

The **SciLifeLab** Data Repository from a user perspective

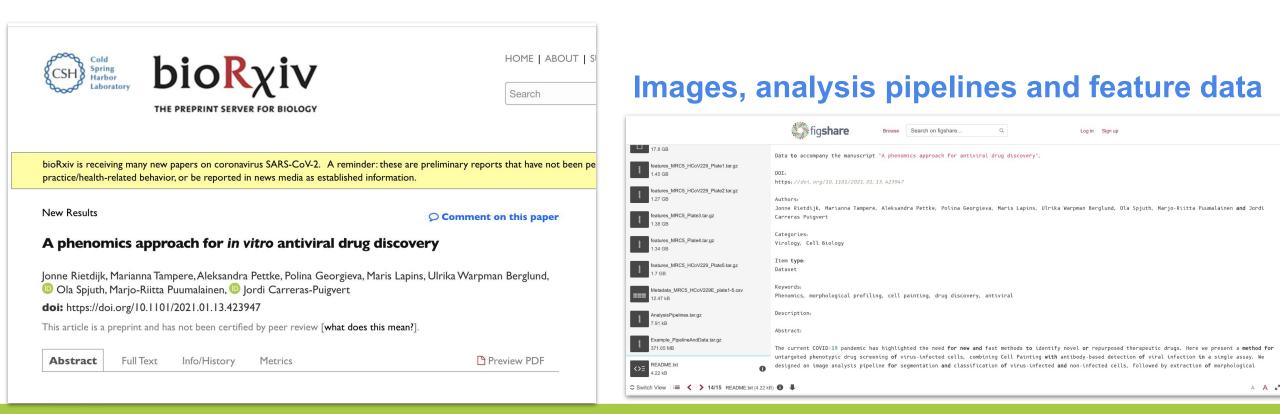
Jonne Rietdijk





Dataset accompanying manuscript:

"A phenomics approach for in vitro antiviral drug discovery"

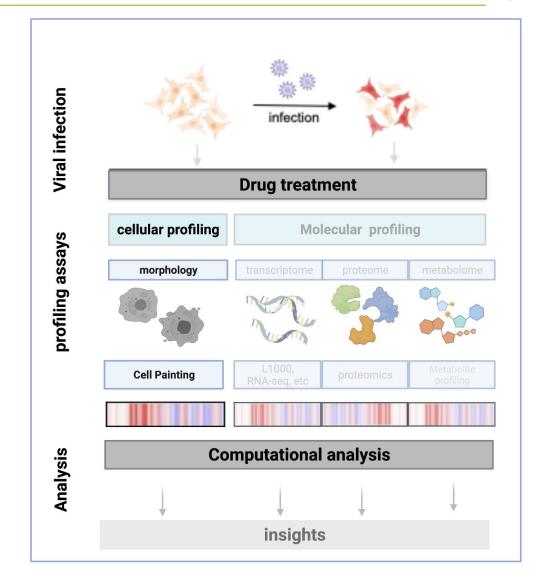


phenomics

- Study changes in cellular morphology
- Image-based morphological profiling of cells using multiplexed fluorescent dyes
- Cell Painting protocol (Bray et al. 2016)



