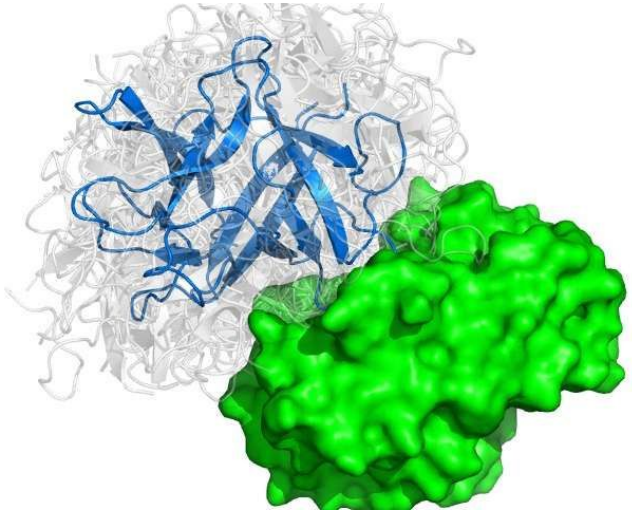
Natural Language Processing in Text Mining for Structural Modeling of Protein Complexes

Varsha D. Badal, Petras J. Kundrotas and Ilya A. Vakser

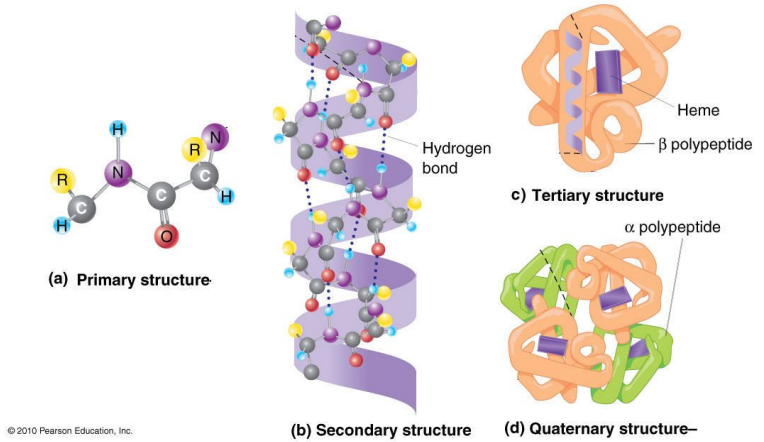


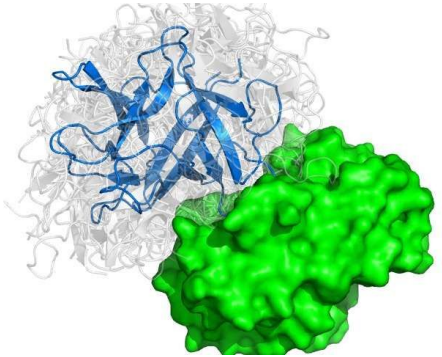
Presented By:
Anmol Sharma
Medical Image Analysis Laboratory
Simon Fraser University
Burnaby, BC Canada

