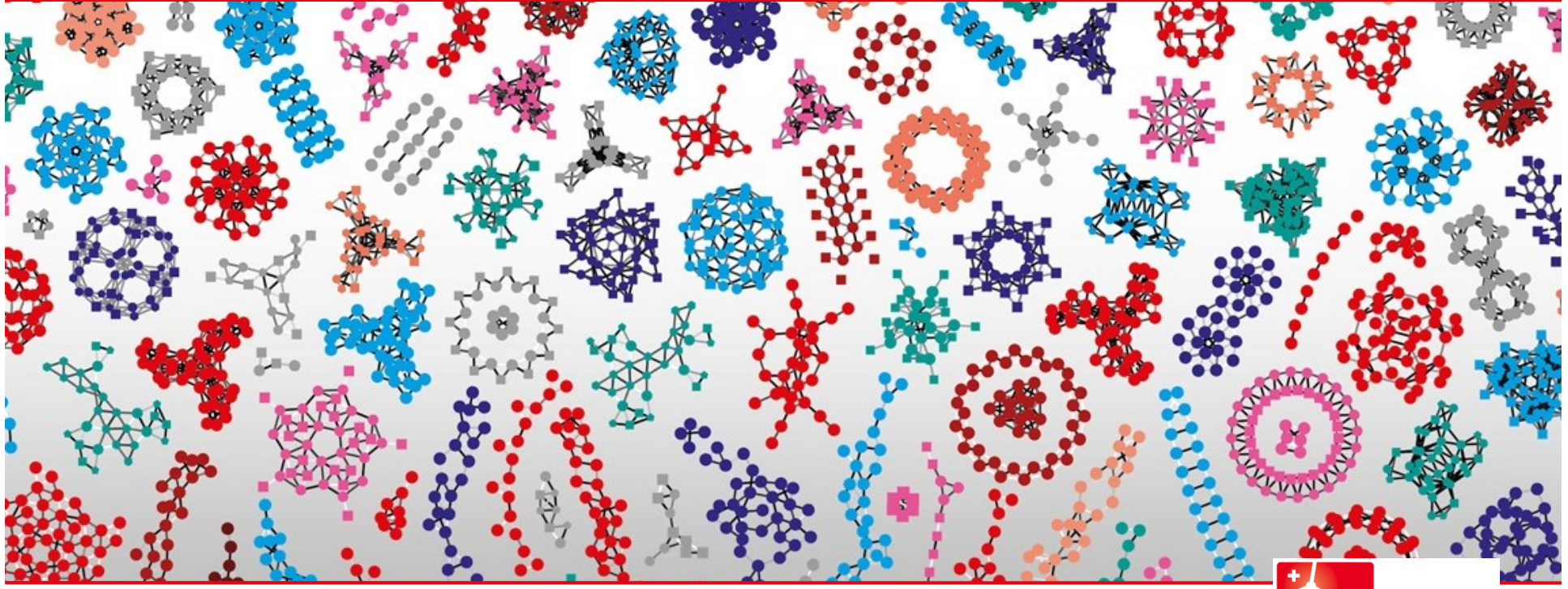
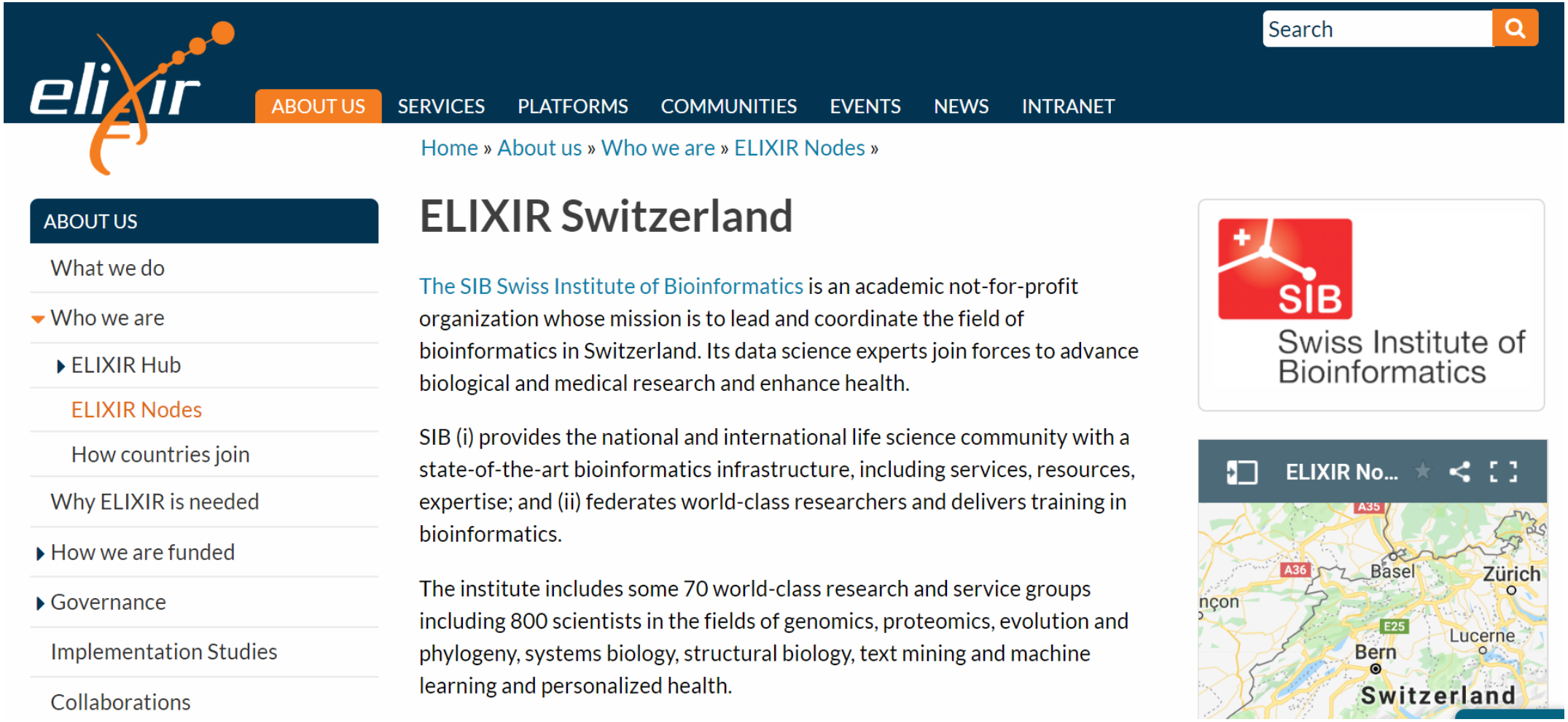


Swiss Institute of Bioinformatics

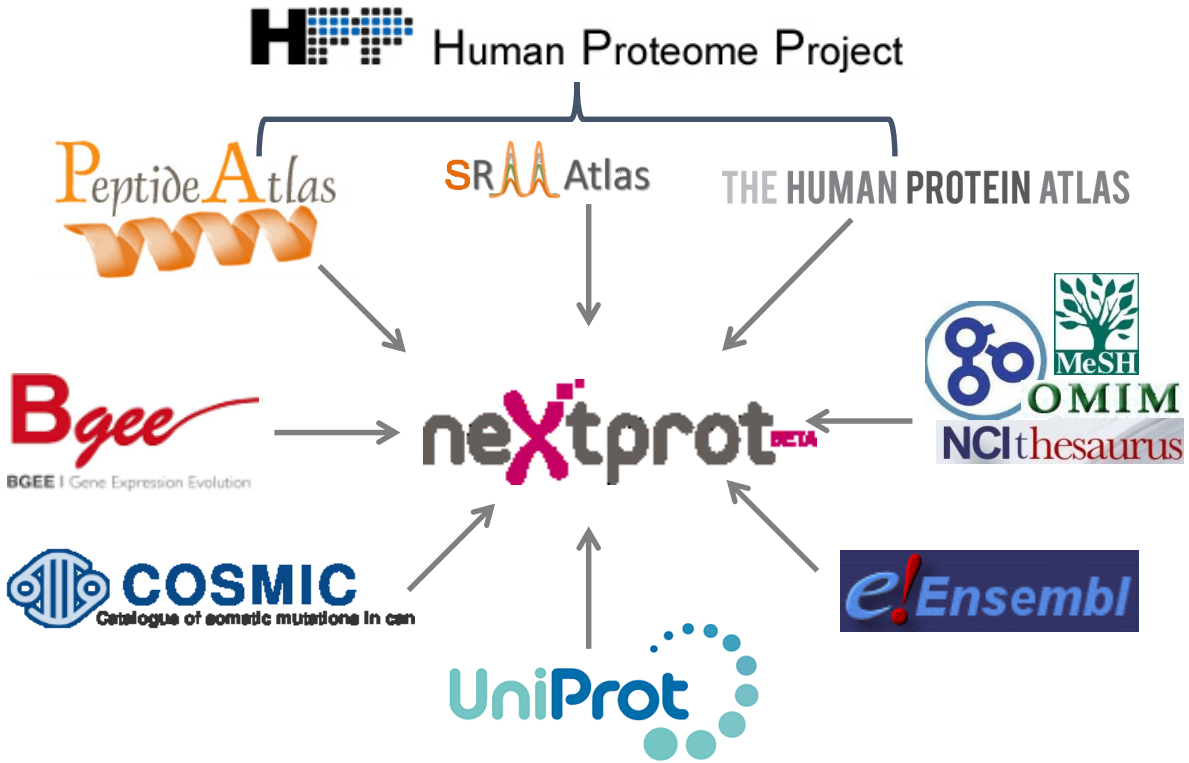
Using SPARQL to explore human protein data in neXtProt and beyond

