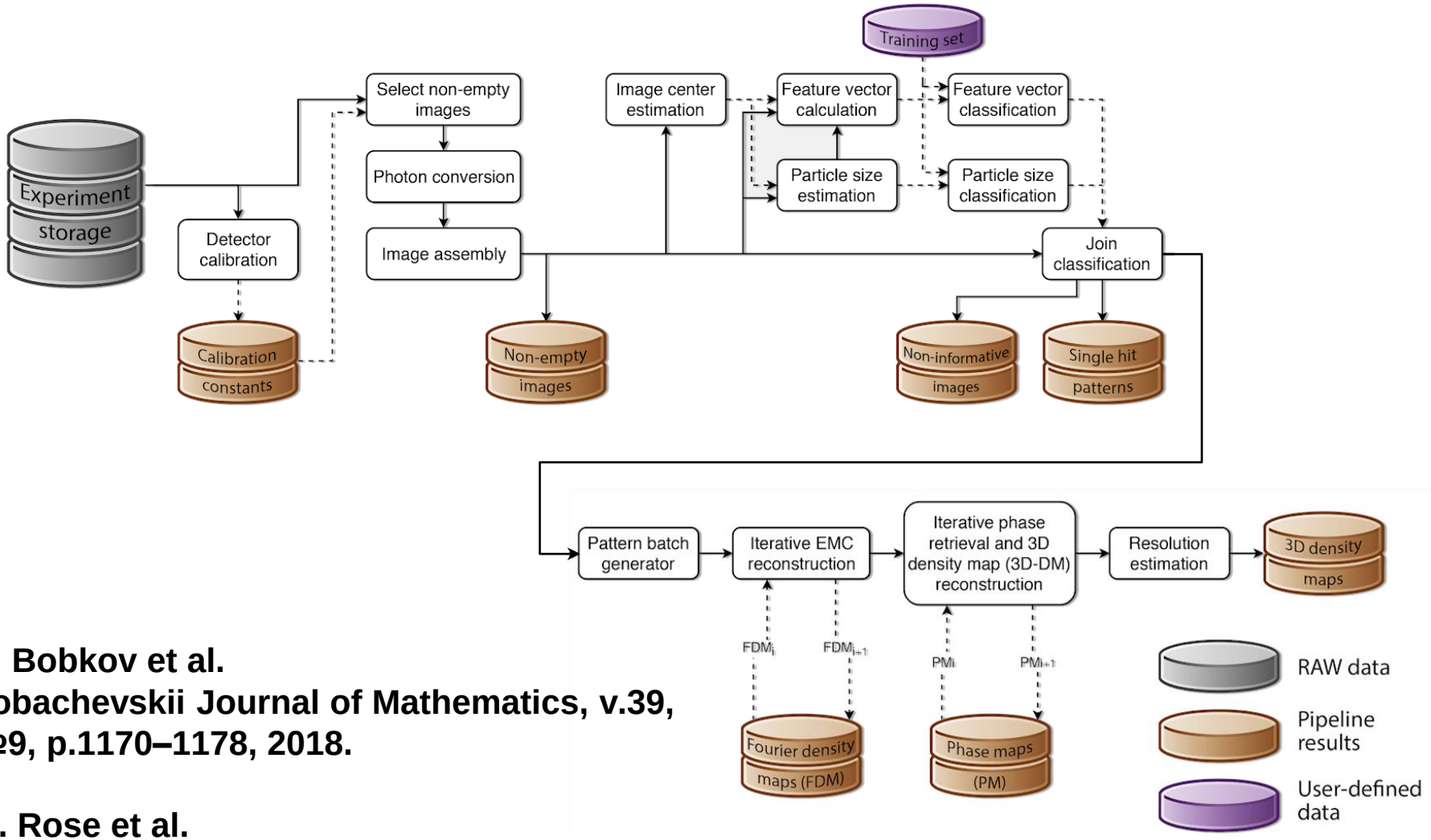


J. Phys.: Conf. Ser. v.955, p. 012005, 2018

NEC'2019, Budva, 1st October 2019

Workflow example: Eu-XFEL SPI data processing

RSF-Helmholtz 18-41-06001

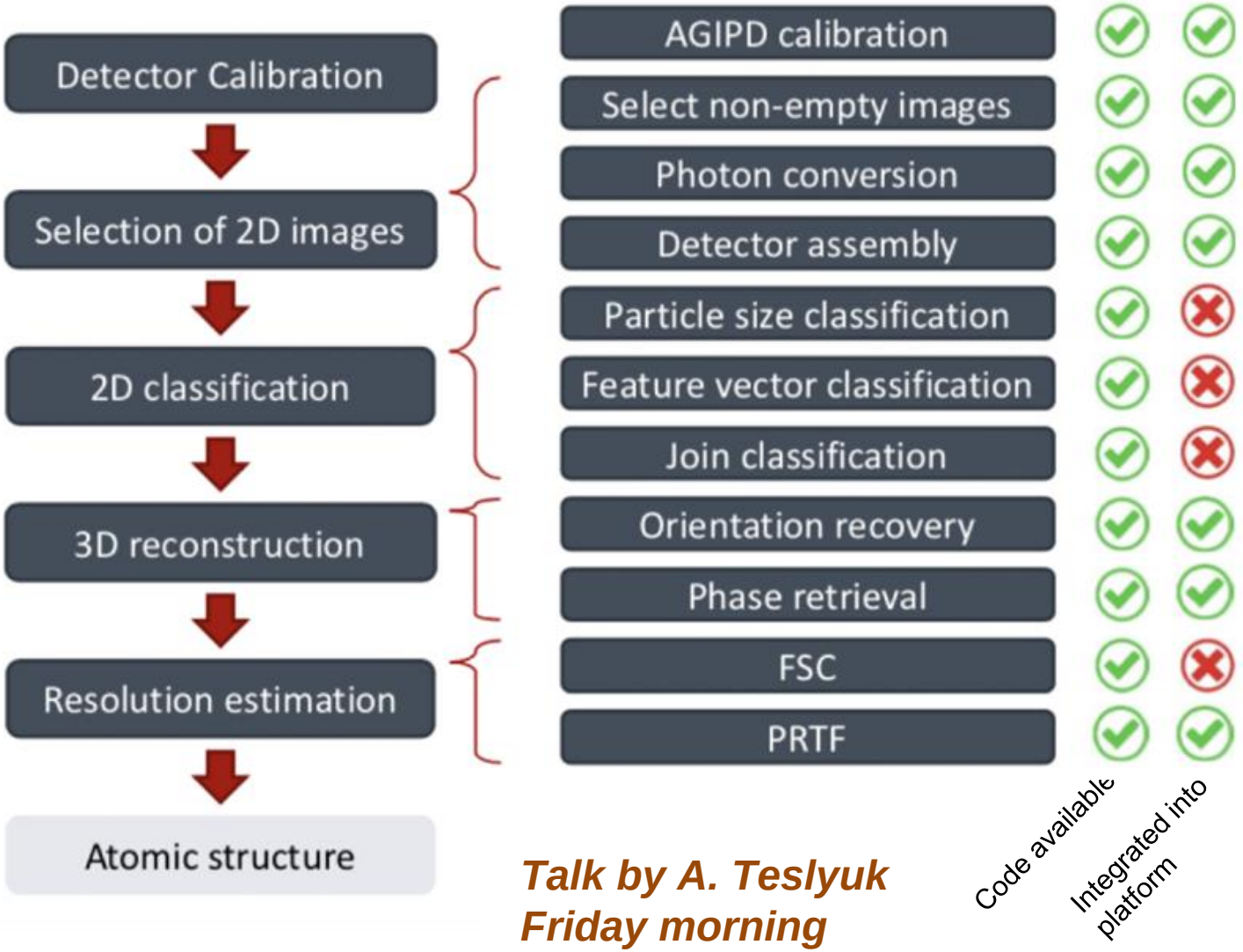


S. Bobkov et al.
Lobachevskii Journal of Mathematics, v.39, №9, p.1170–1178, 2018.

M. Rose et al.
International Union of Crystallography Journal (IUCrJ), v.5, p.727-736, 2018.

Toward to quasi-on-line HPDP platform: to base on containerised infrastructure, *today status*

