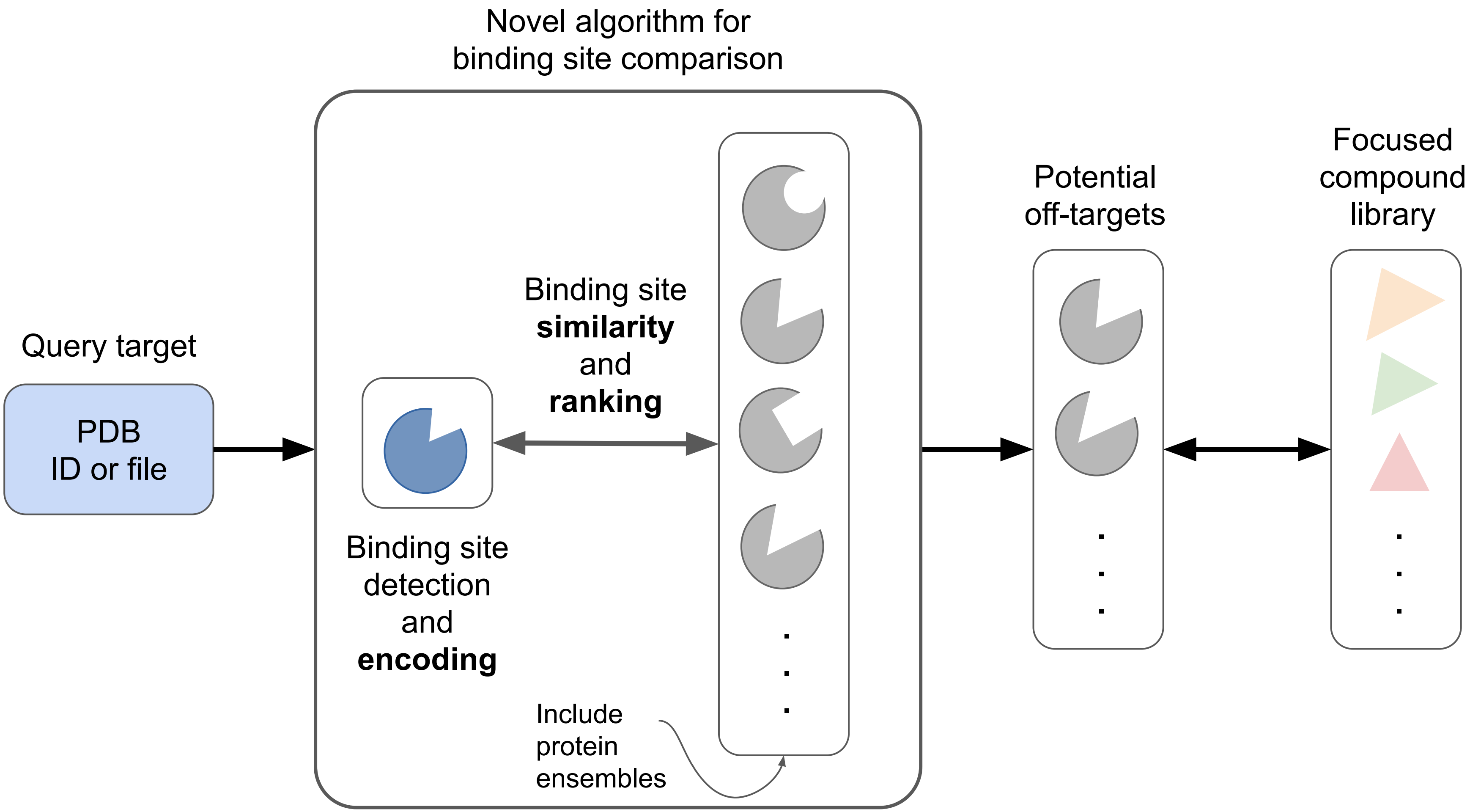
# Read-across the targetome

## Binding site comparison

Group meeting

Dominique Sydow
30.04.2019

# Intro: Read-across the targetome



Novel algorithm for binding site comparison

Query target

PDB ID or file

Binding site detection and **encoding**

Binding site **similarity** and **ranking**

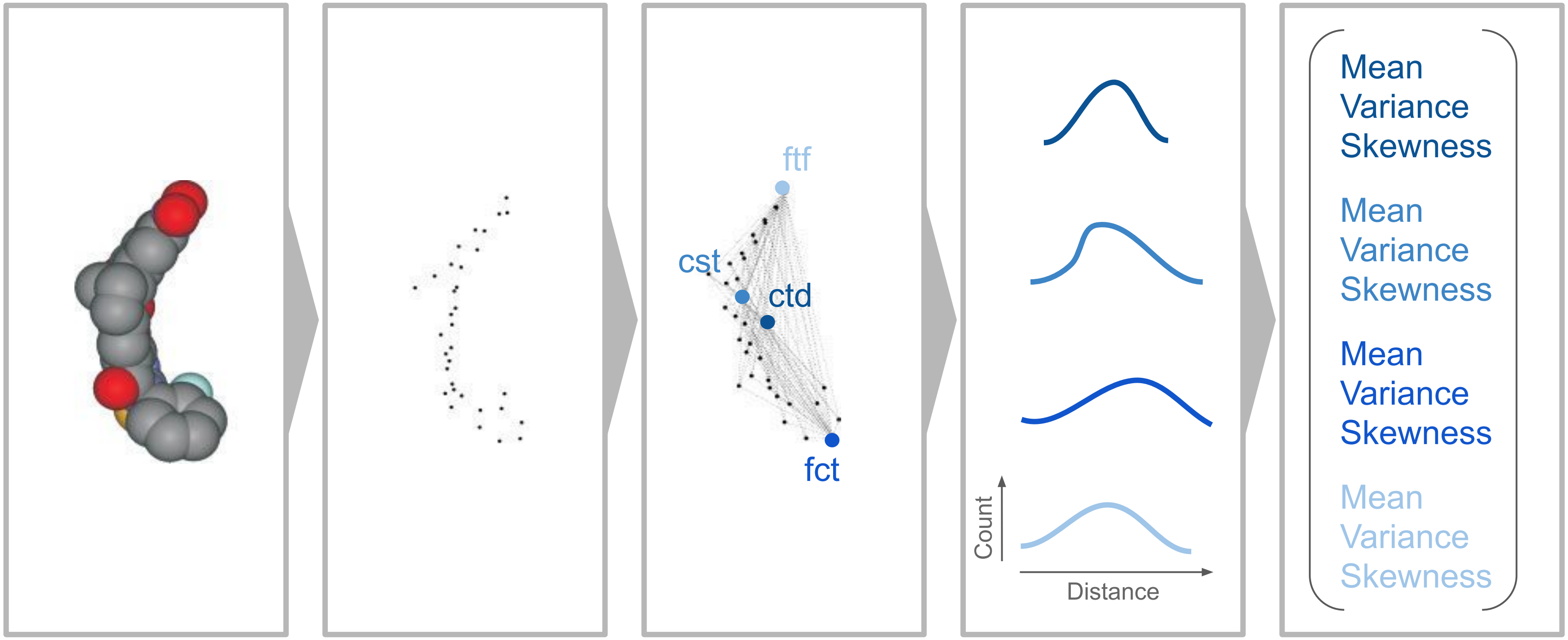
Include protein ensembles

Potential off-targets

Focused compound library

# Intro: USR method for ligand encoding

Molecule | Representatives | Reference points | Distances | Moments



Mean
Variance
Skewness

Mean
Variance
Skewness

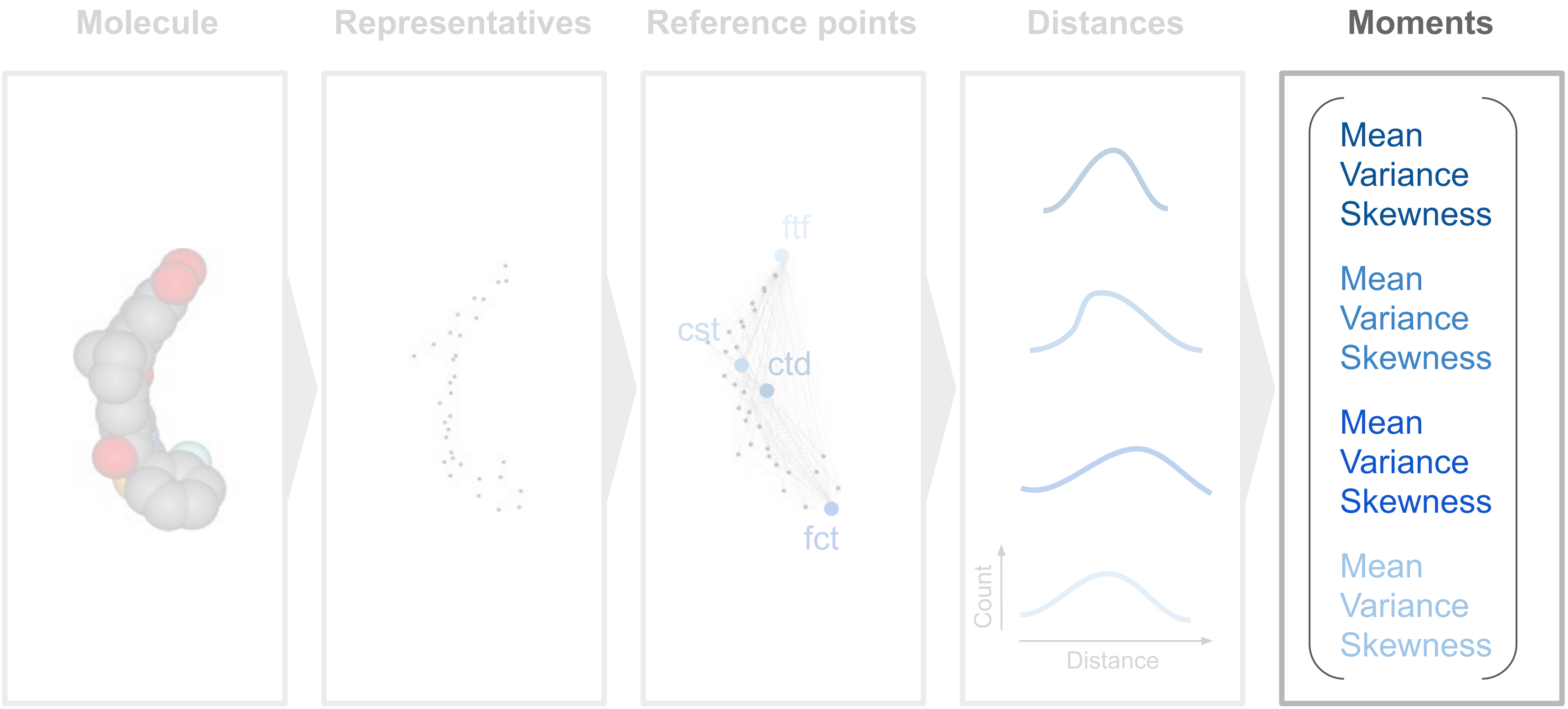
Mean
Variance
Skewness

Mean
Variance
Skewness

Ultrafast shape recognition (USR) - Ballester et al. 2006

Ballester et al. 2006

# Intro: USR method for ligand encoding



Molecule

Representatives

Reference points

Distances

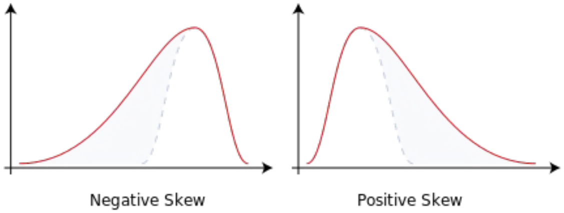
Moments

Mean
Variance
Skewness

Mean
Variance
Skewness

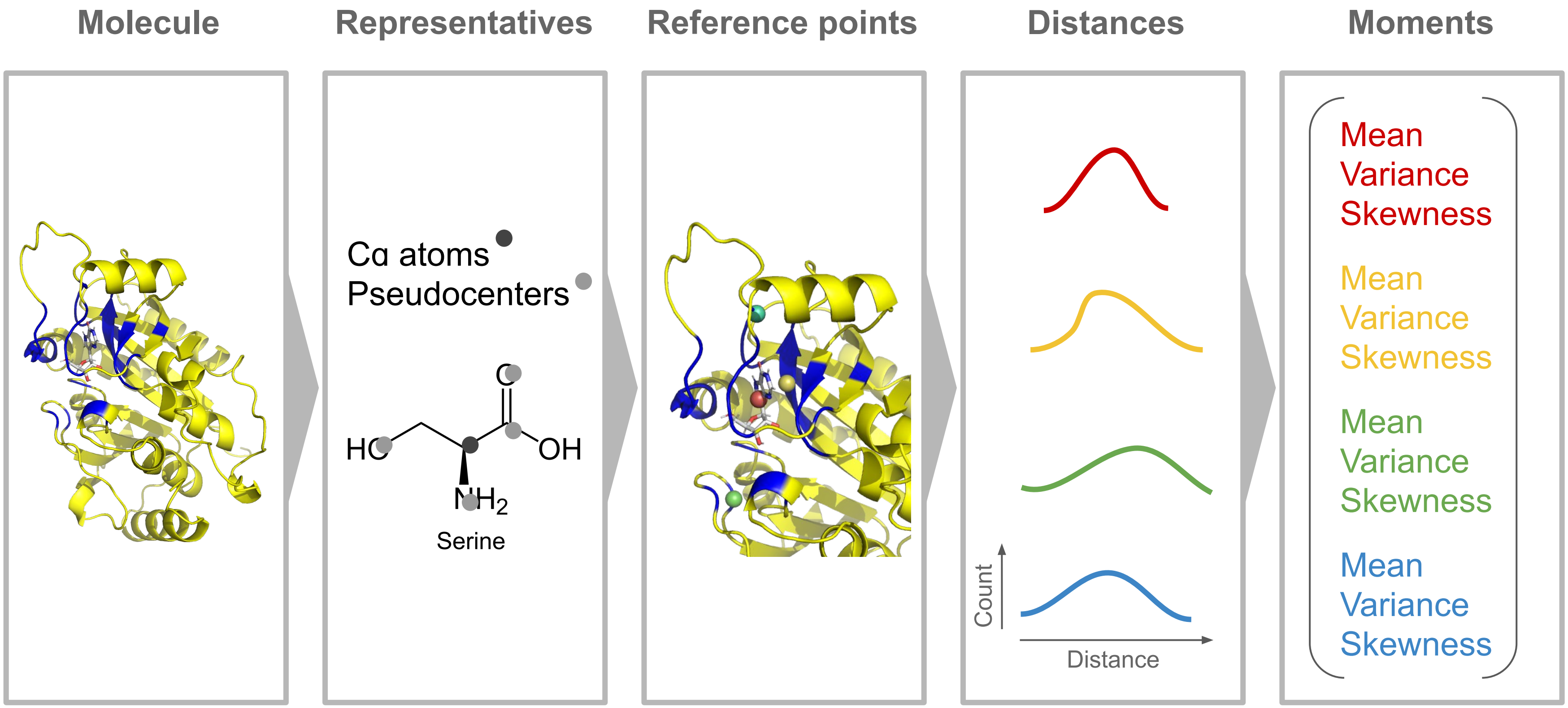
Mean
Variance
Skewness

Mean
Variance
Skewness



https://en.wikipedia.org/wiki/Skewness

# Idea: USR method for binding site encoding?



Molecule

Representatives

Reference points

Distances

Moments

Cα atoms ●
Pseudocenters ●

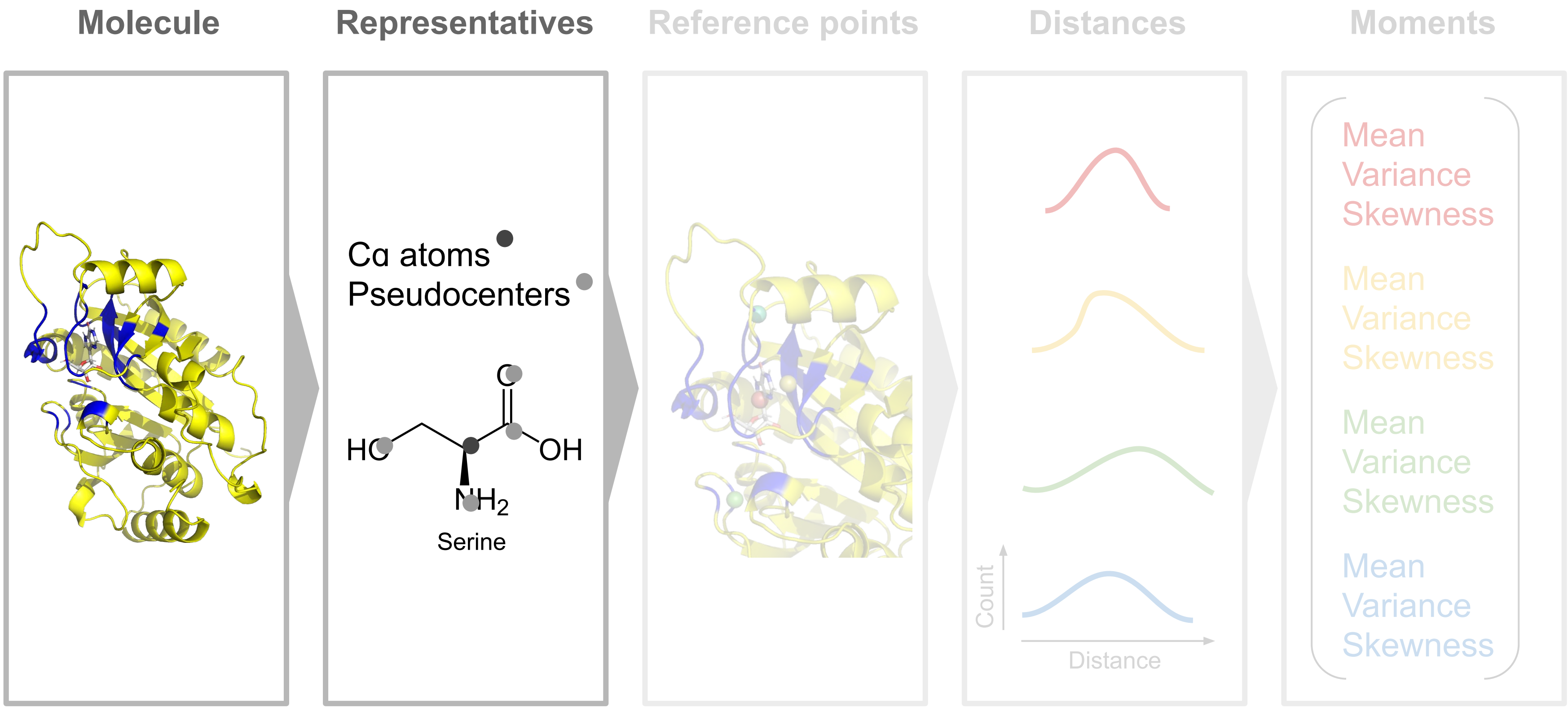
Mean
Variance
Skewness

Mean
Variance
Skewness

Mean
Variance
Skewness

Mean
Variance
Skewness

# Methods: Binding site dimensionality



Molecule

Representatives

Reference points

Distances

Moments

Cα atoms ●

Pseudocenters ●

Serine

Mean
Variance
Skewness

Mean
Variance
Skewness

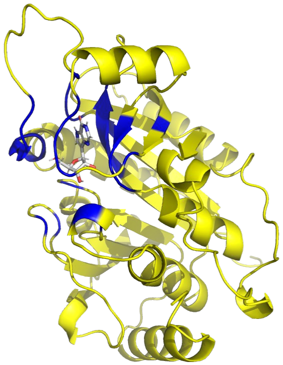
Mean
Variance
Skewness

Mean
Variance
Skewness

How many dimensions per data point?

# Methods: Binding site dimensionality

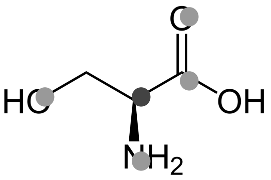
**Molecule**



**Representatives**

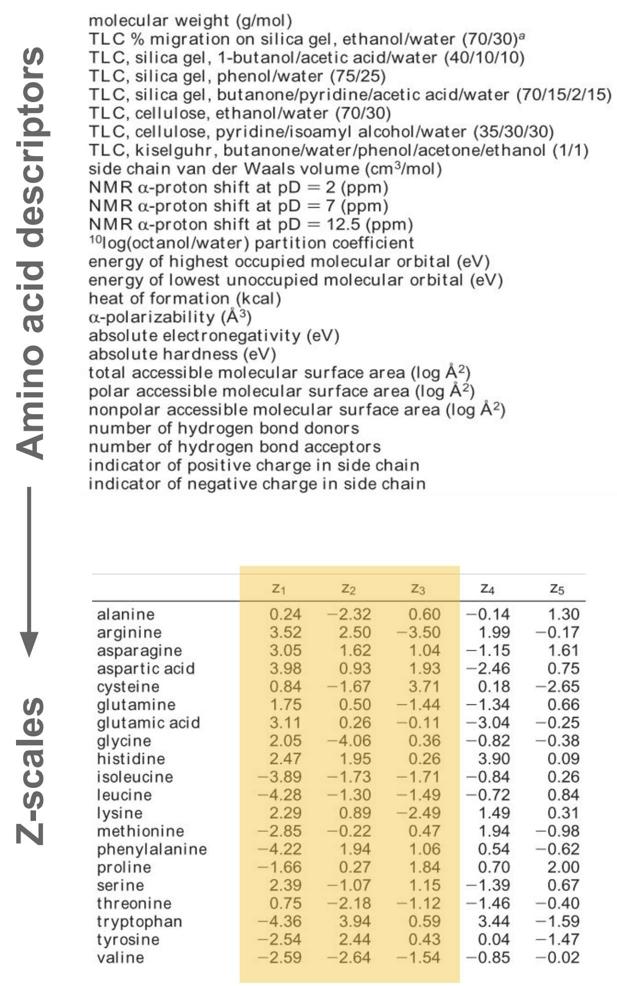
Cα atoms ●

Pseudocenters ●



Serine

How many dimensions per data point?



**Amino acid descriptors**

molecular weight (g/mol)
TLC % migration on silica gel, ethanol/water (70/30)[a]
TLC, silica gel, 1-butanol/acetic acid/water (40/10/10)
TLC, silica gel, phenol/water (75/25)
TLC, silica gel, butanone/pyridine/acetic acid/water (70/15/2/15)
TLC, cellulose, ethanol/water (70/30)
TLC, cellulose, pyridine/isoamyl alcohol/water (35/30/30)
TLC, kiselguhr, butanone/water/phenol/acetone/ethanol (1/1)
side chain van der Waals volume ($cm^3$/mol)
NMR α-proton shift at pD = 2 (ppm)
NMR α-proton shift at pD = 7 (ppm)
NMR α-proton shift at pD = 12.5 (ppm)
