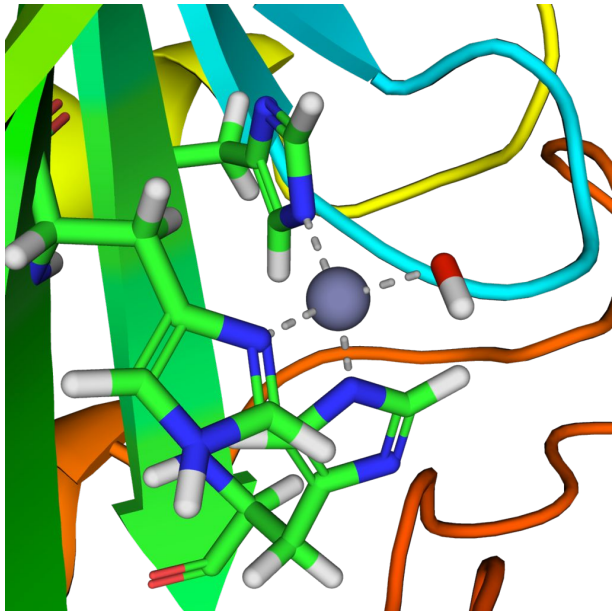
Scaffold Squad:
Scoring metal binding motifs in proteins

ChemE 546 winter 2024

Why design metalloproteins?

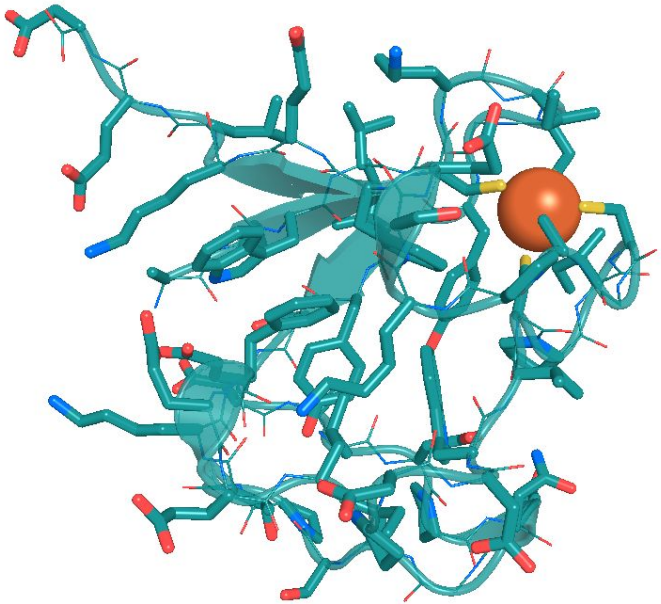
- In nature, metal binding proteins perform many functions
- Catalyze molecular reactions
- Transport electrons
- Storage and transport of other molecules, including oxygen
- Signal-transduction (e.g. transcription factors)



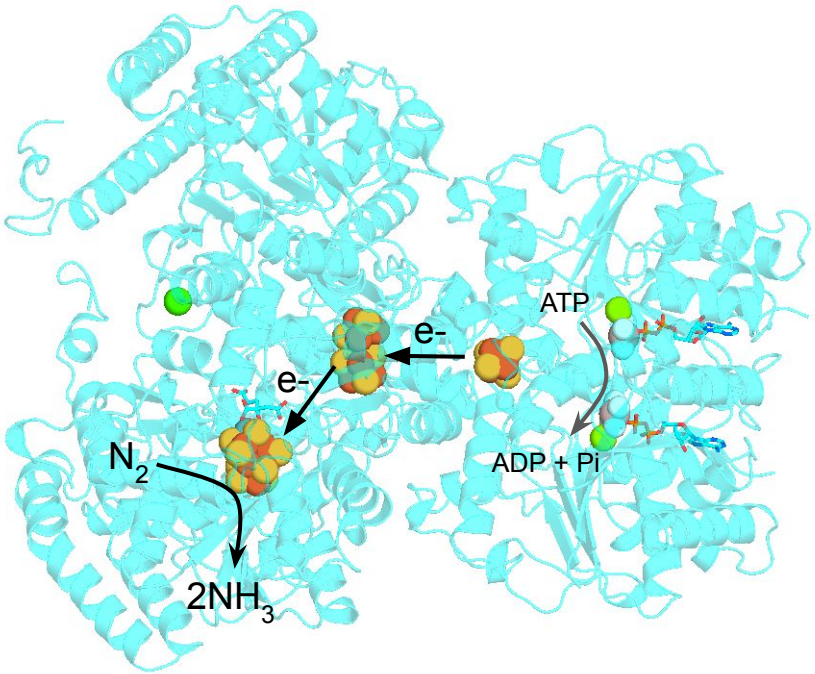
Carbonic anhydrase binds Zn to catalyze the reaction of CO_2 to carbonic acid (H_2CO_3)

Why design metalloproteins?

