CREDO: A Structural Interactomics Database For Drug Discovery

Adrian Schreyer

Department of Biochemistry, University of Cambridge

September 27, 2012

What is CREDO?

(Very) brief summary

- Contains all interactions between molecules found in experimentally-determined **biological assemblies**
- Also contains intramolecular interactions of these molecules
- Contacts are represented as *Structural Interaction Fingerprints* (SIFts)
- Contains a sequence-to-structure mapping to integrate protein sequence data
- External resources are integrated to annotate data in CREDO
- Complete cheminformatics toolkits (OpenEye, RDKit)
- Python Application-Programming Interface (API)

Database statistics

From CREDO release 2012.6

- 79,938 PDB entries
- 104,620 biological assemblies
- 497,403 protein-ligand interactions
- 227,961 protein-protein interfaces, 13,611 protein-nucleic acid grooves
- 20 carbohydrate chains!
- **976,282,974** contacts

Tools used for CREDO

CREDO Tools

- `credovi`: program to create and manage CREDO
- `credoscript`: Python database API
- `credimus`: web interface & RESTful web service

My PostgreSQL extensions

- `pgopeneye`: chemical cartridge using the OpenEye toolkits
- `pgeigen`: numerical extension based on the Eigen library
- `ptree`: extension of the ltree module to mimic the PyMOL selection macros

Data Model used in CREDO

New CREDO data model

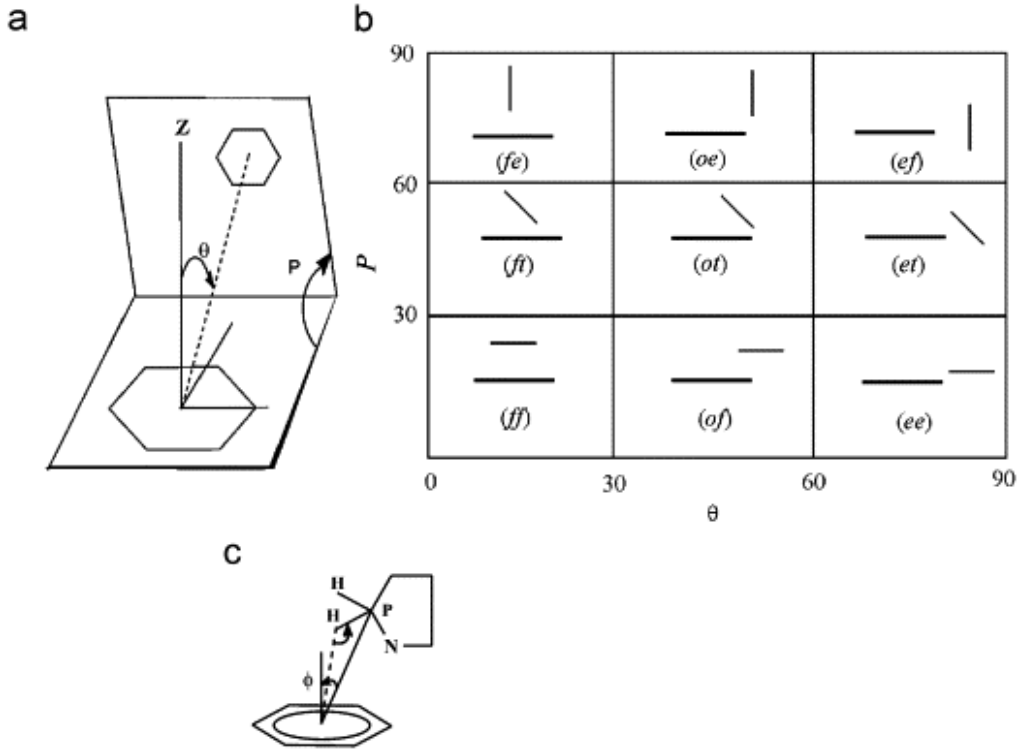
- Since all structural data is retained the resulting data model is simple
- Consist of *Structure*s, *Biomolecule*s, *Chain*s, *Residue*s, *Atom*s, *Contacts* at its core
- **Everything else is just a subset of one of these**
- Example: a ligand is now a set of residues

