

Facilitating protein SAR tables for peptide therapeutics

- A practical approach to protein-naming for researchers and computers

Jan Holst Jensen
CEO and Founder, Biochemfusion

biochemfusion
- Enabling biochemformatics

Bridging the gap

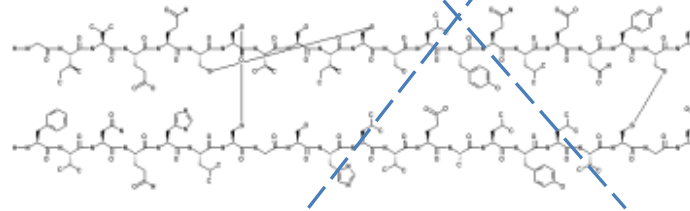
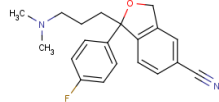
Cheminformatics

Bioinformatics

neither-nor

Molecule graphs

Sequences



```
1 GIVEQVCTSICSLYQLENYC  
21 NFNQHLCGSHLVEALYLVC  
41 GERGFFYTPKT
```

```
1 MVSQALRLLCLLLCLQGCLAACGVAKASGGETRDHPWPKG  
41 PHRVFTQEEAHCVLHRRRRANAFLELLRPGSLERFKEE  
81 QCSFHEAREIFKDAERTRLFWISYSDDGDCASSPCQNGGS  
121 CKDQLQSYICFCLPAFEGRCNCETHKDDQLICVNEGGCCEQ  
161 YCSDHTCTKRSCRCHECYSLLADGVSCTPTVEYPCGKIPI  
201 LEKRNASKPQGRIVGCKVCPRCECCPWQVLLLVNCAQLCCG  
241 TLINTIWVSAAHCFDRIKNWRNLIAVLCEHDLSEHDGDEF  
281 QSRRVAQVIIPSTYVPCTNHDIALLRLHQPVVLDHVVEF  
321 LCLPERTFSERTLAFVRFSLVSGWGCQLLDRGATALELMVL  
361 NVPRLMTQDCLQSRKVGDSPNITEYMFCAGYSDCSKDSCF  
401 KGDSGCPHATHYRGTWYLTGIVSWGCCATVGHFGVYTRV  
441 SQYIEWLQKLMRSEFRPGVLLRAPFF
```

100

10k

1M

MW
Da

Protein representation

Levels of abstraction

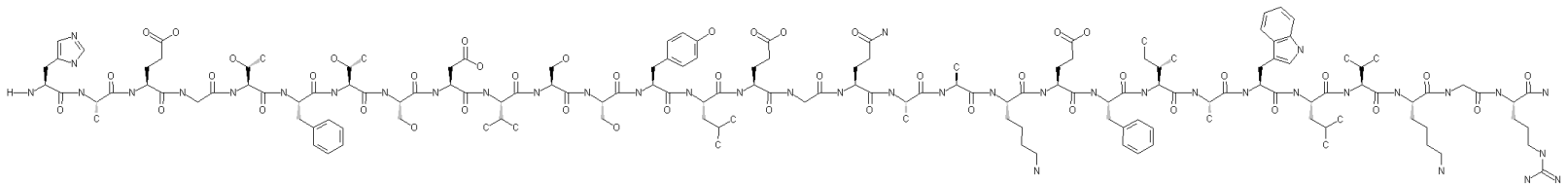
- Name

GLP-1

- Sequence

H-HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR-[NH₂]

- Atoms and bonds

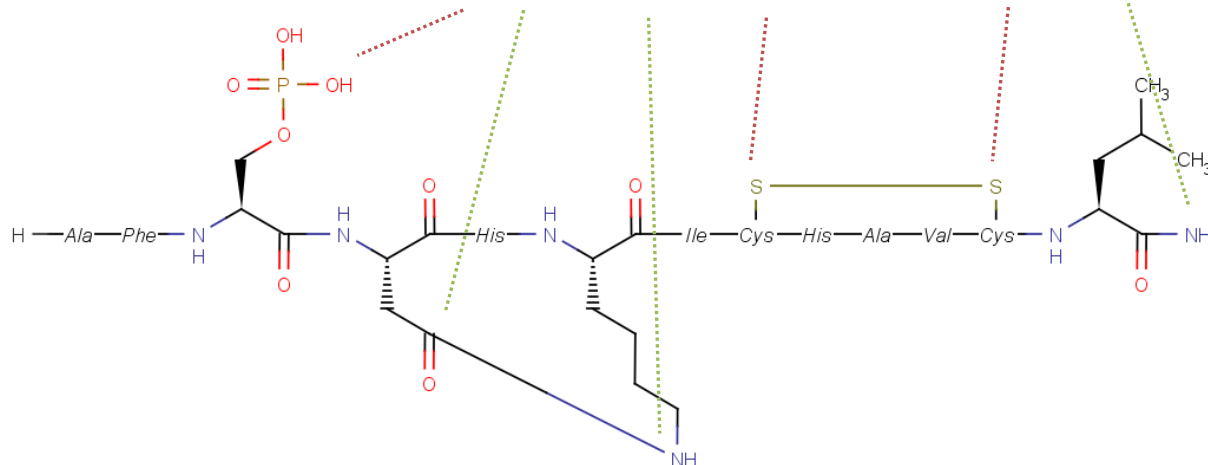


Protein Line Notation (PLN)

