CAS SciFinder^{n}

“Search yourself” guide for Pharma industry

Please let us know if you have any questions

Sergey Zubkevich, szubkevich@acs-i.org

Paul Peters, ppeters@acs-i.org

Contents

I. Predictive Retrosynthesis

Generic Nodes

In a lot of predicted retrosynthesis plans you will see the following generic nodes:

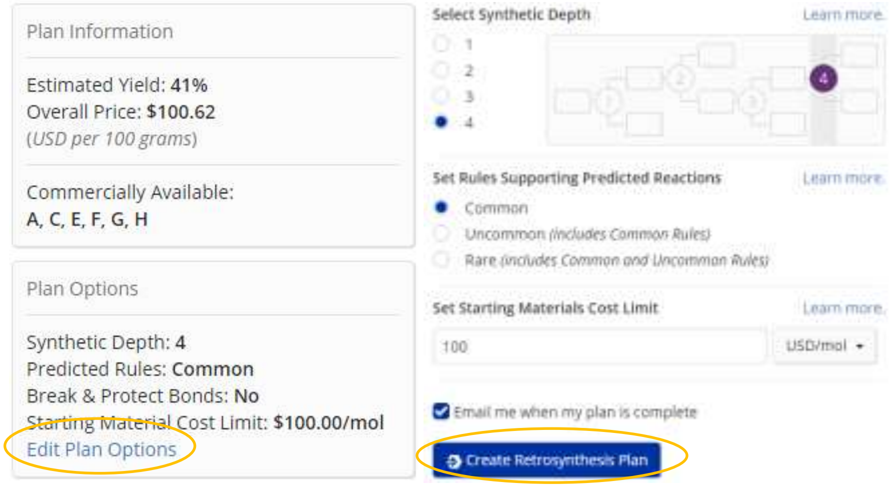
- Cx: Cl, Br or I
- Bx: Br or I
- X: F, Cl, Br or I

These generic nodes will be substituted with the most common corresponding halogen for further calculation or display when you click on the substance.

The differentiation between these nodes is important, because the different halogens can behave very differently, depending on the reaction type.

Accessing Following Plans from Your Retrosynthesis History

In order to reproduce the following retrosynthetic schemes, please following the following steps:



Plan Information

Estimated Yield: **41%**
Overall Price: **$100.62**
(USD per 100 grams)

Commercially Available:
A, C, E, F, G, H

Plan Options

Synthetic Depth: **4**
Predicted Rules: **Common**
Break & Protect Bonds: **No**
Starting Material Cost Limit: **$100.00/mol**

Select Synthetic Depth
○ 1
○ 2
○ 3
● 4

Set Rules Supporting Predicted Reactions
● Common
○ Uncommon *(includes Common Rules)*
○ Rare *(includes Common and Uncommon Rules)*

Set Starting Materials Cost Limit
100 USD/mol

☑ Email me when my plan is complete

1) Click on the link provided in this document
2) When the Retrosynthesis scheme opens, click on “Edit Plan Options”
3) When the plan options open, click on “Create Retrosynthesis Plan”
4) The plan will now be in your history. It can also be saved and given a name

Alternatively, draw the structures by their CAS REGISTRY Numbers and start the retrosynthesis yourself.

When to Use Uncommon or Rare Rules

The standard options for the calculation of predicted retrosynthesis plans use the set of common rules. Rules describing preparations of more complex structures, that by nature have less evidence, e.g. heterocycles or polycyclic moieties are triggered within the common rule set. In some cases, you will see that those are not sufficient to retrieve satisfying results. In those cases you can try the calculation with rare rules.

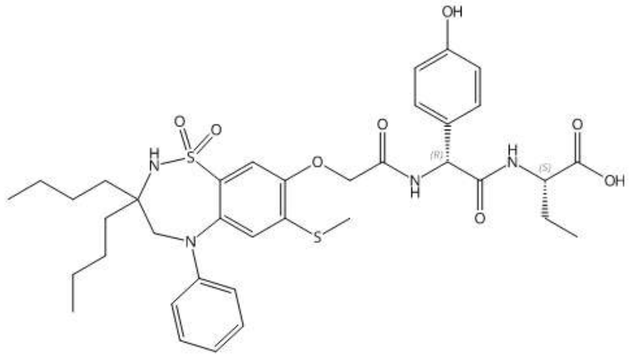
Change the plan options to uncommon or rare rules when you see

- Only small changes within a large fragment that seems to be inefficient regarding complexity reduction
- Disconnections of small side groups, though you think that the molecule could be split into more equally sized pieces
- Functional group interchanges like movement of a double bond that don’t lead to further disconnections
- When the plan stops after only 1 step and other alternatives don’t provide better alternatives

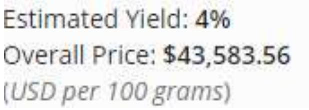
Due to the smart rule selection it is recommended to calculate the plan with common rules. In most cases you will find good plans already. If not, go ahead and calculate the plan again with uncommon or rare rules. Please note, that you can get very long list of alternatives with rare rules. Grouping alternatives is currently under development (as of 08/06/2021).

There are some classes of compounds that would allow you to set up a predictive retrosynthetic plan but it would not give very meaningful results, not even with rare rules. These would include coordination compounds, cyclic peptides, large fused ring systems and radioisotopically labelled compounds.

How to customize retrosynthesis plan

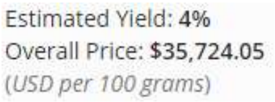


When we try and create a retrosynthesis plan with rare rules for **Odevixibat** - a newly approved medication for the treatment of progressive familial intrahepatic cholestasis (PFIC), we receive very high cost estimations and very low yield, based on the reported reactions.



Estimated Yield: **4%**
Overall Price: **$43,583.56**
(USD per 100 grams)

Original plan: [SciFinder^n Retrosynthesis Plan (cas.org)]

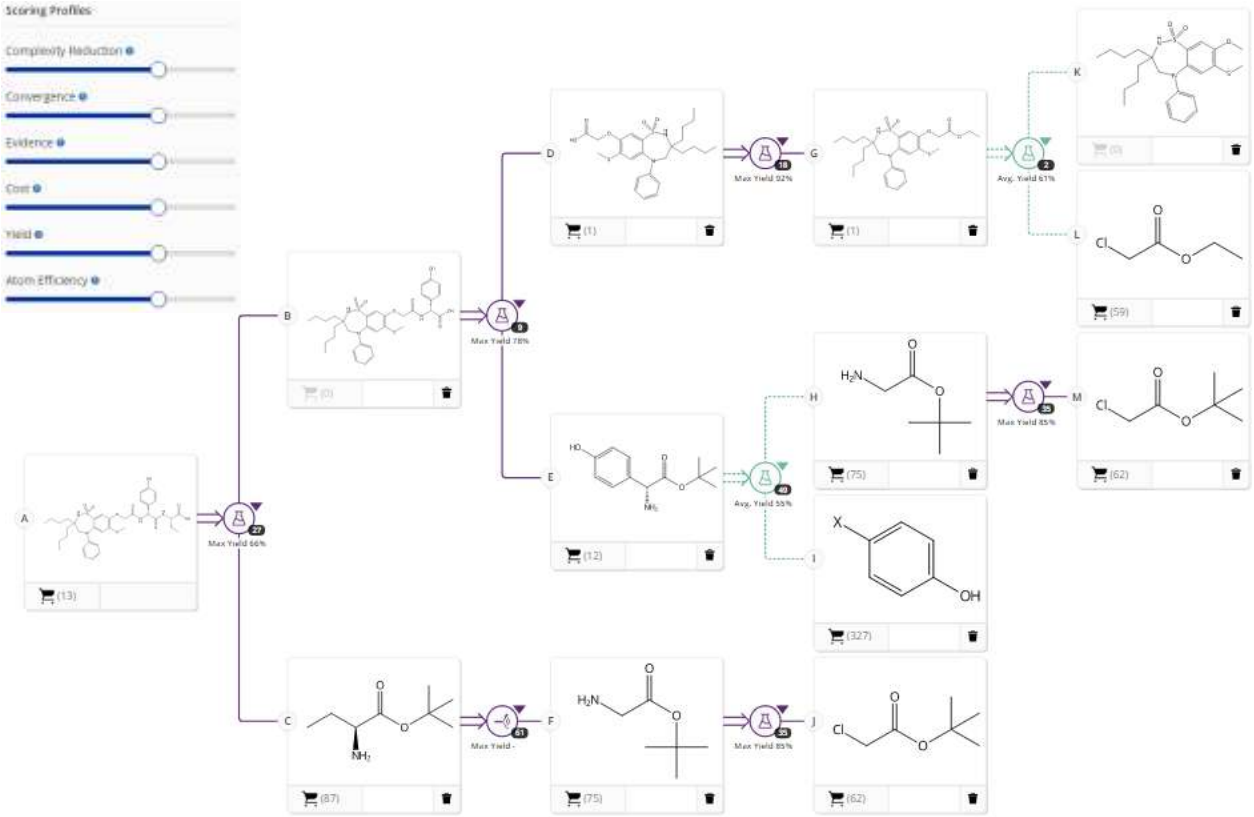


Estimated Yield: **4%**
Overall Price: **$35,724.05**
(USD per 100 grams)

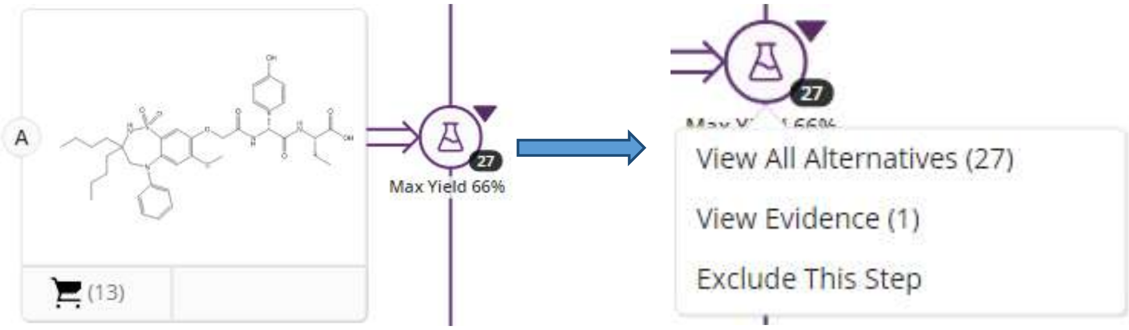
We can slightly improve the results by enabling the predicted steps in our plan.