RSF-Helmholtz 18-41-06001



Talk by A. Teslyuk
Friday morning

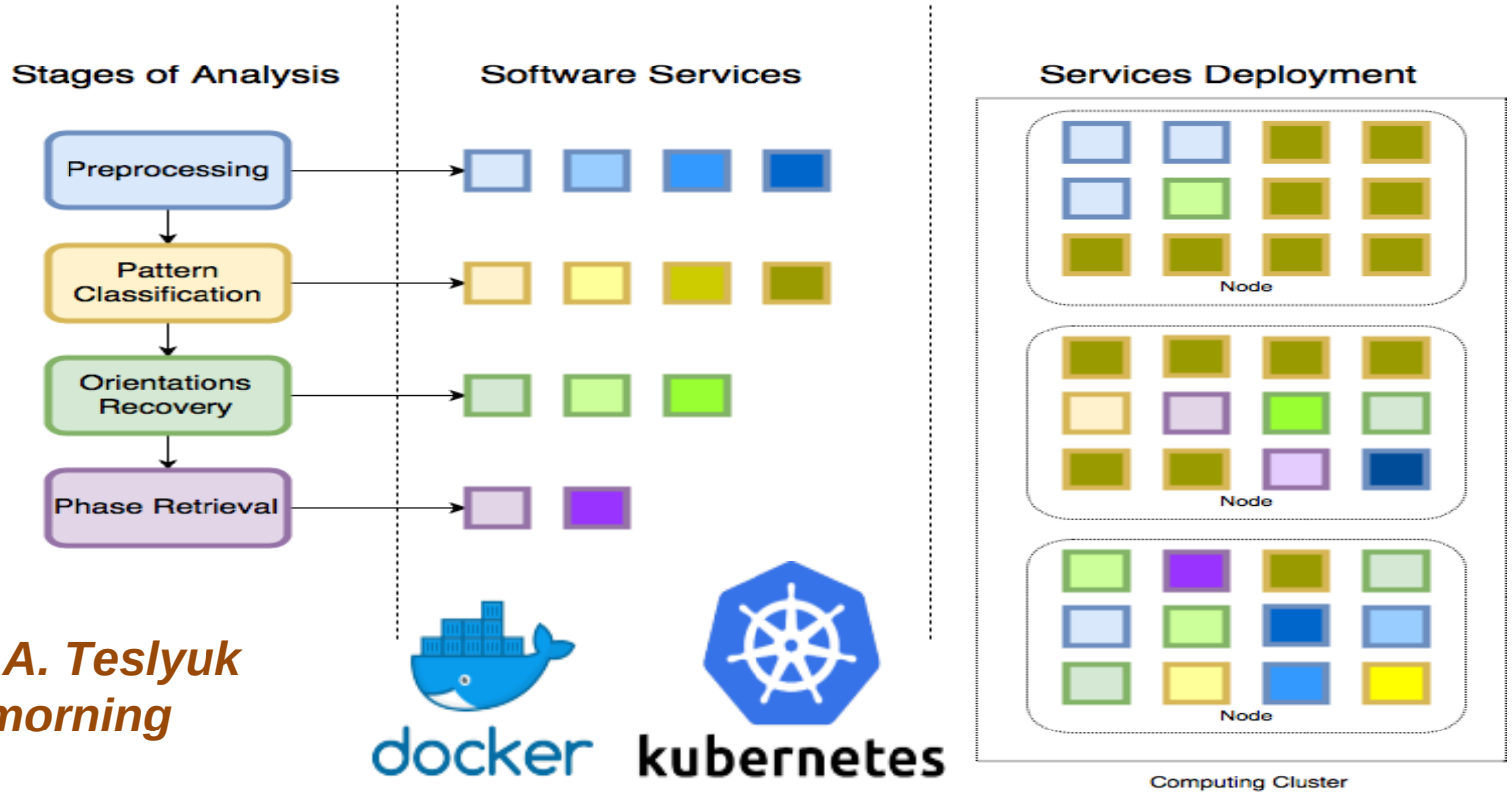
NEC'2019, Budva, 1st October 2019

Toward to quasi-on-line HPDP platform: to base on containerised infrastructure

RSF-Helmholtz 18-41-06001

Key motivations:

- **easy deployment** on heterogeneous computing resources
- easy realization of **different workflows**
- **parallelization**



Talk by A. Teslyuk
Friday morning

Key parameter – resolution in the 3D structure reconstructed

RSF-Helmholtz 18-41-06001

The most important parameter for structural biology is spatial resolution – *what details could be distinguishable in the 3D structure of the biological objects*.

The highest SPI resolution by CryoEM *today* is 1.62 Å

