# $K_{DEEP}$: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks

Jose Jimenez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis

Computational Biophysics Laboratory, Universitat Pompeu Fabra, Barcelona, Spain

Presenter: Eli Draizen

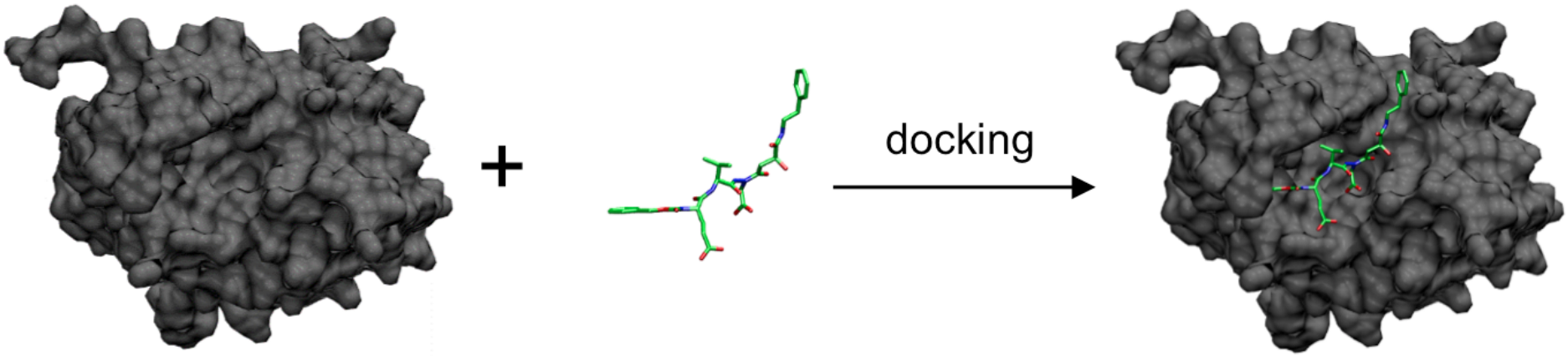
`https://qdata.github.io/deep2Read`

# Outline

# Introduction

## Computational Drug Design

- How will small molecules (drugs) bind to a protein?
- Where will the drug bind to the protein? DeepSite
- How strong of an interaction will it be? This paper