Predicted Results ON

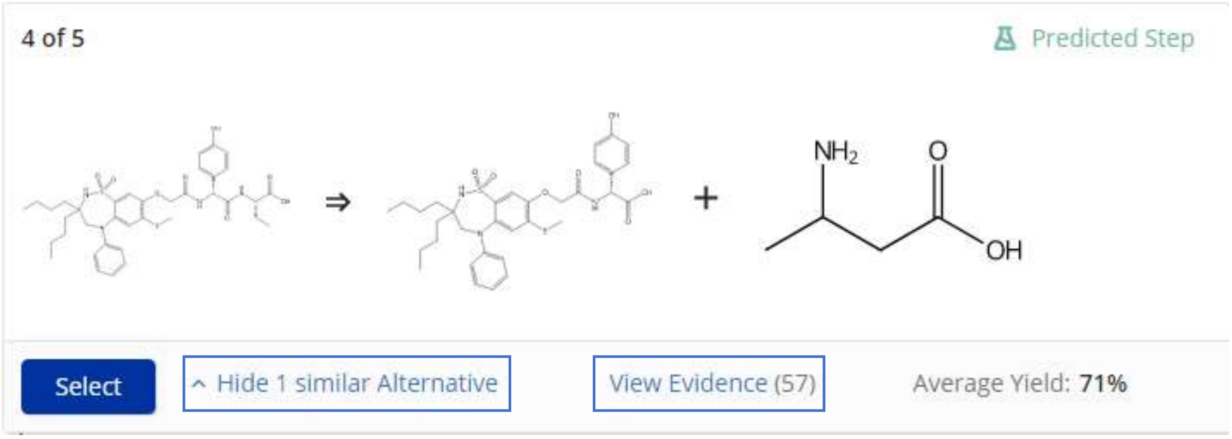
Scoring profiles are not so useful in this case, since we have a major part of experimental steps that are considered preferential by the scoring mechanism. But we can modify the plan manually to achieve the desired results.



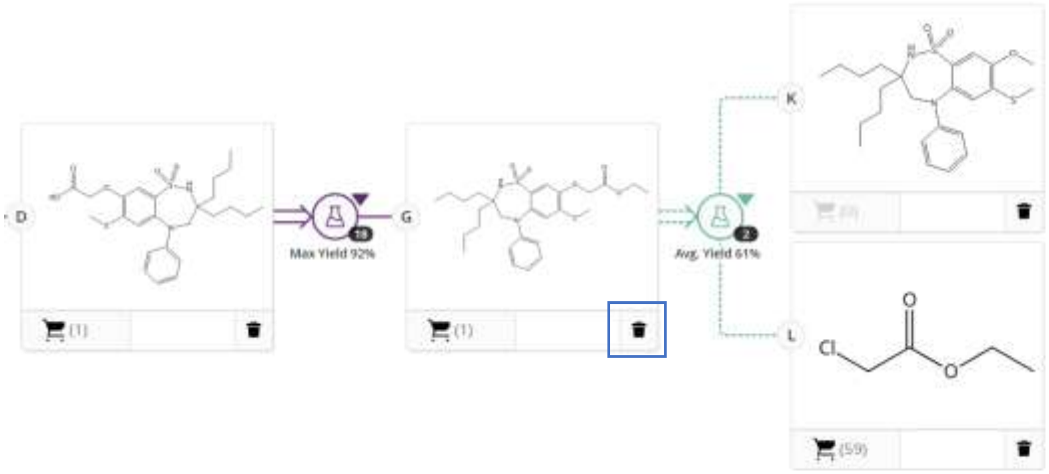
It seems that the first disconnection that leads to the final product provides rather low yield, so we can try and alter it by clicking on the step with left mouse button and browsing the available alternatives.



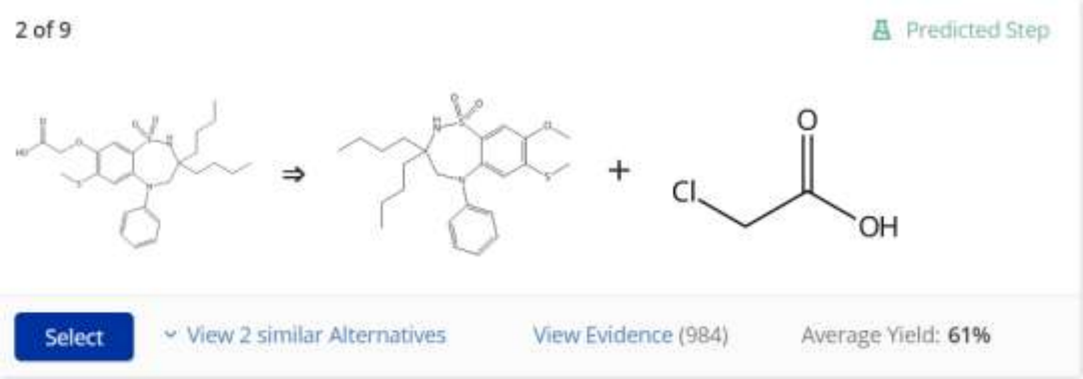
You can notice that with the latest update the alternatives are grouped by similarity, so you don't have to browse through all 27, but instead you have only 5 groups of reactions. From the first impression, the top alternative from group 4 can seem an ideal solution as it increases the yield and reduces the complexity of amino-acid intermediate. However, if you check the Evidence reactions behind this rule, you can see that none of them provides the desired stereoselectivity.



But we can alter the D -> G -> K + L row of disconnections since they provide unnecessary low complexity reduction and can be further modified. We can try and remove the two experimental intermediates (use delete function 2 times) for G to indicate the system that we want other alternatives.

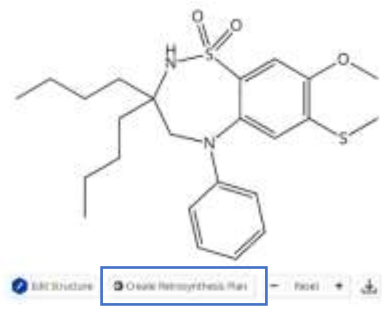


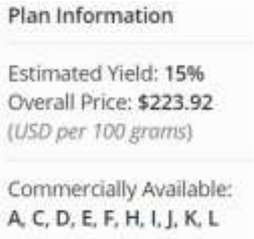
Still the result isn't sufficient and we can manually select the following predicted alternative step for D -> G disconnection as it covers all the above-mentioned transformations in one step. Moreover, it is supported with a significant number of relevant evidence reactions.



Average Yield: 61%

Now we have a synthetic plan with sufficiently better cost and yield estimation. However, there still is room for further improvement, for example, we have starting compound G, that is not commercially available and therefore is not counted towards the final cost. So, it will be a good option to create the second retrosynthesis plan for this compound to further evaluate our options.




Plan Information

Estimated Yield: 15%
Overall Price: $223.92
(USD per 100 grams)

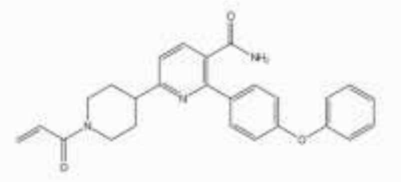
Commercially Available:
A, C, D, E, F, H, I, J, K, L

Final plan: SciFinder^n Retrosynthesis Plan (cas.org)

Examples

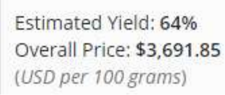
You can browse and study the selected examples or even try and beat our best answers in selecting the most effective retrosynthetic approach in term of target substance cost.

I. **Orelabrutinib** (CAS Registry Number 1655504-04-3)



Background: Orelabrutinib (®) is an orally administered, potent, irreversible and highly selective BTK-inhibitor being developed by InnoCare Pharma for the treatment of B cell malignancies and autoimmune diseases. In December 2020, orelabrutinib received its first approval in China for the treatment of patients with mantle cell lymphoma (MCL) or chronic lymphocytic leukaemia (CLL)/small lymphocytic lymphoma (SLL), who have received at least one treatment in the past.

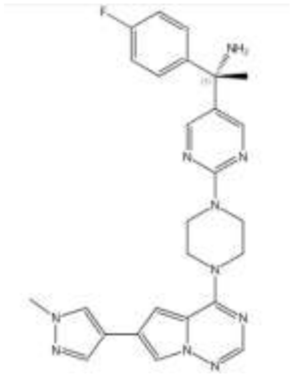
Retrosynthesis tips: Original plan in the absence of predicted steps provides a scheme with high overall cost. However, you can reduce it, enabling predicted steps and changing scoring profiles or editing the steps manually.