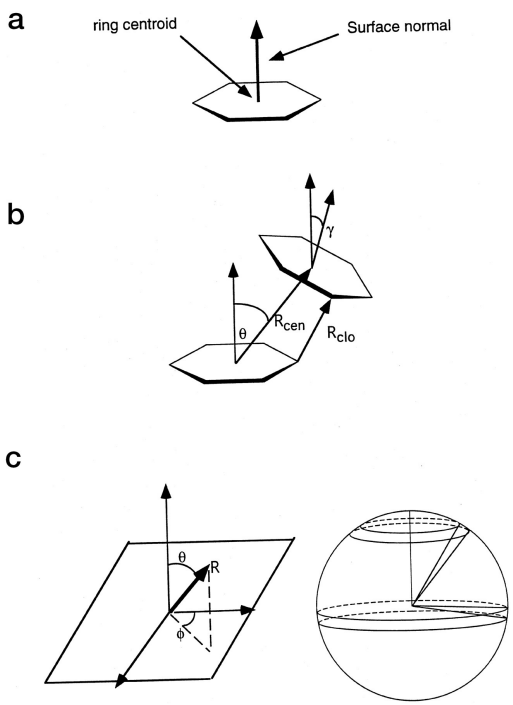
Structural Interaction Fingerprints (SIFts)

Atom and contact types

- Atom types are identified using SMARTS patterns
- Contact types are assigned based on a combination of atom types and geometrical constraints which have to be fulfilled
- Charges (ionisation states) are not required to determine ionic contacts
- Multiple contact types possible but at least one type must be present
- **12** interatomic interaction types
- **9** ring-ring interaction geometries
- **4** ring-atom interaction types

