



Genomics Institute of the
Novartis Research
Foundation

Chemical-text Hybrid Search Engines

Yingyao Zhou, Bin Zhou, Shumei Jiang, Fred King

Genomics Institute of the Novartis Research Foundation

Additional information is available at

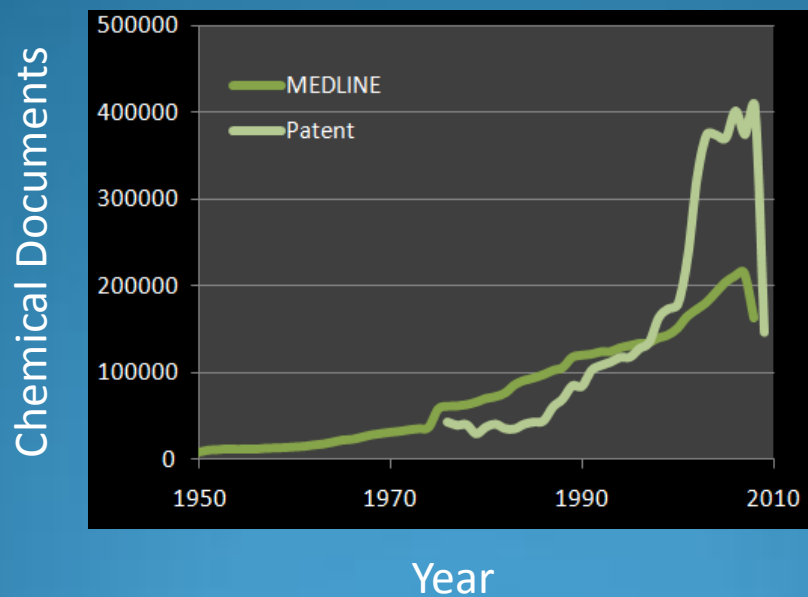
Zhou et al. J Chem Inf Model. 2010 Jan;50(1):47-54.

<http://pubs.acs.org/doi/abs/10.1021/ci900380s>

Chemical Information Explosion

SureChem database already contains over 5.1 million patents and 4.7 million MEDLINE articles. There are over 400,000 patents and 200,000 MEDLINE articles added annually in the past few years.

Are we able to find what we want?



Existing Search Solutions

Existing Search Solutions can be segregated into two categories: text (Google, Bing) and chemical search engines (SciFinder).

Solutions based on these alone are rather limited because of false negative or false positive hits in search.

Example: identify all documents that describe certain associations between a chemical compound (e.g., a Gleevec analog) and a therapeutic application (e.g., chronic myelogenous leukemia (CML))

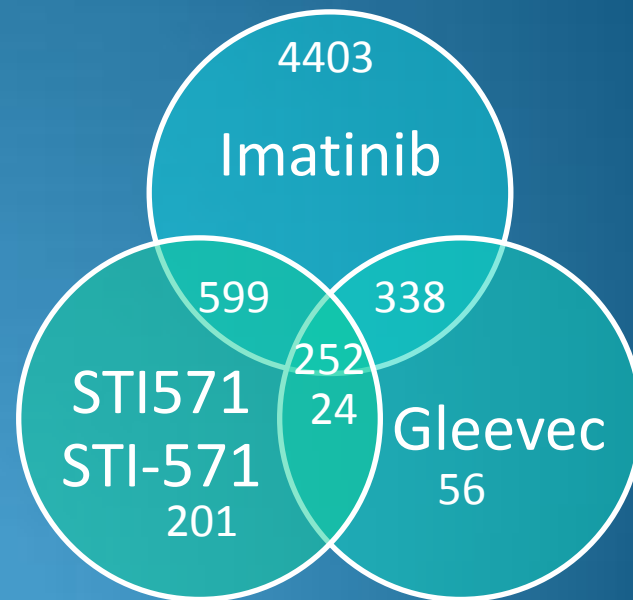
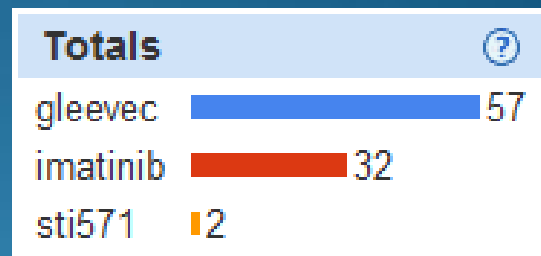
Text Search Engines - False Negative Problem

“Gleevec” is a more popular search term in Google search: 57% of Google searches uses the term “Gleevec”, 32% uses “Imatinib”

“Imatinib” is a more popular identifier in scientific literature: only 11% of PubMed articles uses the term “Gleevec”.