# Introduction

## Structure Representations

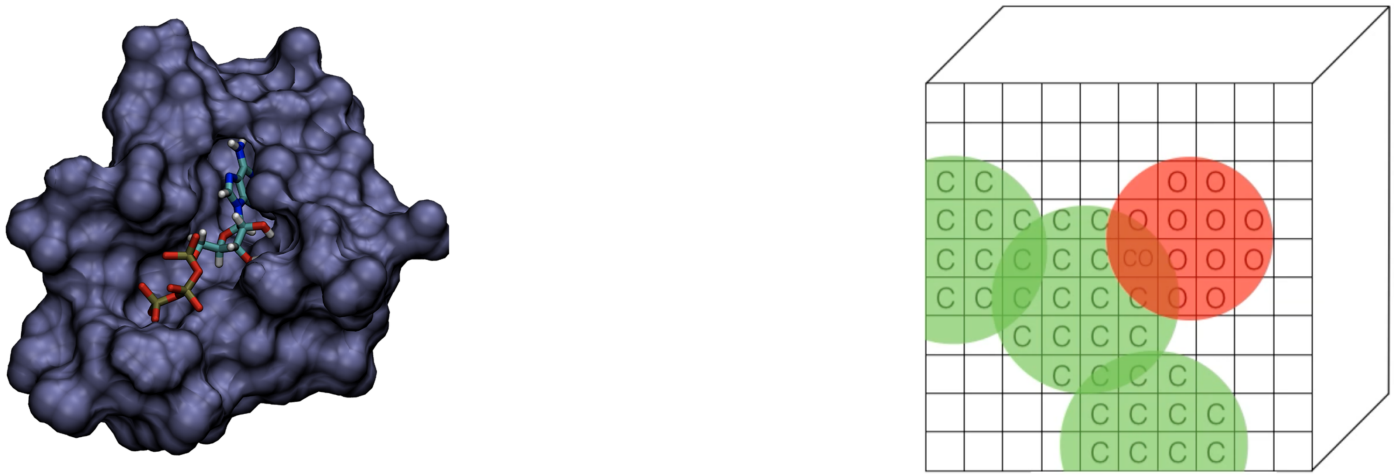


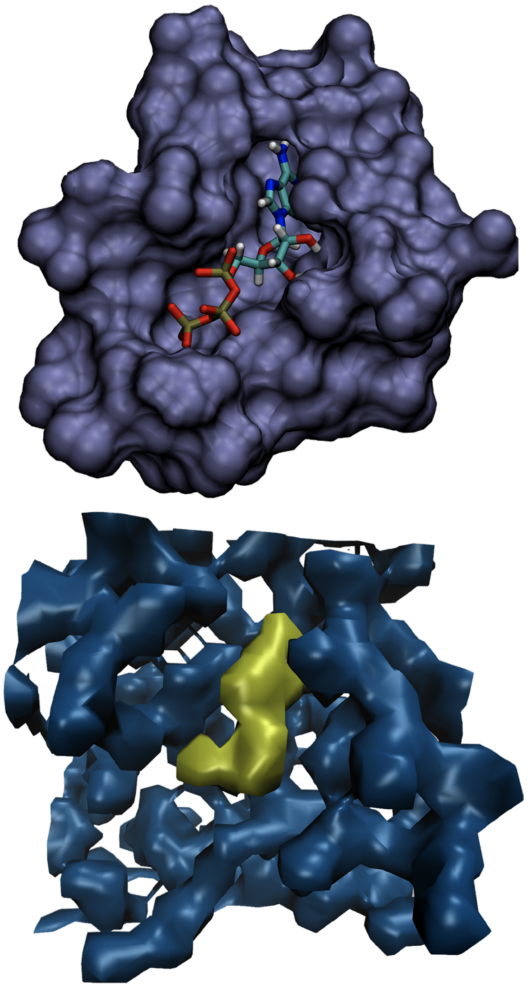
Figure: Left: Protein Structure. Right: Voxel Space (from David Koes)

# Method

## Voxelization: Lennard-Jones Potential

- Extract a 24$\AA$ Volume Around Known Binding Site
- Calculate the contribution, n, of every atom to every voxel in the 24$\AA$ Volume using the repulsive force for L-J Potential
- For every voxel, take the maximum feature value from all atoms that contribute to it

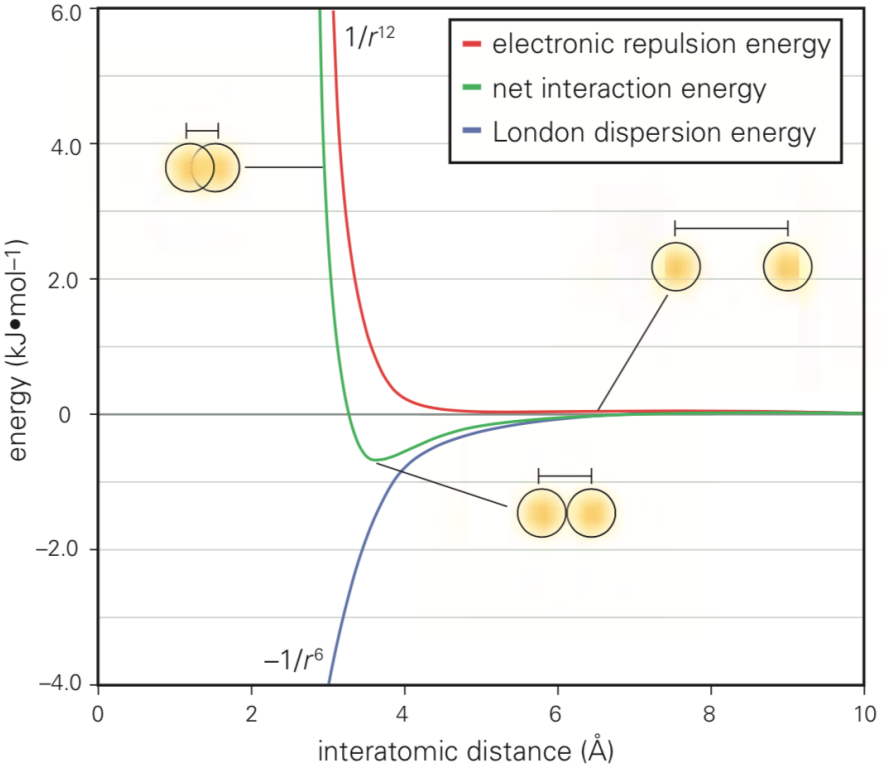
$$n(r) = 1 - exp\left( - \left( \frac{r_{vdw}}{r} \right)^{12} \right)$$



