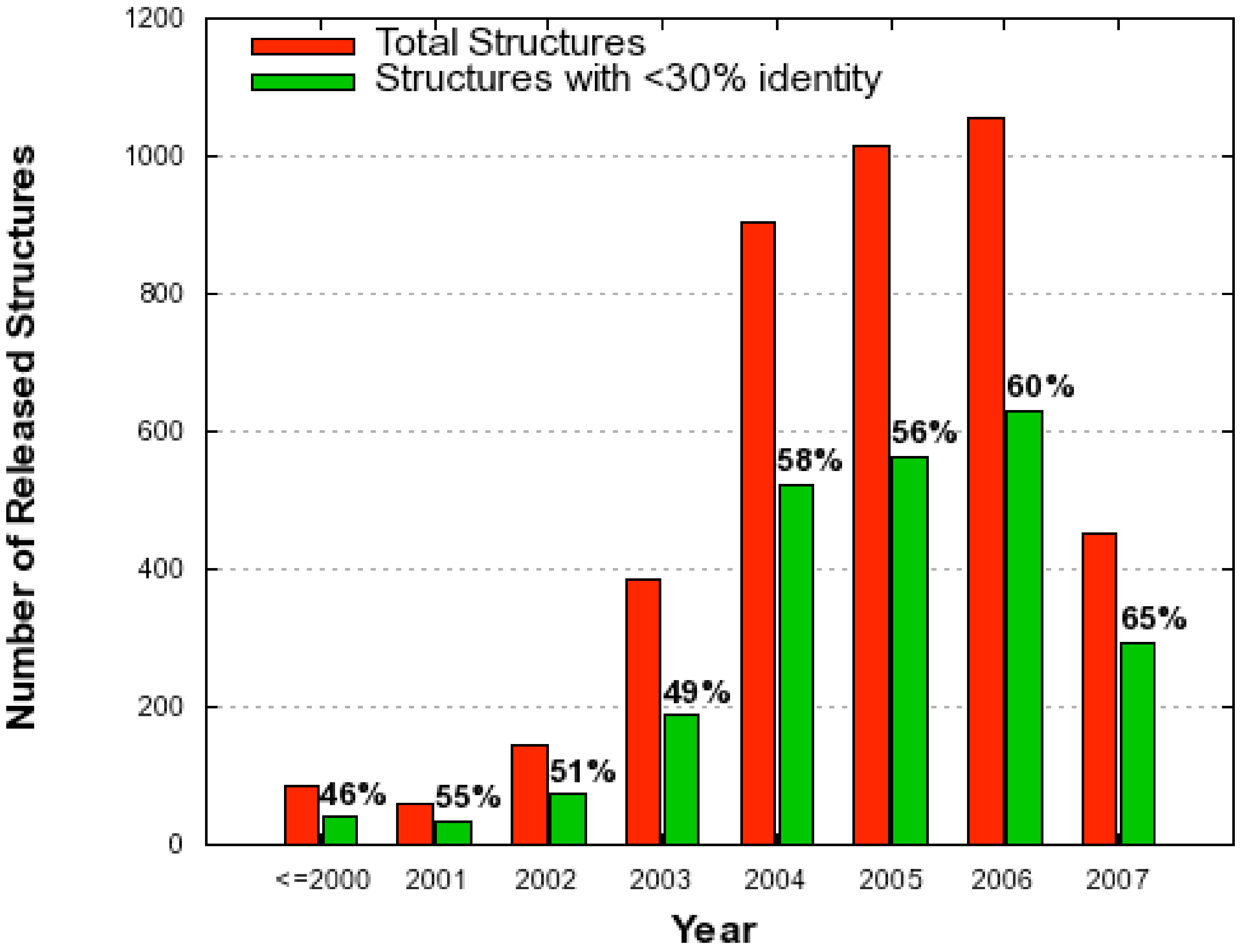
# Combining simulation and machine learning to recognize function in 4D

Russ B. Altman, MD, PhD

Departments of Bioengineering, Genetics, Medicine, Computer Science
Stanford University

# Biological Motivations

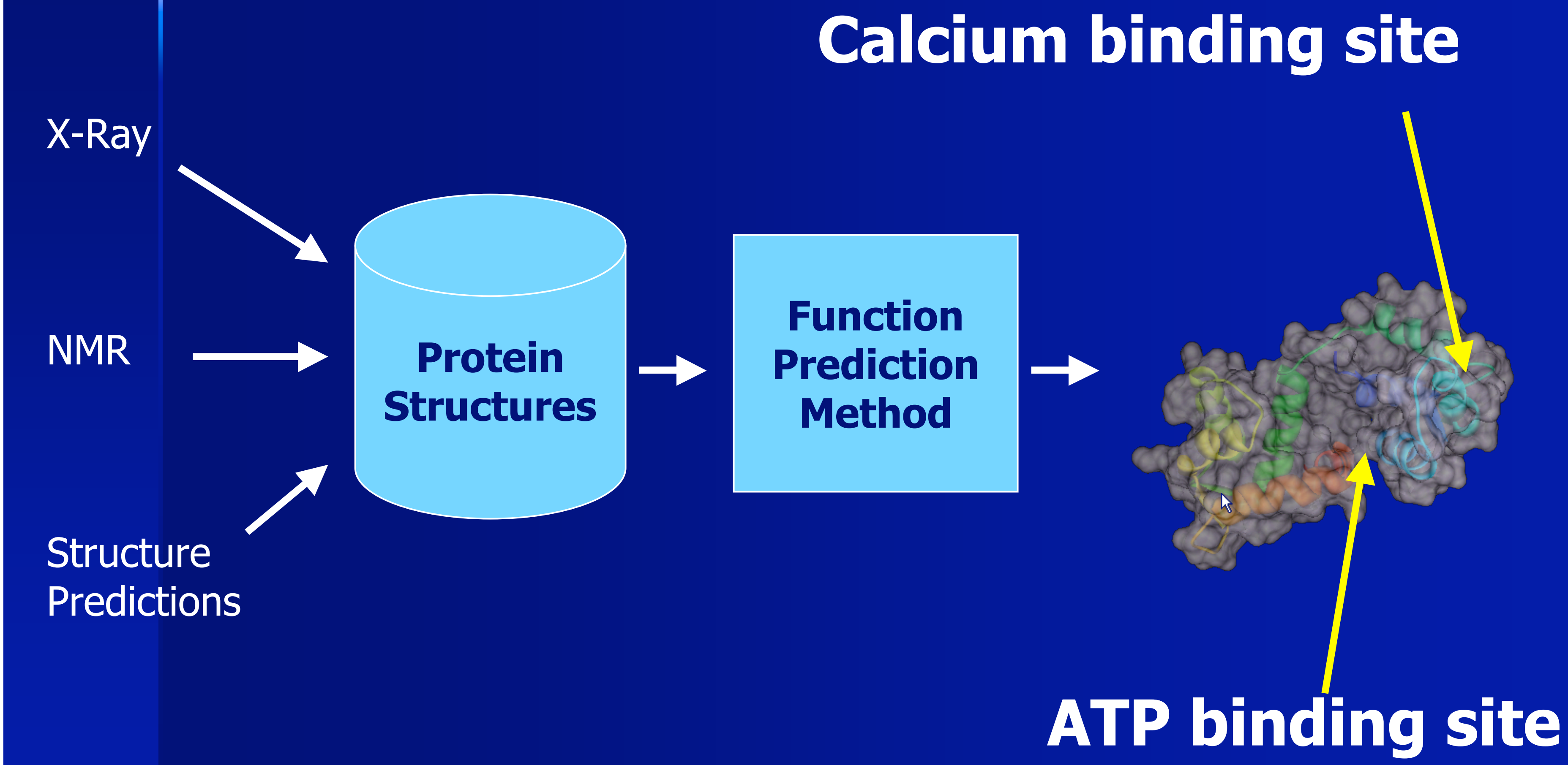
- Structural genomics and prediction (structures without functions)
- Need to label putative molecular functions of new structures
- <u>Need to discover new molecular functions</u>
- Increasing examples of polyfunctional proteins.