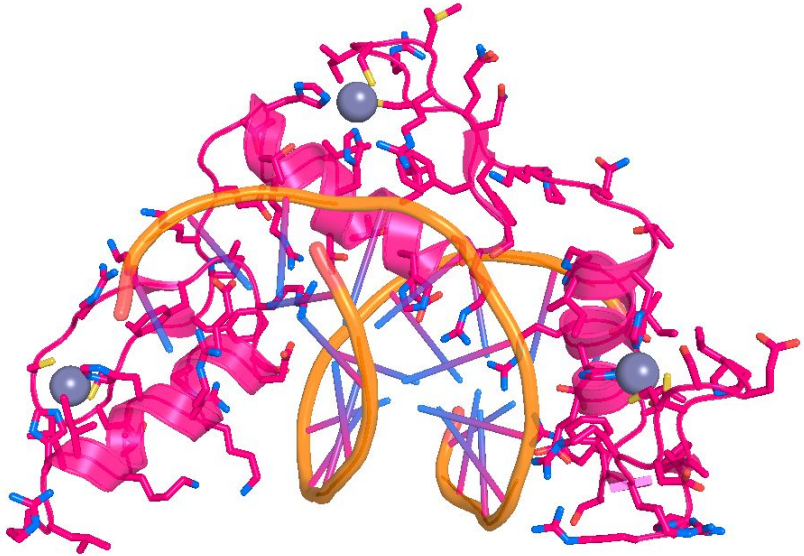
- In nature, metal binding proteins perform many functions
- Catalyze molecular reactions
- Transport electrons
- Storage and transport of other molecules, including oxygen
- Signal-transduction (e.g. transcription factors)



Rubredoxin binds iron to function as an electron transport protein



Nitrogenase enzyme contains metal centers that drive ATP hydrolysis, transfer electrons, and reduce N_2 to NH_3

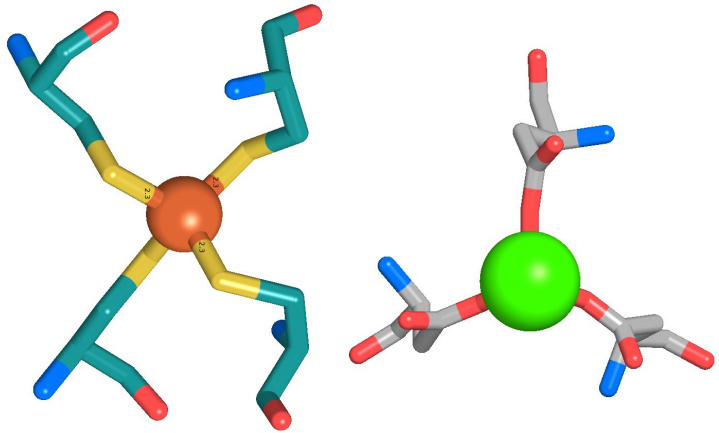


Zinc finger proteins are stabilized by the presence of Zn to form stable folds that bind DNA to modulate transcription

Current limitation: small “viable scaffold” space

- Currently, we only design and screen metalloproteins using “idealized” coordination geometries