Chemical synonyms are not understood by text search engines: search “Gleevec” does not hit “Imatinib”.

Text-based similarity & substructure search are not feasible.



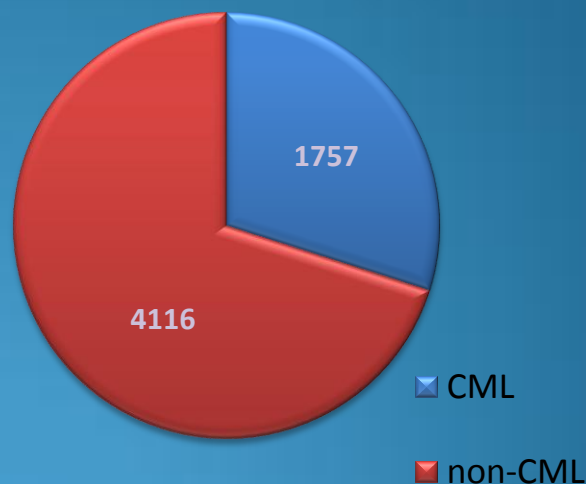
PubMed articles by synonyms

Chemical Search Engines - False Positive Problem

Structure search alone return non-specific hits:

60% of Gleevec-related PubMed articles are not relevant to CML. No support for “Gleevec NEAR CML”.

No ready intranet solution: file sources, file formats, file permissions.



How Text Search Engine Works (MOSS Search)

Three documents:

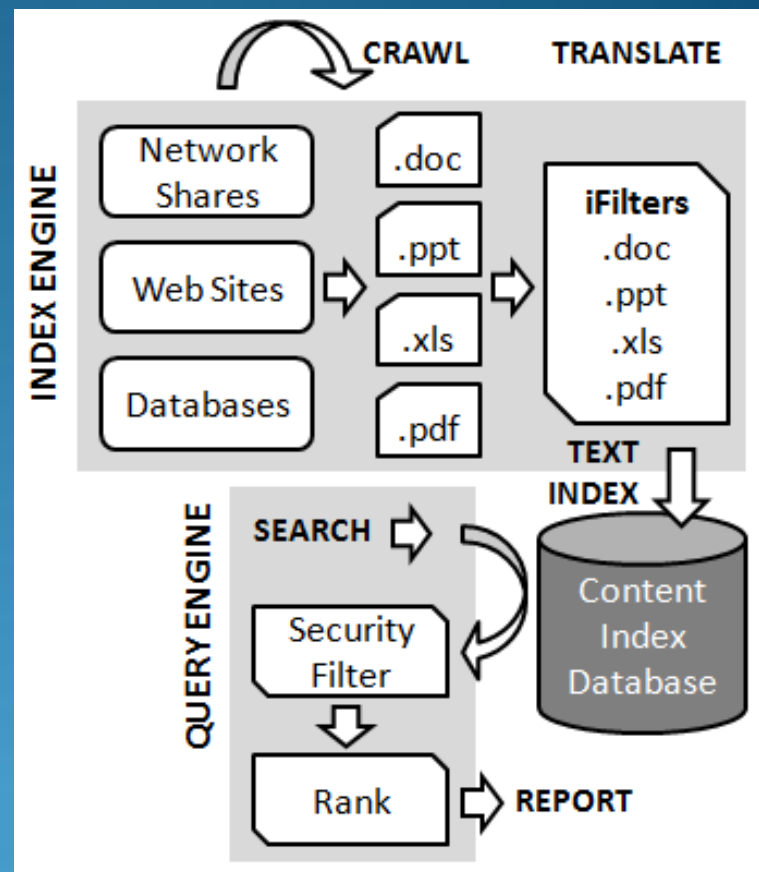
- #1. **STI571** is a Bcr-Abl inhibitor.
- #2. **Gleevec** is a CML drug.
- #3. **Imatinib** is a Novartis drug.

Pros

- Crawler, iFilter, Page Ranker (suitable for Intranet)
- Proximity search: Gleevec NEAR CML.

