

# Deep Learning Applied to Genomics, Deep Semantic Protein Representation

Ariel Schwartz, Director of Computational Biology, Synthetic Genomics, Inc.

Zach Dwiell, Senior Data Scientist, Intel AI Lab

AIDC 2018



# Safe Harbor Statement and Copyright

This presentation may contain “forward-looking statements”. Forward-looking statements relate to expectations, beliefs, projections, future plans, anticipated events, and similar statements and matters that are not historical facts. The viewer is cautioned not to rely on these forward-looking statements. If underlying assumptions prove inaccurate or known or unknown risks or uncertainties materialize, actual results could vary materially from the expectations herein. Risks and uncertainties include, but are not limited to, economic factors, competition, challenges and uncertainties inherent in product research and development, uncertainty of commercial success for new and existing products, the ability of the company to successfully execute strategic plans, the impact of business combinations and divestitures, the impact of patent expirations, market conditions, significant adverse litigation or government action, changes to applicable laws and regulations, and the potential failure to meet obligations in compliance agreements with government bodies.

Any forward-looking statement made in this presentation speaks only as of the date of the presentation. Synthetic Genomics does not update any forward-looking statements as a result of new information or future events or developments.

## **Copyright © 2018 Synthetic Genomics, Inc. All rights reserved.**

This document is the property of Synthetic Genomics, Inc. The copying or disclosure in part or whole of this document without the prior written approval of Synthetic Genomics, Inc. is strictly prohibited. This document is not intended to, and does not, constitute an offer or commitment to enter into a transaction or any other legally binding obligation. Use of this document or any information or data it contains may only occur under the terms of a written secrecy agreement with, or as authorized by, Synthetic Genomics, Inc.

This document is for discussion purposes only, and all information, statements, data, projections, forecasts, predictions, and estimates contained herein are provided only for the purpose of facilitating discussions. No representations or warranties of any kind are made in this document or are made about the accuracy or completeness of this document and its contents, and all warranties, express or implied, are excluded. Synthetic Genomics, Inc. and its affiliates accept no liability for losses caused by reliance on or any use of this information. Each party is responsible for its own evaluation of the proposals in this document, including the conduct of its own financial analysis, financial modeling, legal review, and review of the information, statements, data, projections, forecasts, predictions, and estimates contained herein.

# D-SPACE: Deep Sematic Protein Annotation Classification and Exploration



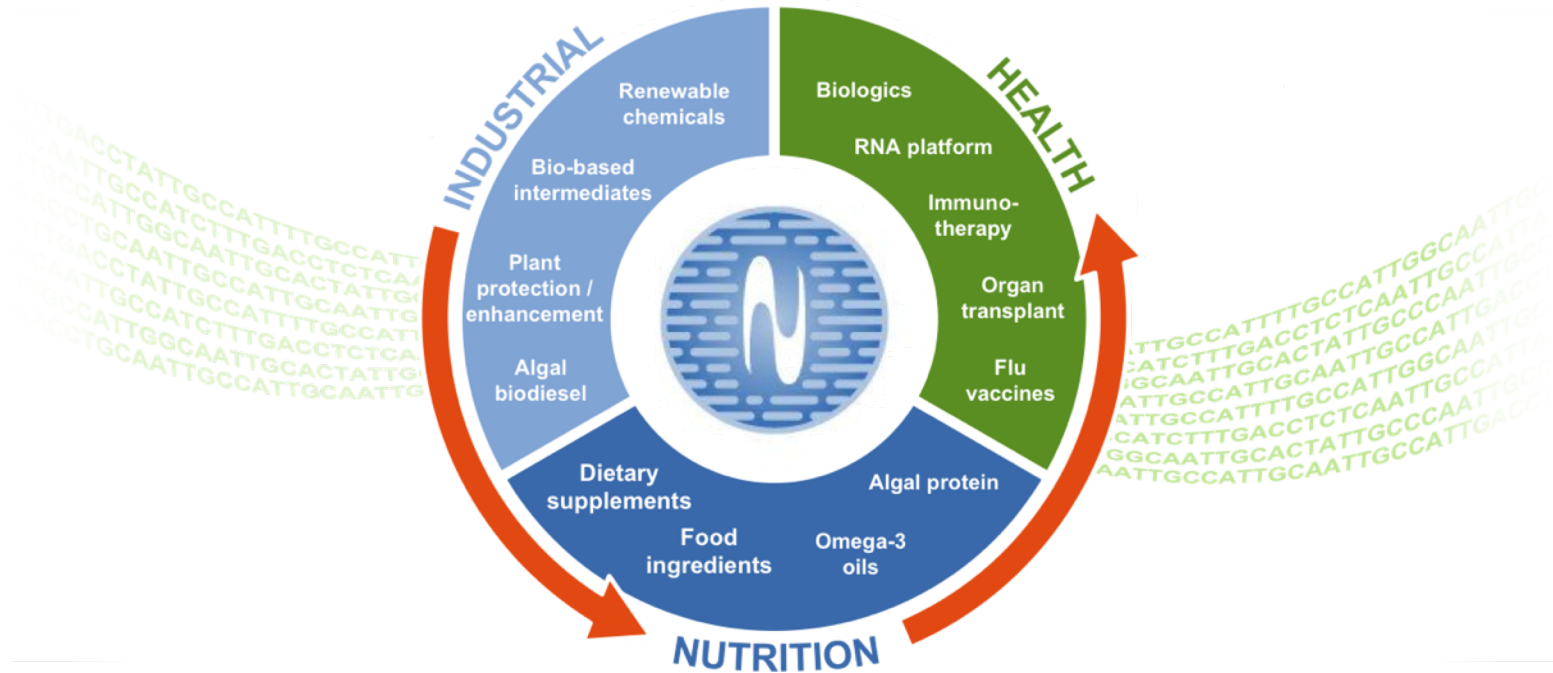
SYNTHETIC GENOMICS®



SYNTHETIC GENOMICS®

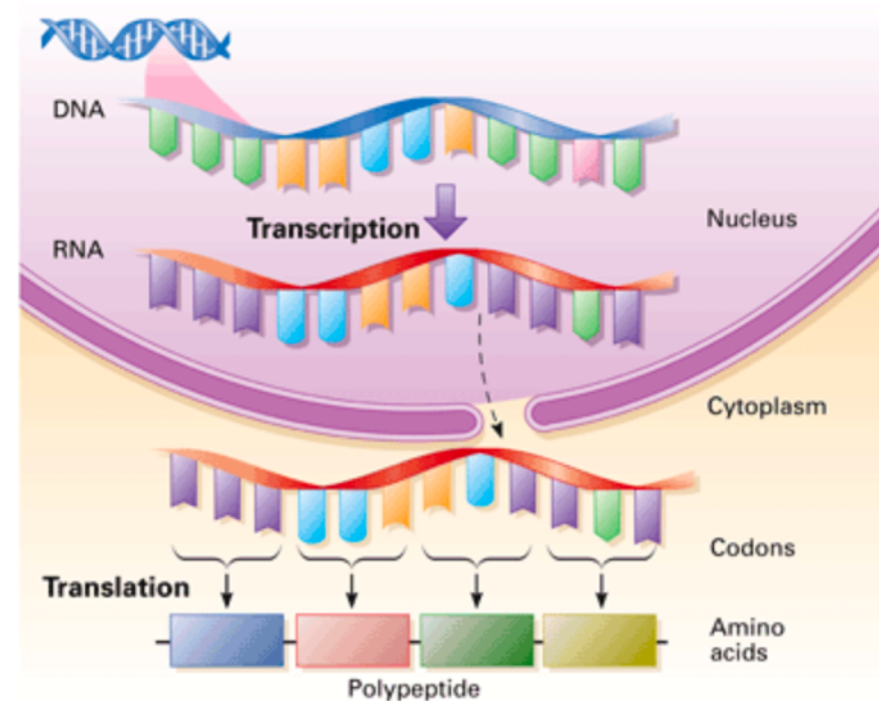
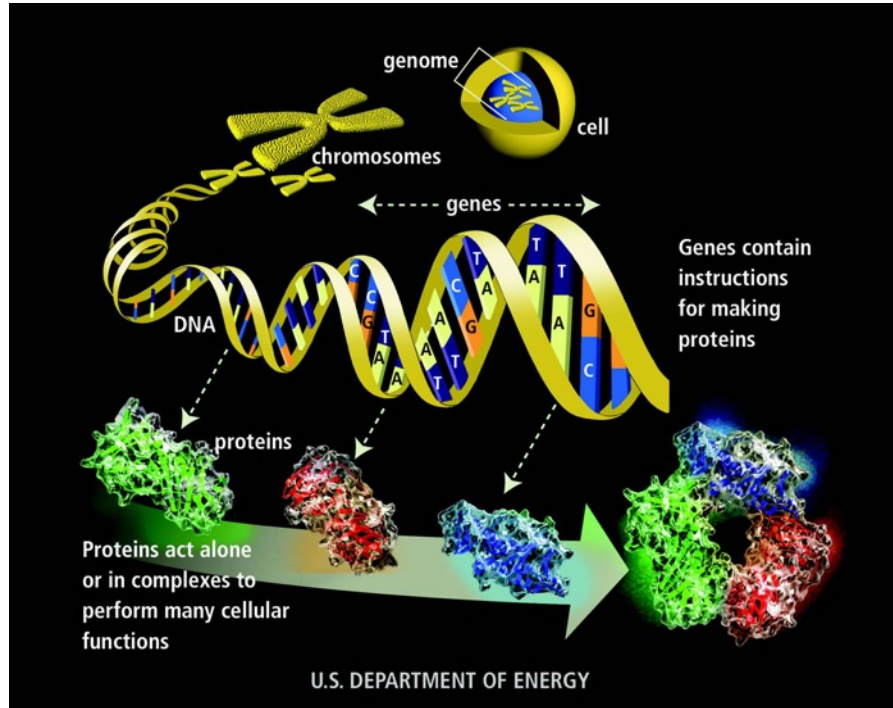