# Method

## Voxelization: Lennard-Jones Potential

The Lennard-Jones potential describes the energy potential between two non-bonded atoms based on their distance to each other.




$$V = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

V = intermolecular potential

$\sigma$ = distance where V is 0

r = distance between atoms, measured from one center to the other

$\epsilon$ = interaction strength

# Method

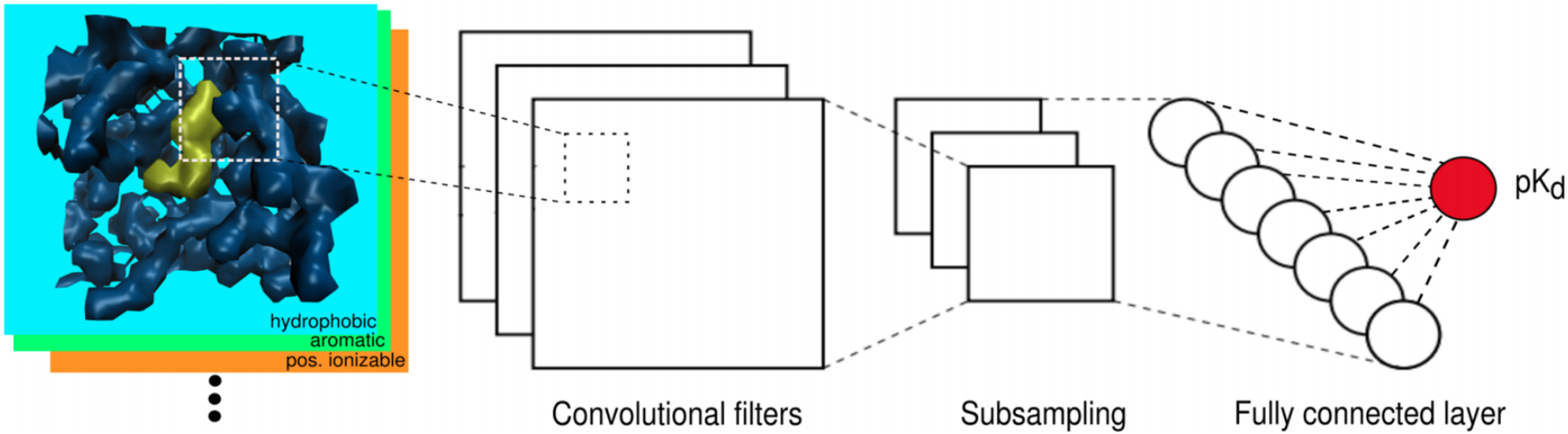
## Featurization

- 16 Features/Channels for every voxel
- 8 descriptors doubled for protein and ligand

| Property | Rule |
|---|---|
| Hydrophobic | atom type C or A |
| Aromatic | atom type A |
| Hydrogen bond acceptor | atom type NA or NS or OA or OS or SA |
| Hydrogen bond donor | atom type HD or HS with O or N partner |
| Positive ionizable | atom with positive charge |
| Negative ionizable | atom with negative charge |
| Metal | atom type MG or ZN or MN or CA or FE |
| Excluded volume | all atom types |

