**Bridge2AI Cell Maps for Artificial Intelligence (CM4AI) - Standards Module**

**Deliverable 1.3: Master Data Dictionary Version 1.0**

Object Metadata, Provenance, & Dataset Schemas

CM4AI Standards Module – University of Virginia – Contact: Tim Clark [twc8q@virginia.edu](mailto:twc8q@virginia.edu) – 2202.12.31

**Abstract**

This document presents Version 1.0 of the Cell Maps for Artificial Intelligence (CM4AI) Master Data Dictionary. It describes the object metadata, including detailed provenance and dataset schemas, for the Cell Maps for Artificial Intelligence project’s Music 1.0 cell architecture pipeline. It is accompanied by a graphical **Provenance Map** and a **Machine-readable Representation** in linkML.

**Introduction**

The Cell Maps for Artificial Intelligence (CM4AI) project is a component of the NIH Bridge2AI program (National Institutes of Health 2022), whose objective is to move biomedical research toward large-scale adoption of artificial intelligence (AI) and thereby significantly improve tractability of complex biomedical investigations “beyond human intuition” . The CM4AI Standards Module is charged with producing, among other outputs, a master Data Dictionary (DD) for the project to enable interoperability and intelligent re-use of CM4AI tools and data. Version 1 of this data dictionary is an early-stage deliverable in month four of the project, and therefore based on preliminary data and software.

The data, software, and processing steps modelled here, originated in the Multi-Scale Integrated Cell (MuSIC) map described in (Qin et al. 2021) and consists of analytic processing of embeddings of AP-MS and subcellular imaging data from the Gygi Lab’s BioPlex 3.0 project (Huttlin et al. 2021) and the Lundberg Lab’s Human Protein Atlas work (Thul et al. 2017; Mahdessian et al. 2021). From this analysis, the MuSIC 1.0 pipeline to produced a machine-processable, hierarchical cellular architecture model. During the CM4AI project, the MuSIC pipeline will be enhanced and improved (MuSIC 2.0) and similar techniques will be applied to select cell types, with the addition of single-cell RNA seq data, to produce a revised cell architecture model. This will then be used to develop a visual deep learning model of cellular interactions to genetic and biochemical perturbation.

As an early-stage deliverable, we necessarily began with the preliminary data available from MuSIC 1.0. A subsequent version of this document and associated software will be developed in collaboration with the MuSIC 2.0 developers in the CM4AI Tools module, and extended deeper into the AP-MS, subcellular imaging, and scRNAseq laboratory and instrumentation processes in collaboration with the Data Acquisition module.

The present document provides the CM4AI Master Data Dictionary, Version 1.0, in human-readable form. It includes supplementary partial descriptions of predecessor objects from processes and computations in the Gygi (BioPlex) and Lundberg (Human Protein Atlas Labs**.** A **Machine-readable Representation** of this material in the linkML (Moxon et al. 2020) knowledge representation language along with a graphical **Provenance Map** of the MuSIC 1.0 pipeline are provided here: <https://github.com/fairscape/B2AI>.

A subsequent version of this document will describe the MuSIC 2.0 pipeline and will be developed during construction of MuSIC 2.0 and in collaboration with its developers, rather than in retrospect as a study of an earlier pipeline version as done here.

**Conventions**

In this document, the labels “Used by” (this icon in the diagram →) and “Generated by” (inverse of this label →→ in the **Provenance Map**, evi:generates) correspond to “WasUsedBy” and “WasGeneratedBy” in W3C PROV and to “UsedBy” and “GeneratedBy” in EVI, or their inverse properties, and are applied when datasets are known to be input to or output by a specific program. The label “Derived from” (inverse of - - -> in the **Provenance Map**, evi:derivedTo, proposed synonym: hasDerivation) corresponds to W3C PROV “WasDerivedFrom” and is used where we unable to determine precisely how the dataset was derived from a predecessor dataset or repository (see **Provenance Map** legend).

Where “TBD” appears associated with any label, our best efforts were used to determine a value, but none was available at the time this document was produced.