- Protein “SMILES” – sequence with chemical annotations

H-AF [PhSer] D (cyclo1) HK (cyclo1) IC (1) HAVC (1) L- [NH2]

1 AF S D H K I C H A V C L



DerNot expressions

- **Derivatives Notation**
 - IUPAC-like protein naming

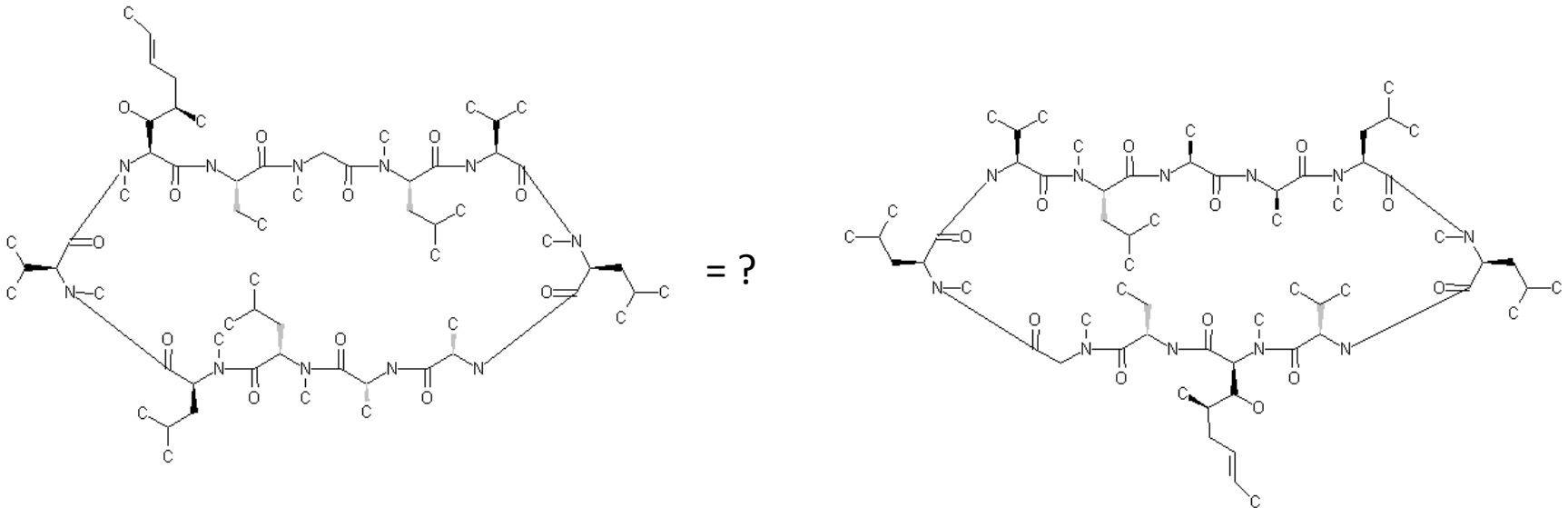
des- (B30) endo-R (B28) K (A4) A (B2) "Human Insulin"