# Method

## 3D-CNN



# Method

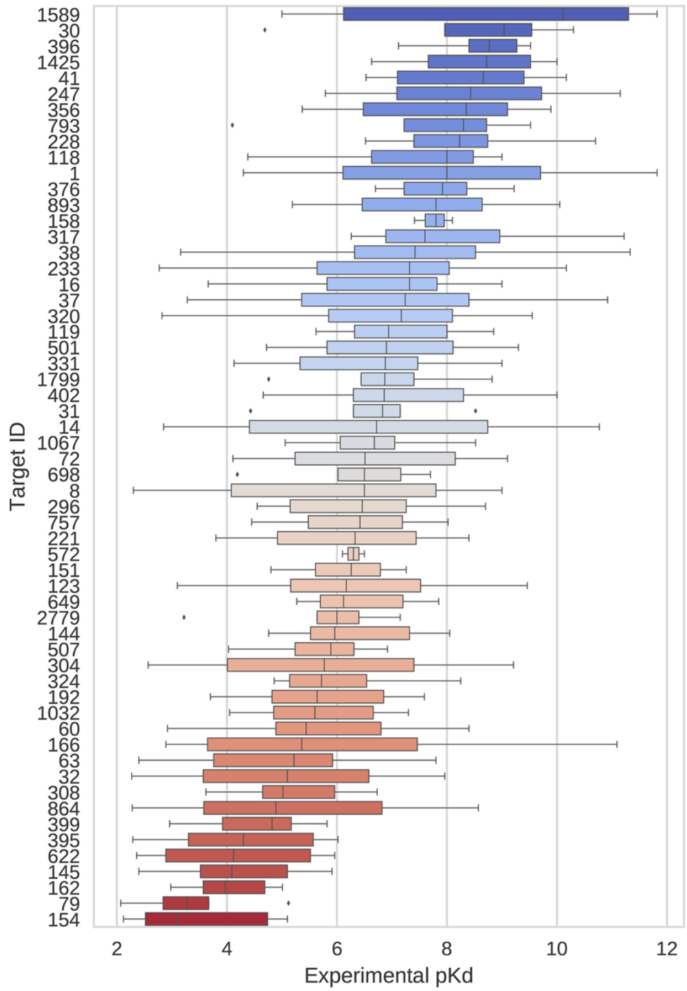
## Data Sources

- **PDBbind (v.2016)**, (n=13,308) complexes with experimental:
  - dissociation constants (Kd)
  - inhibition constants (Ki)
  - half-concentration values (IC50)
- A smaller **refined subset** (nr = 4057) by exluding complexes with:
  - a resolution higher than 2.5 Å
  - an R-factor higher than 0.25
  - ligands bound through covalent bonds
  - ternary complexes or steric clashes
  - affinity not reported either in Kd or Ki
  - falling out of a desired range (Kd ¡ 1pM)
- **A core set** (nc = 290) is a representative non-redundant subset
- But there was no difference in performance: "We tested this by using the PDBbind full minus core set as training and saw no significant performance difference in any test set."

# Method

## Data Sources – Target Set

- **A target set** (n=58) of the core clustered using a 90% sequence similarity
  - average standard deviation of 1.77 pK units.



