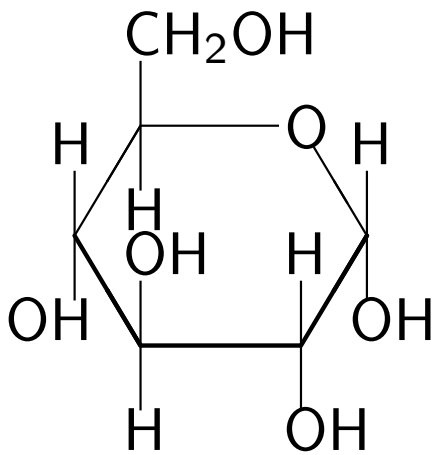
A Pattern Recognition Based Model for Characterizing and Predicting Glucose-Binding Sites

Houssam Nassif

June 7, 2006



Outline

Outline

Outline

Outline

Problem Description

- Glucose is a 6-carbon sugar molecule that plays a key role in many different biochemical pathways.

Problem Description

- Glucose is a 6-carbon sugar molecule that plays a key role in many different biochemical pathways.
- Glucose binds to protein molecules at specific binding sites.

Problem Description

- Glucose is a 6-carbon sugar molecule that plays a key role in many different biochemical pathways.
- Glucose binds to protein molecules at specific binding sites.
- Binding sites are specific to their respective ligands.

Definition

Ligand: The specific molecule that binds to the binding site.

Problem Description

- Glucose is a 6-carbon sugar molecule that plays a key role in many different biochemical pathways.
- Glucose binds to protein molecules at specific binding sites.
- Binding sites are specific to their respective ligands.
- Identifying the ligand docking at a certain binding site remains an open research problem.

Definition

Ligand: The specific molecule that binds to the binding site.

Problem Description

- Glucose is a 6-carbon sugar molecule that plays a key role in many different biochemical pathways.
- Glucose binds to protein molecules at specific binding sites.
- Binding sites are specific to their respective ligands.
- Identifying the ligand docking at a certain binding site remains an open research problem.

Definition

Ligand: The specific molecule that binds to the binding site.

Goal

Identify glucose-binding sites.

Biochemical Background

A glucose-binding site involves:

Biochemical Background

A glucose-binding site involves:

1. A glucose molecule.

Biochemical Background

A glucose-binding site involves:

1. A glucose molecule.
2. A protein.

Biochemical Background

A glucose-binding site involves:

1. A glucose molecule.
2. A protein.
3. Their interaction.

Glucose Structure

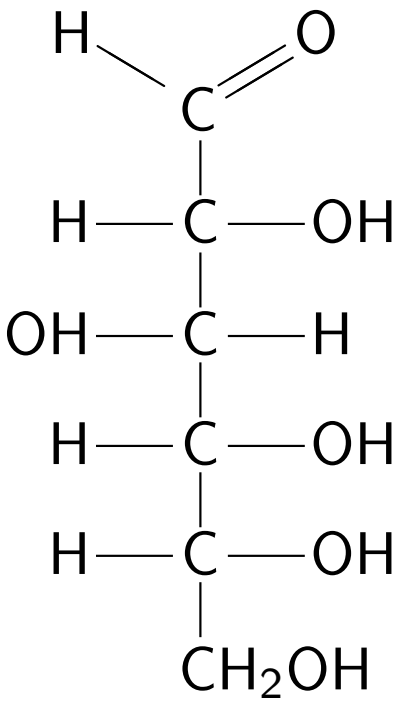


Figure: Glucose

- Glucose is a 6-carbon sugar.

Glucose Structure

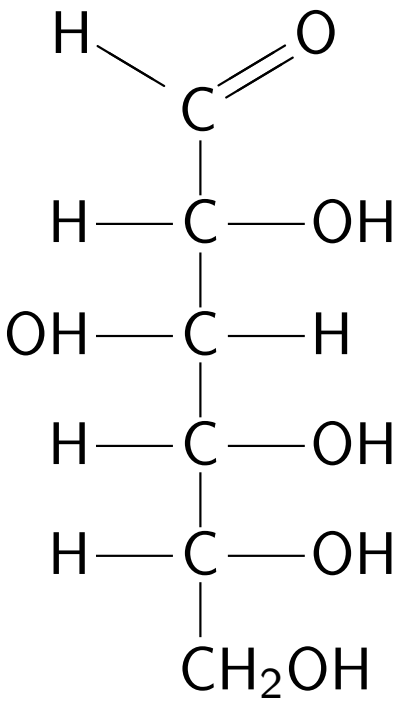


Figure: Glucose

- Glucose is a 6-carbon sugar.
- It contains two functional groups.

Glucose Structure

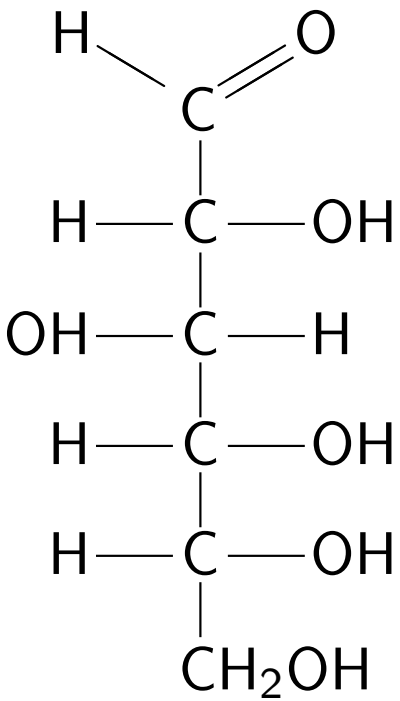
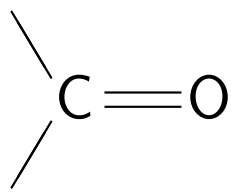


Figure: Glucose

- Glucose is a 6-carbon sugar.
- It contains two functional groups.



(a)
Carbonyl

Glucose Structure

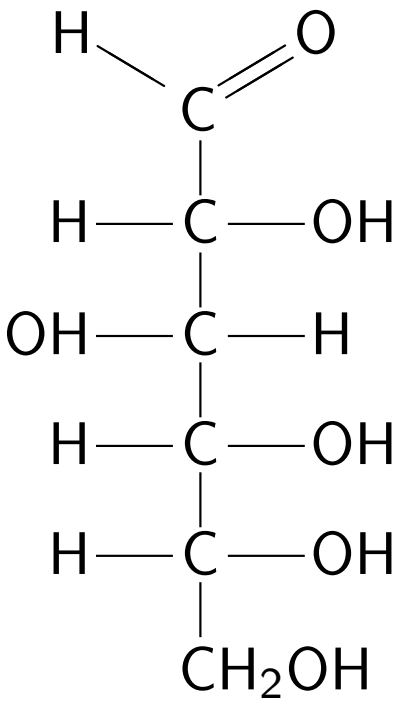
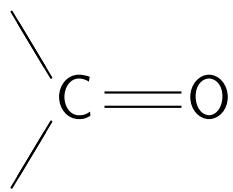
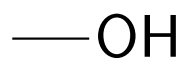


Figure: Glucose

- Glucose is a 6-carbon sugar.
- It contains two functional groups.



(a)
Carbonyl



(b)
Hydroxyl

Glucose Structure

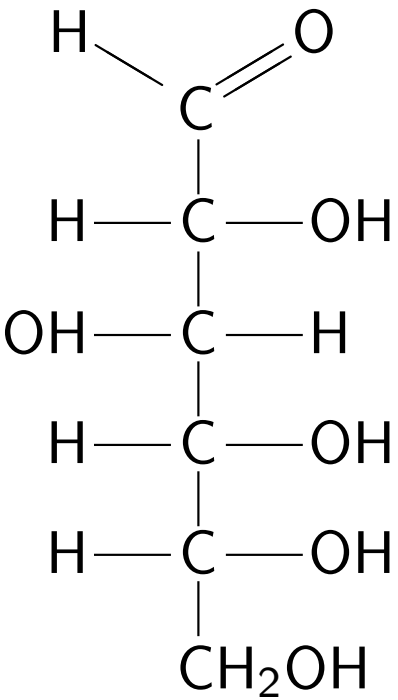
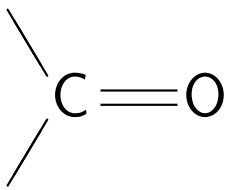
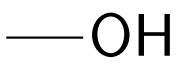


Figure: Glucose

- Glucose is a 6-carbon sugar.
- It contains two functional groups.
- Both groups can interact together.



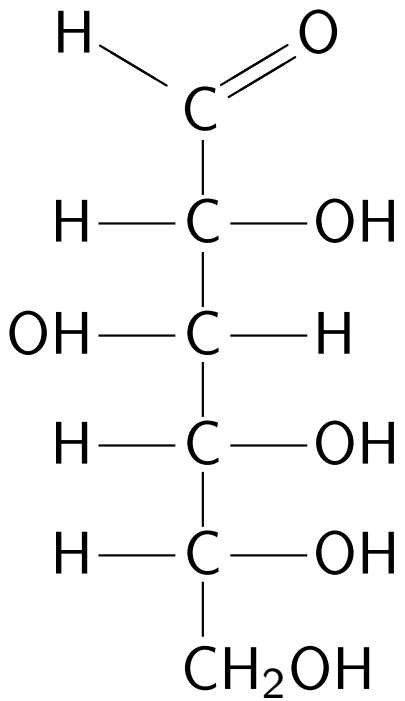
(a)
Carbonyl



(b)
Hydroxyl

Glucose Cyclization

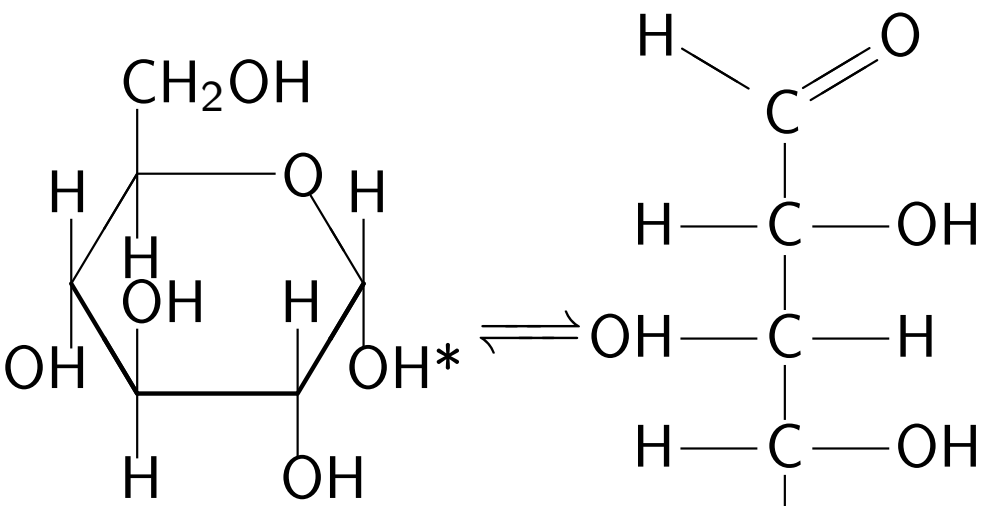
- The molecule folds on itself and forms a *pyranose* ring.



(b) Open chain

Glucose Cyclization

- The molecule folds on itself and forms a *pyranose* ring.

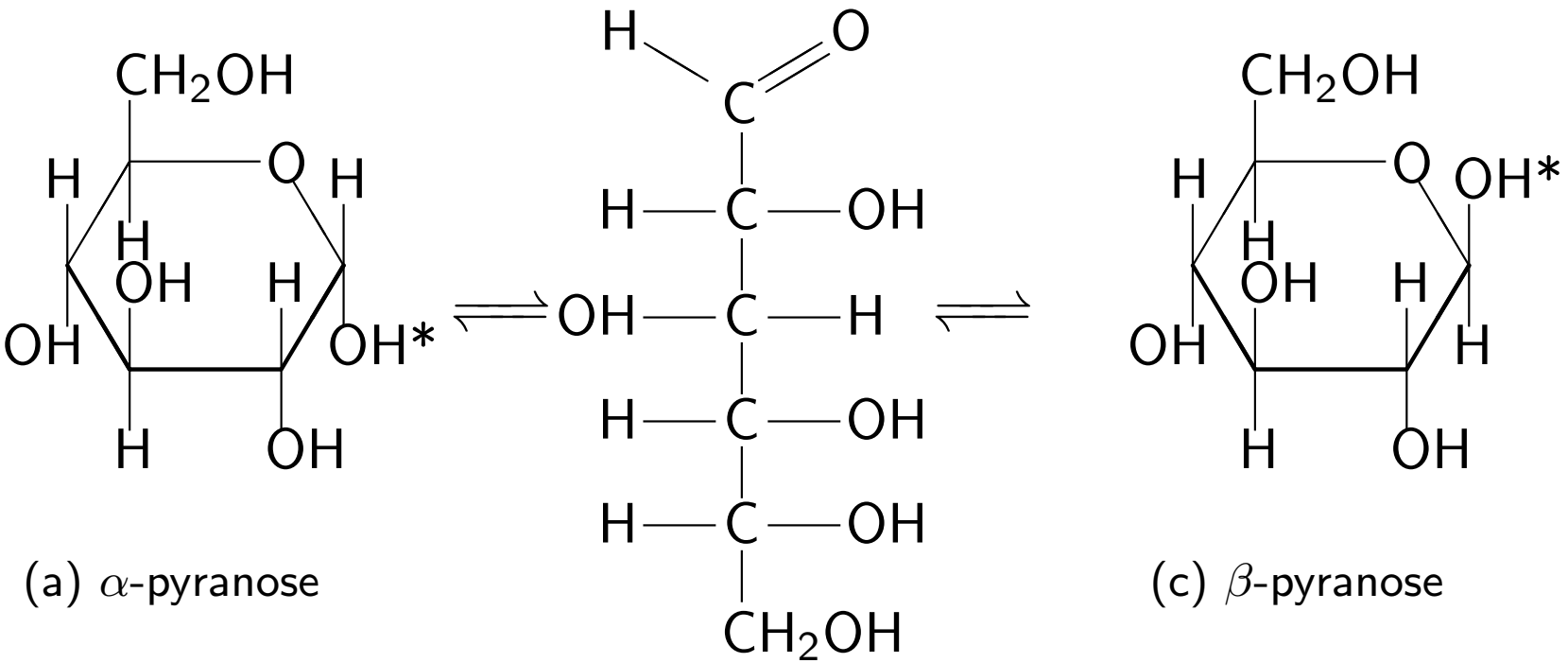


(a) α-pyranose

(b) Open chain

Glucose Cyclization

- The molecule folds on itself and forms a *pyranose* ring.
- In two different ways. Watch the star!



(a) α-pyranose

(b) Open chain

(c) β-pyranose

Conformation Shift

- Glucose readily shifts from one conformation to another.

Conformation Shift

- Glucose readily shifts from one conformation to another.
- In physiological solutions: almost exclusively in the pyranose ring form.

Conformation Shift

- Glucose readily shifts from one conformation to another.
- In physiological solutions: almost exclusively in the pyranose ring form.
  - 36% α-pyranose

Conformation Shift

- Glucose readily shifts from one conformation to another.
- In physiological solutions: almost exclusively in the pyranose ring form.
  - 36% α-pyranose
  - 64% β-pyranose

Amino Acid Structure

- An amino acid consists of:

Amino Acid Structure

- An amino acid consists of:
  - a central carbon atom C, bonded to:

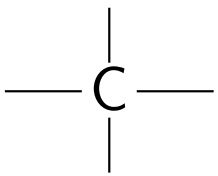


Figure: Amino acid

Amino Acid Structure

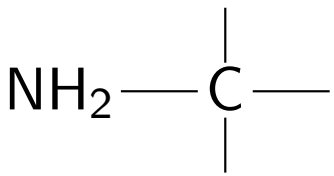


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2

Amino Acid Structure

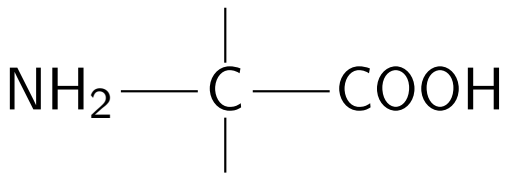


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH

Amino Acid Structure

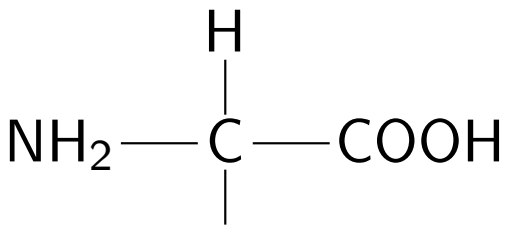


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H

Amino Acid Structure

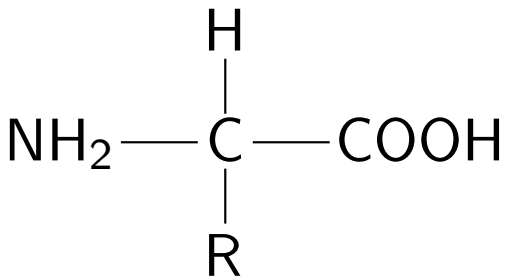


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H
  - and a side chains R.

Amino Acid Structure

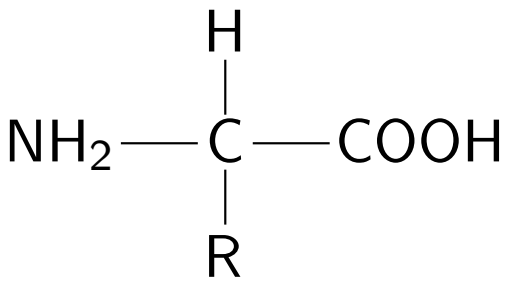


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H
  - and a side chains R.
- Amino acids differ by their side chain R.

Amino Acid Structure

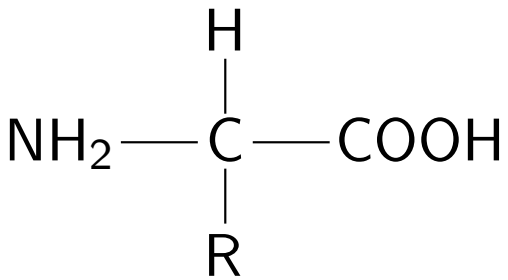


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H
  - and a side chains R.
- Amino acids differ by their side chain R.
- The side chain confers to each amino acid its distinctive properties.

Amino Acid Structure

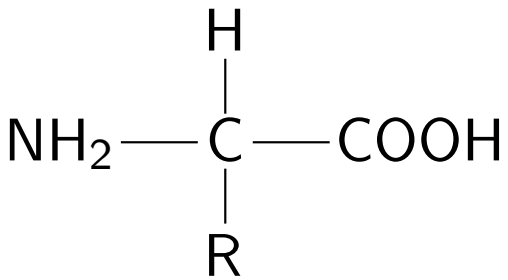


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H
  - and a side chains R.
- Amino acids differ by their side chain R.
- The side chain confers to each amino acid its distinctive properties.
- There are 20 different amino acids.

Amino Acid Structure

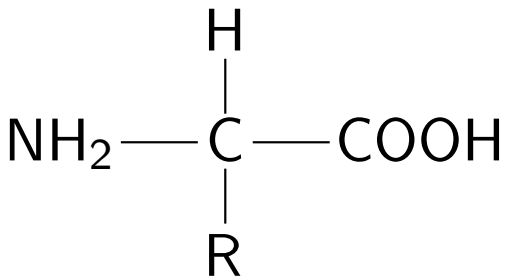


Figure: Amino acid

- An amino acid consists of:
  - a central carbon atom C, bonded to:
  - an amino group NH_2
  - a carboxyl group COOH
  - a hydrogen atom H
  - and a side chains R.
- Amino acids differ by their side chain R.
- The side chain confers to each amino acid its distinctive properties.
- There are 20 different amino acids.

Amino acid properties

Similar R $\Leftrightarrow$ similar properties. Different R $\Leftrightarrow$ different properties.

Protein Structure

- Protein: a long chain of amino acids linked together.

Protein Structure

- Protein: a long chain of amino acids linked together.

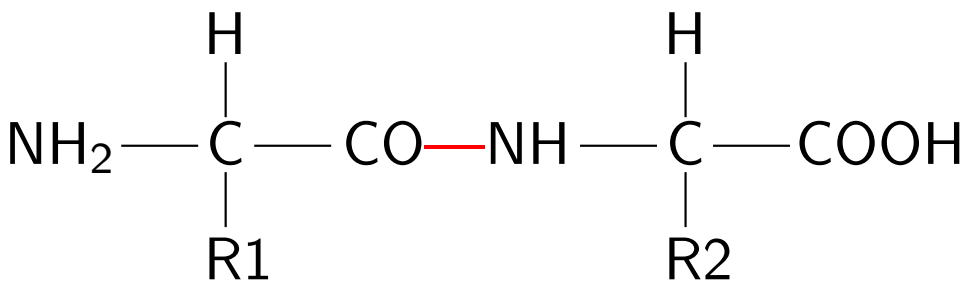


Figure: Linked residues

Protein Structure

- Protein: a long chain of amino acids linked together.
- Amino acids in a protein are also called residues.

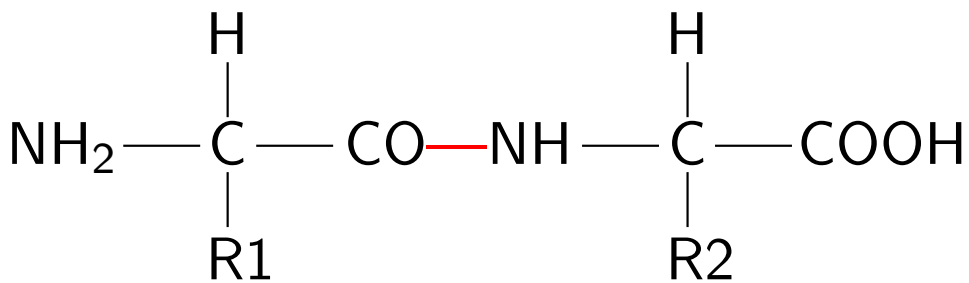


Figure: Linked residues

Protein Structure

- Protein: a long chain of amino acids linked together.
- Amino acids in a protein are also called residues.
- Residues sequence and properties determine the protein shape and function.