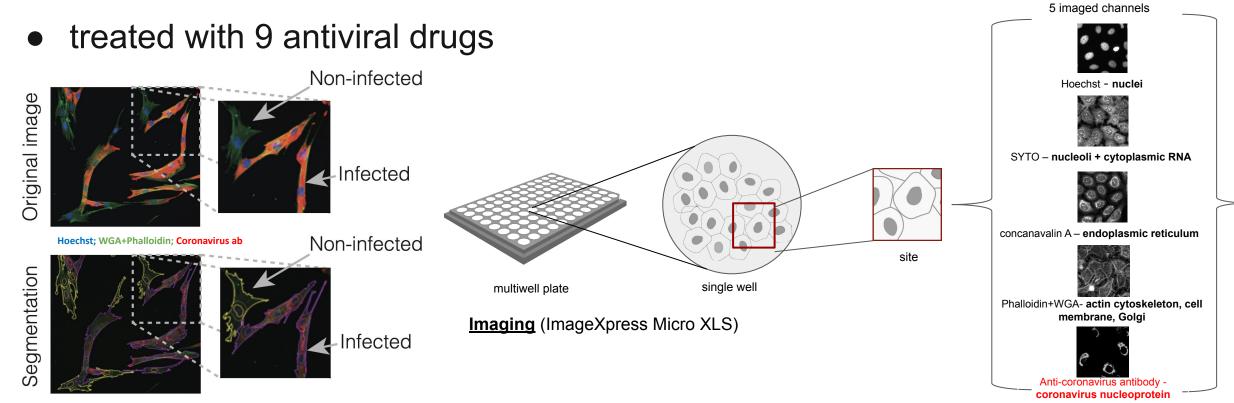
https://carpenterlab.broadinstitute.org/blog/



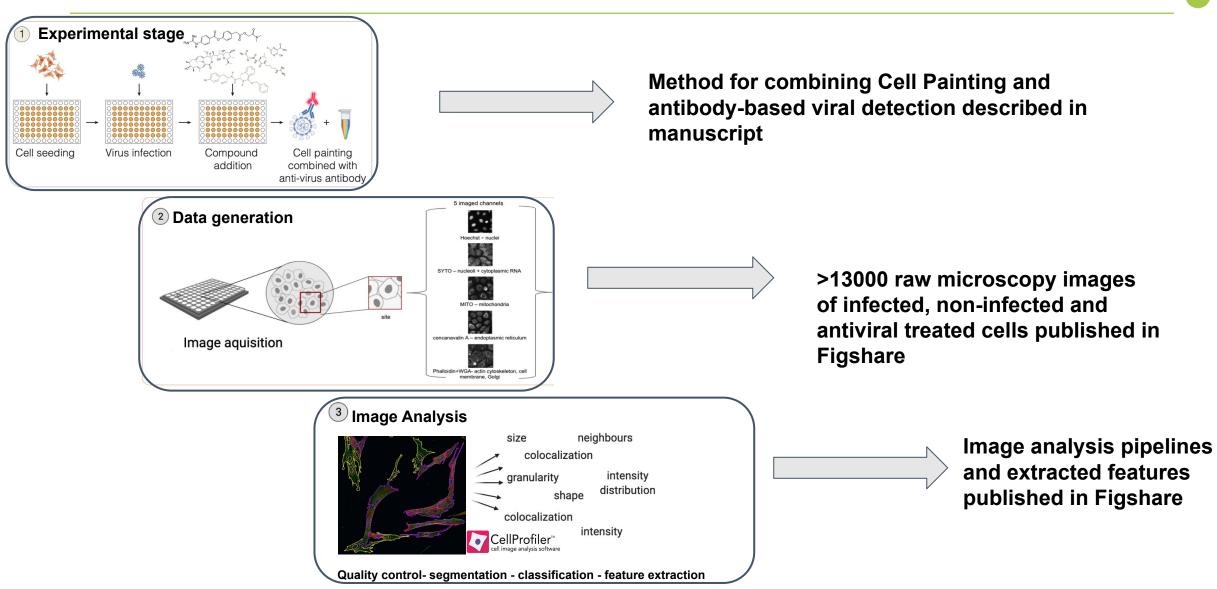


Morphological profiling of infected cells

- human lung fibroblasts (MRC-5) cells infected with human coronavirus 229E
- stained for 7 important cell components + virus nucleoprotein



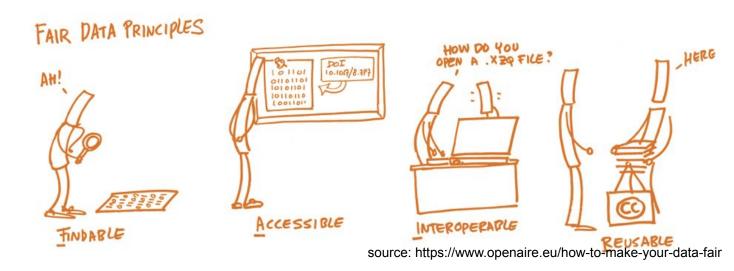
Images, pipelines and features



Why we use the repository



- FAIR
- High-content data
 - multiple questions can be asked
 - Different analysis methods can give new insights
- Flexibility to upload big files and different types of files