Total Structures
Structures with <30% identity
Number of Released Structures
1200
1000
800
600
400
200
0
46%
55%
51%
49%
58%
56%
60%
65%
<=2000
2001
2002
2003
2004
2005
2006
2007
Year

# Outline

- Introduction to FEATURE & WebFEATURE
- Using molecular simulation to improve FEATURE

# Protein Function Prediction & Annotation



X-Ray

NMR

Structure Predictions

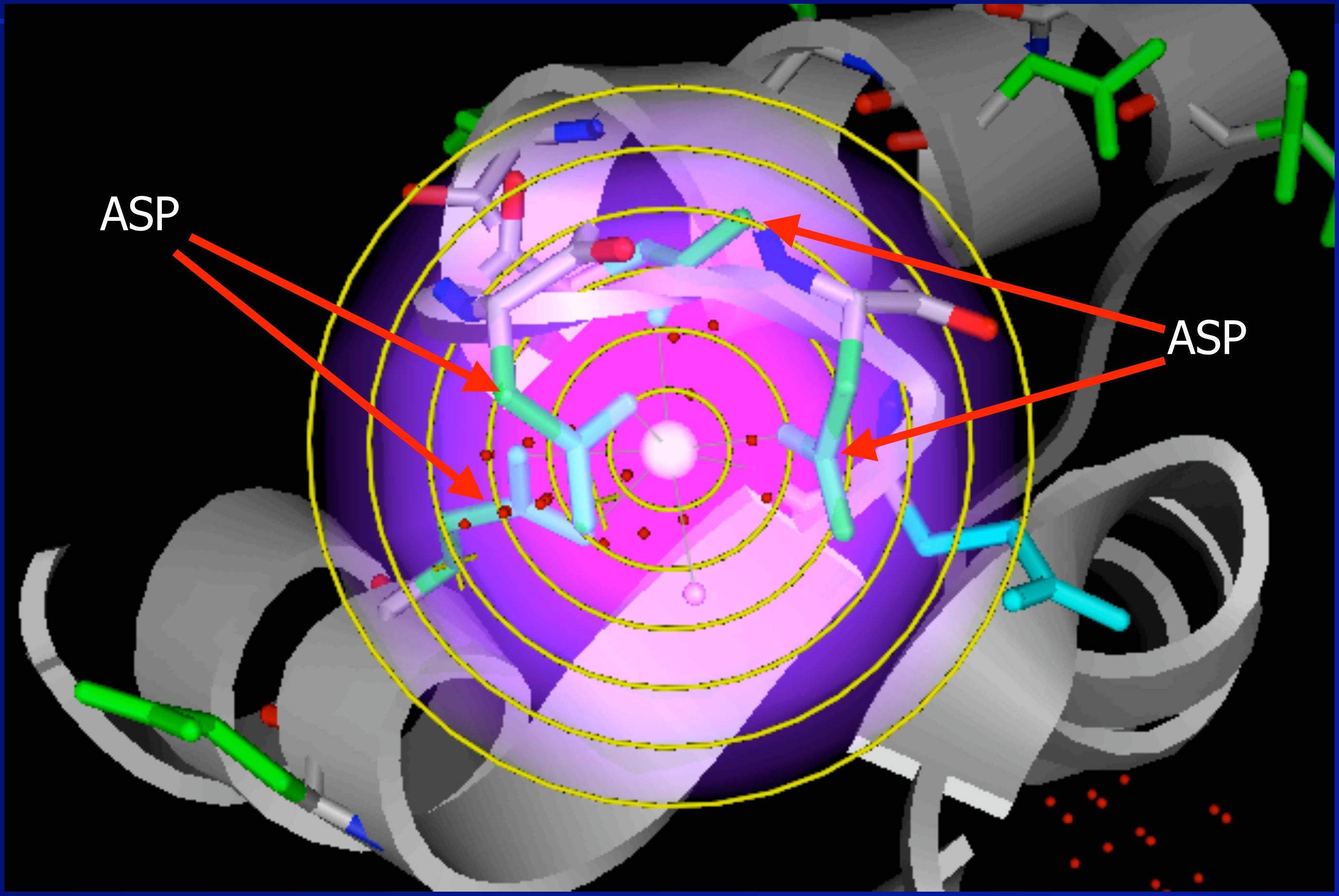
Protein Structures

Function Prediction Method

Calcium binding site

ATP binding site

# Radial Microenvironment



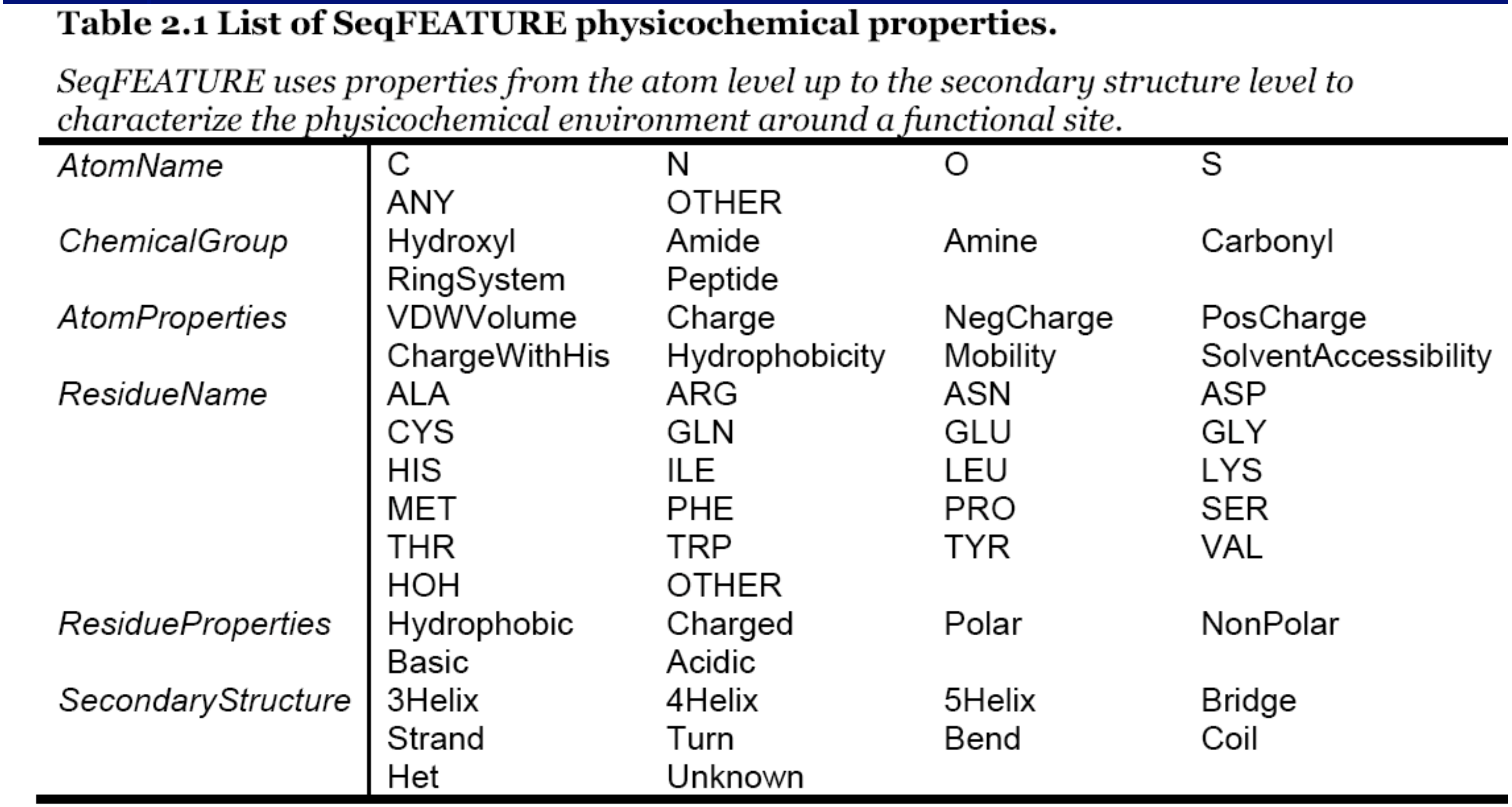
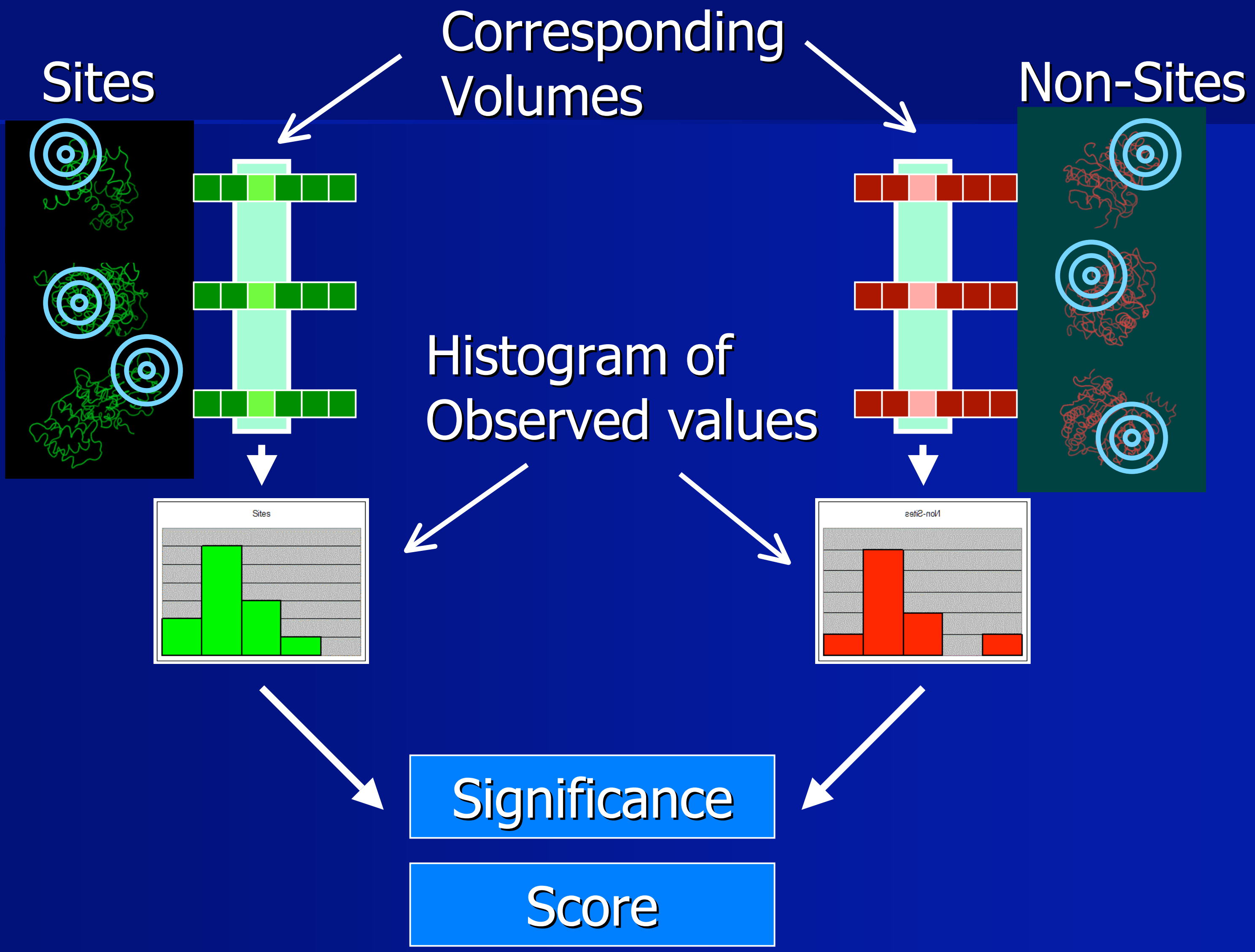