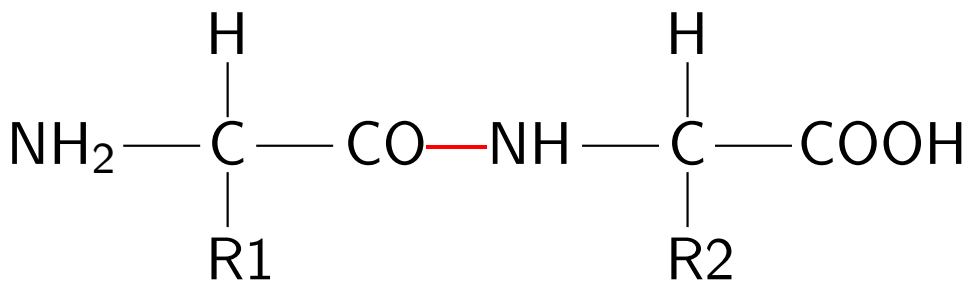


Figure: Linked residues

Protein Structure

- Protein: a long chain of amino acids linked together.
- Amino acids in a protein are also called residues.
- Residues sequence and properties determine the protein shape and function.

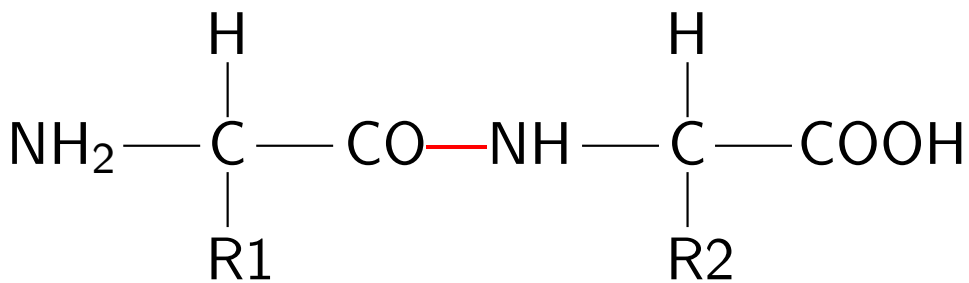


Figure: Linked residues

Important Fact

Similar residues can be easily interchanged in a protein.

Atomic Interactions

Atoms interact together through:

Atomic Interactions

Atoms interact together through:

1. Chemical bonds

Atomic Interactions

Atoms interact together through:

1. Chemical bonds

2. Forces and Interactions

Atomic Interactions

Atoms interact together through:

1. Chemical bonds
    1. Covalent Bonds
2. Forces and Interactions

Atomic Interactions

Atoms interact together through:

1. Chemical bonds
    1. Covalent Bonds
    2. Hydrogen Bonds
2. Forces and Interactions

Atomic Interactions

Atoms interact together through:

1. Chemical bonds
   1. Covalent Bonds
   2. Hydrogen Bonds
2. Forces and Interactions
   1. Van der Waals Forces

Atomic Interactions

Atoms interact together through:

1. Chemical bonds
   1. Covalent Bonds
   2. Hydrogen Bonds
2. Forces and Interactions
   1. Van der Waals Forces
   2. Hydrophobic Interactions

Covalent Bonds

- Close and strong interaction.

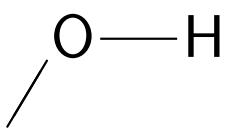


Figure: Covalent bond

Covalent Bonds

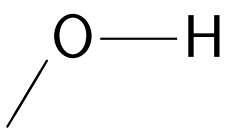


Figure: Covalent bond

- Close and strong interaction.
- Form a molecule.

Covalent Bonds

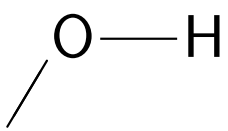


Figure: Covalent bond

- Close and strong interaction.
- Form a molecule.
- Atoms share electrons.

Covalent Bonds

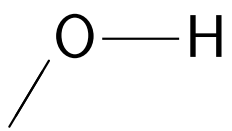


Figure: Covalent bond

- Close and strong interaction.
- Form a molecule.
- Atoms share electrons.
- Electronegativity.

Definition

Electronegativity: A measure of an atom's attraction for electrons

Covalent Bonds

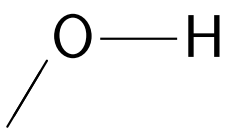


Figure: Covalent bond

- Close and strong interaction.
- Form a molecule.
- Atoms share electrons.
- Electronegativity.
  - Equal $\Rightarrow$ nonpolar

Definition

Electronegativity: A measure of an atom's attraction for electrons

Covalent Bonds

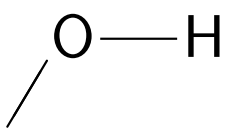


Figure: Covalent bond

- Close and strong interaction.
- Form a molecule.
- Atoms share electrons.
- Electronegativity.
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar

Definition

Electronegativity: A measure of an atom's attraction for electrons

Covalent Bonds

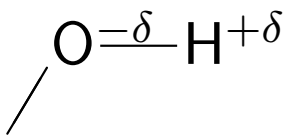


Figure: Covalent polar bond

- Close and strong interaction.
- Form a molecule.
- Atoms share electrons.
- Electronegativity.
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges.

Definition

Electronegativity: A measure of an atom's attraction for electrons

Hydrogen Bonds

- The most relevant in protein interaction.

Hydrogen Bonds

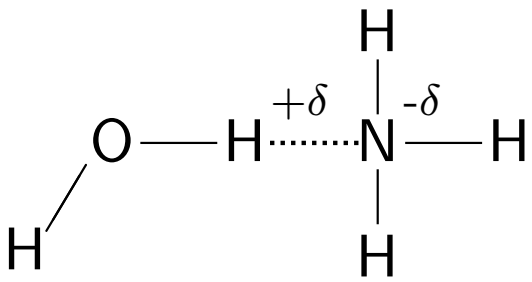


Figure: Hydrogen bond

- The most relevant in protein interaction.
- Attraction between a positively charged H and a negatively charged atoms.

Hydrogen Bonds

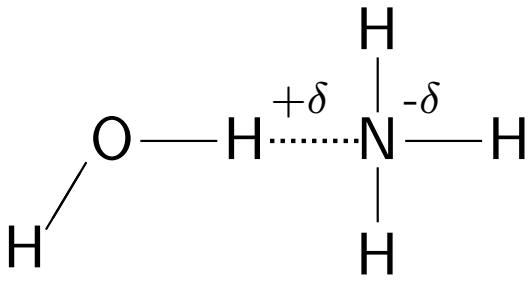


Figure: Hydrogen bond

- The most relevant in protein interaction.
- Attraction between a positively charged H and a negatively charged atoms.
- Glucose attaches to the protein using hydrogen bonds.

Forces and Interactions

1. Van der Waals Forces

Forces and Interactions

1. Van der Waals Forces

Characteristics

Weak electrostatic attraction and repulsion forces.

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions

Definition

Hydrophobic: water hating. *Hydrophilic*: water loving.

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions

Definition

Hydrophobic: water hating. *Hydrophilic*: water loving.

Characteristics

- Hydrophobic atoms tend to gather together.

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions

Definition

Hydrophobic: water hating. *Hydrophilic*: water loving.

Characteristics

- Hydrophobic atoms tend to gather together.
- Hydrophilic atoms tend to gather together.

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions
3. The pyranose ring is hydrophobic

Definition

Hydrophobic: water hating. *Hydrophilic*: water loving.

Forces and Interactions

1. Van der Waals Forces
2. Hydrophobic Interactions
3. The pyranose ring is hydrophobic

Definition

Hydrophobic: water hating. *Hydrophilic*: water loving.

Characteristics

$\Rightarrow$ attracted by hydrophobic residues

Pattern Recognition

Pattern recognition

Definition

Pattern recognition: The computer learns how to correctly classify an object.

Pattern Recognition

Pattern recognition includes:

1. Classifier

Pattern Recognition

Pattern recognition includes:

1. Classifier

Classifier phases

- *Training phase*: Learns how to partition the feature space.

Pattern Recognition

Pattern recognition includes:

1. Classifier

Classifier phases

- *Training phase*: Learns how to partition the feature space.
- *Testing phase*: Predicts the class of an unknown data.

Pattern Recognition

Pattern recognition includes:

1. Classifier
   - k-Nearest-Neighbor (kNN)

Classifier phases

- *Training phase*: Learns how to partition the feature space.
- *Testing phase*: Predicts the class of an unknown data.

Pattern Recognition

Pattern recognition includes:

1. Classifier
   - k-Nearest-Neighbor (kNN)
   - Support Vector Machines (SVM)

Classifier phases

- *Training phase*: Learns how to partition the feature space.
- *Testing phase*: Predicts the class of an unknown data.

Pattern Recognition

Pattern recognition includes:

1. Classifier
   - k-Nearest-Neighbor (kNN)
   - Support Vector Machines (SVM)
2. Feature selection

Classifier phases

- *Training phase*: Learns how to partition the feature space.
- *Testing phase*: Predicts the class of an unknown data.

Pattern Recognition

Pattern recognition includes:

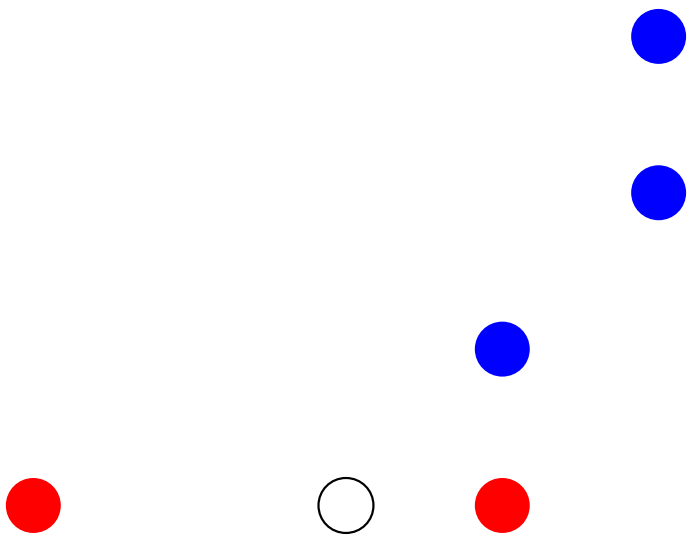
1. Classifier
   - k-Nearest-Neighbor (kNN)
   - Support Vector Machines (SVM)
2. Feature selection
   - Random Forest (RF)

Classifier phases

- *Training phase*: Learns how to partition the feature space.
- *Testing phase*: Predicts the class of an unknown data.

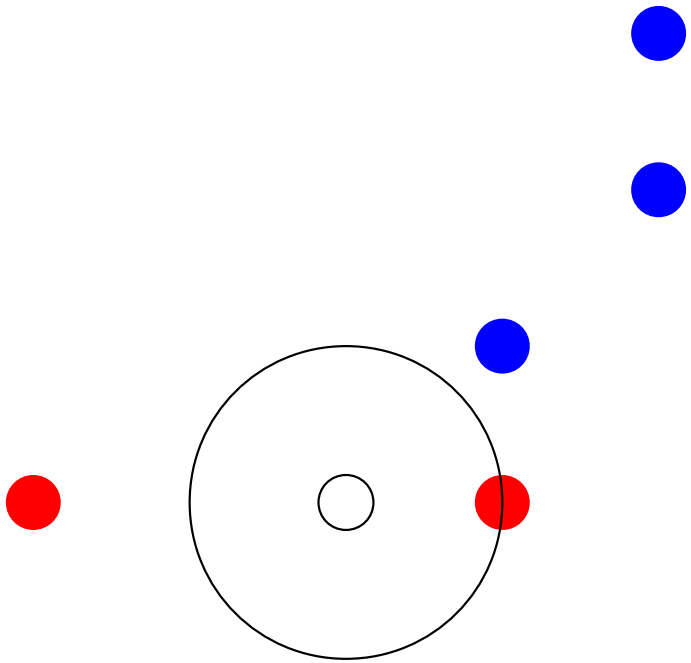
k-Nearest-Neighbors (kNN)

Simple kNN technique



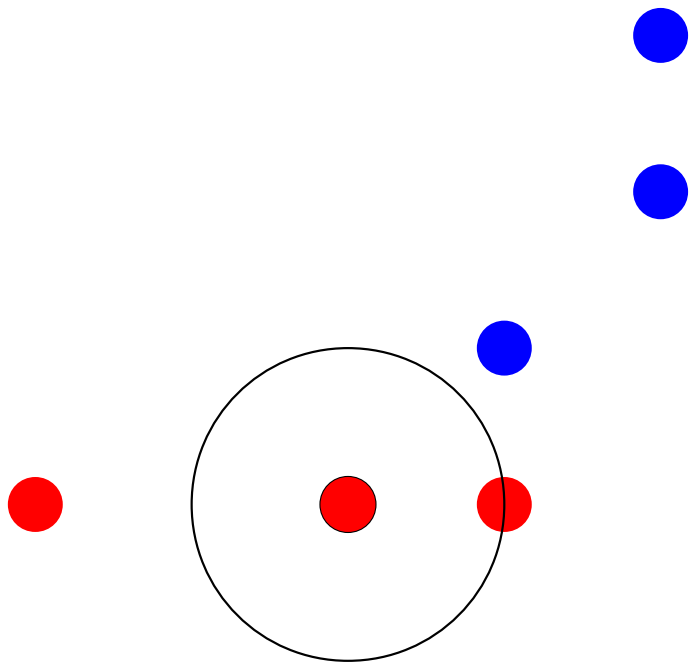
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 1$



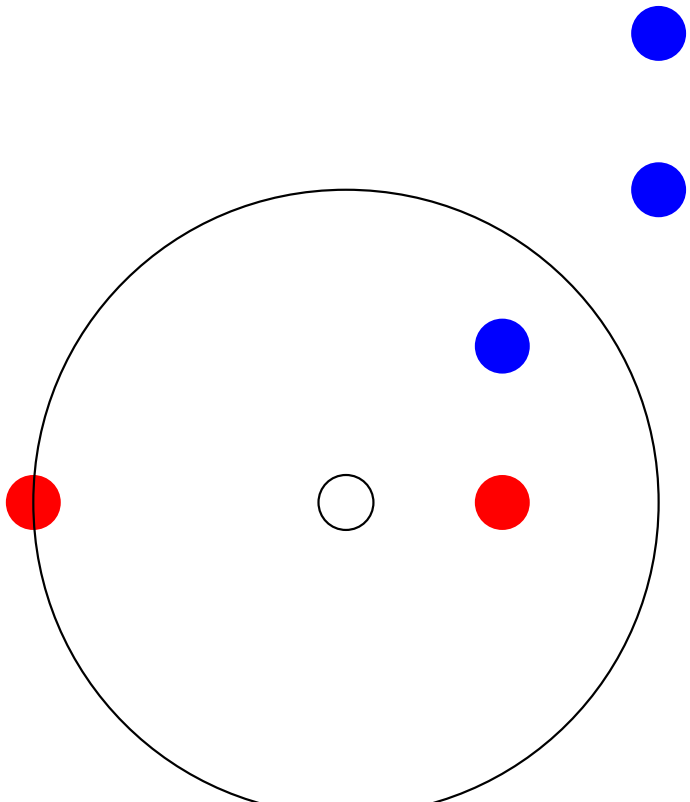
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 1$



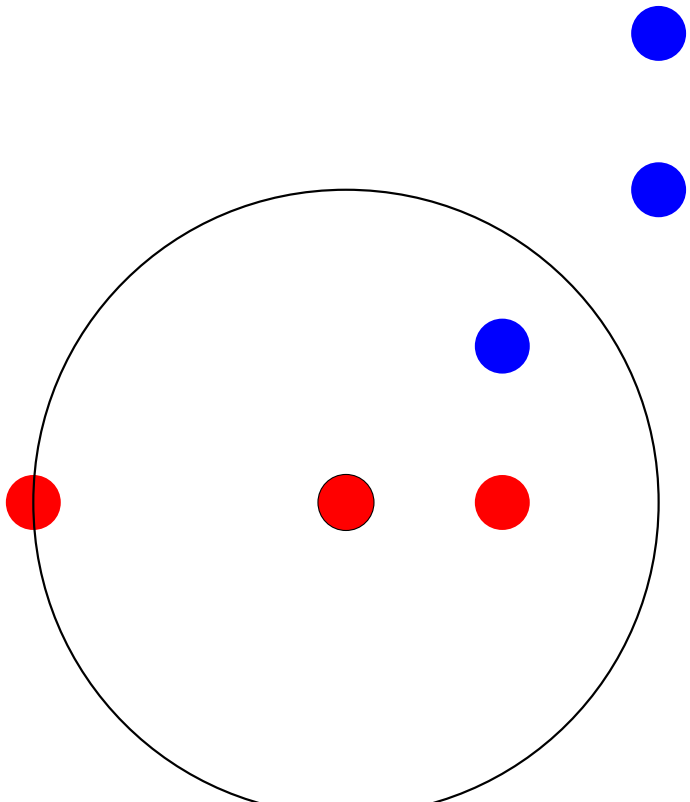
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 3$



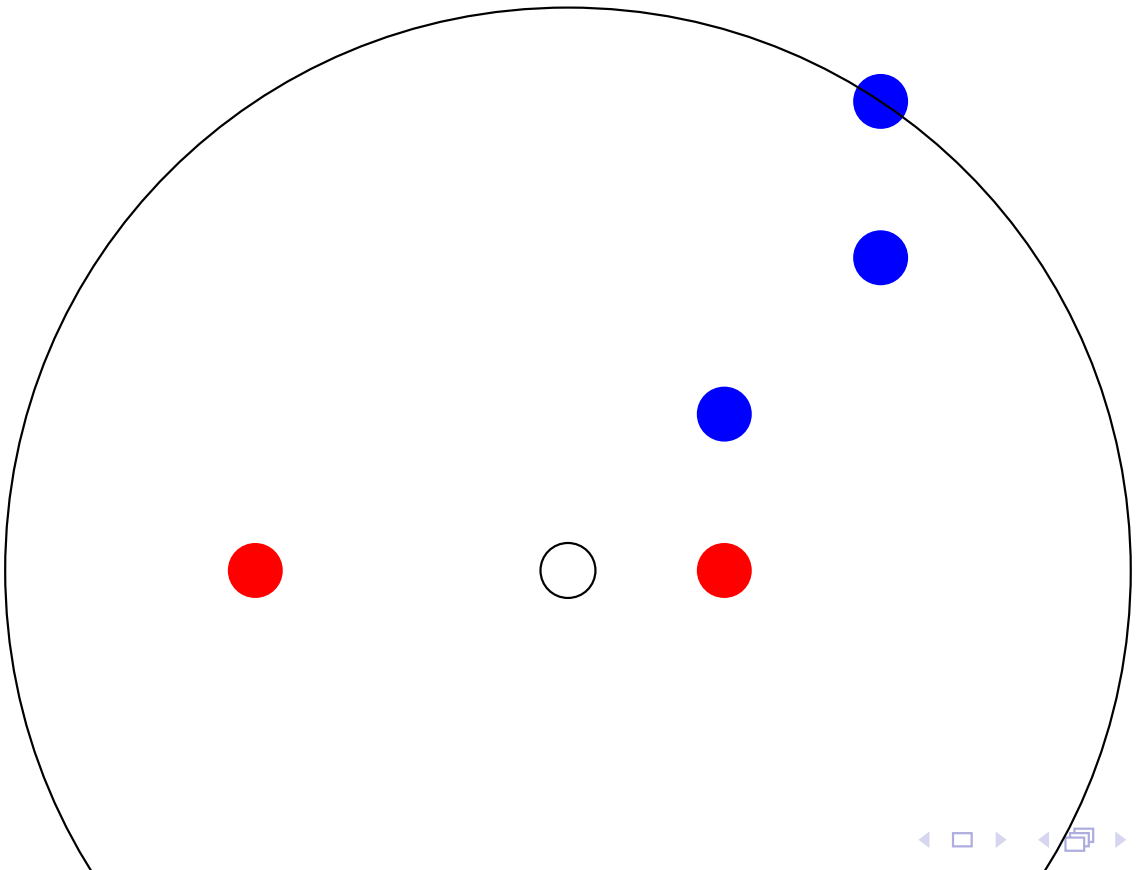
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 3$



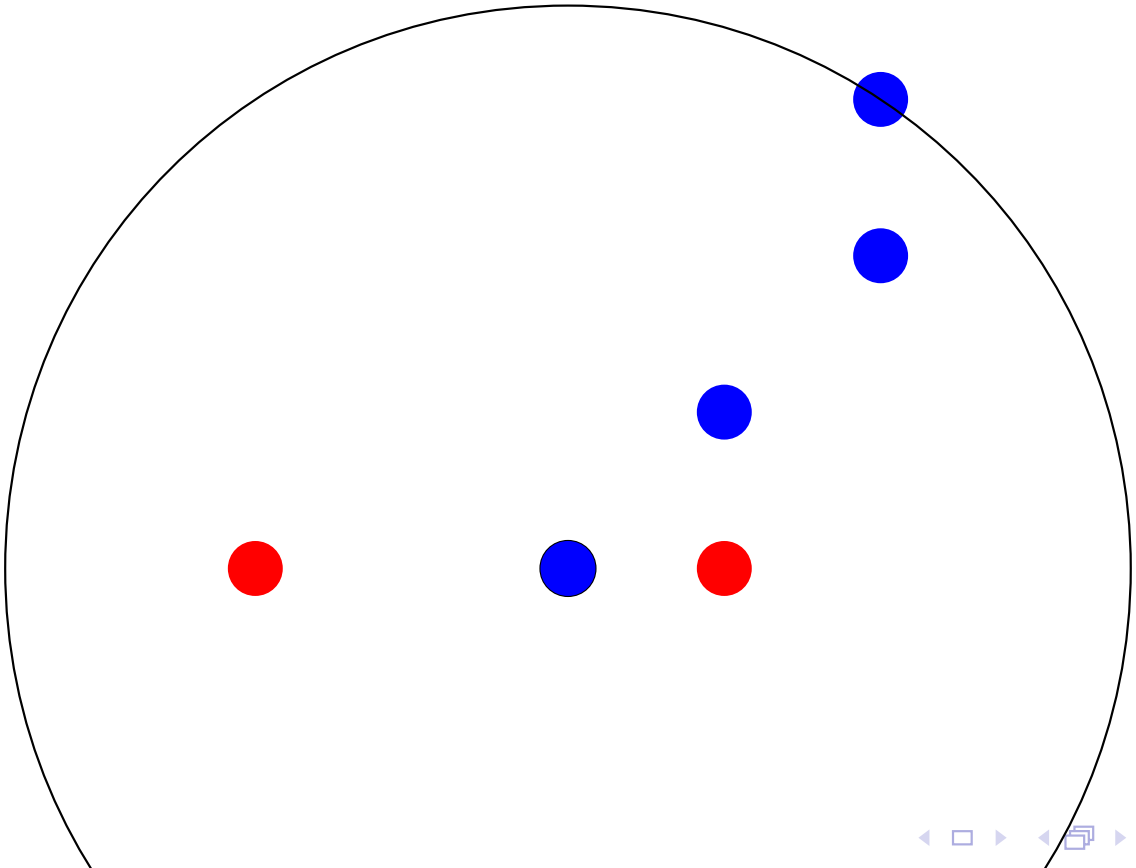
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 5$