All referenced files for ContentURL, where given, are from the Example in <https://idekerlab.ucsd.edu/music/> - not from the actual datasets used in or generated by the MuSIC 1.0 analysis, which were unavailable to us at the time of publication due to a labor dispute.

All proteins are identified by their corresponding human gene symbols.

**MuSIC 1.0 Data, Software, and Provenance**

1. **GygI Lab - BioPlex Computations**

Computation: *node2vec predict*

Run by: TBD

Date: TBD

Description: Create 1024 dimension embeddings based on BioPlex 2.0 HEK293 cell interaction data using trained node2vec model.

Associated publication: Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (ACM, 2016). doi:10.1145/2939672.2939754.

More Info: Huttlin EL, et al. (2017) Architecture of the human interactome defines protein communities and disease networks, Nature 545): 505-509 https://doi.org/10.1038/nature22366.

Used software: node2vec

Used datasets:

BioPlex interaction data

node2vec trained model

Generated datasets:

AP-MS embeddings

Computation: CompPASS-Plus

Run by: TBD

Date: TBD

Description:.

Associated publication: Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864 (ACM, 2016). doi:10.1145/2939672.2939754.

More Info: Huttlin EL, et al. (2017) Architecture of the human interactome defines protein communities and disease networks, Nature 545): 505-509 <https://doi.org/10.1038/nature22366>.

Used software: CompPASS-Plus (https://github.com/HMSBioPlex/CompPASS-Plus-CLI)

Used datasets:

BioPlex interaction data

node2vec trained model

Generated datasets:

AP-MS embeddings (see MuSIC 1.0 Datasets)

1. **GygI Lab - BioPlex Datasets**

*Bioplex interaction data*

Dataset: Bioplex\_interaction\_data

Author: Gygi Lab

Date: TBD

Version: 2.0

Description: Protein Biophysical Associations - (Affinity Purification Mass Spectrometry - AP-MS)

More info: Huttlin EL, et al. (2017) Architecture of the human interactome defines protein communities and disease networks, Nature 545): 505-509 <https://doi.org/10.1038/nature22366>.

Format: TSV

Schema: Name str, Represents str, SecondaryID str, type str, prob(W) float, p(NI) float, p(int) float

Generated by: CompPASS-Plus

Used By: node2vec predict

1. **Lundberg Lab - Human Protein Atlas Computations**

Computation: Densenet predict (predict\_d121.py)

Run By: ( Lundberg Lab @ KTH-Royal Institute of Technology and Stanford)

Date: TBD

Description: compute 1024 feature vector for each protein image

Associated publication: Le, T. *et al.* Analysis of the Human Protein Atlas Weakly Supervised Single-Cell Classification competition. *Nat Methods*

**19**, 1221–1229 (2022).

Associated publication: Wisnes, C. On computational methods for spatial mapping of the human proteome. Doctoral Thesis in Biotechnology. Stockholm: KTH Royal Institute of Technology, 2022. https://www.diva-portal.org/smash/get/diva2:1711195/FULLTEXT02.pdf

Used software: Densenet predict (predict\_d121.py)

Used dataset:

Densenet model (<https://github.com/CellProfiling/densenet/releases/tag/v0.1.0>)

MuSIC images ([resized HPA images])

Generated dataset:

Image embeddings (see MuSIC 1.0 Datasets)

1. **Lundberg Lab - Human Protein Atlas Datasets**

*HPA Image subset*

Dataset: Human\_Protein\_Atlas\_images

Author: Lundberg Lab @ KTH-Royal Institute of Technology and Stanford

Date: [must be 2020.03.06 or earlier, please specify]

Version: [HPA 19.3 or earlier, please specify]

Description: 1,451 iImmunofluorescence microscopy images for 661 proteins

License: CC-BY-SA 3.0 International “The Human Protein Atlas is licensed under the Creative Commons Attribution-ShareAlike 3.0 International License for all copyrightable parts of our database, specifically indicated in the downloadable XML format with 'source="HPA".”