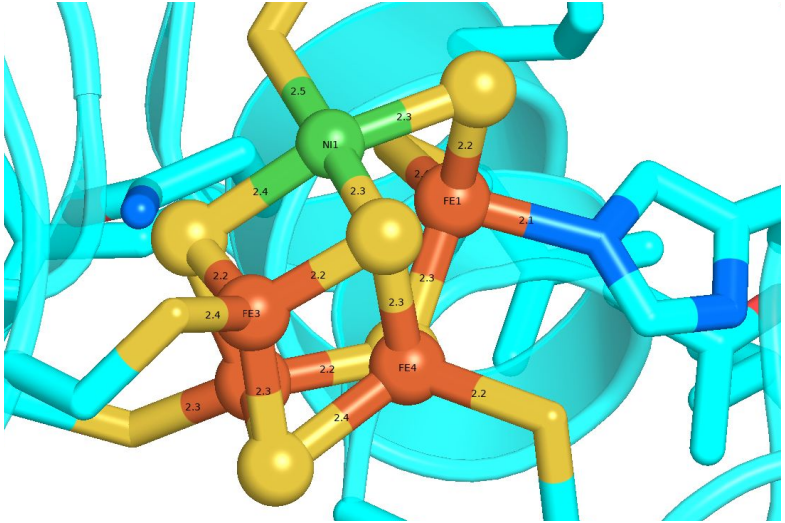
- Protein are not bricks! Natural proteins often bind metals with distorted, or “non-ideal” coordination, leading to dynamic structure and function



Protein design scaffolds use idealized, highly symmetric coordination geometries not (always) reflective of natural proteins

Current limitation: small “viable scaffold” space

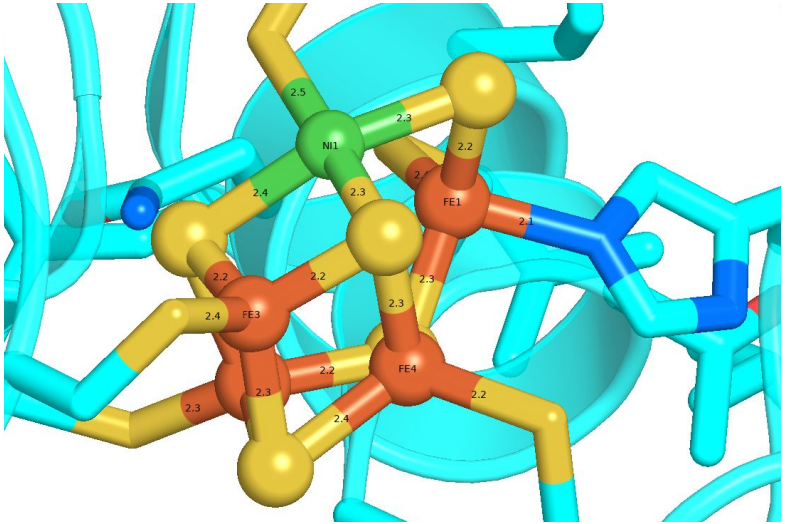
- Distorted coordination distance/geometry can tune ability of protein to **select for specific metal binding**, perform chemistry, transport metal, etc.



Carbon monoxide dehydrogenase binds metal centers with distorted geometry. This likely helps tune its catalytic and redox properties, as well as preference for Ni or Fe.

Current limitation: small “viable scaffold” space

- Learning from examples in nature could allow us to identify scaffolds and designs that 1) bind metal with specificity and 2) bind metal despite having “non-ideal” coordination geometry



Carbon monoxide dehydrogenase binds metal centers with distorted geometry. This likely helps tune its catalytic and redox properties, as well as preference for Ni or Fe.

The Project

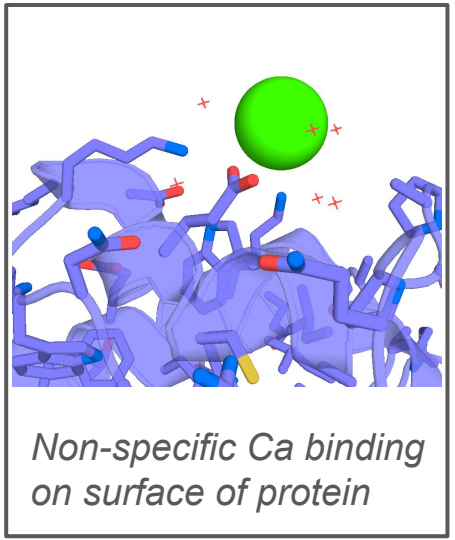
- Design software to predict metal binding ability of a given protein scaffold
- Input: file containing coordinates of metal and ligating atoms
- Output: probability that the scaffold binds that metal
- Target users: Baker Lab and larger metalloprotein design community

First step: where do we get our data?

