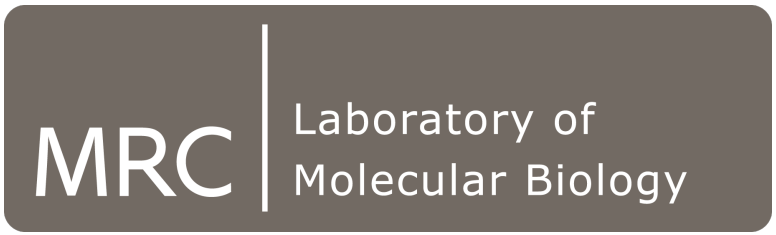
Systematic and Statistical Analysis to Aid Improvement and Validation of ACEDRG

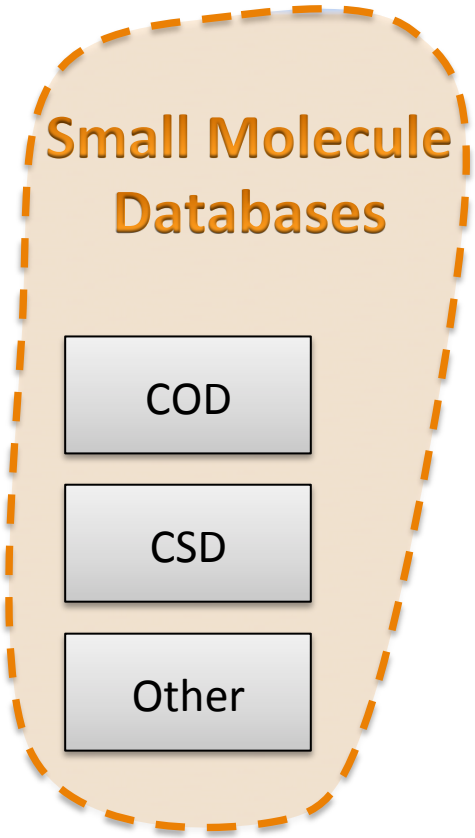