**Table 2.1 List of SeqFEATURE physicochemical properties.**

*SeqFEATURE uses properties from the atom level up to the secondary structure level to characterize the physicochemical environment around a functional site.*

| | | | | |
|---|---|---|---|---|
| *AtomName* | C | N | O | S |
| | ANY | OTHER | | |
| *ChemicalGroup* | Hydroxyl | Amide | Amine | Carbonyl |
| | RingSystem | Peptide | | |
| *AtomProperties* | VDWVolume | Charge | NegCharge | PosCharge |
| | ChargeWithHis | Hydrophobicity | Mobility | SolventAccessibility |
| *ResidueName* | ALA | ARG | ASN | ASP |
| | CYS | GLN | GLU | GLY |
| | HIS | ILE | LEU | LYS |
| | MET | PHE | PRO | SER |
| | THR | TRP | TYR | VAL |
| | HOH | OTHER | | |
| *ResidueProperties* | Hydrophobic | Charged | Polar | NonPolar |
| | Basic | Acidic | | |
| *SecondaryStructure* | 3Helix | 4Helix | 5Helix | Bridge |
| | Strand | Turn | Bend | Coil |
| | Het | Unknown | | |

# FEATURE Training



Sites

Corresponding Volumes

Non-Sites

Histogram of Observed values

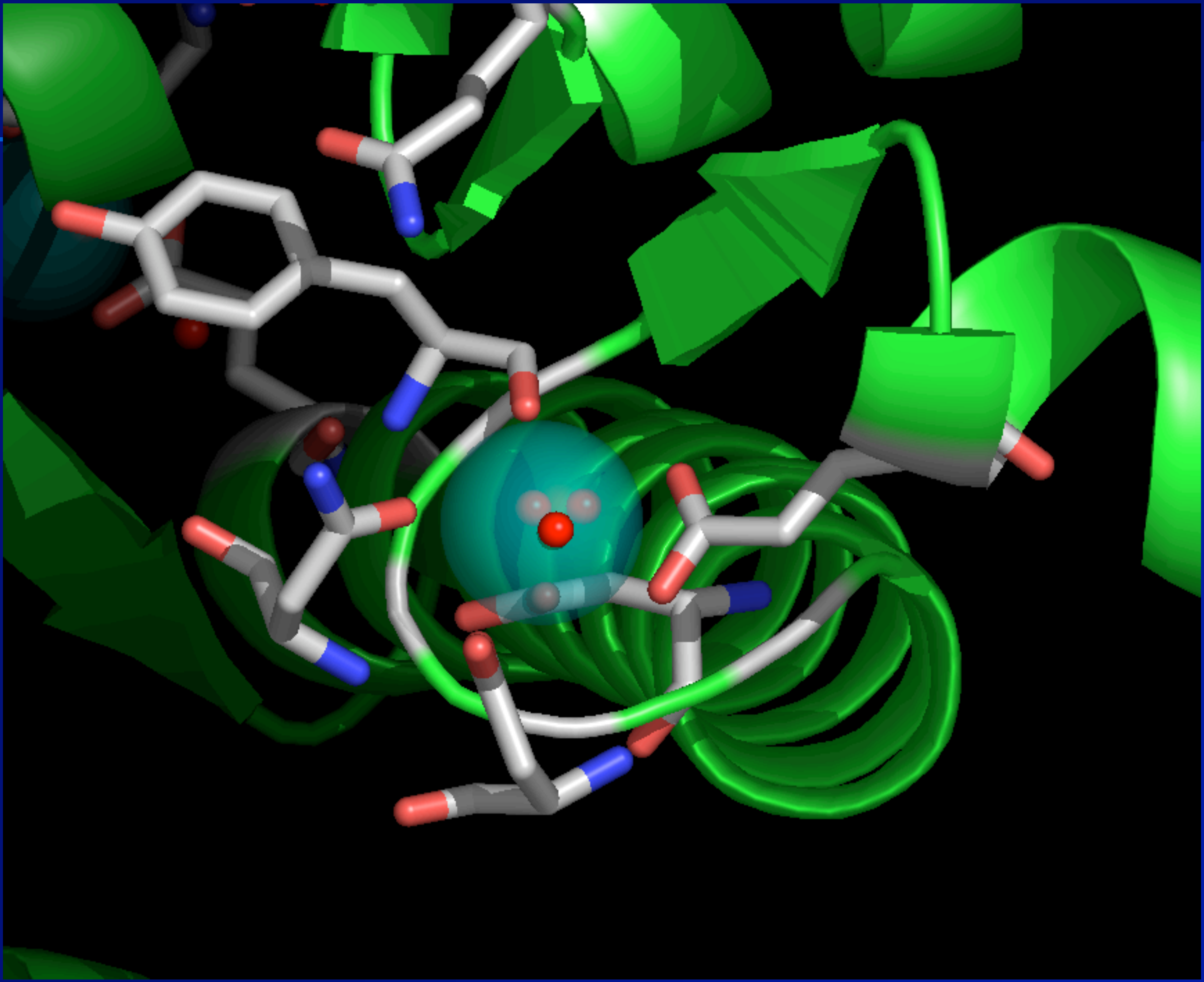
Significance

Score

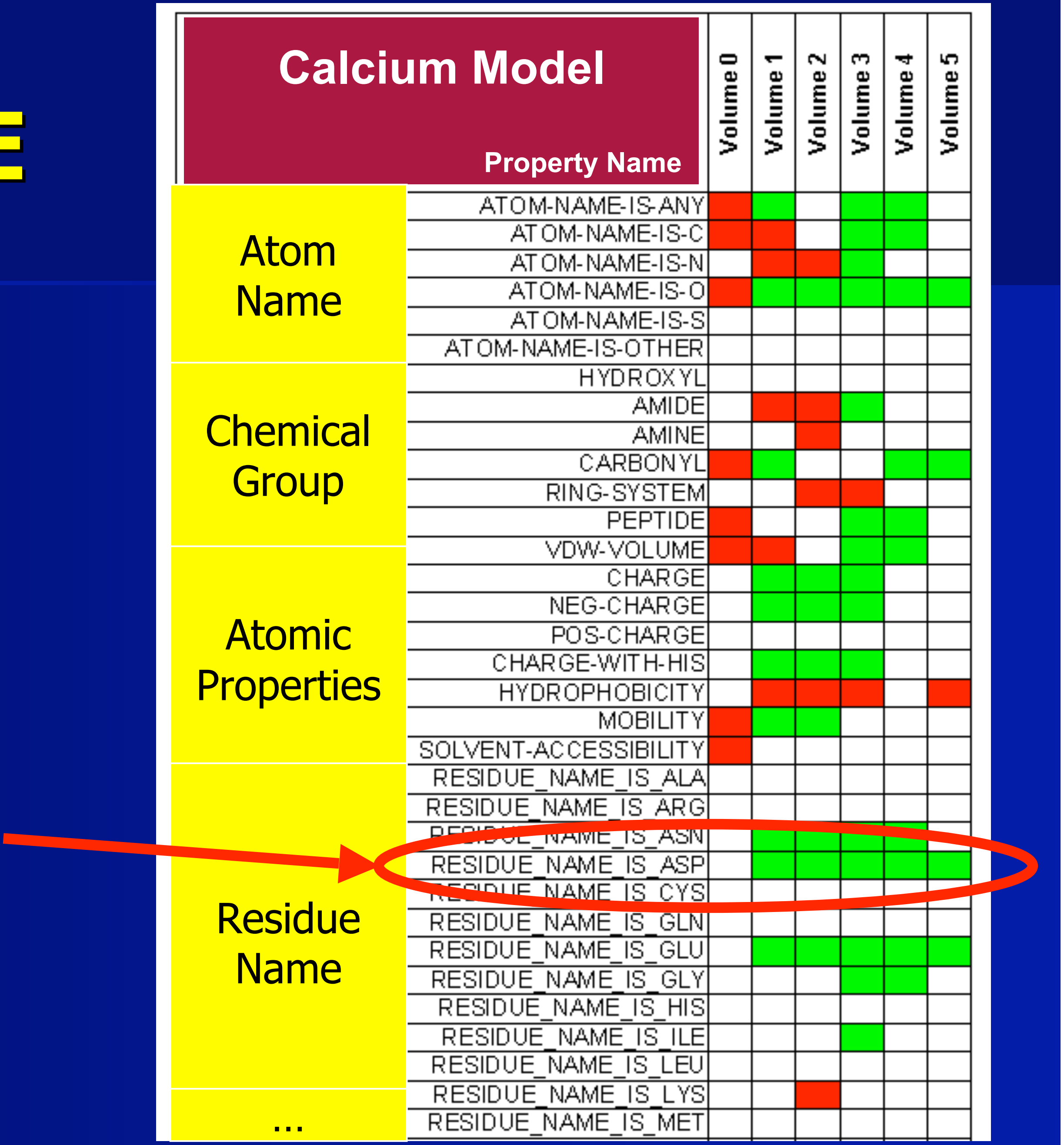
# FEATURE vectors

- 44 vectors of physical/chemical features
- Counted in 6 shells
- 44 x 6 = 264 feature-shell combinations that summarize the abundance of a feature in a shell (e.g. # of oxygens in shell that is 2-3 Angstroms from center)
- Vector = 264 continuous or discrete values describing environment around site center.