# Writing genomes to address sustainability challenges from industrial to health

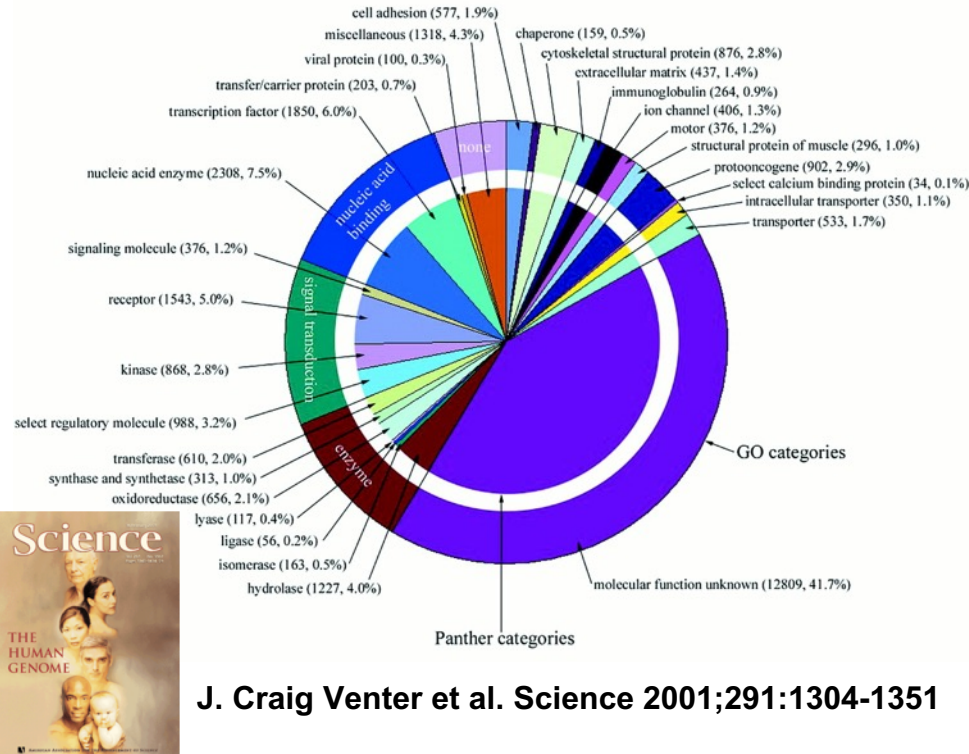




# The Central Dogma of Molecular Biology



# Many proteins have unknown molecular functions



J. Craig Venter et al. Science 2001;291:1304-1351

## Design and synthesis of a minimal bacterial genome

Clyde A. Hutchison III,<sup>1\*†</sup> Ray-Yuan Chuang,<sup>1†‡</sup> Vladimir N. Noskov,<sup>1</sup>  
 Nacyra Assad-Garcia,<sup>1</sup> Thomas J. Deerinck,<sup>2</sup> Mark H. Ellisman,<sup>2</sup> John Gill,<sup>3</sup>  
 Krishna Kannan,<sup>3</sup> Bogumil J. Karas,<sup>1</sup> Li Ma,<sup>1</sup> James F. Pelletier,<sup>4§</sup> Zhi-Qing Qi,<sup>3</sup>  
 R. Alexander Richter,<sup>1</sup> Elizabeth A. Strychalski,<sup>4</sup> Lijie Sun,<sup>1||</sup> Yo Suzuki,<sup>1</sup>  
 Billyana Tsvetanova,<sup>3</sup> Kim S. Wise,<sup>1</sup> Hamilton O. Smith,<sup>1,3</sup> John I. Glass,<sup>1</sup>  
 Chuck Merryman,<sup>1</sup> Daniel G. Gibson,<sup>1,3</sup> J. Craig Venter<sup>1,3\*</sup>

We used whole-genome design and complete chemical synthesis to minimize the 1079-kilobase pair synthetic genome of *Mycoplasma mycoides* JCVI-syn1.0. An initial design, based on collective knowledge of molecular biology combined with limited transposon mutagenesis data, failed to produce a viable cell. Improved transposon mutagenesis methods revealed a class of quasi-essential genes that are needed for robust growth, explaining the failure of our initial design. Three cycles of design, synthesis, and testing, with retention of quasi-essential genes, produced JCVI-syn3.0 (531 kilobase pairs, 473 genes), which has a genome smaller than that of any autonomously replicating cell found in nature. JCVI-syn3.0 retains almost all genes involved in the synthesis and processing of macromolecules. **Unexpectedly, it also contains 149 genes with unknown biological functions.** JCVI-syn3.0 is a versatile platform for investigating the core functions of life and for exploring whole-genome design.

Clyde A. Hutchison III et al. Science 2016;351:aad6253

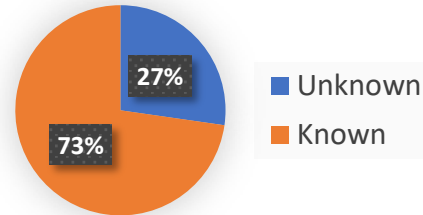
# The Problem

*BLAST: “the google search engine of biology”*

The screenshot shows the BLAST website interface. At the top, it says "Basic Local Alignment Search Tool". Below this, a paragraph explains that BLAST finds regions of similarity between biological sequences. To the right, there is a "NEWS" section with a headline "BLAST+ 2.8.0-alpha released" and a sub-headline "BLAST+ now has a better database." Below the news section, there is a "Web BLAST" section with three main options: "Nucleotide BLAST" (nucleotide to nucleotide), "blastx" (translated nucleotide to protein), and "tblastn" (protein to translated nucleotide). To the right of these is "Protein BLAST" (protein to protein). At the bottom, there is a "BLAST Genomes" section with a search bar and a "Search" button. Below the search bar, there are links for "Human", "Mouse", "Rat", and "Microbes".

BLAST is so pervasive it has become both noun and verb:  
"I BLASTed my sequence"

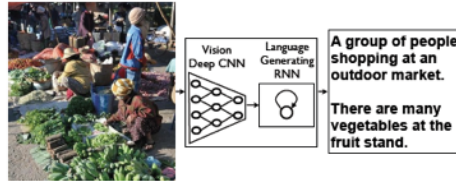
- The amount of sequencing data has outpaced the ability to analyze it
- Protein annotation and search have long been dominated by sequence homology-based methods such as BLAST and HMM
- Current solutions are based on old technology... Blast was invented in 1990!