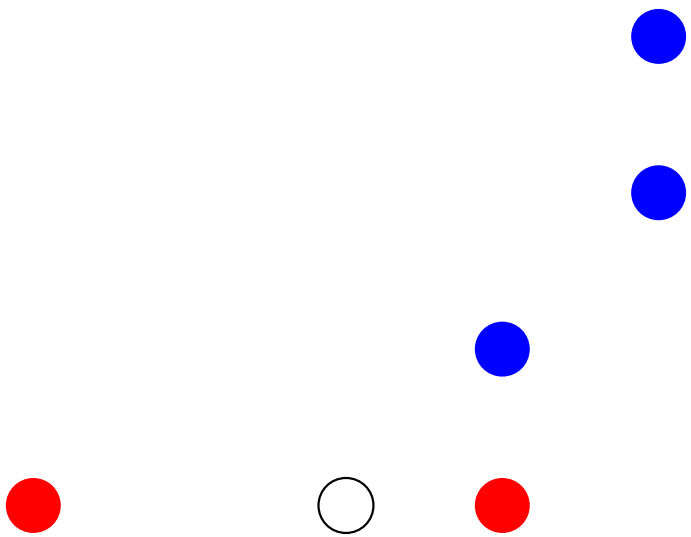
k-Nearest-Neighbors (kNN)

Simple kNN technique, $k = 5$



k-Nearest-Neighbors (kNN)

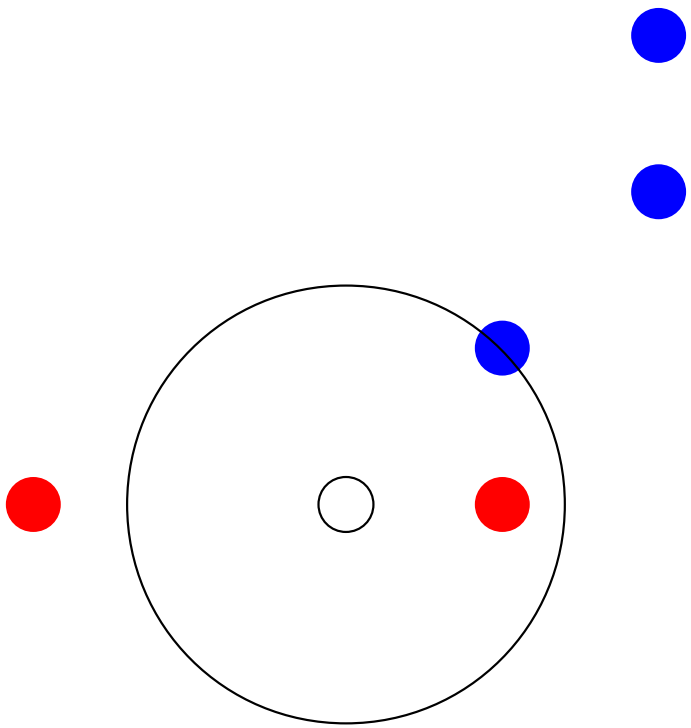
Simple kNN technique limitations



k-Nearest-Neighbors (kNN)

Simple kNN technique limitations

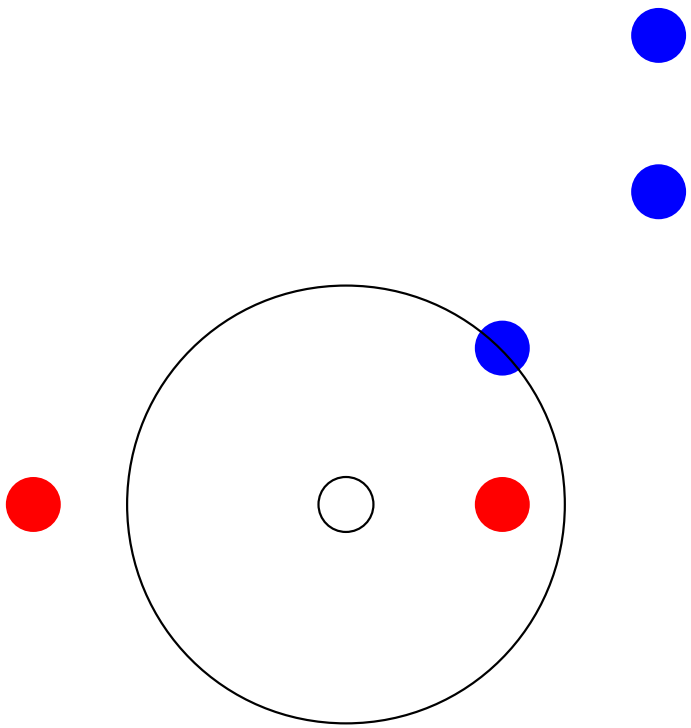
- Cannot handle even k's.



k-Nearest-Neighbors (kNN)

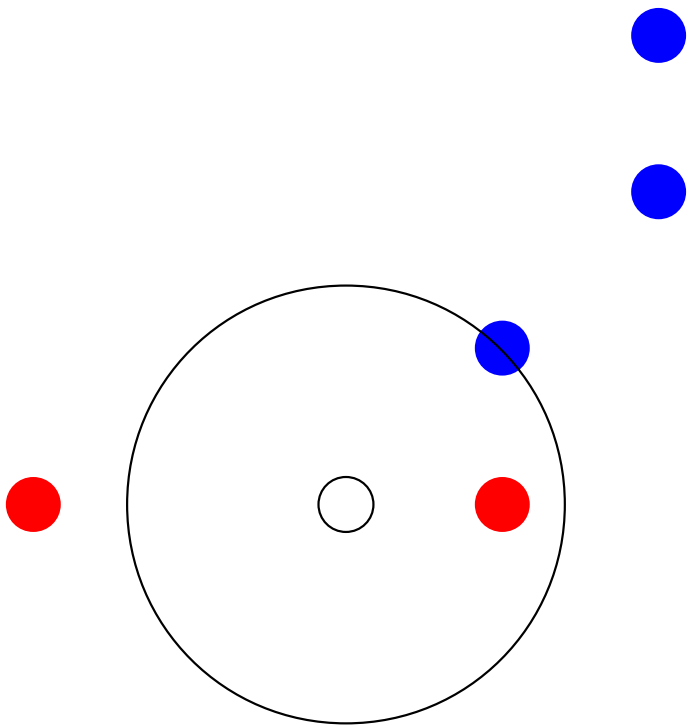
Simple kNN technique limitations

- Cannot handle even k's.
- Ignores distance!



k-Nearest-Neighbors (kNN)

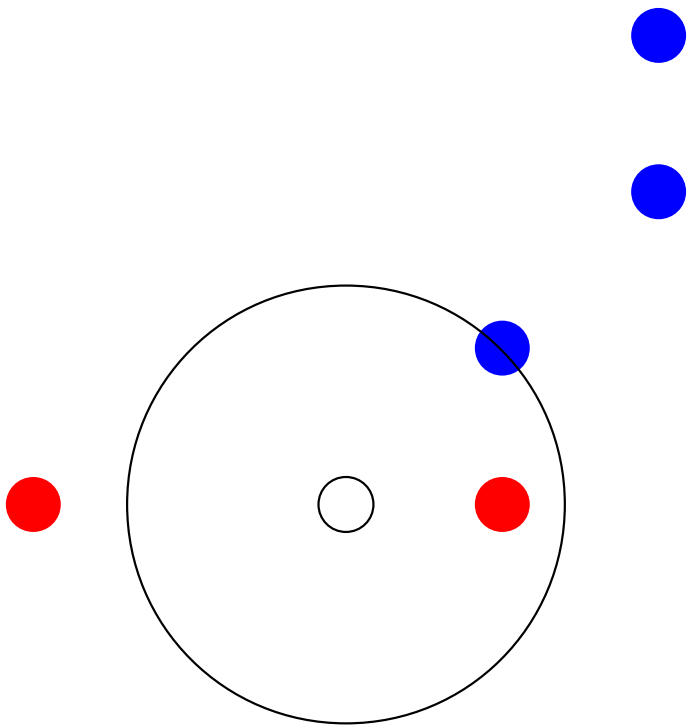
Solution: Weighted kNN



k-Nearest-Neighbors (kNN)

Solution: Weighted kNN

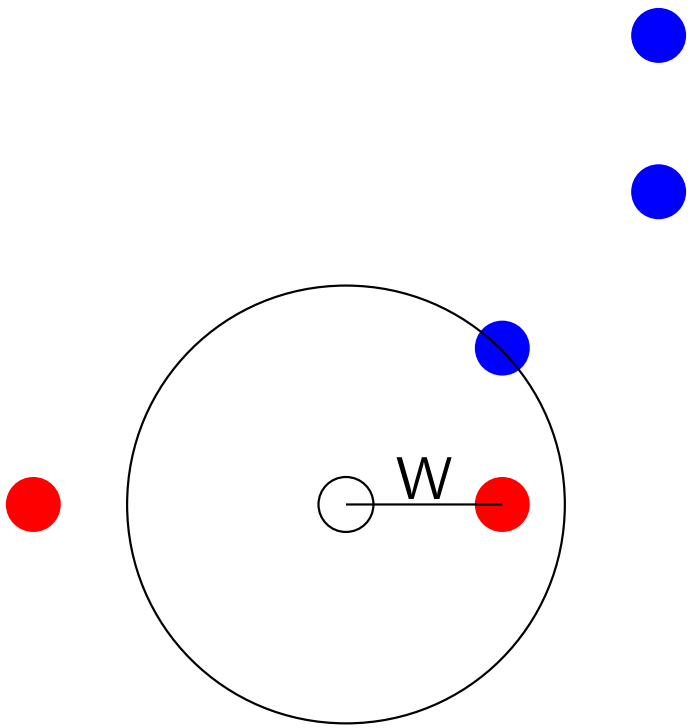
- Puts a higher weight on lesser distances



k-Nearest-Neighbors (kNN)

Solution: Weighted kNN

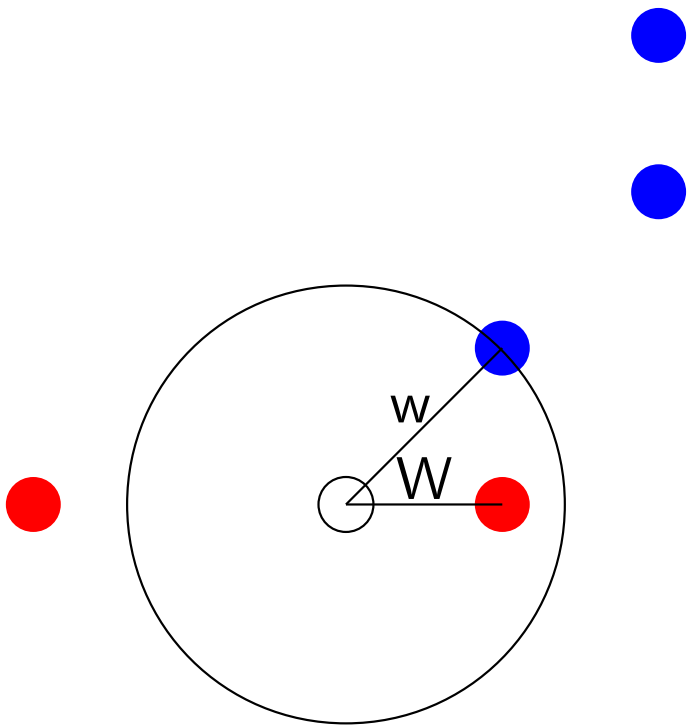
- Puts a higher weight on lesser distances



k-Nearest-Neighbors (kNN)

Solution: Weighted kNN

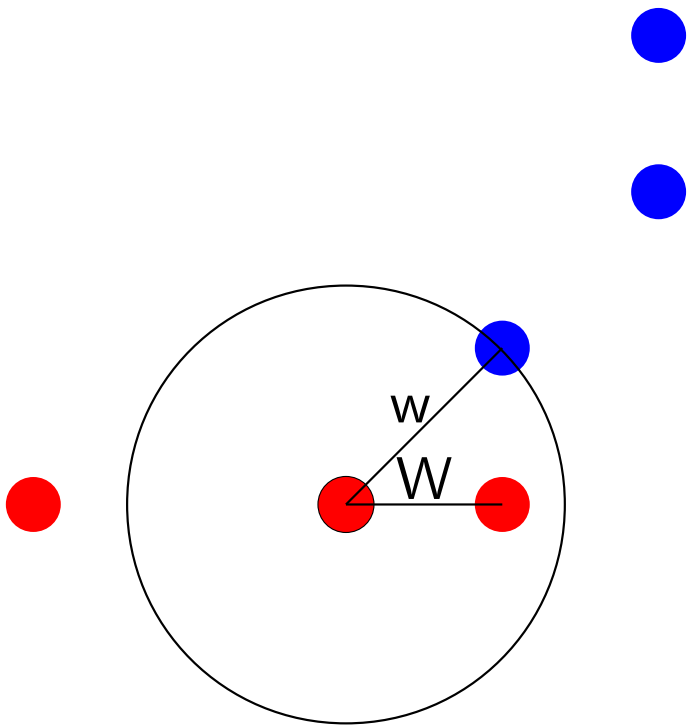
- Puts a higher weight on lesser distances



k-Nearest-Neighbors (kNN)

Solution: Weighted kNN

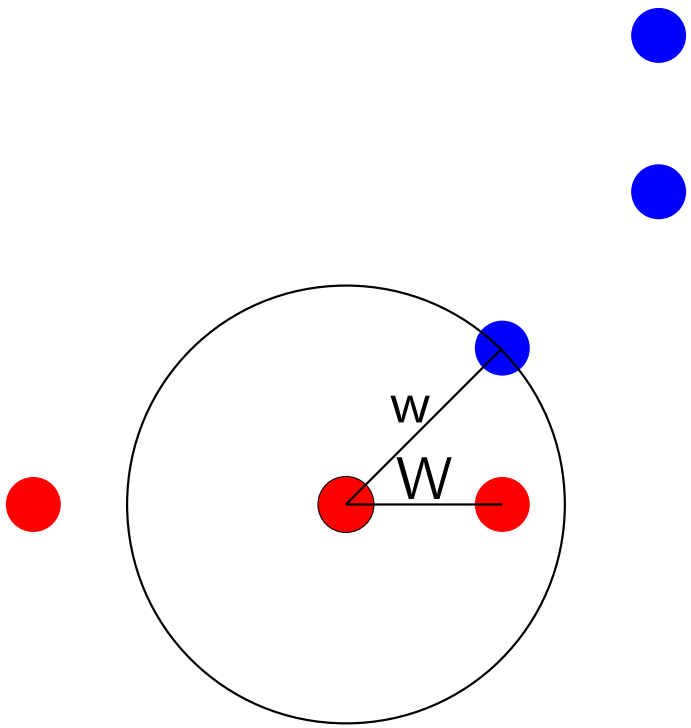
- Puts a higher weight on lesser distances



k-Nearest-Neighbors (kNN)

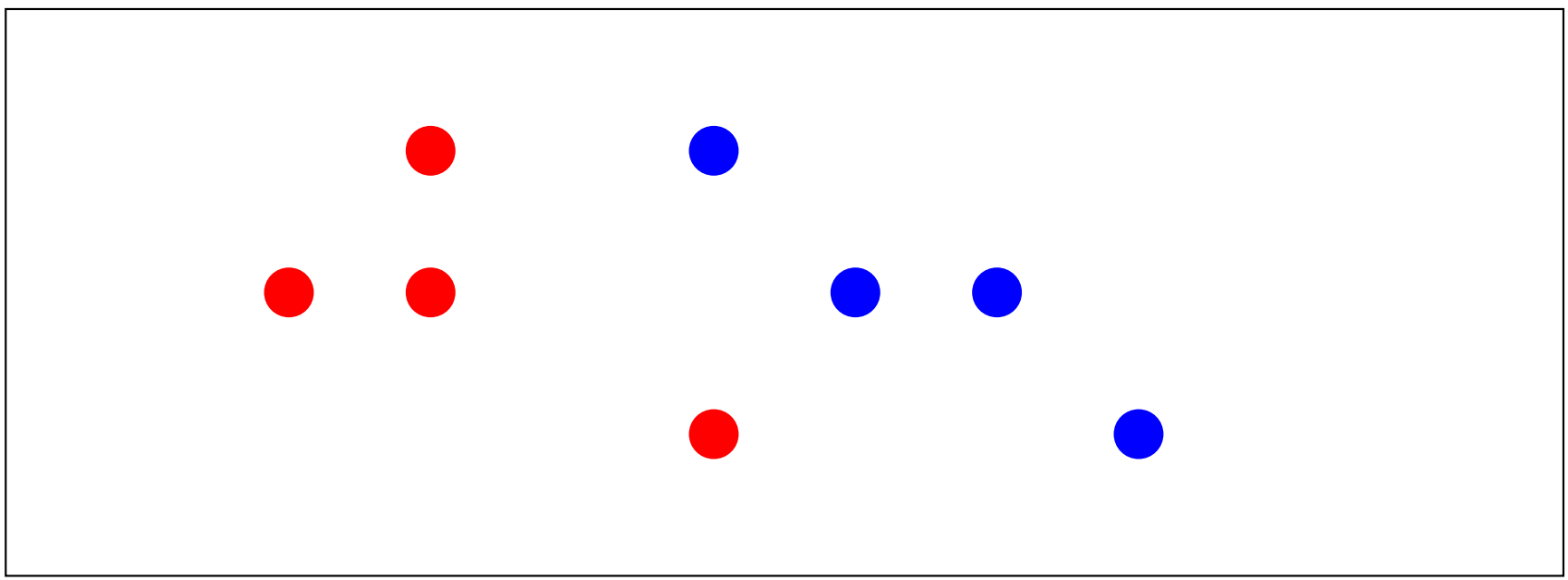
Solution: Weighted kNN

- Puts a higher weight on lesser distances
- Can use any k



Support Vector Machines (SVM)

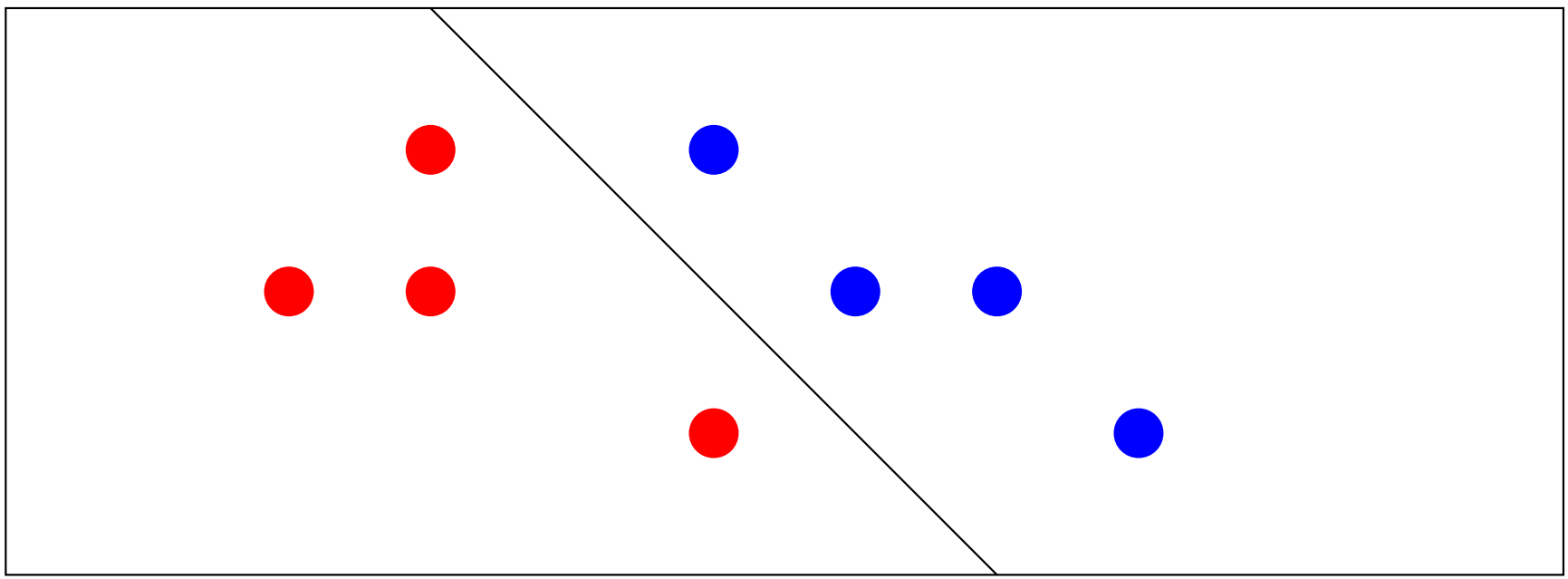
SVM technique



Support Vector Machines (SVM)

SVM technique

- Construct the optimal separating hyperplane.