$^{10}$log(octanol/water) partition coefficient
energy of highest occupied molecular orbital (eV)
energy of lowest unoccupied molecular orbital (eV)
heat of formation (kcal)
α-polarizability ($Å^3$)
absolute electronegativity (eV)
absolute hardness (eV)
total accessible molecular surface area (log $Å^2$)
polar accessible molecular surface area (log $Å^2$)
nonpolar accessible molecular surface area (log $Å^2$)
number of hydrogen bond donors
number of hydrogen bond acceptors
indicator of positive charge in side chain
indicator of negative charge in side chain

**Z-scales**

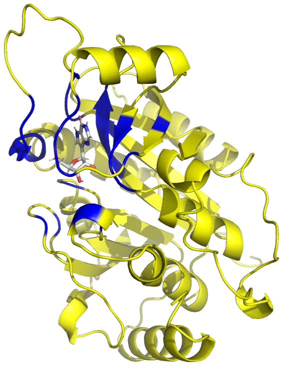
| | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
|---|---|---|---|---|---|
| alanine | 0.24 | −2.32 | 0.60 | −0.14 | 1.30 |
| arginine | 3.52 | 2.50 | −3.50 | 1.99 | −0.17 |
| asparagine | 3.05 | 1.62 | 1.04 | −1.15 | 1.61 |
| aspartic acid | 3.98 | 0.93 | 1.93 | −2.46 | 0.75 |
| cysteine | 0.84 | −1.67 | 3.71 | 0.18 | −2.65 |
| glutamine | 1.75 | 0.50 | −1.44 | −1.34 | 0.66 |
| glutamic acid | 3.11 | 0.26 | −0.11 | −3.04 | −0.25 |
| glycine | 2.05 | −4.06 | 0.36 | −0.82 | −0.38 |
| histidine | 2.47 | 1.95 | 0.26 | 3.90 | 0.09 |
| isoleucine | −3.89 | −1.73 | −1.71 | −0.84 | 0.26 |
| leucine | −4.28 | −1.30 | −1.49 | −0.72 | 0.84 |
| lysine | 2.29 | 0.89 | −2.49 | 1.49 | 0.31 |
| methionine | −2.85 | −0.22 | 0.47 | 1.94 | −0.98 |
| phenylalanine | −4.22 | 1.94 | 1.06 | 0.54 | −0.62 |
| proline | −1.66 | 0.27 | 1.84 | 0.70 | 2.00 |
| serine | 2.39 | −1.07 | 1.15 | −1.39 | 0.67 |
| threonine | 0.75 | −2.18 | −1.12 | −1.46 | −0.40 |
| tryptophan | −4.36 | 3.94 | 0.59 | 3.44 | −1.59 |
| tyrosine | −2.54 | 2.44 | 0.43 | 0.04 | −1.47 |
| valine | −2.59 | −2.64 | −1.54 | −0.85 | −0.02 |

$z_1$ Lipophilicity
$z_2$ Steric bulk/ polarisability
$z_3$ Polarity

Sandberg et al. 1998

# Methods: Binding site dimensionality

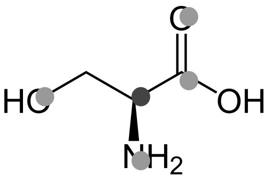
**Molecule**



**Representatives**

Cα atoms ●

Pseudocenters ●



Serine

**6 dimensions per atom**

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Spatial property

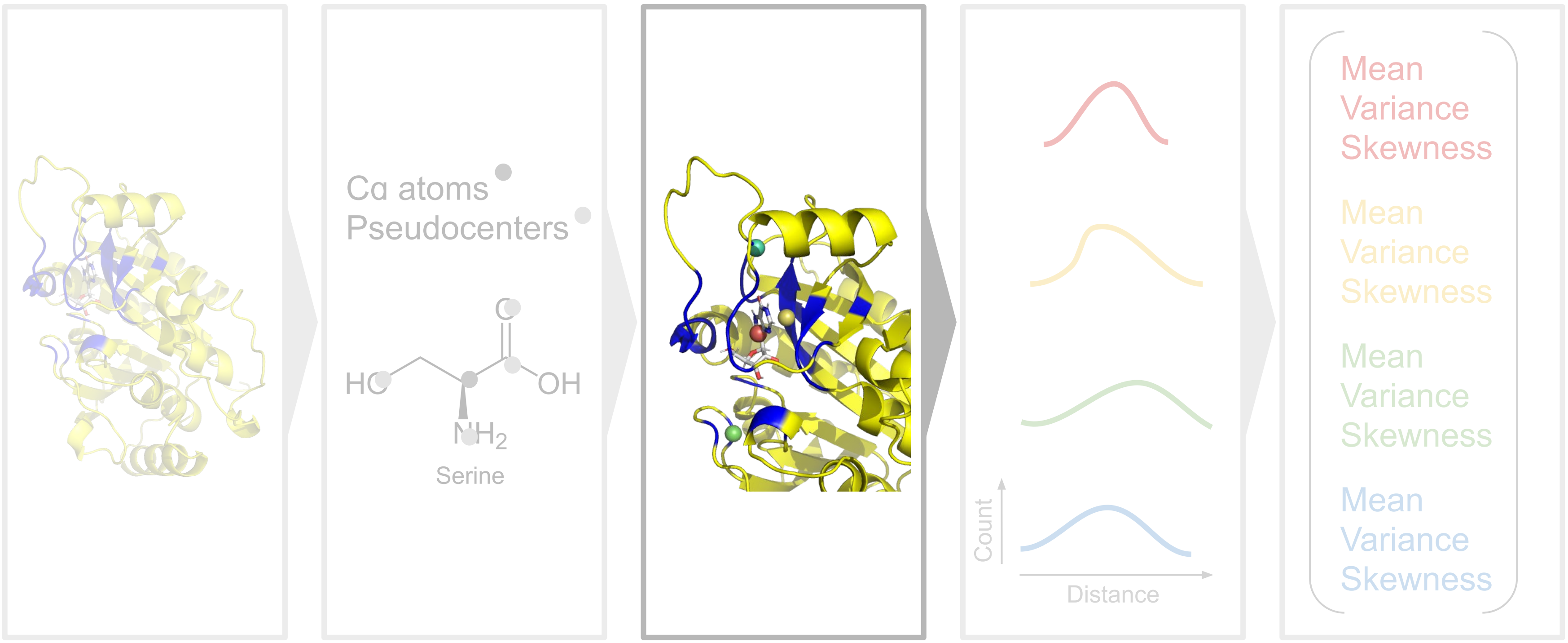
Physicochemical property

How many dimensions per data point?

Sandberg et al. 1998

# Methods: Number of reference points?

Molecule | Representatives | **Reference points** | Distances | Moments



Cα atoms

Pseudocenters

Mean
Variance
Skewness

Mean
Variance
Skewness

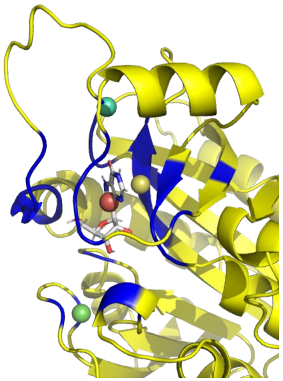
Mean
Variance
Skewness

Mean
Variance
Skewness

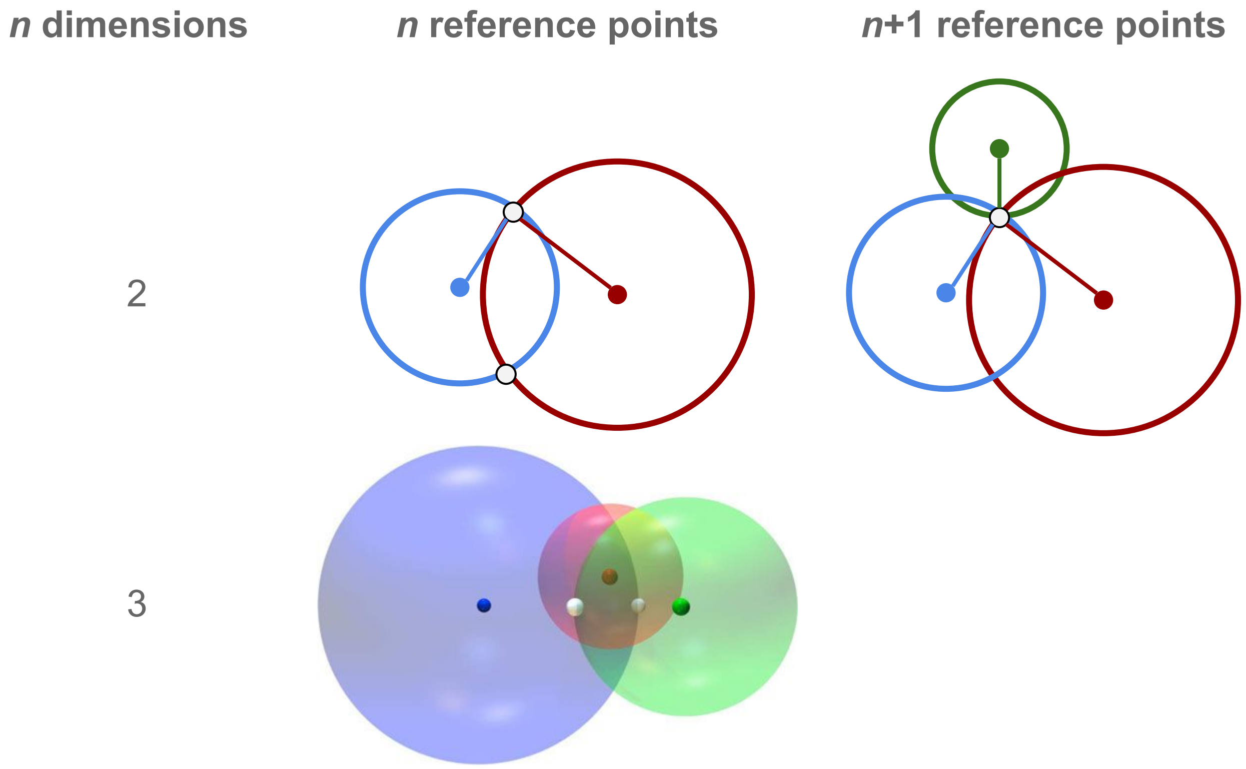
Number of reference points?