Cons

- “Gleevec” only returns #2, misses #1 and #3.
- “Imatinib NEAR CML” misses #2.
- Structures 90% similar to Gleevec, not supported.
- InChi key is not the answer (“KKTUFNOKKBVMGRW-UHFFFAOYSA-N”)



How Chemical Search Engine Works

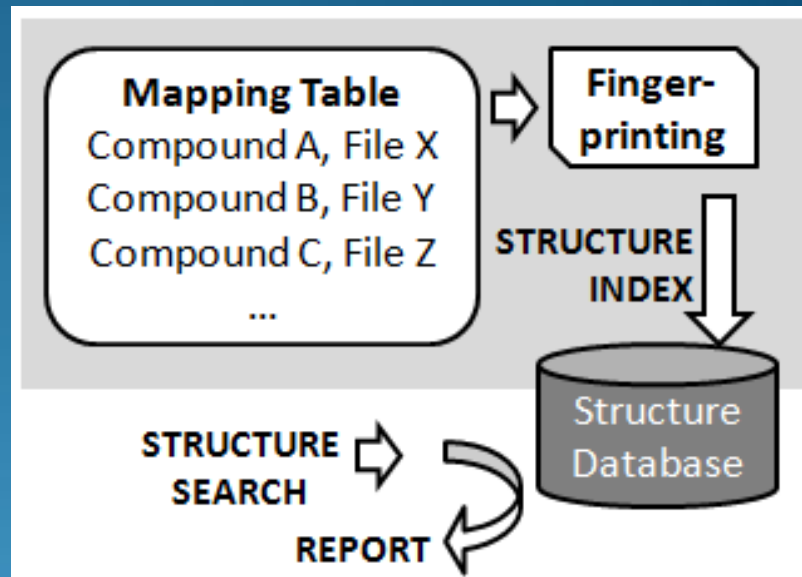
1. STI571, File #1
2. Gleevec, File #2
3. Imatinib, File #3

Pros

- “Gleevec” by structure will return all three documents.
- All documents containing Gleevec analogs (>80% structure similar).

Cons

- Does not support text search
“Gleevec AND CML”
- No proximity search
“Gleevec NEAR CML”
- No Crawler, iFilter, Page Ranker (not suitable for Intranet search)



Text and Chemical Search Engines are Complementary

Text search engines are ideal for Internet and Intranet applications, but lack of chemical intelligence. Chemical search engines are great for structure search, but weak most other aspects. We aim to build chemical-text hybrid engines by introducing chemical intelligence into text search engines.

Features	Text Search Engines	Chemical Search Engines	Hybrid Search Engines	Important to Intranet	Important to Internet
Crawling and indexing	3	1	3	3	3
Support web pages	3	2	3	2	3
Support files in file system, database, SharePoint, etc.	3	1	3	3	1
Support non-text file formats	3	1	3	3	3
Support file meta data	3	1	3	3	2
Page ranking	3	1	3	3	3
Document security	3	1	3	3	1
Understand chemicals names	1	3	3	2	3
Understand proprietary IDs	1	2	3	3	2
Understand chemical drawings	1	2	1	1	3
Search chemicals	1	3	3	3	3
Search chemicals and text	1	1	3	3	3

The Idea of Entity-Canonical Keyword Indexing (ECKI)

Entity

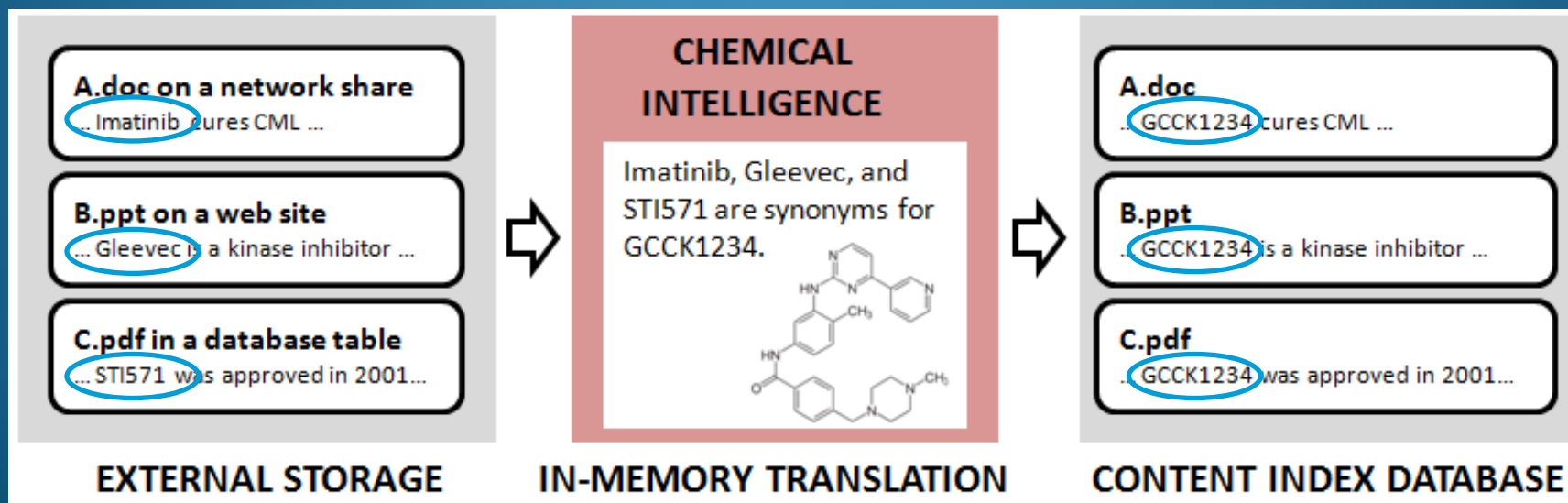
A chemical structure. The entity can be represented in many different forms, e.g., Gleevec, Imatinib, STI571, IUPAC name, etc all represents the same structure.