***6,000 proteins of unknown function in the human genome!***

# The Solution

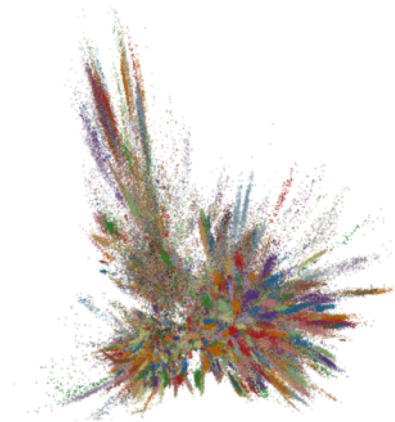
**D-SPACE:** A disruptive platform for protein annotation, discovery and design utilizing the recent advances in deep-learning technology



State of the art deep learning models can now produce high quality image descriptions

D-SPACE can do the same and more for prediction in genomics and for synthetic biology engineering

- D-SPACE utilizes a novel function-based approach to annotation and discovery
- D-SPACE leverages the power of deep learning and high-dimensional embeddings



# The D-SPACE Advantage: Super Fast and Sensitive Protein Annotation, Discovery & Design

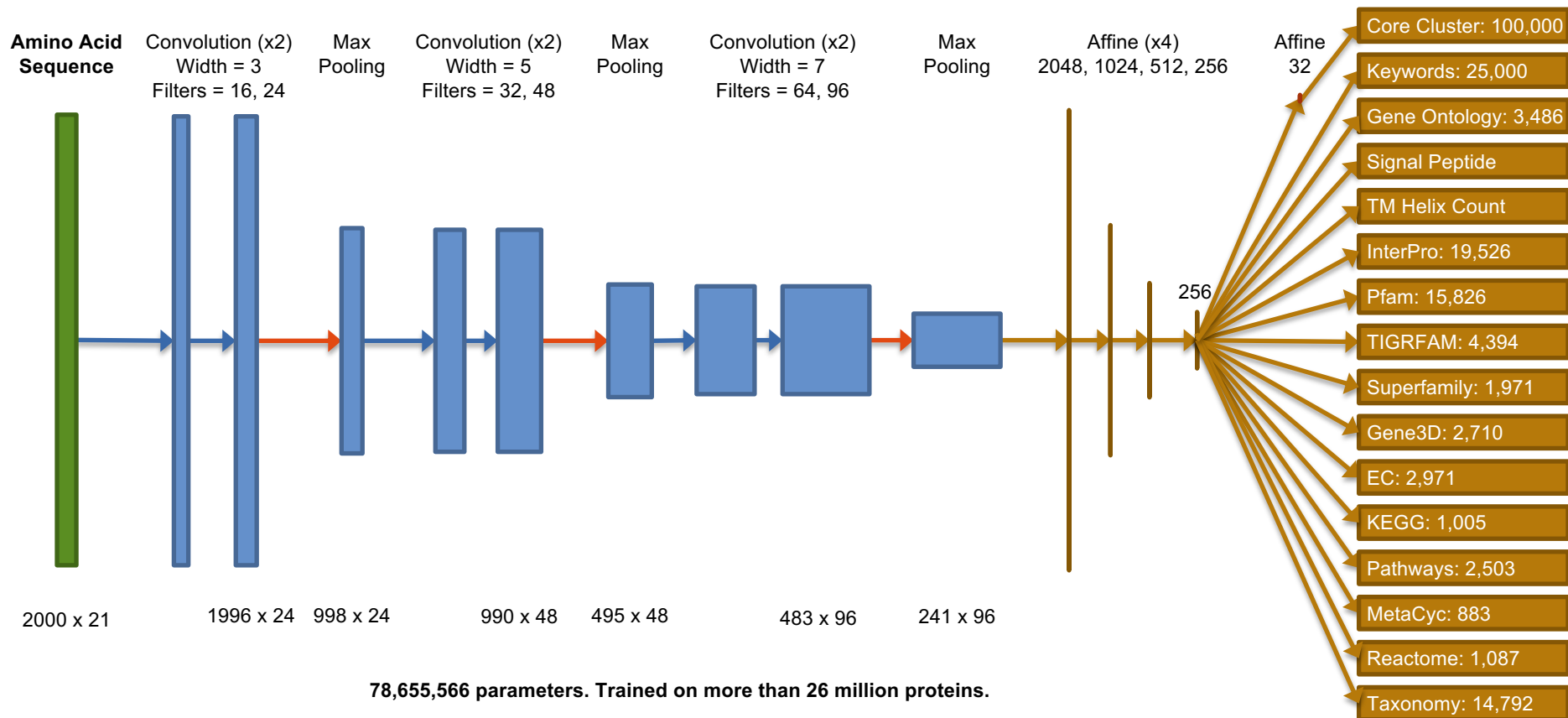
## D-SPACE Innovations

- State-of-the-art protein annotation in milliseconds
- Search based on **function** is bigger than BLAST (much faster and goes beyond sequence homology)
- Intelligent protein optimization for desired features

## D-SPACE Impact v.s. Existing Solutions

- Annotate 1 million genes in under an hour instead of running traditional tools on a 260 CPU cluster for a week
- Assign function to millions of previously uncharacterized proteins
- In a few seconds search over hundred of millions of proteins and discover novel proteins with a desired function instead of waiting minutes to hours for BLAST results only to find no novel hits
- Predict in seconds the impact of every possible AA change on protein function instead of performing expensive and non-comprehensive *in-vitro* alanine scans

# D-SPACE Model C – a CNN architecture





# D-SPACE Model C Validation Performance

Task	F1	Labels	Coverage (%)
Tigrfam	0.89	4,394	86
KEGG	0.85	1,005	94
Metacyc	0.85	883	95
Superfamily	0.84	1,971	89
Interpro	0.84	19,526	76
Gene Ontology	0.82	3,486	85
Gene3D	0.82	2,710	87
Reactome	0.82	1,087	85
EC	0.81	2971	95
Pfam	0.81	15,826	74
Pathway IDs	0.80	2,503	99

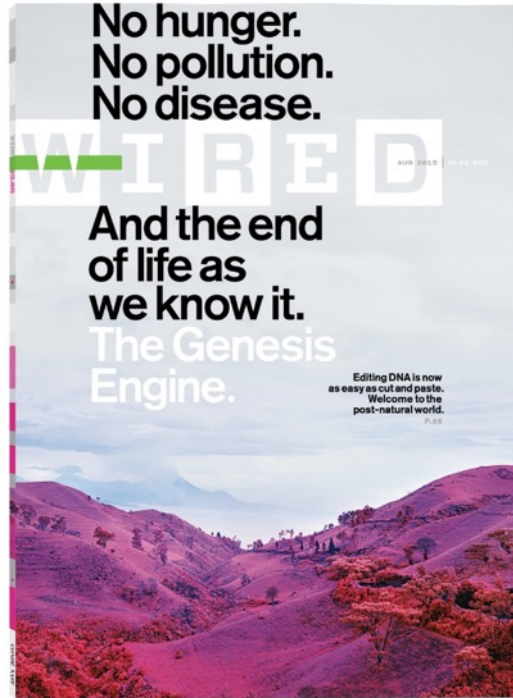
Task	Accuracy
Signal Peptide	0.95

Task	Top 5 Accuracy	Labels
Cluster	0.89	Top 100,000

Task	Mean Square Error
Transmembrane Helix	0.12

Task	F1	Labels	Coverage (%)
Keywords	0.50	Top 25,000	24
Taxonomic Assignment	0.38	14,792	32