Aromatic ring interaction geometries





Atom-aromatic ring interactions

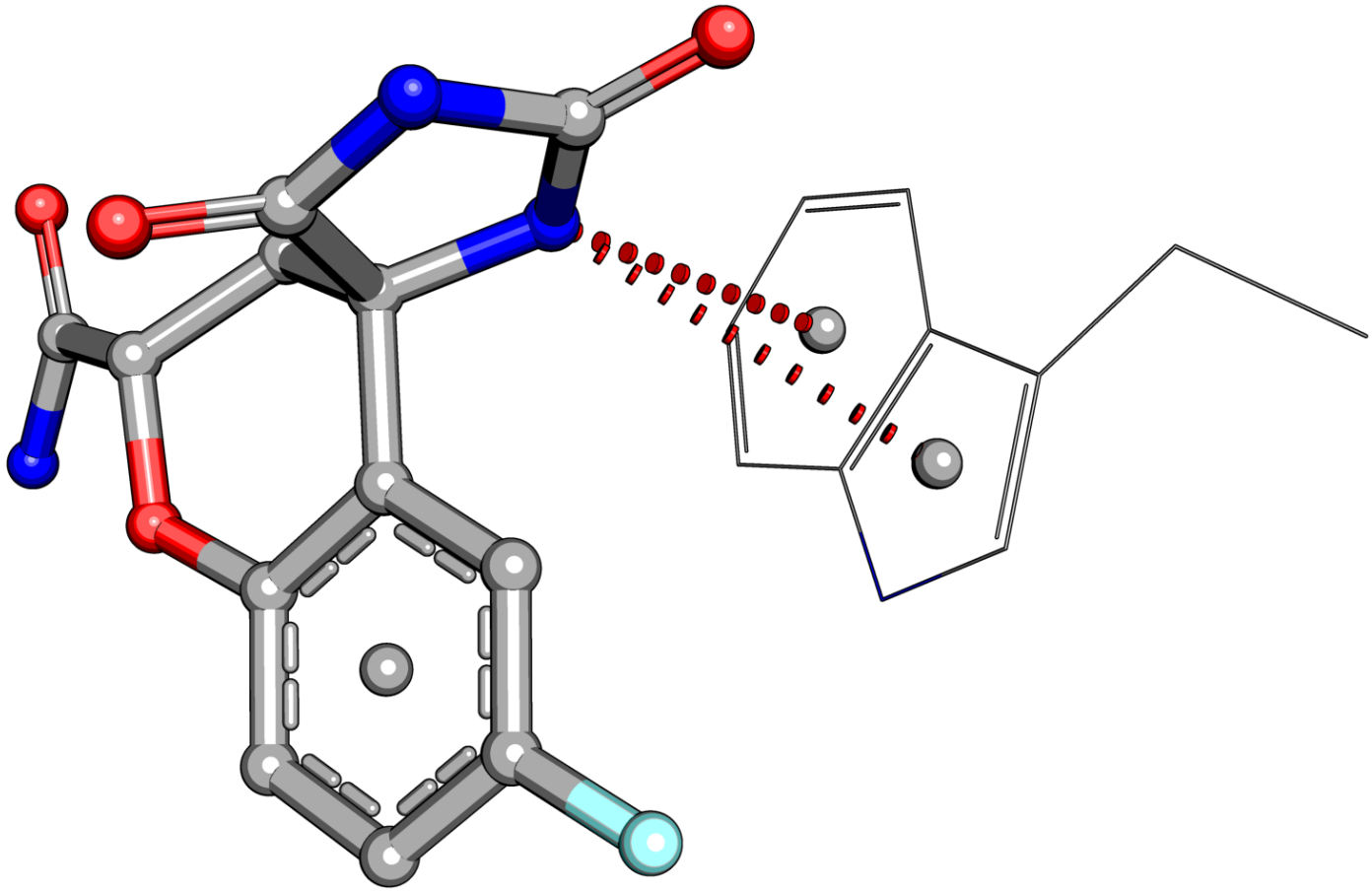
pi-electrons as atom type

- Delocalised π-electron cloud of aromatic ring systems creates negative charge on both faces
- Can act as hydrogen bond acceptor and negatively ionisable group
- Distance- and geometry-dependent

Interaction types

- π-donor: with hydrogen bond donors
- π-cation: with positively ionisable groups
- π-carbon: with weak hydrogen bond donors
- π-halogen: weak hydrogen bonds with halogens in a *head-on* orientation

Atom-aromatic ring interactions



Human aldose reductase mutant V47I complexed with fidarestat (PDB entry: 2PD9)

Data validation in CREDO

Structural properties

- All atomic data is retained (*b-factors*, *occupancies*)
- Boolean flags to identify missing/disordered/clashing residues and atoms
- Boolean flags to identify non-standard, modified and mutated amino acids
- Additional properties from mmCIF: *resolution*, *r-factor*, *r-free*, *pH*
- Ligand geometry (angles) can be problematic

Data validation in CREDO

Diffraction-component precision index (DPI)

- Introduced by *Cruickshank* to estimate the uncertainty of atomic coordinates obtained by structural refinement of protein diffraction data
- Introduced to the virtual screening community by *Goto*

Goto's formula to calculate DPI

$$\sigma(r, B_{avg}) = 2.2 N_{atoms}^{1/2} V_a^{1/2} N_{obs}^{-5/6} R_{free}$$

Goto's formula to calculate theoretical DPI limit

$$\sigma(r, B_{avg}) = 0.22(1+s)^{1/2} V_m^{-1/2} C^{-5/6} R_{free} d_{min}^{5/2}$$

Cross-references to other databases

Linking CREDO to external databases

- Required to annotate structural data
- Also important as entry point into the database
- Can be associated with every CREDO entity: Structure, Chain, ChemComp,...
- Databases include *UniProt*, *EnsEMBL*, *ChEMBL* and many others

Sequence-to-Structure Mapping

Structure integration with function, taxonomy and sequence (SIFTS) initiative

- Maps UniProt sequences onto PDB residue sequences
- Provides further residue level annotation from the *IntEnz*, *GO*, *Pfam*, *InterPro*, *SCOP*, *CATH* and *Pubmed* databases
- Transformed into relational format and linked to all residues in CREDO