Estimated Yield: 64%
Overall Price: $3,691.85
(USD per 100 grams)

Original plan: SciFinder^n Retrosynthesis Plan (cas.org)

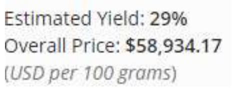
Best answer: SciFinder^{n} Retrosynthesis Plan (cas.org)

II. **Avapritinib** (CAS Registry Number 1703793-34-3)



Background: Avapritinib, sold under the brand name Ayvakit among others, is a medication used for the treatment of advanced systemic mastocytosis and for the treatment of tumors due to one specific rare mutation: it is specifically intended for adults with unresectable or metastatic gastrointestinal stromal tumor (GIST) that harbor a platelet-derived growth factor receptor alpha (PDGFRA) exon 18 mutation. Avapritinib is an orally bioavailable inhibitor of specific mutated forms of platelet-derived growth factor receptor alpha (PDGFR alpha; PDGFRa) and mast/stem cell factor receptor c-Kit (SCFR), with potential antineoplastic activity. Upon oral administration, avapritinib specifically binds to and inhibits specific mutant forms of PDGFRa and c-Kit, including the PDGFRa D842V mutant and various KIT exon 17 mutants. This results in the inhibition of PDGFRa- and c-Kit-mediated signal transduction pathways and the inhibition of proliferation in tumor cells that express these PDGFRa and c-Kit mutants. PDGFRa and c-Kit, protein tyrosine kinases and tumor-associated antigens (TAAs), are mutated in various tumor cell types; they play key roles in the regulation of cellular proliferation.

Retrosynthesis tips: Initial retrosynthesis plan provides unrealistic cost estimation, but it can be reduced along with the complexity of the synthetic pathway, by adjusting scoring profiles and then manually selecting predicted alternative on second disconnection.

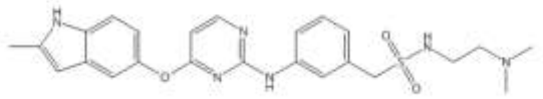


Estimated Yield: 29%
Overall Price: $58,934.17
(*USD per 100 grams*)

Original plan: SciFinder^{n} Retrosynthesis Plan (cas.org)

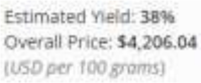
Best answer: SciFinder^{n} Retrosynthesis Plan (cas.org)

III. **Sulfatinib** (CAS Registry Number 1308672-74-3)



Background: An orally bioavailable, small molecule inhibitor of vascular endothelial growth factor receptors (VEGFR) 1, 2, and 3, and the fibroblast growth factor receptor type 1 (FGFR1), with potential antineoplastic and anti-angiogenic activities. Upon oral administration, sulfatinib binds to and inhibits VEGFRs and FGFR1 thereby inhibiting VEGFR- and FGFR1-mediated signal transduction pathways. This leads to a reduction of angiogenesis and tumor cell proliferation in VEGFR/FGFR1-overexpressing tumor cells. Expression of VEGFRs and FGFR1 may be upregulated in a variety of tumor cell types.

Retrosynthesis tips: Experimental retrosynthesis plan provides too high cost estimation. However, enabling predicted options significantly extends the plan and thus reduces the cost of target substance. It can be further reduced by altering the 1st disconnection.



Estimated Yield: 38%
Overall Price: $4,206.04
(*USD per 100 grams*)

Original plan: SciFinder^{n} Retrosynthesis Plan (cas.org)

Best answer: SciFinder^{n} Retrosynthesis Plan (cas.org)

II. CAS Formulus

Approaches to formulation search

You can search for formulations in Sci-Finder using three different approaches:

1) Search relevant publications in Sci-Finder and then limit them using CAS solutions filter -> CAS Formulus.
2) Search for specific ingredient in CAS Formulus and then retrieve all formulations for this compound or browse the **“Commonly formulated with”** section
3) Use text search field to find relevant formulations or create a complex query with advanced search option.

Ingredients

Formulations

Then you can always refine the list of formulations using available filters.

Advanced formulation searching – a case study.

You can familiarize yourself with CAS Formulus functions using this case. Simply recreate all consecutive steps that are listed below and get a nice overview of functions and filters.

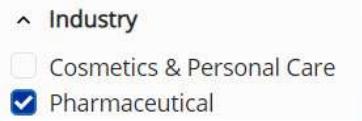
Goal: Detailed formulations with insulin used as antidiabetic agents, suitable for oral drug delivery.

Pathway 1. Most comprehensive approach.

Step 1. Perform a simple text search for **insulin** using field.

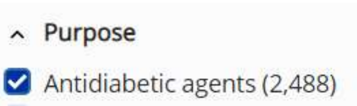
Formulations

Step 2. Then you can consecutively limit you result set by using **industry**, **purpose** and **delivery route** filters on the left:

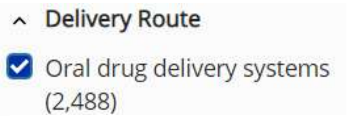

^ Industry

□ Cosmetics & Personal Care

☑ Pharmaceutical

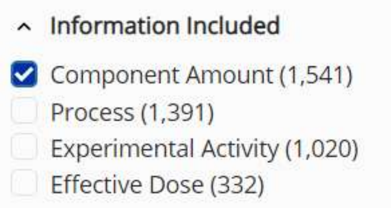

^ Purpose

☑ Antidiabetic agents (2,488)


^ Delivery Route

☑ Oral drug delivery systems (2,488)

Step 3. And then determine the specificity of the data using **information included** and **document type** filters. In this case we decided that formulations with detailed components amount will suffice, but you can select any detail level you prefer. We also decided not to include the information extracted from claims (since it is usually not so comprehensive) and comparative examples (since it is not related directly to the invention) of the patents.

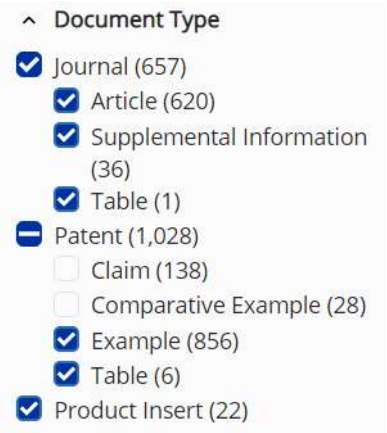

^ Information Included

☑ Component Amount (1,541)

□ Process (1,391)

□ Experimental Activity (1,020)

□ Effective Dose (332)


^ Document Type

☑ Journal (657)

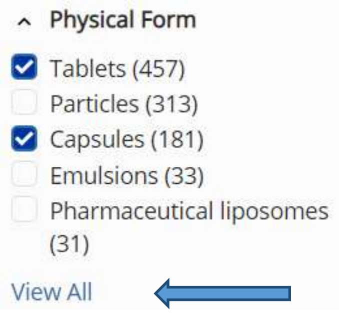
- ☑ Article (620)
- ☑ Supplemental Information (36)
- ☑ Table (1)

⊟ Patent (1,028)

- □ Claim (138)
- □ Comparative Example (28)
- ☑ Example (856)
- ☑ Table (6)

☑ Product Insert (22)

Step 4. You also may need to limit your formulations to specific physical form, using the respective filter.



Physical Form

☑ Tablets (457)

□ Particles (313)

☑ Capsules (181)

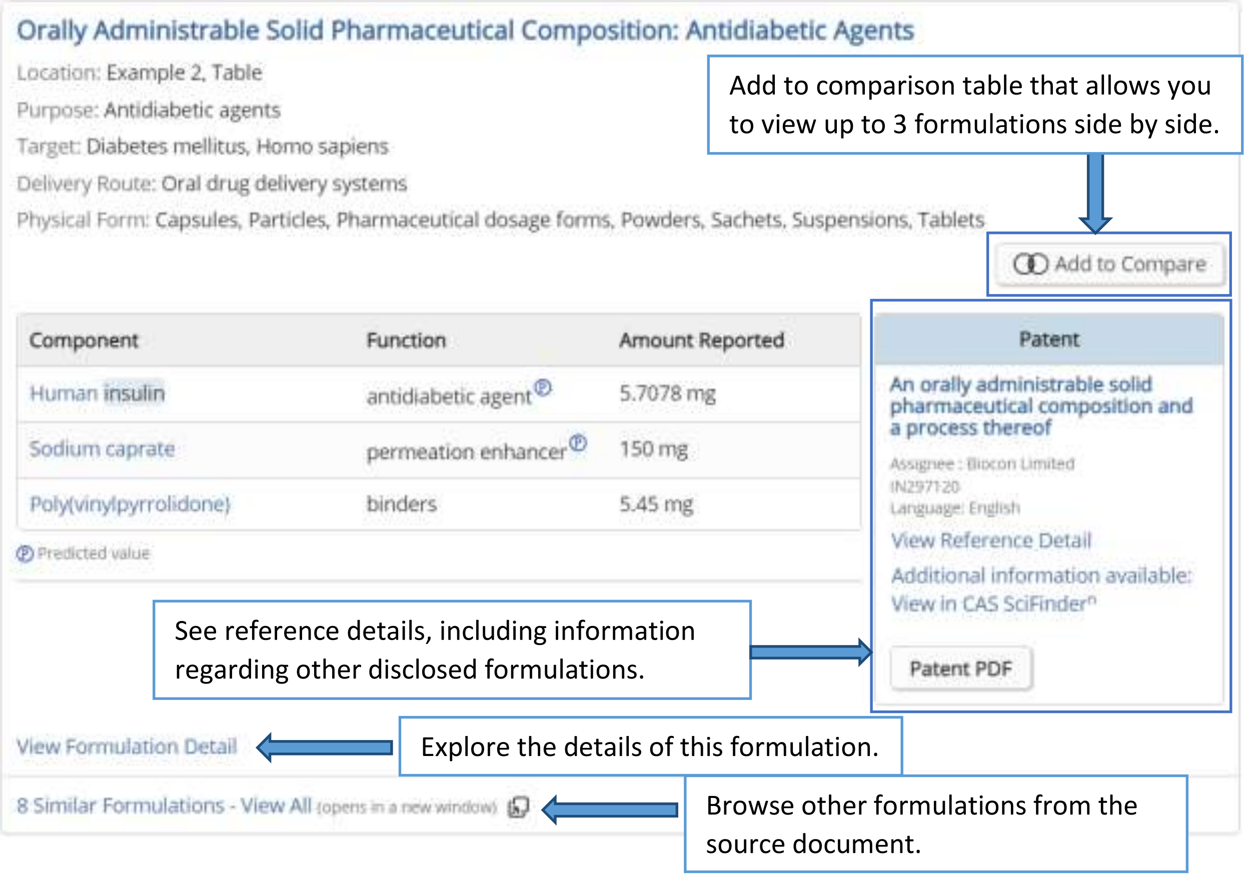
□ Emulsions (33)

□ Pharmaceutical liposomes (31)

View All

do not forget that many other forms are available here

Step 5. Now you can select the formulation of interest and browse the available details.