deletions

insertions

substitutions

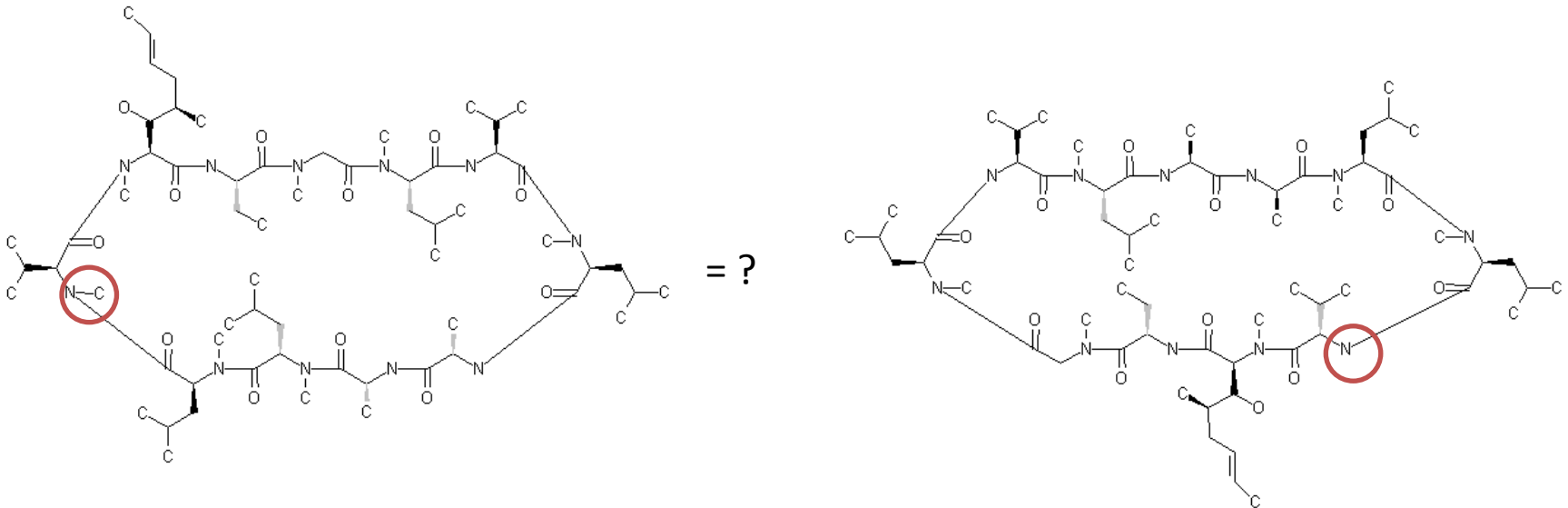
Pop-quiz – spot the difference



(cyclo)-[MeBmt]-[Abu]-[MeGly]-[MeLeu]-
Val-[MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-
[MeVal]-(cyclo) name="Cyclosporin CsA"

(cyclo)-Val-[MeLeu]-Ala-dAla-[MeLeu]-
[MeLeu]-Val-[MeBmt]-[Abu]-[MeGly]-
[MeLeu]-(cyclo)

Pop-quiz – spot the difference



(cyclo)-[MeBmt]-[Abu]-[MeGly]-[MeLeu]-
 Val [MeLeu]-Ala-dAla-[MeLeu]-[MeLeu]-
 [MeVal]-(cyclo) name="Cyclosporin CsA"

(cyclo)-Val [MeLeu]-Ala-dAla-[MeLeu]-
 [MeLeu]-Val-[MeBmt]-[Abu]-[MeGly]-
 [MeLeu]-(cyclo)

Proteax_DerNot_Diff(<right>, <left>) => V(11) "Cyclosporin CsA"