More Info: Thul, PJ, et al. (2017) A subcellular map of the human proteome, Science 356(6340): eaal3321, https://doi.org/10.1126/science.aal3321

Format: .png

Schema: image

Generated By: HPA (<https://www.proteinatlas.org>) query on HEK293 cell data

Used By: densenet predict

1. **Music 1.0 Computations**

***5.1 STEP 1 - Generate gold-standard protein-protein proximity values*** (<https://bit.ly/3iqcnaG>)

Computation: *compute standard proximities*

Run by: Qin Y

Date: TBD

Description: Calculate standard proximities for protein pairs based on number of discrete proteins in smallest GO cellular compartment (=C) and formula derived from regression on a “gold-standard” association of ten compartment diameters (=D) with C. See (<https://github.com/idekerlab/MuSIC/blob/master/Figures/GitHub_calibration.png>) for the calibration function. Compartment diameters input from calibration file (calibration.txt).

This is STEP 1 in the Step By Step Guide (<https://github.com/idekerlab/MuSIC/wiki/A-Step-By-Step-Guide-to-Building-a-MuSIC-Map>)

Used software: calibrate\_pariwise\_distance.py (<https://github.com/idekerlab/MuSIC/blob/master/calibrate_pairwise_distance.py>)

Used datasets:

calibrated protein-protein proximity training labels (<https://github.com/idekerlab/MuSIC/blob/master/data/calibration.txt>)

Generated datasets:

calibrated protein-protein proximity training labels (test.calibrated\_distance.csv, URL not given)

***5.2 STEP 2 - Build random forest to predict protein-protein proximity from data embeddings*** (<https://bit.ly/3ONRQZu>)

***Step 2.1 - Create the input and output files used to train the random forests***

Computation: *create labeled training & test sets*

Run by: Qin, Y.

Date: TBD

Description: Create labeled random forest regressor training and test sets.

Used software: random\_forest\_samples.py (<https://github.com/idekerlab/MuSIC/blob/master/random_forest_samples.py>)

Used datasets:

List of MuSIC proteins (<https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_proteins.txt>)

GO proteins X compartment ([https://raw.githubusercontent.com/idekerlab/MuSIC/master/data/GO\_CC\_human\_no\_hpa.tx](https://raw.githubusercontent.com/idekerlab/MuSIC/master/data/GO_CC_human_no_hpa.txt)t)

GO annotated proteins (<https://github.com/idekerlab/MuSIC/blob/master/data/GO_annotated_proteins.txt>)

image embeddings (<https://github.com/idekerlab/MuSIC/blob/master/Examples/IF_image_embedding.csv> )

AMPS embeddings (<https://github.com/idekerlab/MuSIC/blob/master/Examples/toy/toy_APMS_embedding.csv> )

Generated datasets:

IF & AMPS training sets ()

IF & APMS test sets ()

ContentURL: TBD

**5.3 Step 2.2 - Train random forest regressors**

Computation: *train and run RF models*

Run by: Qin, Y.

Date: TBD

Description: Create labeled random forest regressor training and test sets.

Used software: run\_random\_forest.py: <https://github.com/idekerlab/MuSIC/blob/master/run_random_forest.py>

Used datasets:

IF & AMPS training sets ()

IF & APMS test sets ()

Generated datasets:

RF models (Examples/output/test\_trained\_models/\*)

model-predicted protein proximities

Fold 1 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_1.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_1.pkl

Fold 2 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_2.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_2.pkl

Fold 3 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_3.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_3.pkl

Fold 4 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_4.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_4.pkl

Fold 5 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_5.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_5.pkl

ContentURL: TBD

**5.4 Step 2.3 - Average predicted proximities**

Computation: *average predicted protein proximities*

Run by: Qin, Y.