Orally Administrable Solid Pharmaceutical Composition: Antidiabetic Agents

Location: Example 2, Table

Purpose: Antidiabetic agents

Target: Diabetes mellitus, Homo sapiens

Delivery Route: Oral drug delivery systems

Physical Form: Capsules, Particles, Pharmaceutical dosage forms, Powders, Sachets, Suspensions, Tablets

Add to comparison table that allows you to view up to 3 formulations side by side.

Add to Compare

Component	Function	Amount Reported
Human insulin	antidiabetic agent Ⓟ	5.7078 mg
Sodium caprate	permeation enhancer Ⓟ	150 mg
Poly(vinylpyrrolidone)	binders	5.45 mg

Ⓟ Predicted value

Patent

An orally administrable solid pharmaceutical composition and a process thereof

Assignee : Biocon Limited

IN297120

Language: English

View Reference Detail

Additional information available: View in CAS SciFinderⁿ

Patent PDF

See reference details, including information regarding other disclosed formulations.

View Formulation Detail

Explore the details of this formulation.

8 Similar Formulations - View All (opens in a new window)

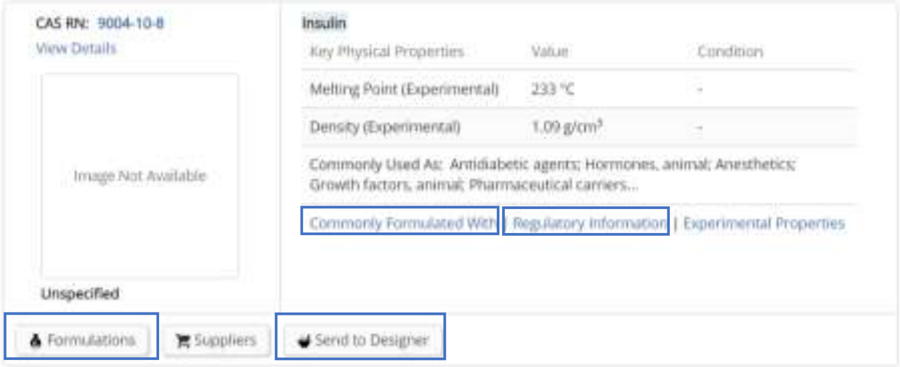
Browse other formulations from the source document.

Pathway 2. If you want to browse available ingredients and then select only one – most relevant for you search, then you can start from individual ingredient and then proceed to formulations with it.

Step 1. Search for **insulin** using

field.

You will get 4 individual ingredients – insulin, porcine insulin, bovine insulin and human insulin. You can then click of **Formulations** button to retrieve all compositions with this ingredient or browse additional info on this compound, other substances it is usually formulated with, details on regulatory information and inventory lists or send it to formulation designer.



CAS RN: 9004-10-8
View Details

Image Not Available

Unspecified

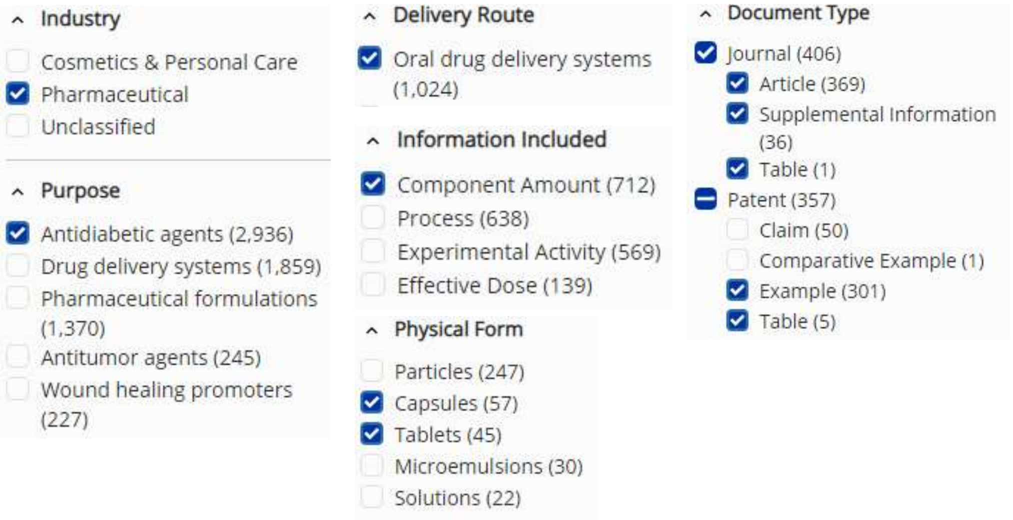
Insulin

Key Physical Properties	Value	Condition
Melting Point (Experimental)	233 °C	-
Density (Experimental)	1.09 g/cm³	-

Commonly Used As: Antidiabetic agents; Hormones, animal; Anesthetics; Growth factors, animal; Pharmaceutical carriers...

Commonly Formulated With | Regulatory Information | Experimental Properties

Step 2. Refine your results.



Industry
□ Cosmetics & Personal Care
☑ Pharmaceutical
□ Unclassified

Purpose
☑ Antidiabetic agents (2,936)
□ Drug delivery systems (1,859)
□ Pharmaceutical formulations (1,370)
□ Antitumor agents (245)
□ Wound healing promoters (227)

Delivery Route
☑ Oral drug delivery systems (1,024)

Information Included
☑ Component Amount (712)
□ Process (638)
□ Experimental Activity (569)
□ Effective Dose (139)

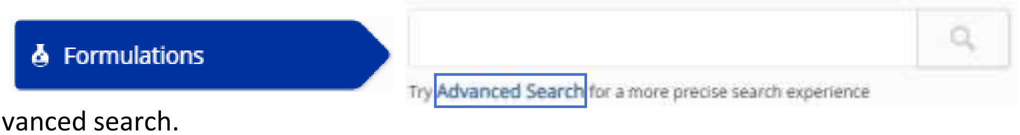
Physical Form
□ Particles (247)
☑ Capsules (57)
☑ Tablets (45)
□ Microemulsions (30)
□ Solutions (22)

Document Type
☑ Journal (406)
☑ Article (369)
☑ Supplemental Information (36)
☑ Table (1)
⊟ Patent (357)
□ Claim (50)
□ Comparative Example (1)
☑ Example (301)
☑ Table (5)

Step 3. Browse and compare formulations.

Pathway 3. Informed advanced search.

Step 1.

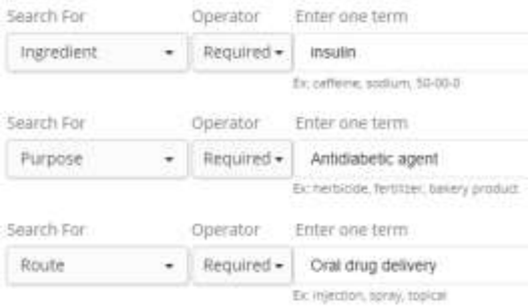


Formulations

Try Advanced Search for a more precise search experience

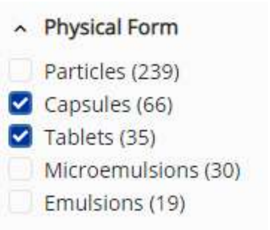
Go to advanced search.

Step 2. Create a detailed query, covering ingredients, purpose and delivery routes.



Search For	Operator	Enter one term
Ingredient	Required	insulin Ex: caffeine, sodium, 50-00-0
Purpose	Required	Antidiabetic agent Ex: herbicide, fertilizer, bakery product
Route	Required	Oral drug delivery Ex: injection, spray, topical

Step 3. Refine the results using filters on the left, for example by selecting physical form and the required information to be present in formulations:



Physical Form

- □ Particles (239)
- ☑ Capsules (66)
- ☑ Tablets (35)
- □ Microemulsions (30)
- □ Emulsions (19)

Step 4. Browse and compare formulations.

Examples for practice

You can try various search strategies in CAS Formulus using the examples mentioned below.

Example 1.

Goal: Search for **fluvoxamine**-containing formulations used as antidepressants or antipsychotics.

Note: there are two forms of fluvoxamine registered in CAS Formulus.

Example 2.

Goal: Search for antibacterial compositions containing **cefotaxime** and **sulbactam** that can be used as injections in humans.

Note1: you can limit some parameters only using advanced search.

Note2: you can also start with the ingredient search for **sulbactam** and then go to commonly formulated with, sort by active ingredient and find **cefotaxime**.

Example 3.

Goal: Search for injection formulations containing **cabotegravir** and **rilpivirine**, used as anti-HIV treatment.