Problem Statement and Background

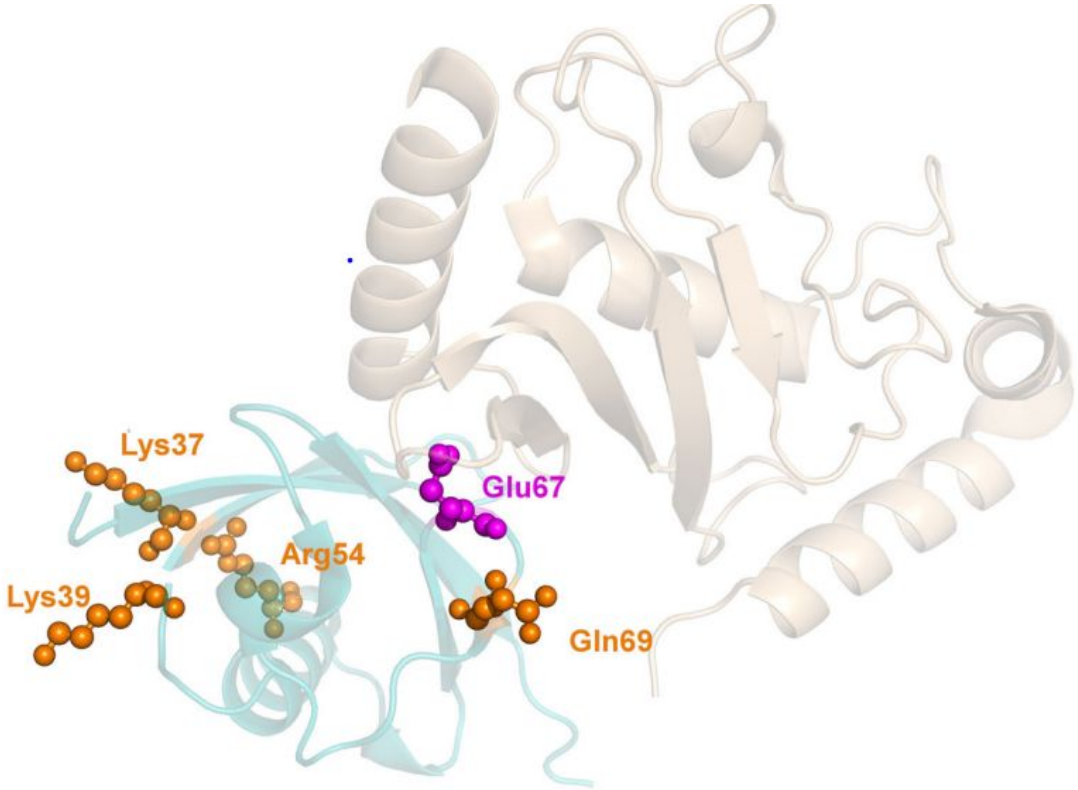
Proteins and Protein-Interactions

- Proteins control all biological systems in a cell
- Many proteins perform their functions independently,
- The vast majority of proteins interact with others for proper biological activity.





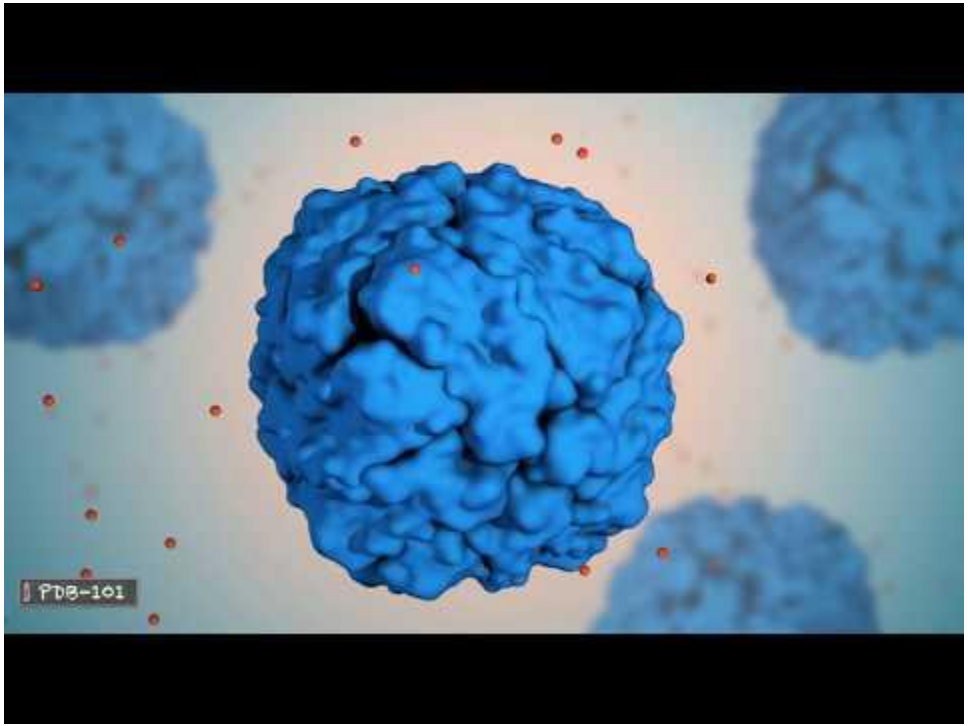
Protein-Protein Interaction



Why Study Protein-Protein Interactions?

- **Cell signalling** is part of the molecular biology system that controls and coordinates the actions of cells.
- **Signalling proteins are intrinsic to all biological processes**
- Characterization of **protein binding can help** to **elucidate protein function within signalling pathways.**
- We gain a more **comprehensive knowledge of cellular networks** which **can then be used to develop new therapeutic strategies for disease.**

Protein-Protein Interaction



Protein-Protein Interaction

- Protein-Protein Interaction is **hard** to model and predict.
- Given two protein structures, the task is to
  - Predict how the protein structures interact
  - Predict the final 3D structure of the complex once interaction is complete
- Many techniques are used for modelling this problem.
- Most commonly used one is Docking.
  - Predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex

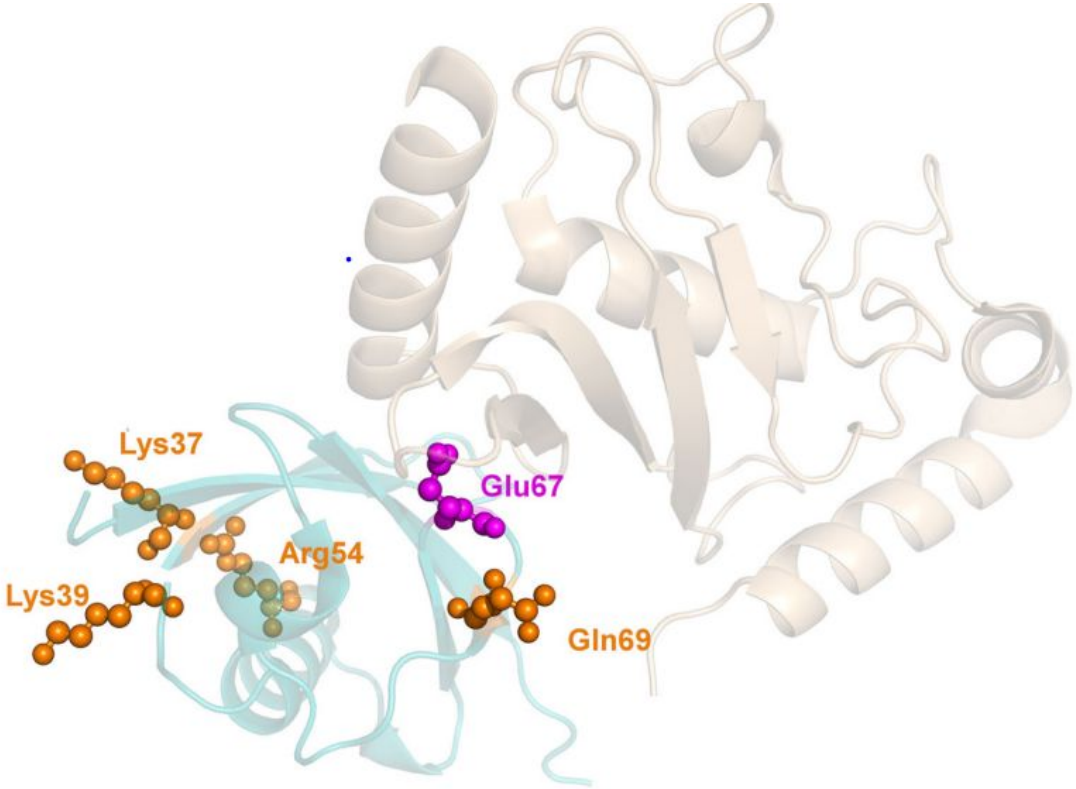
Protein-Protein Interaction

- However, Docking methods are not perfect.

Question: How can we make existing docking methods better?

Answer: Add useful constraints in terms of **binding-site residues**!

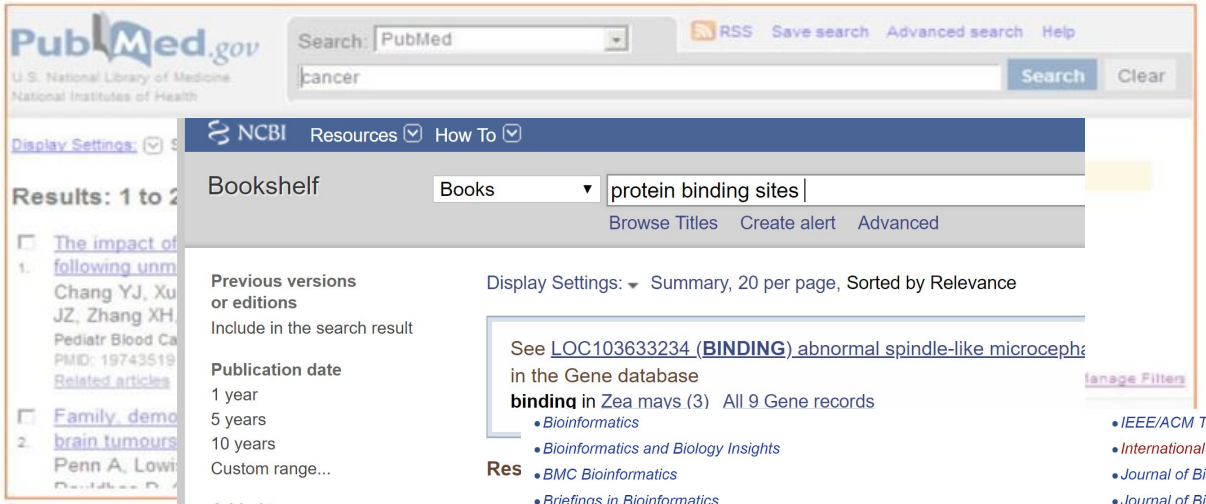
Binding Site Residues



These residues form temporary bonds with the substrate (binding site) and residues that catalyse a reaction of that substrate (catalytic site).

If we know the usual binding-site residues of each protein, and know how to they interact with other binding sites, we can predict if two proteins will interact or not and/or their final 3D structure.

Where does this information comes from?



Published Research!