Structural Variations

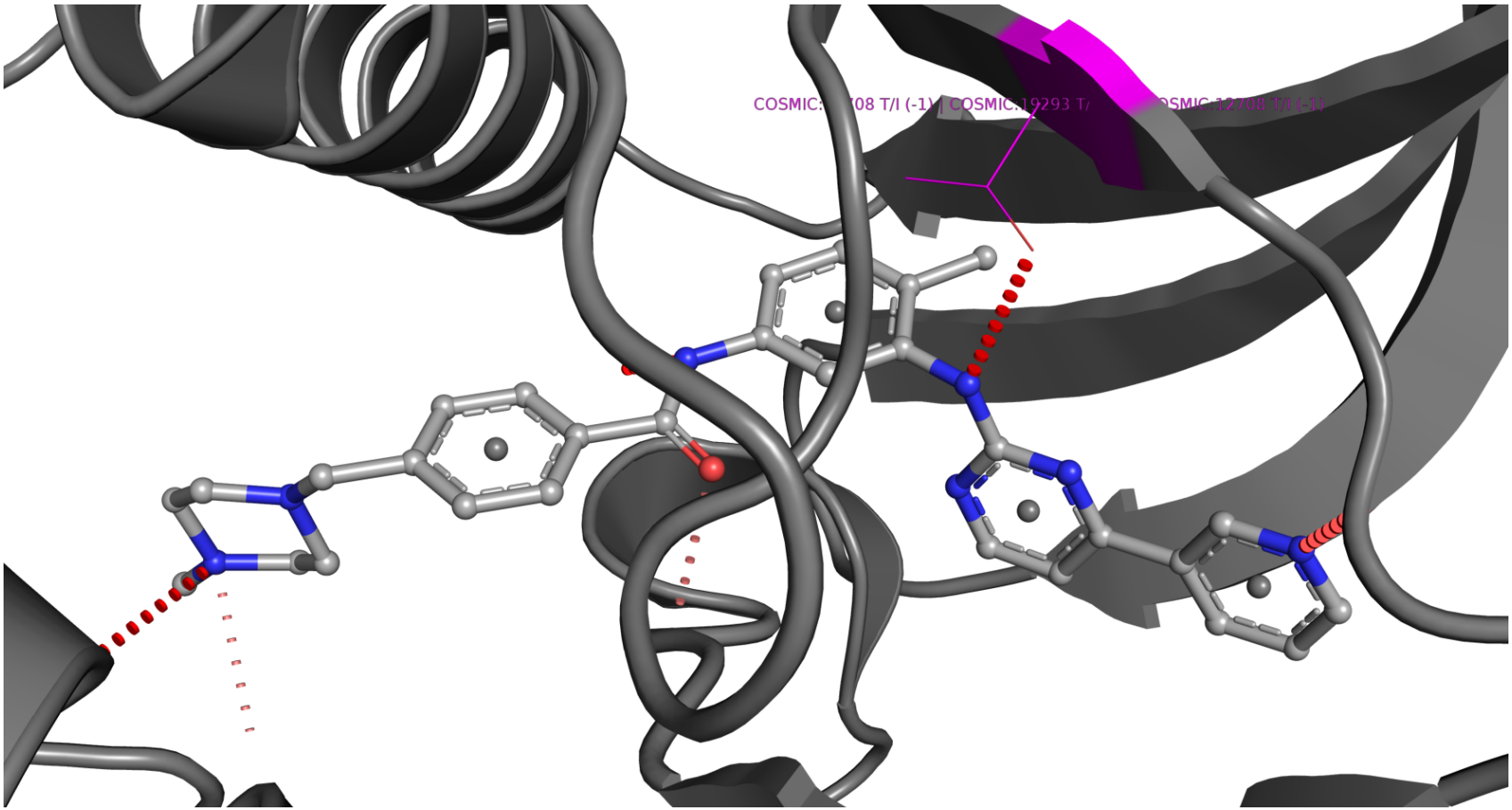
Identifying variations in protein structures

- Mapped onto residues in CREDO through sequence-to-structure mapping
- Can be easily queried and combined with other parameters
- Linked to EnsEMBL **phenotypes**