Implementing it

Technology stack



FF3.5+



Chrome



Android/iPhone



IE9+

HTML5-capable browser



Plug-ins

Apache + mod_python + pycopg2

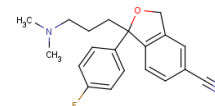
PostgreSQL

Proteax

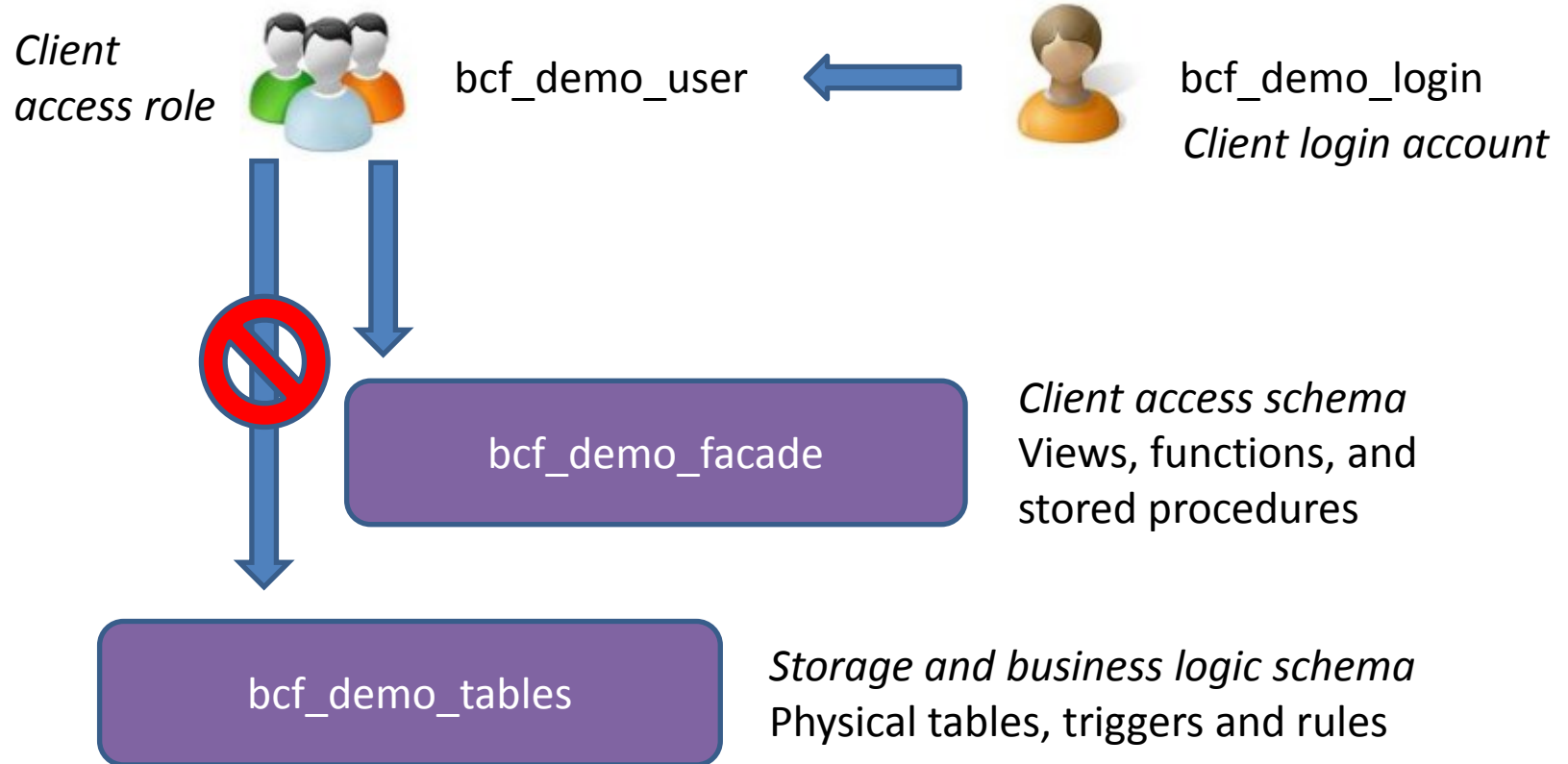
RDKit

(Proteax runs nicely
inside Oracle® too)

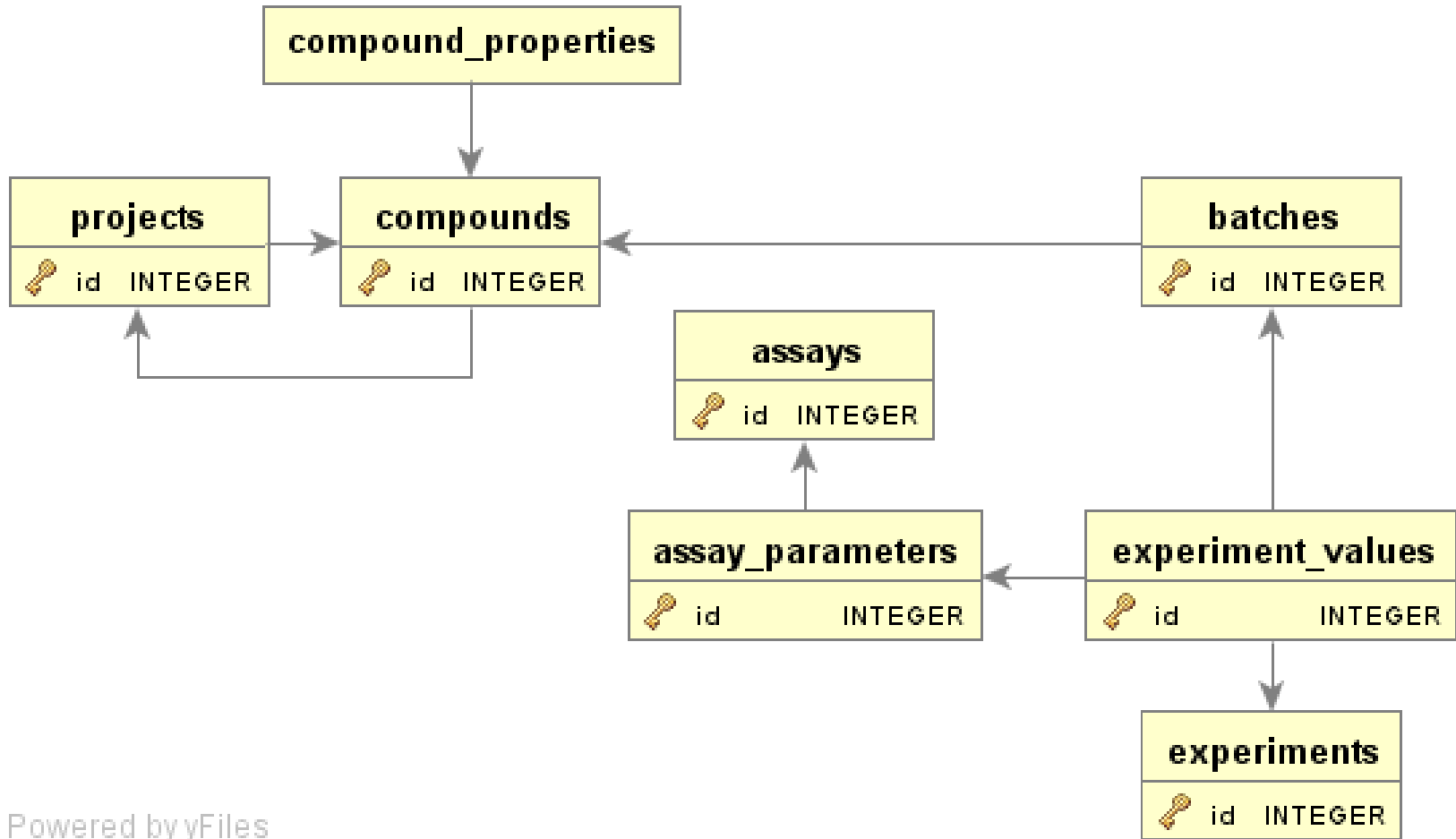
```
1  GIVEQQCTSICSLYQLENYC  
21  NFVNQHLCGSHLVEALYLVC  
41  GERGFFYTPKT
```