Date: TBD

Description: Average the predicted proximities

Used software: random\_forest\_output (<https://github.com/idekerlab/MuSIC/blob/master/random_forest_output.py>)

Used datasets:

predicted protein proximities:

Fold 1 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_1.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_1.pkl

Fold 2 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_2.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_2.pkl

Fold 3 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_3.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_3.pkl

Fold 4 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_4.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_4.pkl

Fold 5 proximities:

IF\_emd\_1\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_5.pkl

IF\_emd\_2\_APMS\_emd\_1.RF\_maxDep\_30\_nEst\_1000.fold\_5.pkl

Generated datasets:

averages of predicted protein proximities (<https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_predicted_proximity.txt>)

ContentURL: TBD

**5.5 Step 3: Analyze proximity data to identify protein communities at progressive resolutions** (<https://bit.ly/3OQL0Tc>)

Computation: *multi-resolution community detection*

Run by: Qin, Y.

Date: TBD

Description: Run CLiXO community detection at various stringencies to predict communities. See diagram at (<https://bit.ly/3ueaBMg>).

Used software: community detection (<https://github.com/idekerlab/MuSIC/blob/master/community_detection.py> )

Used datasets: averages of predicted protein proximities

(<https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_predicted_proximity.txt>)

Generated datasets:

System hierarchical relationships (/Examples/output/test.louvain.ddot)

Specific protein-system assignments (/Examples/output/louvain.termStats

ContentURL: TBD

Computation: *CLiXO*

Run by: Qin, Y.

Date: TBD

Description: Run CLiXO community detection at single stringency to predict a set of communities.

Associated publication: Singhal A, et al. Multiscale community detection in Cytoscape. PLoS Comput Biol. 2020 Oct 23;16(10):e1008239. <https://doi.org/10.1371/journal.pcbi.1008239>.

Associated publication: M. Kramer et al. Inferring gene ontologies from pairwise similarity data. Bioinformatics 30: i34-i42 (2014). <https://doi.org/10.1093/bioinformatics/btu282>.

Used software: CLiXO (<https://github.com/idekerlab/MuSIC/blob/master/CliXO_MuSIC.zip>)

Called by: *multi-resolution community detection*

Used datasets: averages of predicted protein proximities

(<https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_predicted_proximity.txt>)

Generated datasets:

System hierarchical relationships (/Examples/output/test.louvain.ddot)

Specific protein-system assignments (/Examples/output/louvain.termStats)

ContentURL: TBD

1. **Datasets**
   1. *List of MUSIC proteins*

Dataset: MuSIC\_proteins.txt

Author: Qin et al.

Date: Apr 23, 2021

Version: 1.0

Description: Proteins with data from BioPlex 2.0 and Human Protein Atals for HEK 293 cells in both databases. This is the set of proteins analyzed in MuSIC 1.0, consisting of 661 proteins. All proteins are indicated by their corresponding human gene names.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV no header

Schema: protein str

Derived from: BioPlex 2.0 & Human Protein Atlas

Used by: compute standard proximities calibrate\_pairwise\_distance.py

create training & test sets random\_forest\_samples.py

ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_proteins.txt>

* 1. *Calibration file*

Dataset: calibration.txt

Author: Qin et al.

Date: Apr 25, 2021

Description: Calibration file created by the author, relating benchmark cellular compartment diameters in nm to computed number of distinct protein subunits in the compartment, according to GO. The benchmark compartments in this figure https://idekerlab.ucsd.edu/music/ are:

cell, nucleolus, synapse, lysosome, centrosome, centriole, nuclear pore complex, large ribosomal subunit, proteasome, dynein complex. Sizes are stated but not referenced. The file is used by calibrate\_pairwise\_distance.py to compute “standard protein proximities” of any two proteins given the number of distinct proteins in the smallest fewest distinct proteins cellular compartment shown in GO.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV

Schema: C, D; where

C is a the number of distinct proteins in the compartment per GO

D is compartment size in nm

Used By: compute standard proximities calibrate\_pairwise\_distance.py

Generated By: Human annotation based on literature search.

Example: the complete calibration file has ten rows, as follows -

C D

18395 13000

503 2750

1153 1240

665 520

132 220

444 200

89 120

65 23

23 15

20 14

ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/data/calibration.txt>

* 1. *GO proteins X compartment*

Dataset: GO\_cc\_human\_no\_hpa.txt

Author: Qin et al.