Rob Nicholls
nicholls@mrc-lmb.cam.ac.uk



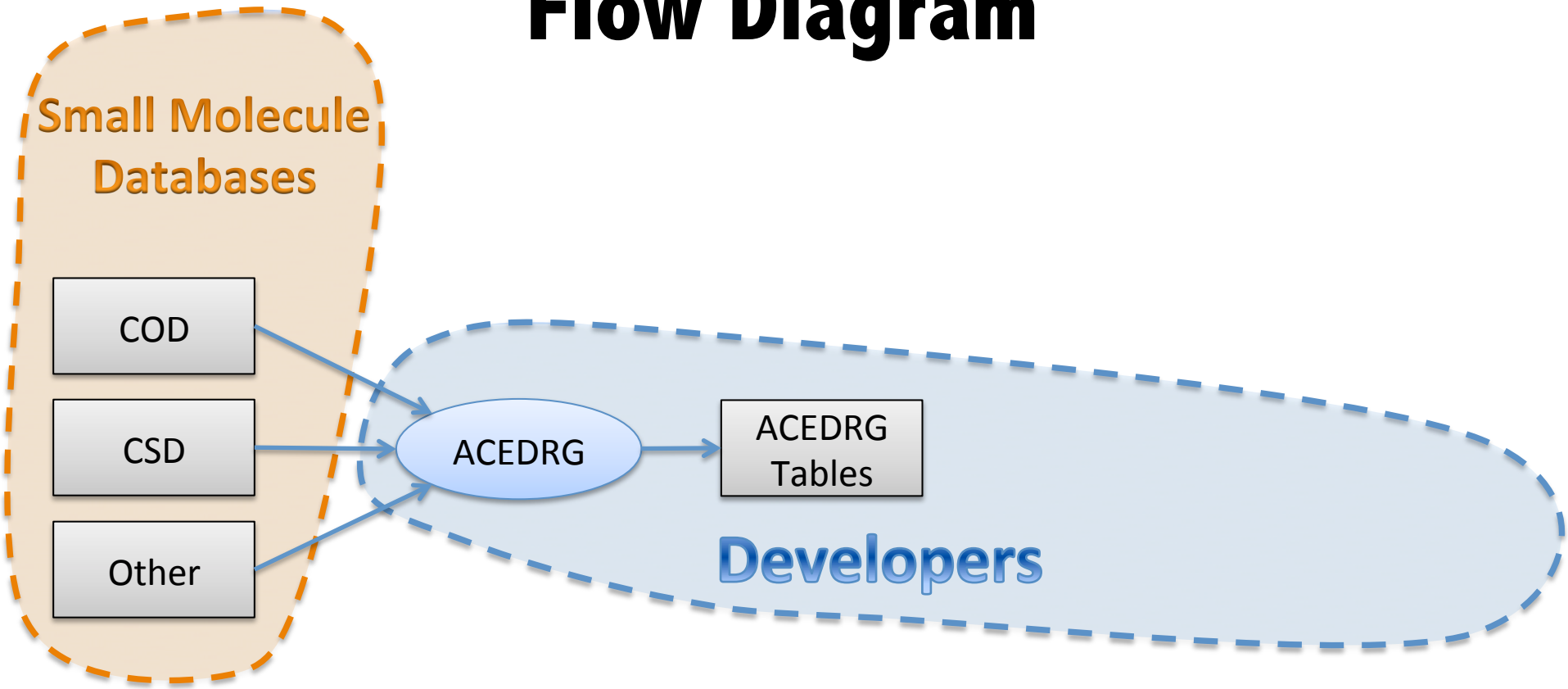
Contents

(1) Flow diagram
(2) Quality control in ACEDRG