# Methods: Number of reference points?

## Reference points



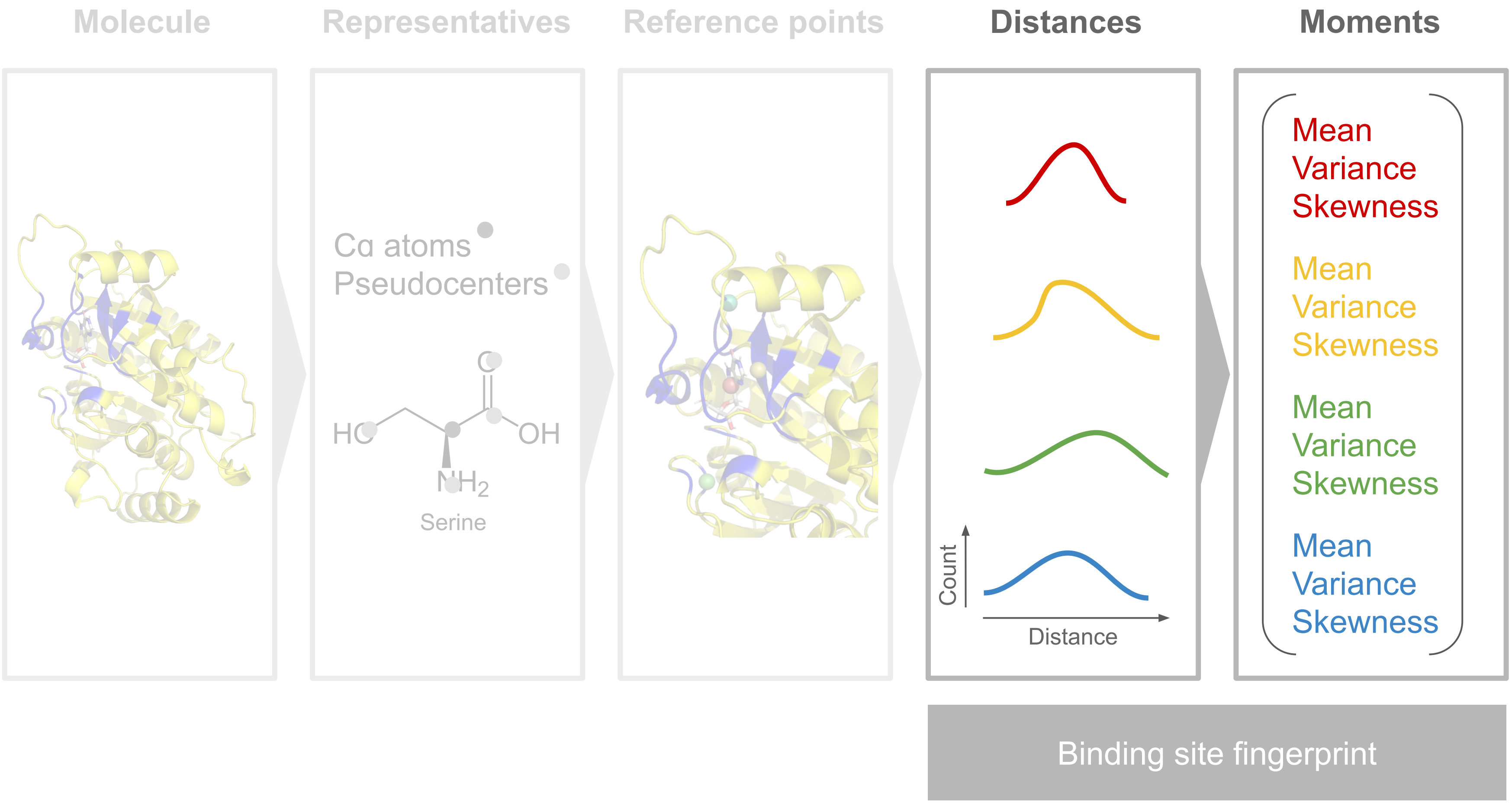
Number of reference points?



To find the exact position of a **point** in $\mathbf{R}^n$,
distances to ***n*+1 fixed (reference) points** are needed.

Armstrong et al. 2010
(ElectroShape)

# Methods: Binding site fingerprints



Molecule

Representatives

Reference points

Distances

Moments

Cα atoms

Pseudocenters

Mean
Variance
Skewness

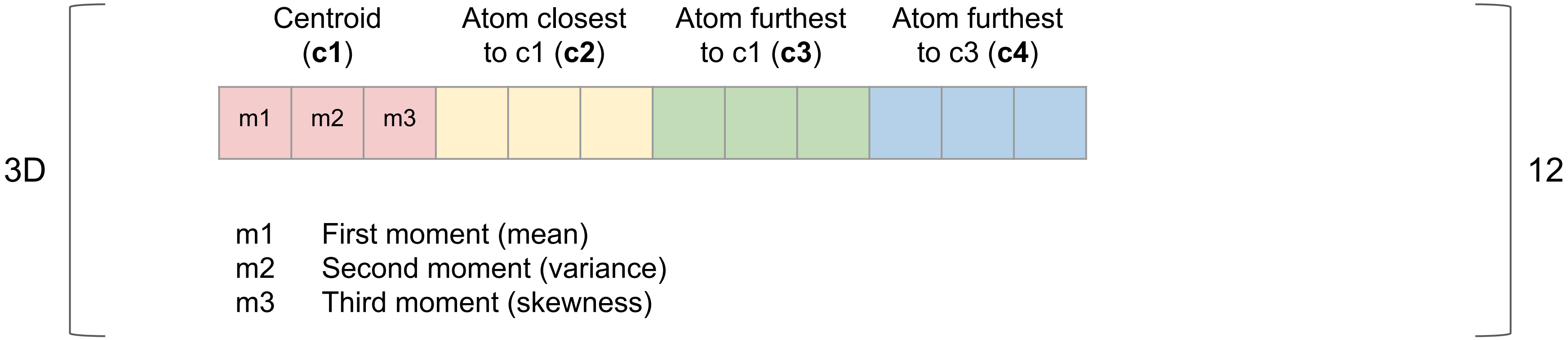
Mean
Variance
Skewness

Mean
Variance
Skewness

Mean
Variance
Skewness

Binding site fingerprint

# Methods: Binding site fingerprints



3D

Centroid (**c1**)

Atom closest to c1 (**c2**)

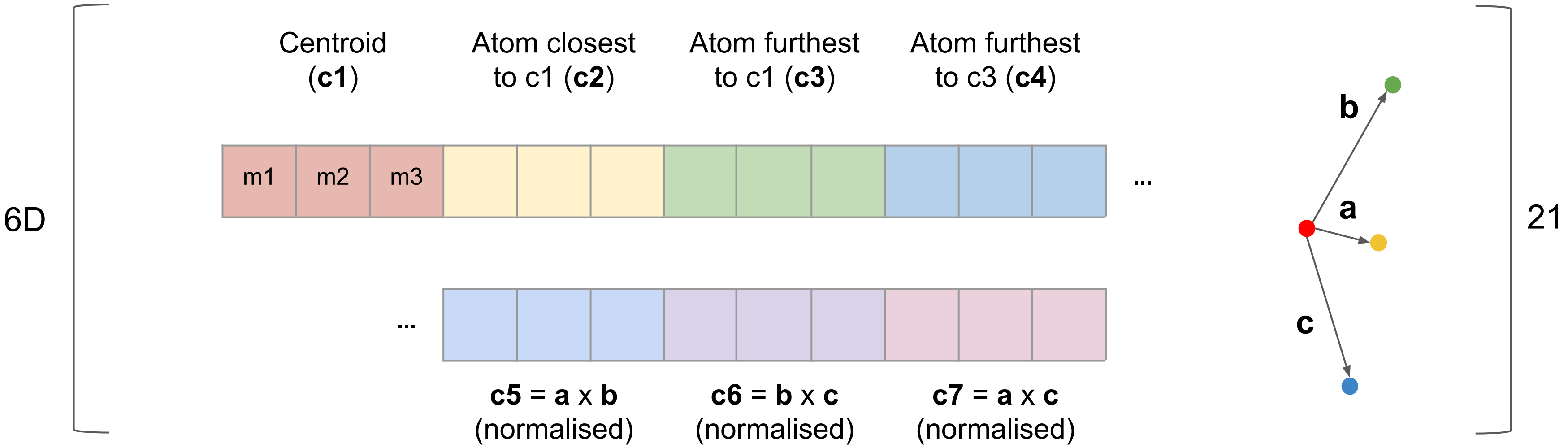
Atom furthest to c1 (**c3**)

Atom furthest to c3 (**c4**)

m1 m2 m3

12

m1 First moment (mean)
m2 Second moment (variance)
m3 Third moment (skewness)



6D

Centroid (**c1**)

Atom closest to c1 (**c2**)

Atom furthest to c1 (**c3**)

Atom furthest to c3 (**c4**)

m1 m2 m3 ...

...

**c5** = **a** x **b** (normalised)

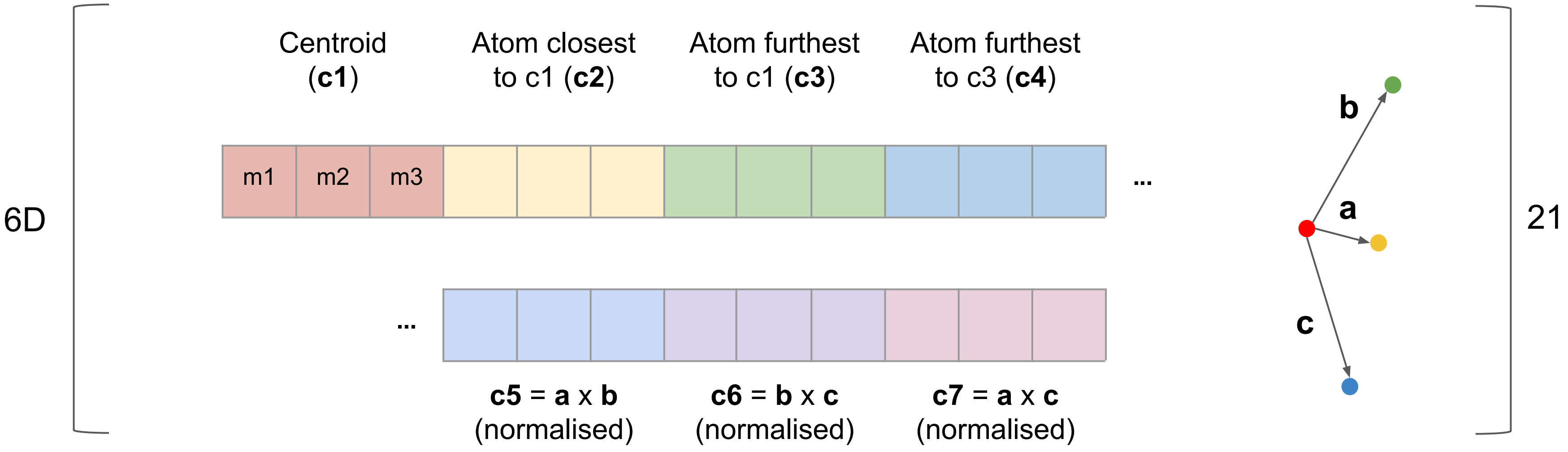
**c6** = **b** x **c** (normalised)

**c7** = **a** x **c** (normalised)

21

# Methods: Binding site fingerprints

Outlook: Calculate cross product in 6 dimensions?



6D

Centroid (**c1**)

Atom closest to c1 (**c2**)

Atom furthest to c1 (**c3**)

Atom furthest to c3 (**c4**)

| m1 | m2 | m3 | | | | | | | | | | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

| ... | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|

**c5** = **a** x **b** (normalised)

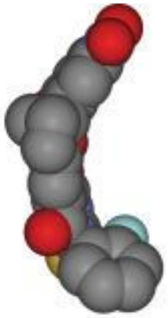
**c6** = **b** x **c** (normalised)

**c7** = **a** x **c** (normalised)

21

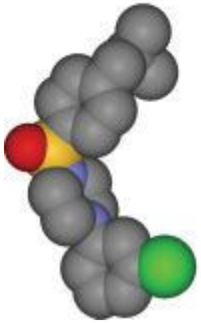
# Methods: Binding site similarity measure

| Centroid (**c1**) | | | Atom closest to c1 (**c2**) | | | Atom furthest to c1 (**c3**) | | | Atom furthest to c3 (**c4**) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m1 | m2 | m3 | | | | | | | | | |



$$\vec{M}^q = (4.44,\ 2.98,\ 1.04,\ 4.55,\ 4.70,\ 0.23,\ 8.30,\ 16.69,\ -22.97,\ 7.37,\ 15.64,\ 0.51)$$

$$S_{qi} = \frac{1}{1+\frac{1}{12}\sum_{l=1}^{12}\left|M_l^q - M_l^i\right|} \boxed{\in (0,1]} \qquad S_{qi} = 0.812$$



$$\vec{M}^i = (4.39,\ 3.11,\ 1.36,\ 4.50,\ 4.44,\ 0.09,\ 8.34,\ 16.78,\ -23.20,\ 7.15,\ 16.52,\ 0.13)$$

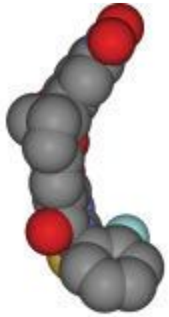
**Similarity measure**

Inverse of the translated and scaled **Manhattan distance**

Ballester et al. 2006

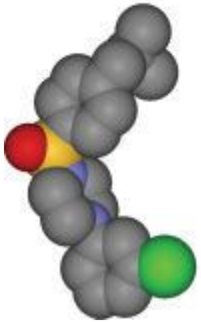
# Methods: Binding site similarity measure

| Centroid (**c1**) | | | Atom closest to c1 (**c2**) | | | Atom furthest to c1 (**c3**) | | | Atom furthest to c3 (**c4**) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m1 | m2 | m3 | | | | | | | | | |



$$\vec{M}^q = (4.44,\ 2.98,\ 1.04,\ 4.55,\ 4.70,\ 0.23,\ 8.30,\ 16.69,\ -22.97,\ 7.37,\ 15.64,\ 0.51)$$

$$S_{qi} = \frac{1}{1 + \frac{1}{12}\sum_{l=1}^{12}\left|M_l^q - M_l^i\right|} \in (0,1] \qquad S_{qi} = 0.812$$