Date: Apr 25, 2021

Version: 1.0

Description: For each GO cellular compartment, a list of the gene symbols for human proteins in that compartment. Omit HPA.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV no header

Schema: GO\_ID str, Num\_proteins int, Proteins comma separated array of str

Derived from: GO download Sept 25, 2018 [see <http://release.geneontology.org/2018-11-15/index.html>]

Used by: compute standard proximities calibrate\_pairwise\_distance.py

create training & test sets random\_forest\_samples.py

ContentURL: <https://raw.githubusercontent.com/idekerlab/MuSIC/master/data/GO_CC_human_no_hpa.txt>

* 1. *GO Annotated Proteins*

Dataset: GO\_annotated\_proteins.txt

Author: Qin et al.

Date: Apr 25, 2021

Version: 1.0

Description: List of 18395 gene symbols representing proteins with annotation in GO.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV no header

Schema: protein str

Derived from: UniProt & GO

Used by: compute standard proximities calibrate\_pairwise\_distance.py

create training & test sets random\_forest\_samples.py

ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/data/GO_annotated_proteins.txt>

* 1. *Calibrated protein-protein proximity training labels*

Dataset: test.calibrated\_distance.csv

Author: Qin et al.

Date: TBD

Version: 1.0

Description: Protein-protein proximity labels for each protein pair generated by applying the formula log10(D) = 1.05 \* log10(C) -0.14 to the number of distinct proteins of the smallest Gene Ontology component shared by the gene pair, from a precomputed matrix derived from GO.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: CSV

Schema: geneA, geneB, C, D, log10D, P

where

geneA := human gene name corresponding to a protein,

geneB := human gene name corresponding to a protein,

C := # of proteins in smallest GO cellular component to which both are annotated,

D := protein-protein distance calibrated from GO see calibration file,

log10D := log10D, and

P := protein-protein proximity = -log10D

Generated by: compute standard proximities (calibrate\_pairwise\_distance.py)

Used by: create labeled training and test sets (random\_forest\_samples.py)

ContentURL: TBD

* 1. *Image embeddings*

Dataset: IF\_image\_embedding.csv

Author: Lundberg lab (<https://cellprofiling.org/people_members.html>)

Date: Apr 23, 2021

Version: 1.0

Description: Immunoflorescence (IF) image embeddings for each protein in the study, as computed by Densenet prediction. IF embeddings are paired per protein, e.g.

IF\_1,HLA-DPA1,0.022881096228957176,-0.060293108224868774, … and

IF\_2,HLA-DPA1,0.022881096228957176,-0.11374042928218842, … etc.,

while APMS embeddings are for a single protein.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: CSV

Schema: IF\_ID str, “IF\_1, IF\_2, …”, protein, embedding array of float X 1024

Generated By: Densenet predict (predict\_d121.py)

Used By: create training & test sets random\_forest\_samples.py

Comment: ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/Examples/IF_image_embedding.csv>

* 1. *AP-MS embeddings*

Dataset: APMS\_embedding.MuSIC.csv

Author: Gygi lab (<https://gygi.hms.harvard.edu/team.html>)

Date: Apr 23, 2021

Version: 1.0

Description: Affinity purification mass spectrometer (APMS) embeddings for each protein in the study, generated by node2vec predict.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: CSV

Schema: APMS\_ID str, “APMS\_1, APMS\_2, …”, protein, embedding array of float X 1024

Derived from: node2vec predict

Used By: create labeled training & test sets random\_forest\_samples.py

Comment: APMS embeddings are one per protein

ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/Examples/APMS_embedding.MuSIC.csv>

* 1. *IF & APMS training sets*

Dataset: TBD

Author: TBD

Date: TBD

Version: TBD

Description: IF and APMS embedding random forest regressor training sets generated by *random\_forest\_samples.py*.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TBD

Schema: TBD

Generated By: create training & test sets (random\_forest\_samples.py)

Used By: train & run RF models run\_random\_forest.py

Comments: check outputs from toy example for format & schema; multiple files

ContentURL: TBD

* 1. *IF & APMS test sets*

Dataset: TBD

Author: TBD

Date: TBD

Version: TBD

Description: IF and APMS embedding random forest regressor training sets generated by *random\_forest\_samples.py*.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TBD