# Example: Calcium binding site



# FEATURE Calcium Model

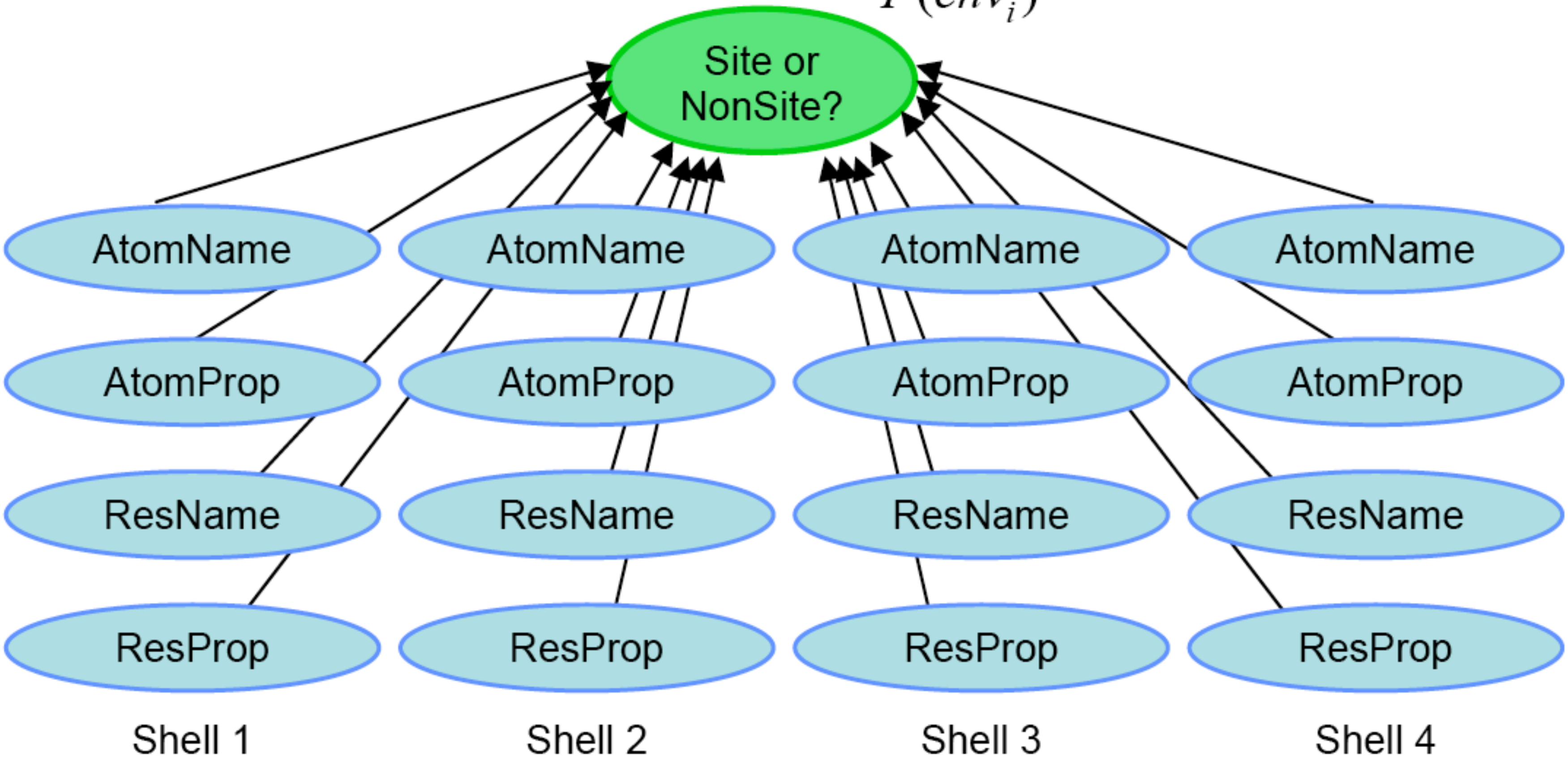


| Calcium Model<br>Property Name | | Volume 0 | Volume 1 | Volume 2 | Volume 3 | Volume 4 | Volume 5 |
|---|---|---|---|---|---|---|---|
| Atom Name | ATOM-NAME-IS-ANY | 🟥 | 🟩 | | 🟩 | 🟩 | |
| | ATOM-NAME-IS-C | 🟥 | 🟥 | | 🟩 | 🟩 | |
| | ATOM-NAME-IS-N | | 🟥 | 🟥 | 🟩 | | |
| | ATOM-NAME-IS-O | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| | ATOM-NAME-IS-S | | | | | | |
| | ATOM-NAME-IS-OTHER | | | | | | |
| Chemical Group | HYDROXYL | | | | | | |
| | AMIDE | | 🟥 | 🟥 | 🟩 | | |
| | AMINE | | | 🟥 | | | |
| | CARBONYL | 🟥 | 🟩 | | | 🟩 | 🟩 |
| | RING-SYSTEM | | | 🟥 | 🟥 | | |
| | PEPTIDE | 🟥 | | | 🟩 | 🟩 | |
| Atomic Properties | VDW-VOLUME | 🟥 | 🟥 | | 🟩 | 🟩 | |
| | CHARGE | | 🟩 | 🟩 | 🟩 | | |
| | NEG-CHARGE | | 🟩 | 🟩 | 🟩 | | |
| | POS-CHARGE | | | | | | |
| | CHARGE-WITH-HIS | | 🟩 | 🟩 | 🟩 | | |
| | HYDROPHOBICITY | | 🟥 | 🟥 | 🟥 | | 🟥 |
| | MOBILITY | 🟥 | 🟩 | 🟩 | | | |
| | SOLVENT-ACCESSIBILITY | 🟥 | | | | | |
| Residue Name | RESIDUE_NAME_IS_ALA | | | | | | |
| | RESIDUE_NAME_IS_ARG | | | | | | |
| | RESIDUE_NAME_IS_ASN | | 🟩 | 🟩 | 🟩 | 🟩 | |
| | RESIDUE_NAME_IS_ASP | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| | RESIDUE_NAME_IS_CYS | | | | | | |
| | RESIDUE_NAME_IS_GLN | | | | | | |
| | RESIDUE_NAME_IS_GLU | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| | RESIDUE_NAME_IS_GLY | | | | 🟩 | 🟩 | |
| | RESIDUE_NAME_IS_HIS | | | | | | |
| | RESIDUE_NAME_IS_ILE | | | | 🟩 | | |
| | RESIDUE_NAME_IS_LEU | | | | | | |
| | RESIDUE_NAME_IS_LYS | | | 🟥 | | | |
| ... | RESIDUE_NAME_IS_MET | | | | | | |

Abundance of ASP amino acid

| | |
|---|---|
| 🟩 | Significant presence |
| 🟥 | Significant absence |

$$P(Site \mid env) = \prod \frac{P(env_i \mid site)}{P(env_i)} P(site)$$



Site or NonSite?