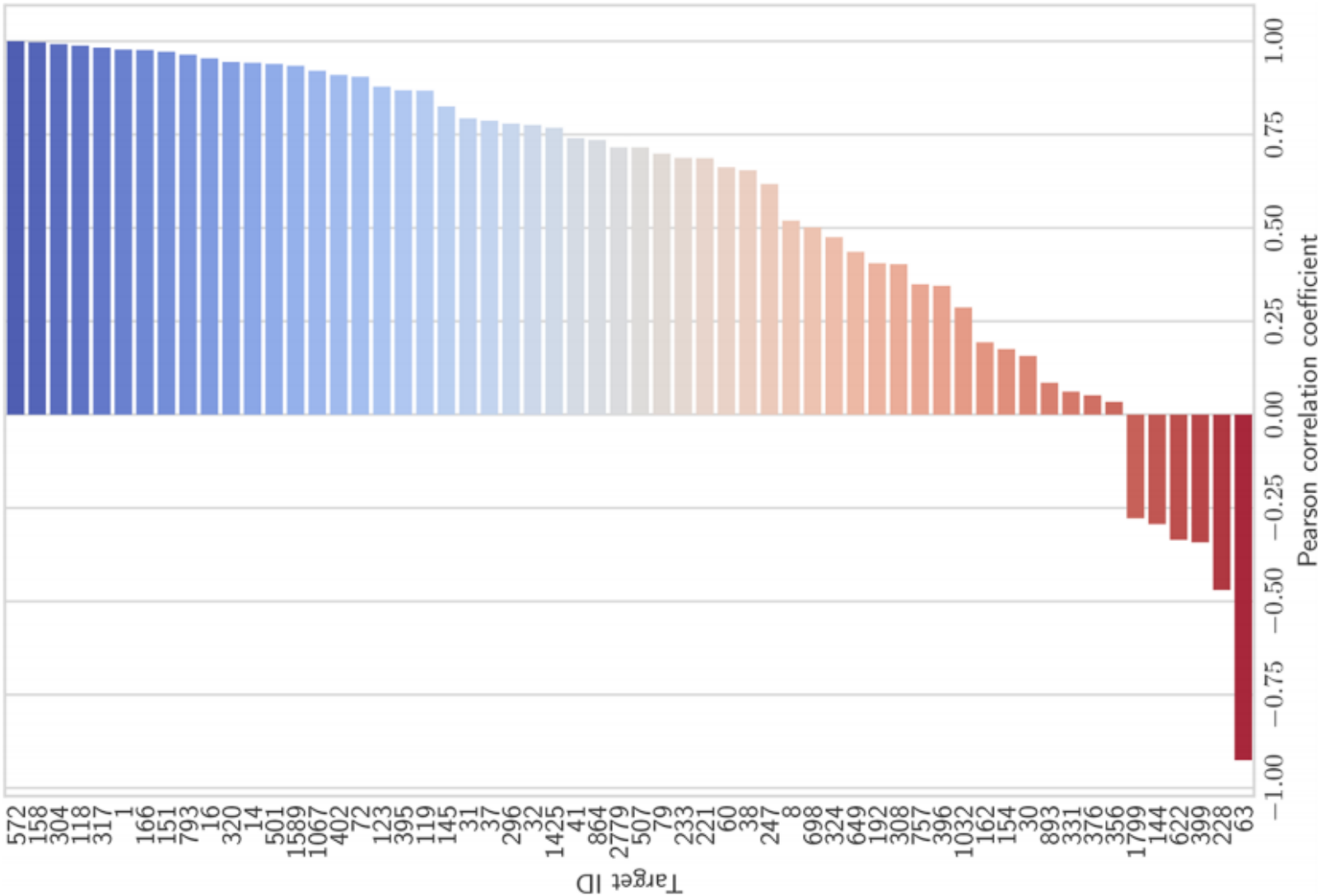
# Comparison

## Other Protein:Ligand Binding Affinity Predictors and Datasets

- Predictors
  - RF-Score: Random forest
  - X-Score: Linear Regression
  - Cy-Score: Linear Regression using a curvature feature
- Datasets
  - CSAR NRC-HiQ (4 data sets: 55, 55, 49, 49
  - CSAR2012 57 proteinligand complexes from the D3R challenge
  - CSAR2014, with 47 complexes from by the D3R consortium