Schema: TBD

Generated By: create training & test sets random\_forest\_samples.py

Used By: train & run RF models run\_random\_forest.py

Comments: check outputs from toy example for format & schema

ContentURL: TBD

* 1. *model-predicted protein proximities*

Dataset: predicted protein proximities

IF\_emd\_[i]\_APMS\_emd\_[j].RF\_maxDep\_[depth]\_nEst\_[estimators].fold\_[fold].pkl

where

i = 1,2; j = 1; depth=30; estimators=1000; fold = 1,2,3,4,5

for the Example files.

Author: Qin et al.

Date: TBD

Version: 1.0

Description: Dataset of ten .pkl files with RF-predicted protein-protein proximities.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: PKL

Schema: TBD

Generated By: train & run RF models run\_random\_forest.py

Used By: avg predicted protein proximities random\_forest\_output.py

ContentURL: TBD

* 1. *averages of predicted protein proximities*

Dataset: text\_predicted\_proximity.txt

Author: Qin et al

Date: Jun 10, 2021

Version: 1.0

Description: Dataset of averaged with RF-predicted protein-protein proximities computed as average of model-predicted proximities.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV

Schema: proteinA str, proteinB str, avg\_predicted\_proximity float(4), ytrue

Generated by: avg predicted protein proximities (random\_forest\_output.py)

Used by: community detection community\_detection.py

Example:

MARCKSL1 RANBP17 0.1465 0.0600

MARCKSL1 FAM160A2 0.1381 0.0600

MARCKSL1 MAP2K7 0.1566 0.0600

MARCKSL1 TMLHE 0.1318 0.0600

ContentURL: <https://github.com/idekerlab/MuSIC/blob/master/Examples/MuSIC_predicted_proximity.txt>

* 1. *precomputed protein min GO size matrix*

Dataset: all\_protein\_min\_GO-size.txt

Author: Qin et al

Date: 4/25/2021

Version: 1.0

Description: Matrix showing # of distinct proteins in the GO cellular compartment with smallest # distinct proteins for each protein pair.

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: TSV

Schema: 18395 X 18395 matrix of minimum protein-protein GO cellular compartment sizes

Derived from: GO download Sept 25 2018 → GO\_cc\_human\_no\_HPA.txt

Used by: community detection community\_detection.py

Example: [tabs inserted for readability]

Line 1: ,GJA1, A0A1C7CYX0, COX11, DDR1, PRMT8, …

Line 2: GJA1, 18395.0, 5341.0, 670.0, 1600.0, 1679.0, …

Line 3 A0A1C7CYX0, 5341.0, 18395.0, 5341.0, 5341.0, 5390.0, …

ContentURL: [https://www.dropbox.com/s/w9lxpnw39g64zs8/all\_protein\_min\_GO\_size.txt](https://www.dropbox.com/s/w9lxpnw39g64zs8/all_protein_min_GO_size.txt?dl=0)

* 1. *System hierarchical relationships*

Dataset: /Examples/output/test.louvain.ddot

Author: Qin et al.

Date: TBD

Version: 1.0

Description: File showing the system hierarchical relationships as determined by community\_detection.py

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: ddot

Schema: TBD

Generated by: community detection community\_detection.py

Used by: Hiview visualization

More info: Yu MK et al. 2019 “DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies”. Cell Syst. 2019 Mar 27;83:267-273.e3. doi: 10.1016/j.cels.2019.02.003

ContentURL: TBD

* 1. *Specific protein-system assignments*

Dataset: /Examples/output/louvain.termStats

Author: Qin et al.