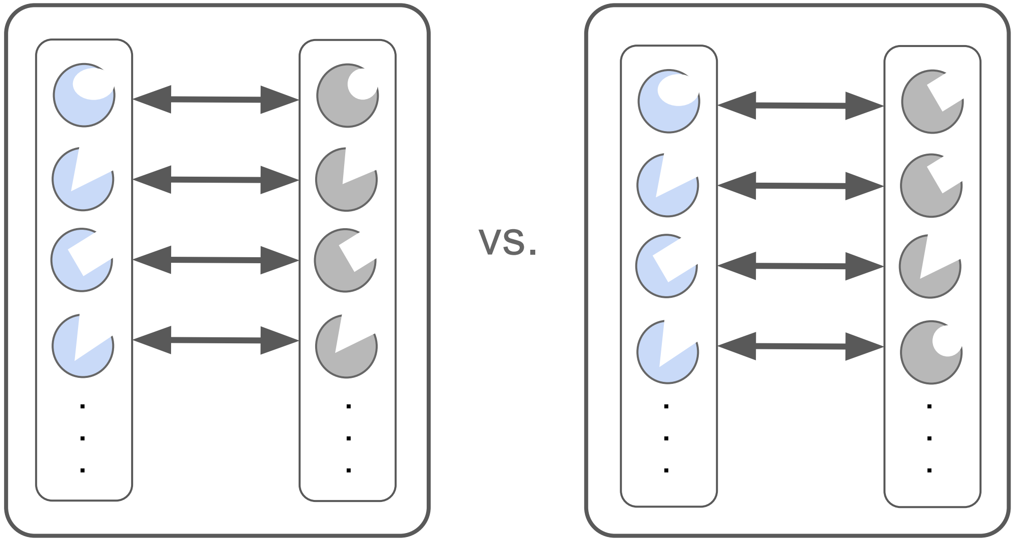
$$\vec{M}^i = (4.39,\ 3.11,\ 1.36,\ 4.50,\ 4.44,\ 0.09,\ 8.34,\ 16.78,\ -23.20,\ 7.15,\ 16.52,\ 0.13)$$

**Similarity measure**

Inverse of the translated and scaled
**Manhattan distance**

Ballester et al. 2006

# Evaluation: Similar vs. dissimilar pairs (FuzCav)



769 pairs of non-redundant **similar binding sites**

769 pairs of non-redundant **dissimilar binding sites**

Data set generation

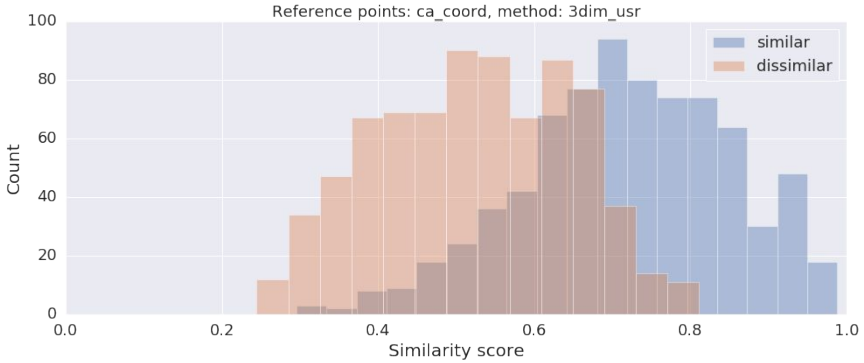
1. Cluster scPDB by UniProt name (911 clusters and 1204 singletons)
2. All-against-all comparison of all active sites within each cluster (SiteAlign)
3. Define *cutoff* for similarity measure discriminating between similar/dissimilar binding sites
4. Choose pairs
   - Similar pairs: select randomly two entries per cluster (considering *cutoff*)
   - Dissimilar pairs: select two entries from clusters differering at the first level of their EC numbers

Weill et al. 2010

# Evaluation: Similar vs. dissimilar pairs (FuzCav)

## Ca atoms

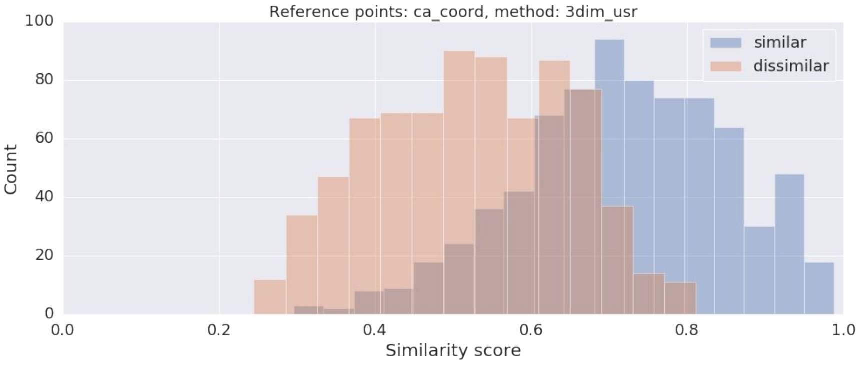
3D



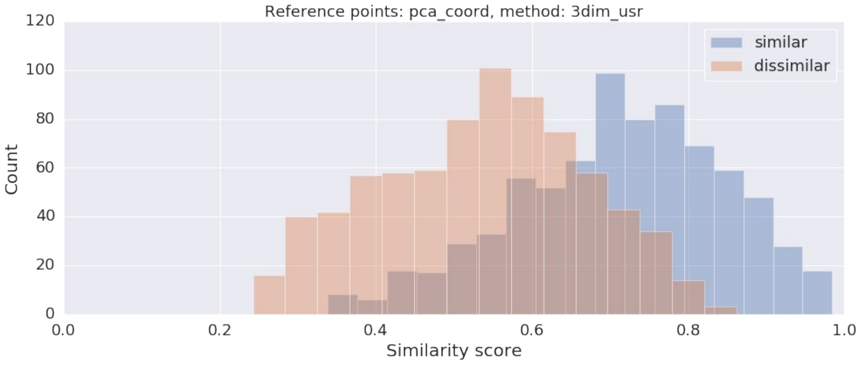
# Evaluation: Similar vs. dissimilar pairs (FuzCav)

3D

## Ca atoms



## Pseudocenter atoms



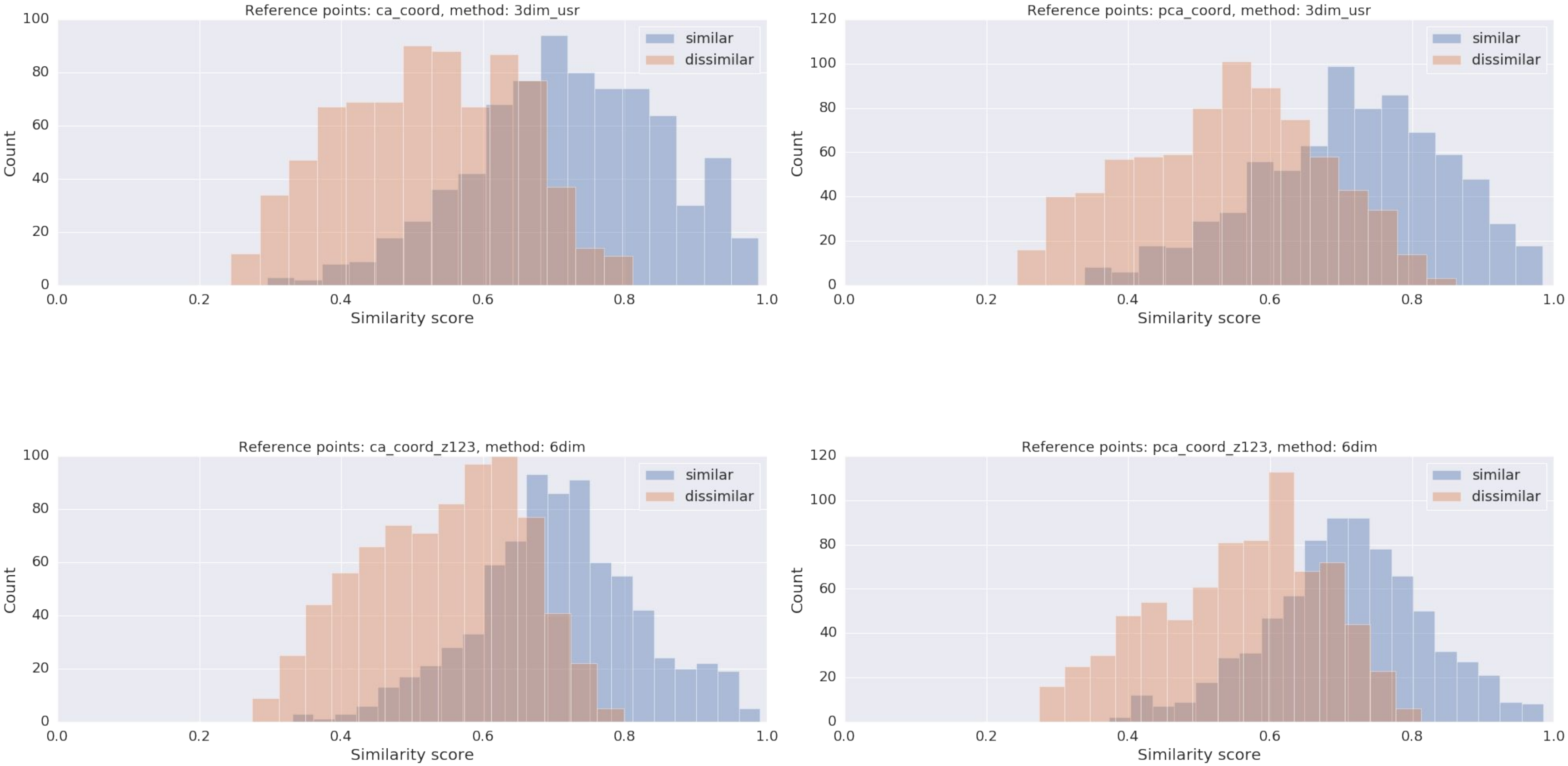
# Evaluation: Similar vs. dissimilar pairs (FuzCav)