Source databases included in EnsEMBL Variation

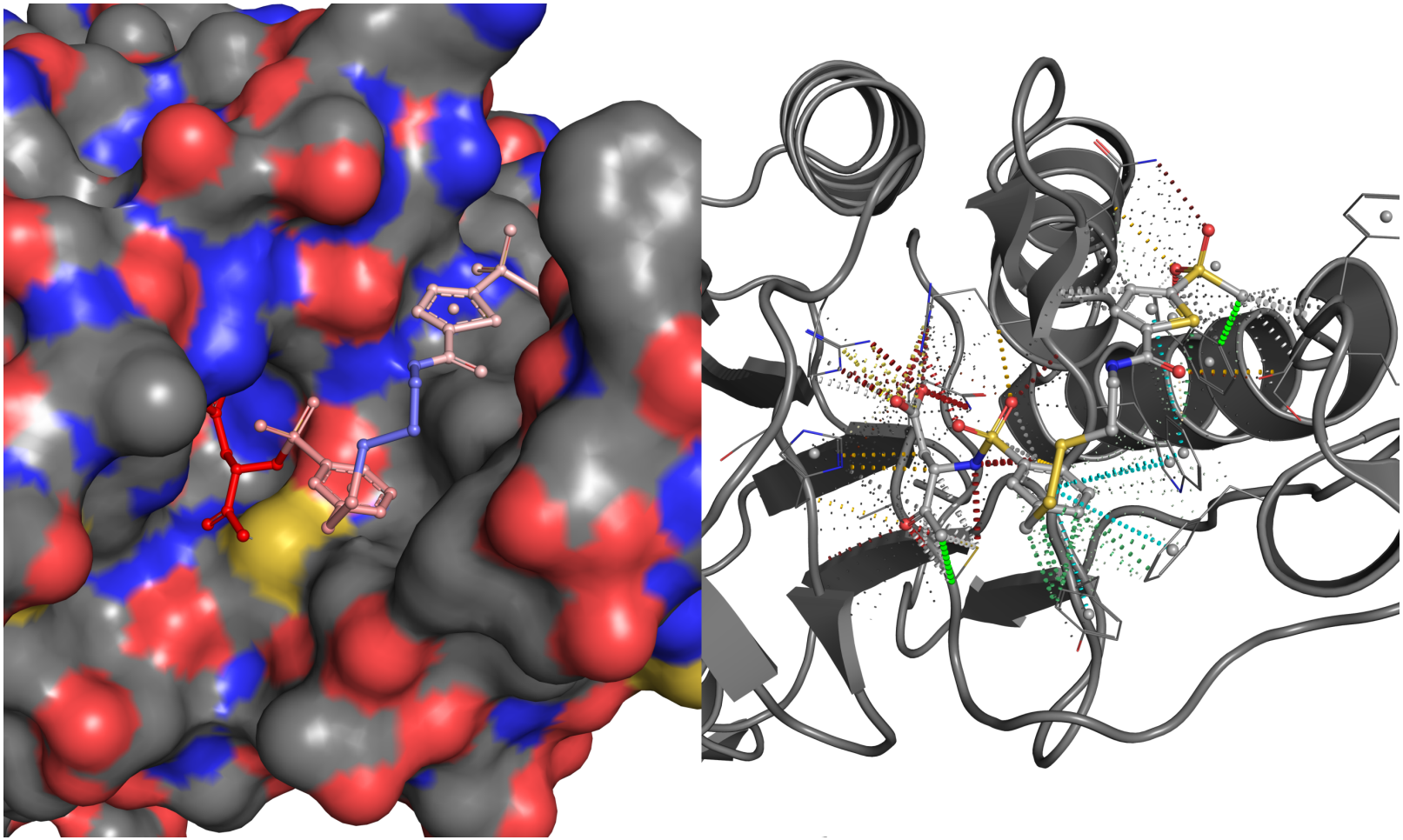
- *dbSNP*
- Catalogue Of Somatic Mutations In Cancer (*COSMIC*)
- Online Mendelian Inheritance in Man (*OMIM*)
- 1000 Genomes

Structural Variations



C-KIT tyrosine kinase in complex with Imatinib (PDB entry: 1T46) with T670I Imatinib-resistant mutation.