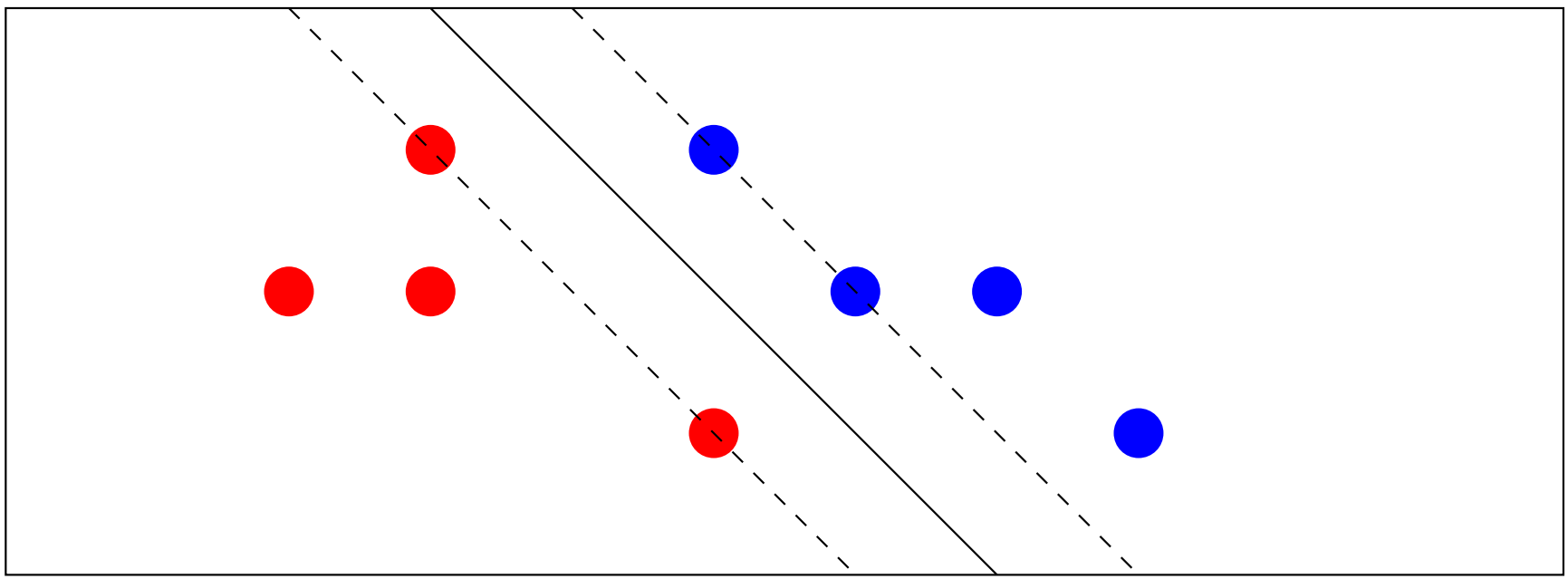
Support Vector Machines (SVM)

SVM technique

- Construct the optimal separating hyperplane.
- Maximize the margins.

Definition

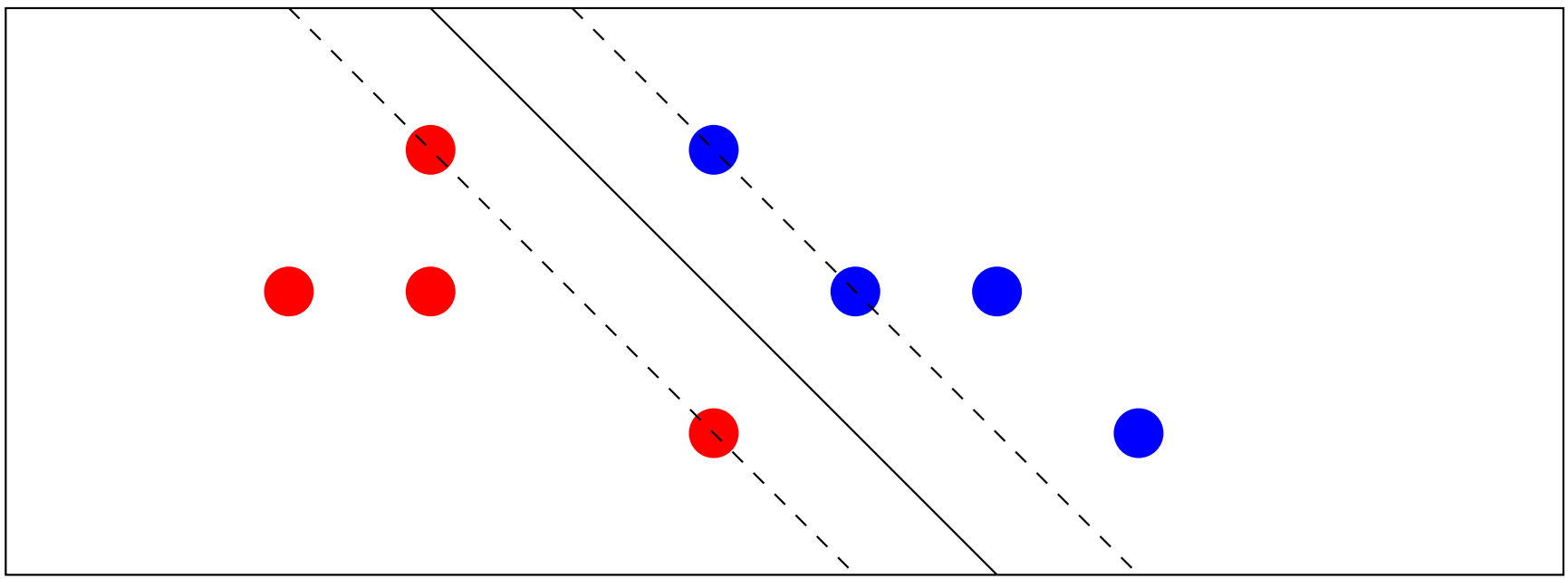
Margin: Minimal distance away from the hyperplane.



Support Vector Machines (SVM)

SVM technique

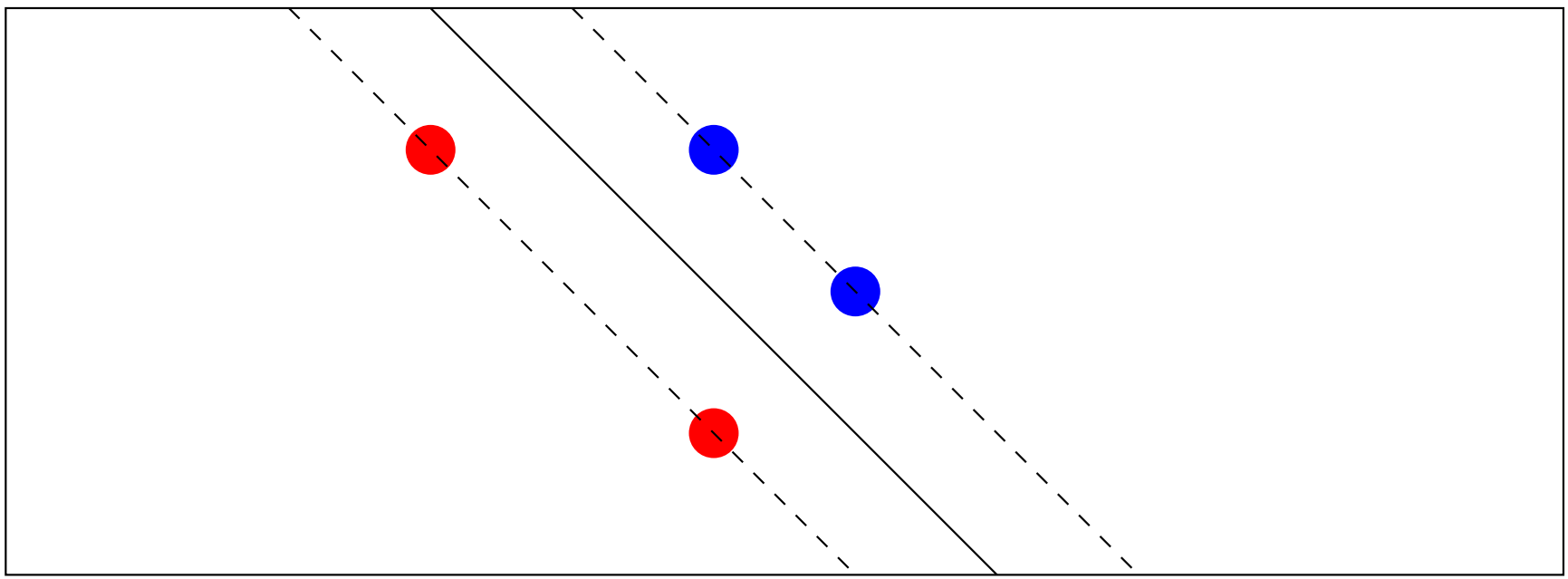
- Construct the optimal separating hyperplane.
- Maximize the margins.
- A small set of samples specifies the hyperplane.



Support Vector Machines (SVM)

SVM technique

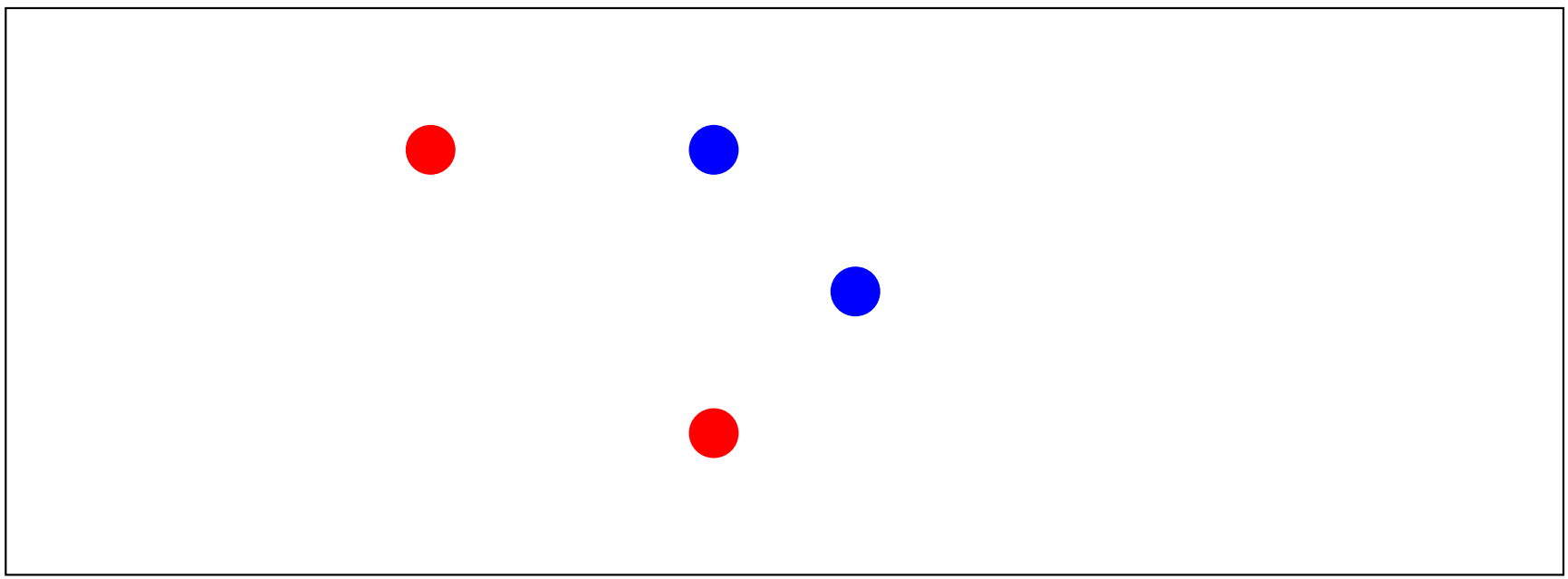
- Construct the optimal separating hyperplane.
- Maximize the margins.
- A small set of samples specifies the hyperplane.



Support Vector Machines (SVM)

SVM technique

- Construct the optimal separating hyperplane.
- Maximize the margins.
- A small set of samples specifies the hyperplane.
- Called Support Vectors (SV).



SVM Characteristics

- Based on a kernel function.

SVM Characteristics

- Based on a kernel function.

Definition

Kernel function: A function that maps input data to a higher dimensional feature space and computes their scalar product.

$$K(\overrightarrow{a}, \overrightarrow{b}) = \phi(\overrightarrow{a})^t \cdot \phi(\overrightarrow{b}) \qquad \text{for} \quad \overrightarrow{a}, \overrightarrow{b} \in X$$

SVM Characteristics

- Based on a kernel function.
- Performs scalar computation in input space $\Rightarrow$ efficient.

Definition

Kernel function: A function that maps input data to a higher dimensional feature space and computes their scalar product.

$$K(\overrightarrow{a}, \overrightarrow{b}) = \phi(\overrightarrow{a})^t \cdot \phi(\overrightarrow{b}) \qquad \text{for} \quad \overrightarrow{a}, \overrightarrow{b} \in X$$

SVM Characteristics

- Based on a kernel function.
- Performs scalar computation in input space $\Rightarrow$ efficient.
- Maps input to higher feature space $\Rightarrow$ generalization problem.

Definition

Kernel function: A function that maps input data to a higher dimensional feature space and computes their scalar product.

$$K(\vec{a}, \vec{b}) = \phi(\vec{a})^t \cdot \phi(\vec{b}) \qquad \text{for} \quad \vec{a}, \vec{b} \in X$$

SVM Characteristics

- Based on a kernel function.
- Performs scalar computation in input space $\Rightarrow$ efficient.
- Maps input to higher feature space $\Rightarrow$ generalization problem.

Definition

Kernel function: A function that maps input data to a higher dimensional feature space and computes their scalar product.

$$K(\vec{a}, \vec{b}) = \phi(\vec{a})^t \cdot \phi(\vec{b}) \qquad \text{for} \quad \vec{a}, \vec{b} \in X$$

Definition

Generalization: Ability to correctly classify new data.

SVM Characteristics

- Based on a kernel function.
- Performs scalar computation in input space $\Rightarrow$ efficient.
- Maps input to higher feature space $\Rightarrow$ generalization problem.
- Small number of support vectors $\Leftrightarrow$ good generalization.

Definition

Kernel function: A function that maps input data to a higher dimensional feature space and computes their scalar product.

$$K(\vec{a}, \vec{b}) = \phi(\vec{a})^t \cdot \phi(\vec{b}) \qquad \text{for} \quad \vec{a}, \vec{b} \in X$$

Definition

Generalization: Ability to correctly classify new data.

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

- Based on multiple classification trees

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

- Based on multiple classification trees
- Provides direct feature importance measure

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise

Random Forest (RF)

- High number of features $\Rightarrow$ *curse of dimensionality*.
- *Feature selection*: select the best subset of the input features.

Random Forest feature selection tool:

- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise
- Low bias and low variance

Problem Representation

A good classifier must be trained on:

1. Representative data

Problem Representation

A good classifier must be trained on:

1. Representative data
2. The right features

Problem Representation

A good classifier must be trained on:

1. Representative data
2. The right features

This sections explains:

1. How the data is obtained

Problem Representation

A good classifier must be trained on:

1. Representative data
2. The right features

This sections explains:

1. How the data is obtained
2. How the features are represented

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures
- $36,837$ entries on May 30, 2006.

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures
- 36,837 entries on May 30, 2006.

Perl:

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures
- $36,837$ entries on May 30, 2006.

Perl:

- A leading language for processing structured flat files

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures
- 36,837 entries on May 30, 2006.

Perl:

- A leading language for processing structured flat files
- *BioPerl*: bioinformatics data processing and analysis

Database Availability

The Protein Data Bank (PDB):

- Repository of 3-D structures of biological molecules
- Follows a strict format
- Stored as a sequence of records in a flat file
- Contains many redundant, obsolete, caveat or hypothetical structures
- $36,837$ entries on May 30, 2006.

Perl:

- A leading language for processing structured flat files
- *BioPerl*: bioinformatics data processing and analysis
- Parse a PDB file, represent it as an object and query its fields.

Data Mining

In a PDB file, glucose can be:

- Covalently bonded to the protein (glycoprotein)
- Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- Covalently bonded to the protein (glycoprotein)
- Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- √ Docking to a binding-site

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- √ Docking to a binding-site

PDB contains:

- 28,353 entries in October 2004

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- √ Docking to a binding-site

PDB contains:

- 28,353 entries in October 2004
- 331 entries containing glucose

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- √ Docking to a binding-site

PDB contains:

- 28,353 entries in October 2004
- 331 entries containing glucose
- 59 resolved protein-glucose binding sites

Data Mining

In a PDB file, glucose can be:

- X Covalently bonded to the protein (glycoprotein)
- X Floating around in the medium
- √ Docking to a binding-site

PDB contains:

- 28,353 entries in October 2004
- 331 entries containing glucose
- 59 resolved protein-glucose binding sites
- 29 distinct protein-glucose binding sites

Solution Approach

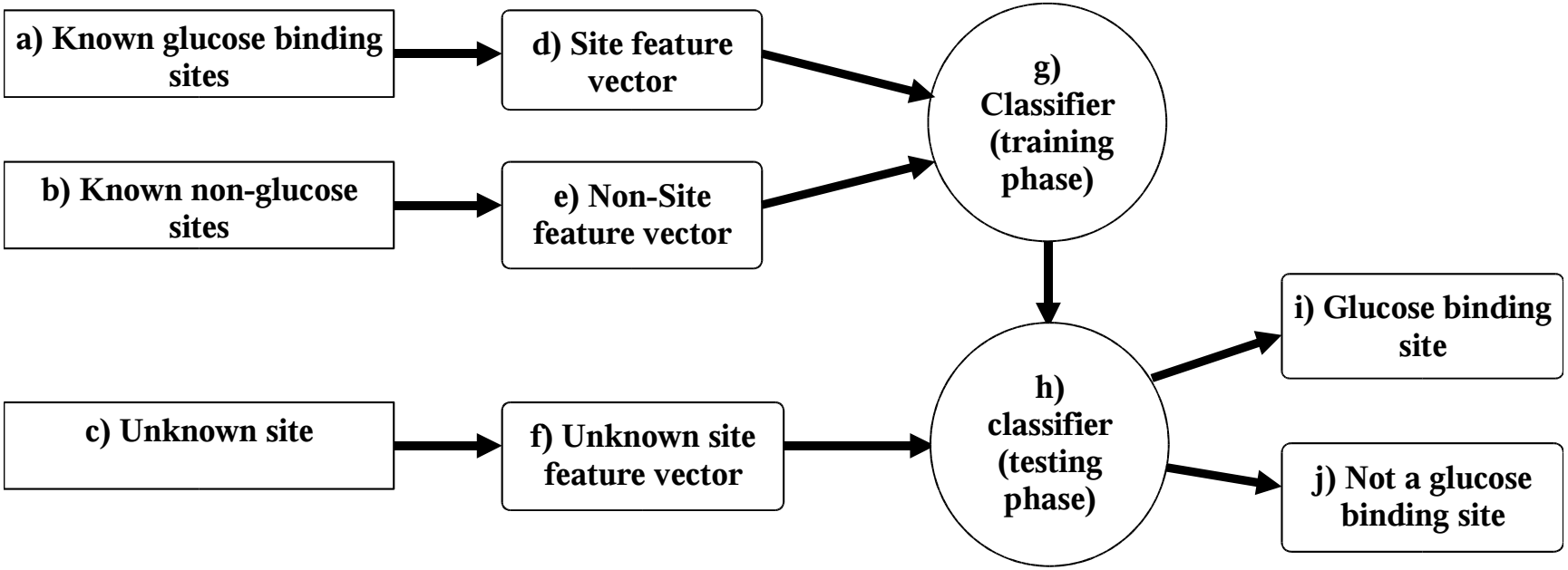


Figure: The classifier algorithm outline.

Solution Approach

Non-glucose sites

1. Binding sites that do not bind glucose

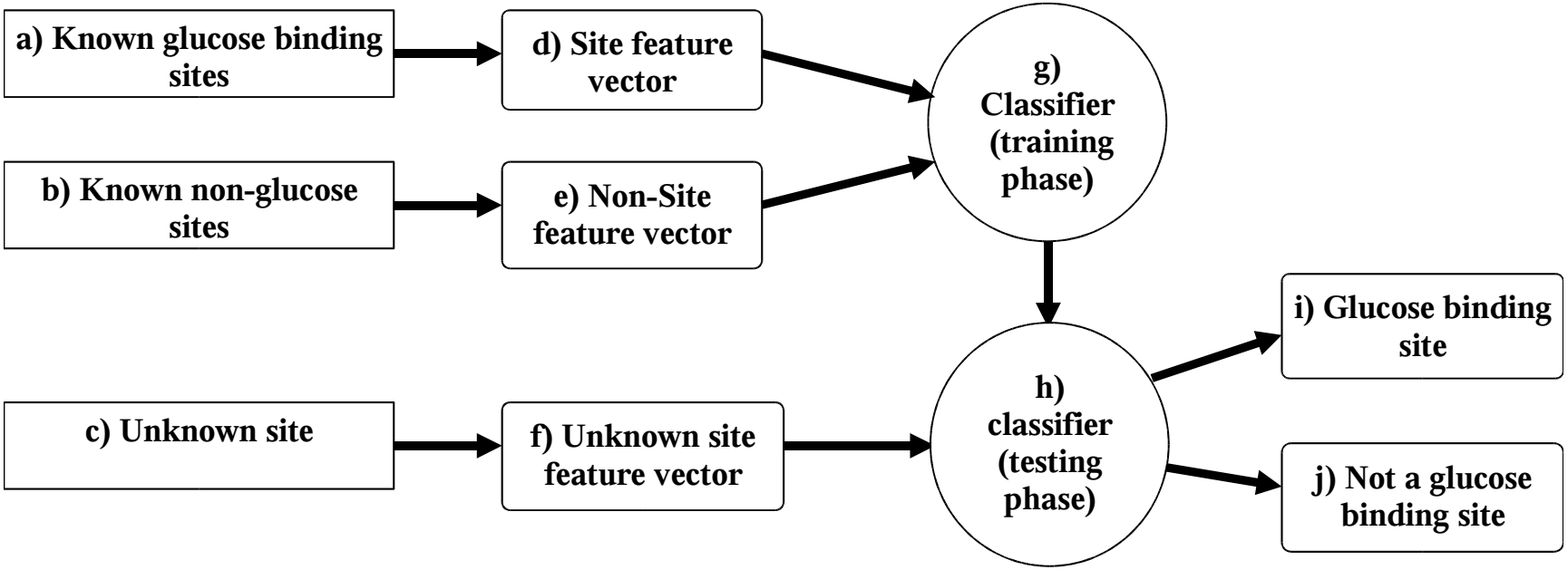


Figure: The classifier algorithm outline.