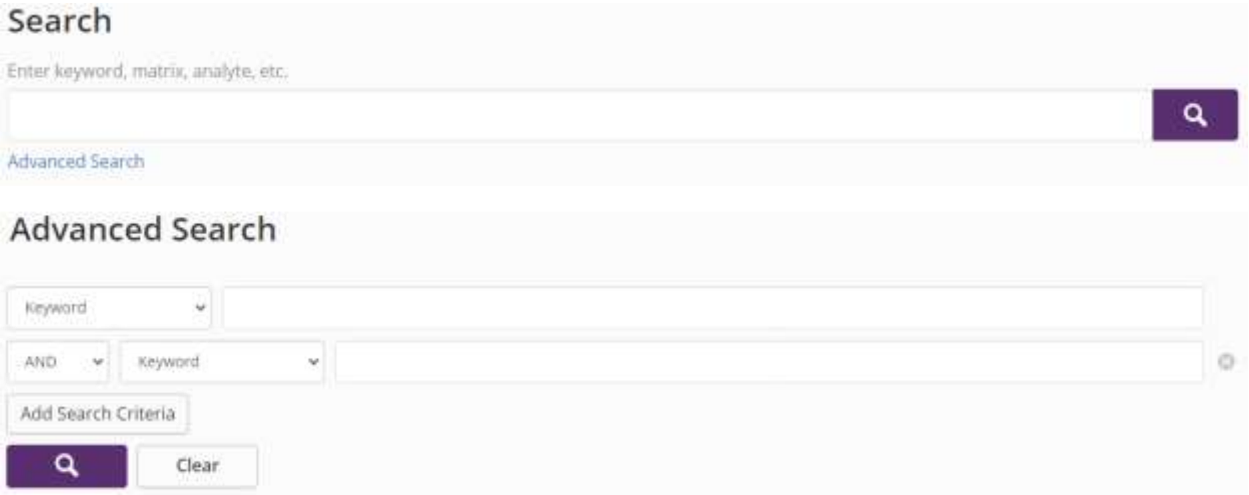
Formulation designer

The Formulation Designer, which is available from the main search page, is an option to create an editable template for a specific composition based on a few initial choices. It uses the CAS databases to provide the most relevant components and compositions based on your selection of formulation area, formulation purpose and physical form. You can try it on your own using examples mentioned above or your own queries to understand whether it is useful for your research purposes.

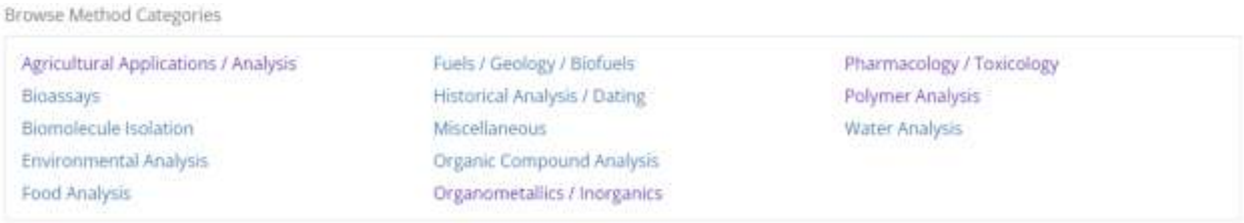
III. CAS Analytical methods

You can perform two different type of searches in CAS Analytical methods:

1) Keyword search or advanced keyword search that allow you to determine the precision of your query and thus retrieve the most relevant results.



2) Browsing by method categories that allows you to get a broad impression of the available content. You can further refine you search using available filters to achieve the required precision of the results.



Browse Method Categories

Agricultural Applications / Analysis
Bioassays
Biomolecule Isolation
Environmental Analysis
Food Analysis

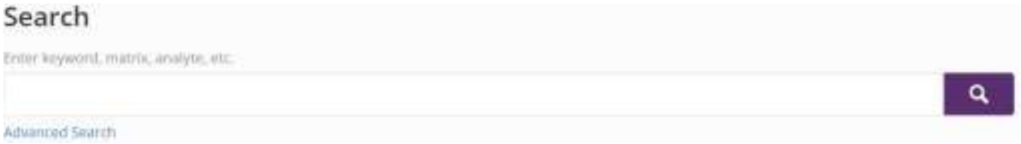
Fuels / Geology / Biofuels
Historical Analysis / Dating
Miscellaneous
Organic Compound Analysis
Organometallics / Inorganics

Pharmacology / Toxicology
Polymer Analysis
Water Analysis

Analyzing analytical techniques – a case study.

If we are interested in methods for analysis of levocetirizine in pharmaceutical tablets we can create one of the following queries:

✓ **Query 1:** Search **levocetirizine in pharmaceutical tablets** or **levocetirizine and pharmaceutical and tablets** in the query box.

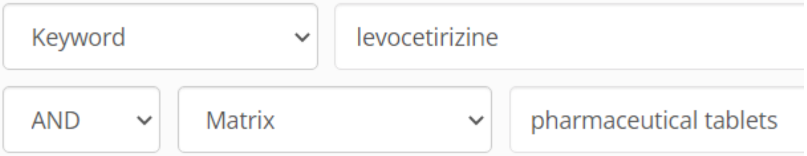


Search

Enter keyword, matrix, analyte, etc.

Advanced Search

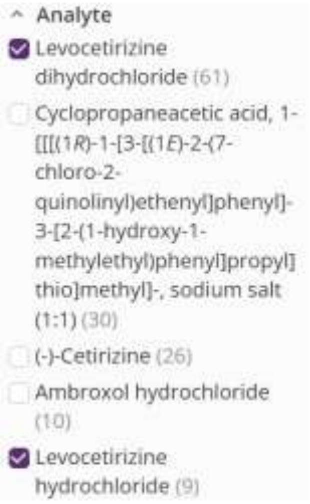
✓ **Query 2:** Go to the advanced search and create a more precise query.



Keyword | levocetirizine

AND | Matrix | pharmaceutical tablets

In this example we search levocetirizine as keyword to cover all possible analytes like levocetirizine hydrochloride or levocetirizine dihydrochloride. However we still receive some noise in the analyte section. Alternatively we can construct a query that will provide us with exact answers.



Analyte

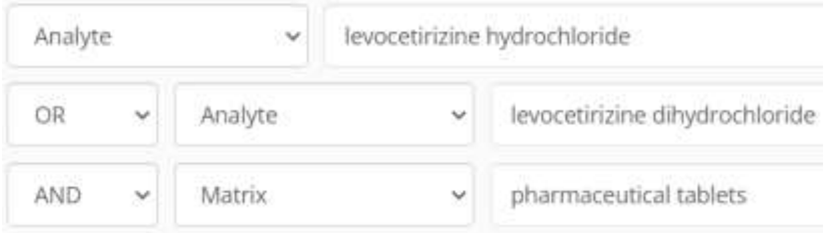
☑ Levocetirizine dihydrochloride (61)

☐ Cyclopropaneacetic acid, 1-[[[(1*R*)-1-[3-[(1*E*)-2-(7-chloro-2-quinolinyl)ethenyl]phenyl]-3-[2-(1-hydroxy-1-methylethyl)phenyl]propyl]thio]methyl]-, sodium salt (1:1) (30)

☐ (-)-Cetirizine (26)

☐ Ambroxol hydrochloride (10)

☑ Levocetirizine hydrochloride (9)

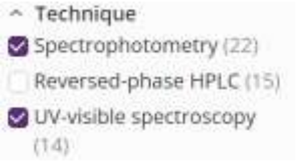


Analyte | levocetirizine hydrochloride

OR | Analyte | levocetirizine dihydrochloride

AND | Matrix | pharmaceutical tablets

Then we can refine the results using the **Technique** filter and browse the details of selected method or add up to 3 analytical methods to side-by-side comparison.

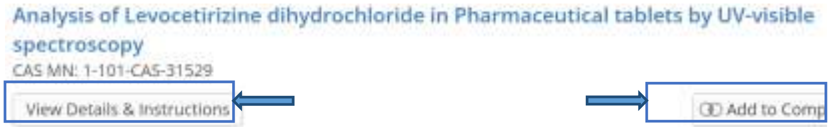


Technique

☑ Spectrophotometry (22)

☐ Reversed-phase HPLC (15)

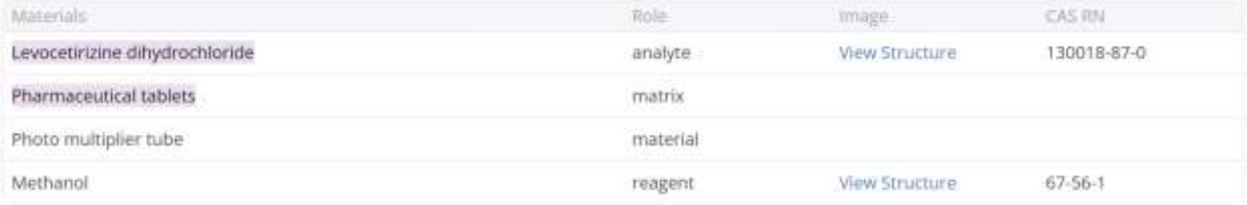
☑ UV-visible spectroscopy (14)



Analysis of Levocetirizine dihydrochloride in Pharmaceutical tablets by UV-visible spectroscopy

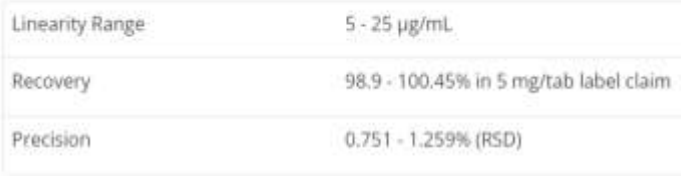
CAS MN: 1-101-CAS-31529

The detailed analytical method besides the list of materials



Materials	Role	Image	CAS RN
Levocetirizine dihydrochloride	analyte	View Structure	130018-87-0
Pharmaceutical tablets	matrix		
Photo multiplier tube	material		
Methanol	reagent	View Structure	67-56-1

and bibliographic information contains details on equipment used, conditions, instructions and **validation**.



Linearity Range	5 - 25 µg/mL
Recovery	98.9 - 100.45% in 5 mg/tab label claim
Precision	0.751 - 1.259% (RSD)

All this information can be viewed in comparison table conveniently, so you don't have to go to each single method for additional details.

Examples for practice

Goal1: Search for analytical methods for cefotaxime detection in blood, blood plasma or blood serum using HPLC or UPLC. *Notes:* check various types of spelling for method category (i.e. HPLC or high-performance liquid chromatography).