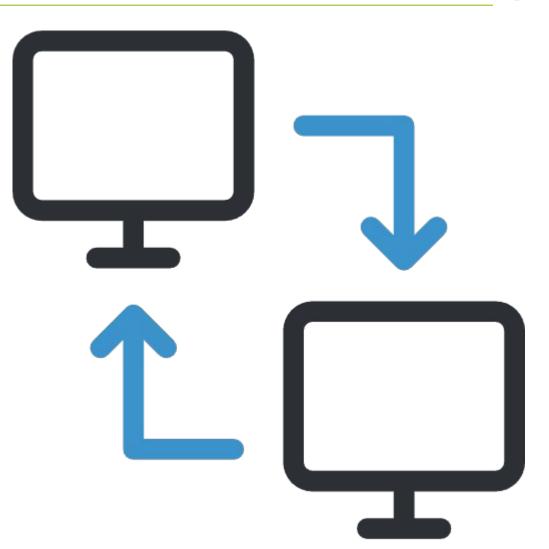
Uploading big data

- Good documentation on how to upload big datasets

https://www.scilifelab.se/data/repository/submission/

https://github.com/ScilifelabDataCentre/SciLifeLab-Data-Repository-API-examples

 Easy to do streaming uploads using a FTP uploader, Python API or bash scripts. Also it is possible to upload from S3 storage backends