Database layers



The data model



Powered by yFiles

compound_properties	
compound_id	INTEGER
protein_id	TEXT
name	TEXT
render_info	TEXT
full_sequence	TEXT
expressed_sequence	TEXT
expressed_mw	DOUBLE PRECISION
expressed_fm1a	TEXT
norm_seq_chksum	TEXT
norm_prot_chksum	TEXT

projects	
id	INTEGER
name	TEXT
is_locked	CHARACTER(1)
reference_compound_id	INTEGER

compounds	
id	INTEGER
no	TEXT
project_id	INTEGER
protein_text	TEXT

batches	
id	INTEGER
compound_id	INTEGER
no	INTEGER
comments	TEXT

assays	
id	INTEGER
name	TEXT

assay_parameters	
id	INTEGER
assay_id	INTEGER
name	TEXT
unit	TEXT

experiment_values	
id	INTEGER
experiment_id	INTEGER
assay_parameter_id	INTEGER
batch_id	INTEGER
value	DOUBLE PRECISION

experiments	
id	INTEGER
name	TEXT

Using it

Selecting SAR source data

SAR report generation

The simple SAR report will compare all compounds within a given project. Only be included.

First, select a project. This will automatically select the project's reference compound of the final report is based on. You are free to choose a different reference comparison - you will then need the compound no which you can find by [brow](#)

Next, select an assay and an assay parameter. You are then ready to press t

Projects

id	Name	Ref. compound no	Ref. compound name
1	UniProt	None	None
4	GLP-1 analogues	None	None
5	Dup check demo	None	None
6	Sandbox	None	None
2	Insulins	C00001	Human insulin
3	Cyclosporins	C00007	Cyclosporin CsA

Project Ref. compound

Selecting SAR source data

+ Projects

Project Ref. compound

- Assays

id	Name	Total result count	Name	Unit	Result count
42	<u>Insulin pharmacokinetics.</u>	25	<u>Tmax</u>		25
75	<u>Cyclosporin toxicity - oral dose, rat.</u>	18			

Assay Assay parameter

```

select
  row_number() over (order by min_value, max_value) as rownum,
  proteax.name(protein_text),
  proteax.dernot_diff(protein_text, reference_protein, '*D'),
  name as parameter,
  min_value || ' - ' || max_value as value,
  unit
from (
  select
    *,
    (select protein_text from bcf_demo_tables.compounds where no = _ref_compound_no
     ) as reference_protein
  from (
    select
      cmp.id as compound_id, apar.id as assay_parameter_id,
      min(expval.value) as min_value, max(expval.value) as max_value
    from bcf_demo_tables.compounds cmp
      inner join bcf_demo_tables.projects proj on proj.id = cmp.project_id
      inner join bcf_demo_tables.batches bat on bat.compound_id = cmp.id
      inner join bcf_demo_tables.experiment_values expval
        on expval.batch_id = bat.id
      inner join bcf_demo_tables.assay_parameters apar
        on apar.id = expval.assay_parameter_id
      inner join bcf_demo_tables.assays assay on assay.id = apar.assay_id
    where proj.name = _project_name
      and assay.name = _assay_name
      and apar.name = _assay_par_name
    group by cmp.id, apar.id
  ) sartable_raw
  inner join bcf_demo_tables.compounds cmp on cmp.id = sartable_raw.compound_id
  inner join bcf_demo_tables.assay_parameters apar
    on apar.id = sartable_raw.assay_parameter_id
  ) sartable
order by min_value, max_value;