Goal2: Search for methods of chiral separation of ketoprofen in drugs or pharmaceutical compositions.

Goal3: Search for methods of DNA analysis in blood, blood plasma or blood serum.

IV. Biosequences

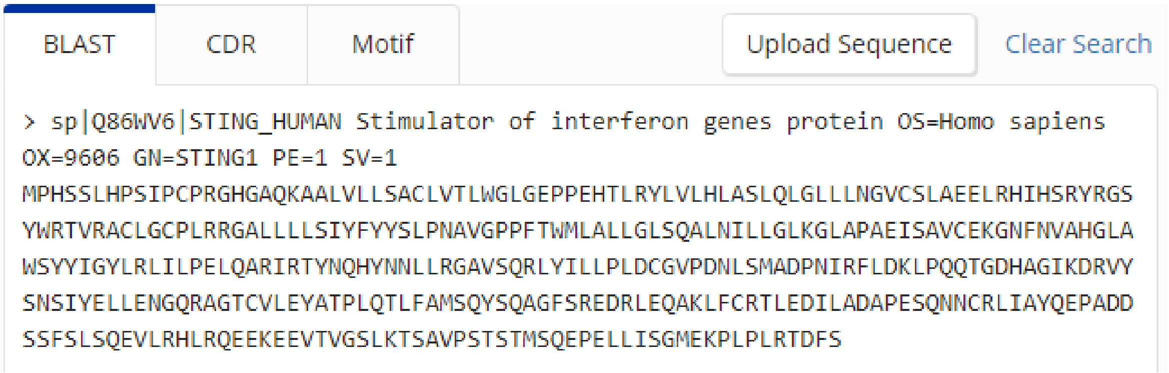
BLAST (Basic Local Alignment Search Tool) – a case study

Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The algorithm compares nucleotide or protein sequences to sequence database and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

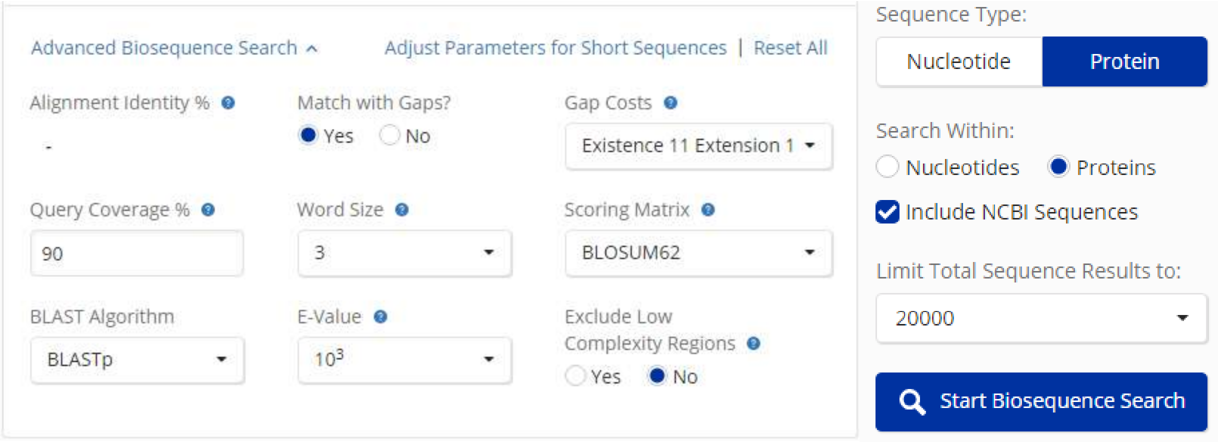
Let us search for patents that are devoted to cancer immunotherapy using analogs of Q86WV6 stimulator of interferon genes protein (STING):

MPHSSLHPSIPCPRGHGAQKAALVLLSACLVTLWGLGEPPEHTLRYLVLHLASLQLGLLLNGVCSLAEELRHIHSRYRGSY
WRTVRACLGCPLRRGALLLLSIYFYYSLPNAVGPPFTWMLALLGLSQALNILLGLKGLAPAEISAVCEKGNFNVAHGLAW
SYYIGYLRLILPELQARIRTYNQHYNNLLRGAVSQRLYILLPLDCGVPDNLSMADPNIRFLDKLPQQTGDHAGIKDRVYSN
SIYELLENGQRAGTCVLEYATPLQTLFAMSQYSQAGFSREDRLEQAKLFCRTLEDILADAPESQNNCRLIAYQEPADDSSF
SLSQEVLRHLRQEEKEEVTVGSLKTSAVPSTSTMSQEPELLISGMEKPLPLRTDFS

You can use the following settings, that provide broader coverage of the related biosequences and bring the highest number of results.



```
> sp|Q86WV6|STING_HUMAN Stimulator of interferon genes protein OS=Homo sapiens
OX=9606 GN=STING1 PE=1 SV=1
MPHSSLHPSIPCPRGHGAQKAALVLLSACLVTLWGLGEPPEHTLRYLVLHLASLQLGLLLNGVCSLAEELRHIHSRYRGS
YWRTVRACLGCPLRRGALLLLSIYFYYSLPNAVGPPFTWMLALLGLSQALNILLGLKGLAPAEISAVCEKGNFNVAHGLA
WSYYIGYLRLILPELQARIRTYNQHYNNLLRGAVSQRLYILLPLDCGVPDNLSMADPNIRFLDKLPQQTGDHAGIKDRVY
SNSIYELLENGQRAGTCVLEYATPLQTLFAMSQYSQAGFSREDRLEQAKLFCRTLEDILADAPESQNNCRLIAYQEPADD
SSFSLSQEVLRHLRQEEKEEVTVGSLKTSAVPSTSTMSQEPELLISGMEKPLPLRTDFS
```



Advanced Biosequence Search ^
Adjust Parameters for Short Sequences | Reset All
Start Biosequence Search

Alignment Identity %: -

Match with Gaps? ● Yes ○ No

Gap Costs: Existence 11 Extension 1

Query Coverage %: 90

Word Size: 3

Scoring Matrix: BLOSUM62

BLAST Algorithm: BLASTp

E-Value: 10^3

Exclude Low Complexity Regions: ○ Yes ● No

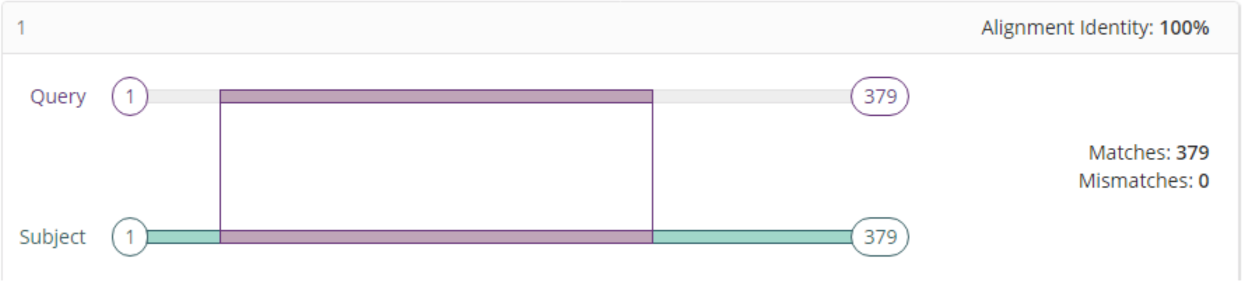
Sequence Type: Nucleotide | Protein (selected)

Search Within: ○ Nucleotides ● Proteins

☑ Include NCBI Sequences

Limit Total Sequence Results to: 20000

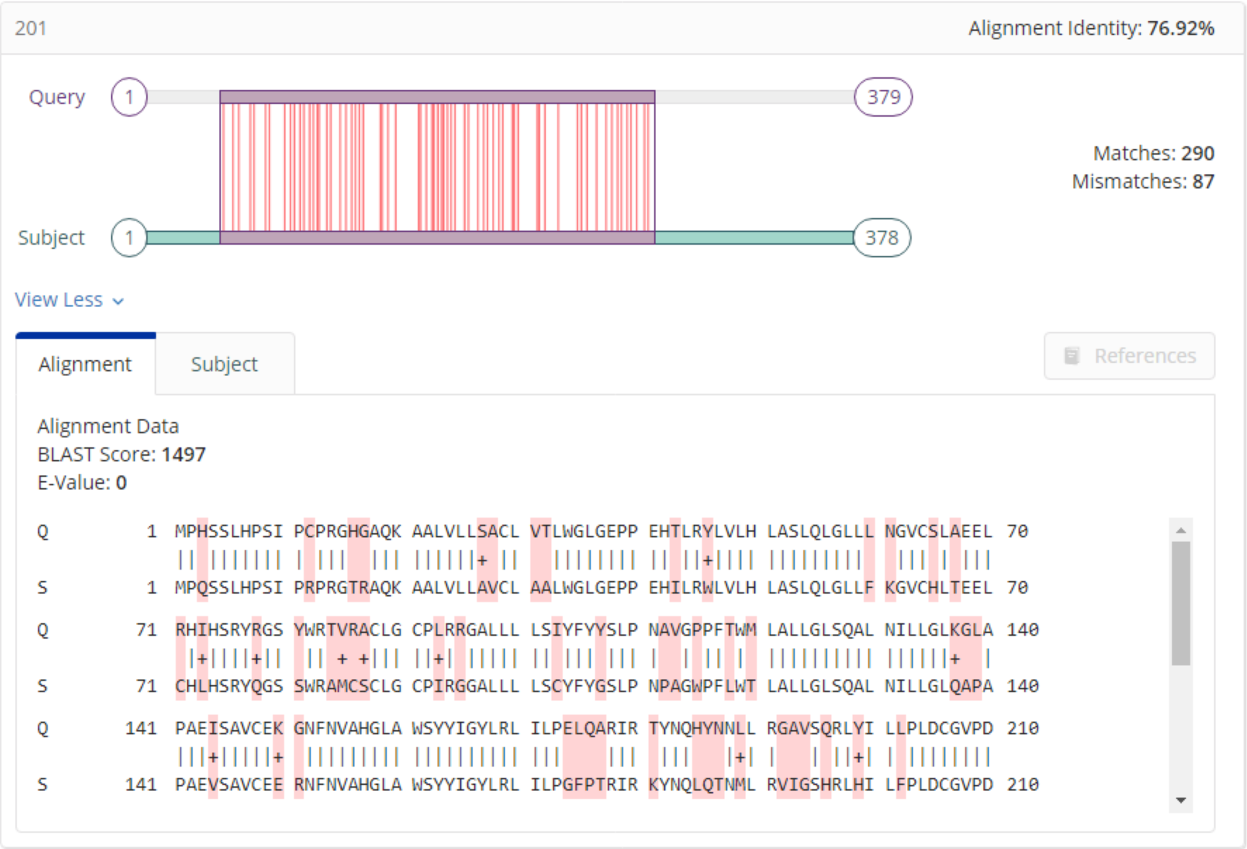
Now you have the set of results that range from exact match



Alignment Identity: **100%**

Matches: **379**
Mismatches: **0**

to sequences with significant number of mismatches, both positive (marked by +, amino acids have related structure and functions) and negative (not marked, rather different amino-acids in terms of structure and function).