## Ca atoms

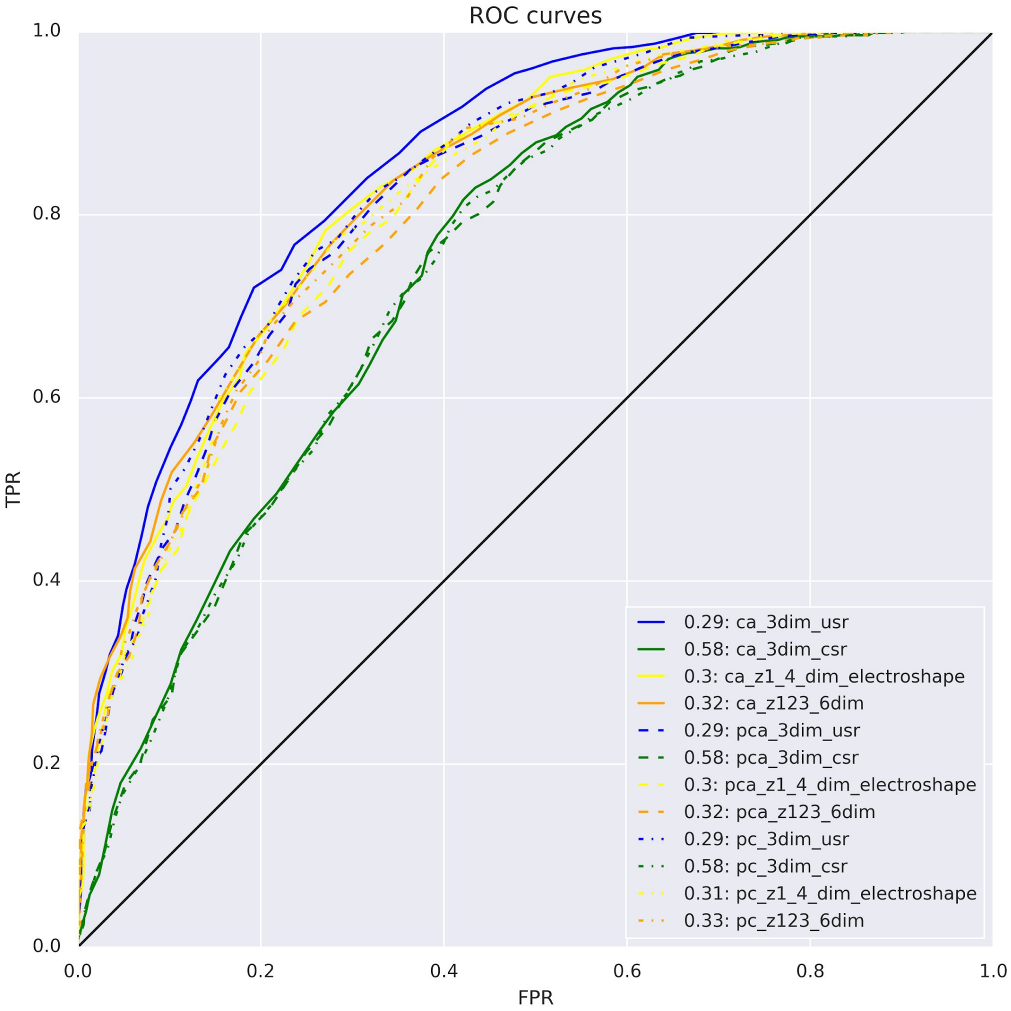
## Pseudocenter atoms

3D

6D



# Evaluation: Similar vs. dissimilar pairs (FuzCav)



# Evaluation: Moment distribution in 6D

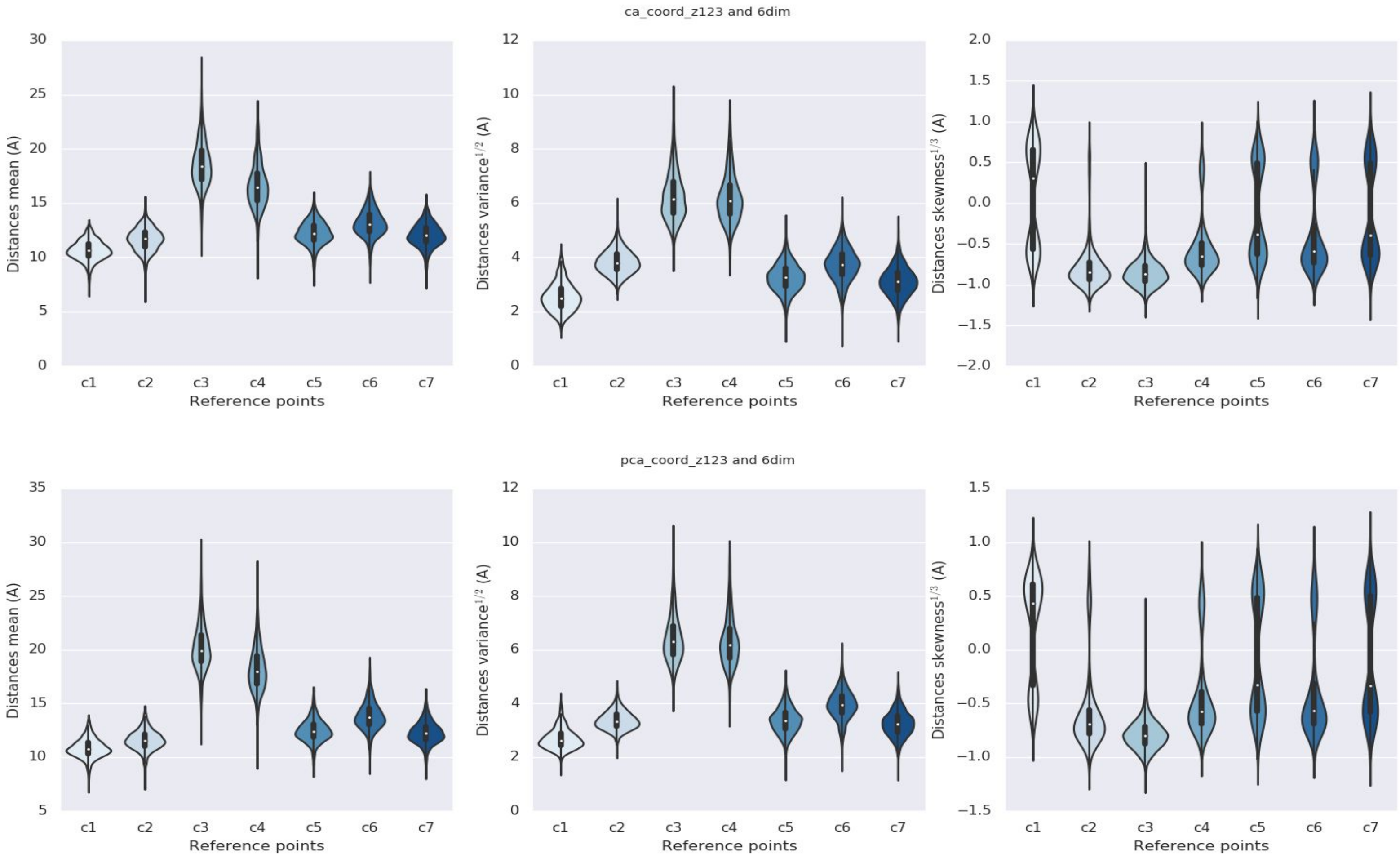
Mean

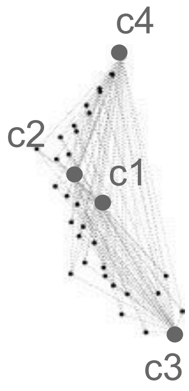
Variance $^{1/2}$

Skewness $^{1/3}$

Cα atoms

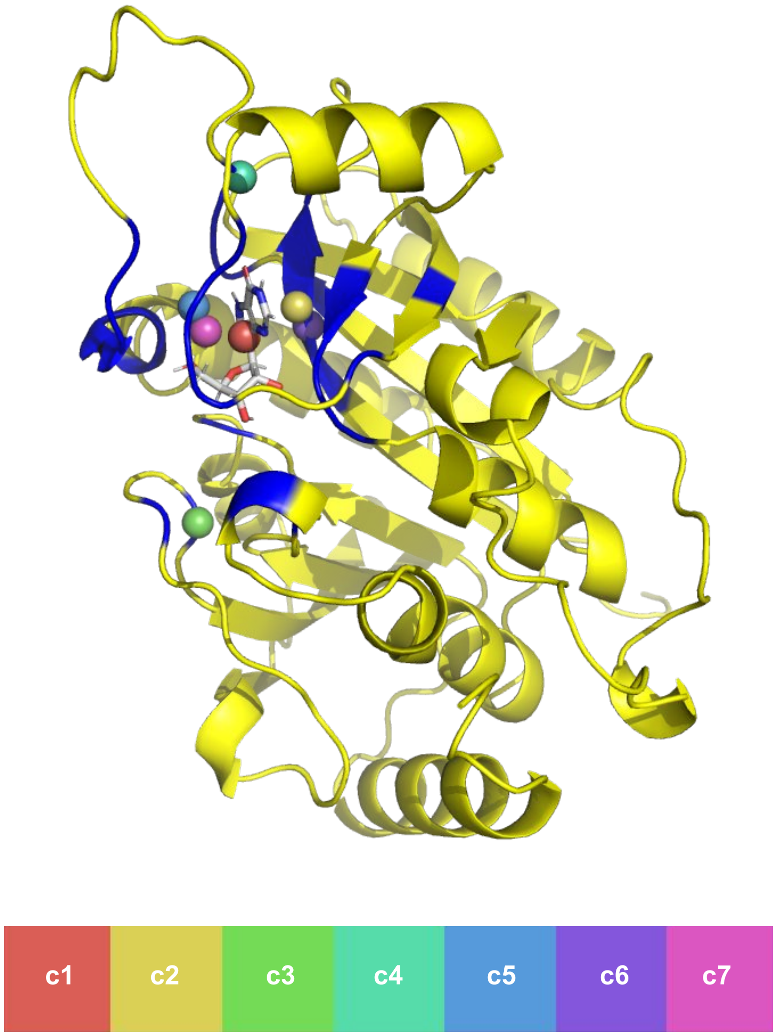
Pseudocenter atoms





# Outlook

- Set reference points based on 6 dimensions
  - Cross product in 6 dimensions
- Instead of Z-scales (per residue) use
  - AutoDock partial charges (per atom) and/or
  - Flexibility information via normal mode analysis (per atom)
- Introduce subpockets/regions
  - Calculate fingerprints for overlapping regions
  - All-against-all fingerprint comparison between binding site regions
  - Find maximal neighboring matches
- Use more information from distance histograms than moments for fingerprint
- Apply method to benchmarking datasets other than FuzCav dataset
  - TOUGH-M1 dataset (Govindaraj et al. 2018)
  - ProSPECCTs datasets (Ehrt et al. 2018)



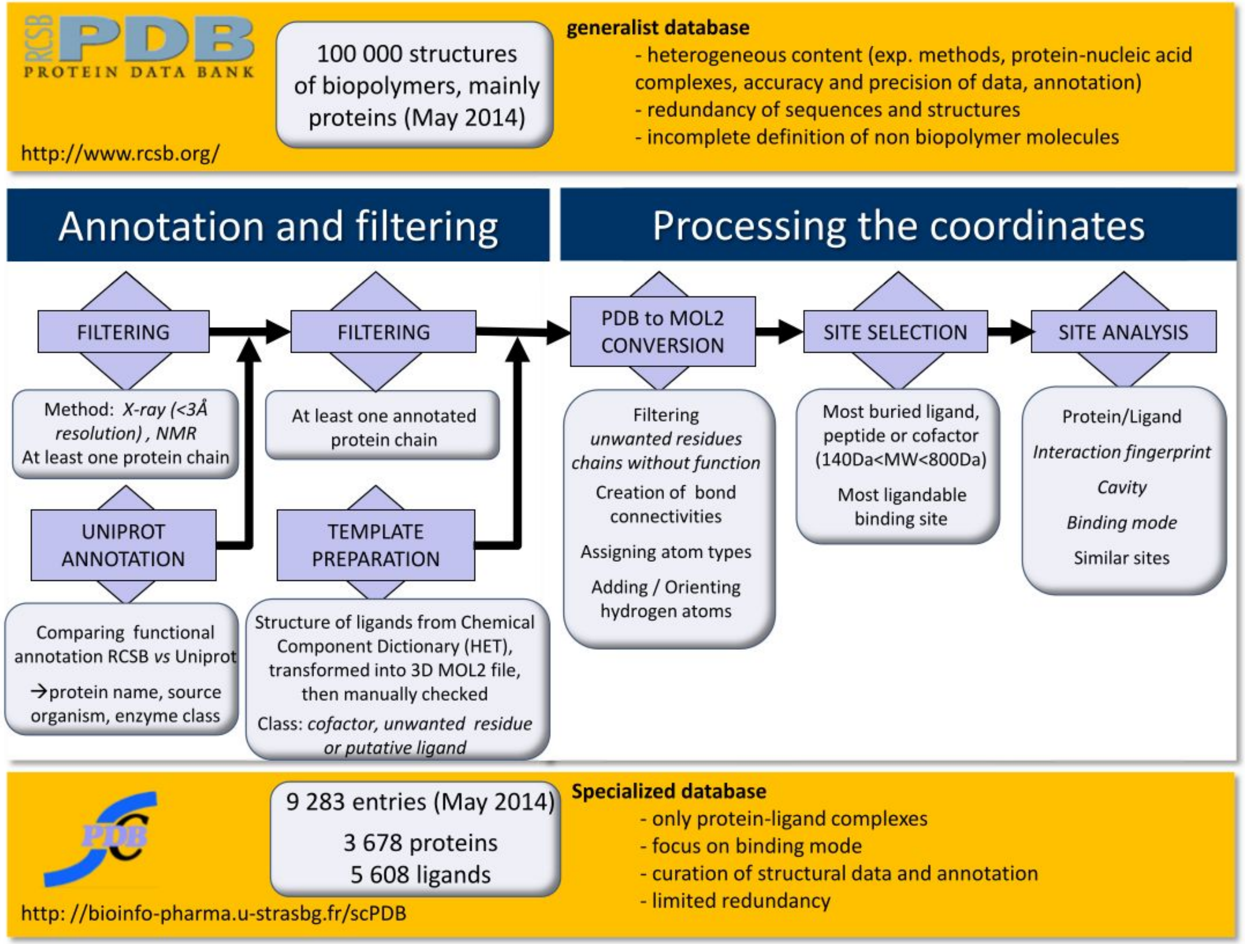
# Code review sessions?

Questions I have...

- How to initialize global variables at start of full program and how to pass them to all associated scripts?
- How to store complex data structures?
  - Dict of dict of Pandas DataFrames
  - Database?
- How to note functions that are only called within other functions (but will not be called by themselves)?
- What is the advantage of Docker over conda - and when is what good to use?
- ...

The end.

# scPDB



RCSB PDB PROTEIN DATA BANK

http://www.rcsb.org/

100 000 structures of biopolymers, mainly proteins (May 2014)

**generalist database**

- heterogeneous content (exp. methods, protein-nucleic acid complexes, accuracy and precision of data, annotation)
- redundancy of sequences and structures
- incomplete definition of non biopolymer molecules

## Annotation and filtering

FILTERING

Method: *X-ray (<3Å resolution)*, *NMR*
At least one protein chain

UNIPROT ANNOTATION

Comparing functional annotation RCSB *vs* Uniprot

→protein name, source organism, enzyme class

FILTERING

At least one annotated protein chain

TEMPLATE PREPARATION

Structure of ligands from Chemical Component Dictionary (HET), transformed into 3D MOL2 file, then manually checked

Class: *cofactor, unwanted residue or putative ligand*

## Processing the coordinates

PDB to MOL2 CONVERSION

Filtering
*unwanted residues*
*chains without function*

Creation of bond connectivities

Assigning atom types

Adding / Orienting hydrogen atoms

SITE SELECTION

Most buried ligand, peptide or cofactor (140Da<MW<800Da)

Most ligandable binding site

SITE ANALYSIS

Protein/Ligand

*Interaction fingerprint*

*Cavity*

*Binding mode*

Similar sites

scPDB

9 283 entries (May 2014)

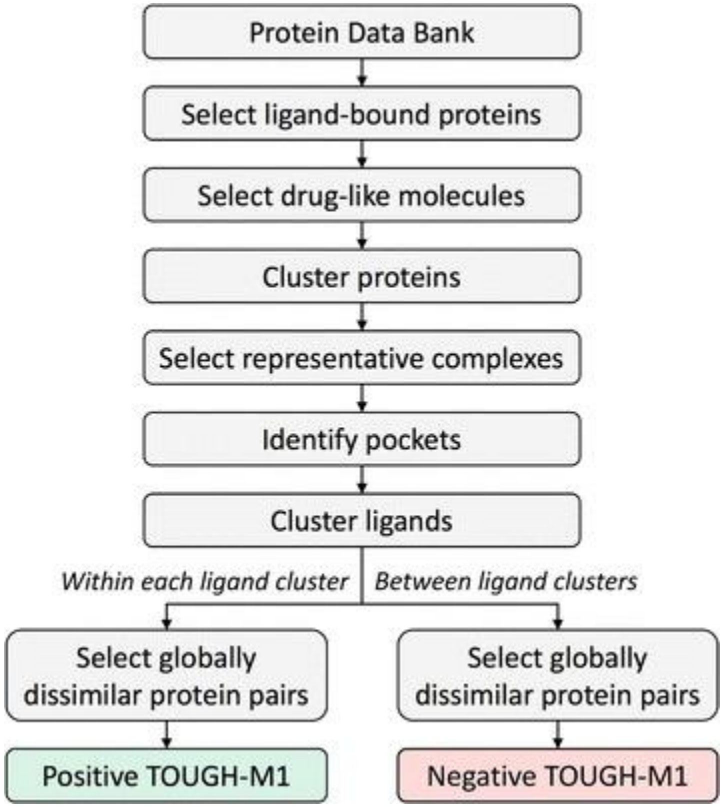
3 678 proteins
5 608 ligands

http: //bioinfo-pharma.u-strasbg.fr/scPDB

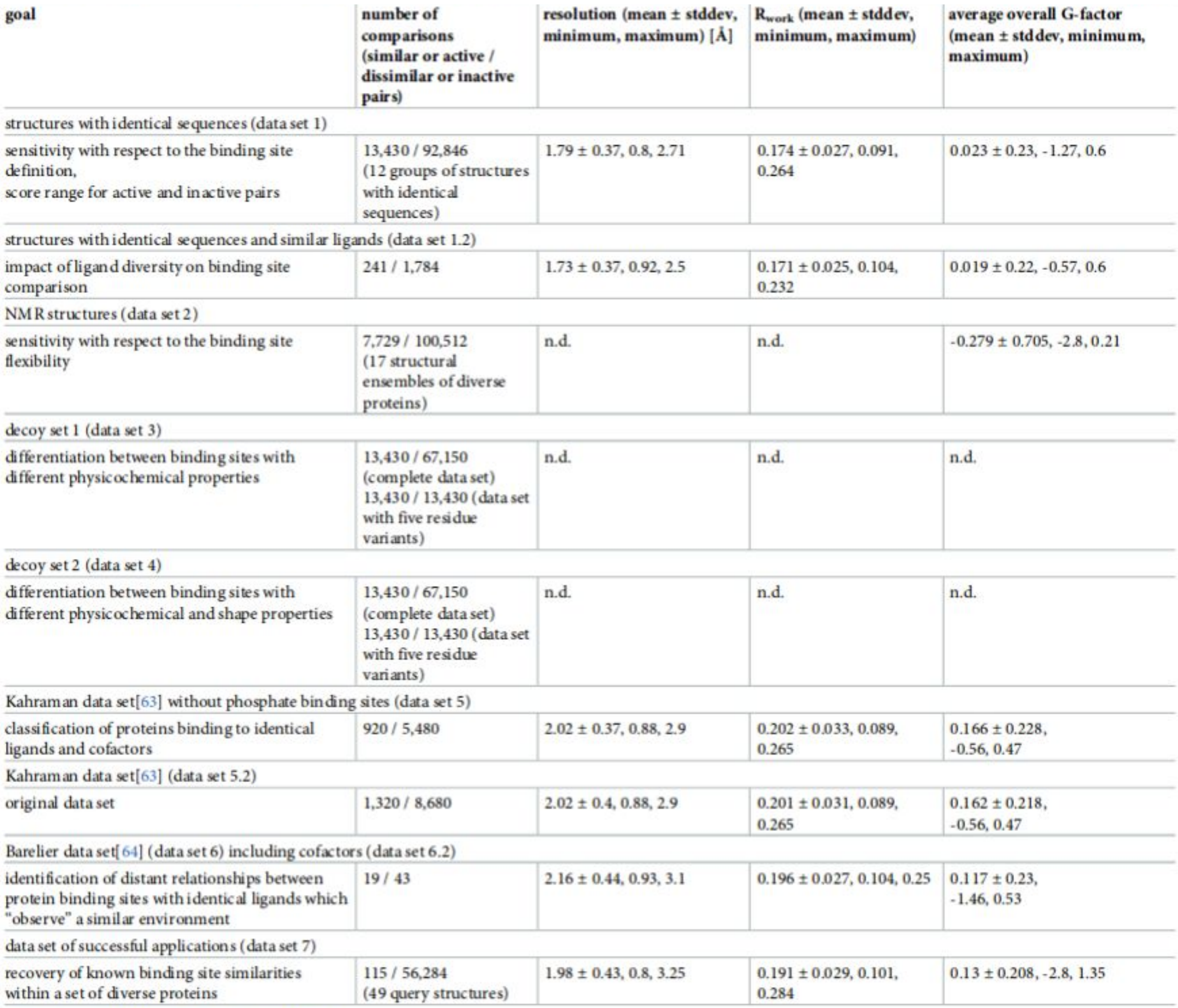
**Specialized database**

- only protein-ligand complexes
- focus on binding mode
- curation of structural data and annotation
- limited redundancy