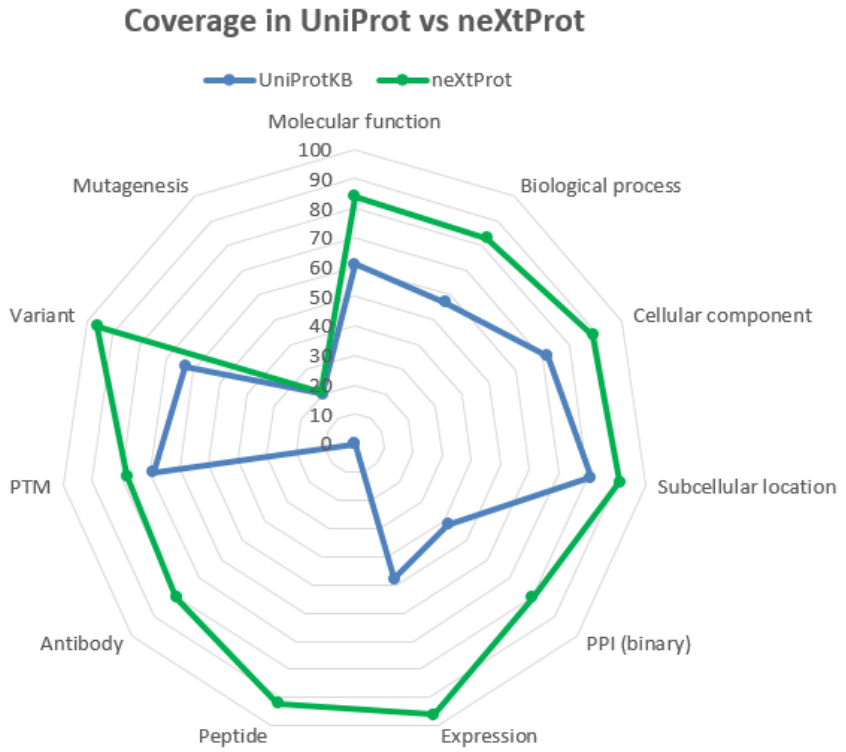
Lydie Lane,
OpenRiskNet final workshop, Amsterdam, Oct 24 2019

www.nextprot.org is the SIB database on human proteins

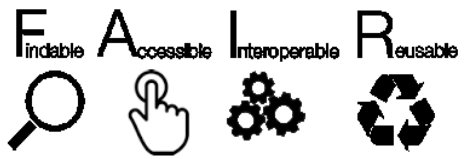


www.nextprot.org collects and represents knowledge on human proteins at genomic, transcriptomic and proteomic levels.





Access to neXtProt data and tools



FINDABLE

- **ENTRIES have a globally unique and persistent ID**

ACCESSIBLE

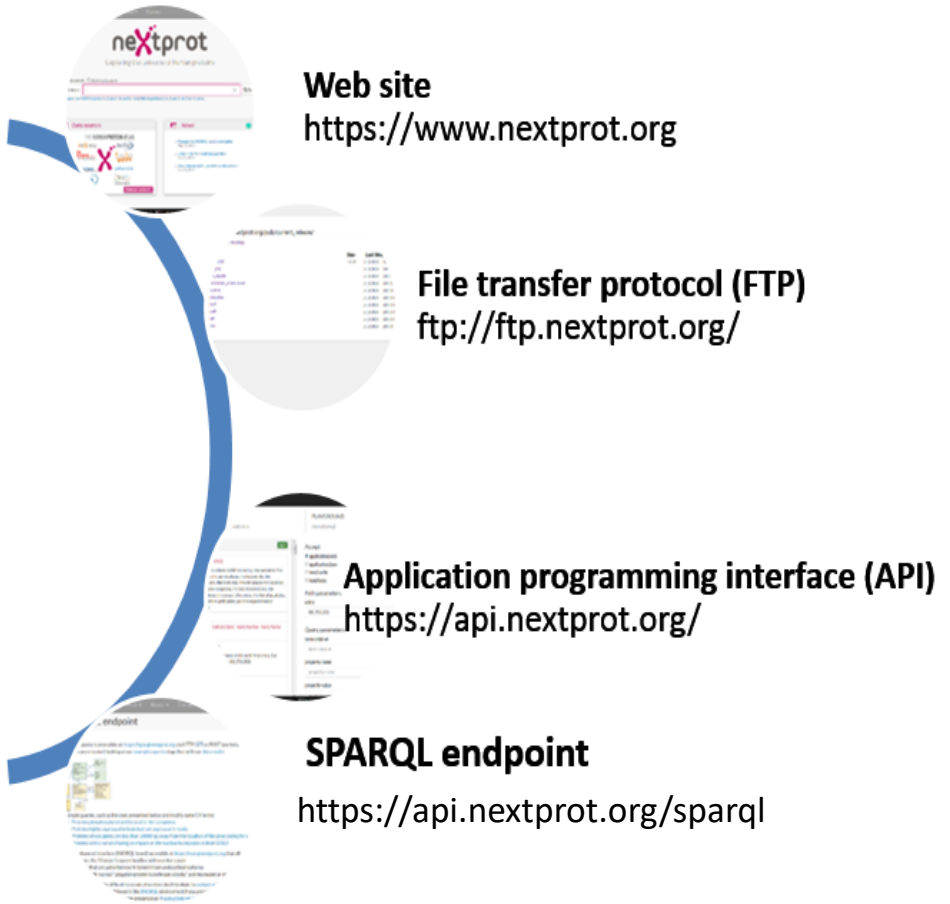
- **Web: human readable**
- **API: machine parsable**
- **FTP: XML, PEFF, RDF/ttl**

INTEROPERABLE

- **Use controlled vocabularies and ontologies**
- **SPARQL end-point**

REUSABLE

- **Detailed provenance**
- **Data license CC BY 4.0**
- **All data since 2011 on FTP**
- **Code on GitHub**



neXtProt in numbers

Release statistics

Search in neXtProt...