201 Alignment Identity: **76.92%**

Query 1 379

Subject 1 378

Matches: **290**
Mismatches: **87**

View Less

Alignment | Subject | References

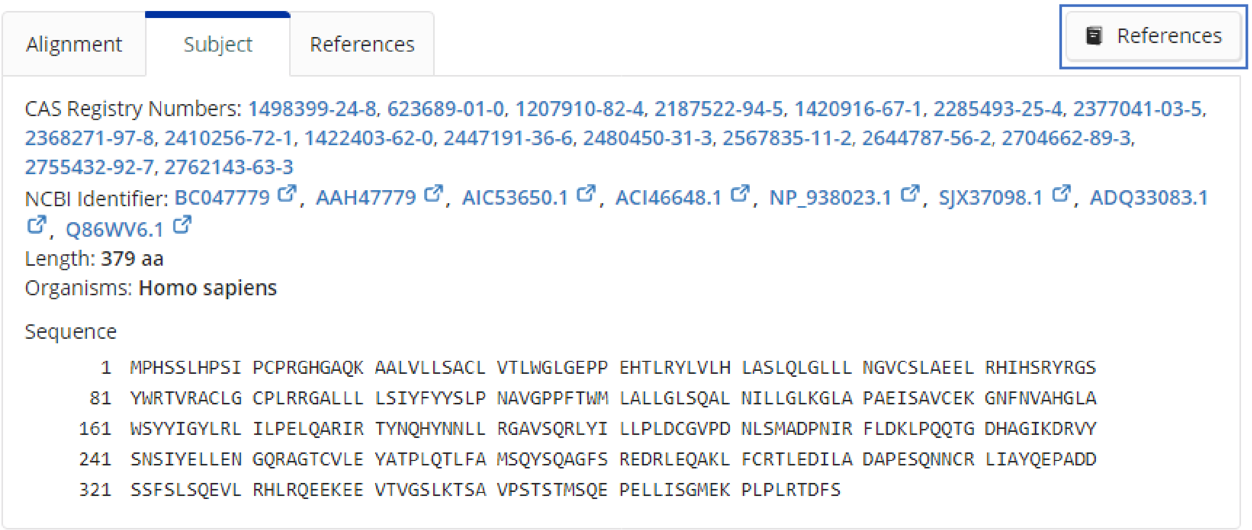
Alignment Data
BLAST Score: **1497**
E-Value: **0**

```
Q          1  MPHSSLHPSI PCPRGHGAQK AALVLLSACL VTLWGLGEPP EHTLRYLVLH LASLQLGLLL NGVCSLAEEL  70
              || ||||||| | |||  ||| ||||||+ ||   |||||||| || ||+|||| |||||||||   ||| | |||
S          1  MPQSSLHPSI PRPRGTRAQK AALVLLAVCL AALWGLGEPP EHILRWLVLH LASLQLGLLF KGVCHLTEEL  70

Q         71  RHIHSRYRGS YWRTVRACLG CPLRRGALLL LSIYFYYSLP NAVGPPFTWM LALLGLSQAL NILLGLKGLA  140
               |+||||+||  || + +||| ||+| ||||| || ||| ||| |  | || |  |||||||||| ||||||+  |
S         71  CHLHSRYQGS SWRAMCSCLG CPIRGGALLL LSCYFYGSLP NPAGWPFLWT LALLGLSQAL NILLGLQAPA  140

Q        141  PAEISAVCEK GNFNVAHGLA WSYYIGYLRL ILPELQARIR TYNQHYNNLL RGAVSQRLYI LLPLDCGVPD  210
              |||+|||||+  ||||||||| |||||||||| |||    |||  |||   |+| |   | ||+| | ||||||||
S        141  PAEVSAVCEE RNFNVAHGLA WSYYIGYLRL ILPGFPTRIR KYNQLQTNML RVIGSHRLHI LFPLDCGVPD  210
```

You can also browse the details for each sequence and find the identifiers of subject sequence and its structure. You can also retrieve references for a specific sequence that contain both patent and non-patent literature.


Alignment | Subject | References

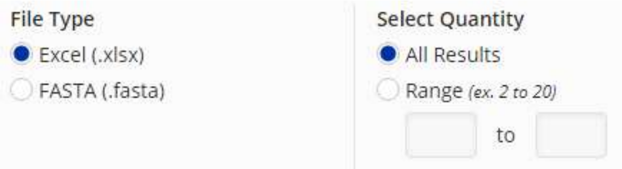
References

CAS Registry Numbers: 1498399-24-8, 623689-01-0, 1207910-82-4, 2187522-94-5, 1420916-67-1, 2285493-25-4, 2377041-03-5, 2368271-97-8, 2410256-72-1, 1422403-62-0, 2447191-36-6, 2480450-31-3, 2567835-11-2, 2644787-56-2, 2704662-89-3, 2755432-92-7, 2762143-63-3
NCBI Identifier: BC047779, AAH47779, AIC53650.1, ACI46648.1, NP_938023.1, SJX37098.1, ADQ33083.1, Q86WV6.1
Length: **379 aa**
Organisms: **Homo sapiens**

Sequence

```
      1  MPHSSLHPSI PCPRGHGAQK AALVLLSACL VTLWGLGEPP EHTLRYLVLH LASLQLGLLL NGVCSLAEEL RHIHSRYRGS
     81  YWRTVRACLG CPLRRGALLL LSIYFYYSLP NAVGPPFTWM LALLGLSQAL NILLGLKGLA PAEISAVCEK GNFNVAHGLA
    161  WSYYIGYLRL ILPELQARIR TYNQHYNNLL RGAVSQRLYI LLPLDCGVPD NLSMADPNIR FLDKLPQQTG DHAGIKDRVY
    241  SNSIYELLEN GQRAGTCVLE YATPLQTLFA MSQYSQAGFS REDRLEQAKL FCRTLEDILA DAPESQNNCR LIAYQEPADD
    321  SSFSLSQEVL RHLRQEEKEE VTVGSLKTSA VPSTSTMSQE PELLISGMEK PLPLRTDFS
```

You can also download the results in either Excel (up to 1000 sequences) or Fasta (up to 100 sequences) formats.



File Type

● Excel (.xlsx)

○ FASTA (.fasta)

Select Quantity

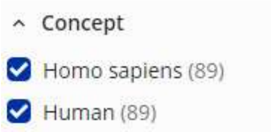
● All Results

○ Range *(ex. 2 to 20)*

to

To retrieve all patent references for the obtained set of sequences using the button on the top of the page.

You can further refine the reference set using concepts and search within functionality - **cancer and (therapy or treatment or immunotherapy)**.



Concept

☑ Homo sapiens (89)

☑ Human (89)

Examples for practice

Goal 1. Using sequence of B-cell receptor CD22 isoform 1 precursor, find patents that describe antibody-drug conjucates using monoclonal antibodies targeting relapsed or refractory CD22-positive B-cell precursor acute lymphoblastic leukemia.

```
> NP_001762.2 B-cell receptor CD22 isoform 1 precursor [Homo sapiens]

MHLLGPWLLLLVLEYLAFSDSSKWVFEHPETLYAWEGACVWIPCTYRALDGDLESFILFHNPEYNKNTSKFDGTRLYESTKDGKVPSE
QKRVQFLGDKNKNCTLSIHPVHLNDSGQLGLRMESKTEKWMERIHLNVSERPFPPHIQLPPEIQESQEVTLTCLLNFSCYGYPIQLQ
WLLEGVPMRQAAVTSTSLTIKSVFTRSELKFSPQWSHHGKIVTCQLQDADGKFLSNDTVQLNVKHTPKLEIKVTPSDAIVREGDSVT
MTCEVSSSNPEYTTVSWLKDGTSLKKQNTFTLNLREVTKDQSGKYCCQVSNDVGPGRSEEVFLQVQYAPEPSTVQILHSPAVEGSQ
VEFLCMSLANPLPTNYTWYHNGKEMQGRTEEKVHIPKILPWHAGTYSCVAENILGTGQRGPGAELDVQYPPKKVTTVIQNPMPIRE
GDTVTLSCNYNSSNPSVTRYEWKPHGAWEEPSLGVLKIQNVGWDNTTIACAACNSWCSWASPVALNVQYAPRDVRVRKIKPLSEI
HSGNSVSLQCDFSSSHPKEVQFFWEKNGRLLGKESQLNFDSISPEDAGSYSCWVNNSIGQTASKAWTLEVLYAPRRLRVSMSPGDQ
VMEGKSATLTCESDANPPVSHYTWFDWNNQSLPYHSQKLRLEPVKVQHSGAYWCQGTNSVGKGRSPLSTLTVYYSPETIGRRVAV
GLGSCLAILILAICGLKLQRRWKRTQSQQGLQENSSGQSFFVRNKKVRRAPLSEGPHSLGCYNPMMEDGISYTTLRFPEMNIPRTGD
AESSEMQRPPPDCDDTVTYSALHKRQVGDYENVIPDFPEDEGIHYSELIQFGVGERPQAQENVDYVILKH
```

Note1: You can allow to include gaps to broaden you BLAST search. Use protein to protein BLAST with NCBI sequences.