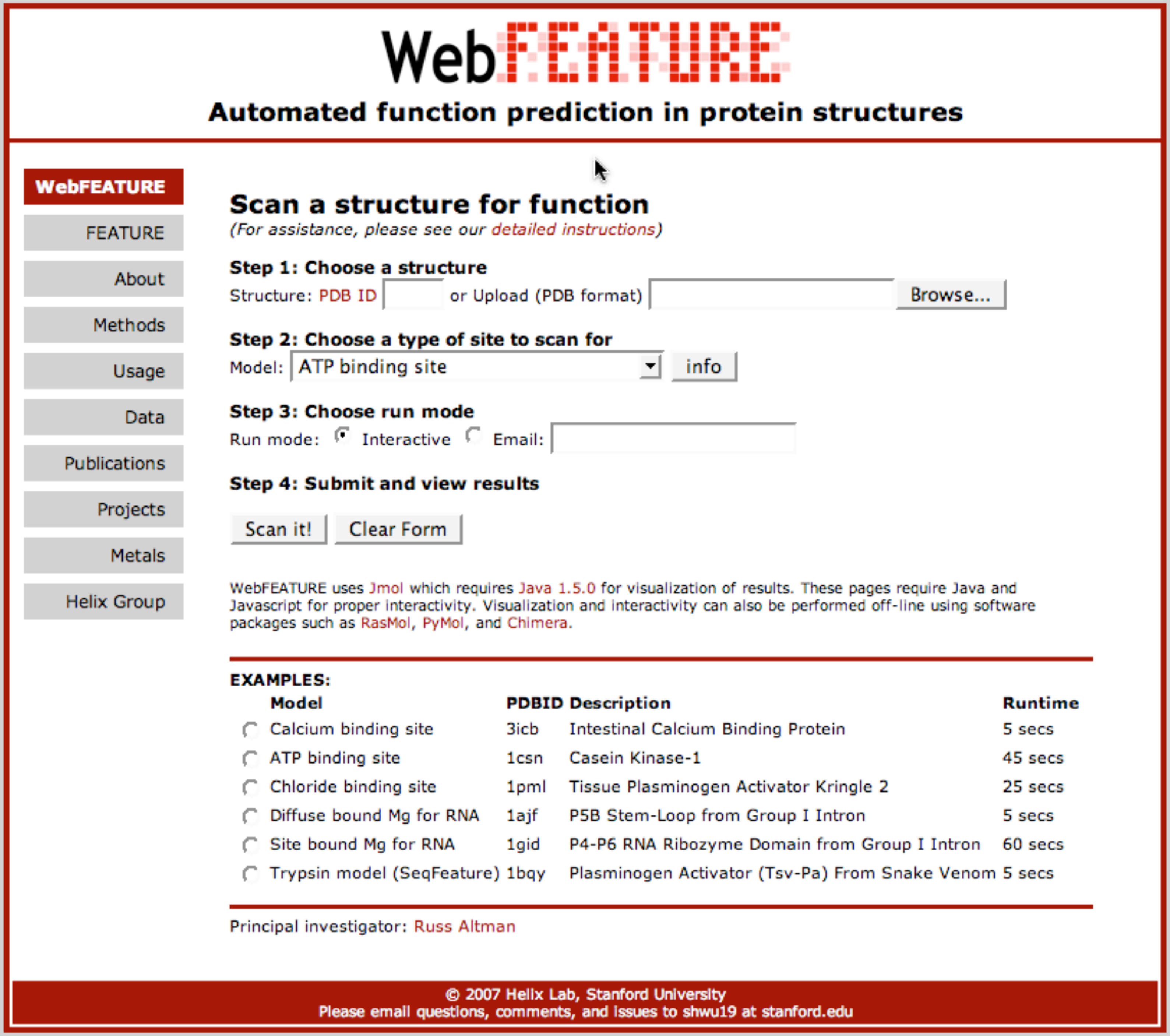
AtomName AtomName AtomName AtomName

AtomProp AtomProp AtomProp AtomProp

ResName ResName ResName ResName

ResProp ResProp ResProp ResProp

Shell 1 Shell 2 Shell 3 Shell 4



# WebFEATURE

**Automated function prediction in protein structures**

## Scan a structure for function

*(For assistance, please see our detailed instructions)*

**Step 1: Choose a structure**

Structure: PDB ID ____ or Upload (PDB format) ____

**Step 2: Choose a type of site to scan for**

Model: ATP binding site

**Step 3: Choose run mode**

Run mode: ● Interactive ○ Email: ____

**Step 4: Submit and view results**

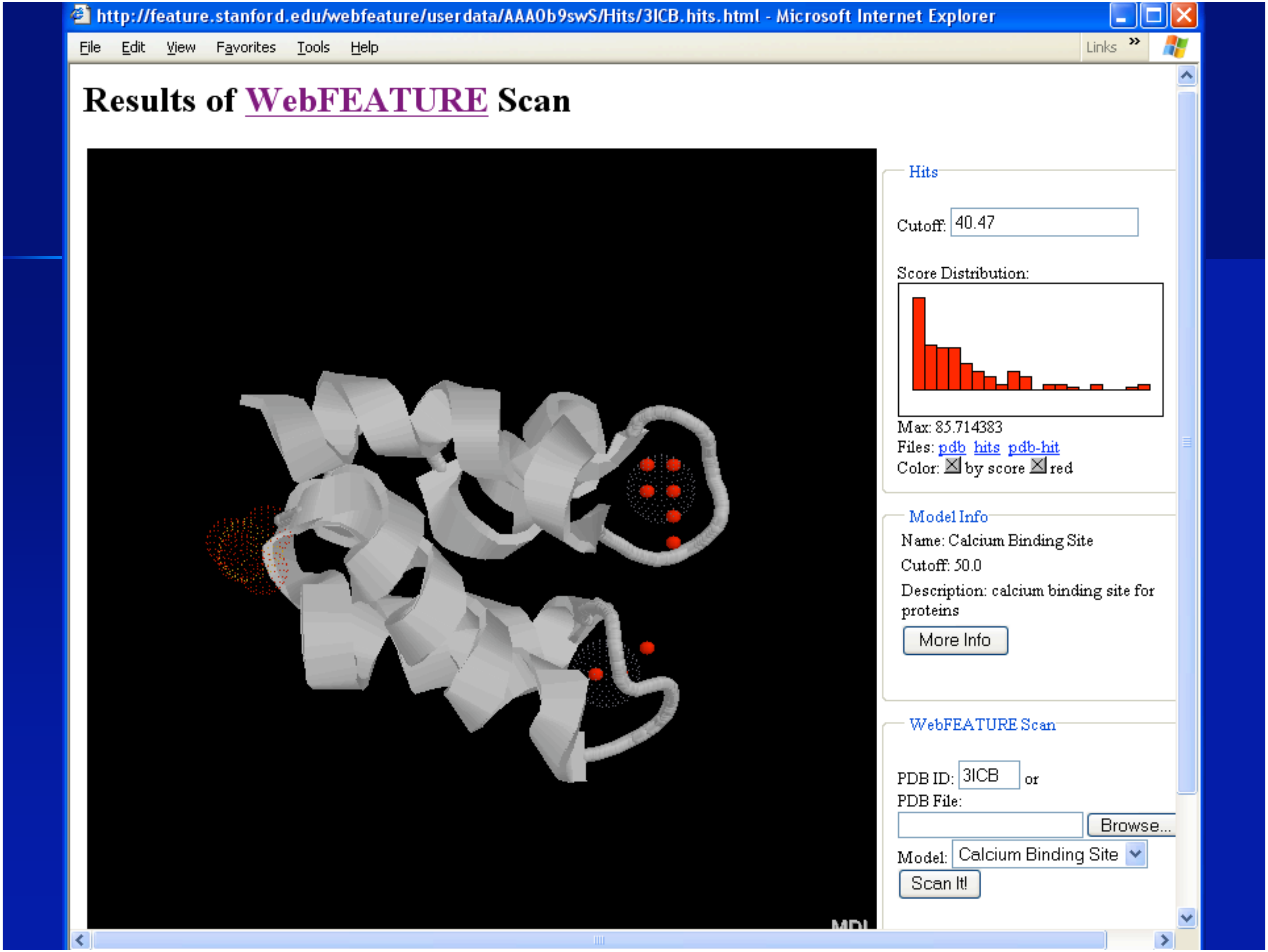
WebFEATURE uses Jmol which requires Java 1.5.0 for visualization of results. These pages require Java and Javascript for proper interactivity. Visualization and interactivity can also be performed off-line using software packages such as RasMol, PyMol, and Chimera.

**EXAMPLES:**

| | Model | PDBID | Description | Runtime |
|---|---|---|---|---|
| ○ | Calcium binding site | 3icb | Intestinal Calcium Binding Protein | 5 secs |
| ○ | ATP binding site | 1csn | Casein Kinase-1 | 45 secs |
| ○ | Chloride binding site | 1pml | Tissue Plasminogen Activator Kringle 2 | 25 secs |
| ○ | Diffuse bound Mg for RNA | 1ajf | P5B Stem-Loop from Group I Intron | 5 secs |
| ○ | Site bound Mg for RNA | 1gid | P4-P6 RNA Ribozyme Domain from Group I Intron | 60 secs |
| ○ | Trypsin model (SeqFeature) | 1bqy | Plasminogen Activator (Tsv-Pa) From Snake Venom | 5 secs |

Principal investigator: Russ Altman

Please email questions, comments, and issues to shwu19 at stanford.edu



http://feature.stanford.edu/webfeature/userdata/AAA0b9swS/Hits/3ICB.hits.html - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Links
MDL

# Results of WebFEATURE Scan

Hits

Cutoff: 40.47

Score Distribution:

Max: 85.714383

Files: pdb hits pdb-hit

Color: ☒ by score ☒ red

Model Info

Name: Calcium Binding Site

Cutoff: 50.0

Description: calcium binding site for proteins

WebFEATURE Scan

PDB ID: 3ICB or

PDB File:

Model: Calcium Binding Site

# Shirley Wu

**Challenge:**

**Create a library of models available to FEATURE**

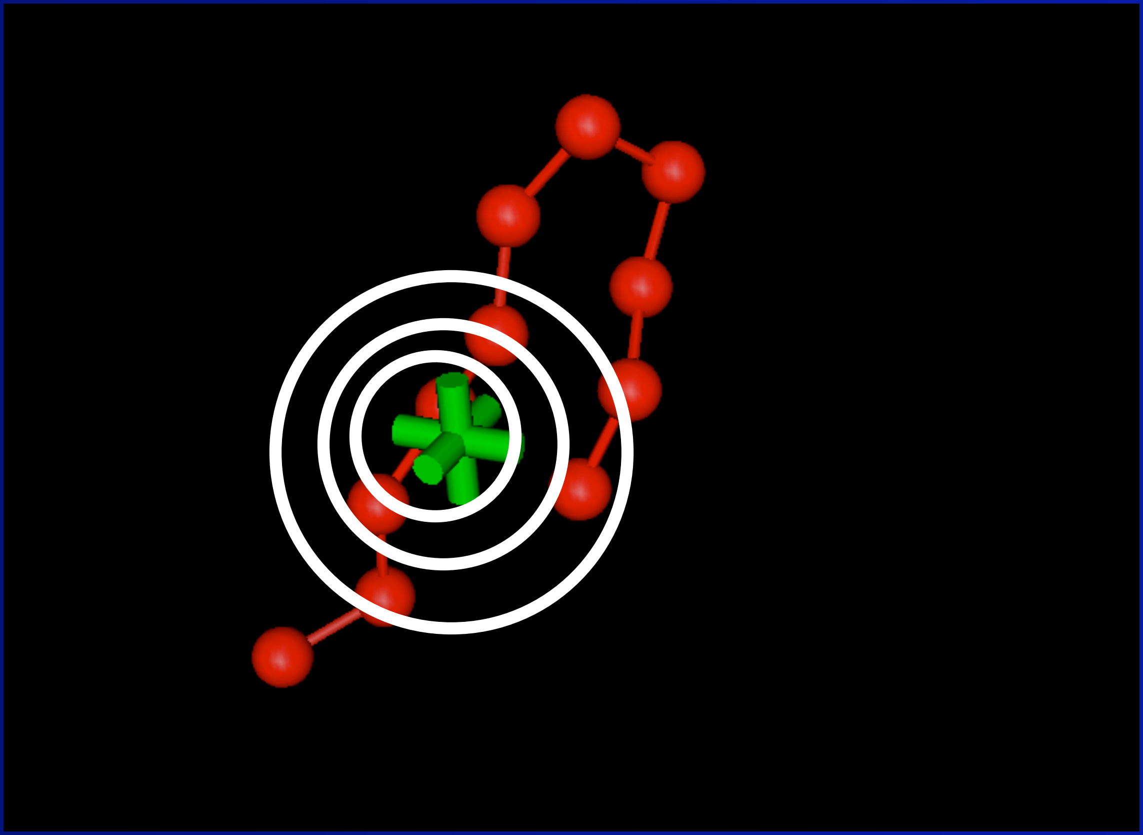
**Solution:**

**Use 1D sequence motifs as “seeds” for 3D motifs**

# SeqFEATURE

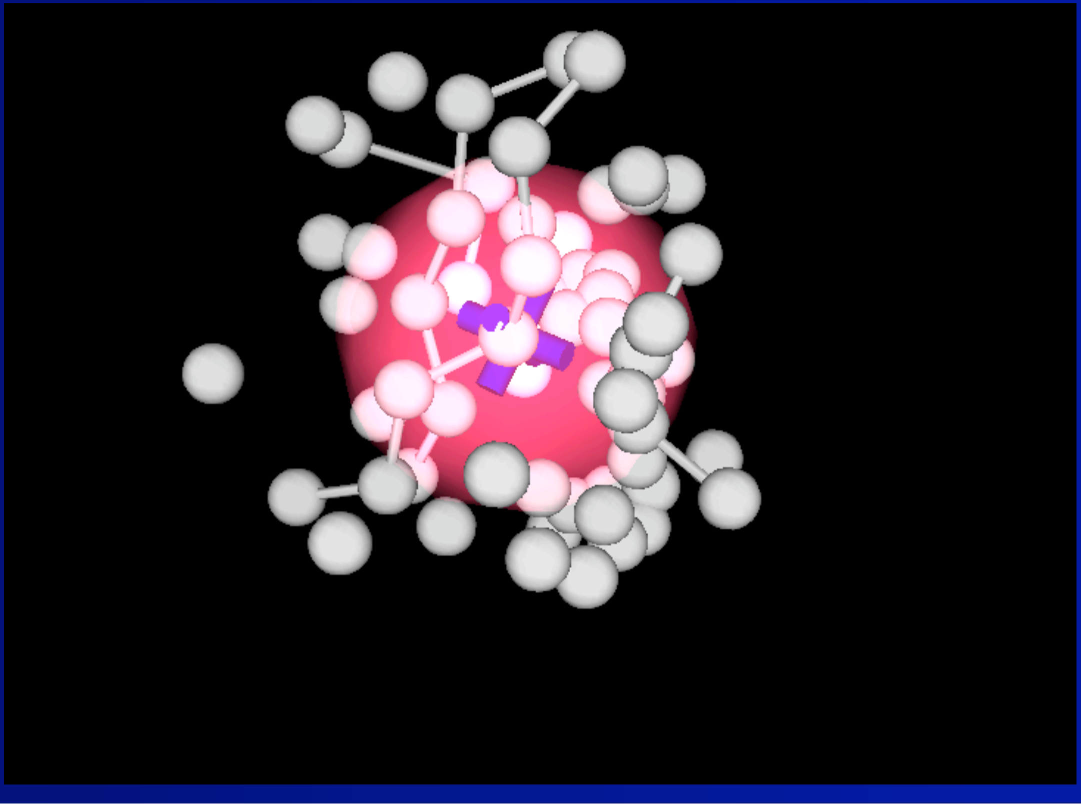
- Build from Sequence Motif Databases
- Automatically creates 3D motifs from 1D sequence motifs
- Hypothesis: 3D motifs perform better than 1D motifs in identifying functional sites

# Extracting 3D motif examples from 1D motif



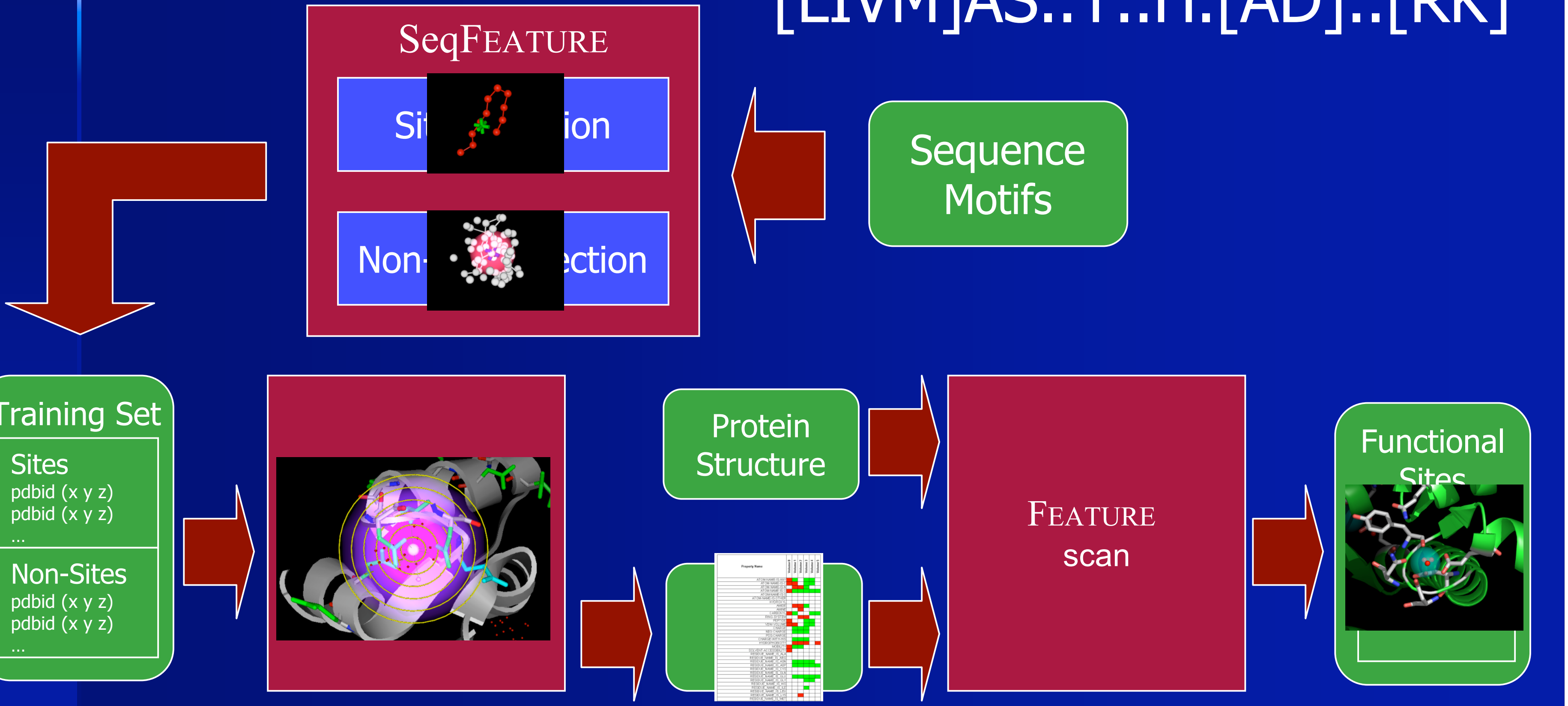
# Extracting non-sites for background statistics

- Random amino acid with similar atom density and outside the site area



# SeqFEATURE

[LIVM]AS..T..H.[AD]..[RK]