Solution Approach

Non-glucose sites

1. Binding sites that do not bind glucose
2. Random protein surface regions

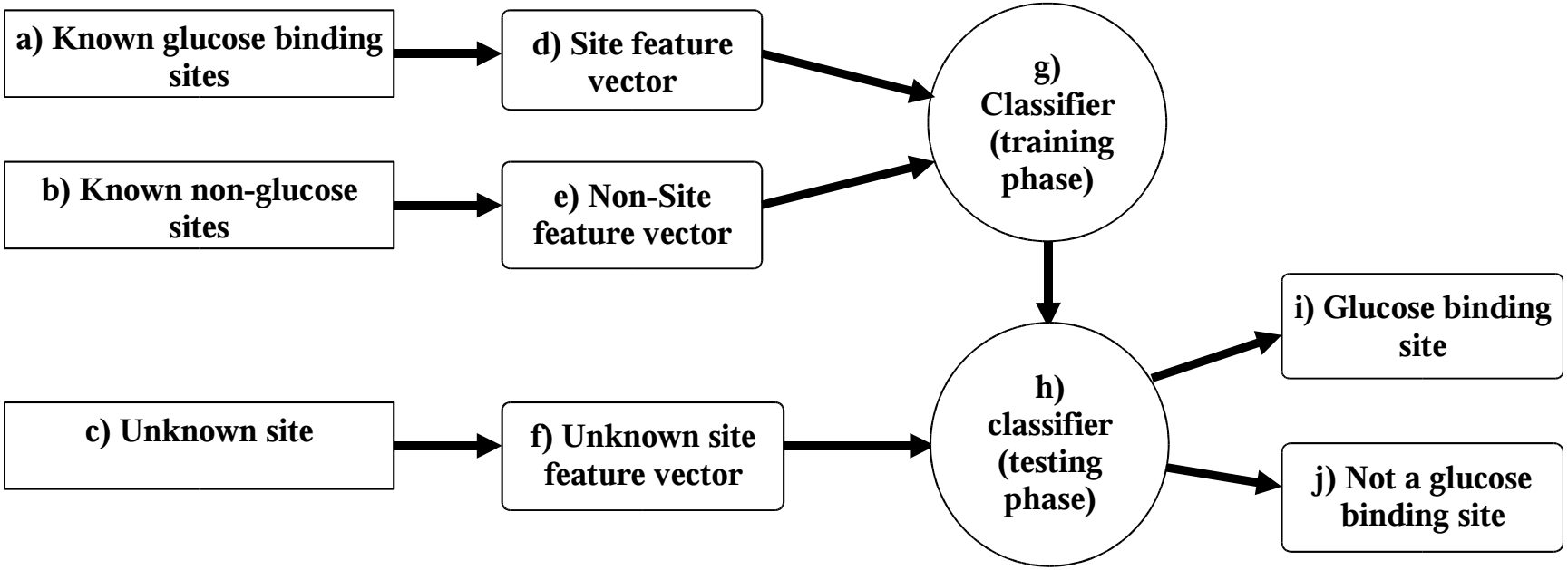


Figure: The classifier algorithm outline.

Binding Site Representation

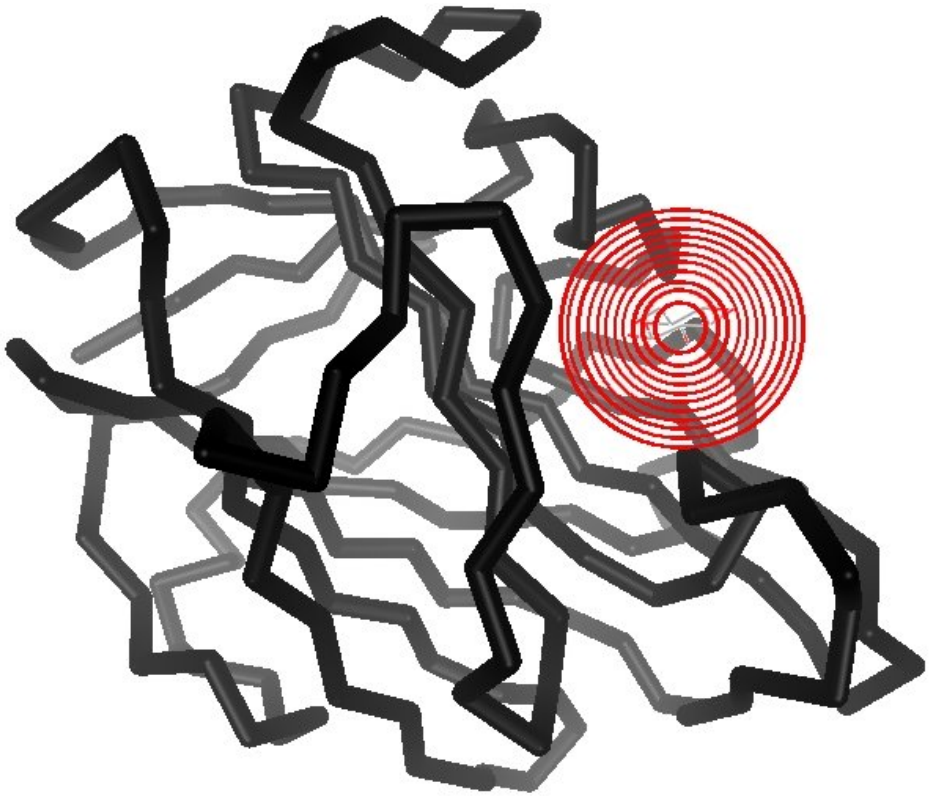


Figure: Glucose bound to a hydrolase, PDB entry 1I8A.

Binding Site Representation

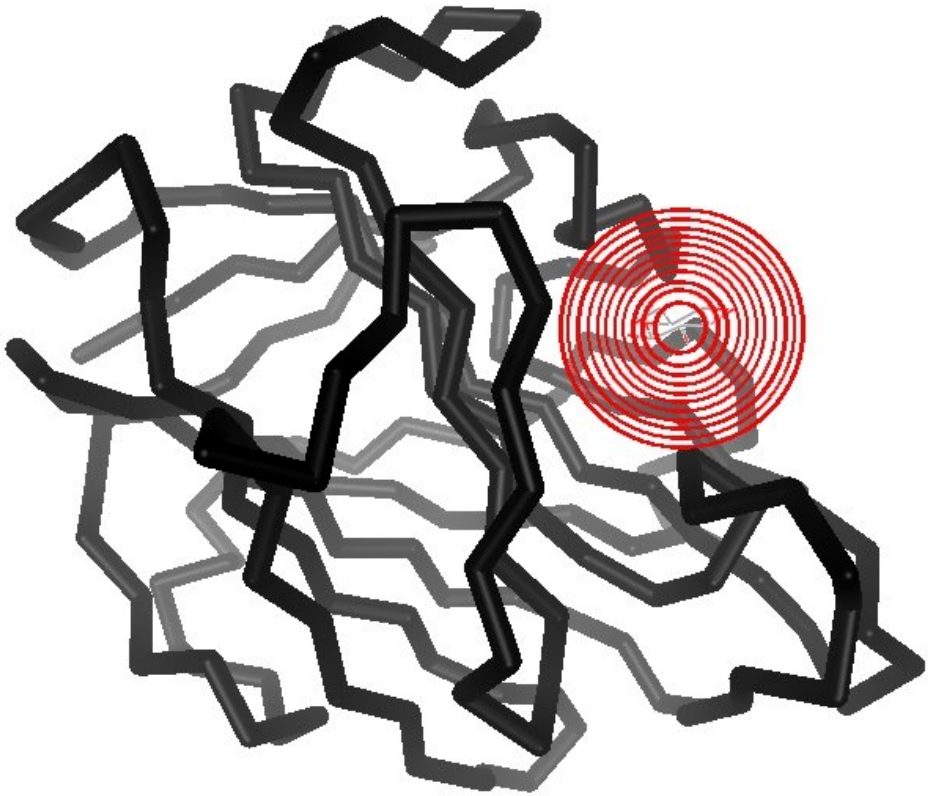


Figure: Glucose bound to a hydrolase, PDB entry 1I8A.

- Sphere of 10 Å centered at the ligand

Binding Site Representation

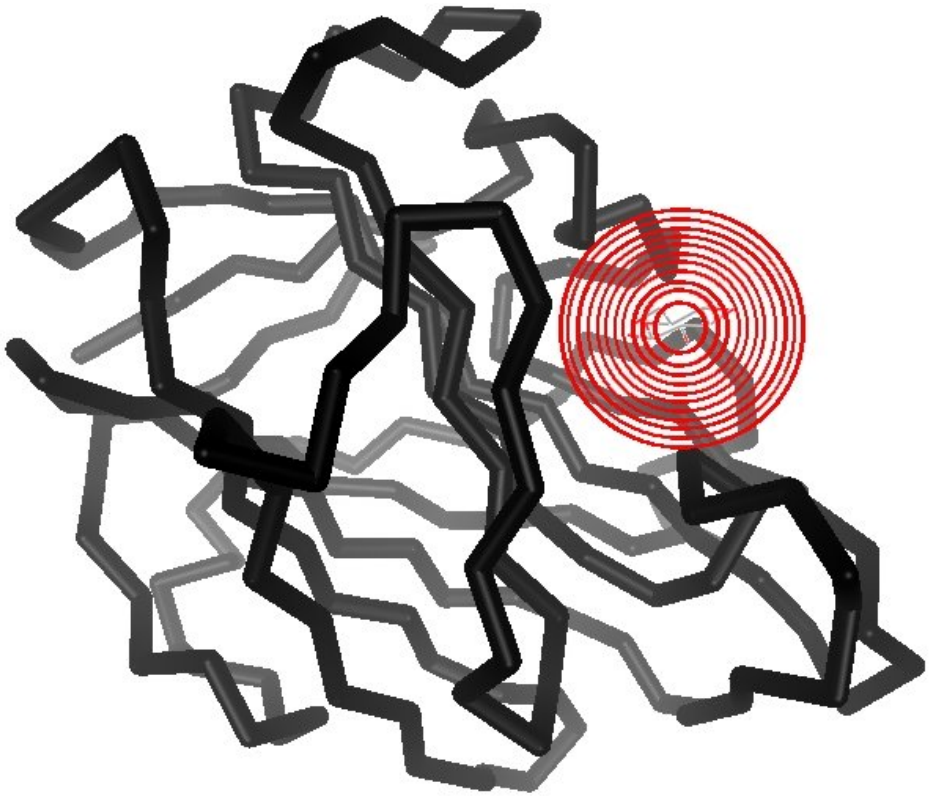


Figure: Glucose bound to a hydrolase, PDB entry 1I8A.

- Sphere of 10 Å centered at the ligand
- Pyranose ring centroid

Binding Site Representation

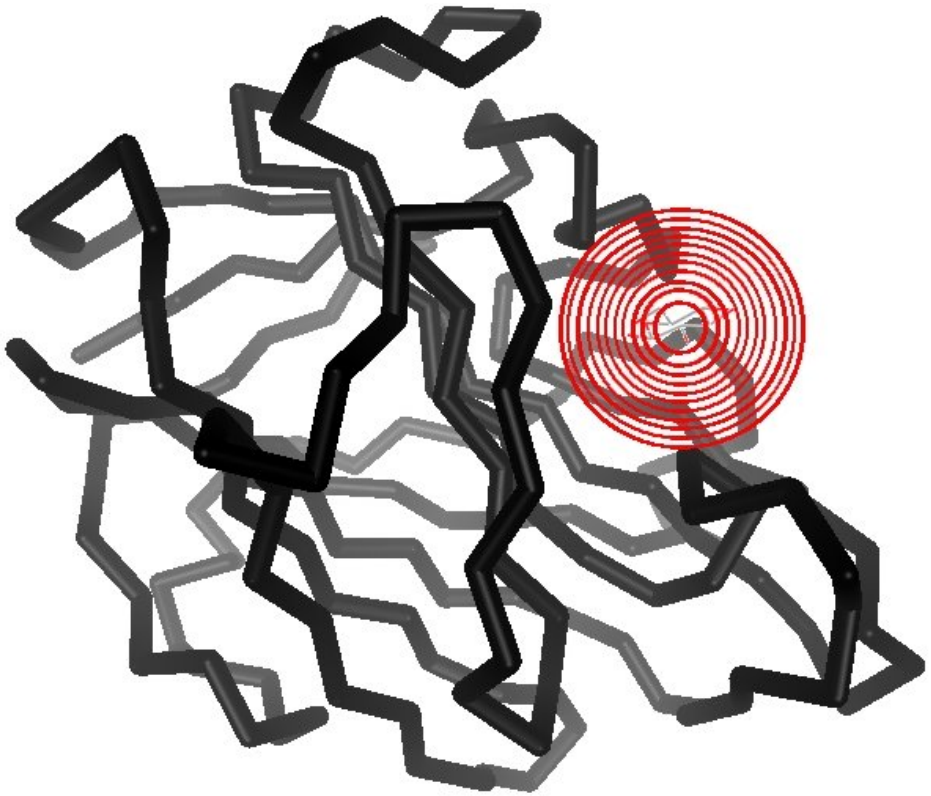


Figure: Glucose bound to a hydrolase, PDB entry 1I8A.

- Sphere of 10 Å centered at the ligand
- Pyranose ring centroid
- Divided into 8 concentric layers

Binding Site Representation

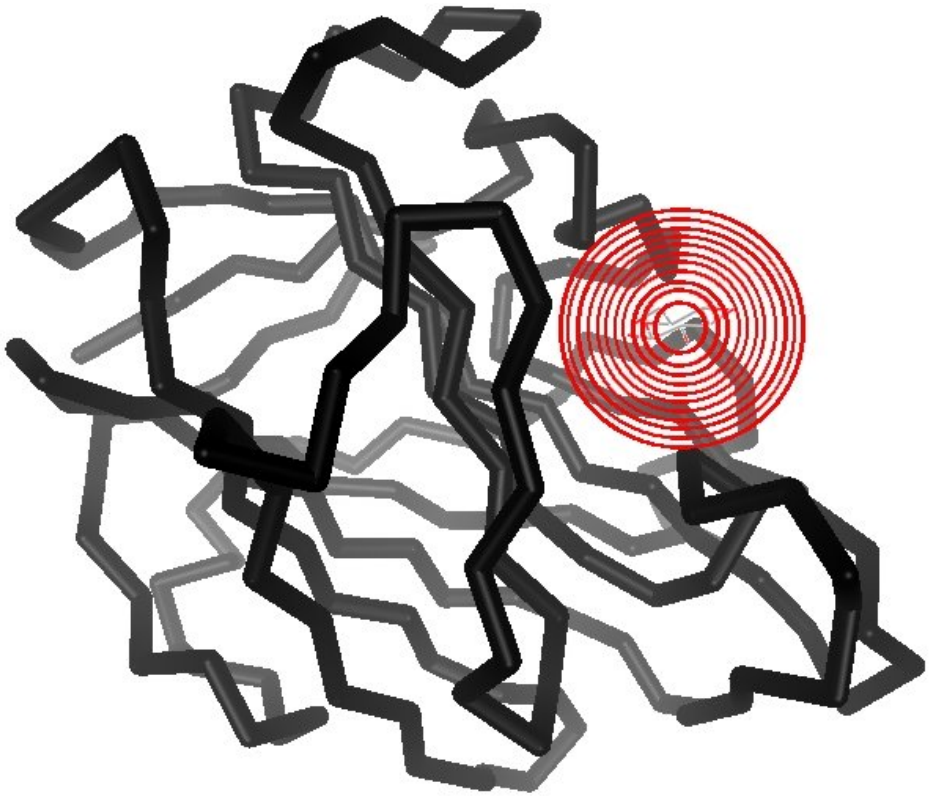


Figure: Glucose bound to a hydrolase, PDB entry 1I8A.

- Sphere of 10 Å centered at the ligand
- Pyranose ring centroid
- Divided into 8 concentric layers
- Gather features for each layer

Atomic Features

1. Charge

Atomic Features

1. Charge
2. Hydrogen Bond

Atomic Features

1. Charge
2. Hydrogen Bond
3. Hydrophobicity

Atomic Features

1. Charge
    1. Positive
2. Hydrogen Bond
3. Hydrophobicity

Atomic Features

1. Charge
   1. Positive
   2. Neutral
2. Hydrogen Bond
3. Hydrophobicity

Atomic Features

1. Charge
    1. Positive
    2. Neutral
    3. Negative
2. Hydrogen Bond
3. Hydrophobicity

Atomic Features

1. Charge
    1. Positive
    2. Neutral
    3. Negative
2. Hydrogen Bond
    1. Hydrogen-bonding
3. Hydrophobicity

Atomic Features

1. Charge
   1. Positive
   2. Neutral
   3. Negative
2. Hydrogen Bond
   1. Hydrogen-bonding
   2. Non hydrogen-bonding
3. Hydrophobicity

Atomic Features

1. Charge
    1. Positive
    2. Neutral
    3. Negative
2. Hydrogen Bond
    1. Hydrogen-bonding
    2. Non hydrogen-bonding
3. Hydrophobicity
    1. Hydrophobic

Atomic Features

1. Charge
   1. Positive
   2. Neutral
   3. Negative
2. Hydrogen Bond
   1. Hydrogen-bonding
   2. Non hydrogen-bonding
3. Hydrophobicity
   1. Hydrophobic
   2. Neutral

Atomic Features

1. Charge
   1. Positive
   2. Neutral
   3. Negative
2. Hydrogen Bond
   1. Hydrogen-bonding
   2. Non hydrogen-bonding
3. Hydrophobicity
   1. Hydrophobic
   2. Neutral
   3. Hydrophilic

Residue Schemes

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified1

- Planar polar residues
- Form an extensive hydrogen-bond network with glucose

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified1

- Planar polar residues
- Form an extensive hydrogen-bond network with glucose
- Remaining residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified2

- Planar polar residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified2

- Planar polar residues
- Aromatic residues
- Ring containing residues
- Stack against pyranose ring thru hydrophobic and van der Waals forces.

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified2

- Planar polar residues
- Aromatic residues
- Ring containing residues
- Stack against pyranose ring thru hydrophobic and van der Waals forces.
- Remaining residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified3

- Planar polar residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified3

- Planar polar residues
- Aromatic residues
- Histidine is reported to interact with glucose in a similar way to aromatic residues.

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Simplified3

- Planar polar residues
- Aromatic residues
- Histidine is reported to interact with glucose in a similar way to aromatic residues.
- Remaining residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues
- Neutral residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues
- Neutral residues
- Carboxylate residues (negatively charged)

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues
- Neutral residues
- Carboxylate residues (negatively charged)
- Positively charged residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues
- Neutral residues
- Carboxylate residues (negatively charged)
- Positively charged residues
- Aliphatic residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed1

- Aromatic residues
- Neutral residues
- Carboxylate residues (negatively charged)
- Positively charged residues
- Aliphatic residues
- Histidine

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Residue Schemes

Detailed2

- Aromatic residues
- Neutral residues
- Carboxylate residues (negatively charged)
- Positively charged residues
- Aliphatic residues

Asparagine	Glutamine	Arginine	Glutamate
Aspartate	Tyrosine	Tryptophan	Phenylalanine
Histidine	Serine	Threonine	Glycine
Proline	Cysteine	Lysine	Alanine
Valine	Leucine	Isoleucine	Methionine

Experimental Results

What determines a binding-site:

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties
2. Amino acids spatial arrangement

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties
2. Amino acids spatial arrangement

The approach consists of:

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties
2. Amino acids spatial arrangement

The approach consists of:

1. Learning phase

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties
2. Amino acids spatial arrangement

The approach consists of:

1. Learning phase
2. Feature selection

Experimental Results

What determines a binding-site:

1. Amino acids biochemical properties
2. Amino acids spatial arrangement

The approach consists of:

1. Learning phase
2. Feature selection
3. Testing phase

Atomic Combinations Results

Table: Atomic properties classification error rates.

Properties	kNN	SVM	SV
Charge	14.55%	14.55%	78.18%
H-Bond	21.82%	16.36%	92.73%
Hydrophobicity	21.82%	20.00%	92.73%
Charge + H-Bond	14.55%	14.55%	89.09%
Charge + Hydro	12.73%	14.55%	47.27%
H-Bond + Hydro	21.82%	18.18%	100%
Charge + H-Bond + Hydro	16.36%	16.36%	60.00%

Atomic Combinations Results

- Charge outperforms hydrophobicity and hydrogen bond

Table: Atomic properties classification error rates.

Properties	kNN	SVM	SV
Charge	14.55%	14.55%	78.18%
H-Bond	21.82%	16.36%	92.73%
Hydrophobicity	21.82%	20.00%	92.73%
Charge + H-Bond	14.55%	14.55%	89.09%
Charge + Hydro	12.73%	14.55%	47.27%
H-Bond + Hydro	21.82%	18.18%	100%
Charge + H-Bond + Hydro	16.36%	16.36%	60.00%

Atomic Combinations Results

- Charge outperforms hydrophobicity and hydrogen bond
- Charge, linked with another property, yields similar or slightly better results

Table: Atomic properties classification error rates.

Properties	kNN	SVM	SV
Charge	14.55%	14.55%	78.18%
H-Bond	21.82%	16.36%	92.73%
Hydrophobicity	21.82%	20.00%	92.73%
Charge + H-Bond	14.55%	14.55%	89.09%
Charge + Hydro	12.73%	14.55%	47.27%
H-Bond + Hydro	21.82%	18.18%	100%
Charge + H-Bond + Hydro	16.36%	16.36%	60.00%

Atomic Combinations Results

- "Charge + Hydro" is the best atomic combination

Table: Atomic properties classification error rates.

Properties	*k*NN	SVM	SV
Charge	14.55%	14.55%	78.18%
H-Bond	21.82%	16.36%	92.73%
Hydrophobicity	21.82%	20.00%	92.73%
Charge + H-Bond	14.55%	14.55%	89.09%
Charge + Hydro	12.73%	14.55%	47.27%
H-Bond + Hydro	21.82%	18.18%	100%
Charge + H-Bond + Hydro	16.36%	16.36%	60.00%

Atomic Combinations Results

- kNN and SVM give similar results

Table: Atomic properties classification error rates.