- MetalPDB: database that contains information on 20,000+ protein-metal binding sites, including pdb accession codes, number and type of ligating residues, and coordination geometry
- PDB (Protein Data Bank): database containing atomic coordinates for over 200,000 proteins with cofactors
- Dataset with electronegativity and atomic radius of atoms (available on github)

Curating data set

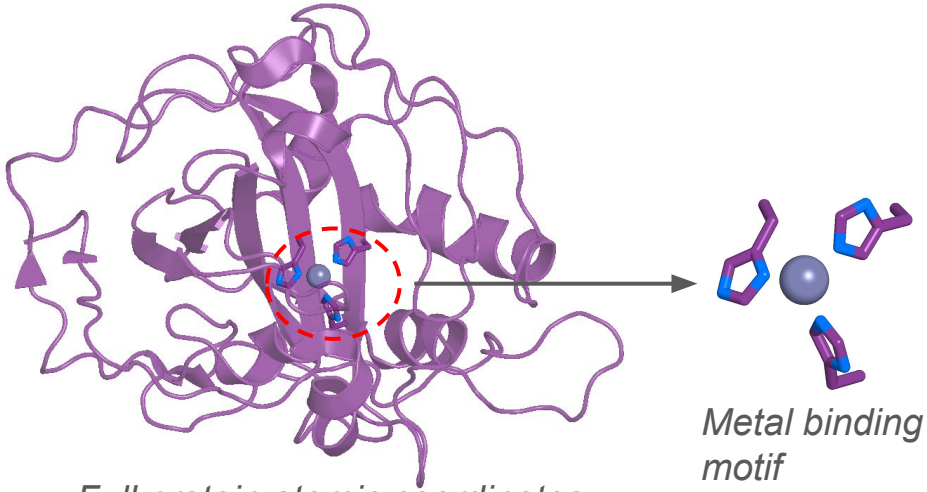
- Dataset was filled with redundancies, such as identical structures of homooligomers with identical sites
- Many examples of weak binding sites with just one ligating residue, or non-protein ligation
- Removed all redundancy from the dataset, as well as any sites with less than 2 amino acids ligating metal



Non-specific Ca binding on surface of protein

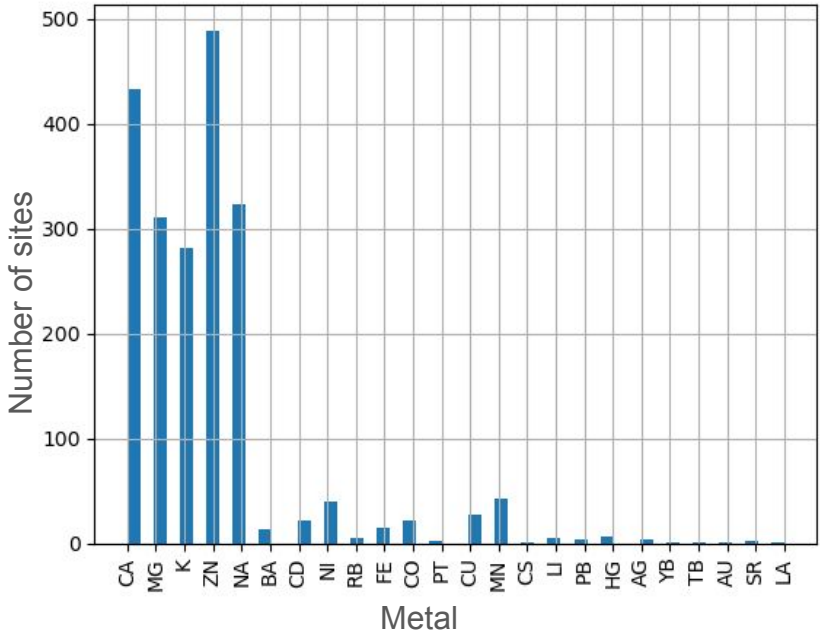
Extracting data features of molecular structures

- Downloaded PDB files corresponding to curated MetalPDB dataset
- Wrote code to identify metal binding motif and extract atomic coordinates of metal and ligating atoms
- Can then calculate data such as bond distances and geometry



Challenges/Limitations

- All transition metals (besides Zn) are underrepresented in the dataset
- Heavy metals and lanthanides have virtually no representation
- For now, Zn will be used as a stand in for other transition metals



Go about choosing training models

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

GODS: Generalized One-class Discriminative Subspaces for Anomaly Detection

Jue Wang*
Australian National University, Canberra
jue.wang@anu.edu.au

Anoop Cherian
Mitsubishi Electric Research Labs, Cambridge, MA
cherian@merl.com

Abstract

One-class learning is the classic problem of fitting a model to data for which annotations are available only for a single class. In this paper, we propose a novel objective for one-class learning. Our key idea is to use a pair of orthonormal frames – as subspaces – to "sandwich" the labeled data via optimizing for two objectives jointly: i) minimize the distance between the origins of the two subspaces, and ii) to maximize the margin between the hyperplanes and the data, either subspace demanding the data to be in its positive and negative orthant respectively. Our proposed objective however leads to a non-convex optimization problem, to which we resort to Riemannian optimization schemes and derive an efficient conjugate gradient scheme on the Stiefel manifold.