SeqFEATURE

Si ion

Non- ection

Sequence Motifs

Training Set

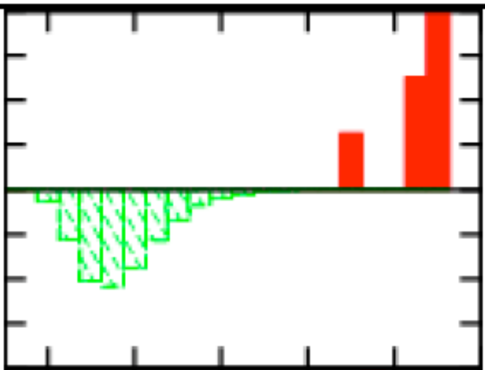
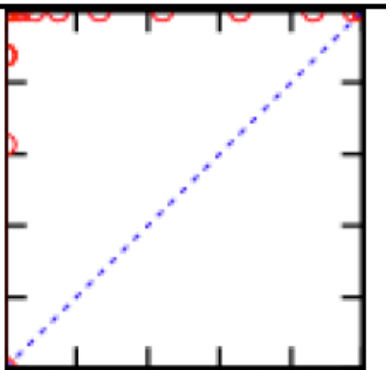
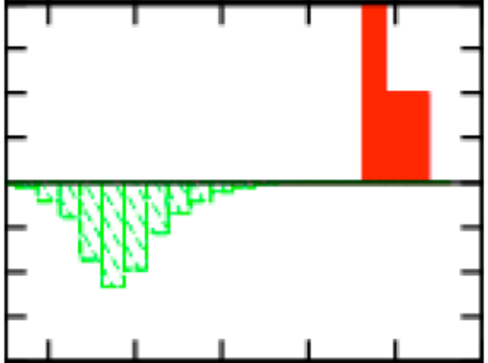
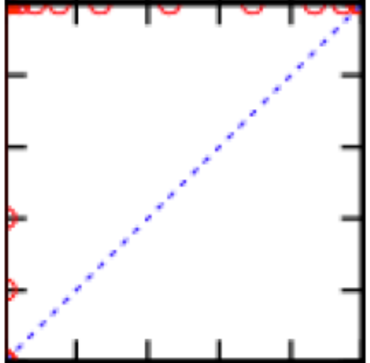
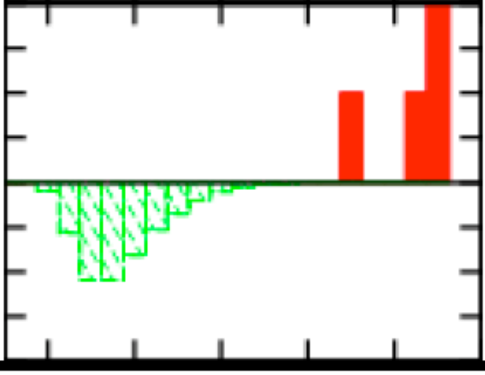
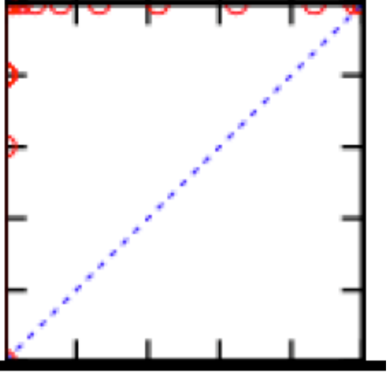
Sites
pdbid (x y z)
pdbid (x y z)
…

Non-Sites
pdbid (x y z)
pdbid (x y z)
…

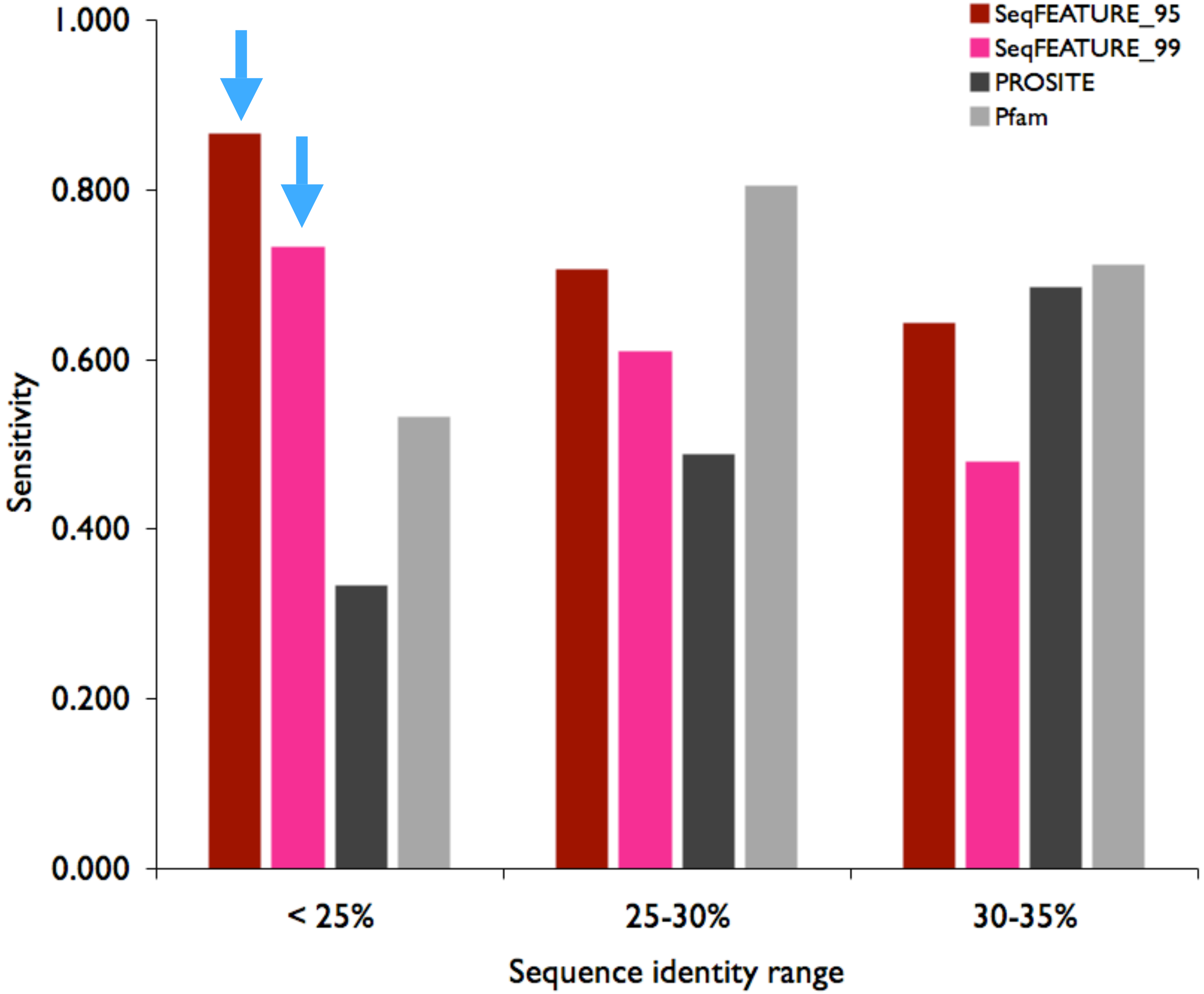
Protein Structure

FEATURE scan

Functional Sites

| *Model Performance* | *z-score Distribution* | *z-score ROC* |
|---|---|---|
| **Model:** TYR_PHOSPHATASE_1.3.CYS.SG<br>**AUC:** 1.0<br>**zAUC:** 1.0<br>**psAUC:** None<br>**psPPV:** 0.9824<br>**psSens:** 0.9076 |  |  |
| **Model:** RNASE_T2_1.4.HIS.NE2<br>**AUC:** 1.0<br>**zAUC:** 1.0<br>**psAUC:** None<br>**psPPV:** 1.0<br>**psSens:** 1.0 |  |  |
| **Model:** HIPIP.7.CYS.SG<br>**AUC:** 0.9999<br>**zAUC:** 1.0<br>**psAUC:** None<br>**psPPV:** 1.0<br>**psSens:** 1.0 |  |  |