Properties	kNN	SVM	SV
Charge	14.55%	14.55%	78.18%
H-Bond	21.82%	16.36%	92.73%
Hydrophobicity	21.82%	20.00%	92.73%
Charge + H-Bond	14.55%	14.55%	89.09%
Charge + Hydro	12.73%	14.55%	47.27%
H-Bond + Hydro	21.82%	18.18%	100%
Charge + H-Bond + Hydro	16.36%	16.36%	60.00%

Hydrogen Bond Property

- Hydrogen bond is a primary factor in glucose binding

Hydrogen Bond Property

- Hydrogen bond is a primary factor in glucose binding
- Hydrogen bond preforms poorly

Hydrogen Bond Property

- Hydrogen bond is a primary factor in glucose binding
- Hydrogen bond preforms poorly

Hypothesis

1. Negative set composed mainly of non-glucose binding sites

Hydrogen Bond Property

- Hydrogen bond is a primary factor in glucose binding
- Hydrogen bond preforms poorly

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds

Hydrogen Bond Property

- Hydrogen bond is a primary factor in glucose binding
- Hydrogen bond preforms poorly

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds
3. Therefore hydrogen bond is a bad discriminating factor

Hydrogen Bond Property

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds
3. Therefore hydrogen bond is a bad discriminating factor

Hydrogen Bond Property

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds
3. Therefore hydrogen bond is a bad discriminating factor

To validate this hypothesis:

Hydrogen Bond Property

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds
3. Therefore hydrogen bond is a bad discriminating factor

To validate this hypothesis:

1. Negative set composed of random protein surface regions

Hydrogen Bond Property

Hypothesis

1. Negative set composed mainly of non-glucose binding sites
2. Most binding sites rely on hydrogen bonds
3. Therefore hydrogen bond is a bad discriminating factor

To validate this hypothesis:

1. Negative set composed of random protein surface regions
2. Hydrogen bond should outperform charge and hydrophobicity

Hydrogen Bond Property

To validate this hypothesis:

1. Negative set composed of random protein surface regions
2. Hydrogen bond should outperform charge and hydrophobicity

Hydrogen Bond Property

To validate this hypothesis:

1. Negative set composed of random protein surface regions
2. Hydrogen bond should outperform charge and hydrophobicity

The results confirm the hypothesis:

Table: Classifier training using an exclusively non-binding sites negative set.

Properties	kNN	SVM	SV
Charge	05.26%	05.26%	73.68%
H-Bond	03.51%	03.51%	61.40%
Hydrophobicity	05.26%	05.26%	68.42%

Residue Schemes Comparison

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- Detailed schemes outperform simplified schemes

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- Detailed schemes outperform simplified schemes
- Simplified schemes have better generalization

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- Detailed schemes yield similar results to the atomic properties

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- Detailed schemes yield similar results to the atomic properties
- Incorporate in themselves some of the atomic knowledge?

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- "Simplified2" scheme gives similar results to "simplified1" while adding the aromatic subgroup

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- "Simplified2" scheme gives similar results to "simplified1" while adding the aromatic subgroup
- Aromatic residues play an important role in glucose binding

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- "Simplified2" scheme gives similar results to "simplified1" while adding the aromatic subgroup
- Aromatic residues play an important role in glucose binding
- Aromatic residues do not seem to contribute to the classification accuracy

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- "Simplified3" improves the "simplified2" classification rate by one hit

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Residue Schemes Comparison

- "Simplified3" improves the "simplified2" classification rate by one hit
- Adding histidine to the aromatic subgroup gives slightly better results

Table: Comparison of the different residue schemes.

Residue scheme	kNN	SVM	SV
Simplified1	16.36%	18.18%	76.36%
Simplified2	18.18%	18.18%	70.91%
Simplified3	16.36%	16.36%	70.91%
Detailed1	16.36%	14.55%	81.82%
Detailed2	21.82%	14.55%	81.82%

Water and Ion inclusion

- Ordered water molecules and ions play a role in glucose binding

Table: Testing the importance of water and ions to glucose specificity

Mode	Properties	kNN	SVM	SV
Water + ions	Charge + Hydro	12.73%	14.55%	47.27%
	Residue + Charge + Hydro	20.00%	09.09%	100%
Ions	Charge + Hydro	14.55%	16.36%	83.64%
	Residue + Charge + Hydro	20.00%	10.91%	100%
Water	Charge + Hydro	16.36%	18.18%	83.64%
	Residue + Charge + Hydro	20.00%	10.91%	100%
	Charge + Hydro	14.55%	16.36%	52.73%
	Residue + Charge + Hydro	20.00%	12.73%	100%

Water and Ion inclusion

- Ordered water molecules and ions play a role in glucose binding
- We test and validate this hypothesis

Table: Testing the importance of water and ions to glucose specificity

Mode	Properties	*k*NN	SVM	SV
Water + ions	Charge + Hydro	12.73%	14.55%	47.27%
	Residue + Charge + Hydro	20.00%	09.09%	100%
Ions	Charge + Hydro	14.55%	16.36%	83.64%
	Residue + Charge + Hydro	20.00%	10.91%	100%
Water	Charge + Hydro	16.36%	18.18%	83.64%
	Residue + Charge + Hydro	20.00%	10.91%	100%
	Charge + Hydro	14.55%	16.36%	52.73%
	Residue + Charge + Hydro	20.00%	12.73%	100%

Atomic Properties Feature Selection

Table: Classifiers performance on atomic data using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensitivity	Specificity	SV
Charge	false	24	14.55%	14.55%	96.55%	73.08%	78.18%
	true	6	09.09%	05.45%	93.10%	96.15%	41.82%
H-Bond	false	16	21.82%	16.36%	86.21%	80.77%	92.73%
	true	5	09.09%	07.27%	96.55%	88.46%	16.36%
Hydro	false	24	21.82%	20.00%	79.31%	80.77%	92.73%
	true	5	09.09%	10.91%	96.55%	84.62%	34.55%

Atomic Properties Feature Selection

Definition

Sensitivity: Ability to detect true positives (TP/P)
Specificity: Ability to reject true negatives (TN/N)

Table: Classifiers performance on atomic data using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensitivity	Specificity	SV
Charge	false	24	14.55%	14.55%	96.55%	73.08%	78.18%
	true	6	09.09%	05.45%	93.10%	96.15%	41.82%
H-Bond	false	16	21.82%	16.36%	86.21%	80.77%	92.73%
	true	5	09.09%	07.27%	96.55%	88.46%	16.36%
Hydro	false	24	21.82%	20.00%	79.31%	80.77%	92.73%
	true	5	09.09%	10.91%	96.55%	84.62%	34.55%

Atomic Properties Feature Selection

- Classification error decreases

Table: Classifiers performance on atomic data using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensitivity	Specificity	SV
Charge	false	24	14.55%	14.55%	96.55%	73.08%	78.18%
	true	6	09.09%	05.45%	93.10%	96.15%	41.82%
H-Bond	false	16	21.82%	16.36%	86.21%	80.77%	92.73%
	true	5	09.09%	07.27%	96.55%	88.46%	16.36%
Hydro	false	24	21.82%	20.00%	79.31%	80.77%	92.73%
	true	5	09.09%	10.91%	96.55%	84.62%	34.55%

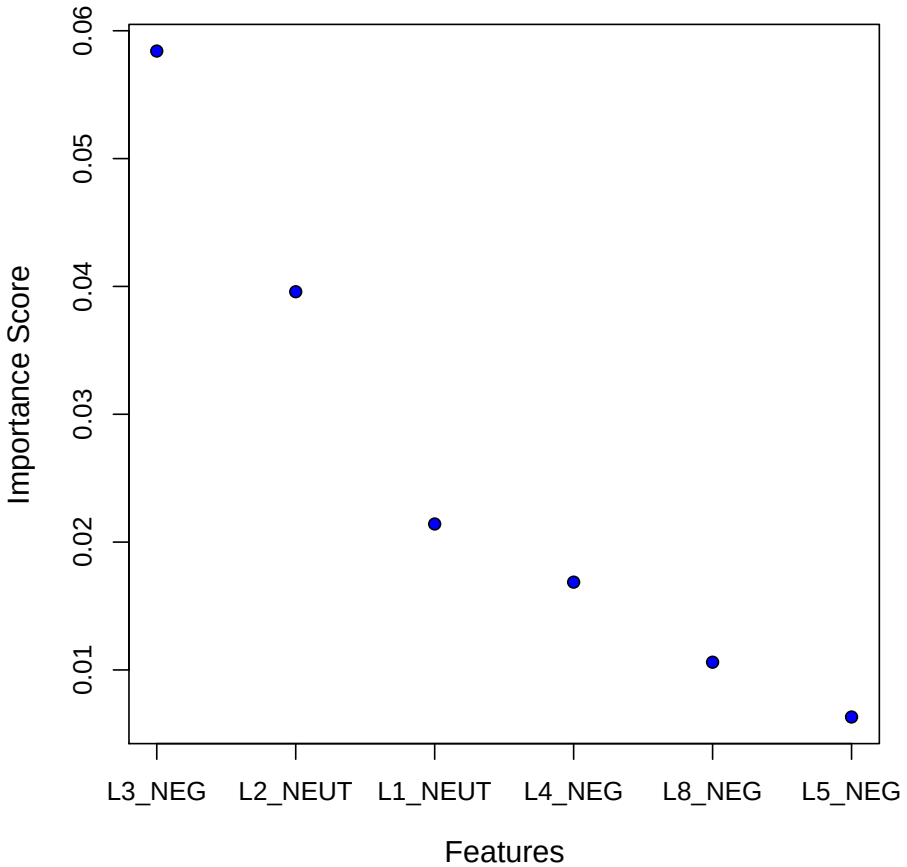
Atomic Properties Feature Selection

- Classification error decreases
- SV percentage decreases (generalization capacity increases)

Table: Classifiers performance on atomic data using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensitivity	Specificity	SV
Charge	false	24	14.55%	14.55%	96.55%	73.08%	78.18%
	true	6	09.09%	05.45%	93.10%	96.15%	41.82%
H-Bond	false	16	21.82%	16.36%	86.21%	80.77%	92.73%
	true	5	09.09%	07.27%	96.55%	88.46%	16.36%
Hydro	false	24	21.82%	20.00%	79.31%	80.77%	92.73%
	true	5	09.09%	10.91%	96.55%	84.62%	34.55%

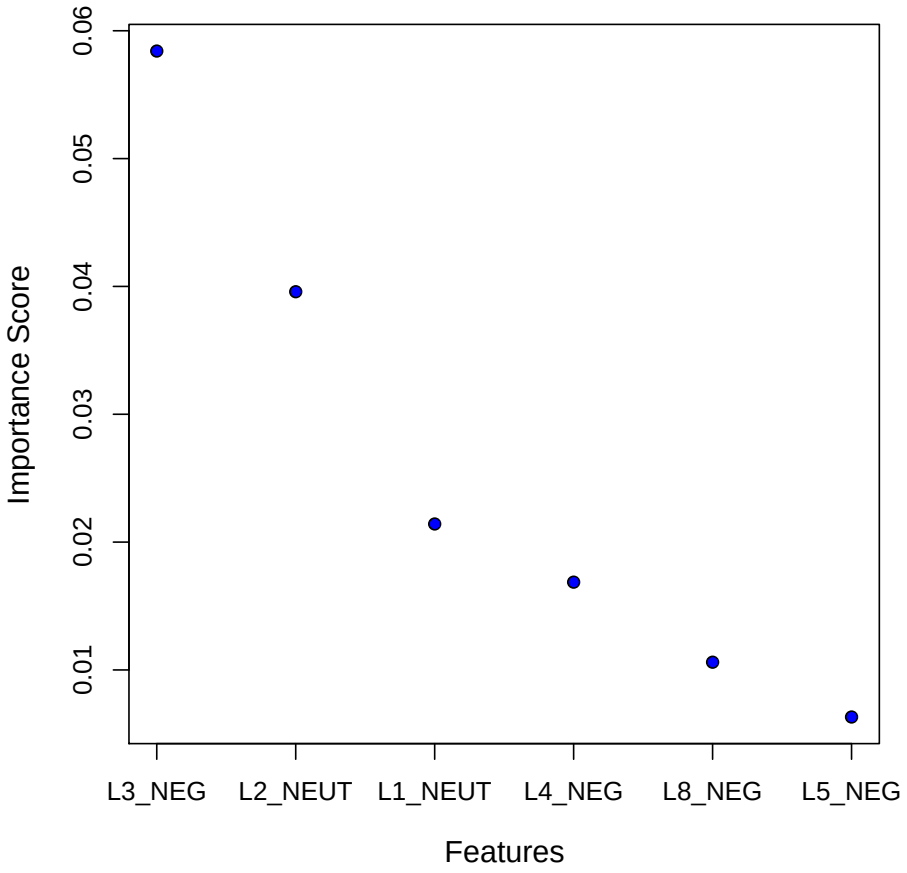
Charge Feature Selection



- Remains the best discriminating factor

Figure: Importance of charge features

Charge Feature Selection



- Remains the best discriminating factor
- Negatively charged glucose binding site

Figure: Importance of charge features

Charge Feature Selection

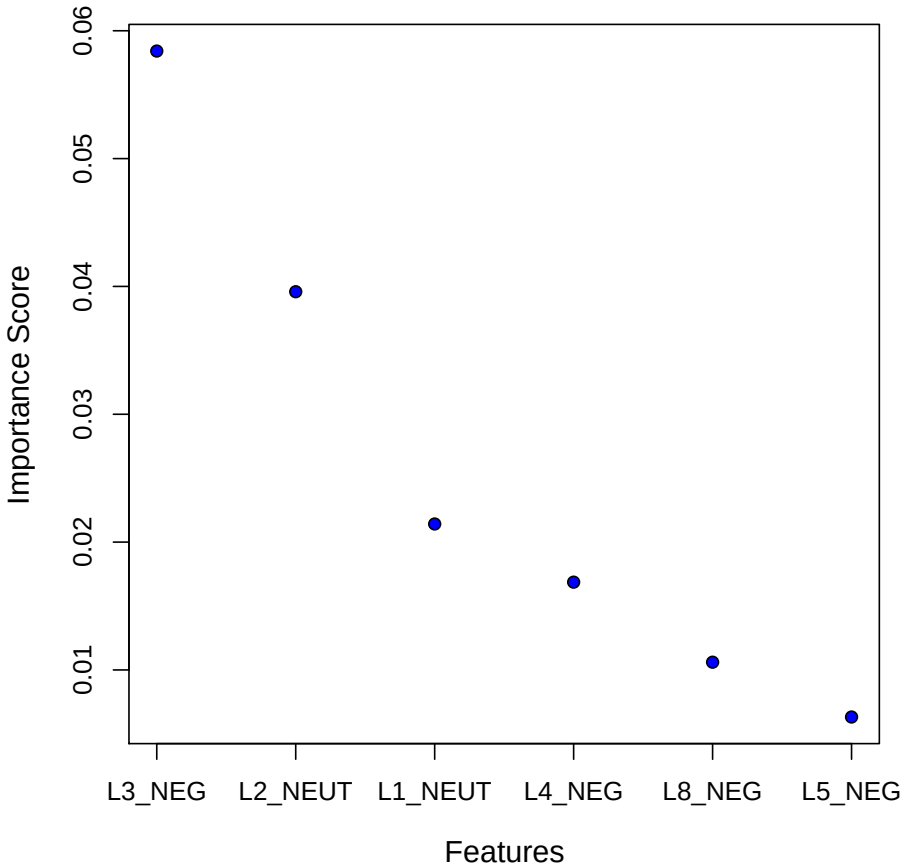
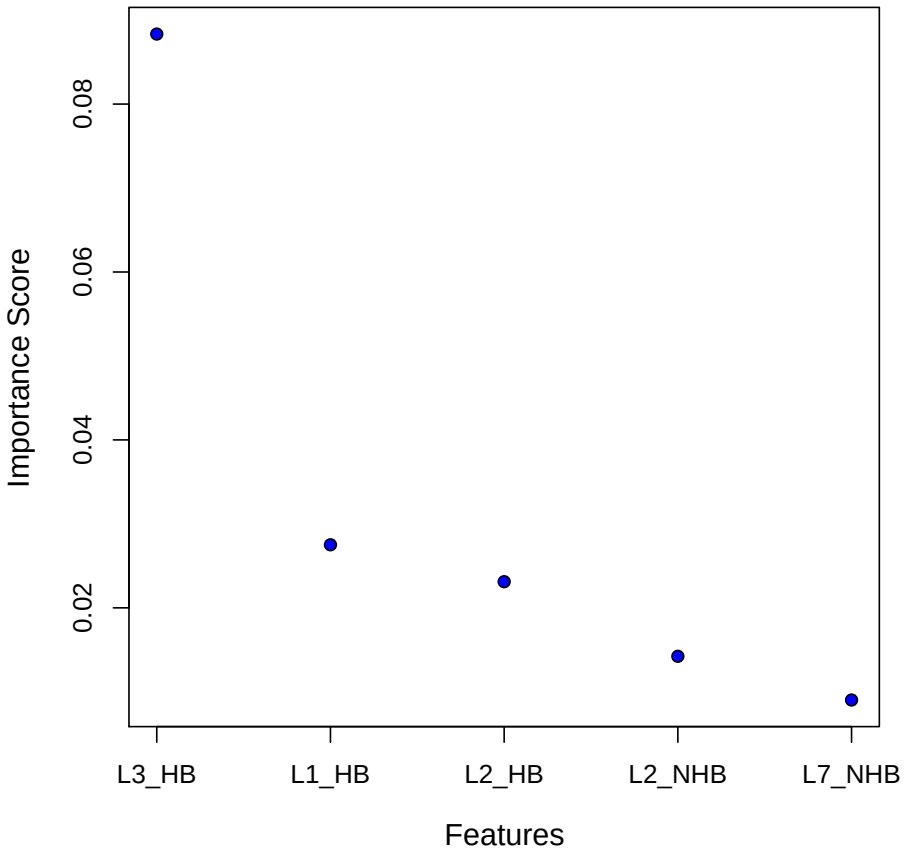


Figure: Importance of charge features

- Remains the best discriminating factor
- Negatively charged glucose binding site
- Layers 1 and 2: overlap with glucose

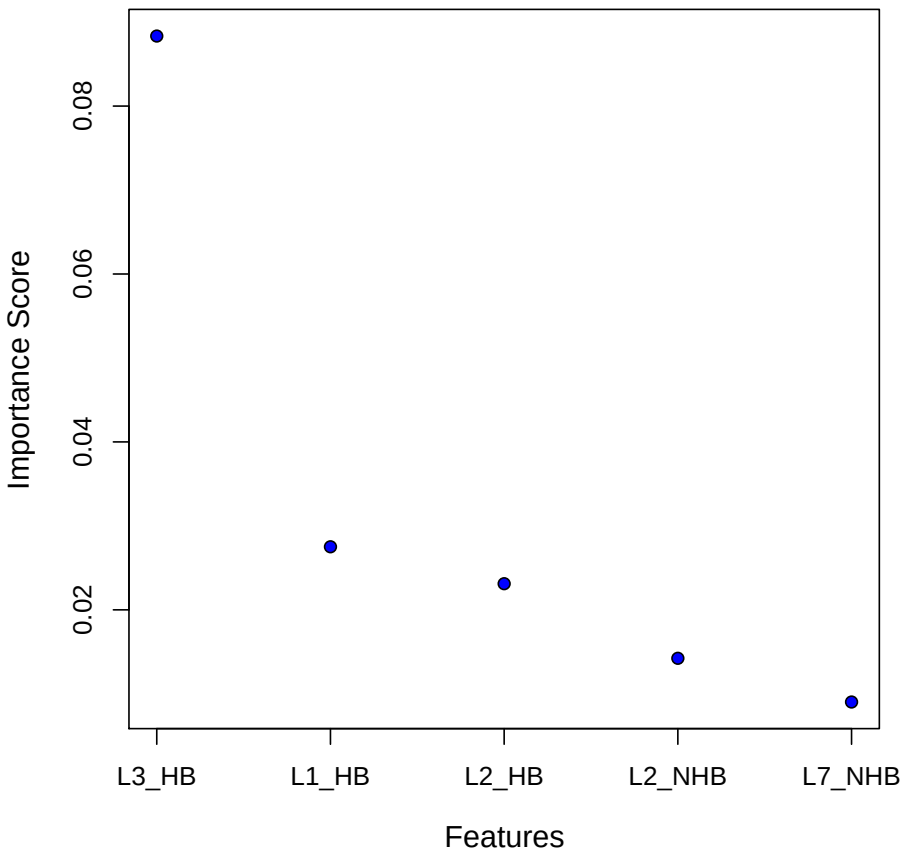
Hydrogen Bond Feature Selection



- A record low SV percentage

Figure: Importance of hbond features

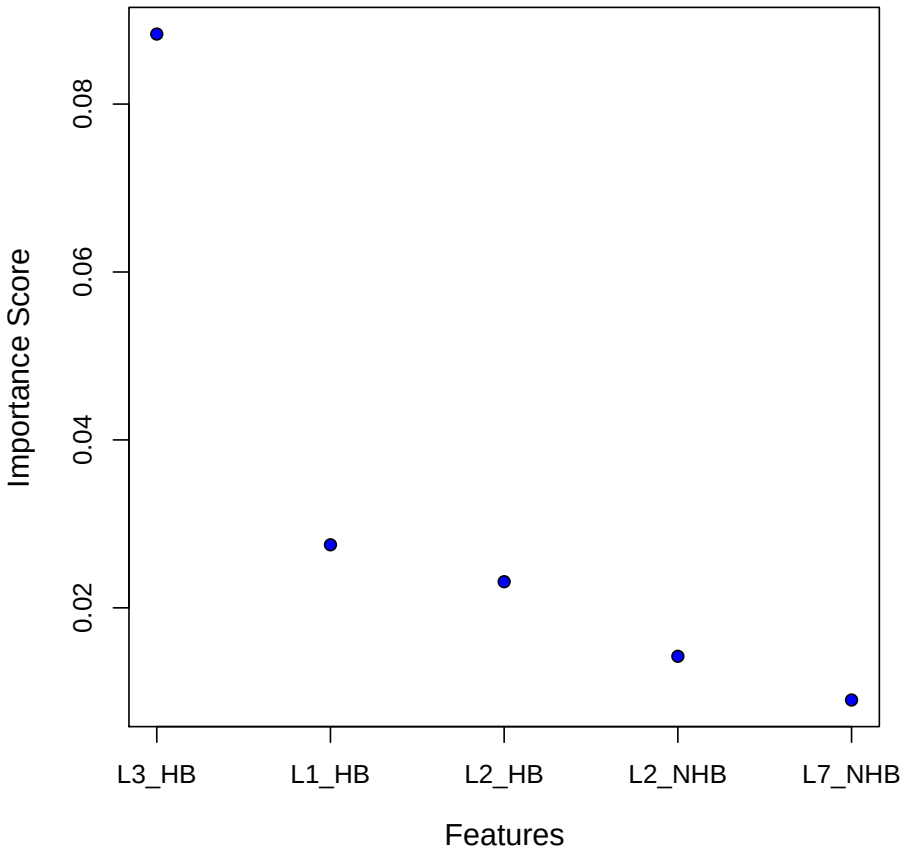
Hydrogen Bond Feature Selection



- A record low SV percentage
- Both positive and negative sets rely on H-Bond

Figure: Importance of hbond features

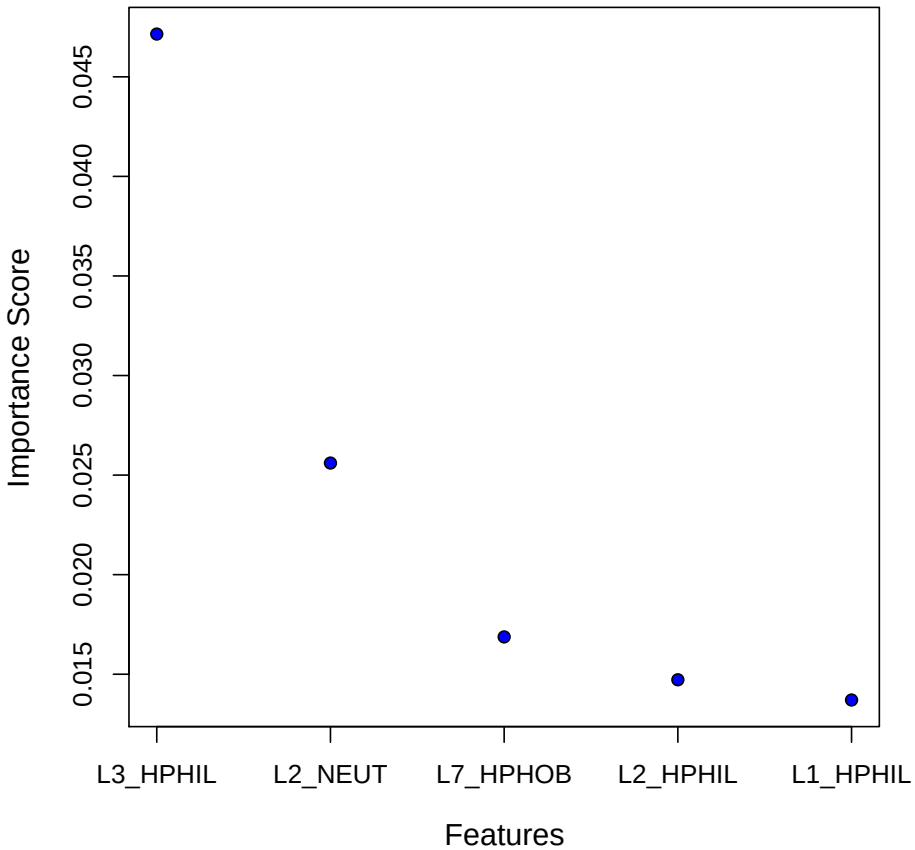
Hydrogen Bond Feature Selection



- A record low SV percentage
- Both positive and negative sets rely on H-Bond
- Differ by their distribution