- *Bioinformatics*
- *Bioinformatics and Biology Insights*
- *BMC Bioinformatics*
- *Briefings in Bioinformatics*
- *Bulletin of Mathematical Biology*
- *Cancer Informatics*
- *Computational and Structural Biotechnology Journal*
- *Computational Biology and Chemistry*
- *Computers in Biology and Medicine*
- *Current Bioinformatics*
- *Database*
- *EMBnet.journal*
- *Evolutionary Bioinformatics*
- *IEEE/ACM Transactions on Computational Biology and Bioinformatics*
- *International Journal of Functional Informatics and Personalized Medicine*
- *Journal of Bioinformatics and Computational Biology*
- *Journal of Biomedical Informatics*
- *Journal of Computational Biology*
- *Journal of Mathematical Biology*
- *Journal of Theoretical Biology*
- *PLoS Computational Biology*
- *Rapid Communications in Mass Spectrometry*
- *Source Code for Biology and Medicine*
- *Statistical Applications in Genetics and Molecular Biology*
- Open Computer Science (open access journal)

Protein Binding Sites

- **Question**: How do we consolidate this knowledge and incorporate this into our system?
- **Answer:** Two methods:
  a. Read all the papers, manually scrape through all information and find relevant ones.
  b. Design an automated system to scrape through the material and get the information curated
- **LARGE** number of publications on this topic makes **a)** unfeasible.
- **Our only option is b).**

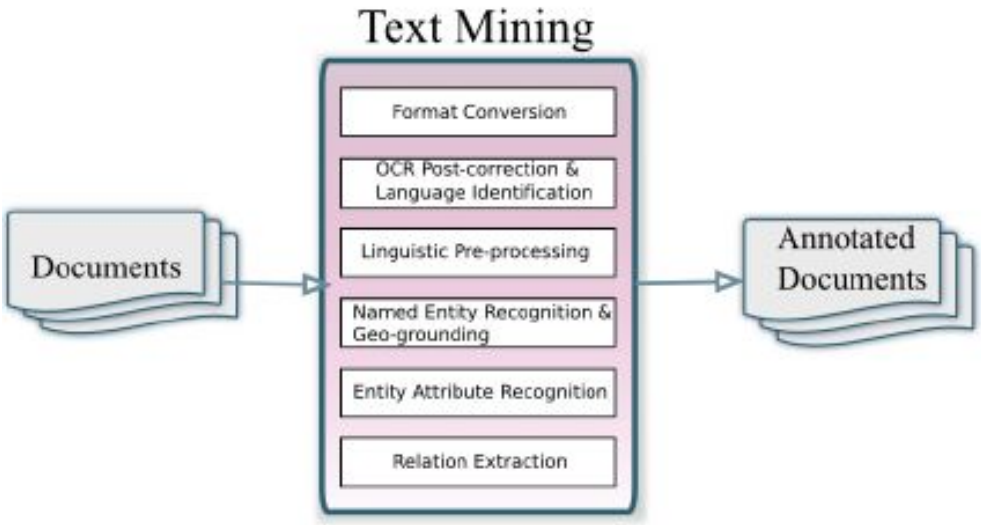
What does this entail?

- Need a system that:
  - Parses through unstructured text:
    - **“This is a bioinformatics class with lectures at BC Cancer Agency building. “**
      - Should answer the questions:
        - **What is the course?**
          - **Bioinformatics**
        - **Where is it held?**
          - **BC Cancer Agency Building**

- Where can we find this system?