Canonical Keyword (CK)

An indexable unique identifier for an entity. E.g., the CK for Gleevec is GCCK1234.

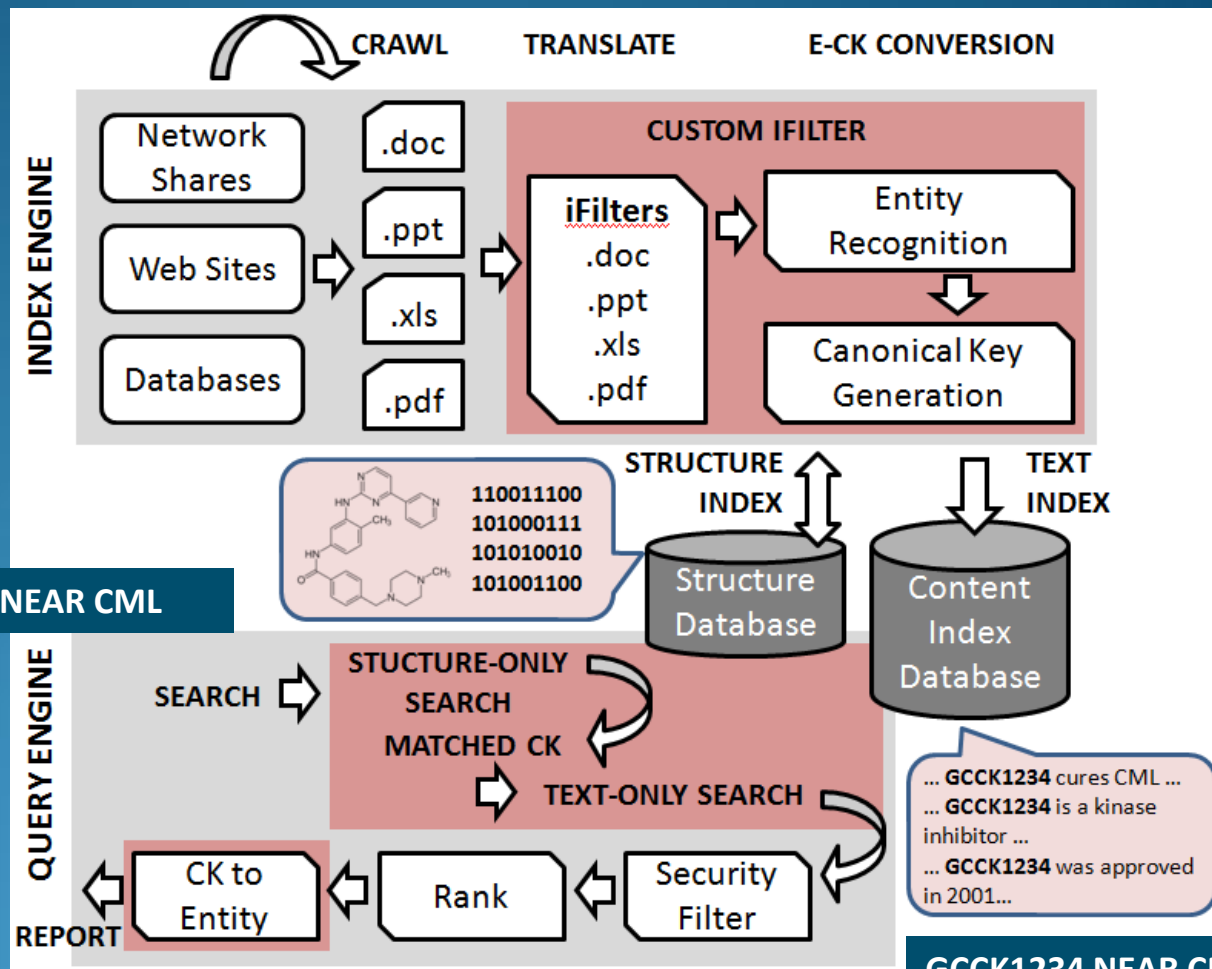
ECKI

No matter what synonyms an entity used in the original document, it appears as if the corresponding CK were used for the text search engine.



GNF Implementation using MOSS + ECKI

1. STI571 cures CML.
2. Gleevec is a kinase inhibitor.
3. Imatinib was approved in 2001.

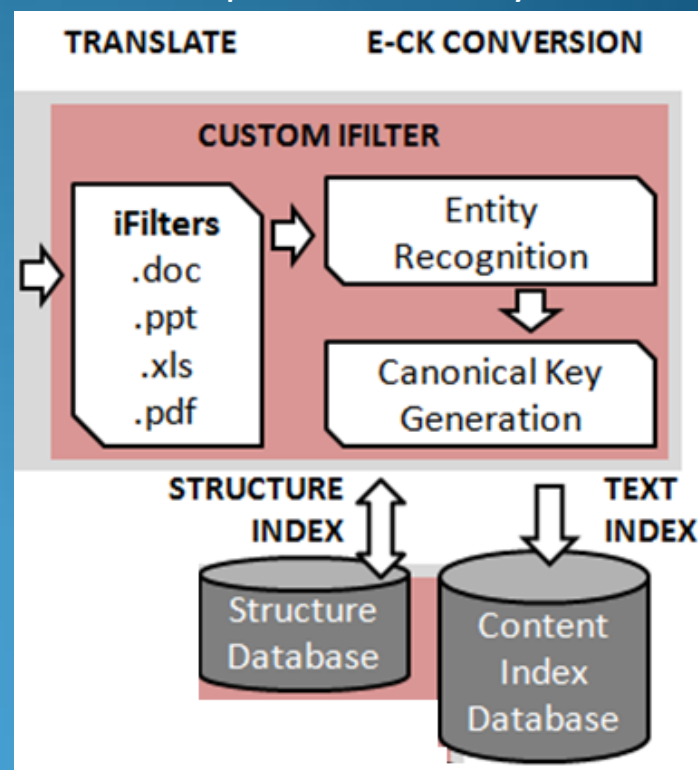


Query “[Gleevec] NEAR CML” is transformed into “(GCCK1234 NEAR CML)”

Query “[Gleevec > 90%] NEAR CML” is transformed into “(GCCK1234 NEAR CML) OR (GCCK5678 NEAR CML)”

GNF Custom iFilter

1. Act as the proxy for existing iFilters, total transparency in formation conversion.
2. Recognizes chemical entities, including proprietary corporate ID (customized) drug names dictionary (SureChem, ChemAxon), IUPAC-to-structure conversion library (ChemAxon), etc.
3. Canonical Key generation uses ChemAxon cartridge, can be replaced with any other key generation service.



SharePoint Search Interface

Query: “[S1] NEAR CML” or “[Gleevec > 0.9] NEAR CML”

The screenshot displays the ChemAxon software interface. At the top, a search bar contains the text "Search key words: [S1] NEAR CML" and a "Submit Search" button. Below the search bar is a menu bar with options: File, Edit, View, Insert, Atom, Bond, Structure, Tools, and Help. A toolbar with various icons is located below the menu bar. The main workspace shows a chemical structure of a complex molecule with multiple rings and functional groups. To the right of the workspace is a vertical toolbar with buttons for H, C, N, O, S, F, and P. Below the workspace is a horizontal toolbar with various icons. At the bottom of the workspace are several ring templates. On the right side of the interface, there is a "Smiles/SDF" section with "Select All" and "Clear" buttons. Below this are buttons for "Put File", "Get Mol", and "Get Smiles". At the bottom right, there is a "Search type:" section with three radio buttons: "Exact", "Substructure", and "Similarity 0.90" (which is selected).