```

```

select
  row_number() over (order by min_value, max_value) as rownum,
  proteax.name(protein_text),
  proteax.dernot_diff(protein_text, reference_protein, '*D'),
  name as parameter,
  min_value || ' - ' || max_value as value,
  unit
from (
  select
    *,
    (select protein_text from bcf_demo_tables.compounds where no = _ref_compound_no
     ) as reference_protein
  from (

```

select

```

    cmp.id as compound_id, apar.id as assay_parameter_id,
    min(expval.value) as min_value,
    max(expval.value) as max_value
  from bcf_demo_tables.compounds cmp
    inner join bcf_demo_tables.projects proj on ...
    inner join bcf_demo_tables.batches bat on ...
    inner join bcf_demo_tables.experiment_values expval ...
    inner join bcf_demo_tables.assay_parameters apar ...
    inner join bcf_demo_tables.assays assay on ...
  where proj.name = _project_name
    and assay.name = _assay_name
    and apar.name = _assay_par_name
  group by cmp.id, apar.id
) satable_raw
  inner join bcf_demo_tables.compounds cmp on cmp.id = satable_raw.compound_id
  inner join bcf_demo_tables.assay_parameters apar
    on apar.id = satable_raw.assay_parameter_id
) satable
order by min_value, max_value;

```

```

select
  row_number() over (order by min_value, max_value) as rownum,
  proteax.name(protein_text),
  proteax.dernot_diff(protein_text, reference_protein, '*D'),
  name as parameter,
  min_value || ' - ' || max_value as value,
  unit
from (
  select
    *,
    (select protein_text from bcf_demo_tables.compounds
      where no = _ref_compound_no
    ) as reference_protein
  from (
select
  cmp.id as compound_id, apar.id as assay_parameter_id,
  min(expval.value) as min_value, max(expval.value) as max_value
from bcf_demo_tables.compounds cmp
  inner join bcf_demo_tables.projects proj on proj.id = cmp.project_id
  inner join bcf_demo_tables.batches bat on bat.compound_id = cmp.id
  inner join bcf_demo_tables.experiment_values expval
    on expval.batch_id = bat.id
  inner join bcf_demo_tables.assay_parameters apar
    on apar.id = expval.assay_parameter_id
  inner join bcf_demo_tables.assays assay on assay.id = apar.assay_id
where proj.name = _project_name
  and assay.name = _assay_name
  and apar.name = _assay_par_name
group by cmp.id, apar.id
    ) sartable_raw
    inner join bcf_demo_tables.compounds cmp on ...
    inner join bcf_demo_tables.assay_parameters apar ...
  ) sartable
  order by min_value, max_value;

```

select

```
row_number() over (order by min_value, max_value) as rownum,  
proteax.name(protein_text),  
proteax.dernot_diff(protein_text, reference_protein, '*D'),  
name as parameter,  
min_value || ' - ' || max_value as value,  
unit
```

from (

```
select  
*,  
(select protein_text from bcf_demo_tables.compounds where no = _ref_compound_no  
 ) as reference_protein  
from (  
  select  
    cmp.id as compound_id, apar.id as assay_parameter_id,  
    min(expval.value) as min_value, max(expval.value) as max_value  
  from bcf_demo_tables.compounds cmp  
    inner join bcf_demo_tables.projects proj on proj.id = cmp.project_id  
    inner join bcf_demo_tables.batches bat on bat.compound_id = cmp.id  
    inner join bcf_demo_tables.experiment_values expval  
      on expval.batch_id = bat.id  
    inner join bcf_demo_tables.assay_parameters apar  
      on apar.id = expval.assay_parameter_id  
    inner join bcf_demo_tables.assays assay on assay.id = apar.assay_id  
  where proj.name = _project_name  
    and assay.name = _assay_name  
    and apar.name = _assay_par_name  
  group by cmp.id, apar.id  
) sartable_raw  
  inner join bcf_demo_tables.compounds cmp on cmp.id = sartable_raw.compound_id  
  inner join bcf_demo_tables.assay_parameters apar  
    on apar.id = sartable_raw.assay_parameter_id
```

) sartable

```
order by min_value, max_value;
```

Resulting SAR table

Biochemfusion Demo DB

SAR report generation

The simple SAR report will compare all compounds within a given project. Only compounds that have associated experiment data will be included.

First, select a project. This will automatically select the project's reference compound, which is the compound that the DEPNOT_DIFF column of the final report is based on. You are free to choose a different reference compound as you like to re-base the compound comparison - you will then need the compound no which you can find by [browsing all compounds](#).

Next, select an assay and an assay parameter. You are then ready to press the "Create report" button and see the final SAR report.