Note2: After retrieving references you can use both concepts and search within to determine relevant publications.

Goal 2. Using this partial HIV-1 GP120 envelope glycoprotein find all patents considering HIV vaccines.

```
CTRPNNNTRKSIHIGPGRAFYTTGEIIGDIRQAHC
```

Note1: Use BLAST-p-short algorithm for short-sequence.

Note2: Try using subject coverage filter to exclude long and less relevant sequences.

Note3: After retrieving references you can use both concepts and search within to determine relevant publications.

Goal 3. Search for peptides closely related to the “human anti-(human immunodeficiency virus 1 envelope glycoprotein gp120env) immunoglobulin G1 γ1-chain-specifying” coded by the following DNA part. Get references that describe antibodies that target HIV GP120.

```
caggtccacttgtctcaatctggcgccgctgtgacaaagcctggcgcttctgtcagagtgtcttgcgaggcctccggctacaacatccgggactactttatccactggtg
gcggcaggctccaggacagggattgcaatgggtcggatggatcaaccctaagaccggccagcctaacaaccctagacagttccagggcagagtgtccctgaccaga
cacgcctcttgggacttcgacaccttcagcttctacatggacctgaaggccgtgcggagcgacgacaccgctatctacttttgcgccagacagagatccgactactggg
atttcgatgtgtggggctctggcacccaagtgaccgtgtcctctgcttctaccaagggaccctctgtgttccctctggctccttccagcaagtctacctctggtggaaccg
ctgctctgggctgcctggtcaaggattactttcctgagcctgtgacagtgtcctggaactctggtgctctgacctccggcgtgcacacatttccagctgtgctgcagtcct
ccggcctgtactctctgtcctctgtcgtgaccgtgccttctagctctctgggcacccagacctacatctgcaacgtgaaccacaagccttccaacaccaaggtggacaag
aaggtggaacccaagtcctgcgacaagacccacacctgtcctccatgtcctgctccagaactgctggctggccccgatgtctttctgttccctccaaagcctaaggaca
ccctgatgatctctcggacccctgaagtgacctgcgtggtggtggatgtgtctcacgaggatcccgaagtgaagttcaattggtacgtggacggcgtggaagtgcaca
acgccaagaccaagcctagagaggaacagtacaactccacctacagagtggtgtccgtgctgaccgtgctgcaccaggattggctgaacggcaaagagtacaagtg
caaggtgtccaacaaggccctgcctctgcctgaggaaaagaccatctctaaggctaagggccagcctcgcgagcctcaggtttacacactgcctccatctcgggaag
agatgaccaagaaccaggtgtcactgacctgcctcgtgaagggcttctacccttccgatatcgccgtggaatgggagtccaatggccagcctgagaacaactacaag
acaacccctcctgtgctggactccgacggctcattcttcctgtactccaagctgacagtggacaagtctcggtggcagcagggcaacgtgttctcttgtagtgtgctgca
cgaggccctgcactcccactatacccagaagtccctgtctctgtcccctggcaaa
```

Note1: In this case the subject sequence of exact match is 3-times shorter because 3 nucleotides of the query DNA code only one amino acid of respective peptide.

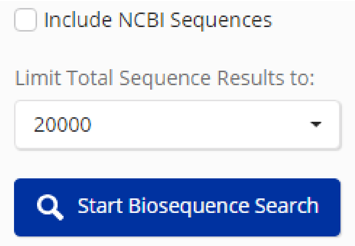
Note2: After retrieving references you can use both concepts and search within to determine relevant publications.

CDR (Complementarity-Determining Region) – a case study

Complementarity-determining regions (CDRs) are part of the variable chains in immunoglobulins (antibodies) and T cell receptors, generated by B-cells and T-cells respectively, where these molecules bind to their specific antigen. A set of CDRs constitutes a paratope. As the most variable parts of the molecules, CDRs are crucial to the diversity of antigen specificities generated by lymphocytes.

In this case we will search for antibodies specifically binding to HER2 (a protein overexpressed in a certain cancer cell lines) that can carry a payload of drug – tubulin inhibitor (antibody-drug conjugate).

First, create a search for these 3 CDRs listed below. Usually, it is always valuable to include NCBI sequences in the search results, but as we are interested in combining the obtained results with structures of specific drugs we omit them in this case.



☐ Include NCBI Sequences

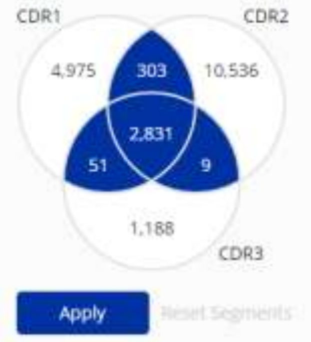
Limit Total Sequence Results to: 20000

CDR1: DTYIH

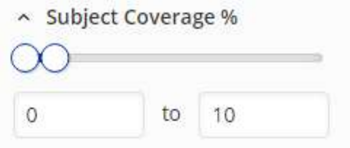
CDR2: RIYPTNGYTRYADSVKG

CDR3: WGGDGFYAMDY

Then, using the Vann diagram limit the results to sequences, containing at least 2 of the target CDRs. Antibodies have normally 2 chains, light chain has around 220 amino acids, heavy chains up to 550 amino acids.

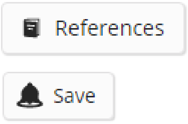


To limit results only to heavy chains we can limit the substance coverage to <10%.

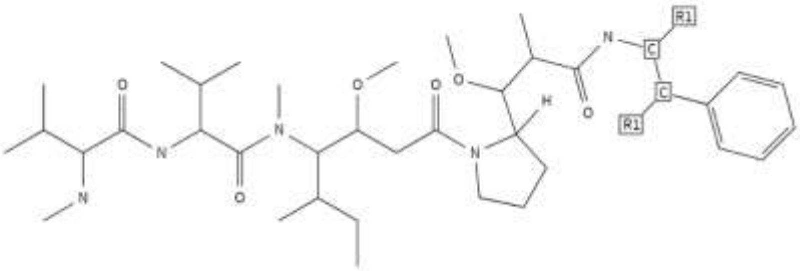


0 to 10

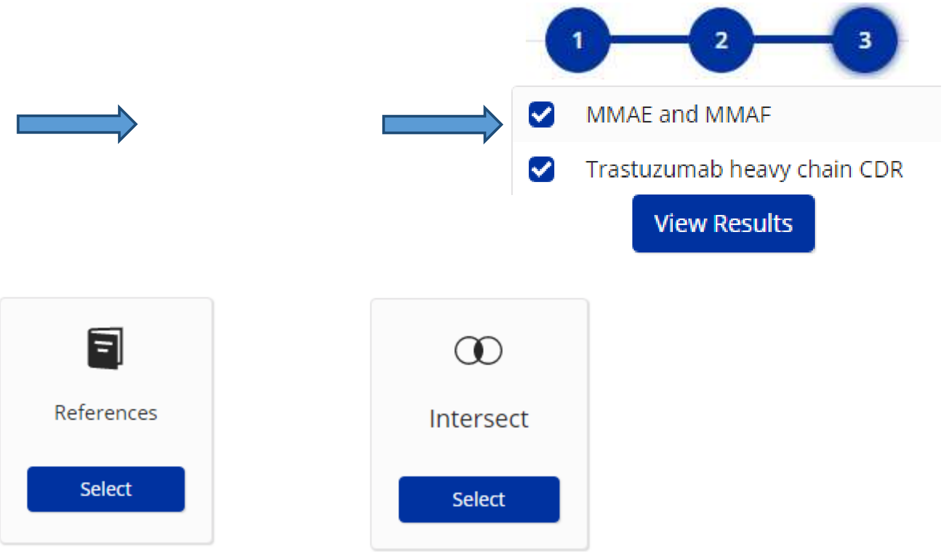
As we have 33 amino acids in the original query this limitation will bring target sequences with >330 amino acids. Then we can retrieve all patent references for the obtained set of sequences using the button on the top of the page and save the resulting references.



Next, we perform a substructure search for modified structure of CAS RN: 745017-94-1 with no stereo and some variations (R1 = H, N or COOH). Substructure search is required in this case to include molecules linked to sequences. Then we retrieve all references and save these results.



Finally, we go to saved 🔔 results Combine and use the function for 2 sets of references to find the references that have both parts of our search.



Examples for practice

Goal 1. Starting from CDRs of SARS-COV-2 Immunoglobulin G1, anti-(severe acute respiratory syndrome coronavirus 2 spike glycoprotein) (human monoclonal REGN10933 γ1-chain) find all publications related to SARS-COV-2.

CDR1: DYYMS

CDR2: YITYSGSTIYYADSVKG

CDR3: DRGTTMVPFDY

Note1: Use Vann diagram to limit results to sequences that contain at least 2 CDRs.

Goal 2. Use CDRs of H1/H5 cross-reactive influenza antibody heavy chain VDJ region immunoglobulin to find all publications related to CD3 antigens.

CDR1: TYAIS

CDR2: GIIAIFGTTNYAQKFQG

CDR3: GNGYYHNYFDF

Note1: Use Vann diagram to limit results to sequences that contain at least 2 CDRs.