To study the effectiveness of our scheme, we propose a new dataset Dash-Cam-Pose, *consisting of clips with skeleton poses of humans seated in a car, the task being to classify the clips as normal or abnormal; the latter is when any human pose is out-of-position with regard to say an airbag deployment. Our experiments on the proposed Dash-Cam-Pose dataset, as well as several other stan-*

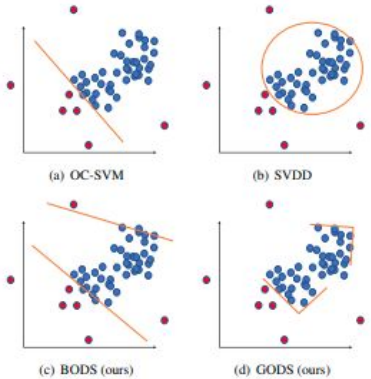
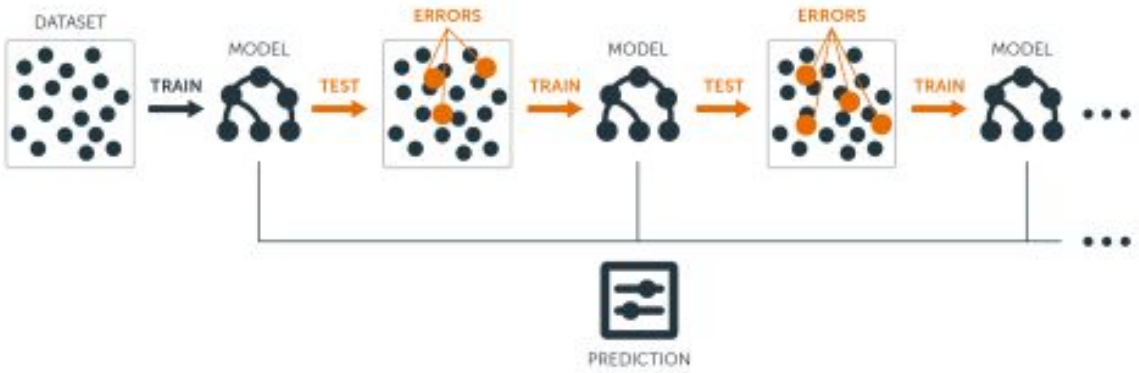


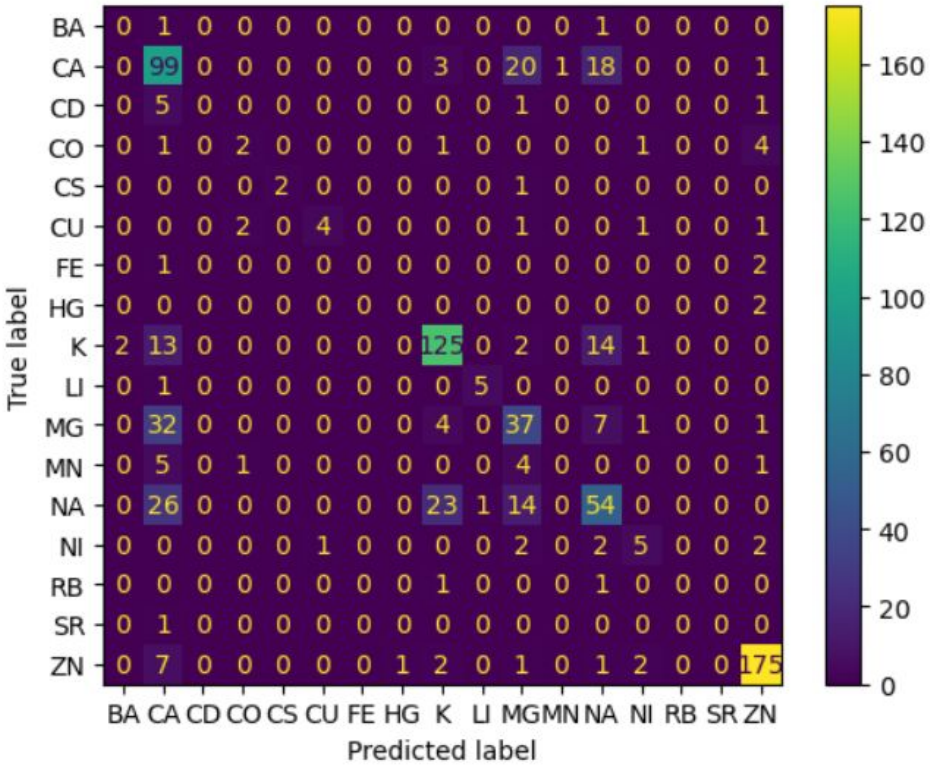
Figure 1. A graphical illustration of classical OC-SVM and SVDD in relation to our proposed BODS and GODS schemes. The blue points show the given one-class data, the red-points are outliers, and decision boundary of each method is shown by orange curves/lines.