***Today* resolution in Eu-XFEL SPI experiments is two-order worse. While starting the operation in the end of 2017 Eu-XFEL will come to the designed parameters 2-3 years later. The designed spacial resolution in SPI experiments would be ~ 1 Å.**

1 Å (0.1 nm) resolution corresponds to typical inter-atomic distance

Examples of ML use:
first stages of the analysis of SPI LCLS data (sept. 2018)

RSF-Helmholtz 18-41-06001

Motivation to use ML methods in HPDP platform in SPI experiment data analysis (as in cryo-EM):

great progress in image analysis by use of (deep) ML;

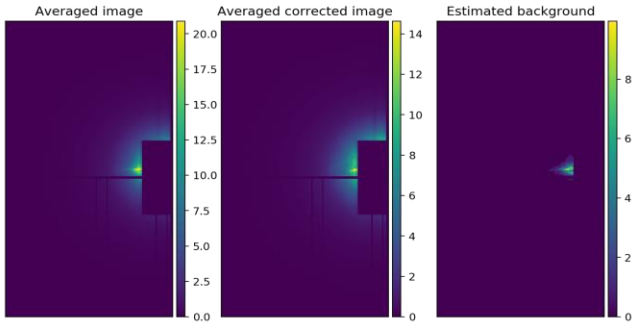
large number of poor understandable parameters (beam, detector, samples etc.);

high dimension data; …

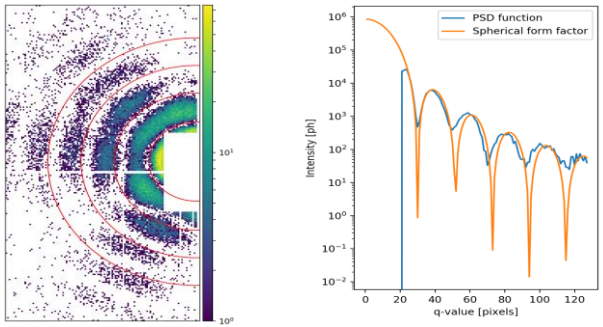
see, e.g., **S. Bobkov, PhD dissertation (MEPhI, Moscow, oct. 2018)**