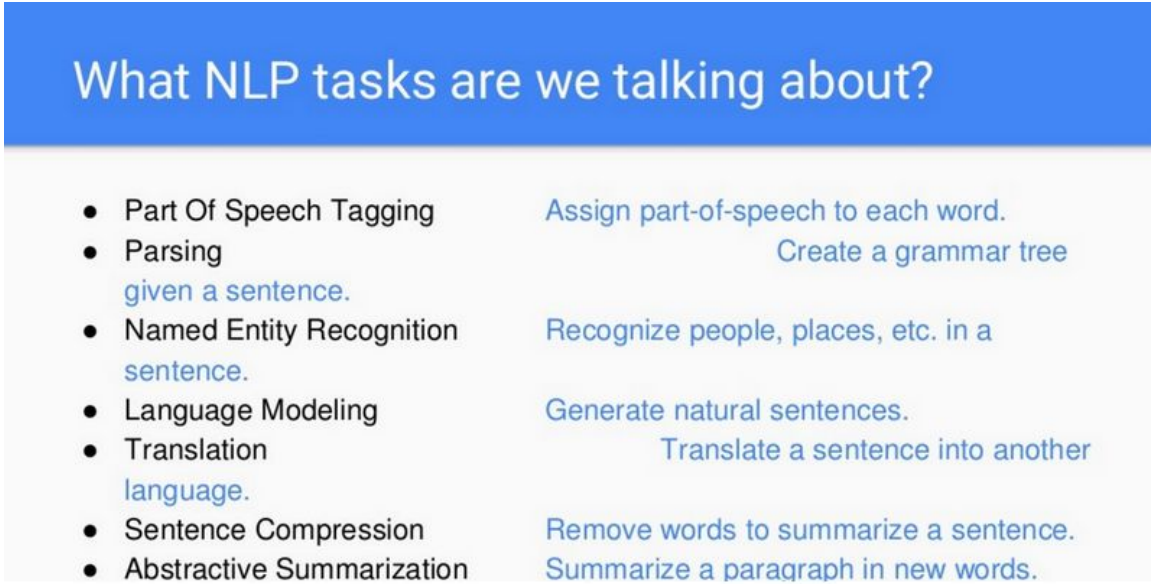
Text Mining

- Thankfully, computer science has a dedicated subfield of research that works on making sense from natural language and unstructured text.
- The umbrella term is Text Mining



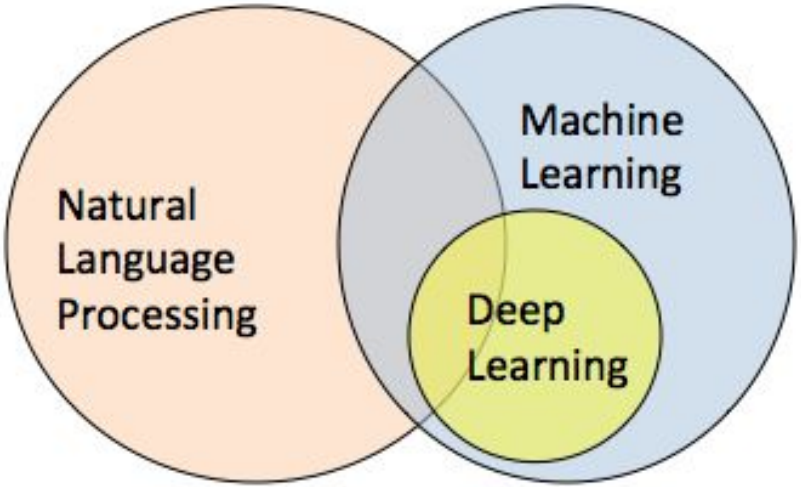
Natural Language Processing

- Under Text Mining, there is another subfield called Natural Language Processing, which performs further analysis on the parsed text.



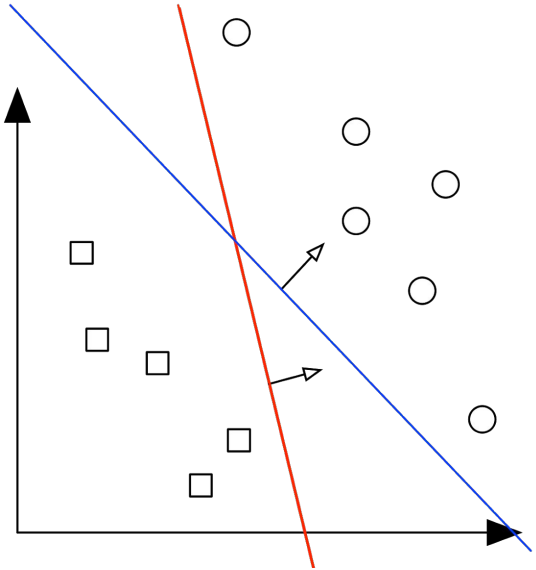
Machine Learning in NLP

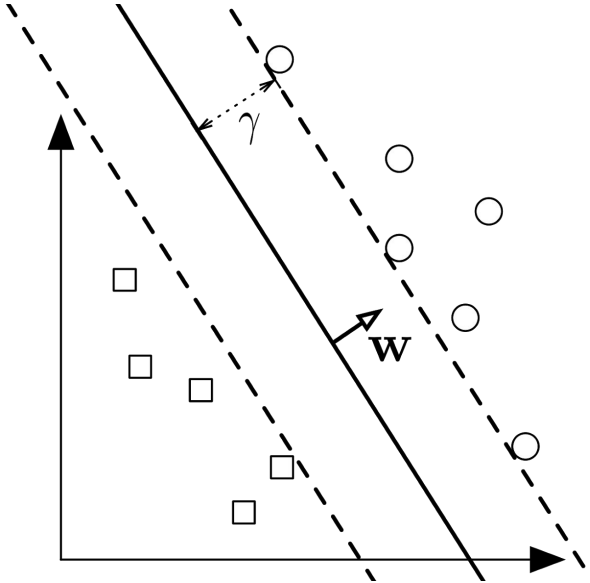
- NLP utilizes Machine Learning, which is a subfield under Artificial Intelligence to perform many tasks.



Machine Learning in NLP

- A “Classifier” in Machine Learning is any agent that can learn from to make decisions given previous data.