# Protein Discovery using D-SPACE



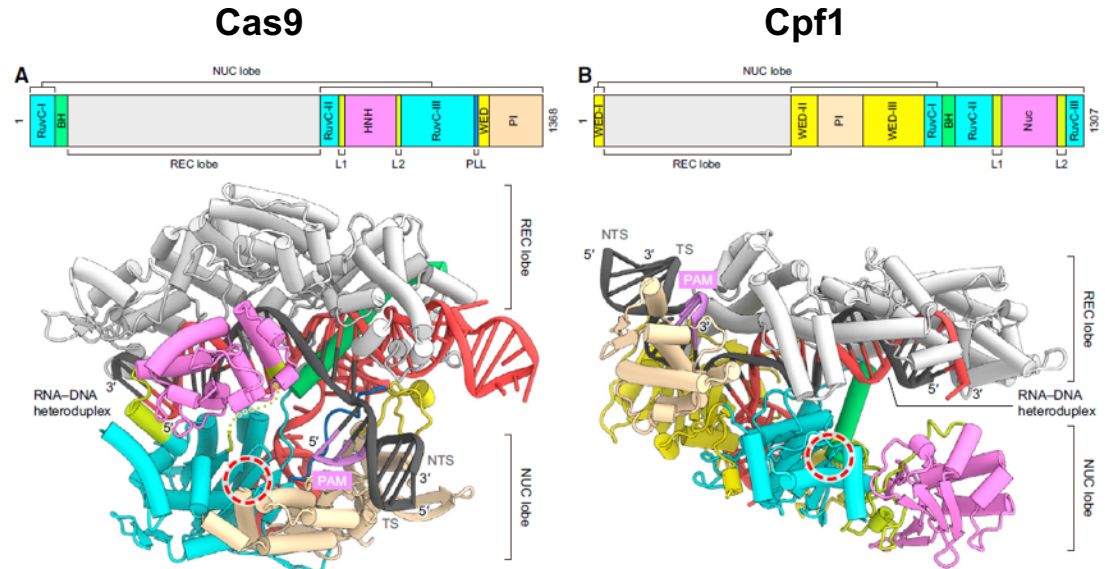
- Genome-editing is revolutionized biology
- Currently a \$3.2B market and on pace to reach \$12B by 2025
- There's a frenetic race to find novel gene-editing systems

# Protein Discovery using D-SPACE

Cas9 and Cpf1 are both Class II CRISPR effector proteins with very similar functions

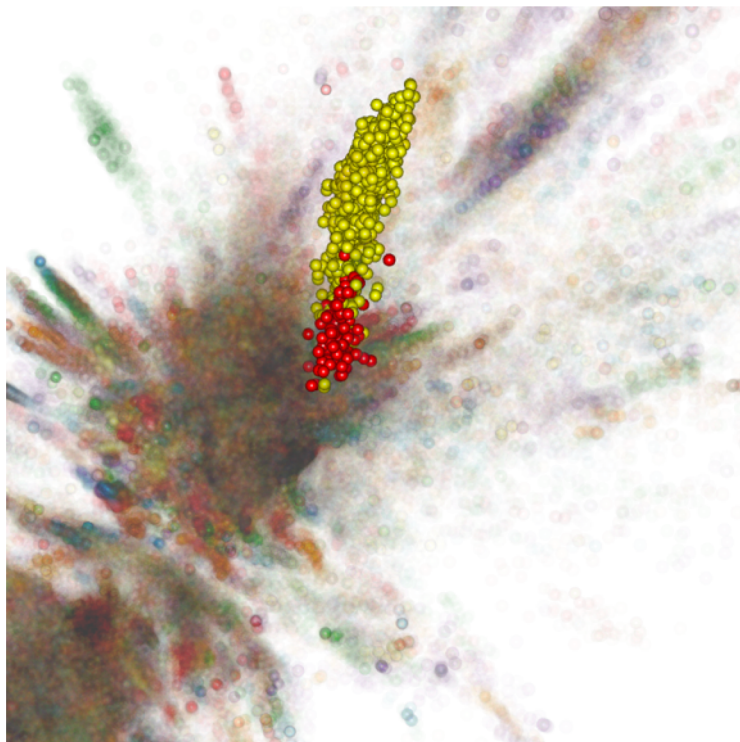
Cas9 and Cpf1 have very different domain architecture and share only ~15% amino acid identity

Cas9 and Cpf1 have very different 3D crystal structures



A BLAST search of Cas9 does not reveal Cpf1 and vice versa

# Protein Discovery using D-SPACE



While **Cas9** and **Cpf1** are very different in sequence space they are quite similar in D-SPACE's embedding space, allowing discovery of **Cpf1** using **Cas9** as a query.

*Other putative gene-editing systems  
have been identified!*

# D-SPACE demo

Learn how to use D-SPACE for:

- Annotating thousands of proteins in seconds
- Assigning function to uncharacterized proteins
- Searching for remote functional homologs in Archetype and public databases
- Performing advanced semantic profile-based searches
- Identifying critical residues
- Interpreting the effects of mutations on protein function
- And more ...

# Demo: using D-SPACE annotation & search functionality



Welcome Ariel | Logout

Discovery | Biotools | Pathways | Omics | Metadata | **Analysis Apps** | Support

D-SPACE



Deep Semantic Protein Annotation Classification and Exploration

Data

Data Manager

Dataset

Protein

Ac3H11\_1015

Tools

Protein Annotation

Protein Discovery

Protein Design

Advanced Settings

Send Feedback

Create dataset

Text Input

File Input

Dataset name

RandomExample

Submit!

I'm Feeling Lucky!

Dataset variance

Available datasets

Refresh

Show 10 entries

Search:

ID	Entries	Creation	Model	Type
All	All	All	All	All
ArielSchwartz-18_05_18_10:56:24	1	2018-05-18 10:56:33	modelC.run-20171024131729.best	Basic
ArielSchwartz-18_05_17_22:16:17	1	2018-05-17 22:16:38	modelC.run-20171024131729.best	Basic
sp_ABL3V1_RF1_FRASN_Metropolis_scan	49851	2018-05-01 09:21:37	modelC.run-20171024131729.best	Metropolis_scan
sp_ABL3V1_RF1_FRASN_Full_scan	6765	2018-05-01 06:45:16	modelC.run-20171024131729.best	Full_scan
ArielSchwartz-18_05_01_05:51:55	1	2018-05-01 05:53:04	modelC.run-20171024131729.best	Basic



# Demo: using D-SPACE advanced search

Archetype<sup>™</sup> Welcome Ariel | Logout

Discovery | Biotools | Pathways | Omics | Metadata | **Analysis Apps** | Support

D-SPACE ≡ Deep Semantic Protein Annotation Classification and Exploration

**Data**

- Data Manager
- Dataset**  
Escherichia\_coli\_ATCC\_8739
- Protein**  
P3M1CT1670720

**Tools**

- Protein Annotation
- Protein Discovery**
- Protein Design
- Advanced Settings
- Send Feedback

**D-SPACE Search**

☒ Start with my selected protein

Keywords ? +

Dataset signatures ? +

**Length range** ?

1 2,000

**Number of top hits** ?

100 2,000

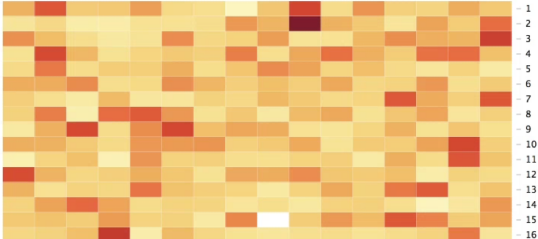
Select all ☐ **Search Database** ?

☒ UniProtKB ☐ PUBLIC ☐ SG

Search

**Advanced options** +

**Combined embedding** ?



**Discovered proteins word cloud** ?

factor  
elongation

# Demo: using D-SPACE advanced search



Welcome Ariel | Logout

Discovery

Biobase

Pathways

Omics

Metadata

**Analysis Apps**

Support

D-SPACE

Deep Semantic Protein Annotation Classification and Exploration

Data

Data Manager

Dataset

Protein

Tools

Protein Annotation

Protein Discovery

Protein Design

Advanced Settings

Send Feedback

Create dataset

Dataset variance

Available datasets

ID	Entries	Creation	Model	Type
Escherichia_coli_ATCC_8739	4197	2018-05-18 12:38:25	modelC.run-20171024131729.best	Basic
sp_Q48D34_EFTU_PSE14_Full_scan	7544	2018-05-18 11:56:07	modelC.run-20171024131729.best	Full_scan
ArielSchwartz-18_05_18_10:56:24	1	2018-05-18 10:56:33	modelC.run-20171024131729.best	Basic
ArielSchwartz-18_05_17_22:16:17	1	2018-05-17 22:16:38	modelC.run-20171024131729.best	Basic
sp_ABL3V1_RF1_FRASN_Metropolis_scan	49851	2018-05-01 09:21:37	modelC.run-20171024131729.best	Metropolis_scan