Projects
Project: Insulin, Ref. compound: C00001

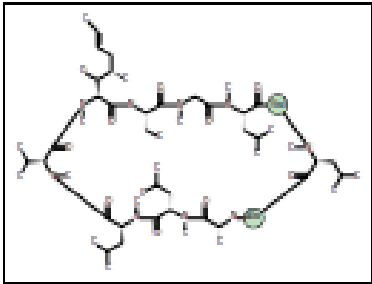
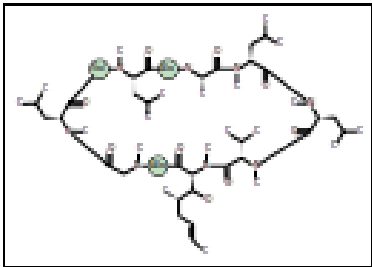
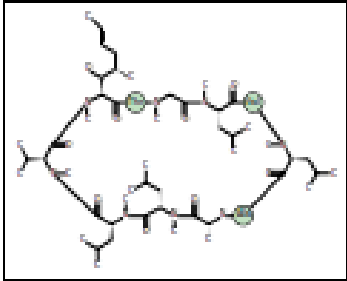
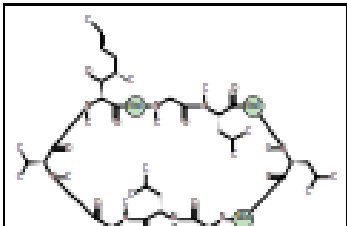
Assays
Assay: Insulin pharmacokinetic, Assay parameter: Tmax

Resulting SAR table

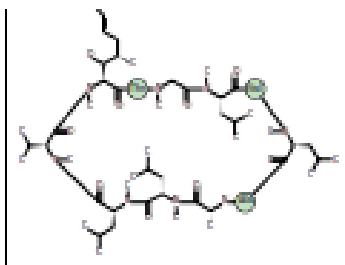
Cmp.No	Sequence	Name	DerNot diff.	Param.	Value	Unit
C00001	1 GIVEQCCTSIICSLYQIENYC 21 NFFVNOHLCGSHLVEALYLIVG 41 GERGFFYYTKPT	Human insulin	*	Tmax	1 - 3	hrs
C00002	1 GIVEQCCTSIICSLYQIENYC 21 NFFVNOHLCGSHLVEALYLIVG 41 GERGFFYYTKPT	Lispro	endo-P(B29) des-P(B28) *	Tmax	0.5 - 1.5	hrs
C00003	1 GIVEQCCTSIICSLYQIENYC 21 NFFVNOHLCGSHLVEALYLIVG 41 GERGFFYYTKPT	Aspart	D(B28) *	Tmax	1 - 1.5	hrs



Cmp.No	Sequence	Name	DerNot diff.	Param.
C00001	<pre> 1 GIVEQCCTSICSLYQLENYC 21 NFNQHLVCGSHLVEALYLVC 41 GERGFFYTPKT </pre>	Human insulin *		Tmax 1
C00002	<pre> 1 GIVEQCCTSICSLYQLENYC 21 NFNQHLVCGSHLVEALYLVC 41 GERGFFYTPKT </pre>	Lispro	endo-P(B29) des-P(B28) *	Tmax 0
C00003	<pre> 1 GIVEQCCTSICSLYQLENYC 21 NFNQHLVCGSHLVEALYLVC 41 GERGFFYTDKT </pre>	Aspart	D(B28) *	Tmax 1
C00004	<pre> 1 GIVEQCCTSICSLYQLENYC 21 NFVKQHLVCGSHLVEALYLVC 41 GERGFFYTPET </pre>	Glulisine	K(B3)E(B29) *	Tmax 1
C00005	<pre> 1 GIVEQCCTSICSLYQLENYC 21 GFVNQHLVCGSHLVEALYLVC 41 GERGFFYTPKTRR </pre>	Glargine	G(A21) * -RR-(B)	Tmax 2
C00006	<pre> 1 GIVEQCCTSICSLYQLENYC 21 NFNQHLVCGSHLVEALYLVC 41 GERGFFYTPK </pre>	Detemir	des-T(B30) [N6-C14fattyacid-lysine](B29) *	Tmax 6

Cmp.No	Molecule	Name	DerNot diff. Param.	Value	Unit
C00007		Cyclosporin CsA *	LD50	1400 - 1510	mg/kg
C00008		Cyclosporin CsB A(2) *	LD50	1600 - 1820	mg/kg
C00009		Cyclosporin CsC T(2) *	LD50	800 - 1200	mg/kg
C00010		Cyclosporin CsD V(2) *	LD50	990 - 1200	mg/kg