S. Bobkov. Journal of Synchrotron Radiation, vol.22, №6, p. 1345-1352, 2015.

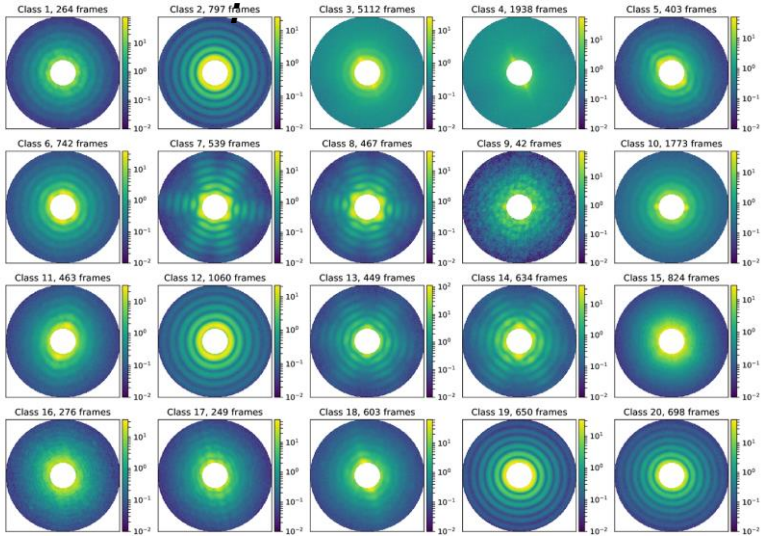
Background correction:



Particle size estimation:



Classification



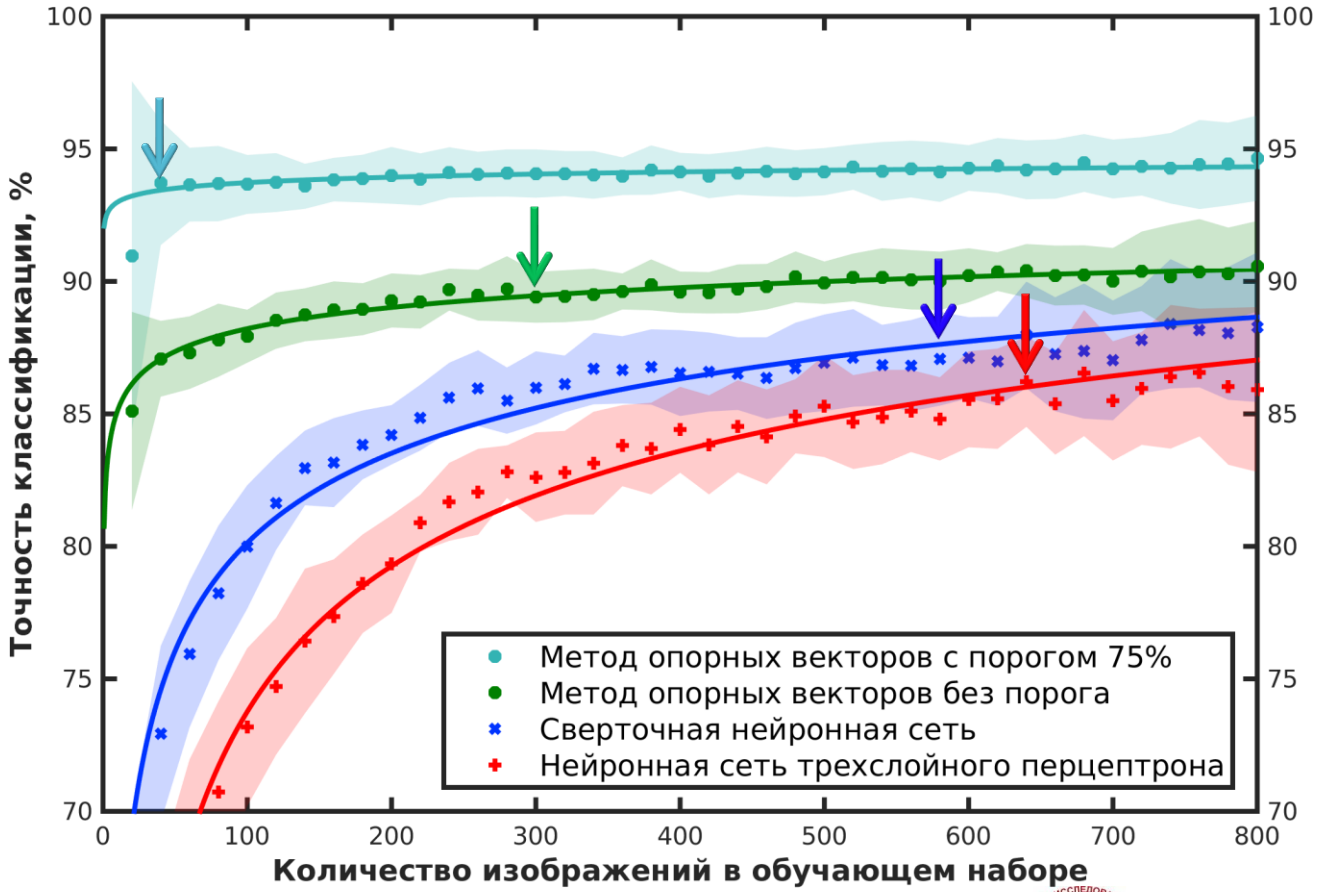
Результаты точности разработанных ML методов на блоках экспериментальных данных, полученных на LCLS