Data release: 2019-08-22

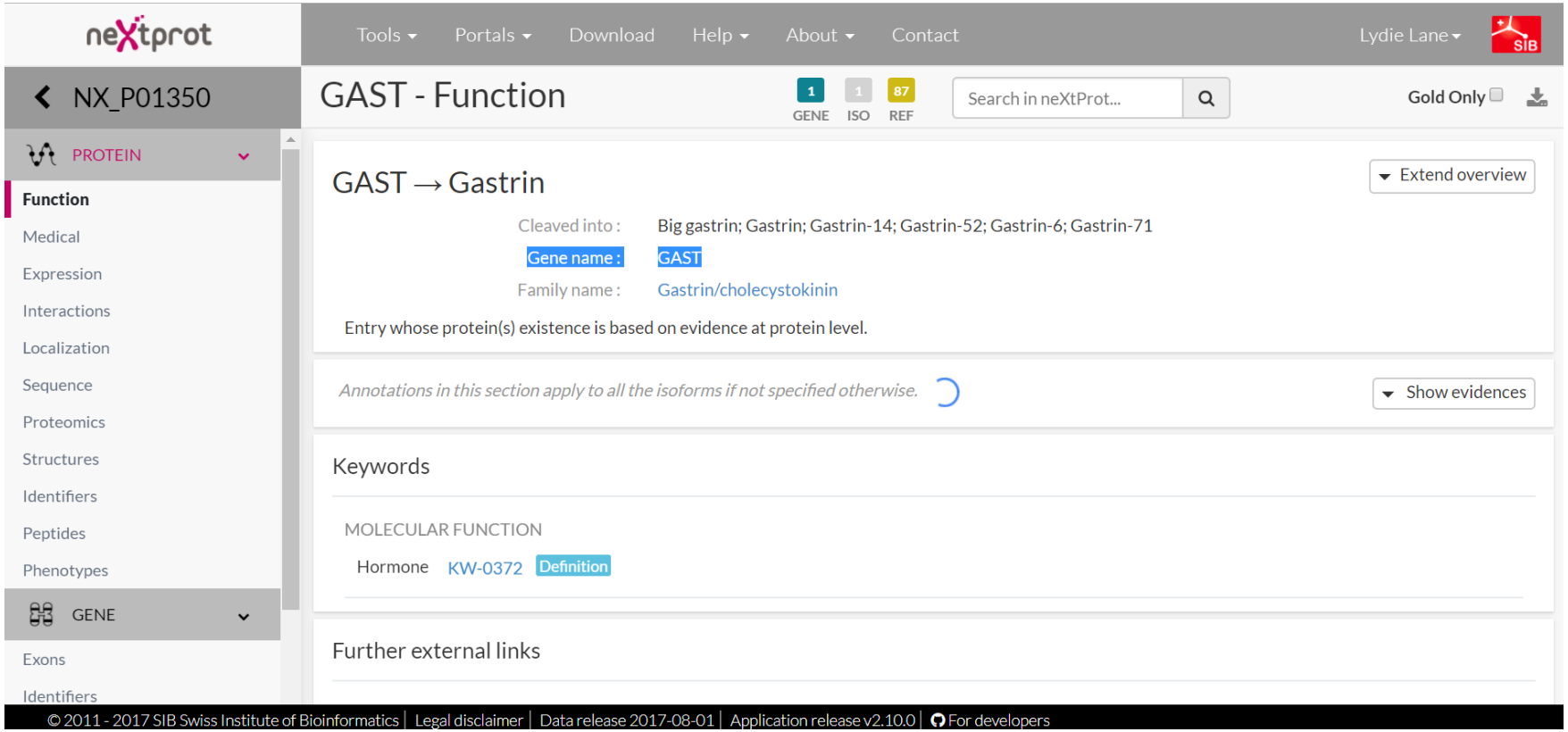
Application release: v2.23.2

Data statistics

ENTRIES	Protein entries	20399
	Isoforms (produced by splicing)	42410
	Identifiers	878422
	Binary interactions	240010
	Post-translational modifications (PTMs)	190921
	Natural peptides (seen by MS)	1772990
	Variants (including disease mutations)	6019871
	Controlled vocabularies and ontology terms	252243
	Total number of publications	495928
	Links between protein entries and terms	3977548