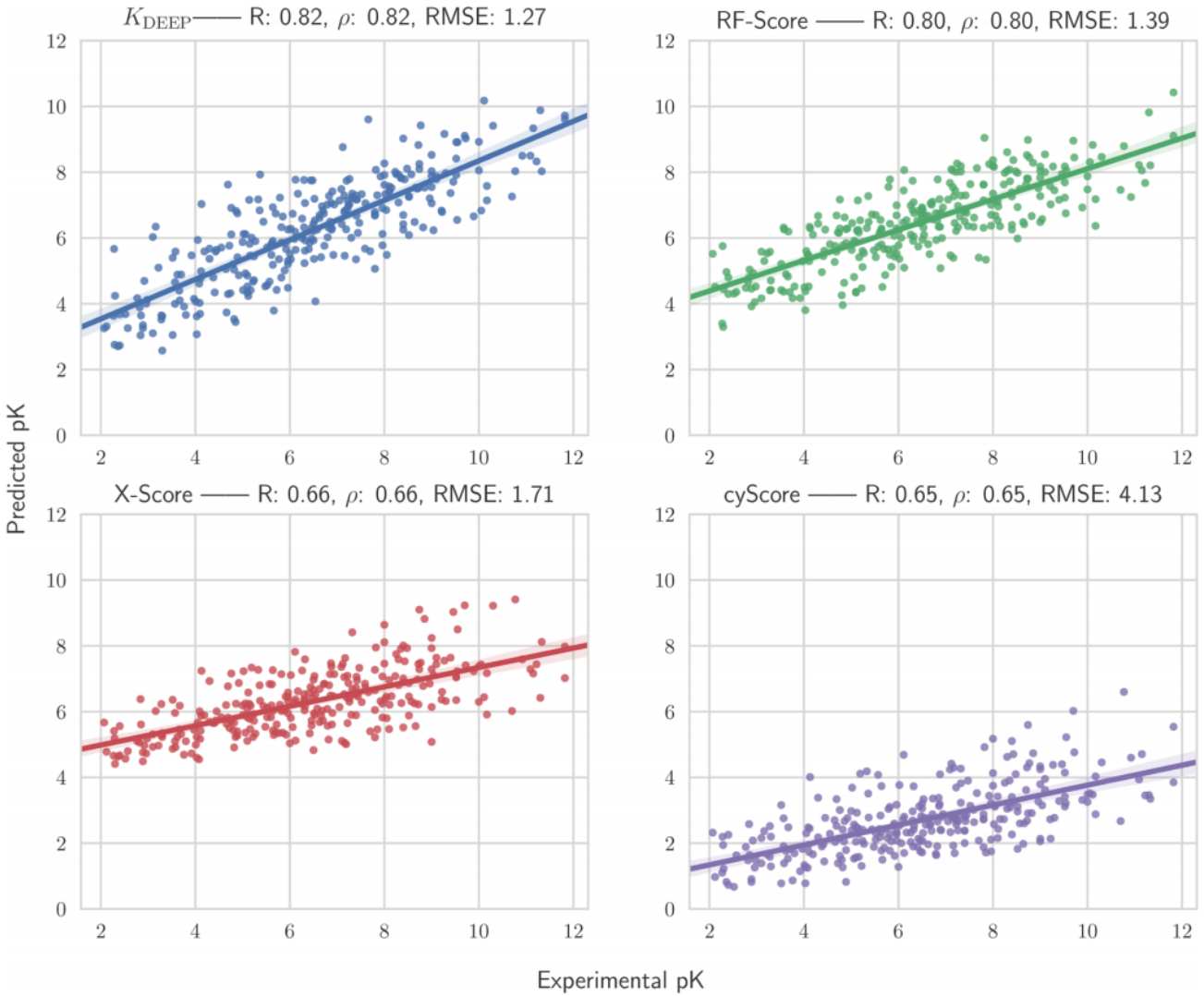
# Results

## Pearson Correlations coefficients