Visualisation with PyMOL using the credoscript API



Cysteine aspartyl protease-3 (caspase-3) in complex with a non-peptidic inhibitor

CREDO on the web

Web interface

- Can be used to browse & search the most important data in CREDO
- Protein-ligand complexes can be visualised directly, including visualisation of contacts and highlighting of mutations (WebGL through a fork of GLMol)

RESTful Web service

- Most resources of the service support can be queried programmaticly through GET or POST requests
- Can be used to run `credovi` on uploaded structures (local users only)

pgopeneye: database cartridge for cheminformatics

Cheminformatics extension based on the OpenEye toolkits

- Implements commonly used cheminformatics routines
- Substructure, topological similarity, SMARTS, Murcko scaffolds, etc.
- Supports I/O of SMILES, SDF, OEB, IUPAC
- Fingerprint similarity metrics use SSE (POPCNT)
- Fingerprints can be indexed (GIST): 1M fingerprints, result in less than 500 ms
- Very fast MCS search: 6500 structures $<$ 90 ms (great with ChEMBL)

USRCAT: real-time USR with pharmacophoric constraints

USRCAT: an extension of USR

- USRCAT is an extension of Ultrafast Shape Recognition (USR) that includes pharmacophoric information into the moments
- Outperforms USR significantly in a virtual screening benchmark (using DUD-E)
- Implemented natively into the database: can be used in any SQL query (limit to specific family | include chemical graph similarity)
- Average screening performance of 5.3M conformers (moments) per second (including sorting)
- Currently used with all PDB chemical components and ZINC drug-like set (12M compounds, 200M+ conformers)

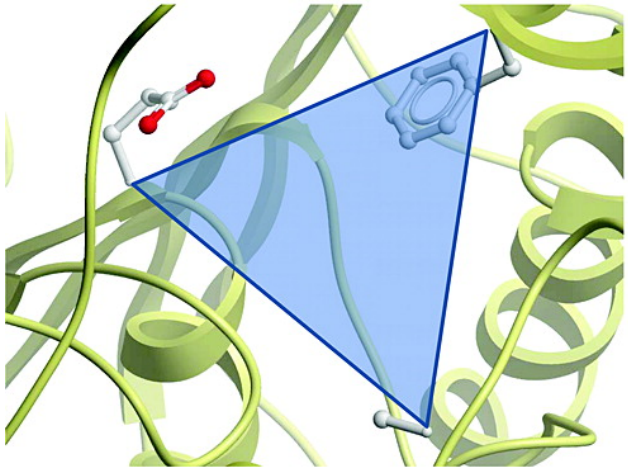
FuzCav: Binding site similarity

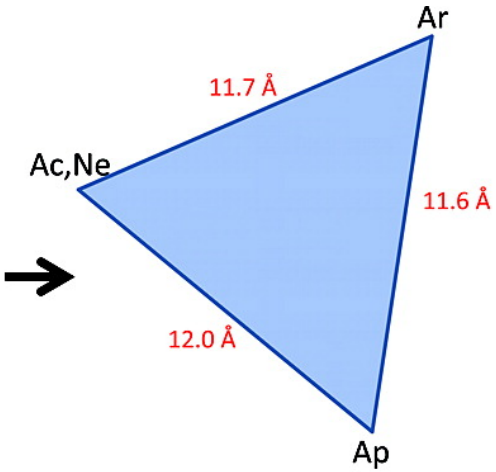
The FuzCav algorithm

- Alignment-free and very easy to calculate
- Based on pharmacophore triplet count to describe a ligand binding site
- Can detect local similarities between binding sites
- Performed *natively* on the server-side with PostgreSQL using numerical extension (`pgeigen`)
- Various similarity metrics can be used
- Calculated for all binding sites in CREDO
- *Ultrafast version is currently work in progress (GIST index for sparse vectors/matrices)*

Journal of Chemical Information and Modeling 2010 50 (1), 123-135

FuzCav: Binding site similarity





6 Pharm features
Ap,Do,Ac,Ar,Po,Ne

5 distance ranges
A: < 4.8 Å; B: 4.8-7.2 Å;
C: 7.2-9.5 Å; D: 9.5-11.9;
E:11.9-14.3 Å.

Triplet	Distances	Box	Count
Ap-Ap-Ap	A-A-A	1	0
Ap-Ap-Do	A-A-A	2	0
...			
Ap-Ar-Ac	E-D-D	4555	1
...			
Ne-Ne-Ne	E-E-E	4833	0

FuzCav Fingerprint: 4833 integers