C00009



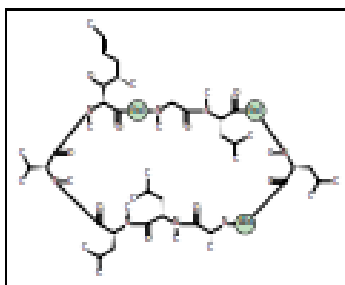
Cyclosporin CsC T(2) *

LD50

800 - 1200

mg/kg

C00010



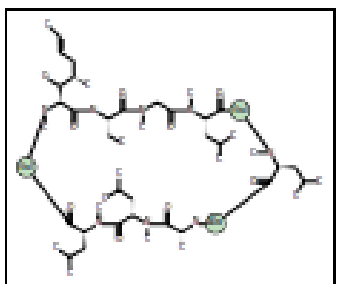
Cyclosporin CsD V(2) *

LD50

990 - 1200

mg/kg

C00011



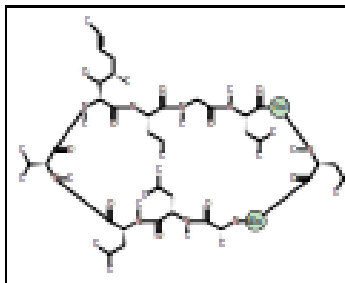
Cyclosporin CsE V(11) *

LD50

1240 - 1500

mg/kg

C00012



Cyclosporin CsG [Nva](2) *

LD50

2500 - 3005

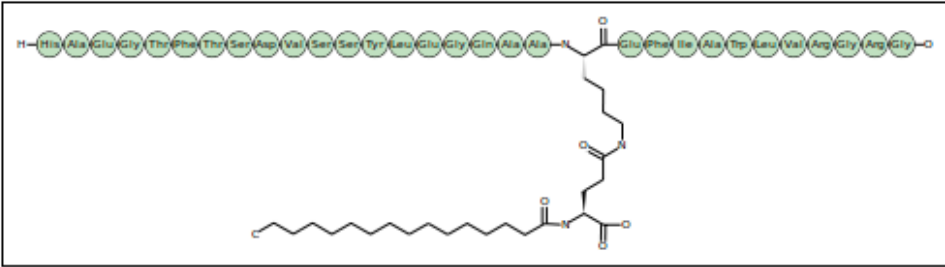
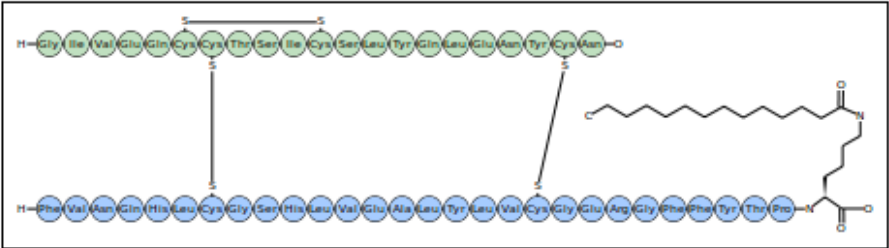
mg/kg

SSS search results

(SSS query pane collapsed)

Query compounds

☒ - SSS query pane

Compound no.	Project	Compound name	Molecule
C00015	GLP-1 analogues	Liraglutide according to http://www.ama-assn.org/ama1/pub/upload/mm/365/liraglutide.pdf	 The image shows the chemical structure of Liraglutide. It consists of a linear peptide chain of 30 amino acids: H-His-Ala-Gln-Gly-Thr-Phe-Thr-Ser-Met-Val-Ser-Ser-Tyr-Ile-His-Pro-Phe-Gln-NH ₂ . The C-terminus is modified with a long-chain fatty acid (C16) attached via a thioether linkage to the side chain of the 26th residue, Lysine. The structure is shown in a 2D representation with the peptide backbone as a series of circles and the fatty acid chain as a zigzag line.
C00006	Insulins	Detemir	 The image shows the chemical structure of Detemir, a long-acting insulin analog. It features two polypeptide chains: an A-chain (21 residues: H-Phe-Val-Met-Gln-His-Leu-Cys-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cys-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-NH ₂) and a B-chain (30 residues: H-Gly-Ile-Val-Gln-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Lys-Tyr-Gln-Leu-Gln-Met-Tyr-Cys-Asn). The chains are linked by two inter-chain disulfide bridges (A6-B7 and A11-B20) and one intra-chain disulfide bridge in the A-chain (A6-A11). The C-terminus of the B-chain is modified with a long-chain fatty acid (C18) attached via a thioether linkage to the side chain of the 29th residue, Lysine. The structure is shown in a 2D representation with the peptide backbone as a series of circles and the fatty acid chain as a zigzag line.

Acknowledgements

- Thanks for great discussions
 - Thomas Dörner, Independent consultant
 - Gerd Blanke, StructurePendium GmbH
- "I now declare this demo database opened"
 - http://www.proteax.dk/demo_db/

(my humble apologies to Mr. Winterbottom)