for each target



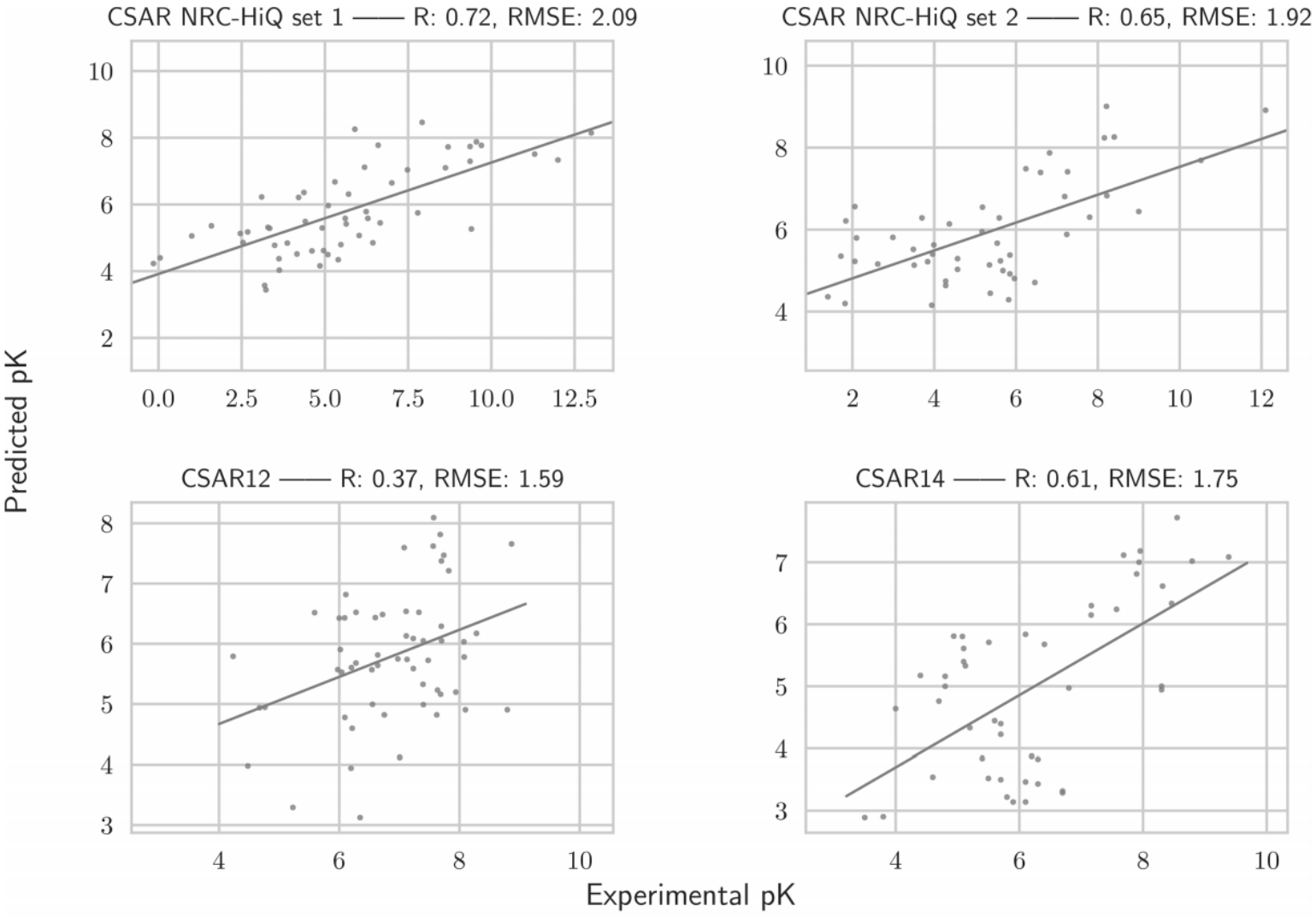
# Results (Core Set)