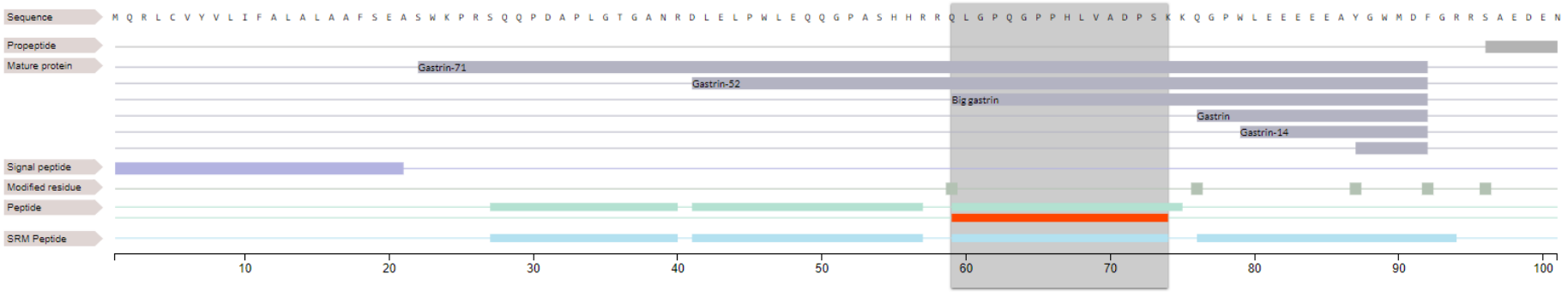
neXtProt website

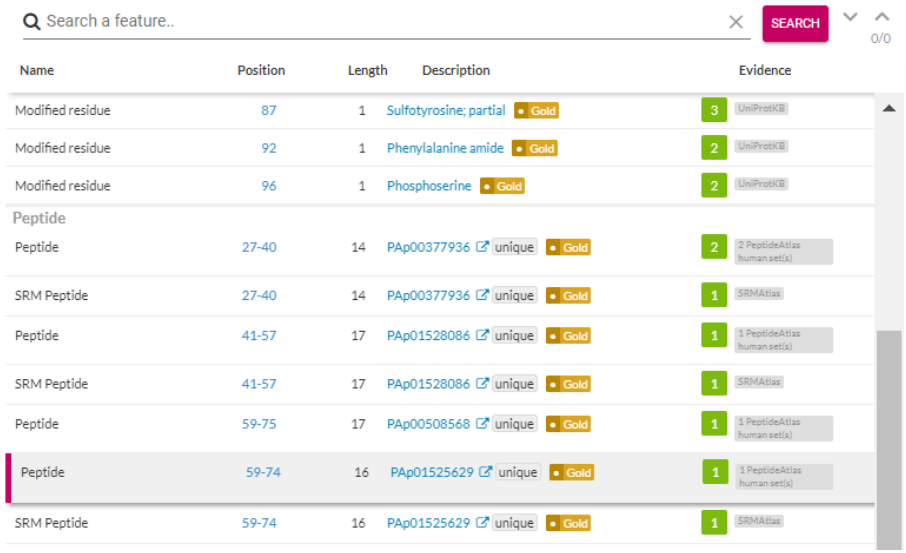
Each protein entry has multiple views

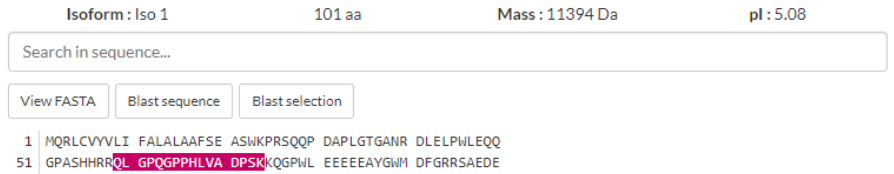


Proteomics view

- *Which MS/MS peptides from my protein have been found ?*
- *Which PTM?*

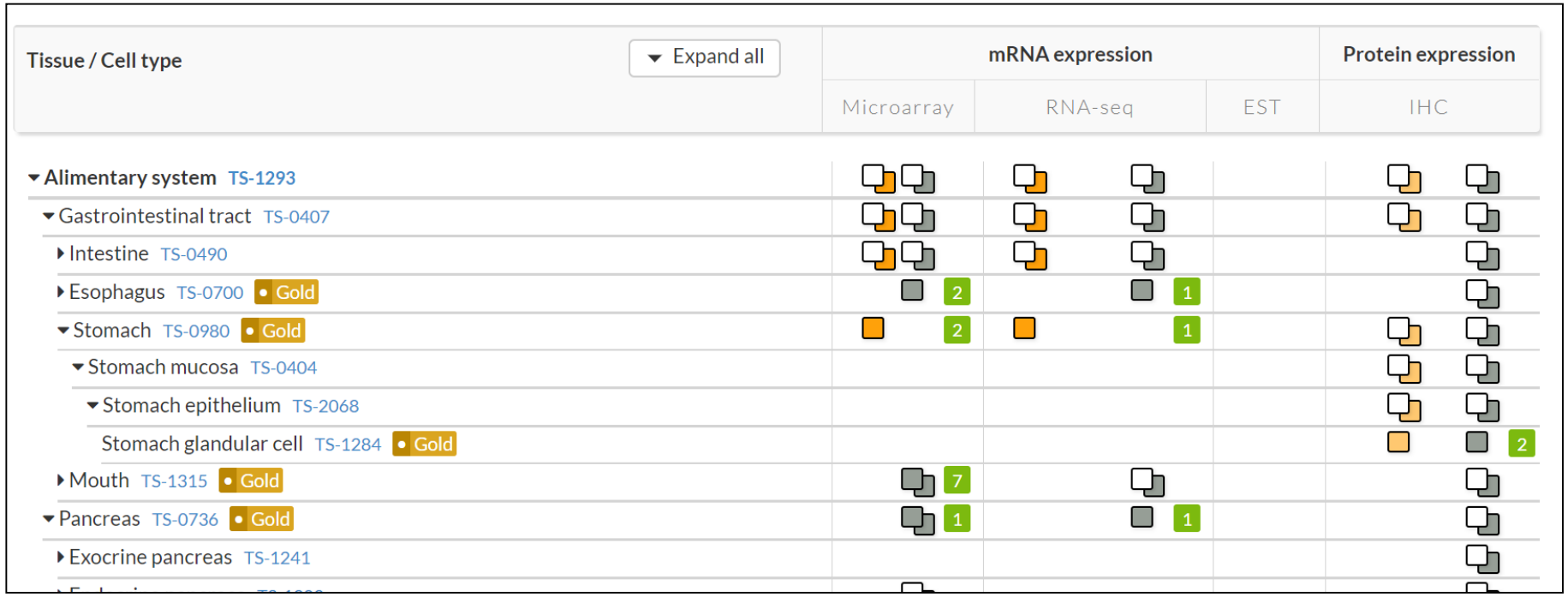






Expression view

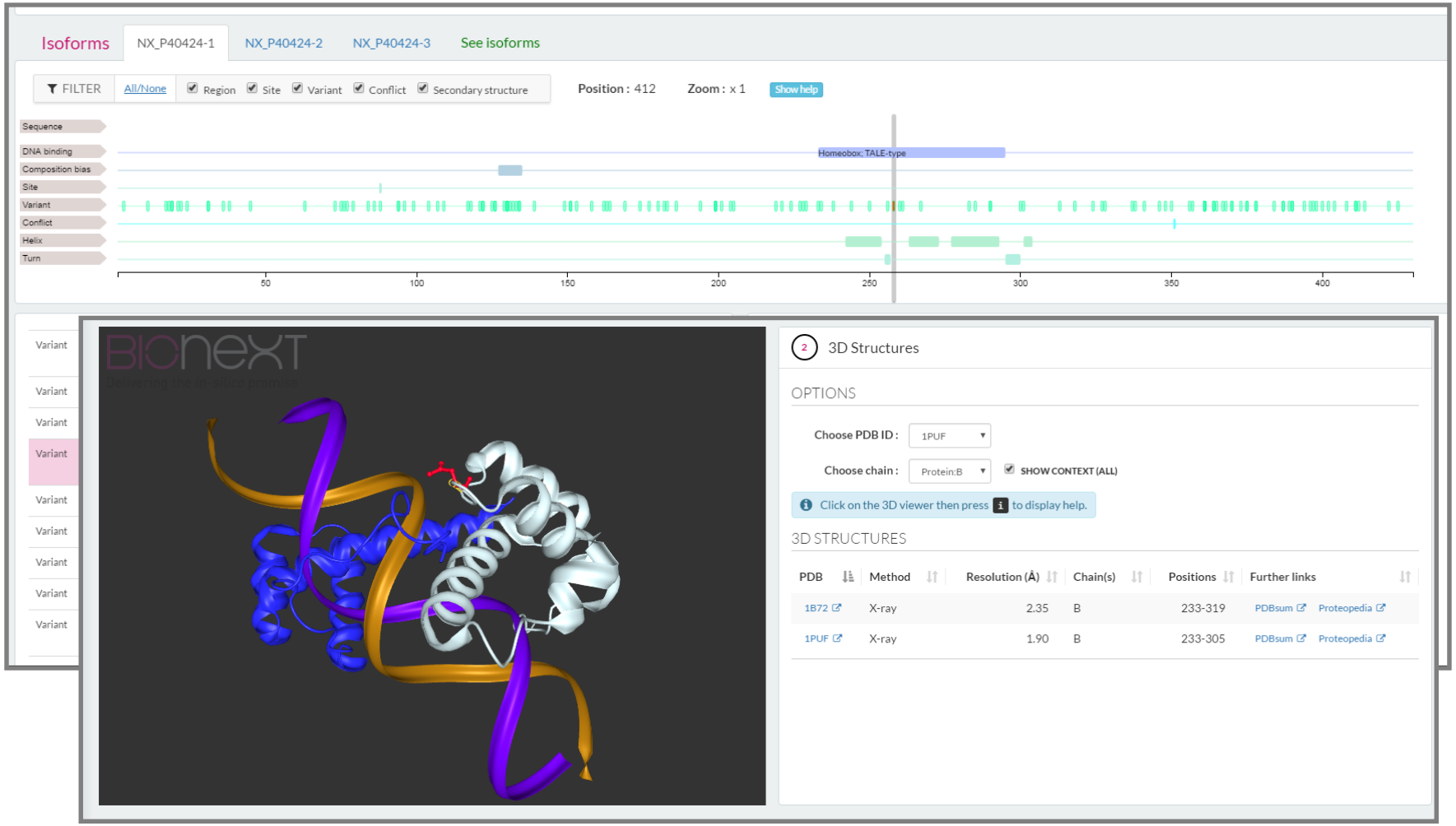
✓ *Where is this protein expressed, at mRNA and protein levels?*



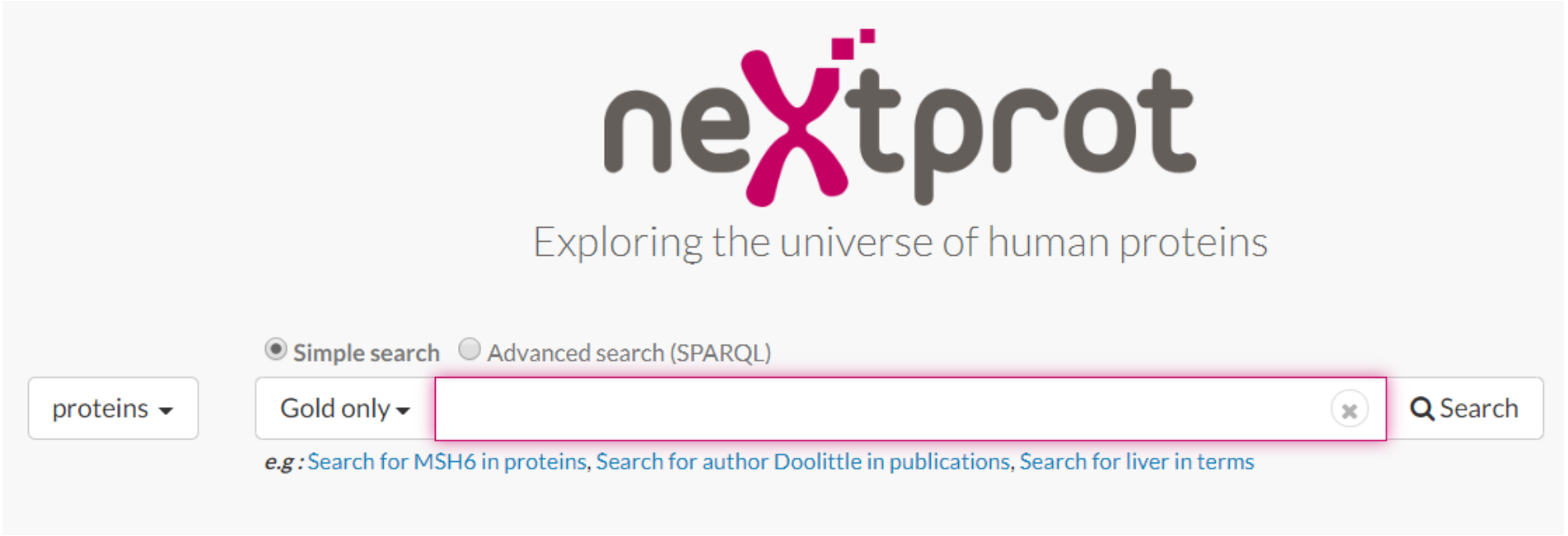
Expression information is integrated from Bgee (MicroArray and EST) and from HPA (IHC protein and RNA-seq)

Structure view

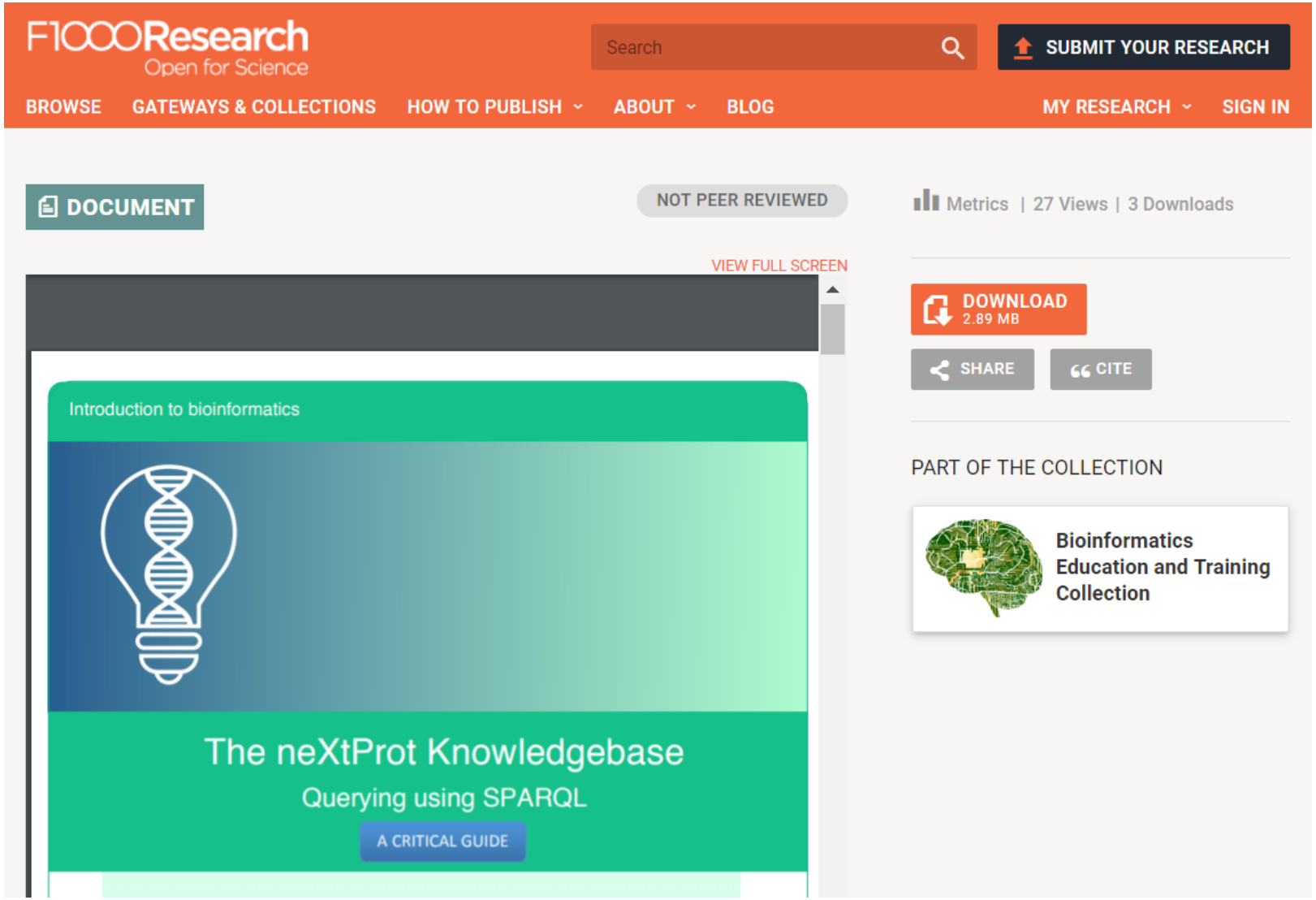
✓ *Is this particular variant externally accessible or internal to the structure?*