Proposed Method

Method Overview

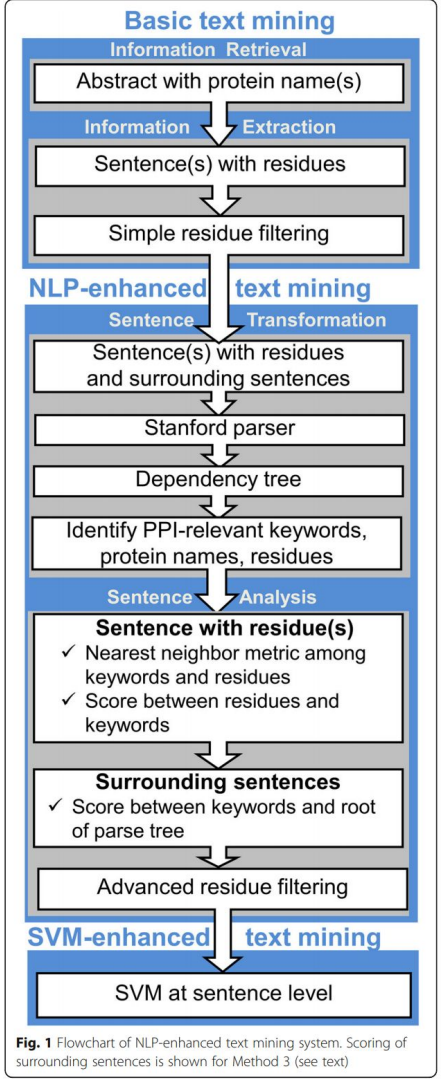
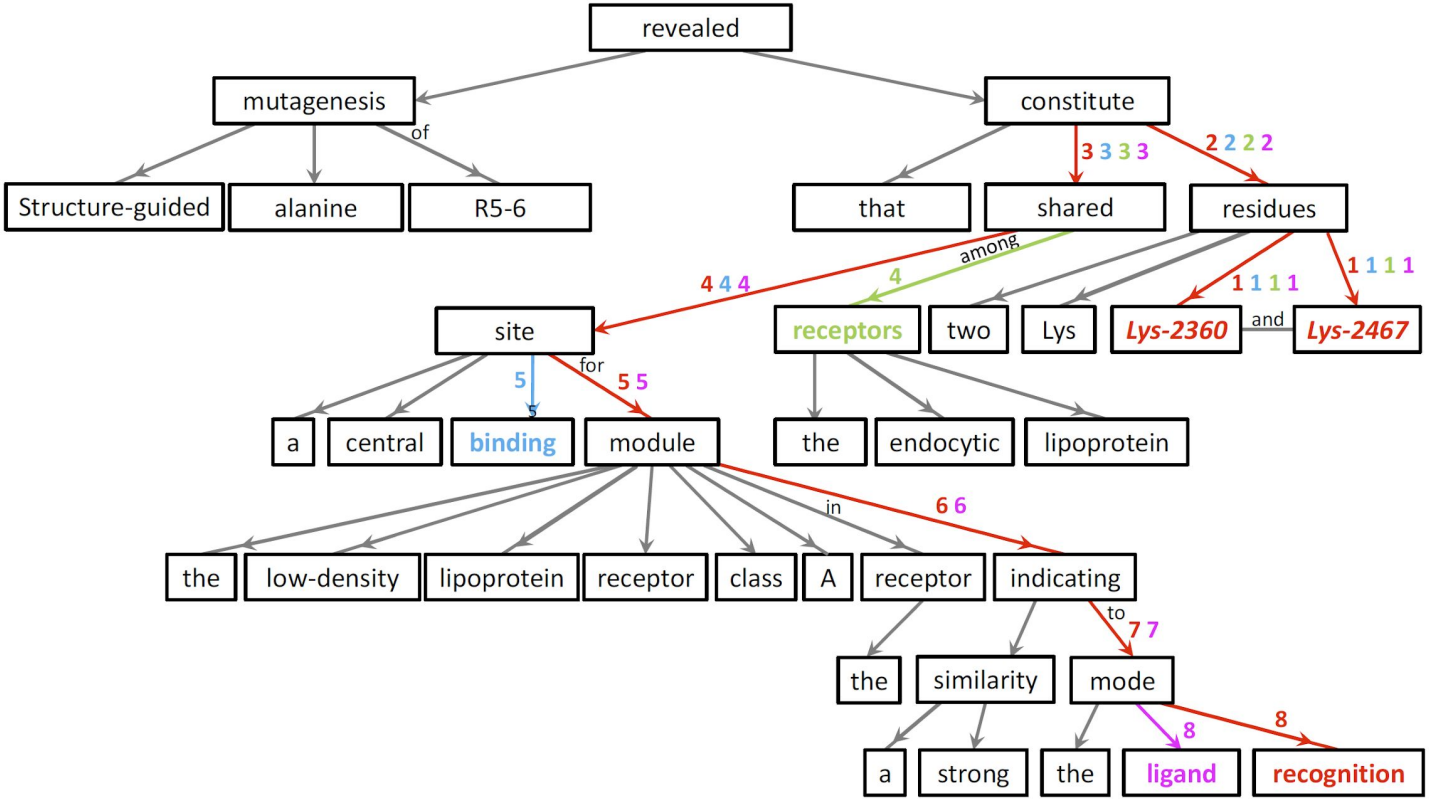