Also take a look at these



Select an option

- datasets
- posters
- presentation
- soon: Deep learning models

{⊘}pharm<mark>b.io</mark>

Research group website: http://pharmb.io

LNP drug delivery image data		Synergy Conformal Prediction for Reg Predicting target profiles with c	onfidence as a service using
Cite Download (3.5 GB) Share Embed + Collect	273 31 0 views downloads citations	docking scores	
		Cite Download (297.65 kB) Share Embed + Collect 20	downloads22 0itation
Dataset posted on 24.06.2020, 19:18 by Philip Harrison, Håkan Wieslander, Alan Sabirs Johan Karlsson, Victor Malmsjö, Andreas Hellender, Carolina Wählby, Ola Spjuth	h,	+ Follow Posted on 16.10.2020 - 05:33	views citations
Johan Kansson, victor Mainisjo, Andreas Heilender, Garonna Wahiby, Ola Spjuri	CATEGORIES	Abstract Background Identifying and assessing ligand-target bindi	na is a core component
Data to accompany the manuscript "Deep learnings models for lipid-nanoparticle-based	Cell Biology	Presentation posted on 07.02.2021, 20:08 by Ola Spjuth, nihal in early drug discovery as one or more unwanted interactions may	be associated with
drug delivery"	KEYWORDS	safety issues. Contributions We present an open-source, extendal predicting larget profiles with confidence using machine learning fr	ble web service for Read the peer-revi
Abstract:	lipid-nanoparticles	Gene Expression and Phenotypic Profiling Data materia	
Large-scale time-lapse microscopy experiments are useful to understand delivery and expression in RNA-based therapeutics. The resulting data has high dimensionality and	time-lapse microscopy	Presentation by Ola Spjuth at the conference International Committee Internati	dock chemiting of a 1 scores 0 0
high (but sparse) information content, making it challenging and costly to store and		Recognition Applications and Methods ICPR (M202) - http://www.adatan.com/adatan/com/adatan	pound basis. Results downloads cita
process. Early prediction of experimental outcome enables intelligent data management and decision making. We start from time-lapse data of HepG2 cells exposed to lipid-	drug delivery	Dataset posted on 11.09.2019/15.07 by Maris Lepins, Ota Sportk ^{1/V} II. The model pro	dictions showed SPRINGER NA
nanoparticles loaded with mRNA for expression of green fluorescent protein (GFP). We	LICENCE	Large and distributed data sets pose many challenges for maccomparable performance to molecular docking scores against an	external validation set. etes ensureCATEGORIES AUTHORS (9)
hypothesize that it is possible to predict if a cell will express GFP or not based on cell morphology at time-points prior to GFP expression. Here we present results on per-cell	CC BY 4.0	requirements on computational resources an Data sets of manuscript "Evaluation of Gene Expression and Phenotypic Profiling Data	ta as • Cell Biology A Laeeg Ahmed
classification (GFP expression/no GFP expression) and regression (level of GFP		multiple models in parallel on subsets of date Quantitative Descriptors for predicting Drug Targets and Mechanisms of Action" by L	
expression) using three different approaches. In the first approach we use a convolution neural network extracting per-cell features at each time point. We then utilize the same	Predictive models for o	off-target binding profiles generation and Spluth	Cite other genome-wide approaches)
features combined with: a long-short-term memory (LSTM) network encoding temporal	Select an option V) the hand ed the base of the detailed be avoided. Com 98 History 6 is from the one 0 and the point mouse over the clatter text to the one of the detailed by the avoided by the detailed by t	Staffan Arvidsso
dynamics (approach 2); and time-series feature extraction using the python package tsfresh followed by principal component analysis and gradient boosting machines	Cite Download (79.53 MB	synergy rules has been shown to have adval + 11.09.2019 - First online date. Posted date	sson, Jonathan Alvars
(approach 3), to reach a final classification or regression result. Application of the three		synergy rules has been shown to have adval a 11.09.2019 - First online date, Posted date al. (2020): Predicting target profiles with	confidence as a gene expression Anvid Berg
approaches to a previously unanalyzed test set of cells showed good predictive performance of all three approaches but that accounting for the temporal dynamics via		by Lampa, Samuel, Alvarsson, Jonathan, Arvidsson https://doi.org/10.608/instrance.com/10.608/instran	cial Intelligence
LSTMs or tsfresh led to significantly improved performance. The predictions made by th	Mc Shane, Staffan, Berg, Arvid, Ahlbe		Phenotypic Profilin Anders Larsson
LSTM and tsfresh applications were not significantly different. The results highlight the benefit of accounting for temporal dynamics when studying drug delivery using high		and cross-Transfer learning with deep convolutional neural network	is for classifying c
content imaging.		ig, built with Conformal Prediction, and the CPSign	
Didhan andar		coming publication (Manuscript in preparation), the second problem data, such as when working in	
https://github.com/pharmbig/phil_LNP_modelling	which will provide more details. The dataset is a GZipped Tar archive	with the models as Java Archive (JAR) files. Ford (976.2748) Share Embed + Collect	E 125 25
		conding audit log, with the extension ".audit.json", • Genetics	views downloads
		SciPipe) used to train the models. This audit file Molecular Biology	CC BY 4.0
	contains all the shell commands used	in the workflow that produced the models. 08.2019 5 Ecology Biological Sciences not elsewhere billip Harrison, Ola Spjuth	
		Biological Sciences not elsewhere Export Export Export	
armb.io	HISTORY	Information Systems not elsewhere	CATEGORIES
	• 09.06.2018 - First online date, Pub	lication date, Posted date study we applied pre-trainclassified is to predict cell mechanisms of action (MoAs) in le	t an option formatics
2		response to chemical perturbation Cancer vo cell profiling datasets from the Broad	
		pe Benchmark Collection (bbbc) and obtained higher predictive accuracy than	
		sly reported, between 95 and 97%.	KEYWORDS
			high-content imaging, ce
		Y Y	
ALL & MARY MARY	THE -101	3.2019 - First online date. Posted date	LICENCE
			CC BY 4.0
			CC BY 4.0
The second se			EXPORTS
A REAL PROPERTY AND A REAL			EAFORIS
THE DESIGNATION AND DESIGNATION OF THE OWNER.			Colort on option