# (seq)FEATURE performs better when sequence and structural similarity are low

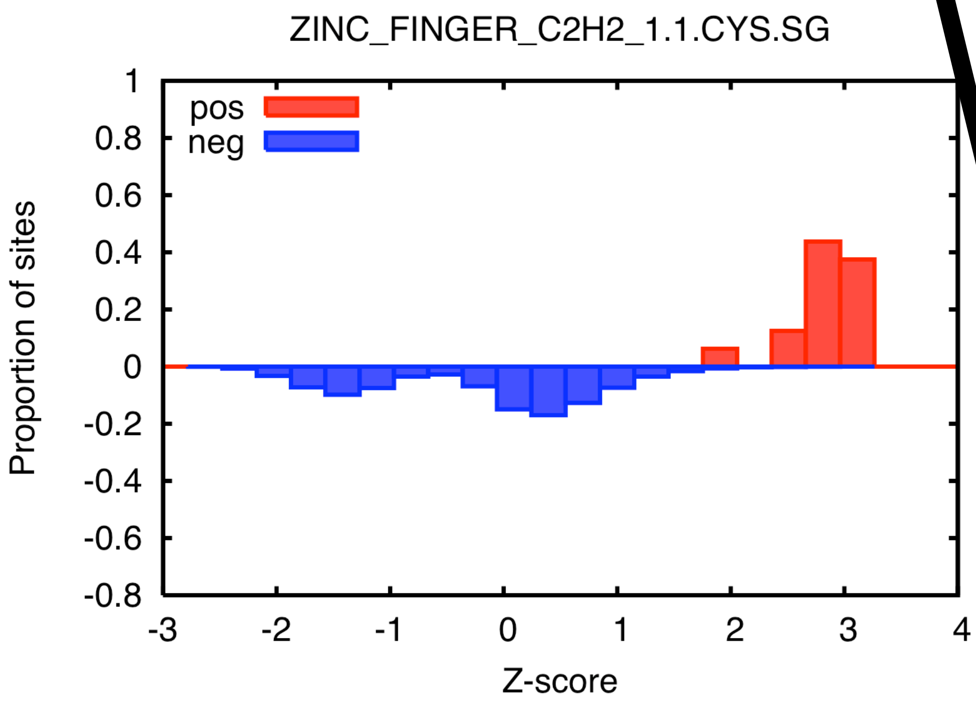


PROSITE

other

# 1Z84: Galt-like protein

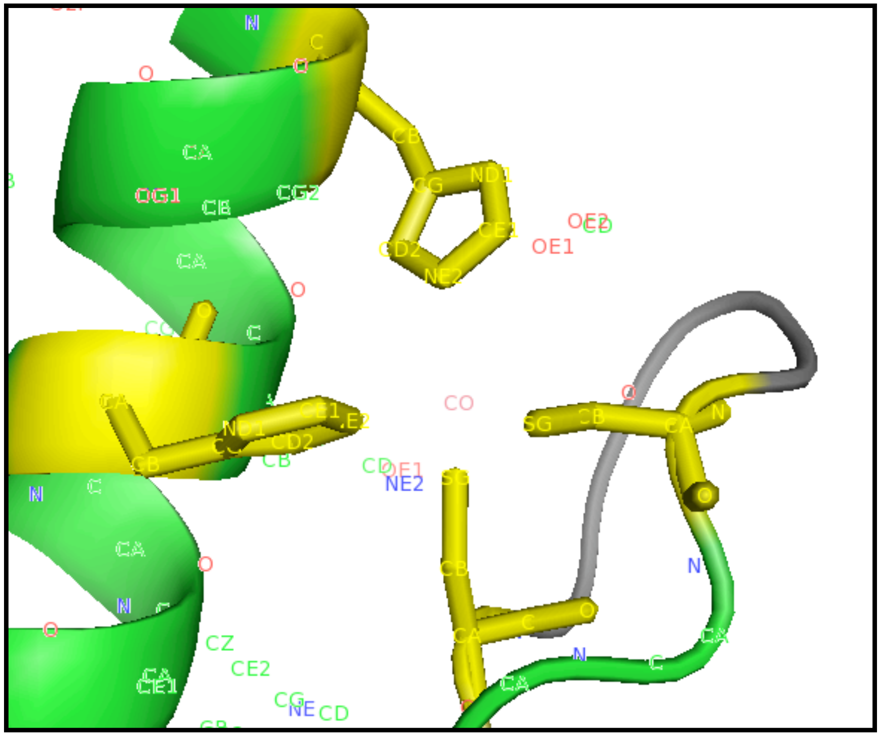
| SeqFEATURE Model | Site | z-score | Cutoff |
|---|---|---|---|
| ZINC_FINGER_C2H2_1.1.CYS.SG | CYS 63 | 4.713 | 3.177 |
| ZINC_FINGER_C2H2_1.3.CYS.SG | CYS 66 | 2.795 | 3.071 |



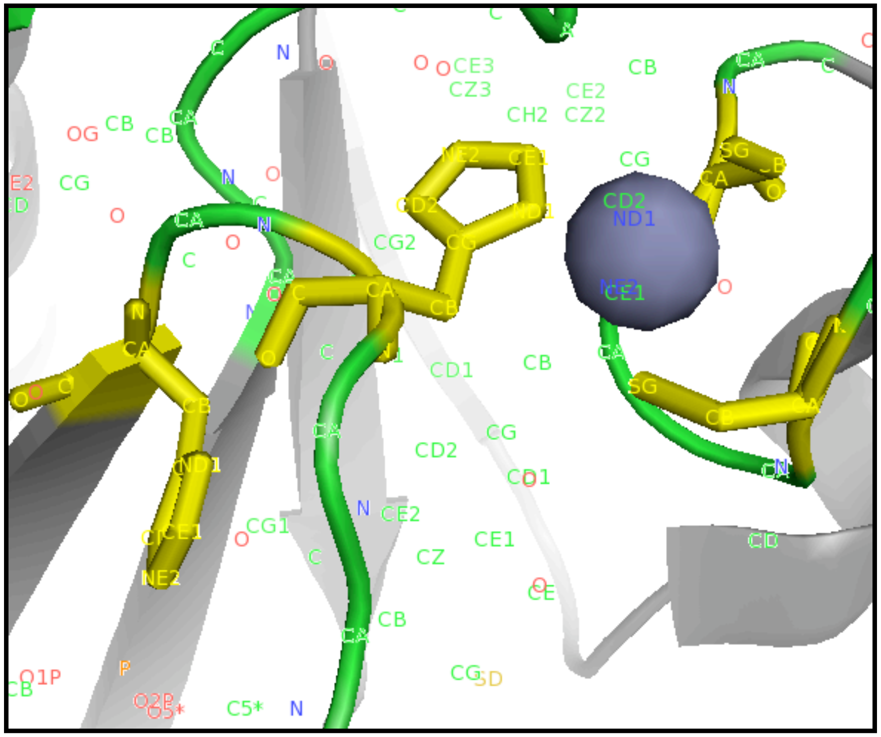
| Meth | |
|---|---|
| PROS | |
| Pfam | |
| SSM | ylase |
| 3-D te | |
| SeqFI | |

# 1Z84: Galt-like protein

2GLI
(zinc-finger protein)

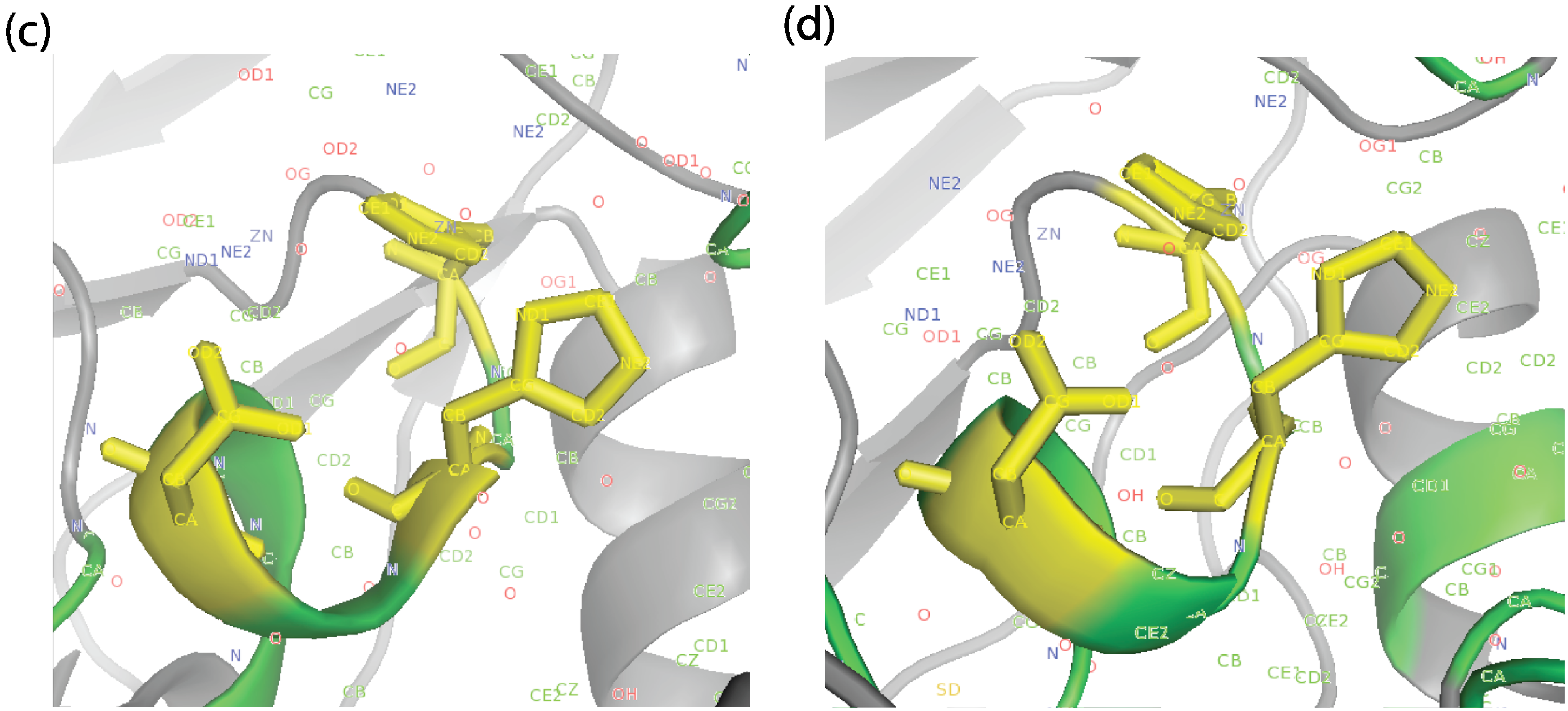


1Z84
(prediction)



back

# Beta-lactamase prediction (left)



# Go to other slides