Searching for information in neXtProt



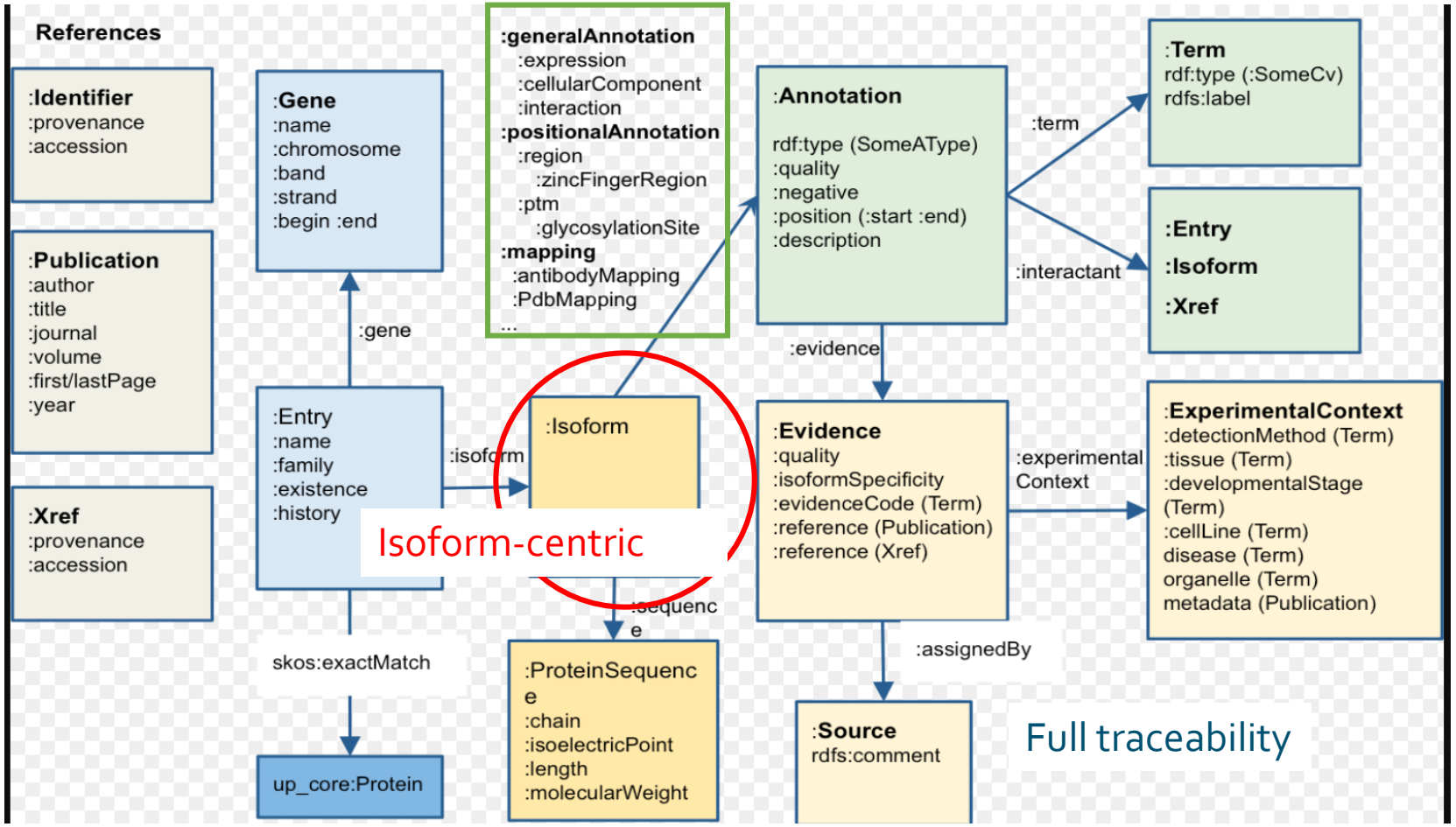
neXtProt SPARQL for beginners



www.nextprot.org RDF data model

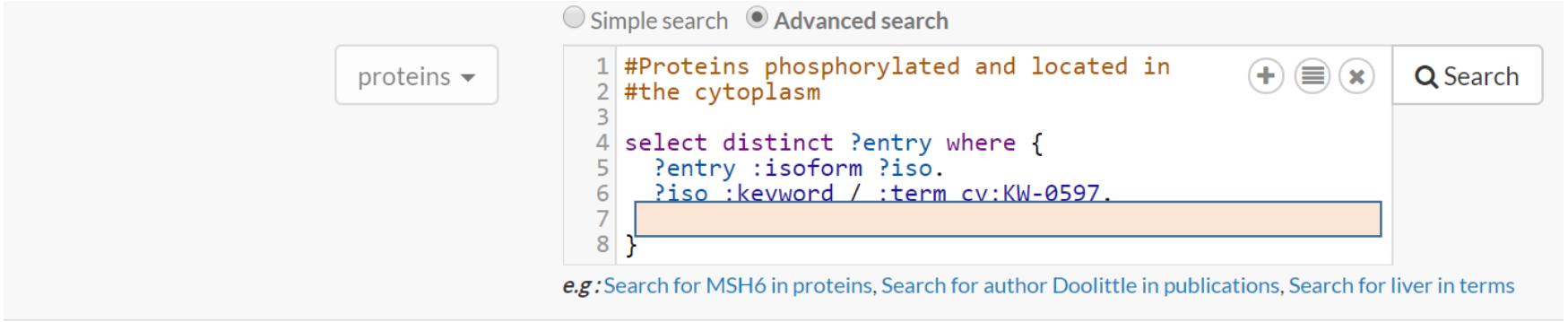
A wide variety of experimental and predicted annotations