Date: TBD

Version: 1.0

Description: File showing the putative protein-system assignments as determined by community\_detection.py

Associated publication: Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 2021

Additional documentation: https://idekerlab.ucsd.edu/music/

Format: ddot

Schema: TBD

Generated by: community detection community\_detection.py

Used by: TBD

More Info: Yu MK et al. 2019 “DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies”. Cell Syst. 2019 Mar 27;83:267-273.e3. doi: 10.1016/j.cels.2019.02.00

ContentURL: TBD

1. **Validation Rules**
   1. ***Common Metadata***
      1. Primary Identifier

Primary identifiers in CM4AI are ARK-form persistent IDs. Valid ARKs are defined by (Kunze and Bermès 2022) and are structured as shown in Figure 1, with implementation-specific details for ARKs used in CM4AI supplied by a configuration file.

1. NMA: resolving host name, e.g. “<https://example.org>”
2. label: “ark:”
3. NAAN: name assigning authority number, e.g. “99999”, followed by “/”
4. shoulder: the organization and project labels, = “B2AI/CM4AI/”
5. pre-qualifier: the schema name, e.g. “music\_proteins”
6. qualifier: a valid UUID

According to Bioregistries.org, the local ID portion of an ark should match this regex:

^/\*[0-9A-Za-z]+(?:/[\w/.=\*+@\$-]\*)?(?:\?.\*)?$

* + 1. Object Name - string up to 64 characters
    2. Object Type: Software / Dataset / Computation
    3. Author - string of up to 64 characters
    4. Description - string of up to 2056 characters
    5. Date - valid date in YMD format
    6. Version - string
    7. Associated publication - string of up to 2056 characters;
    8. Format - string, a valid registered format name, e.g. “CSV”, “TSV”, “XML”, “JPEG”, “OME”, etc.
    9. Schema Name: - string, corresponding to a valid registered file type name (lookup), e.g. “calibration.txt”
    10. Schema Description: See ***Dataset Schemas*** below.
    11. Generated By: (a) Object type in 3. above must be “Dataset”; (b) Object Name for GeneratedBy must be valid registered Computation.
    12. Used By: (a) Object type in 3. above must be “Dataset”; (b) Object name for GeneratedBy must be valid registered Computation.
    13. Used Software: (a) Object type in 3. above must be “Computation: (b) Object name for UsedSoftware must be valid registered Software.
    14. Used Dataset: (a) Object type in 3. above must be “Computation: (b) Object name for UsedSoftware must be valid registered Dataset.

Base Compact Name Qualifiers

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_

/ \/ \

https://example.org/ark:12345/x6np1wh8k/c3/s5.v7.xsl

\\_\_\_\_\_\_\_\_\_/ \\_\_/\\_\_\_/\\_/\\_\_\_\_\_/\\_\_\_\_/\\_\_\_\_\_/

NMA Label NAAN | Blade Parts Variants

Shoulder

\\_\_\_\_\_\_\_\_\_\_\_\_\_/

Check Zone

**Figure 1a. ARK Anatomy. Adapted from Kunze 2022.**

Base (NMA): [resolver hostname, e.g. “https://cm4ai.org”]

Label: ark:

NAAN: 59853 [University of Virginia] /

Organization /

Project /

(optional) Group /

Schema.version /

UUID-fragment (final 17 bytes).

Generic: NMA/ark:NAAN/Organization/Project/Group (optional)/Schema.version/UUID-fragment

Example 1: [https://example.org/ark:59853/B2AI/CM4AI/GOproteins.1/8540-897e9086f162.txt](https://id.cm4ai.org/ark:59853/B2AI/CM4AI/GOproteins.1.0/8540-897e9086f162.txt)

Example 2: <https://example.org/ark:59853/B2AI/CM4AI/IF_image_embedding.1/8594-a041c6ea612a.tsv>

Example 3: [https://example.org/ark:59853/B2AI/CM4AI/PROX/protein\_proximity.1/8af7-8358fb4b46a7.pkl](https://example.org/ark:59853/B2AI/CM4AI/proximities/protein_proximity.1/8af7-8358fb4b46a7.pkl)