Training Model: Amino Acids

- Utilized a gradient boosted classifier
- Test accuracy was 67%
- Input data: Identity of coordinating amino acids, number of coordinating amino acids, whether or not its a homodimer, and the dimer number
- Predicts which metal from the data set it believes will bind



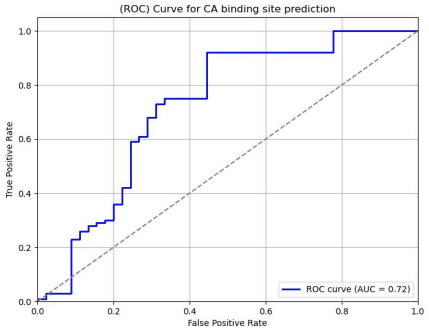
GBC:AA model results

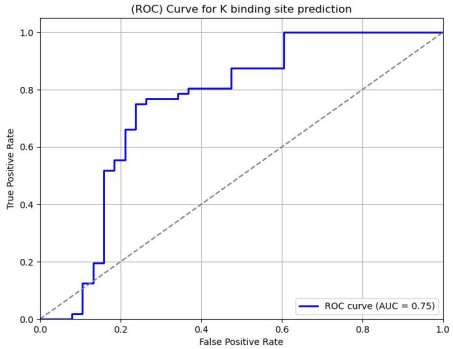


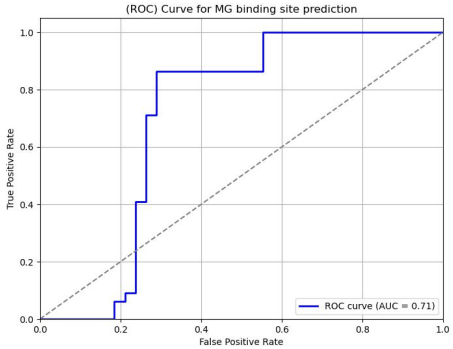
- Model has several non-idealities
  - There's artificial bias towards proteins that use more AAs to bind their target ion
  - Predictions are more unstable at lower AA coordinators
- Best at predicting binding to Zn^{2+} and K^{+}, poor at Ca^{2+} and Na^{+}
- Model is fairly computationally expensive to use
  - This could likely be ameliorated by hyperparameter tuning/pruning
  - Other classifiers also worked well (RF and DT were also tested), so model could be changed to save comp. time

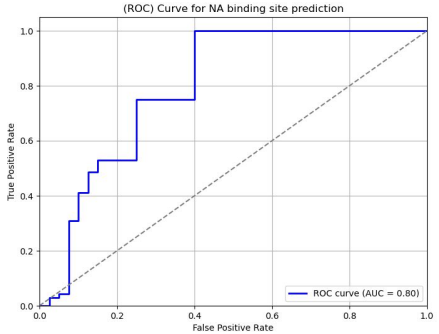
Training Model: Tip Atoms - One Class SVM

- Utilized metal - interactive atoms interaction
- Only expose to positive data (binding = 1)
- ROC curve look greats