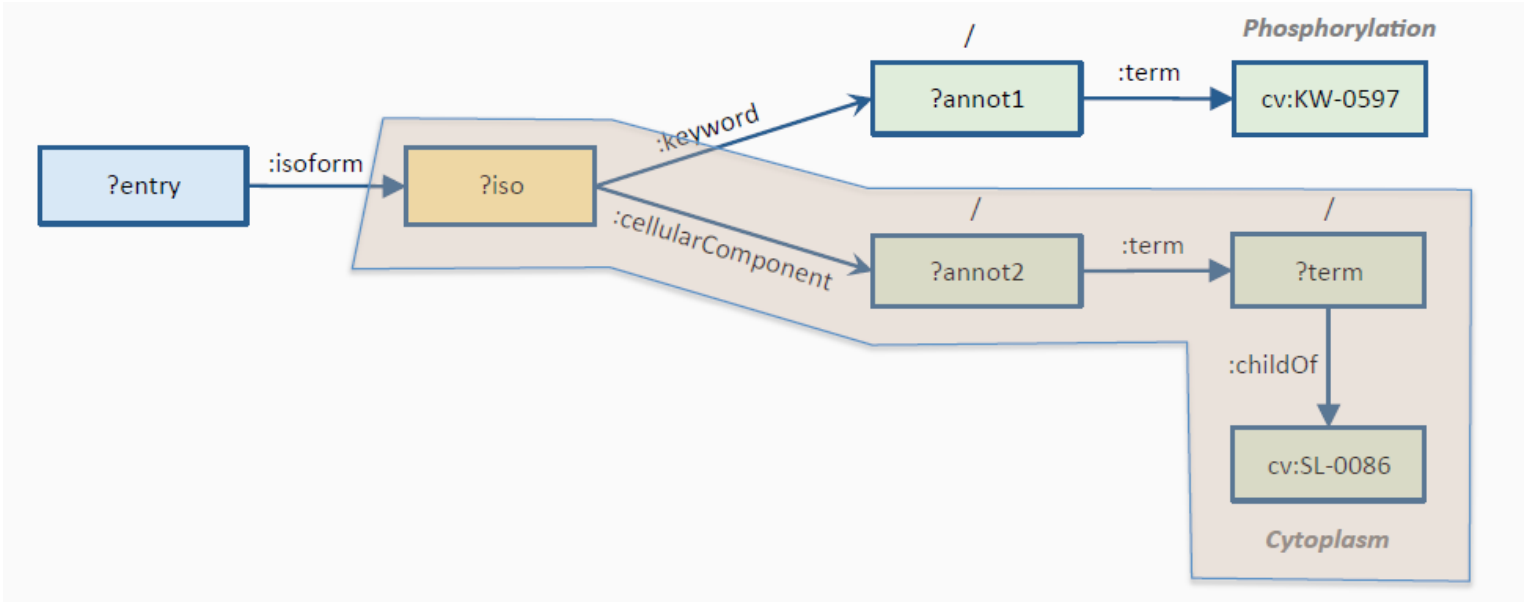
Extensive use of CV



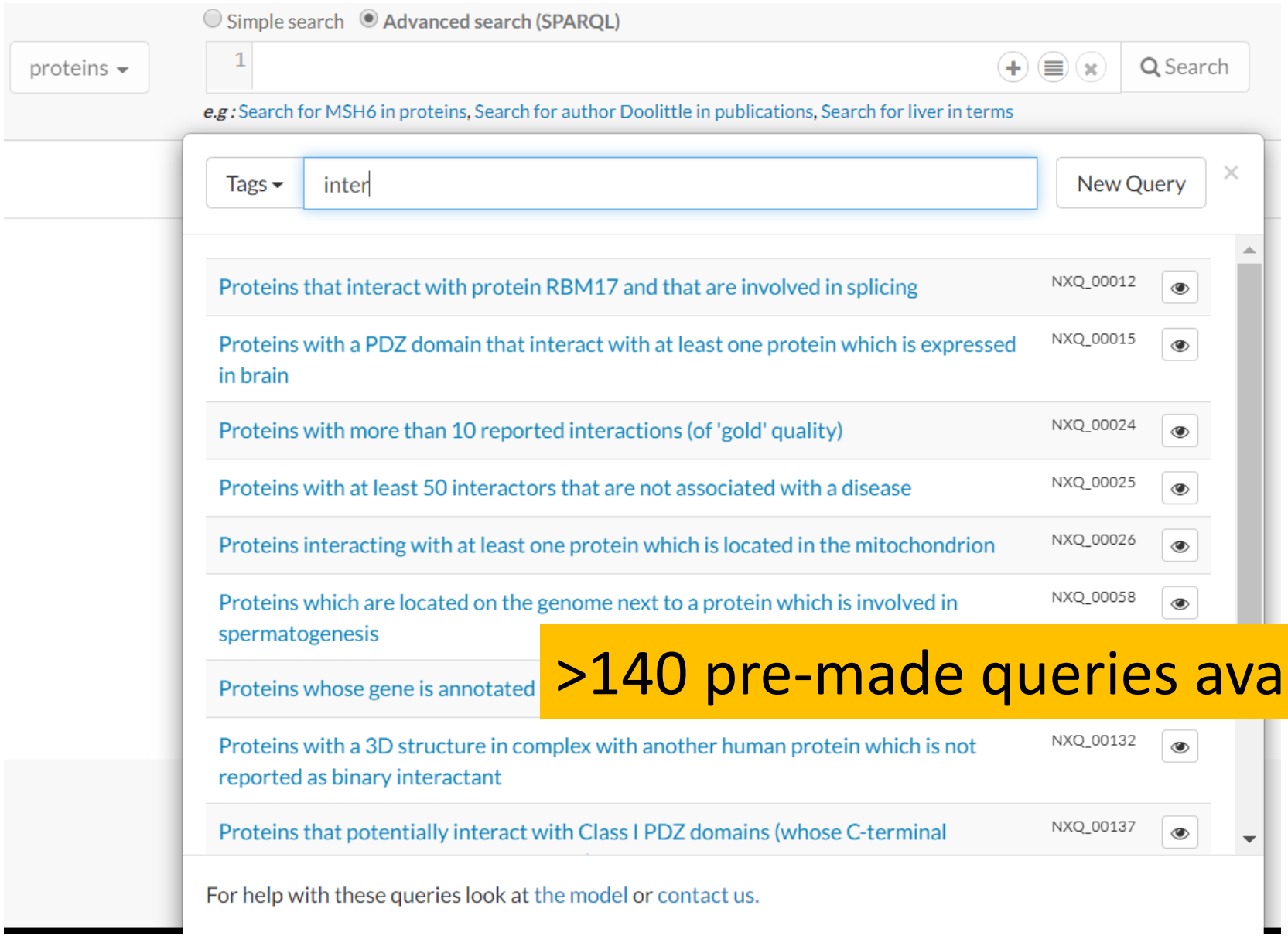
Advanced search (SPARQL) interface 1/2

Retrieves lists of entries with associated information





Advanced search (SPARQL) interface 2/2



>140 pre-made queries available

SNORQL interface https://snorql.nextprot.org

Retrieves any data as a table

NXQ_00124 - What are the 25 most frequent families with member count

family snorql-only tutorial

>170 pre-made queries available

NXQ_00256 - Variants identified in exome datasets in a frequent homozygote state

evidence sequence variation snorql-only tutorial variant

What are the 25 most frequent families with member count

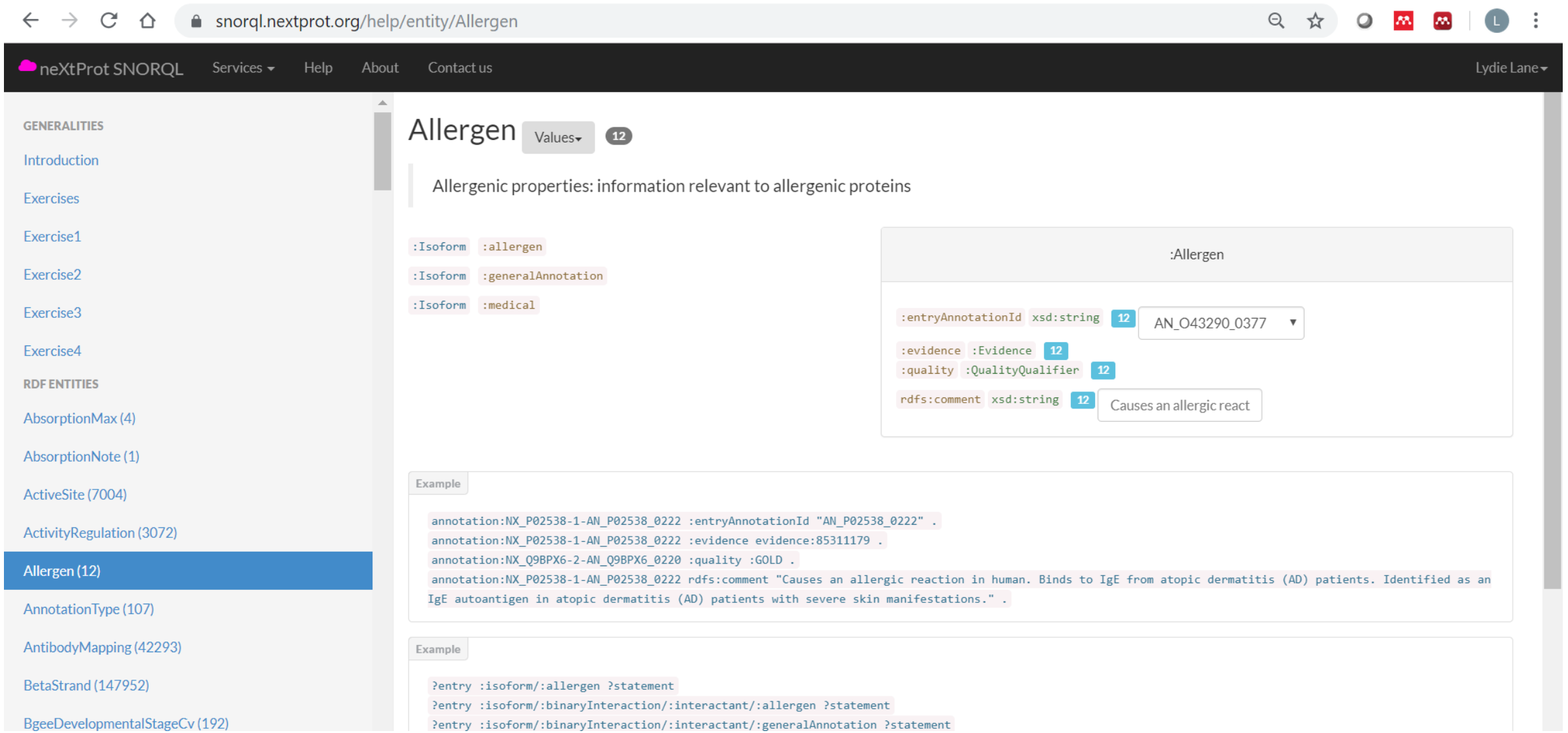
html Go Reset eg. peroxisome, liver

Found a bug? Improve this query!