Text Input

File Input

>sp|A0Q7Q2|CS12A\_FRATN CRISPR-associated endonuclease Cas12a OS=Francisella tularensis subsp. novicida (strain U112)  
OX=401614 GN=cas12a PE=1 SV=1  
MSIQEFVNYKSLKTLRFELIPQKTLLENIKARGLILDEKRAKDYKKAKIIDKYHQF  
PIEELSSVCISEDILQNYSDVYFKLKSDDDNLQKDFKSAKDTIKKQISEYKDKSEFK  
NLFNQNLIDAKKGQESDILWLKQKDNGLFKANSDDITDIDEALIEIKSFKGWTTYK  
GFHENRKNYSSNDIPTSIYRIVDDNLKPFLENKAKYESKDKAPEAINEYKDKDLAE  
ELTFDIDYKTEVNRQVFSLDEVFEIANFNLYLNQSGITKNTIGGKFNVGENTKRRGI  
NEYINLYSQINDKTLKKYMSVLFKQILSDTESKSFVIDKLEDDSDVTTMQSFYEQIA  
AFKTVEEKSIKETLSLLFDDKAKQLDLKIFYKNDKSLTDSQQVFDYVIGTAVLEY  
ITQIAPKNLNDNPSKKEQELIAKTEKAKYLSLETIKLALFEFNKHRDIDKQCFEEILA

Dataset name

A0Q7Q2\_Cpf1

Submit!

I'm Feeling Lucky!

Dataset variance

Available datasets

Refresh

Show 10 entries

Search:

ID	Entries	Creation	Model	Type
All	All	All	All	All
Escherichia_coli_ATCC_8739	4197	2018-05-18 12:38:25	modelC.run-20171024131729.best	Basic
sp_Q48D34_EFTU_PSE14_Full_scan	7544	2018-05-18 11:56:07	modelC.run-20171024131729.best	Full_scan
ArielSchwartz-18_05_18_10:56:24	1	2018-05-18 10:56:33	modelC.run-20171024131729.best	Basic
ArielSchwartz-18_05_17_22:16:17	1	2018-05-17 22:16:38	modelC.run-20171024131729.best	Basic
sp_ABL3V1_RF1_FRASN_Metropolis_scan	49851	2018-05-01 09:21:37	modelC.run-20171024131729.best	Metropolis_scan



SYNTHETIC GENOMICS



# Demo: towards protein design with D-SPACE

The screenshot displays the D-SPACE web application interface. The top navigation bar includes the Archetype logo, a user welcome message, and links for Discovery, Biotools, Pathways, Omics, Metadata, Analysis Apps, and Support. The main header reads "D-SPACE | Deep Semantic Protein Annotation Classification and Exploration".

**Left Sidebar:**

- Data:** Data Manager
- Dataset:** recd
- Protein:** P0A7G6(RECA\_ECOLI)
- Tools:**
  - Protein Annotation
  - Protein Discovery
  - Protein Design**
  - Advanced Settings
  - Send Feedback

**Main Content Area:**

**Sequence scan options**

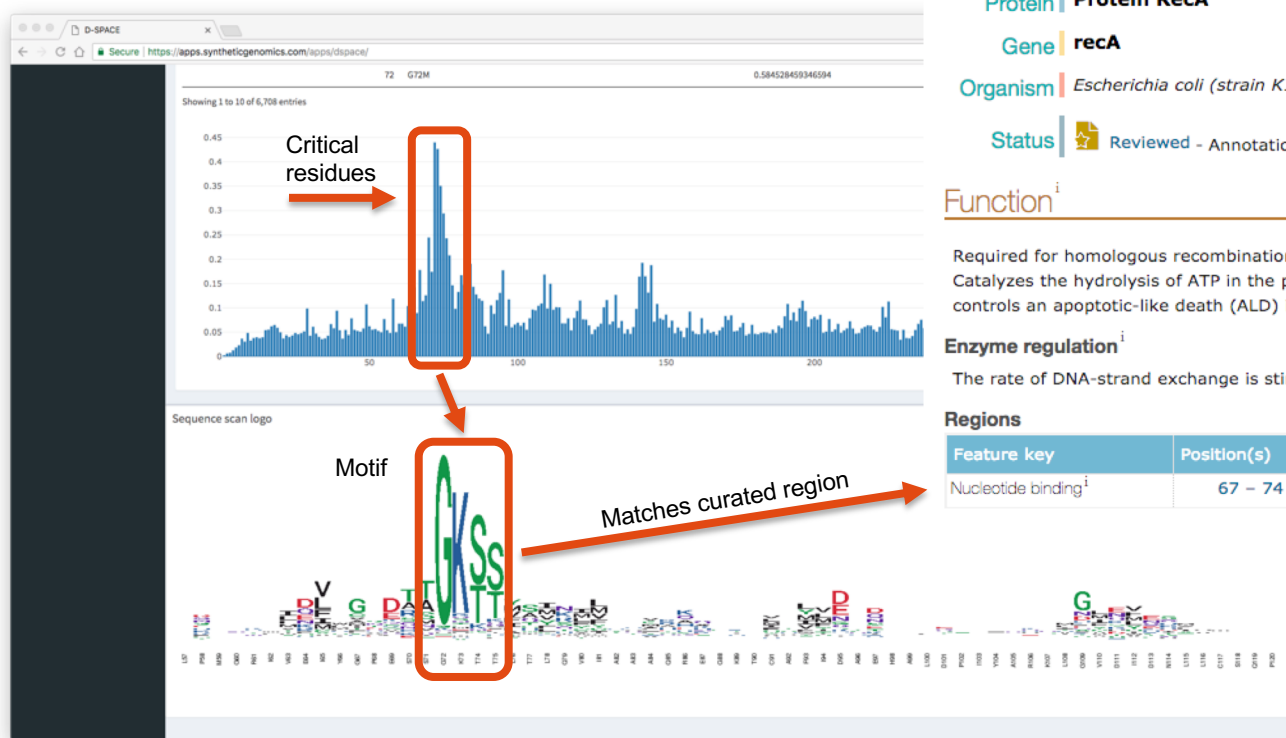
- Scan type:** Alanine, Full, Metropolis, Saved
- Feature type:** Embedding
- Feature to use for comparison:** Embedding Similarity
- ☒ In-silico mutagenesis scanning
- Insertions:** ☐
- Deletions:** ☐
- Maximum Mutations (Metropolis):** A slider set to 10.

**In-silico mutagenesis scanning results**

Show: 10 entries

	Position	Mutation	Embedding_similarity	Feature_score	Abs_delta
	All	All	All	All	All
	72	G72I	0.559978169555559	0.559978169555559	0.440017291286422
	72	G72F	0.560147676159381	0.560147676159381	0.4398477846826
	72	G72P	0.567658572523888	0.567658572523888	0.432233688318093
	72	G72W	0.570954161403615	0.570954161403615	0.429041299438366
	72	G72L	0.57114430930319	0.57114430930319	0.428651151538791
	73	K73P	0.573495292006078	0.573495292006078	0.426500168835903
	72	G72V	0.574440921813981	0.574440921813981	0.425554539028
	72	G72Y	0.574520648590964	0.574520648590964	0.425474812251017

# Demo: towards protein design with D-SPACE



Protein | **Protein RecA**

Gene | **recA**

Organism | *Escherichia coli (strain K12)*

Status | Reviewed - Annotation score: - Experimental evidence at protein level<sup>i</sup>

## Function<sup>i</sup>

Required for homologous recombination and the bypass of mutagenic DNA lesions by the SOS response. Catalyzes the hydrolysis of ATP in the presence of single-stranded DNA, the ATP-dependent uptake of ssDNA, and controls an apoptotic-like death (ALD) induced (in the absence of the mazE-mazF toxin-antitoxin module).