Figure: Importance of hbond features

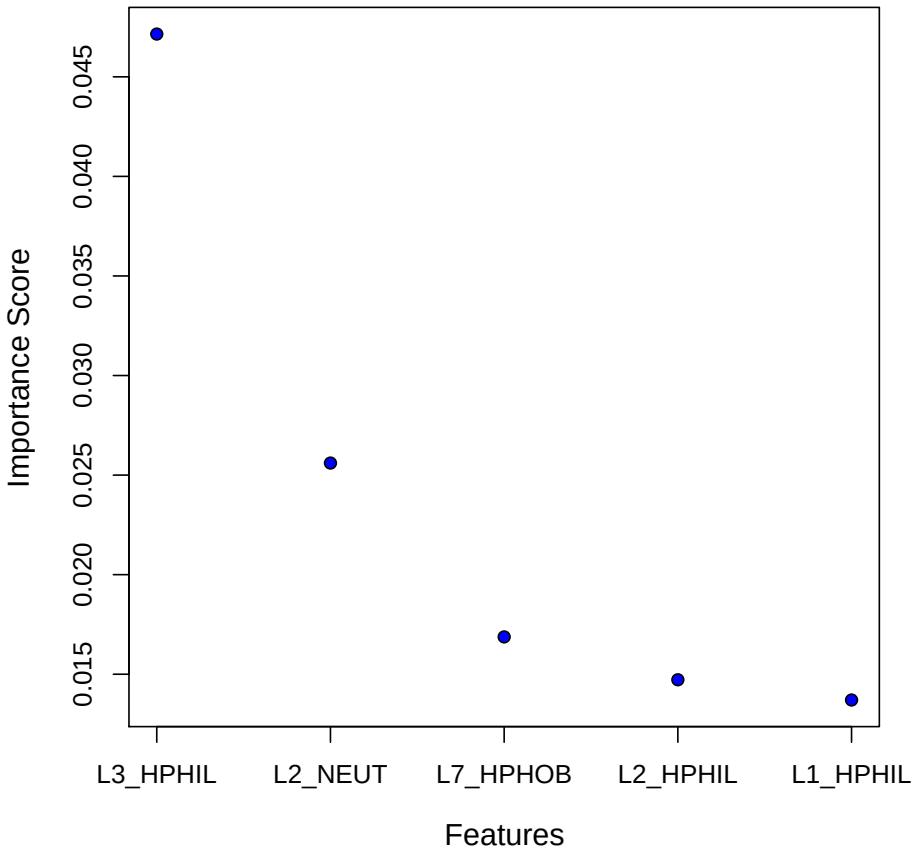
Hydrophobicity Feature Selection



- Remains the worst discriminating factor

Figure: Importance of hydro features

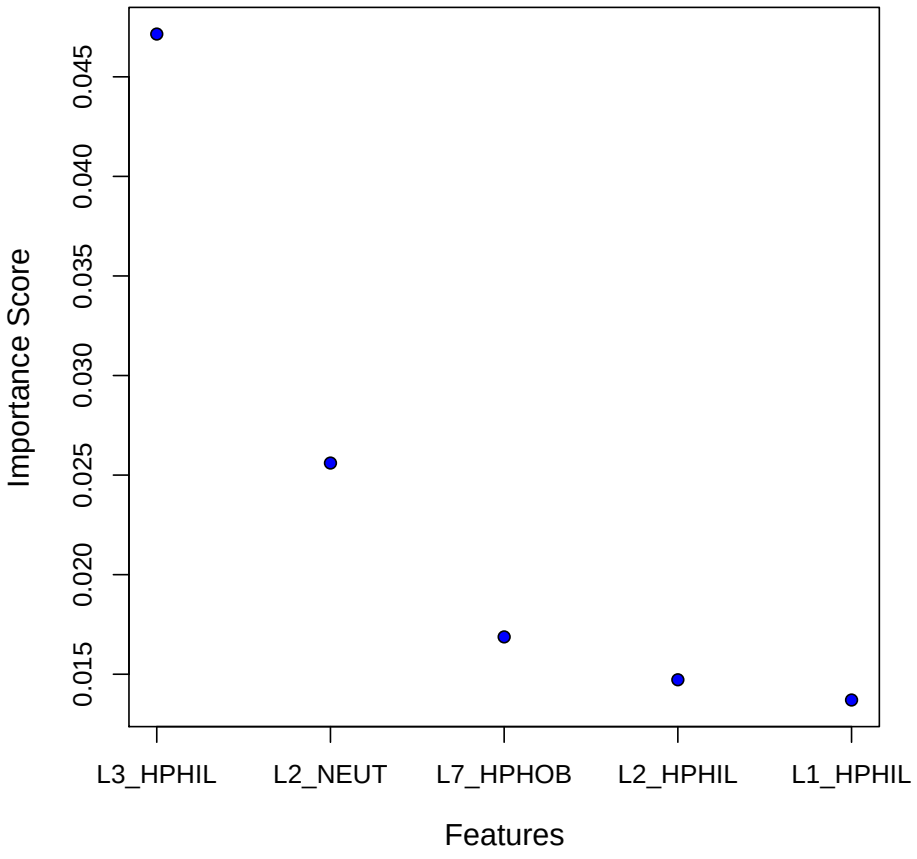
Hydrophobicity Feature Selection



- Remains the worst discriminating factor
- Positive set: high concentration of hydrophilic atoms in layer 3

Figure: Importance of hydro features

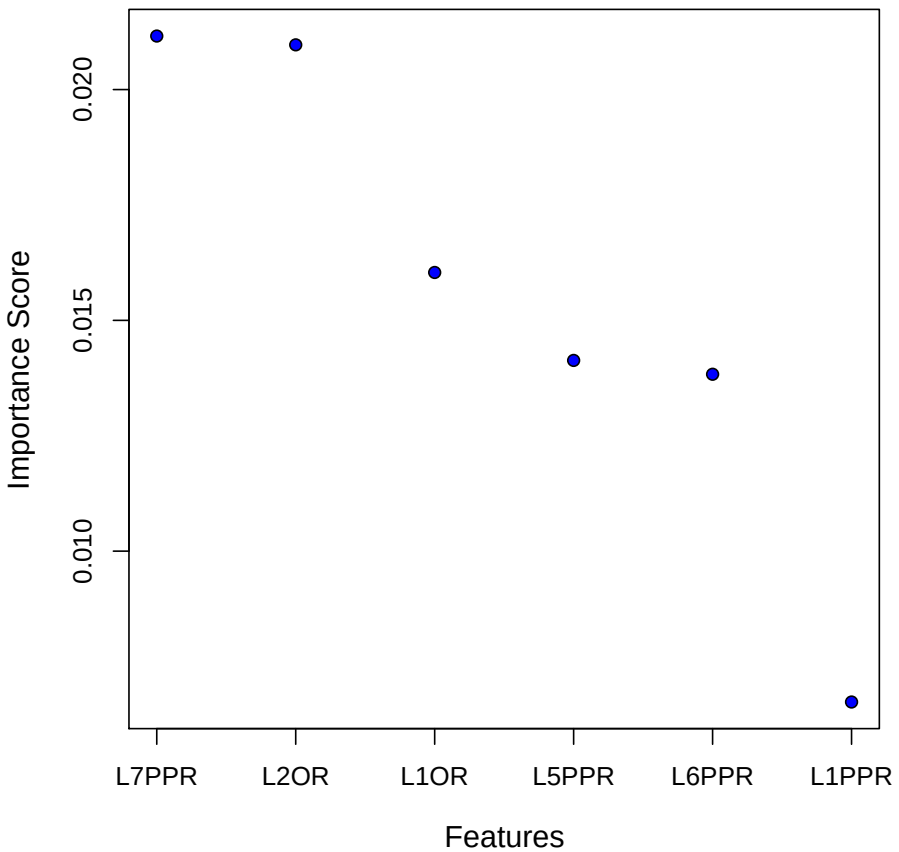
Hydrophobicity Feature Selection



- Remains the worst discriminating factor
- Positive set: high concentration of hydrophilic atoms in layer 3
- Low importance score of hydrophobic features

Figure: Importance of hydro features

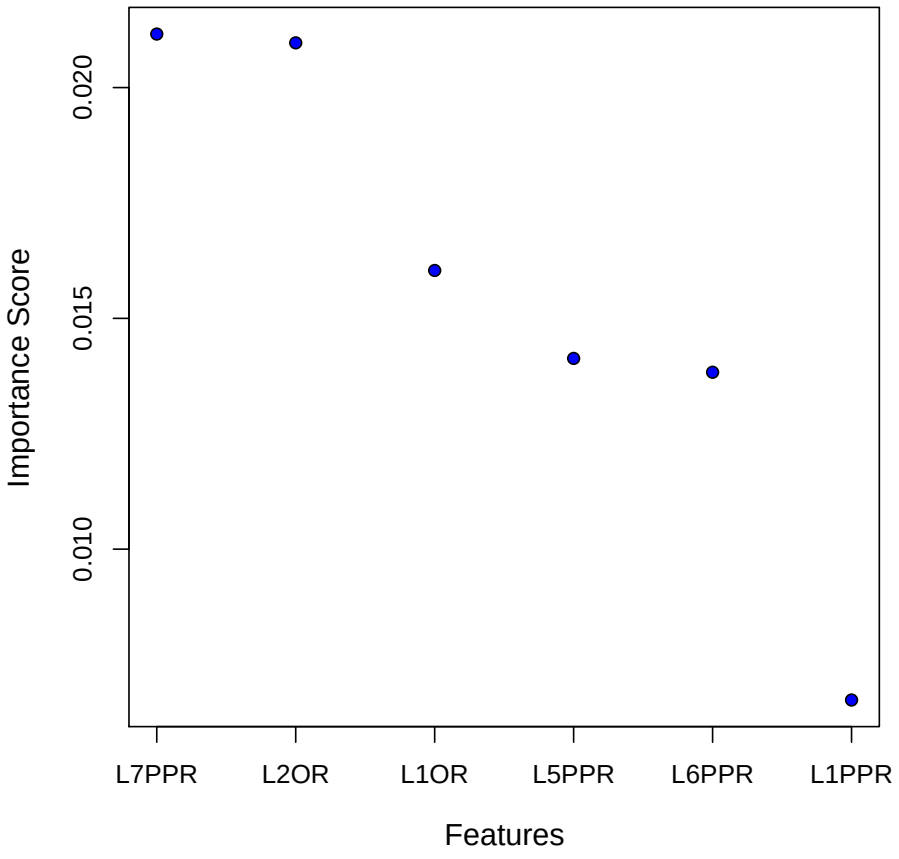
Simplified Residue Schemes Feature Selection



- Similar results for the 3 simplified schemes

Figure: Importance of simplified features

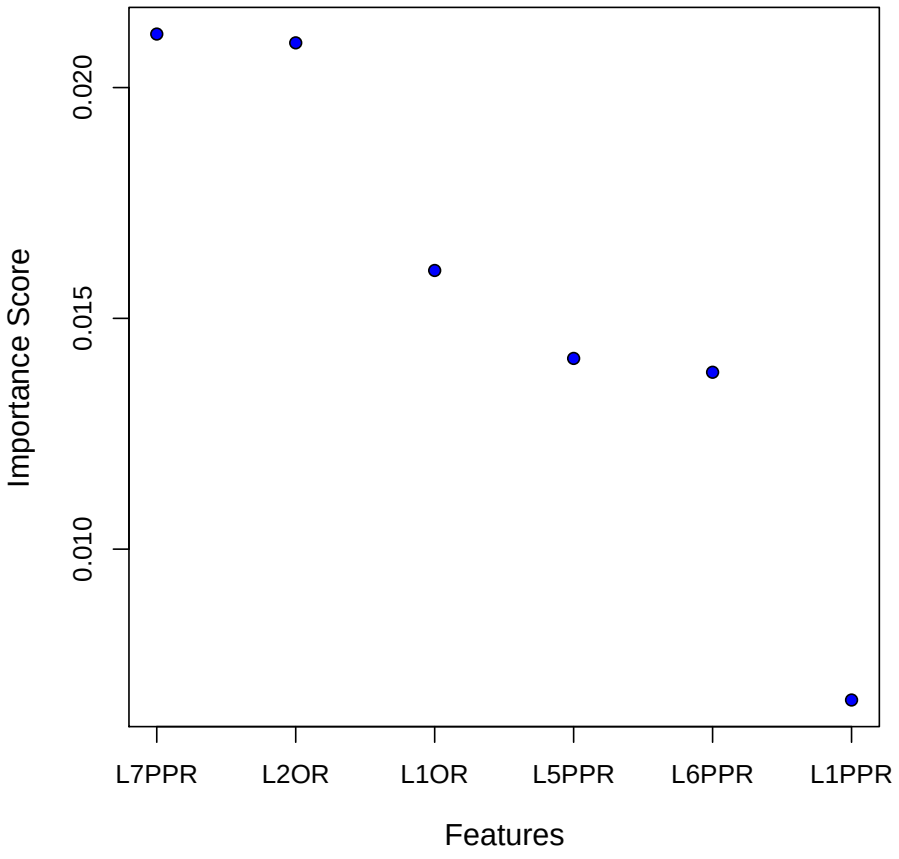
Simplified Residue Schemes Feature Selection



- Similar results for the 3 simplified schemes
- Importance of planar polar residues

Figure: Importance of simplified features

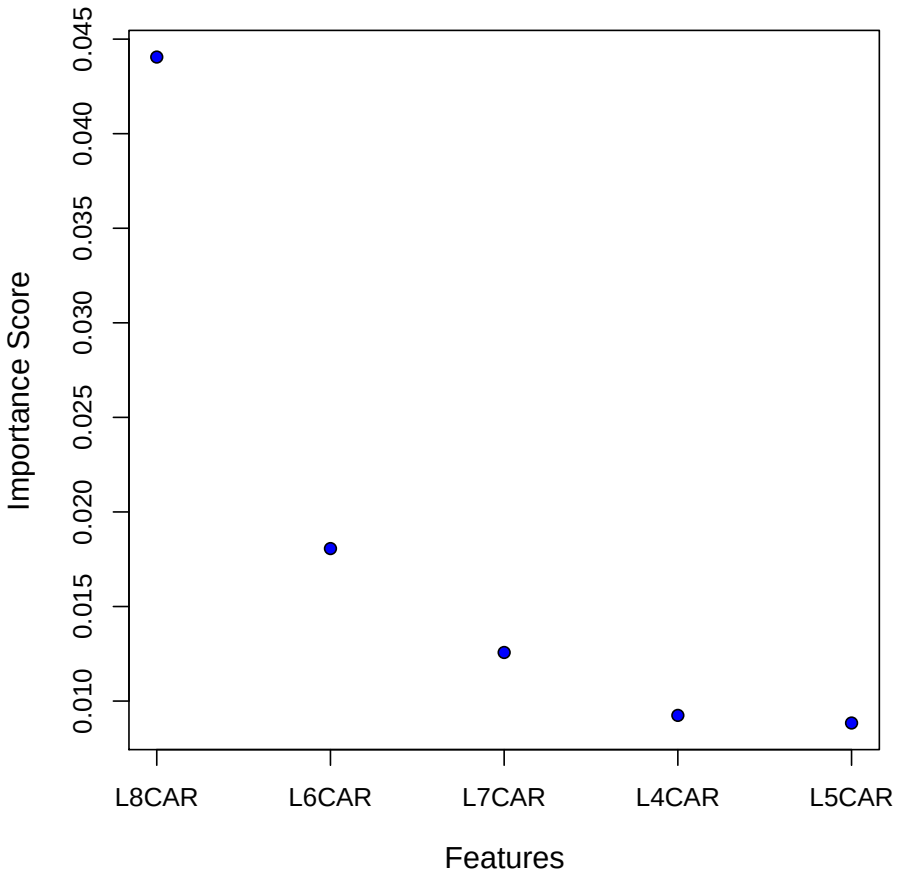
Simplified Residue Schemes Feature Selection



- Similar results for the 3 simplified schemes
- Importance of planar polar residues
- Low importance score of aromatic residues

Figure: Importance of simplified features

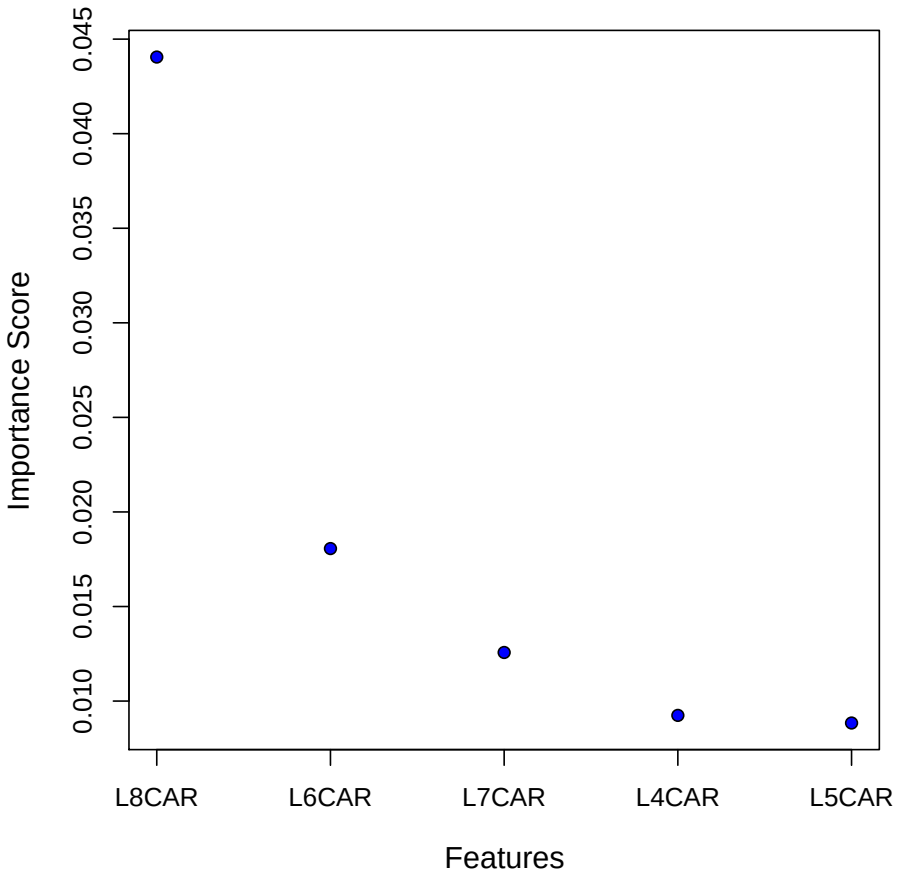
Detailed Residue Schemes Feature Selection



- Similar results for both detailed schemes

Figure: Importance of detailed features

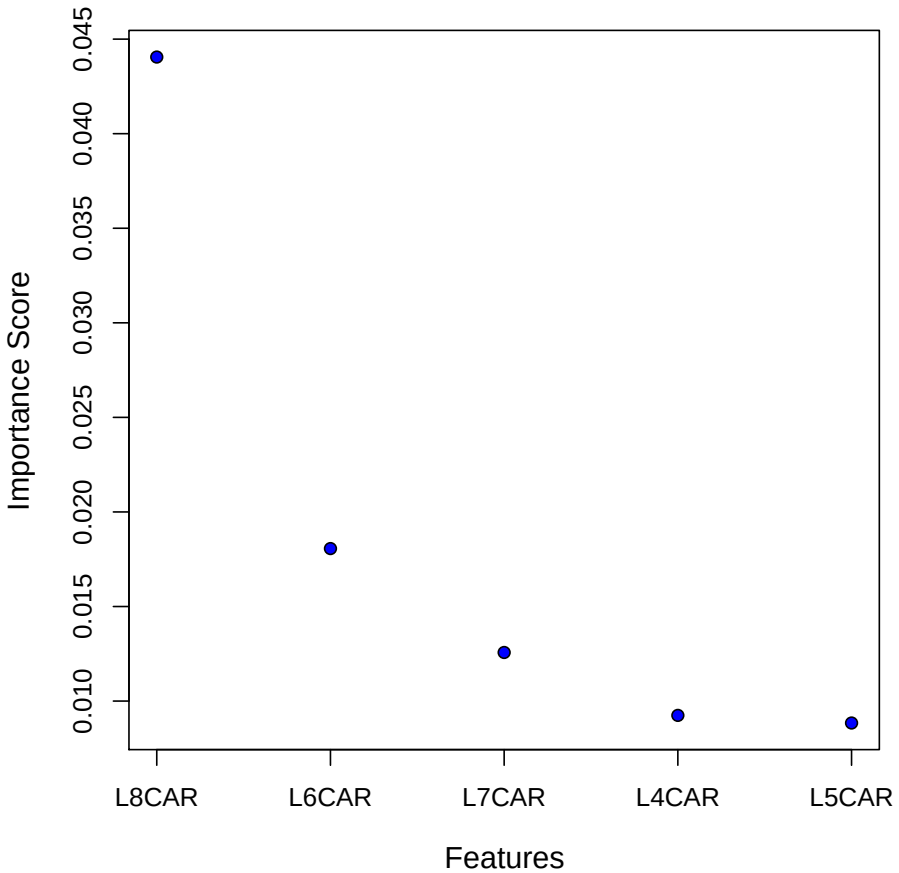
Detailed Residue Schemes Feature Selection



- Similar results for both detailed schemes
- Still outperform simplified schemes.