## Predicted vs. Truth



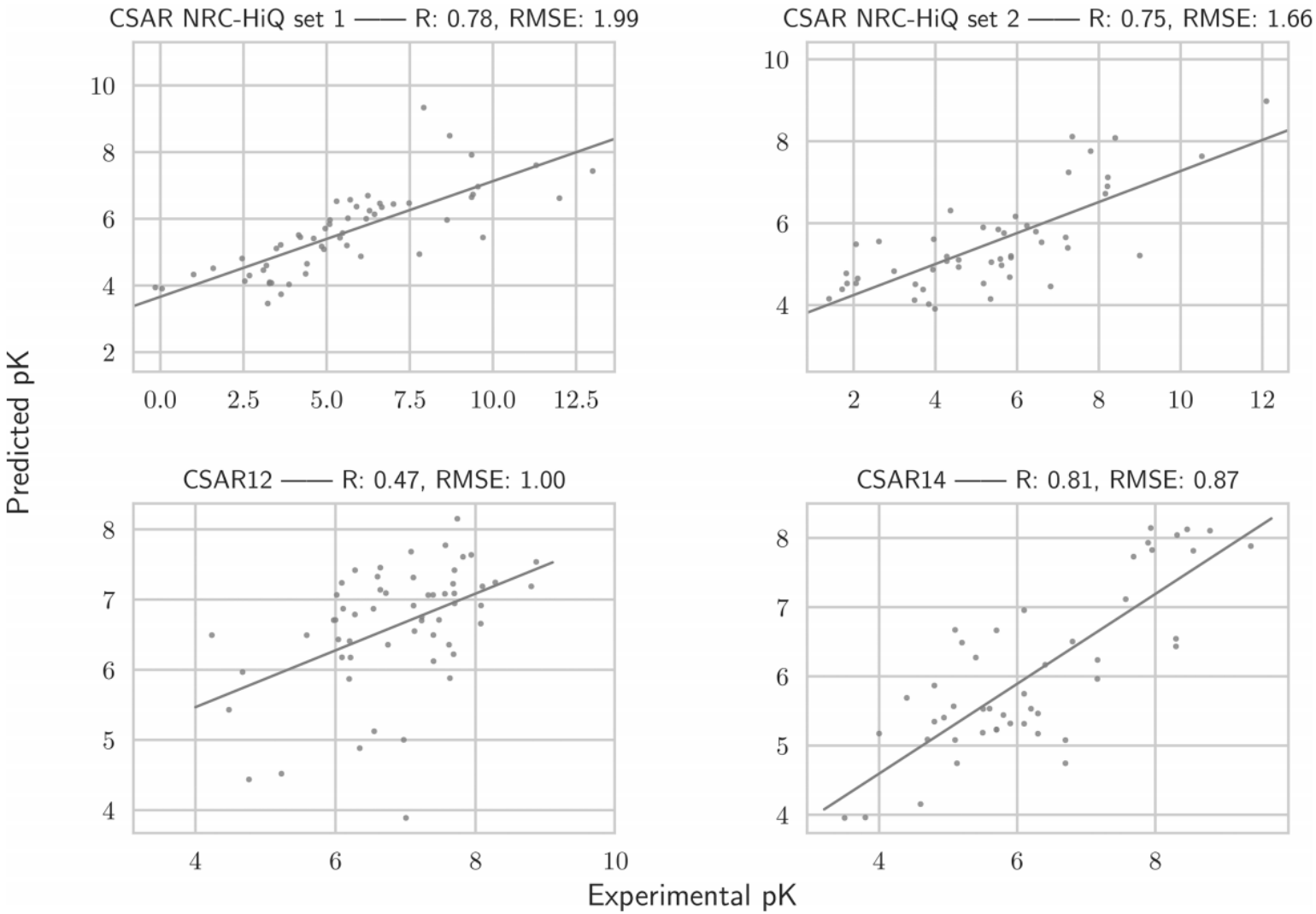
# Results (Core Set)

## KDEEP prediction for CSAR Datasets



# Results (Core Set)

## RF-Score prediction for CSAR Datasets



# Results (Core Set)

## Correlations for all programs against CSAR Datasets

**Table 3. Pearson's Correlation ($R$) between Experimental and Predicted Binding Affinity in Four CSAR Data Sets**

| | $K_{DEEP}$ | RF-Score | cyScore | X-Score | log $P$ | mol. weight |
|---|---|---|---|---|---|---|
| CSAR NRC-HiQ set 1 | 0.72[a] | 0.77[a] | 0.65[a] | 0.6[a] | 0.33 | 0.28 |
| CSAR NRC-HiQ set 2 | 0.65[a] | 0.75[a] | 0.64[a] | 0.65[a] | 0.44[a] | 0.44[a] |
| CSAR12 | 0.37[a] | 0.46[a] | 0.26 | 0.48[a] | 0.17 | 0.4[a] |
| CSAR14 | 0.61[a] | 0.8[a] | 0.67[a] | 0.82[a] | 0.22 | 0.82[a] |
| **Weighted average** | 0.58 | **0.69** | 0.55 | 0.63 | 0.29 | 0.47 |
| **Simple average** | 0.59 | **0.7** | 0.56 | 0.64 | 0.29 | 0.54 |

[a]Correlation significant at $\alpha = 0.01$.

# Conclusion

## Problems

- Input Volumes ($24\mathring{A}$) seems too small
  - It may ignore too much of the binding site and/or other important surface residues.
  - They say it is because of memory constraints, but do not discuss sparse 3DCNNS.
- L-J Voxelization method is confusing and will be slow for full proteins
- No alternate poses considered
- No evolutionary information was included
  - How do small changes around the binding site effect it?
- Small datasets
- Binding Site and Pose Predictions not incorporated nor re-added back into the training set (David Koes' research https://www.youtube.com/watch?v=jPg-4rkjZKc)

# Conclusion

## Summary

- They created a new binding affinity predictor using 3DCNNs on 24$\AA$ volumes
- $K_{DEEP}$ performs well and is competitive against other binding site predictors when compared just against PDBbind
- However, no method outperforms the others on the other datasets