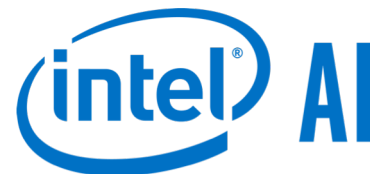
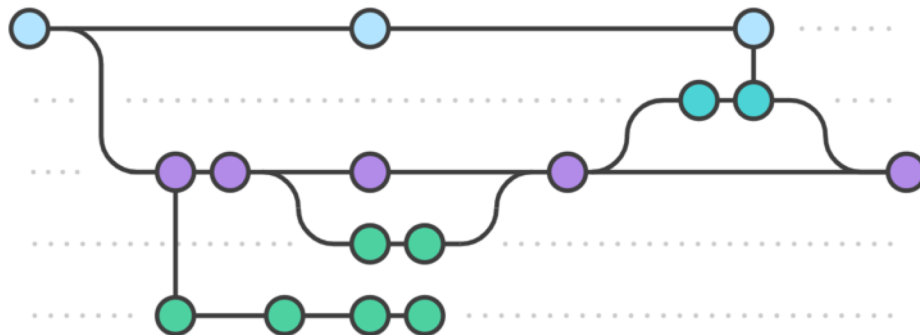
## Enzyme regulation<sup>i</sup>

The rate of DNA-strand exchange is stimulated by RadA. 1 Publication

## Regions

Feature key	Position(s)	Description
Nucleotide binding <sup>i</sup>	67 – 74	ATP  UniRule annotation   1 Publication

# Engagement Model

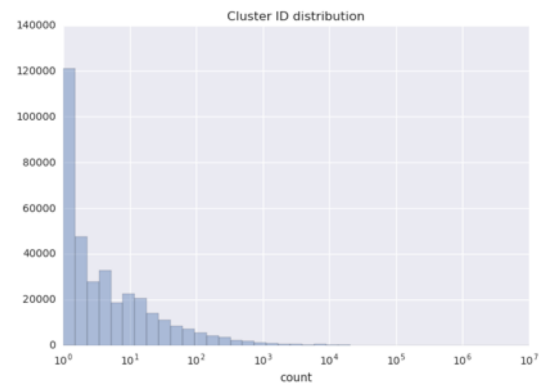
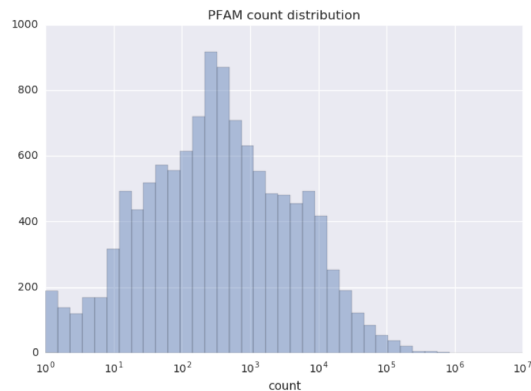
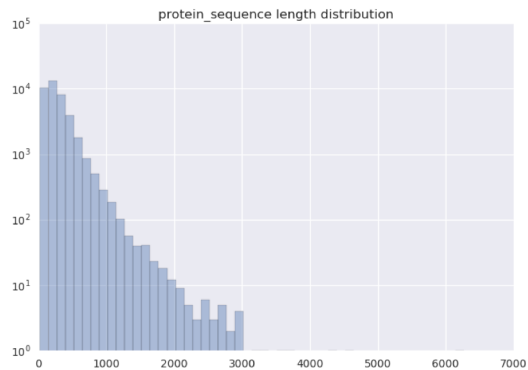