Figure: Importance of detailed features

Detailed Residue Schemes Feature Selection



- Similar results for both detailed schemes
- Still outperform simplified schemes.
- High density of the negatively charged carboxylate residues

Figure: Importance of detailed features

Combined Atomic and Residue Feature Selection

- 3.64% error for both classifiers.

Table: Classifiers performance using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensi-tivity	Speci-ficity	SV
Detailed2	false	104	14.06%	07.81%	93.10%	91.43%	53.13%
+ Charge							
+ H-Bond	true	15	03.64%	03.64%	96.55%	96.15%	69.09%
+ Hydro							

Selected Feature Subset

Table: The selected feature subset.

Property	Features	L1	L2	L3	L4	L5	L6	L7	L8
Charge	Negative			X					
	Neutral	X	X						
H-Bond	H-Bonding			X					
Hydro	Hydrophilic	X	X	X					
	Neutral		X						
	Hydrophobic							X	X
Residues	Neutral						X		
	Carboxylate					X	X		X
	Aliphatic			X					

Feature Selection Findings

- The difference in spatial configuration

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3
- Negatively charged layer 3

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3
- Negatively charged layer 3
- High density of negatively charged carboxylate residues

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3
- Negatively charged layer 3
- High density of negatively charged carboxylate residues
- The relevance of planar polar residues

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3
- Negatively charged layer 3
- High density of negatively charged carboxylate residues
- The relevance of planar polar residues
- Low discrimination capacity of hydrophobic interactions and aromatic residues

Feature Selection Findings

- The difference in spatial configuration
- Charge is the best discriminant property
- Discriminating atomic properties on layer 3
- Negatively charged layer 3
- High density of negatively charged carboxylate residues
- The relevance of planar polar residues
- Low discrimination capacity of hydrophobic interactions and aromatic residues
- Possibly due to their one-correctly-placed relationship

Testing Phase

- Parse PDB for entries newer than October 2004

Testing Phase

- Parse PDB for entries newer than October 2004
- 7 distinct glucose binding sites

Testing Phase

- Parse PDB for entries newer than October 2004
- 7 distinct glucose binding sites

Table: Testing phase results.

Classifier	TP	FP	TN	FN	Error	Sensitivity	Specificity	Precision
SVM	6	0	15	1	04.55%	85.71%	100%	100%
*k*NN	6	2	13	1	13.64%	85.71%	86.67%	75.00%
COTRAN	94	27	633	12	05.09%	88.68%	95.91%	77.69%

Testing Phase

- Parse PDB for entries newer than October 2004
- 7 distinct glucose binding sites
- Compare with COTRAN, a galactose-binding site classifier

Table: Testing phase results.

Classifier	TP	FP	TN	FN	Error	Sensitivity	Specificity	Precision
SVM	6	0	15	1	04.55%	85.71%	100%	100%
kNN	6	2	13	1	13.64%	85.71%	86.67%	75.00%
COTRAN	94	27	633	12	05.09%	88.68%	95.91%	77.69%

Testing Phase

- Parse PDB for entries newer than October 2004
- 7 distinct glucose binding sites
- Compare with COTRAN, a galactose-binding site classifier
- *Precision:* ability to reject false positives.

Table: Testing phase results.

Classifier	TP	FP	TN	FN	Error	Sensitivity	Specificity	Precision
SVM	6	0	15	1	04.55%	85.71%	100%	100%
*k*NN	6	2	13	1	13.64%	85.71%	86.67%	75.00%
COTRAN	94	27	633	12	05.09%	88.68%	95.91%	77.69%

Future Work

- Add more experiments

Future Work

- Add more experiments
- Tune fixed parameters: sphere radius, layers number ...

Future Work

- Add more experiments
- Tune fixed parameters: sphere radius, layers number ...
- Wet lab experiments

Future Work

- Add more experiments
- Tune fixed parameters: sphere radius, layers number ...
- Wet lab experiments
- Web based interface

Future Work

- Add more experiments
- Tune fixed parameters: sphere radius, layers number ...
- Wet lab experiments
- Web based interface
- Extend to a glucose binding-site finder

Future Work

- Add more experiments
- Tune fixed parameters: sphere radius, layers number ...
- Wet lab experiments
- Web based interface
- Extend to a glucose binding-site finder
- Extend this study to other sugars

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad
- Mr. Al-Ali for providing all the data and closely assisting me

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad
- Mr. Al-Ali for providing all the data and closely assisting me
- Dr. Keyrouz for supervising my thesis and supporting me

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad
- Mr. Al-Ali for providing all the data and closely assisting me
- Dr. Keyrouz for supervising my thesis and supporting me
- Dr. Kachfe for his many discussions

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad
- Mr. Al-Ali for providing all the data and closely assisting me
- Dr. Keyrouz for supervising my thesis and supporting me
- Dr. Kachfe for his many discussions
- The committee members

Acknowledgments

Finally, I would like to thank

- Dr. Khuri for initiating and constantly supervising this work from abroad
- Mr. Al-Ali for providing all the data and closely assisting me
- Dr. Keyrouz for supervising my thesis and supporting me
- Dr. Kachfe for his many discussions
- The committee members
- The audience

Amino Acids Table

Table: Standard amino acids names, three- and one-letter abbreviations. They are sorted according to biological convention based on the size and properties of the side chain.

Name	3-L	1-L	Name	3-L	1-L
Glycine	Gly	G	Cysteine	Cys	C
Alanine	Ala	A	Serine	Ser	S
Valine	Val	V	Threonine	Thr	T
Leucine	Leu	L	Aspartate	Asp	D
Isoleucine	Ile	I	Glutamate	Glu	E
Proline	Pro	P	Histidine	His	H
Phenylalanine	Phe	F	Lysine	Lys	K
Tyrosine	Tyr	Y	Arginine	Arg	R
Tryptophan	Trp	W	Asparagine	Asn	N
Methionine	Met	M	Glutamine	Gln	Q

Atomic Features

Atom Type	Functional Group	Location	Residue	PDB Atom Symbol	Chrg	Hydrophob	H Bond
Oxygen	Amide peptide linkage	Backbone	All	O	0	-1	H Bond
Oxygen	Carboxyl – C terminus	Backbone	All	OXT	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE2	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD2	-ve	-1	H Bond
Oxygen	Amide	Side Chain	GLN	OE1	0	-1	H Bond
Oxygen	Amide	Side Chain	ASN	OD1	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	SER	OG	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	THR	OG1	0	-1	H Bond
Oxygen	Hydroxyl - Phenolic	Side Chain	TYR	OH	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	All except PRO	N	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	PRO	N	0	-1	--
Nitrogen	Amide	Side Chain	GLN	NE2	0	-1	H Bond
Nitrogen	Amide	Side Chain	ASN	ND2	0	-1	H Bond
Nitrogen	Amine	Side Chain	LYS	NZ	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NE	+ve	-1	--
Nitrogen	Guanidino	Side Chain	ARG	NH1	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NH2	+ve	-1	H Bond
Nitrogen	Imidazole	Side Chain	HIS	ND1	0	-1	--
Nitrogen	Imidazole	Side Chain	HIS	NE2	0	-1	H Bond
Nitrogen	Indole	Side Chain	TRP	NE1	0	0	--
Carbon	Amide peptide linkage	Backbone	All	C	0	0	--
Carbon	C-alpha	Backbone	All	CA	0	0	--
Carbon	Aliphatic – neutral	Side Chain	Set A (See below)	CB, CG, CD, CE	0	0	--
Carbon	Aliphatic – hydrophobic	Side Chain	LEU, VAL, ILE, MET	CB, CG, CD, CE	0	1	--
Carbon	Aliphatic – Branch	Side Chain	LEU, VAL, ILE	CG1, CG2, CD1, CD2, CD1	0	1	--
Carbon	Phenyl - aromatic	Side Chain	PHE, TYR	CG,CD1, CD2, CE1, CE2, CZ	0	1	--
Carbon	Imidazole	Side Chain	HIS	CG, CD2, CE1	0	1	--
Carbon	Aromatic	Side Chain	TRP	CG,CD1, CD2,	0	1	--
Carbon	Aromatic	Side Chain	TRP	CE2, CE3, CZ2, CZ3, CH2	0	1	--
Sulfur	Sulfhydril	Side Chain	CYS	SG	0	-1	H Bond
Sulfur	Thioether	Side Chain	MET	SD	0	0	--
Oxygen	Sulfate	HET Group	SO4	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Phosphate	HET Group	2HP	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Water	HET Group	HOH	O	0	-1	H Bond
Calcium	Ion	HET Group	CA	CA	+ve	-1	H Bond
Magnesium	Ion	HET Group	MG	MG	+ve	-1	H Bond
Zinc	Ion	HET Group	ZN	ZN	+ve	-1	H Bond

Set A = ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO

Atomic and Residue Combinations Results

Table: Atomic and residue properties classification error rates.

Properties	kNN	SVM	SV
Residue (detailed1 scheme)	16.36%	14.55%	81.82%
Residue + Charge	18.18%	10.91%	100%
Residue + H-Bond	16.36%	10.91%	94.55%
Residue + Hydro	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	16.36%	10.91%	100%
Residue + Charge + Hydro	20.00%	09.09%	100%
Residue + H-Bond + Hydro	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	18.18%	09.09%	100%

Atomic and Residue Combinations Results

- SVM performs better

Table: Atomic and residue properties classification error rates.

Properties	kNN	SVM	SV
Residue (detailed1 scheme)	16.36%	14.55%	81.82%
Residue + Charge	18.18%	10.91%	100%
Residue + H-Bond	16.36%	10.91%	94.55%
Residue + Hydro	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	16.36%	10.91%	100%
Residue + Charge + Hydro	20.00%	09.09%	100%
Residue + H-Bond + Hydro	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	18.18%	09.09%	100%

Atomic and Residue Combinations Results

- SVM performs better
- kNN deteriorates

Table: Atomic and residue properties classification error rates.

Properties	kNN	SVM	SV
Residue (detailed1 scheme)	16.36%	14.55%	81.82%
Residue + Charge	18.18%	10.91%	100%
Residue + H-Bond	16.36%	10.91%	94.55%
Residue + Hydro	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	16.36%	10.91%	100%
Residue + Charge + Hydro	20.00%	09.09%	100%
Residue + H-Bond + Hydro	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	18.18%	09.09%	100%

Atomic and Residue Combinations Results

- Residue property constrains the differences between the various combinations of atomic features

Table: Atomic and residue properties classification error rates.

Properties	kNN	SVM	SV
Residue (detailed1 scheme)	16.36%	14.55%	81.82%
Residue + Charge	18.18%	10.91%	100%
Residue + H-Bond	16.36%	10.91%	94.55%
Residue + Hydro	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	16.36%	10.91%	100%
Residue + Charge + Hydro	20.00%	09.09%	100%
Residue + H-Bond + Hydro	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	18.18%	09.09%	100%

Atomic and Residue Combinations Results

- "Charge + Hydro + Residue " is the best combination

Table: Atomic and residue properties classification error rates.

Properties	kNN	SVM	SV
Residue (detailed1 scheme)	16.36%	14.55%	81.82%
Residue + Charge	18.18%	10.91%	100%
Residue + H-Bond	16.36%	10.91%	94.55%
Residue + Hydro	16.36%	10.91%	90.91%
Residue + Charge + H-Bond	16.36%	10.91%	100%
Residue + Charge + Hydro	20.00%	09.09%	100%
Residue + H-Bond + Hydro	18.18%	10.91%	98.18%
Residue + Charge + H-Bond + Hydro	18.18%	09.09%	100%

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor
- Glucose exhibits a dual hydrophobic-hydrophilic nature

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

- Glucose exhibits a dual hydrophobic-hydrophilic nature
- Both antagonist properties are involved in glucose binding

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

- Glucose exhibits a dual hydrophobic-hydrophilic nature
- Both antagonist properties are involved in glucose binding
  1. Hydrophobic: Aromatic residues stack against pyranose ring

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

- Glucose exhibits a dual hydrophobic-hydrophilic nature
- Both antagonist properties are involved in glucose binding
  1. Hydrophobic: Aromatic residues stack against pyranose ring
  2. Hydrophilic: Planar-polar residues form hydrogen-bonds

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

- Glucose exhibits a dual hydrophobic-hydrophilic nature
- Both antagonist properties are involved in glucose binding
  1. Hydrophobic: Aromatic residues stack against pyranose ring
  2. Hydrophilic: Planar-polar residues form hydrogen-bonds
- May explain why hydrophobicity is the worst discriminant

Hydrophobicity and Charge Properties

- Charge: the best discriminating factor

- Glucose exhibits a dual hydrophobic-hydrophilic nature
- Both antagonist properties are involved in glucose binding
  1. Hydrophobic: Aromatic residues stack against pyranose ring
  2. Hydrophilic: Planar-polar residues form hydrogen-bonds
- May explain why hydrophobicity is the worst discriminant
- Hydrophobicity: secondary discriminating factor

Detailed Schemes Comparison

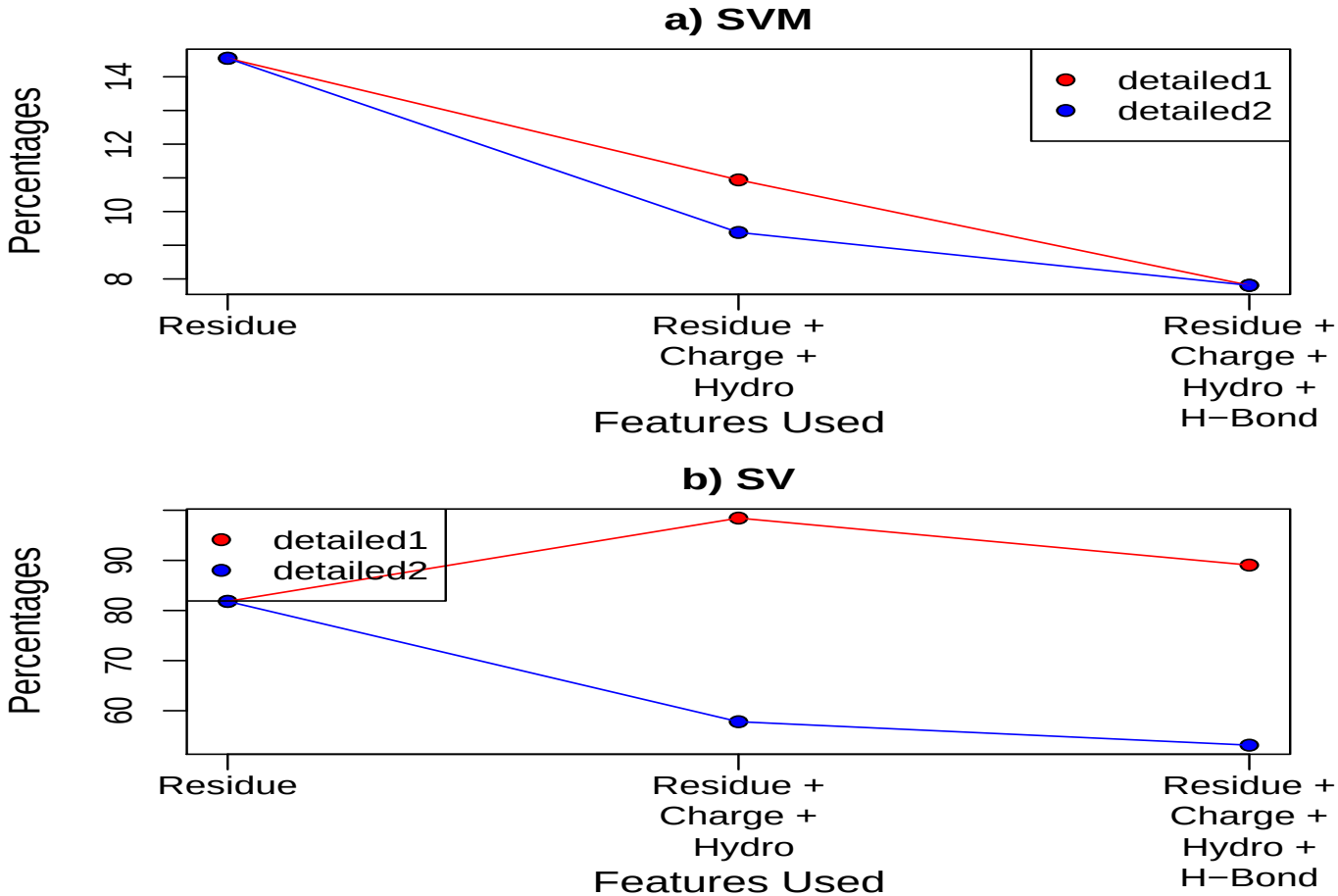


Figure: Comparison of detailed1 and detailed2 schemes

Residue Schemes Feature Selection

Table: Classifiers performance on residue data using feature selection.

Property	RF	Feat. Nb.	kNN Error	SVM Error	Sensitivity	Specificity	SV
Detailed1	false	48	16.36%	14.55%	89.66%	80.77%	81.82%
	true	19	09.09%	12.73%	93.10%	88.46%	56.36%
Detailed2	false	40	21.82%	14.55%	96.55%	73.08%	81.82%
	true	3	10.91%	12.73%	89.66%	88.46%	56.36%
Simplified1	false	16	16.36%	18.18%	96.55%	69.23%	76.36%
	true	6	12.73%	12.73%	93.10%	80.77%	67.27%
Simplified2	false	24	18.18%	18.18%	86.21%	76.92%	70.91%
	true	4	10.91%	10.91%	96.55%	80.77%	54.55%
Simplified3	false	24	16.36%	16.36%	96.55%	69.23%	70.91%
	true	6	12.73%	12.73%	96.55%	76.92%	74.55%

The Selected Features Importance

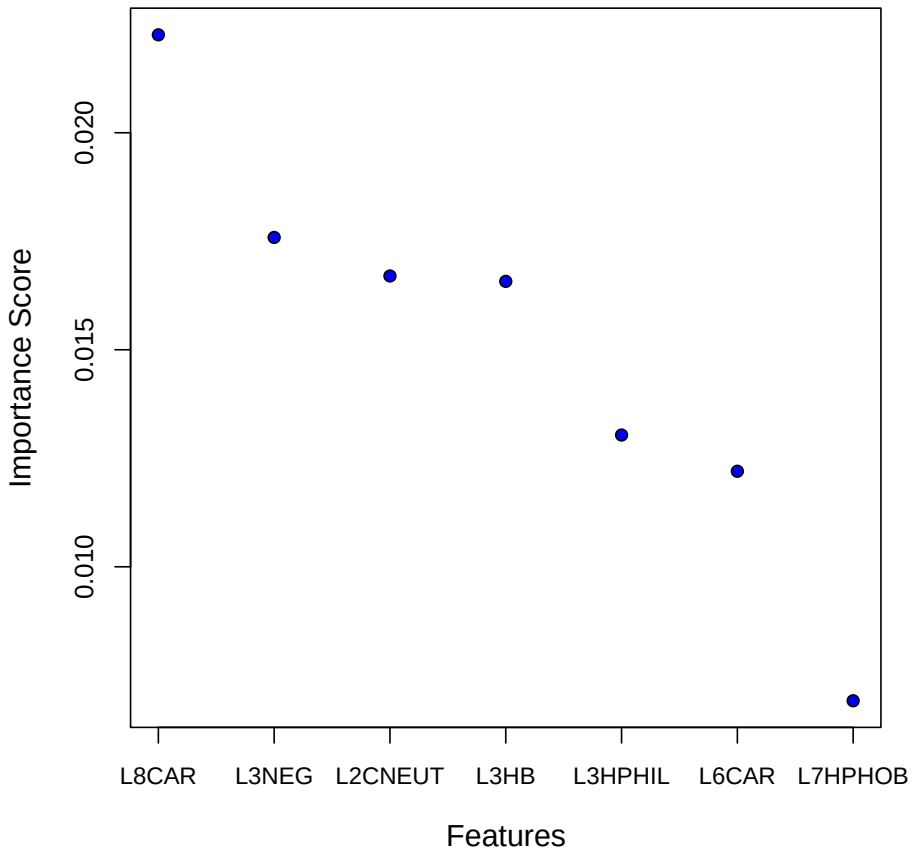


Figure: Importance of the selected features subset