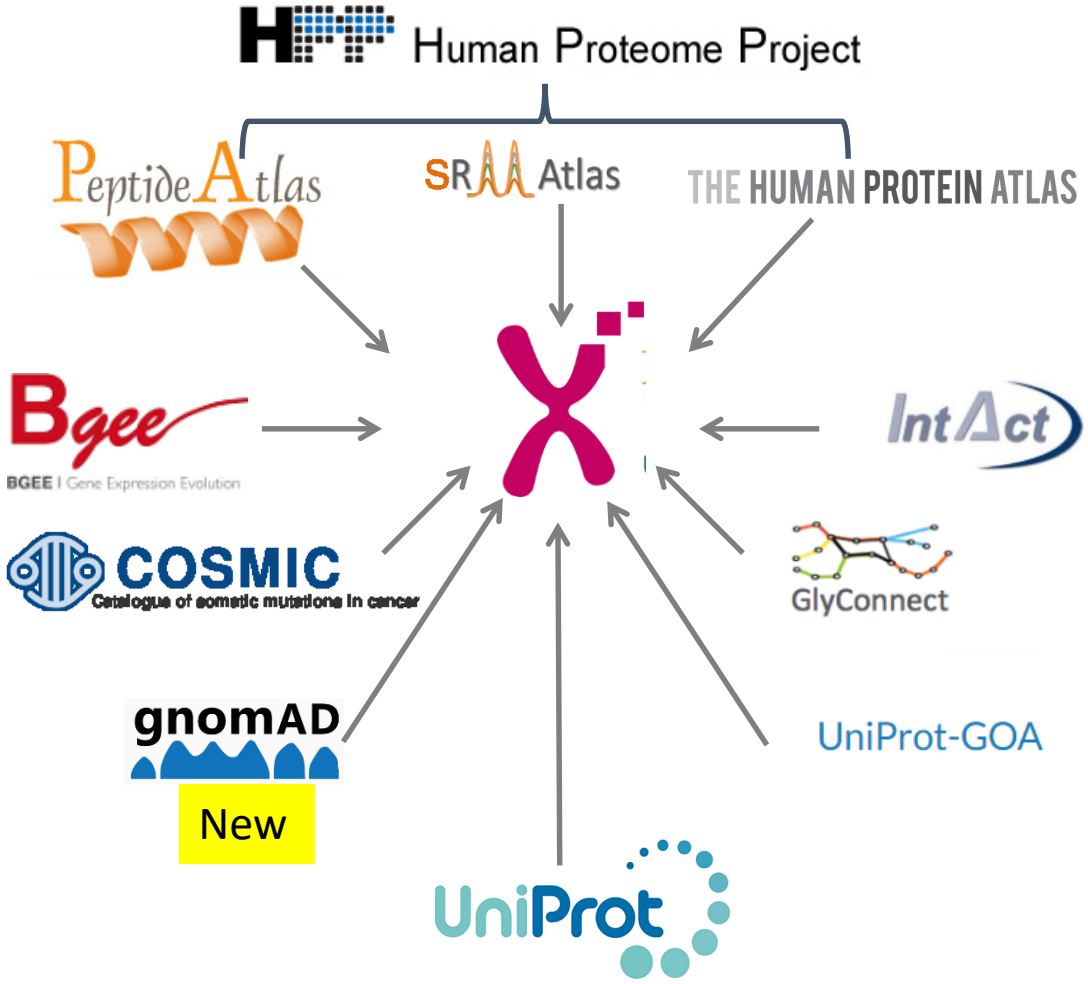
Query time is 0.291[s] for 25 rows

familylabel	membercnt
"G-protein coupled receptor 1 family"	723
"Krueppel C2H2-type zinc-finger protein family"	546
"Protein kinase superfamily"	492
"Small GTPase superfamily"	162
"Immunoglobulin superfamily"	130
"Peptidase S1 family"	121
"Major facilitator superfamily"	99

A detailed help on neXtProt RDF entities is available



www.nextprot.org does not contain everything….



- ✓ MS-data
- ✓ Antibody-based data
- ✓ RNA-seq data
- ✓ PTMs
- ✓ Variants (genomic and somatic)
- ✓ PPIs
- ✓ Functional annotations

- ○ Phylogenetic information
- ○ Data on model organisms
- ○ Pharmacology data
- ○ Toxicology data
- ○ Clinical proteogenomic data
- ○ Structural data
- ○ Protein interactions with pathogens
- ○ …

Going further with federated queries

Tags ▾ federated

NXQ_00094 - Proteins which are targets of antipsychotic drugs and highly expressed in brain
drug expression federated query tutorial

NXQ_00096 - Proteins which are targets of drugs for cardiac therapy
drug federated query tutorial

NXQ_00139 - Protein kinases which are drug targets according to CHEMBL
CHEMBL drug federated query tutorial

NXQ_00140 - Proteins that interact with viral proteins
federated query interaction PPI tutorial UniProt

NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein
expression federated query interaction PDB snorql-only tutorial UniProt

NXQ_00246 - Proteins which are enzymes catalyzing a reaction involving lipids
enzyme federated query tutorial

NXQ_00253 - Human pathways in which at least one protein is mitochondrial GOLD
federated query pathway snorql-only subcellular location tutorial

NXQ_00254 - Proteins with associated pathways in WikiPathways
federated query pathway snorql-only tutorial

And soon to come…

Federating neXtProt and WikiPathways

SPARQL endpoint: https://api.nextprot.org/sparql

PREFIX...

```
PREFIX wp: <http://vocabularies.wikipathways.org/wp#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>

select distinct ?pathwayname group_concat(distinct ?gen , ',') as ?genes where {
service <http://sparql.wikipathways.org/sparql> {
SELECT DISTINCT ?pathwayname ?gen WHERE
{
  { ?geneProduct a wp:GeneProduct .  }
  union
  { ?geneProduct a wp:Protein .  }
  ?geneProduct rdfs:label ?gen .
  filter(!regex(?gen,"[ a-z-]")). # ensures official gene names for subsequent neXtprot matching
  ?geneProduct rdfs:label ?gen .
  ?geneProduct dcterms:isPartOf ?pathway .
  ?pathway a wp:Pathway .
  ?pathway wp:organism ?organism .
  FILTER(contains(str(?organism),"9606"))
  ?pathway dc:title ?pathwayname .
}
}
?entry a :Entry .
?entry :gene / :name ?gen .
?entry :isoform / :cellularComponent ?loc .
values ?mitoloc {cv:SL-0173 cv:GO_0005739 } # SL and GO values for mitochondrion
```

Tags ▾ pathw

NXQ_00022 - Proteins with no function annotated
function tutorial

NXQ_00044 - Proteins involved in the Reactome pathway "Respiratory electron transport"
pathway tutorial

NXQ_00135 - Proteins involved in both Wnt and Hippo signaling pathways
function pathway tutorial

NXQ_00253 - Human pathways in which at least one protein is mitochondrial GOLD
federated query pathway snorql-only subcellular location tutorial

NXQ_00254 - Proteins with associated pathways in WikiPathways
federated query pathway snorql-only tutorial

html ▾ Go Reset

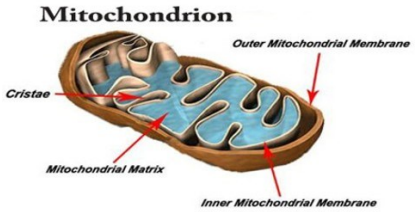
html
json
csv
xml

eg. peroxisome, liver

Improve this query!

is 0.502[s] fo 781 rows

pathwayname	genes
"22q11.2 Deletion Syndrome"@en	"AIFM3,ALDH4A1,GLUD1,MRPL40,OAT,PRODH,RAF1,RTL10,SLC25A1,TXNRD2"
"4-hydroxytamoxifen, Dexamethasone, and Retinoic Acids Regulation of p27 Expression"@en	"MAP2K1,MAP2K2,MAPK1,MAPK3,MTOR,RAF1"
"ABC-family proteins mediated transport"@en	"ABCB10,ABCB6,ABCB8,ABCD1,ABCG1,PSMB3,RNF185,RNF5,UBB"
"ACE Inhibitor Pathway"@en	"CYP11B2"
"AGE/RAGE pathway"@en	"ATF2,CASP3,CYCS,FOXO1,LGALS3,MAP2K1,MAPK1,MAPK3,MAPK8,MMP2,MTOR,MYD88,PRKCA,RAF1,SHC1,SOD1,SRC



Retrieving proteins involved in pathways of interest, and their properties

Example: 106 entries associated with pathways named “nano…” (according to Wikipathways), and their associated subcellular location (GOLD according to neXtProt)