# Benchmarking dataset: TOUGH-M1


Protein Data Bank
Select ligand-bound proteins
Select drug-like molecules
Cluster proteins
Select representative complexes
Identify pockets
Cluster ligands
Within each ligand cluster
Between ligand clusters
Select globally dissimilar protein pairs
Select globally dissimilar protein pairs
Positive TOUGH-M1
Negative TOUGH-M1

# Benchmarking datasets: ProSPECCTs



| goal | number of comparisons (similar or active / dissimilar or inactive pairs) | resolution (mean ± stddev, minimum, maximum) [Å] | $R_{work}$ (mean ± stddev, minimum, maximum) | average overall G-factor (mean ± stddev, minimum, maximum) |
|---|---|---|---|---|
| structures with identical sequences (data set 1) | | | | |
| sensitivity with respect to the binding site definition,<br>score range for active and inactive pairs | 13,430 / 92,846 (12 groups of structures with identical sequences) | 1.79 ± 0.37, 0.8, 2.71 | 0.174 ± 0.027, 0.091, 0.264 | 0.023 ± 0.23, -1.27, 0.6 |
| structures with identical sequences and similar ligands (data set 1.2) | | | | |
| impact of ligand diversity on binding site comparison | 241 / 1,784 | 1.73 ± 0.37, 0.92, 2.5 | 0.171 ± 0.025, 0.104, 0.232 | 0.019 ± 0.22, -0.57, 0.6 |
| NMR structures (data set 2) | | | | |
| sensitivity with respect to the binding site flexibility | 7,729 / 100,512 (17 structural ensembles of diverse proteins) | n.d. | n.d. | -0.279 ± 0.705, -2.8, 0.21 |
| decoy set 1 (data set 3) | | | | |
| differentiation between binding sites with different physicochemical properties | 13,430 / 67,150 (complete data set)<br>13,430 / 13,430 (data set with five residue variants) | n.d. | n.d. | n.d. |
| decoy set 2 (data set 4) | | | | |
| differentiation between binding sites with different physicochemical and shape properties | 13,430 / 67,150 (complete data set)<br>13,430 / 13,430 (data set with five residue variants) | n.d. | n.d. | n.d. |
| Kahraman data set[63] without phosphate binding sites (data set 5) | | | | |
| classification of proteins binding to identical ligands and cofactors | 920 / 5,480 | 2.02 ± 0.37, 0.88, 2.9 | 0.202 ± 0.033, 0.089, 0.265 | 0.166 ± 0.228, -0.56, 0.47 |
| Kahraman data set[63] (data set 5.2) | | | | |
| original data set | 1,320 / 8,680 | 2.02 ± 0.4, 0.88, 2.9 | 0.201 ± 0.031, 0.089, 0.265 | 0.162 ± 0.218, -0.56, 0.47 |
| Barelier data set[64] (data set 6) including cofactors (data set 6.2) | | | | |
| identification of distant relationships between protein binding sites with identical ligands which "observe" a similar environment | 19 / 43 | 2.16 ± 0.44, 0.93, 3.1 | 0.196 ± 0.027, 0.104, 0.25 | 0.117 ± 0.23, -1.46, 0.53 |
| data set of successful applications (data set 7) | | | | |
| recovery of known binding site similarities within a set of diverse proteins | 115 / 56,284 (49 query structures) | 1.98 ± 0.43, 0.8, 3.25 | 0.191 ± 0.029, 0.101, 0.284 | 0.13 ± 0.208, -2.8, 1.35 |

https://doi.org/10.1371/journal.pcbi.1006483.t002

# Binding site encoding: representatives

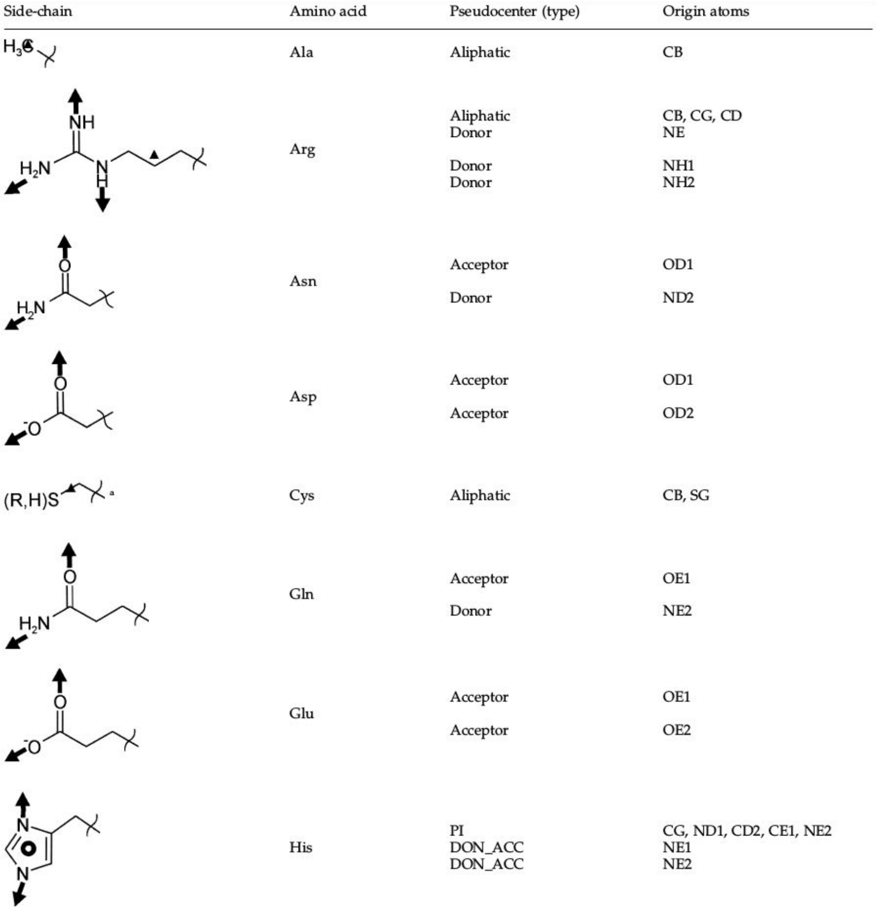
## Point number

- Cα atoms
- Pseudocenters

## Point dimensions *n*

- Spatial information
  - $x_1$, $x_2$, and $x_3$
- Physicochemical information
  - Z-scales $z_1$, $z_2$, and $z_3$ (lipophilicity, steric bulk/polarisability, and polarity)
  - Physicochemical atom subsets (based on pseudocenters: aliphatic, donor, acceptor, aromatic, or donor/acceptor)

Pseudocenters

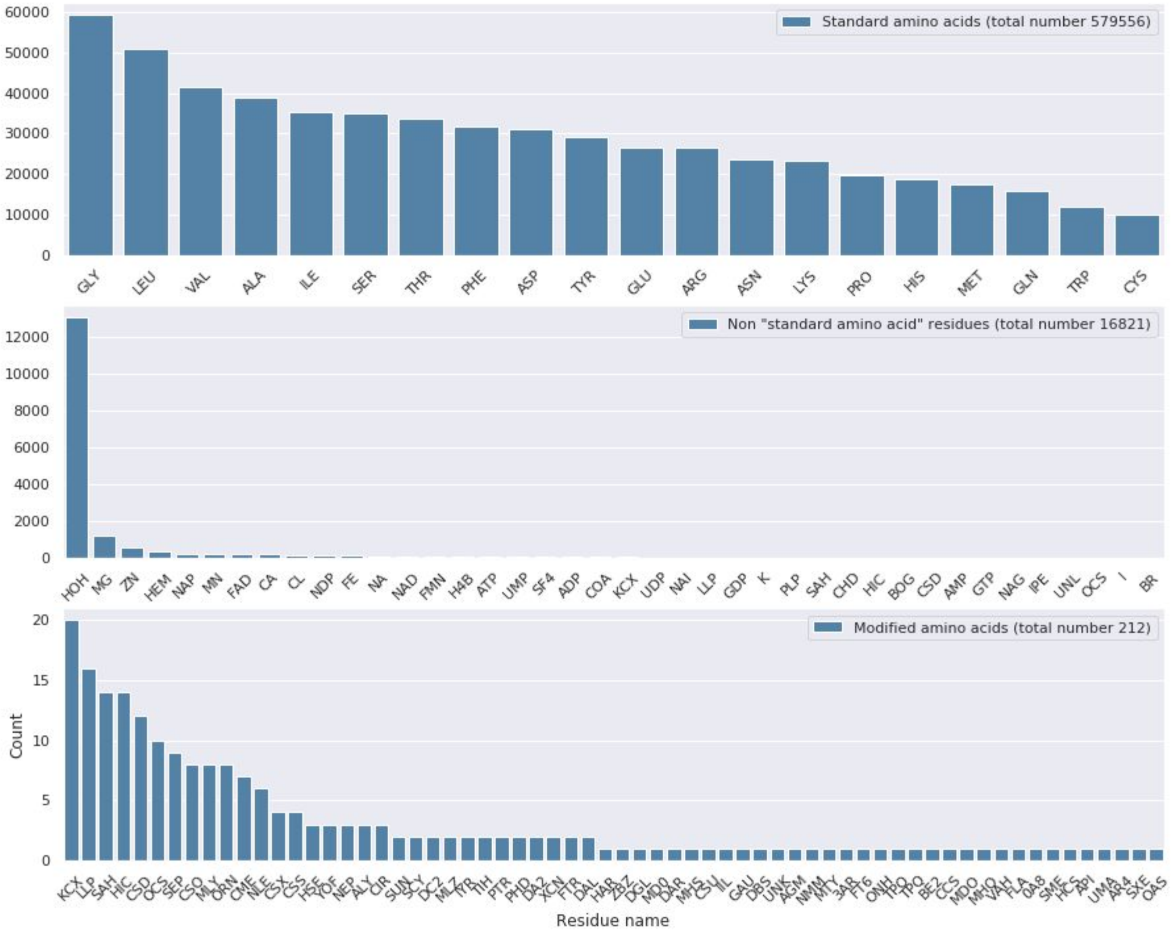


| Side-chain | Amino acid | Pseudocenter (type) | Origin atoms |
|---|---|---|---|
| $H_3C$ | Ala | Aliphatic | CB |
| NH<br>$H_2N$<br>N<br>H | Arg | Aliphatic<br>Donor<br>Donor<br>Donor | CB, CG, CD<br>NE<br>NH1<br>NH2 |
| O<br>$H_2N$ | Asn | Acceptor<br>Donor | OD1<br>ND2 |
| O<br>$^-O$ | Asp | Acceptor<br>Acceptor | OD1<br>OD2 |
| (R,H)S | Cys | Aliphatic | CB, SG |
| O<br>$H_2N$ | Gln | Acceptor<br>Donor | OE1<br>NE2 |
| O<br>$^-O$ | Glu | Acceptor<br>Acceptor | OE1<br>OE2 |
| N<br>N | His | PI<br>DON_ACC<br>DON_ACC | CG, ND1, CD2, CE1, NE2<br>NE1<br>NE2 |

Schmitt et al. 2002 (Cavbase)

# Data set: Residue composition (scPDB)

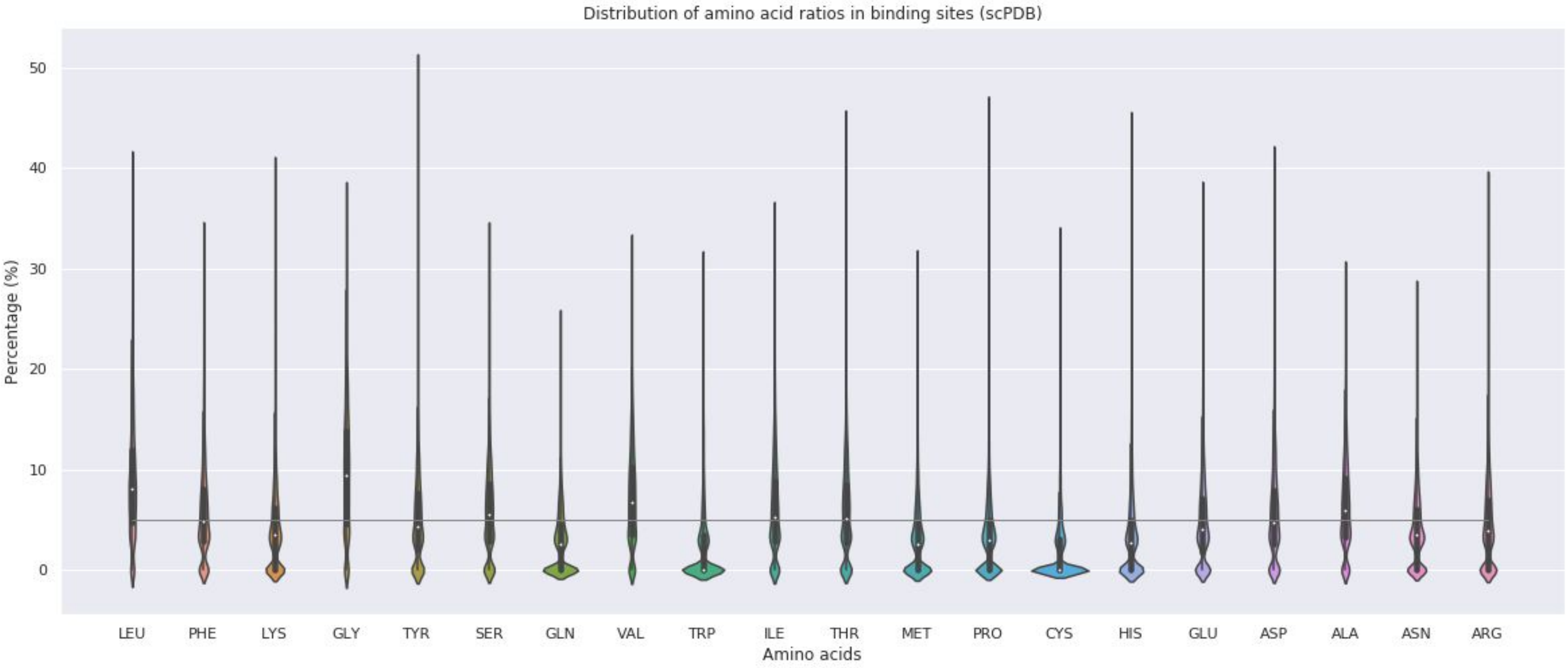
Overall - how often are standard amino acids and other residues?