**Figure 1b. CM4AI-specific ARK anatomy with examples**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **name** | **datatype** | **standard\_name** | **vocabulary** | **definition** |
| GeneA | int | -TBD- | none | -TBD- |
| GeneB | int | -TBD- | none | -TBD- |
| UniProtA | str | bait\_protein | UniProt:  accession | The Uniprot accession number of the specific protein used as bait in an AP-MS experiment. |
| UniProtB | str | prey\_protein | UniProt:  accession | The Uniprot accession number of the specific protein used as bait in an AP-MS experiment. |
| SymbolA | str | bait\_protein\_gene | UniProt:  gene name | The UniProt gene name corresponding to bait\_protein entry in UniProt. |
| SymbolB | str | prey\_protein\_gene | UniProt:  gene name | The UniProt gene name corresponding to prey\_protein entry in UniProt. |
| prob(Wrong) | float | wrong\_call\_probability | none | -TBD- |
| prob (No Interaction) | float | no\_interaction\_probability | none | The probability that the interaction does not occur. |
| prob (Interaction) | float | interaction\_probability | none | The probability that the interaction does occur. |

**Figure 2. Table of Data Elements Example. Where CSV/TSV headers are missing, *name* may be blank. All other columns are mandatory.**

***7.2 Dataset Schemas***

* 1. Valid schema names are defined in a list with corresponding schema descriptions
  2. Valid schema descriptions for TSV and CSV files are given in a table of data elements, as in the Figure 2 example.
  3. Valid schema descriptions for XML files are supplied is an XSD file.
  4. Schema descriptions for image files (JPEG, OME, etc. ) are omitted.

**REFERENCES**

Erickson, H. P. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol Proced Online* **11**, 32–51 (2009).

Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock A, Zarraga G, Tam S, Szpyt J, Gassaway BM, Panov A, Parzen H, Fu S, Golbazi A, Maenpaa E, Stricker K, Guha Thakurta S, Zhang T, Rad R, Pan J, Nusinow DP, Paulo JA, Schweppe DK, Pontano Vaites L, Harper JW, Gygi SP. (2021) Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. Cell 184(11):3022-3040.e28. doi:10.1016/j.cell.2021.04.011

Kunze, J. and Bermès, E. The ARK Identifier Scheme. Internet Draft. Internet Engineering Task Force (2022). <https://datatracker.ietf.org/doc/html/draft-kunze-ark-36>

Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, Zhang C, Le T, Johansson F, Schutten R, Bäckström A, Axelsson U, Thul P, Cho NH, Carja O, Uhlén M, Mardinoglu A, Stadler C, Lindskog C, Ayoglu B, Leonetti MD, Pontén F, Sullivan DP, Lundberg E. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature. 2021 Feb;590(7847):649-654. doi: 10.1038/s41586-021-03232-9. Epub 2021 Feb 24. Erratum in: Nature. 2022 Aug;608(7924):E32. PMID: 33627808.

Moxon, S, Solbrig, H, Unni, D, Jiao, D, Bruskiewich, R, Balhoff, J, Vaidya, G, Duncan, W, Hegde, H, Miller, M, Brush, M, Harris, N, Haendel, M & Mungall, C 2021, 'The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics', CEUR Workshop Proceedings, vol. 3073, pp. 148-151.

National Institutes of Health. Bridge to Artificial Intelligence (Bridge2AI). Accessed Dec 31, 2022. https://commonfund.nih.gov/bridge2ai

Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).

Singhal, A. *et al.* Multiscale community detection in Cytoscape. *PLoS Comput Biol* **16**, e1008239 (2020).

Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, Bäckström A, Danielsson F, Fagerberg L, Fall J, Gatto L, Gnann C, Hober S, Hjelmare M, Johansson F, Lee S, Lindskog C, Mulder J, Mulvey CM, Nilsson P, Oksvold P, Rockberg J, Schutten R, Schwenk JM, Sivertsson Å, Sjöstedt E, Skogs M, Stadler C, Sullivan DP, Tegel H, Winsnes C, Zhang C, Zwahlen M, Mardinoglu A, Pontén F, von Feilitzen K, Lilley KS, Uhlén M, Lundberg E. A subcellular map of the human proteome. Science. 2017 May 26;356(6340):eaal3321. doi: 10.1126/science.aal3321. Epub 2017 May 11. PMID: 28495876.