Fig. 1 Flowchart of NLP-enhanced text mining system. Scoring of surrounding sentences is shown for Method 3 (see text)

Data Gathering or Information Retrieval

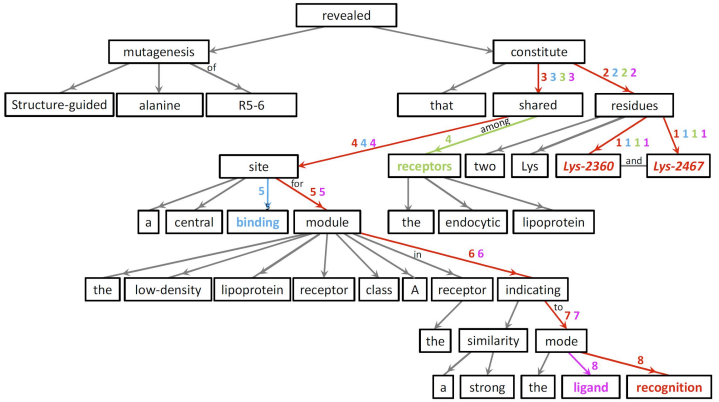
- Gather all papers using two queries:
  - AND Query
    - Abstract contains the name of both proteins of interest.
  - OR Query
    - Abstract contains name of either protein of interest.
- Using NCBI e-Utilities Tool
- Once abstracts are found, the text of the abstracts is processed for residue names.
- **PROBLEM:** Not all residue names will be relevant to the binding process of both proteins!

Perform NLP


revealed
mutagenesis
of
constitute
Structure-guided
alanine
R5-6
that
shared
residues
3 3 3 3
2 2 2 2
among
4
4 4 4
1 1 1 1
1 1 1 1
site
receptors
two
Lys
Lys-2360
and
Lys-2467
for
5
5 5
a
central
binding
module
the
endocytic
lipoprotein
in
6 6
the
low-density
lipoprotein
receptor
class
A
receptor
indicating
to
7 7
the
similarity
mode
8
8
a
strong
the
ligand
recognition

Perform NLP

- Generated parse tree is used to “rank” residues according to a particular score.
- Intuitive Ranking: If a residue name is “close” to a word that resembles interaction (like *bind, interface, complex, recept, cocntact, recog, dock etc.*)
  - Give it a higher score, and hence higher rank
- Otherwise if its close to a Protein-Protein Interaction (PPI) negative word (like *deamidation, dissociat, antibo etc*)
  - Give a lower score



Calculate Features

- Machine Learning models (like Support Vector Machines (SVMs)) cannot be trained directly on text data.
  - They work on numbers, and text != numbers.
- We have to convert each sentence which contains residue name into a *number.*
- These “Numbers” are called feature vectors.
- The scores that we calculated in last slide are used as these “feature vectors” to train an SVM.
- **Question: What does this SVM learn to do then?**

Calculate Features

- **Question: What does this SVM learn to do then?**
- **Answer: Let's visualize this using a simple example.**

Sentence	**Score**	**Label (Contains person name or not?)**
My name is Michael D'Souza	0.9	YES
My daughter is turning 13 this month	0.3	NO

Calculate Features

- **In our application, the table looks something like:**

Sentence	**Score**	**Label (Contains interface residue or non-interface residue?)**
Structure-guided alanine mutagenesis of R5-6 revealed that two Lys residues (Lys-2360 and Lys-2467) constitute a central binding site for the low-density lipoprotein receptor class	0.9	YES
A module in the receptor, indicating a strong similarity to the ligand recognition mode shared among the endocytic lipoprotein receptors	0.3	NO

Calculate Features

- **Train the SVM to predict the label, given the score value of a sentence!**

Sentence	**Score**	**Label (Contains interface residue or non-interface residue?)**
Structure-guided alanine mutagenesis of R5-6 revealed that two Lys residues (Lys-2360 and Lys-2467) constitute a central binding site for the low-density lipoprotein receptor class	0.9	YES
A module in the receptor, indicating a strong similarity to the ligand recognition mode shared among the endocytic lipoprotein receptors	0.3	NO