Query time is 0.663[s] for 106 rows

entry	gen	pathwayname	locs
entry:NX_Q07812 -- (neXtProt link)	"BAX"^^xsd:string	"Nanomaterial induced apoptosis"@en	"BAX complex,Bcl-2 family protein complex,Cytoplasm,Mitochond membrane,cell periphery,cytoplasm,cytosol,endoplasmic reticulun reticulum membrane,mitochondrial outer membrane,mitochondri transition pore complex,mitochondrion,nuclear envelope,nucleus,p
entry:NX_Q14249 -- (neXtProt link)	"ENDOG"^^xsd:string	"Nanomaterial induced apoptosis"@en	"Mitochondrion,mitochondrial inner membrane,mitochondrion,nu
entry:NX_Q16611 -- (neXtProt link)	"BAK1"^^xsd:string	"Nanomaterial induced apoptosis"@en	"BAK complex,Mitochondrion outer membrane,integral componer outer membrane,mitochondrial outer membrane,mitochondrion,p
entry:NX_P10415 -- (neXtProt link)	"BCL2"^^xsd:string	"Nanomaterial induced apoptosis"@en	"Endoplasmic reticulum membrane,Mitochondrion outer membrane,Nucleoplasm,Nucleus membrane,cytoplasm,endoplasm reticulum,endoplasmic reticulum membrane,membrane,mitochon

Federating neXtProt, Wikipathways and DrugBank

169 proteins are part of “cancer pathways” and “druggable” according to DrugBank (35 are known targets of anti-neoplastic agents)

neXtprot

entry	callret-1	cpathways	drugs
entry:NX_O15530 -- (neXtProt link)	"PDPK1"	"Endometrial cancer"	"Celecoxib"
entry:NX_P00352 -- (neXtProt link)	"ALDH1A1"	"Folate-Alcohol and Cancer Pathway Hypotheses"	"Tretinoin"
entry:NX_P00374 -- (neXtProt link)	"DHFR"	"Retinoblastoma Gene in Cancer"	"Methotrexate,Pemetrexed"
entry:NX_P00519 -- (neXtProt link)	"ABL1"	"Apoptosis-related network due to altered Notch3 in ovarian cancer,Integrated Breast	"Dasatinib,Imatinib"

Going further and filling gaps

- **Annotating toxicology data?**

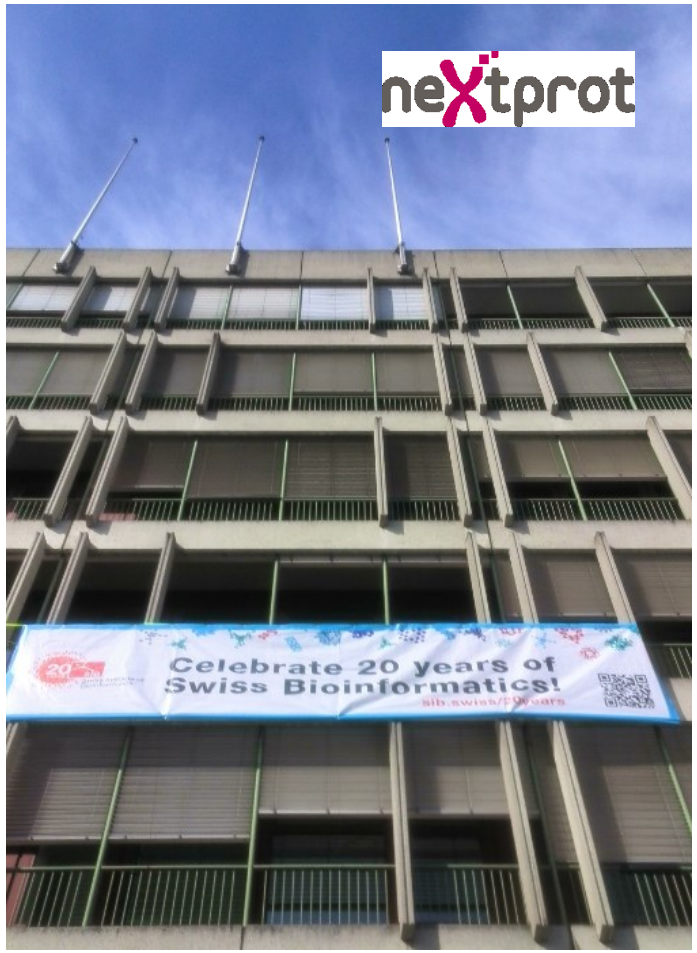
Examples: proteins that bind nanoparticles? Proteins that are differentially expressed upon nanoparticle exposure?

- **Federating with other SPARQL endpoints from OpenRisknet?**

- **Other ideas?**

Happy to discuss use cases and possible applications with you!

The current neXtprot team @ SIB - Geneva



Directors
Amos Bairoch, Lydie Lane

Biocurator
Paula Duek

Developers
Pierre-André Michel, Alain Gateau, Valentine Rech de Laval, Kasun Samarasinghe, Mathieu Schaeffer (PhD student)

Quality assurance
Monique Zahn

Contact: support@nextprot.org

Web: https://www.nextprot.org/
Twitter: @neXtProt_news
ResearchGate: neXtProt project

SWAT4HCLS

12th International SWAT4HCLS Conference

Semantic Web Applications and Tools for Health Care and Life Sciences

December 9-12, 2019
Edinburgh, Scotland

Edinburgh 2019

SWA
venu
of se

Start
expa
and

SWAT4LS intends to provide an open and vibrant discussion space cross cutting across academia, industry and institutional participants.

A one-day tutorial on 9 independent SPARQL endpoints hosted by the SIB, including neXtProt, UniProt , GlyConnect, Rhea, OrthoDB, OMA and Bgee

RCH

LINKS

About SWAT4HCLS

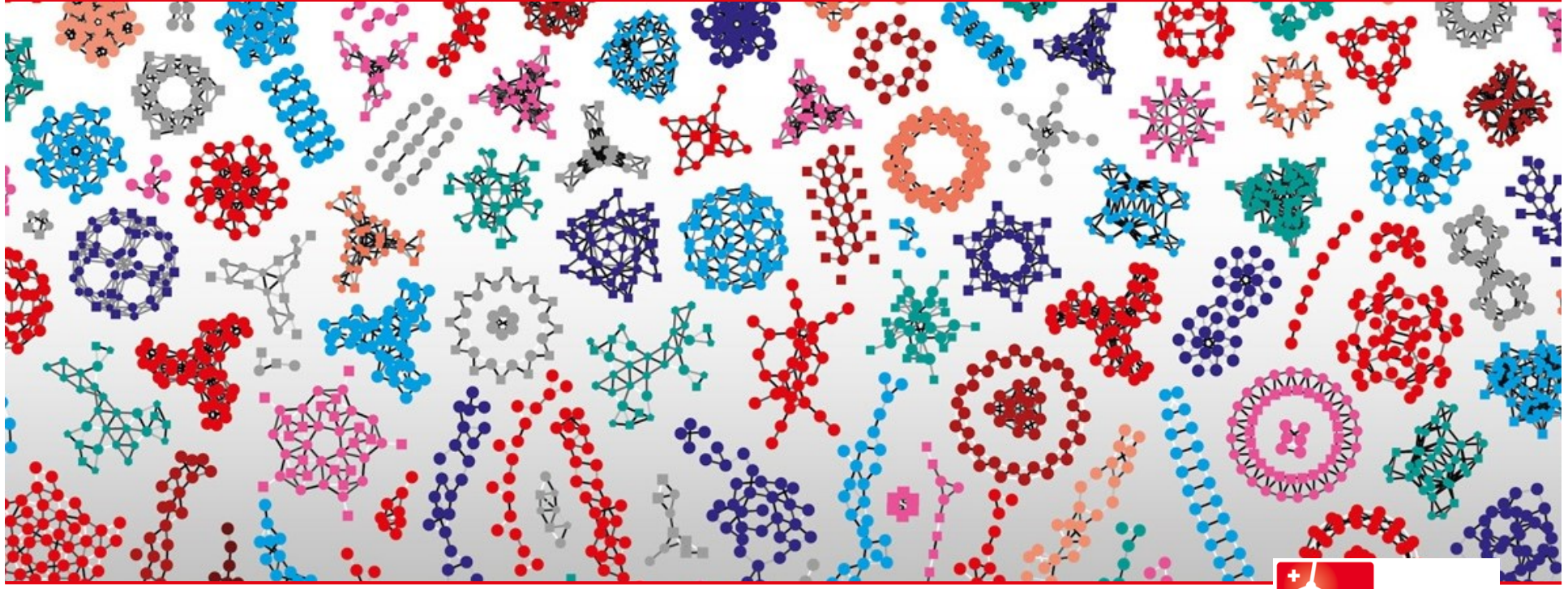
Antwerp, 2018
Rome, 2017
Amsterdam, 2016
Cambridge, 2015

SWAT4LS 2019 IS SPONSORED BY:

Read more about sponsorship opportunities.

LATEST NEWS

Updates from SWAT4HCLS



SIB
Swiss Institute of Bioinformatics

Thank you!

E-mail: lydie.lane@sib.swiss