GitHub



Gitter

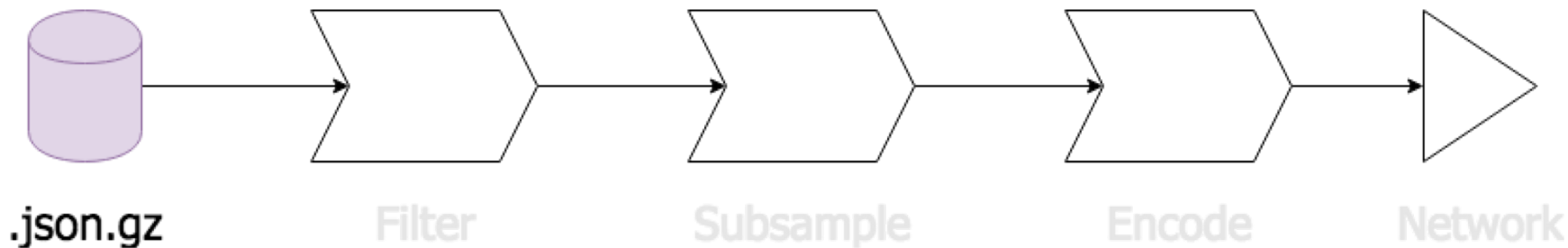
# Initial Understanding





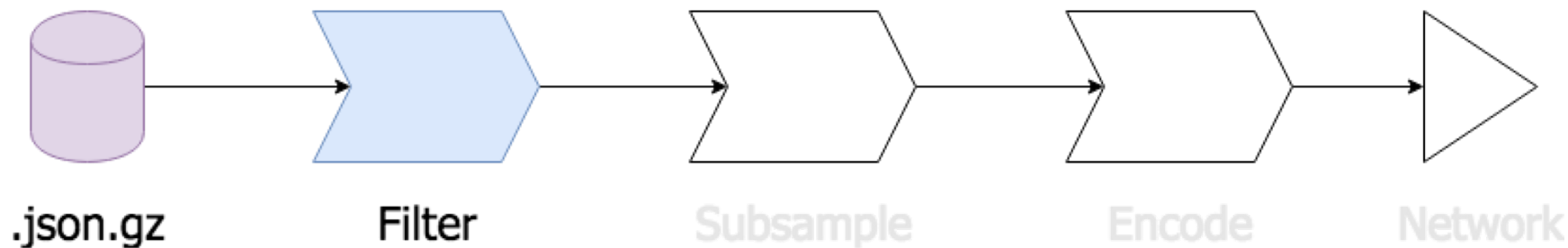
# Data Loading

- .json.gz file format combines disk i/o efficiency with flexibility



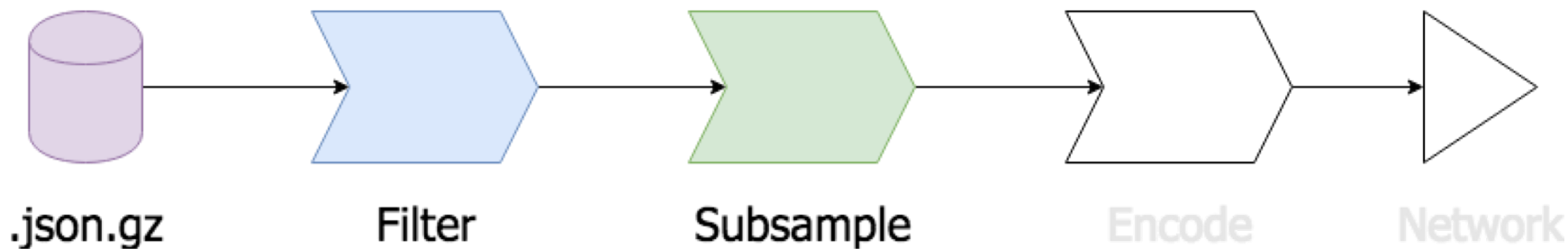
# Data Loading

- On-the-fly filtering allows for rapid iteration



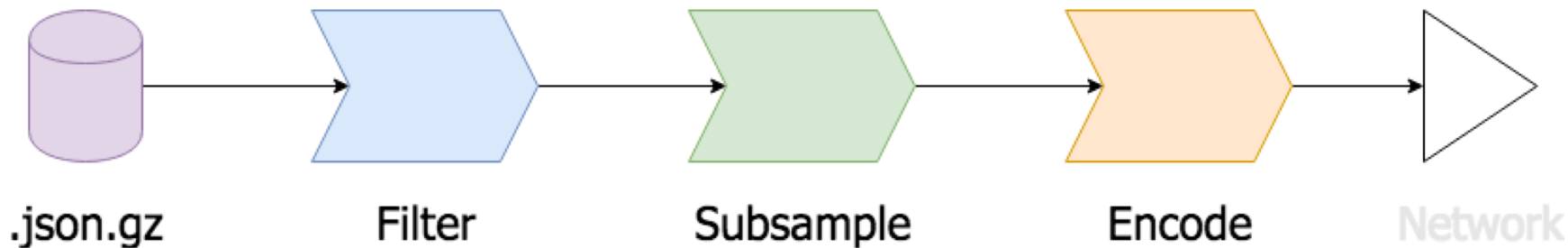
# Data Loading

- On-the-fly subsampling allows for modifying the data distribution

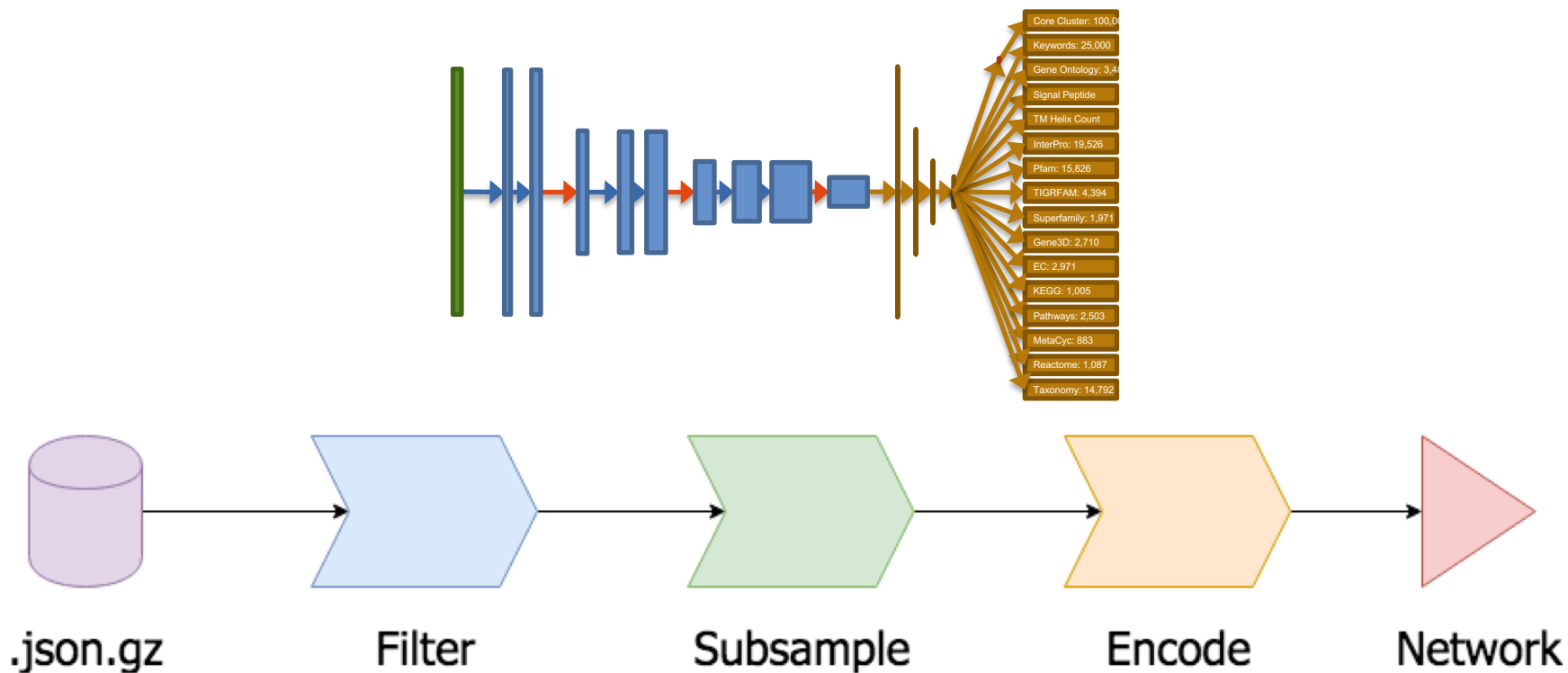


# Data Loading

- Classification: onehot
- Tags: multihot
- Protein sequences:  $21 \times \text{maximum sequence length} \times \text{batch size}$
- Text:  $\text{vocabulary size} \times \text{maximum text length} \times \text{batch size}$



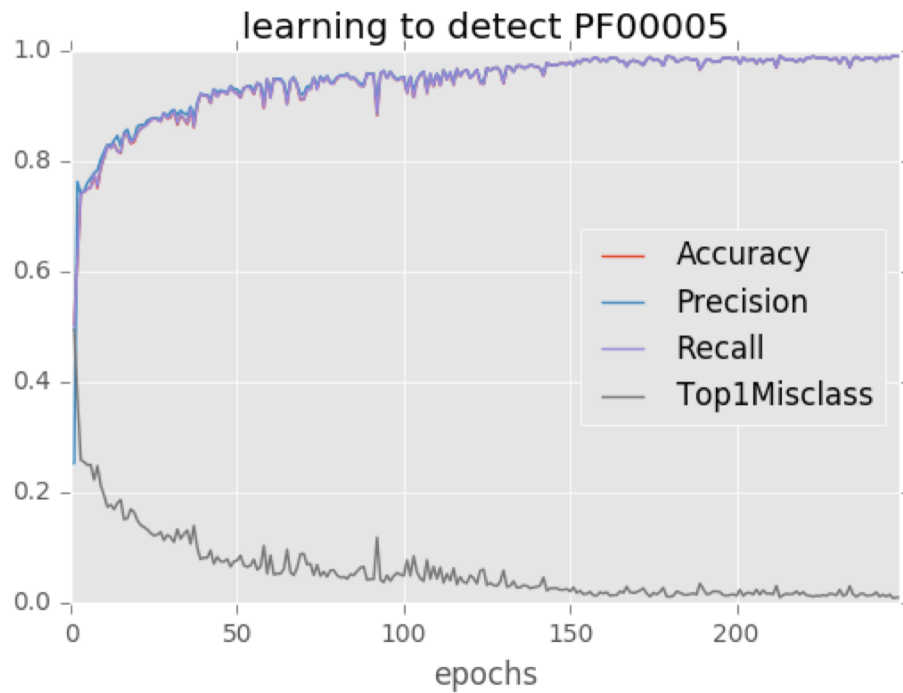
# Data Loading



# Starting Simple

## Single Boolean Classification

- Maximum Sequence Length: 250
- Rebalanced Dataset

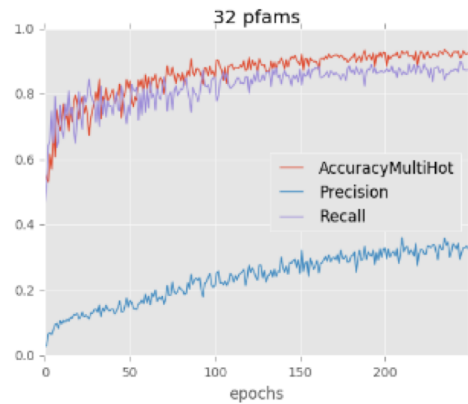
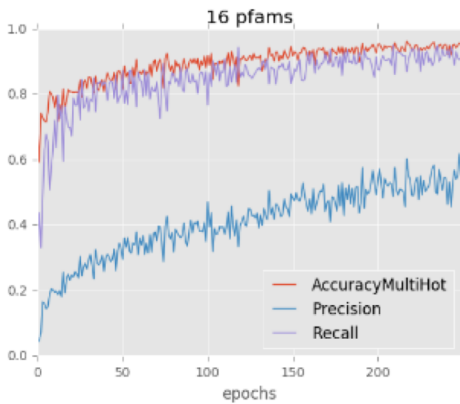
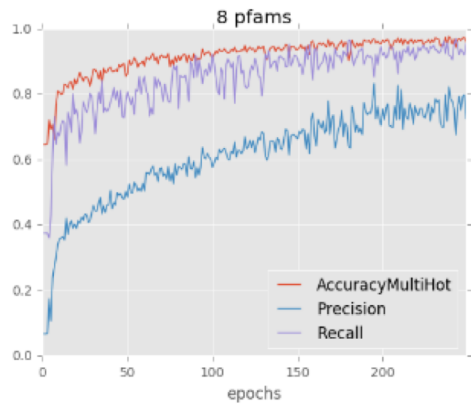