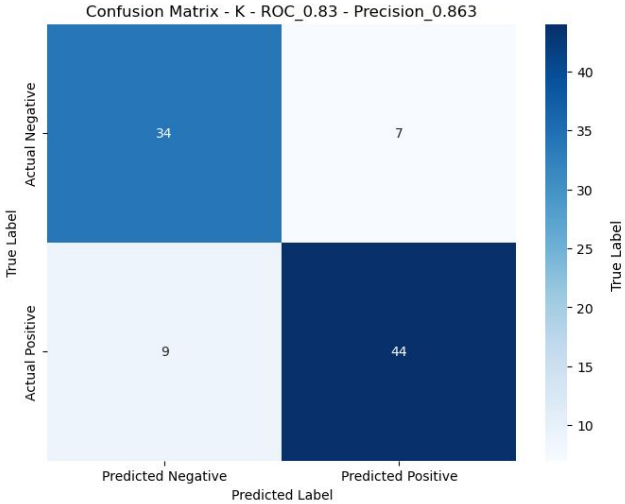


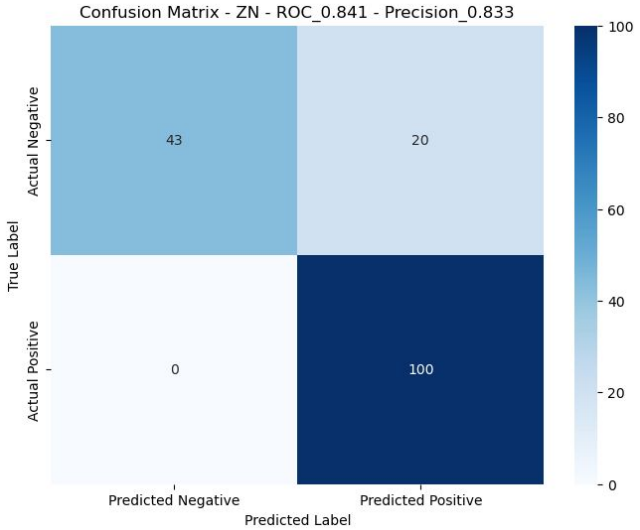
→ potential overfitting, only perform well on positive data in datasets.

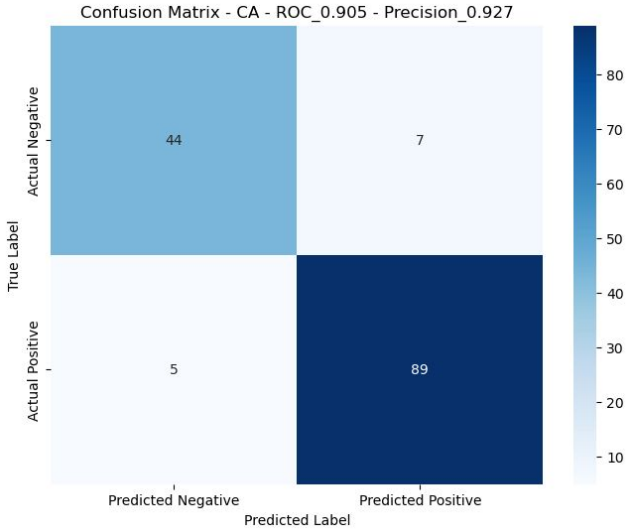
→ cannot replicate data on real pdb

Training Model: Tip Atoms - SVC

- Utilized metal - interactive atoms interaction
- Strategically noise positive data → get pseudo-negative data
- Train on both positive and negative data

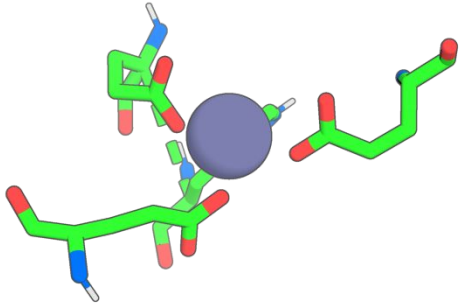






→ able to recapitulate prediction performance on real metal binding pdb

Demo - ZN binding motif



```
2  test_pdb = '/home/lhtran/class/scaffold_squad/test_cases/test_ZN_2.pdb'
3  pdb_data = util.preprocess_tip_atom(test_pdb, metal='ZN',metal_id=1)#, metal_id=15859)
4  pdb_data = np.reshape(pdb_data, (1, -1))
```

[16] ✓ 0.0s Python

```
1  model_ZN.predict(pdb_data), model_NA.predict(pdb_data)
```

[21] ✓ 0.0s Python

```
(array([1.]), array([-1.]))
```

→ please try out with your metal binding protein :)