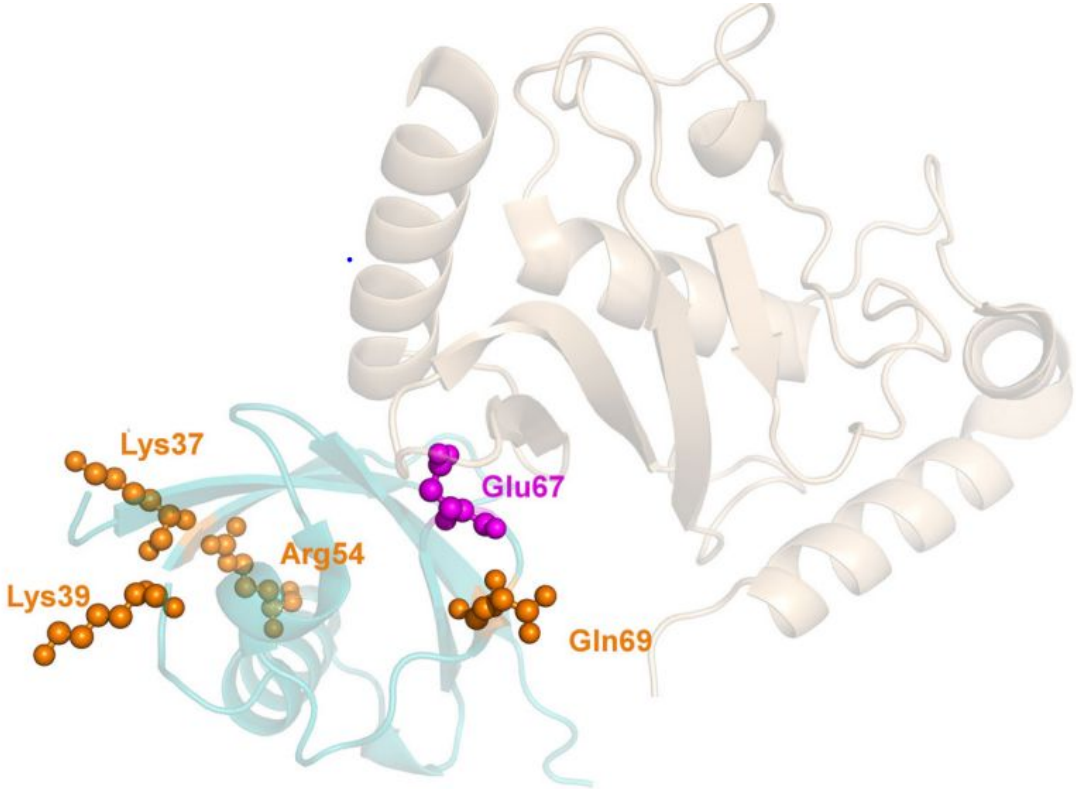
Trained SVM used for Filtering

- Once the SVM is trained it is used to filter the residues.
- Question: How?
- Answer:
  - Given a sentence: “Structure-guided alanine mutagenesis of R5-6 revealed that two Lys residues (Lys-2360 and Lys-2467) constitute a central binding site”
  - Are Lys-2360 and Lys-2467 Interface-Residues OR Non-Interface Residues?

What Next?

- Use this filtering technique to build a database of knowledge that can be used to constrain Docking methods.

What Next?



How Good is this Method?

- Pretty good!
- The authors show good improvement in docking accuracy when there are additional constraints applied that were mined/processed from unstructured text in publications!

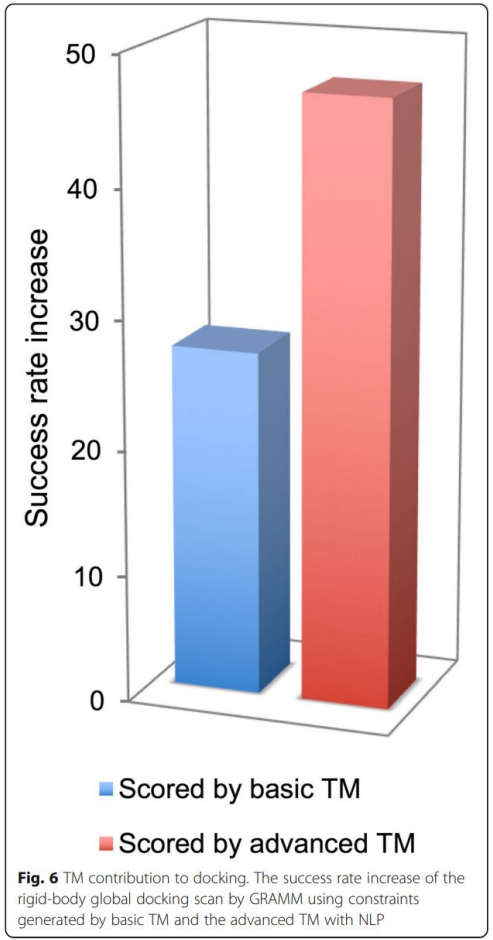


Fig. 6 TM contribution to docking. The success rate increase of the rigid-body global docking scan by GRAMM using constraints generated by basic TM and the advanced TM with NLP

Thank You!

Questions?

Contact:
Anmol Sharma
anmol_sharma@sfu.ca