SharePoint Search Interface (continued)

Query Result Presentation

(GCCK1234 NEAR CML) OR (GCCK5678 NEAR CML)

Results 1-4 of 4. Your search took 0.22 seconds.

GNF Wildcard Search

Results by Relevance | View by Modified Date | Alert Me

- [Dasatinib - Wikipedia, the free encyclopedia](#)
 imatinib ... were seen in 37 of 40 patients with chronic-phase **CML**. ... wer
<file:///depts/cheminfo/yzhou/mosssearch/wiki/2151.html> - 44KB - 3/4/2009
- [Imatinib - Wikipedia, the free encyclopedia](#)
has approved imatinib as first-line treatment for **CML**. ... Imatinib has
mice treated with large doses of imatinib ...
<file:///depts/cheminfo/yzhou/mosssearch/wiki/3482.html> - 58KB - 3/5/2009
- [Nilotinib - Wikipedia, the free encyclopedia](#)
... in cases of **CML** resistant to treatment with ... imatinib ... <http://en.wik>
<file:///depts/cheminfo/yzhou/mosssearch/wiki/4647.html> - 38KB - 3/5/2009
- [Busulfan - Wikipedia, the free encyclopedia](#)
 Imatinib ... **CML**), where it is used as a conditioning drug. ... <http://en.wikipedia.org/wiki/> Busulfan
<file:///depts/cheminfo/yzhou/mosssearch/wiki/1348.html> - 56KB - 3/4/2009

report

GNFTAG13642

The chemical structure of Imatinib is shown. It features a central pyrimidine ring substituted with a 4-piperidinylphenyl group and a 4-(3,4-dimethylphenyl)amino group. The pyrimidine ring is also substituted with a 2-pyridinyl group and a 4-(3,4-dimethylphenyl)amino group.

Use Case #1: Crawling GNF File Share

Goal: For each chemical structure, list all the in-house documents in the drug discovery folder where users describe the compound (e.g., used in e-discovery).

Top 12 most frequently referenced compounds.

Compound	# of Files	Description (corporate annotation removed)
Cpd1	50857	CSP/sPoC
Cpd2	29587	Novartis Drug
Cpd3	28011	CSP/sPoC
Cpd4	22429	CSP/sPoC
Cpd5	20457	patent
Cpd6	20155	patent
Cpd7	16812	patent
Cpd8	16419	GNF Patent
Cpd9	14277	patent
Cpd10	14071	
Cpd11	14001	patent
Cpd12	13223	sPoC

Use Case #2: Wikipedia Search

Drug Wikipedia Search

We downloaded ~7000 drug wiki pages, indexed by our hybrid-MOSS Search Engine.

Question	Query	Matched Wiki Entry	Text Engine
Everolimus-analogs	[Everolimus> 0.95]	Everolimus, Sirolimus	None
Use STI571 to show rational drug design can work	[STI571] NEAR rational NEAR design	Imatinib	False negatives: None (STI-571 was used)
All compounds related to GIST(s)	GCCK* AND GIST*	Imatinib, Sunitinib	False Positives: GIST-containing pages does not describe compounds

Use Case #3: PubMed Search

Search Recent **MEDLINE** Titles and Abstracts.
We downloaded ~250k MEDLINE web pages.

Question	Query	Matched Entry (PubMed ID-Compound)
non-Gleevec CML compounds	"Chronic myelogenous leukemia" AND "GCCK*" AND NOT "GCCK1234"	18537755, nelarabine, forodesine 18705753, vincristine, quinacrine 18644865, doxorubicin
Use STI571 to show rational drug design can work	[STI571] NEAR rational NEAR design	18616236, imatinib
All compounds related to GIST(s)	"GCCK*" AND "gastrointestinal stromal tumor" AND "GIST"	18708414, imatinib 18294292, imatinib 17729245, imatinib, sunitinib

Summary

What is out there

- Text search engines (Google) do not understand compound synonyms (false-negative issue), do not support similarity/substructure searches.
- Chemical search engines (SciFinder) ignore text context. No proximity search, no crawling, security filtering and ranking mechanism.

What ECKI enables

- Adding chemical intelligence to existing text search engines (say MOSS Search), so that chemical search naturally becomes a text search problem.
- Support complex hybrid search such as “[Gleevec > 80%] NEAR CML”.

Corporate Usage

- Develop custom iFilters to recognize proprietary terms/IDs, install it with its existing MOSS Search engine to index corporate file stores.
- A tool for biomedical and IP research (e-discovery).
- The concept of ECKI can be extended to genes, proteins, etc.