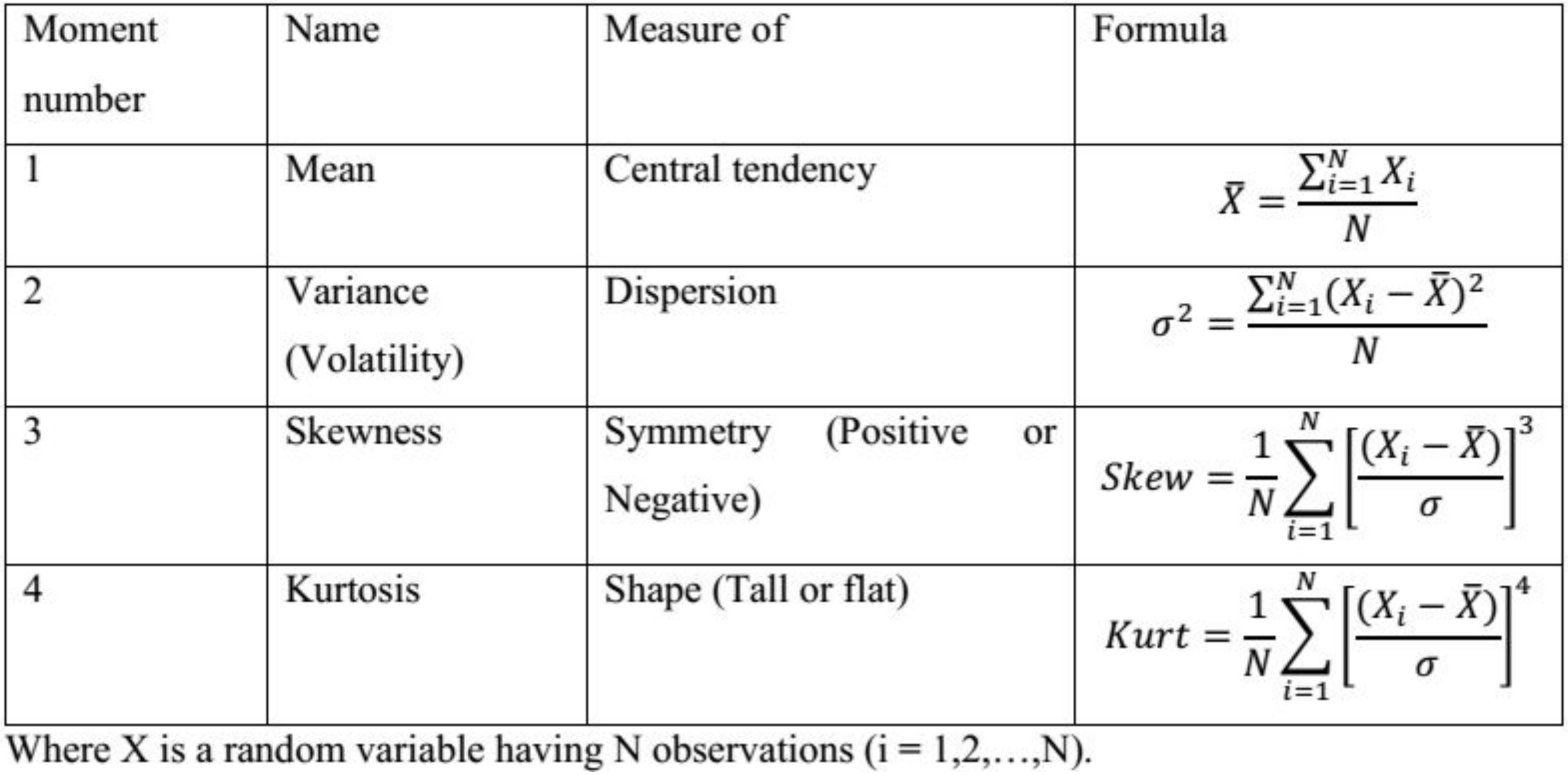
| Mod. aa in scPDB & z-scales | # in scPDB |
|---|---|
| NLE | 6 |
| ISE | 1 |
| ORN | 8 |

# Data set: Residue composition (scPDB)

Per binding site - are standard amino acids somewhat equally distributed?



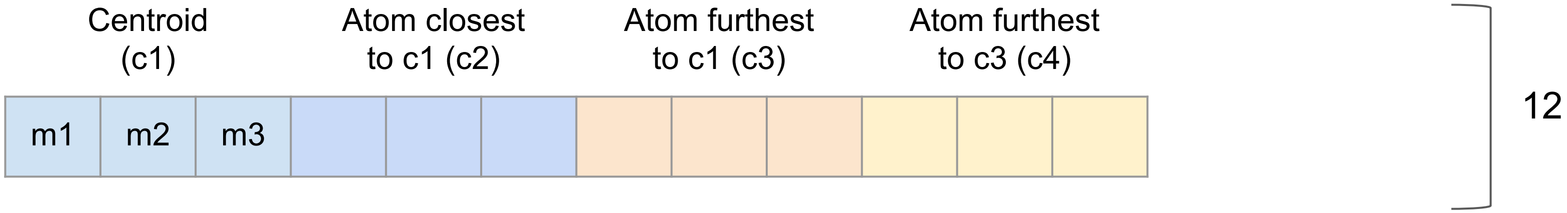
# Moments



| Moment number | Name | Measure of | Formula |
|---|---|---|---|
| 1 | Mean | Central tendency | $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$ |
| 2 | Variance (Volatility) | Dispersion | $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$ |
| 3 | Skewness | Symmetry (Positive or Negative) | $Skew = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{(X_i - \bar{X})}{\sigma} \right]^3$ |
| 4 | Kurtosis | Shape (Tall or flat) | $Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{(X_i - \bar{X})}{\sigma} \right]^4$ |

Where X is a random variable having N observations (i = 1,2,…,N).

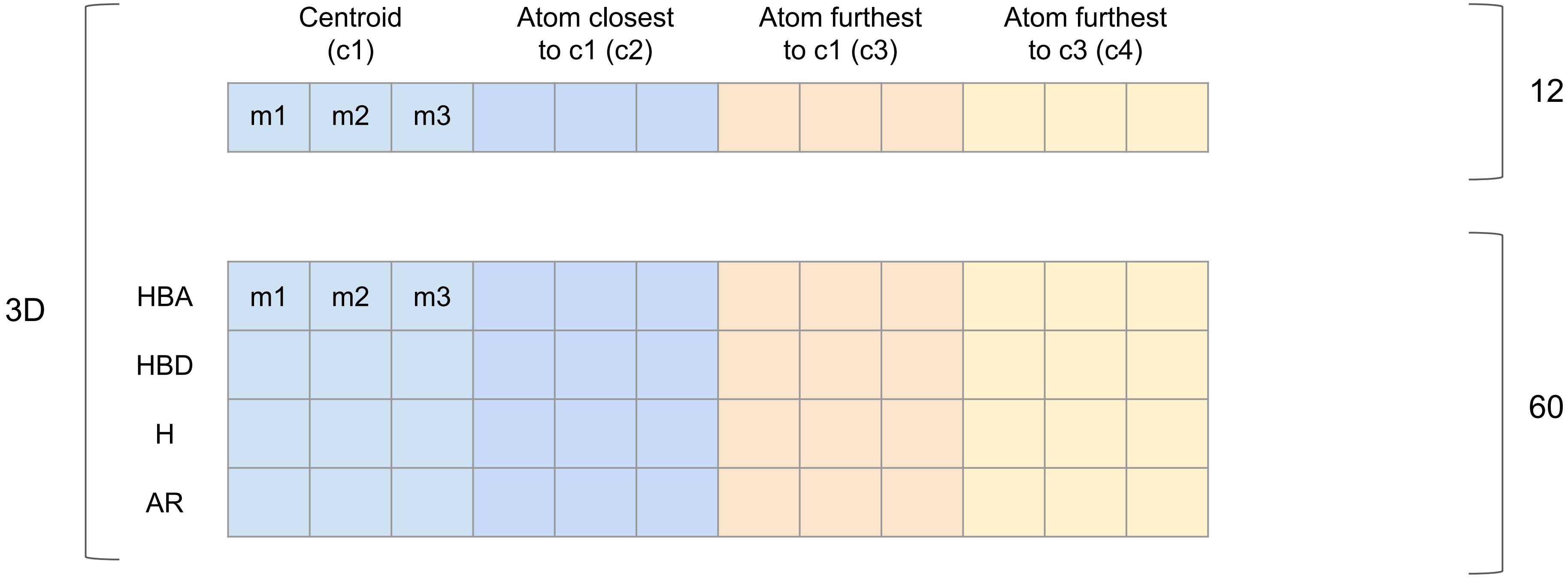
# Binding site “fingerprints” (implemented)



3D

| | |
|---|---|
| m1 | First moment |
| m2 | Second moment |
| m3 | Third moment |

# Binding site “fingerprints” (implemented)



3D

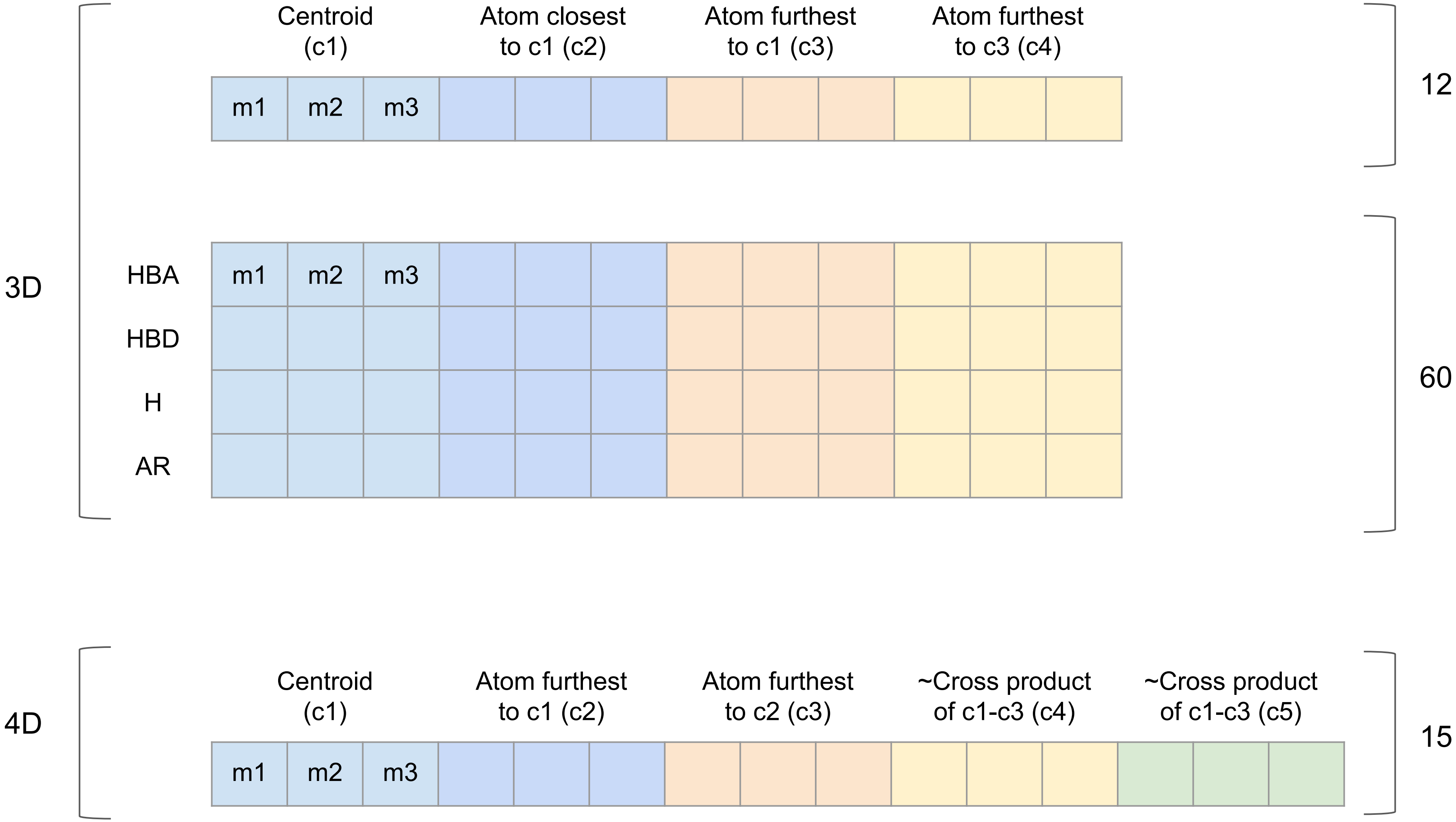
| Centroid (c1) | | | Atom closest to c1 (c2) | | | Atom furthest to c1 (c3) | | | Atom furthest to c3 (c4) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| m1 | m2 | m3 | | | | | | | | | |

12

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HBA | m1 | m2 | m3 | | | | | | | | | |
| HBD | | | | | | | | | | | | |
| H | | | | | | | | | | | | |
| AR | | | | | | | | | | | | |

60

# Binding site “fingerprints” (implemented)



3D

Centroid (c1)

Atom closest to c1 (c2)

Atom furthest to c1 (c3)

Atom furthest to c3 (c4)

m1 m2 m3

12

HBA m1 m2 m3

HBD

H

AR

60

4D

Centroid (c1)

Atom furthest to c1 (c2)

Atom furthest to c2 (c3)

~Cross product of c1-c3 (c4)

~Cross product of c1-c3 (c5)

m1 m2 m3

15

# Binding site “fingerprints” (points with 4 dimensions)

## Points

ElectroShape
$(x_1, x_2, x_3, q)$

ElectroShape for $z_1$
$(x_1, x_2, x_3, z_1)$

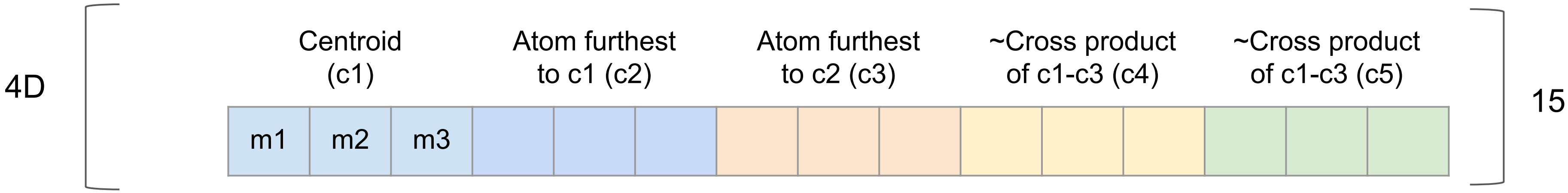
| | |
|---|---|
| $\mathbf{c1}$ | geometric centre, $\mathbb{R}^4$ |
| $\mathbf{c2}$ | atom furthest from $\mathbf{c1}$, $\mathbb{R}^4$ |
| $\mathbf{c3}$ | atom furthest from $\mathbf{c2}$, $\mathbb{R}^4$ |

| | |
|---|---|
| $\mathbf{a} = \mathbf{c2} - \mathbf{c1}$ | $\mathbf{a_s}$ only spatial part, $\mathbb{R}^3$ |
| $\mathbf{b} = \mathbf{c3} - \mathbf{c1}$ | $\mathbf{b_s}$ only spatial part, $\mathbb{R}^3$ |

$$\mathbf{c_s} = \left(\frac{\|\mathbf{a}\|}{2}\right) \frac{\mathbf{a_s} \times \mathbf{b_s}}{\|\mathbf{a_s} \times \mathbf{b_s}\|}$$

$$\mathbf{c4} = \mathbf{c1_s} + \mathbf{c_s} + (0, 0, 0, \mu q_+)$$
$$\mathbf{c5} = \mathbf{c1_s} + \mathbf{c_s} + (0, 0, 0, \mu q_-)$$

