Grand Challenges

Building a Collaborative Computational Project for Volume Electron Microscopy (CCP-volumeEM)

Over the next five to ten years, CCP-volumeEM will help the community to realise the full capabilities of this sought after technology by adopting and disseminating the latest computational methods

The Community

Imaging biological material across functional length scales and at high resolution

VolumeEM (vEM) is a group of techniques that are used to image biological material through continuous depths of at least one micrometre at nanometre resolution. vEM is applied extensively in neuroscience to image brain tissue in sufficient detail to map the wiring of neurons. vEM techniques are now being applied to a wide range of biological systems, bridging the gap between the functional imaging of fluorescence microscopy and the molecular detail of cryoEM. The scientific value of insights that have been made possible via vEM have placed this family of techniques amongst the most important scientific technologies of recent years, having been named as one of the "Seven technologies to watch in 2023" by *Nature*, and as the second most requested technology of 2024 in the Euro-Bioimaging Annual Report. vEM techniques use highly automated instruments that are capable of producing data well into the multi-terabyte regime per sample, leading to many challenges in data management, processing and interpretation. AI techniques are becoming widespread to automate the identification of features in complex data, and there is a pressing need for a collaborative community to coordinate and standardise efforts in the field.

The Challenge

Overcoming sheer scale of data, lack of training, limited data formats, financial and experimental costs and ethical imperatives

Historically, much image analysis performed in biological vEM has involved a great deal of manual effort, since many of the computational methods that work well in other imaging modalities, such as fluorescence microscopy, are unable to perform adequately with the complexities of vEM data. Recent years have seen AI based analysis methods begin to perform well in some applications, particularly in the area of "segmentation", an image analysis process used to delineate objects of interest from their surroundings. However, the typical AI problems of obtaining suitable training data and creating models that are sufficiently generalised to work across the necessary range of data are commonplace and can typically only be overcome by significant community effort. A major challenge to the robust usage of modern computational techniques in vEM is the sheer scale of the data, as well as the broad range of backgrounds of the various stakeholders, with many end-users being biologists with little-to-no formal training in computational methods. Bridging these experimental-to-computational domain boundaries is a familiar problem in bioimage analysis more generally, especially where much of the accumulated knowledge and tooling has grown organically over many years into a very heterogeneous ecosystem deployed across multiple incompatible platforms. Prevailing image data formats in use are generally bitmap-style formats like TIFF and a range of proprietary formats, which need to be loaded directly into RAM, placing hard limits on the size of data that can be analysed without downscaling or cropping. As datasets get larger and more complex, the resulting lack of standardisation or community-established best practices for dealing with larger-than-RAM data has led to a hard-to-navigate landscape, with much missing functionality. These difficulties often result in limited extraction of information from the incredibly rich, dense datasets typical of vEM. As well as the associated experimental costs, there are also ethical imperatives to ensure that data from clinical or animal studies are used effectively, squeezing maximum scientific benefit from each experiment, in accordance with the UKRI principle of Replacement, Refinement and Reduction of animals in research.

The Solution

Harnessing AI technologies, integrating image processing functionalities and adopting standards for multimodal data

The advances in imaging instrumentation and computational techniques used in vEM show no sign of slowing down, and the ability to flexibly adopt new and emerging computational methods will be key to ensure the activities of CCP-volumeEM remain at the forefront. In particular, harnessing the rapid evolution of AI techniques will benefit greatly from increased standardisation and modularity in workflow-building principles and implementation, as well as from the sharing of FAIR datasets in public repositories. A comprehensive integration of next generation file formats (NGFF)-enabled image processing functionalities into end-user accessible software tools and interfaces will allow transparent-to-the-user incorporation of the latest technologies but will require significant developer effort to achieve full coverage. Such endeavours have been largely neglected under traditional funding and publication models due to their long-term nature – however, a CCP is an ideal custodian of such critical development activities. Another rapidly advancing frontier in biology research is the use of multimodal data, whether that's multiple imaging modalities, for example volume correlative light and electron microscopy (vCLEM), or experiments also incorporating genetic, molecular and other data types. In this context, the cross-cutting activities of CoSeC will be crucial, enabling CCP-volumeEM to work with other CCPs and HECs, for example CCP-EM, CCP4, HECBioSim and others, in order to ensure effective cross-domain communication and standardisation to unlock the possibilities that such multimodal methods promise.

Grand Challenges

The Outcome

Bridging experimental to computational domain boundaries with new computational technologies that can enable efficient analysis of large datasets

By fostering a community of stakeholders from across the whole spectrum of domain experts and end-users, CCP-volumeEM will support the development and maintenance of a range of resources that will allow practitioners to access a set of high-quality, interoperable, and benchmarked tools and libraries. Through the promotion of NGFF, which use a chunked storage strategy to enable lazy-loading of data, the ability to efficiently visualise, process and analyse these ever-growing datasets will be enhanced. New functionality to provide standard image processing operations to NGFF data will be available to enable efficient processing and analysis of such datasets. In addition, the establishment of appropriate schemas and ontologies will enable the rich experimental and imaging metadata to be recorded in consistent formats that allow the prospect of much more extensive mining and re-use of these high value datasets.

Key deliverables for this will be the sharing of public resources such as containerised builds of tried-and-tested collections of tools that can be deployed on any platforms, from laptops to HPC and cloud infrastructure. Exemplar workflows demonstrating the use of NGFF will be shared and documented thoroughly to demonstrate how to sustainably build pipelines for other use-cases, using research software engineering and data stewardship best practices. Associated training materials will be created and delivered to provide taught and self-guided resources for the community. The developer community will be grown through a series of focused hackathons, interfacing with related communities where appropriate.

More Information

CoSeC
www.CoSeC.ac.uk
CoSeC@stfc.ac.uk

CCP-vEM
www.ccp- volumeem.ac.uk
(in development)