	Точность (Полнота)			
	CXIDB ID 13,14	CXIDB ID 10,11	CXIDB ID 20,25,37	CXIDB ID 25
Классификация характеристических векторов методом опорных векторов с порогом 75%	**94.1%** 80.8%	**99.8%** 99.8%	**99.9%** 98.9%	**95.0%** 79.6%
Классификация характеристических векторов методом опорных векторов без порога	**90.8%** 90.8%	**99.8%** 99.8%	99.6% 99.6%	**90.7%** 91.5%
Классификация характеристических векторов на основе метода k-средних	82.9% 78.2%	71.2% 53.6%	86.1% 80.1%	64.4% 99.8%
Классификация характеристических векторов на основе метода спектральной кластеризации	85.8% 83.4%	79.0% 53.4%	84.1% 73.7%	68.6% 99.3%
Классификация изображений на основе нейронной сети трехслойного персептрона	87.1% 96.9%	97.7% 97.6%	**99.8%** 99.6%	85.8% 91.2%
Классификация изображений на основе свёрточной нейронной сети	88.5% 88.3%	99.4% 99.3%	**99.8%** 99.8%	85.8% 95.5%

Оптимальный размер обучающего набора для блока CXIDB (ID 13, 14) LCLS

Оптимальный размер обучающего набора – точность классификации достигает 99% от максимума (на правой границе). Всего в блоке 958 изображений.

Метод опорных векторов с порогом 75%	40
Метод опорных векторов без порога	300
Сверточная нейронная сеть	580
Нейронная сеть трехслойного персептрона	640



Сценарий потоковой классификации данных в SPI экспериментах на European XFEL

Возможный сценарий потоковой классификации данных на European XFEL:
(оценки времени для классификации на основе метода опорных векторов с порогом 75%
на приведенных ранее аппаратных ресурсах)

1. Характеристические вектора изображений рассчитываются в режиме онлайн.
2. После начала работы, накапливается обучающий набор оптимального размера. (< 1 минуты)
3. Ручная разметка обучающего набора. (≈ 10 минут)
4. Машинное обучение. (< 5 секунд)
5. После завершения обучения, новые дифракционные изображения классифицируются в режиме онлайн.
6. Сохраненные ранее изображения могут быть классифицированы после завершения эксперимента или параллельно, за счет избыточной производительности аппаратных ресурсов.

Могут использоваться все разработанные методы классификации по типам структуры, однако оптимально – классификация на основе метода опорных векторов с порогом вероятности корректной классификации в 75%, т.к.:

- *достигается высокая точность,*
- *оптимальный обучающий набор имеет минимальный размер.*