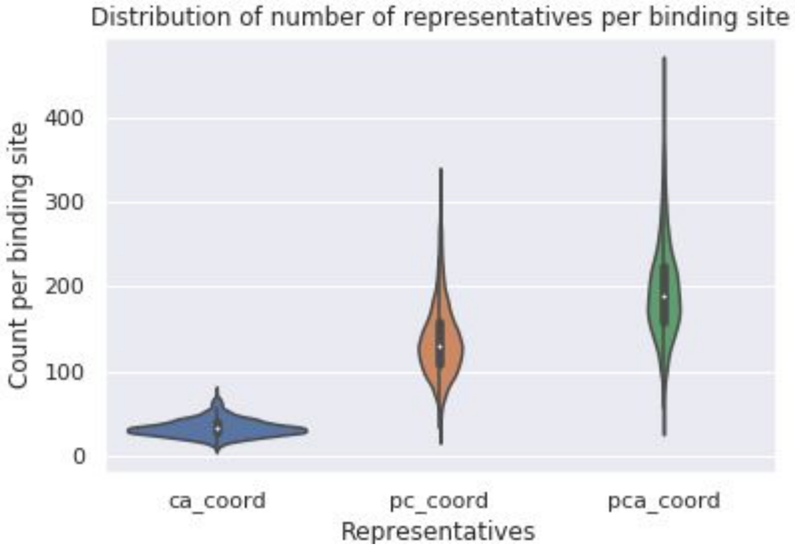
| | |
|---|---|
| $\mu$ | scaling factor |
| $q_{+/-}$ | highest/lowest value of 4th dimension in molecule |

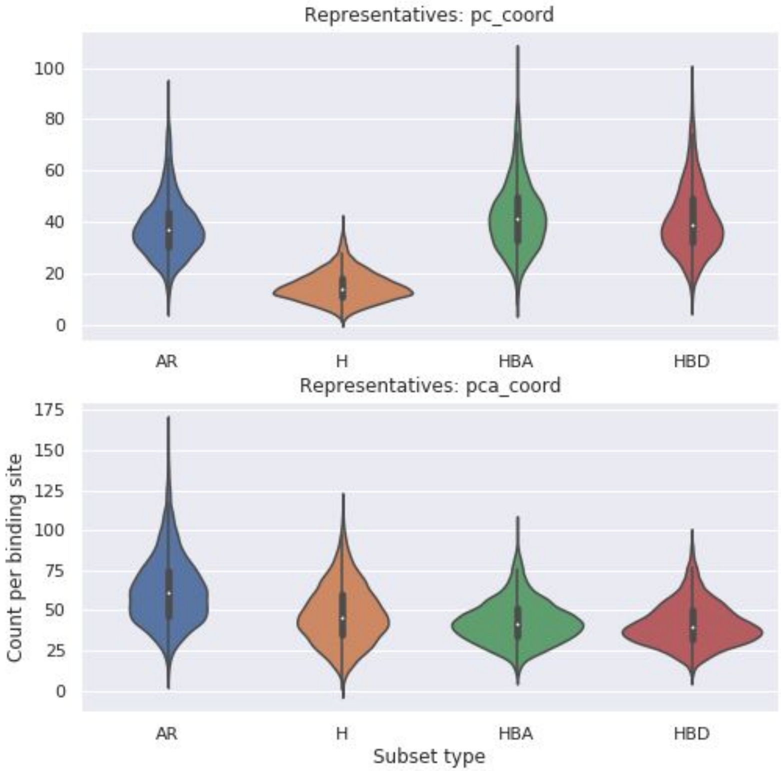


# Cross product

$$\begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} \times \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \begin{pmatrix} a_y b_z - b_y a_z \\ a_z b_x - b_z a_x \\ a_x b_y - b_x a_y \end{pmatrix}$$

# Number of binding site representatives Full scPDB



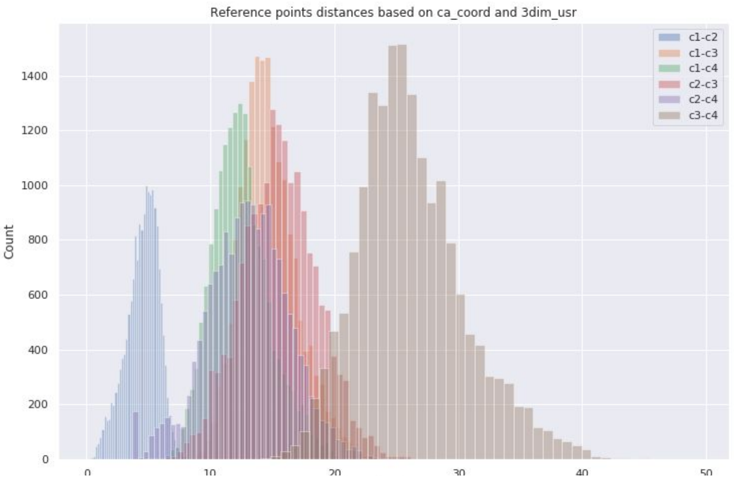


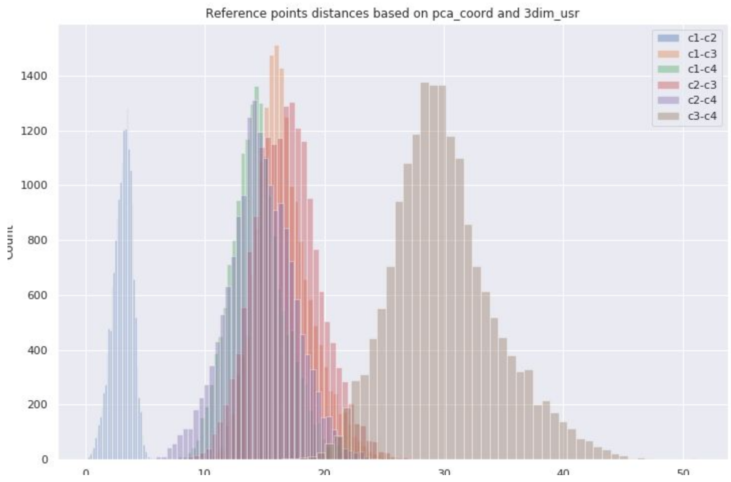
# Reference points distances Full scPDB

Ca atoms

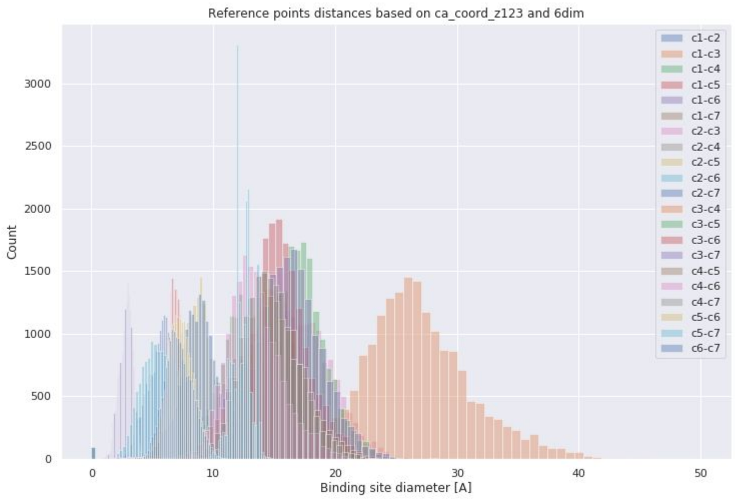
Pseudocenter atoms

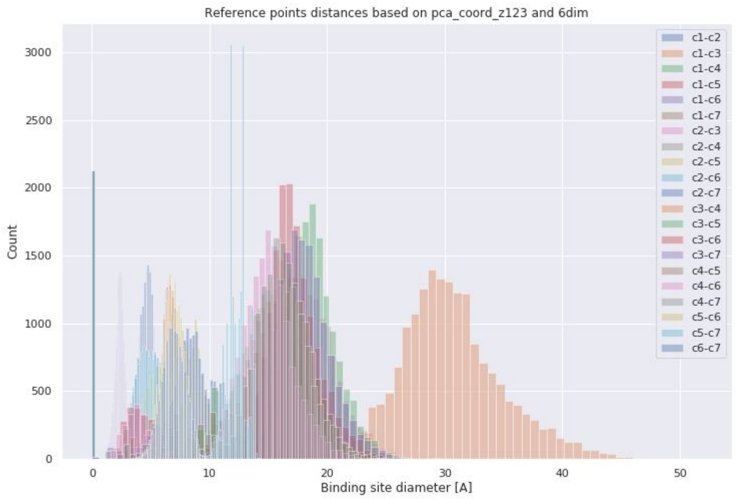
3D (USR)


Reference points distances based on ca_coord and 3dim_usr

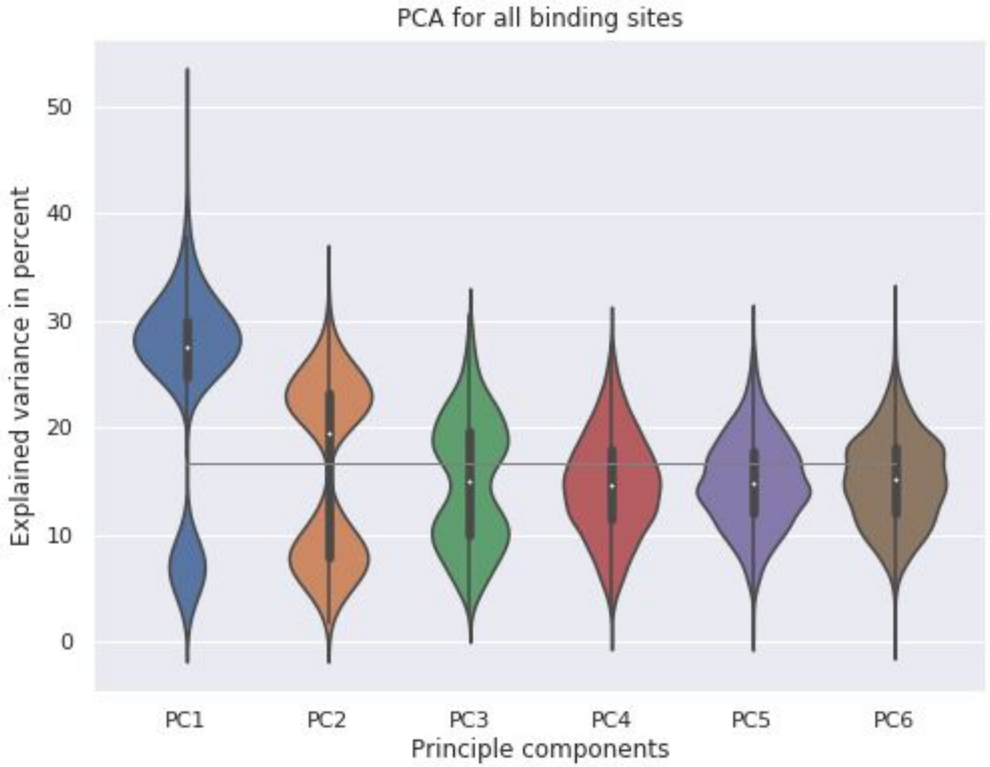

Reference points distances based on pca_coord and 3dim_usr

6D


Reference points distances based on ca_coord_z123 and 6dim


Reference points distances based on pca_coord_z123 and 6dim

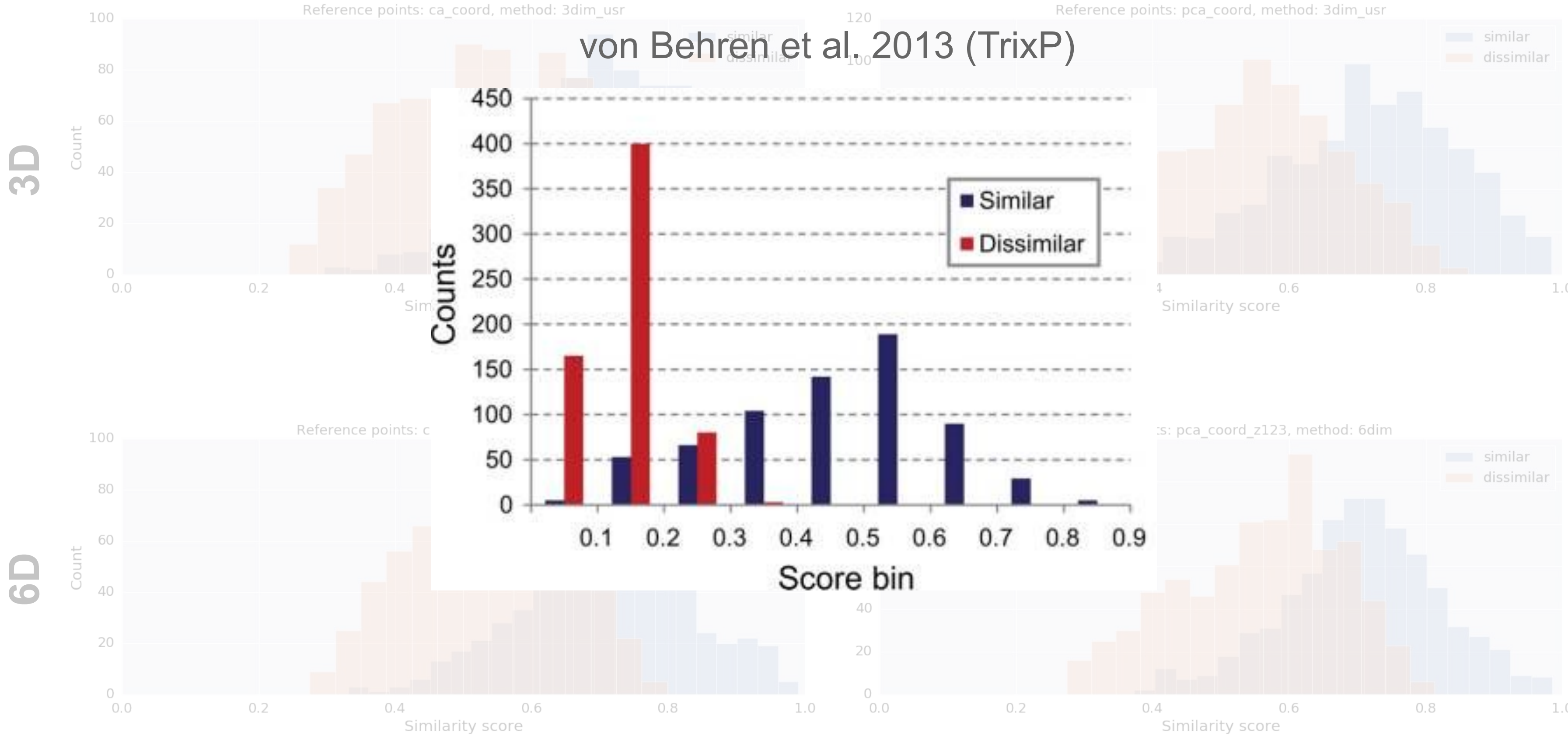
# Encoding: PCA for binding sites (FuzCav: 2061 structures)



# Evaluation: similar vs. dissimilar pairs Weill et al. 2010 (FuzCav)

Ca atoms

Pseudocenter atoms



von Behren et al. 2013 (TrixP)

# Outlook: Introduce subpockets/regions

**Ideas**

- Overlapping/sliding window
  - Adapting von Behren et al. 2013 (TrixP)
- Overlapping subgraphs
  - Adapting Konc et al. 2010 (ProBis)
- Triangulation/Voronoi
  - Adapting Lindow et al. 2011
- Density-based clustering of binding site atoms
  - Adapting Oliver Lempke (group of Bettina Keller)