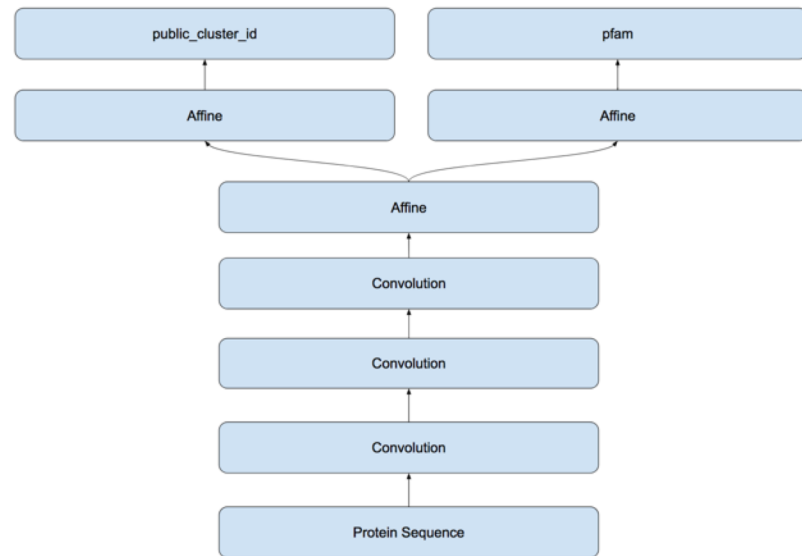
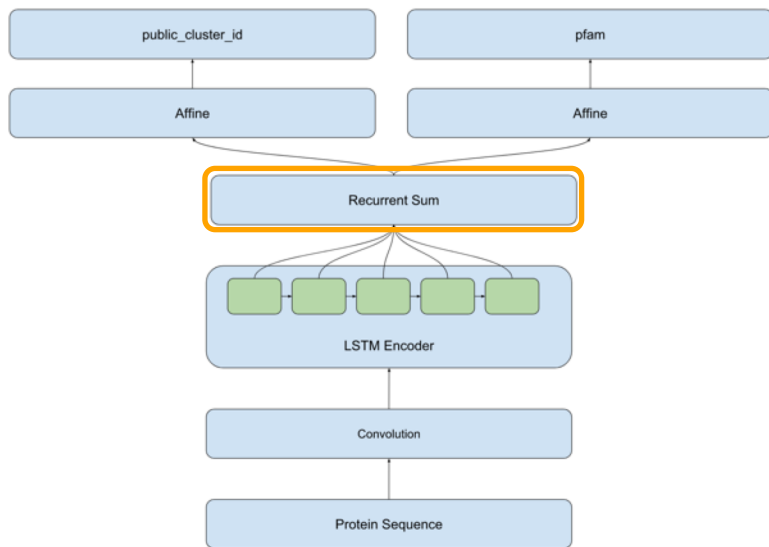
# Increase Complexity

## Multiple Simultaneous Boolean Classification

- Maximum Sequence Length: 250
- Partially Rebalanced Dataset



# Network Architecture : Exploration



# Software Architecture : Multi-task

Supported Output Datatypes:

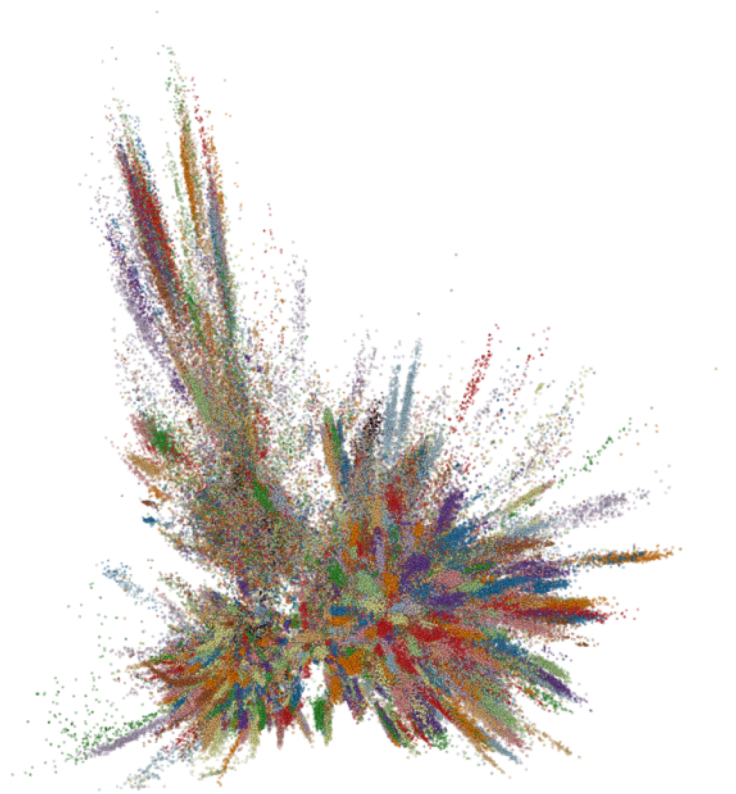
- Boolean
- Numeric
- One-hot : categorical
- Multi-hot : tags
- Text sequence : caption, English description
- Protein sequence : auto-encoder

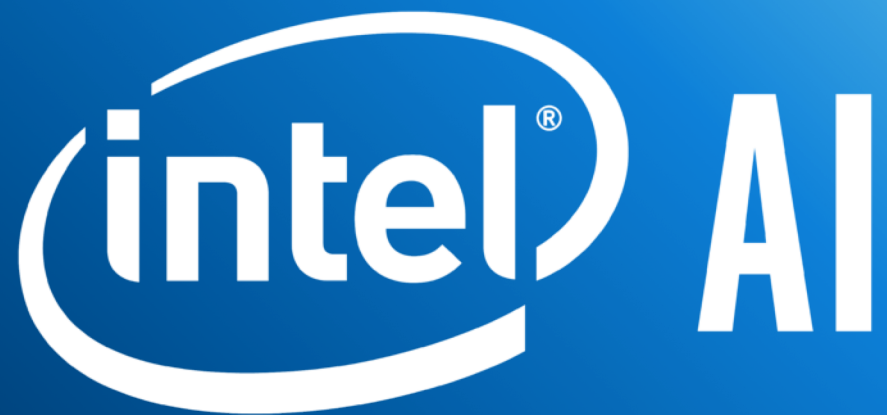
# Software Architecture : Multi-task

FIELD	ACTIVE	SCALE	TYPE	CLASSDICT	CLASSCOUNT	DATADIR	DATAFUN	MINHOT	TEXTLEN
core_public_cluster_id	Y	10	onehot	1610_nervanaPr	100000	.	.	0	.
interproscan	Y	15826	multihot	1610_nervanaPr	15826	.	pfam.model_name	0	.
tm_helix_count	Y	1	numeric	NA	NA	.	.	0	.
signal_peptide	Y	5	boolean	NA	NA	.	.	0	.
ec	Y	2971	multihot	1610_nervanaPr	2971	.	.	0	.
interproscan	Y	4394	multihot	1610_nervanaPr	4394	.	tigrfam.model_name	0	.
sequence_region_tax_hierarchy	Y	14792	multihot	1610_nervanaPr	14792	.	.	0	.
combined_pathway_ids	Y	2530	multihot	1610_nervanaPr	2503	.	.	0	.
translation_description	Y	25000	multihot	1610_nervanaPr	25000	.	keywords	1	.
interproscan	Y	2710	multihot	1610_nervanaPr	2710	.	gene3d.model_name	0	.
interproscan	Y	1971	multihot	1610_nervanaPr	1971	.	superfam.model_name	0	.
interproscan	Y	3486	multihot	1610_nervanaPr	3486	.	go.accession	0	.
interproscan	Y	19526	multihot	1610_nervanaPr	19526	.	interpro.accession	0	.
interproscan	Y	1005	multihot	1610_nervanaPr	1005	.	kegg.accession	0	.
interproscan	Y	883	multihot	1610_nervanaPr	883	.	metacyc.accession	0	.
interproscan	Y	1087	multihot	1610_nervanaPr	1087	.	reactome.accession	0	.

# Summary

- D-SPACE is a result of a very productive collaboration between SGI and Intel AI Lab
- Demonstrated applicability of recent advances in deep-learning to genomics applications
- D-SPACE can annotate previously uncharacterized proteins, discover novel enzymes, and assist in protein design and optimization





# Notices and Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit: <http://www.intel.com/performance>.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks).

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com](http://intel.com).

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel, the Intel logo, Xeon, Xeon Phi and Nervana are trademarks of Intel Corporation in the U.S. and/or other countries.

\*Other names and brands may be claimed as the property of others

© 2018 Intel Corporation. All rights reserved.