(3) Independent systematic validation
(4) Overview of the independent validation tool
(5) Statistical analysis
(6) Discussion

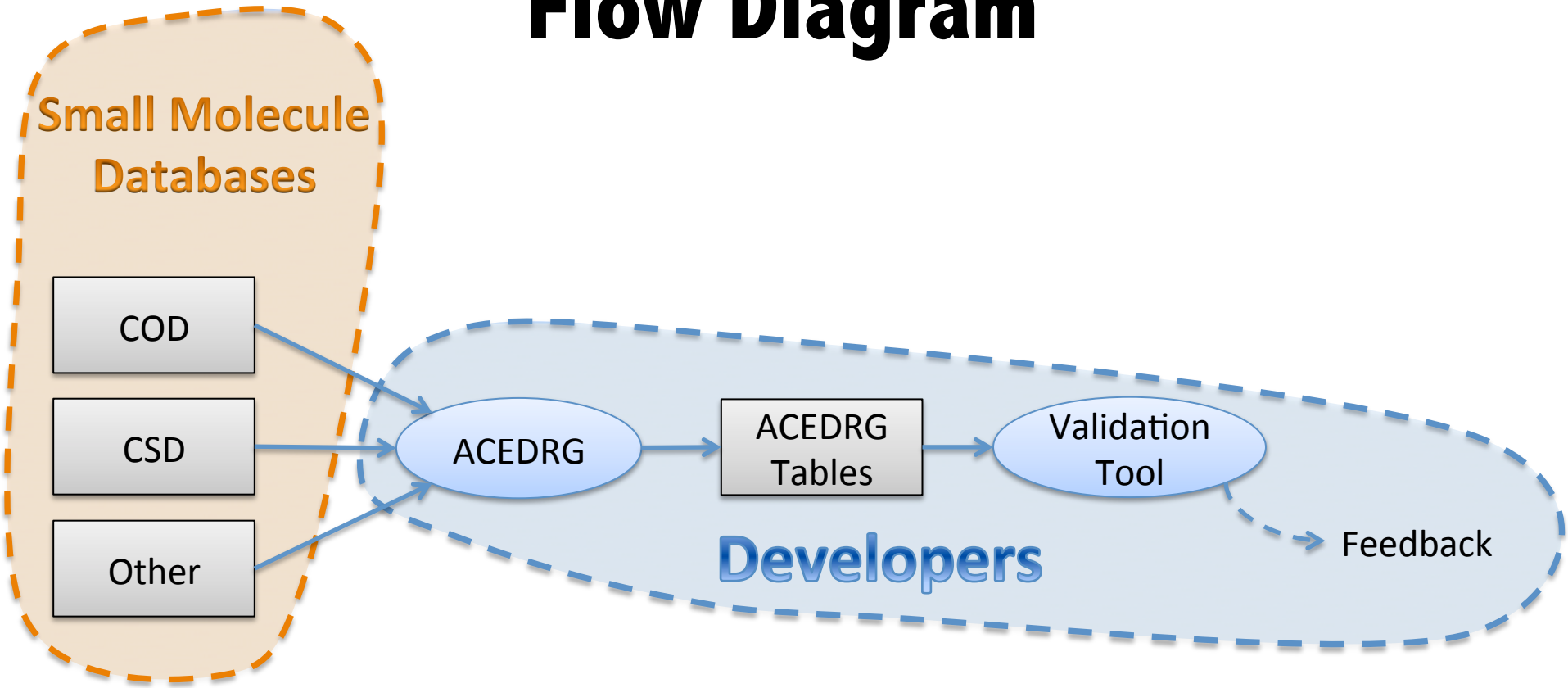
Flow Diagram



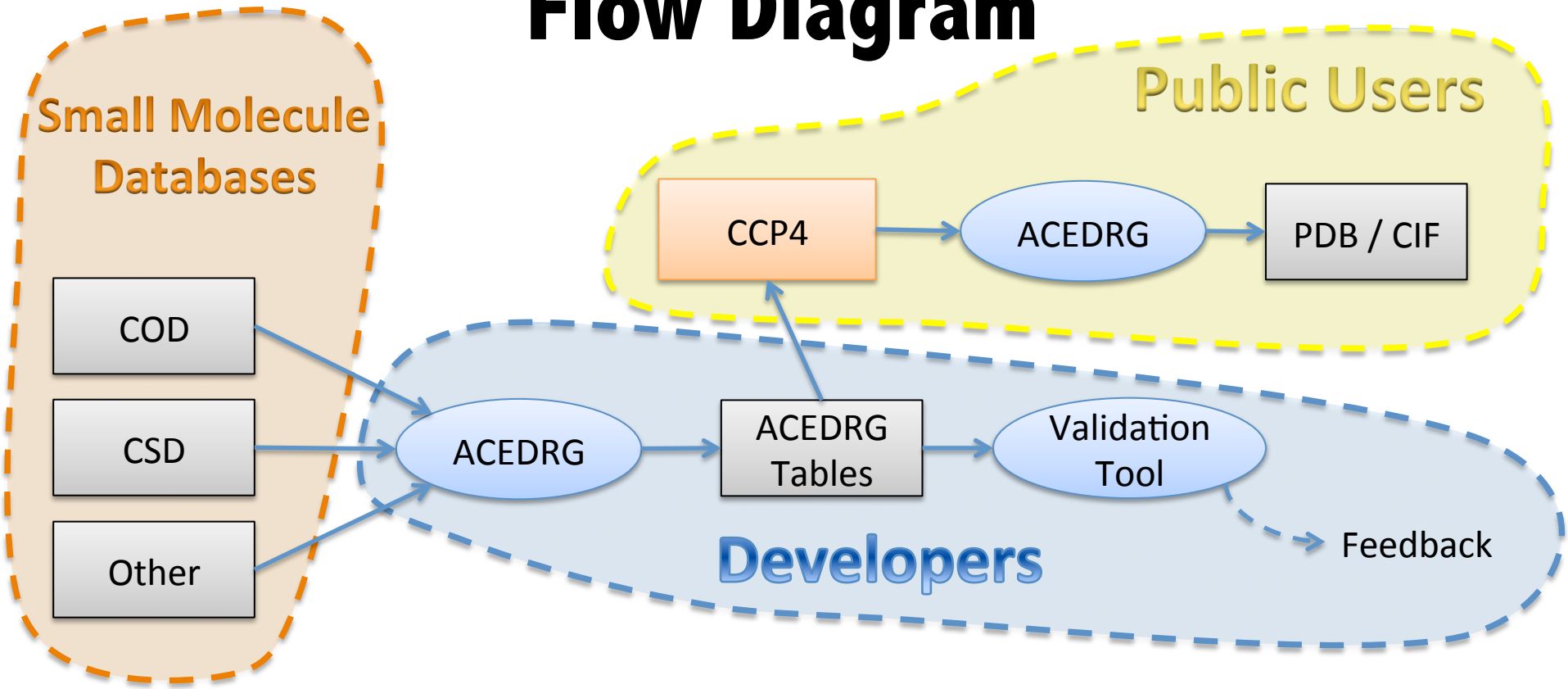
Flow Diagram



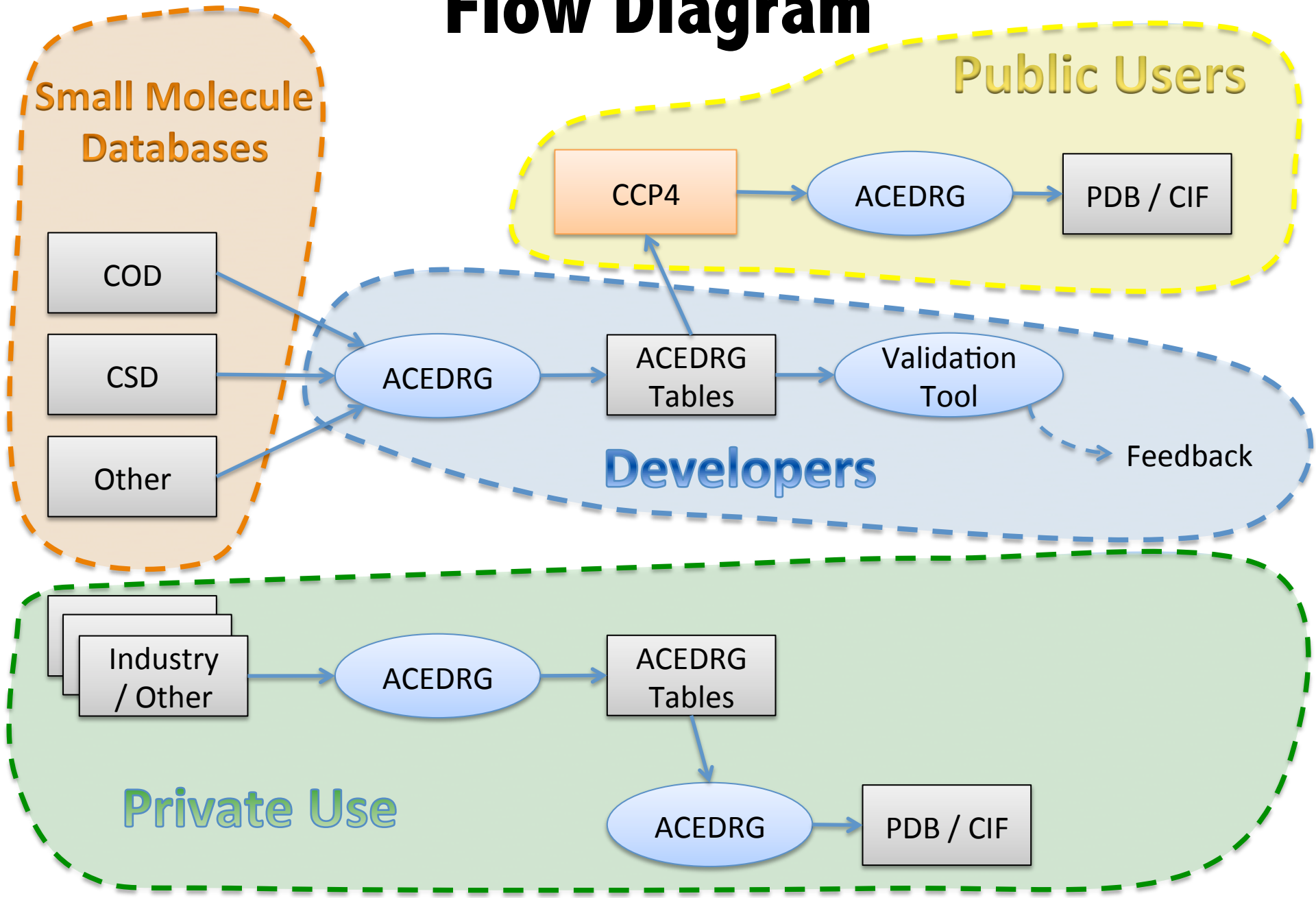
Flow Diagram



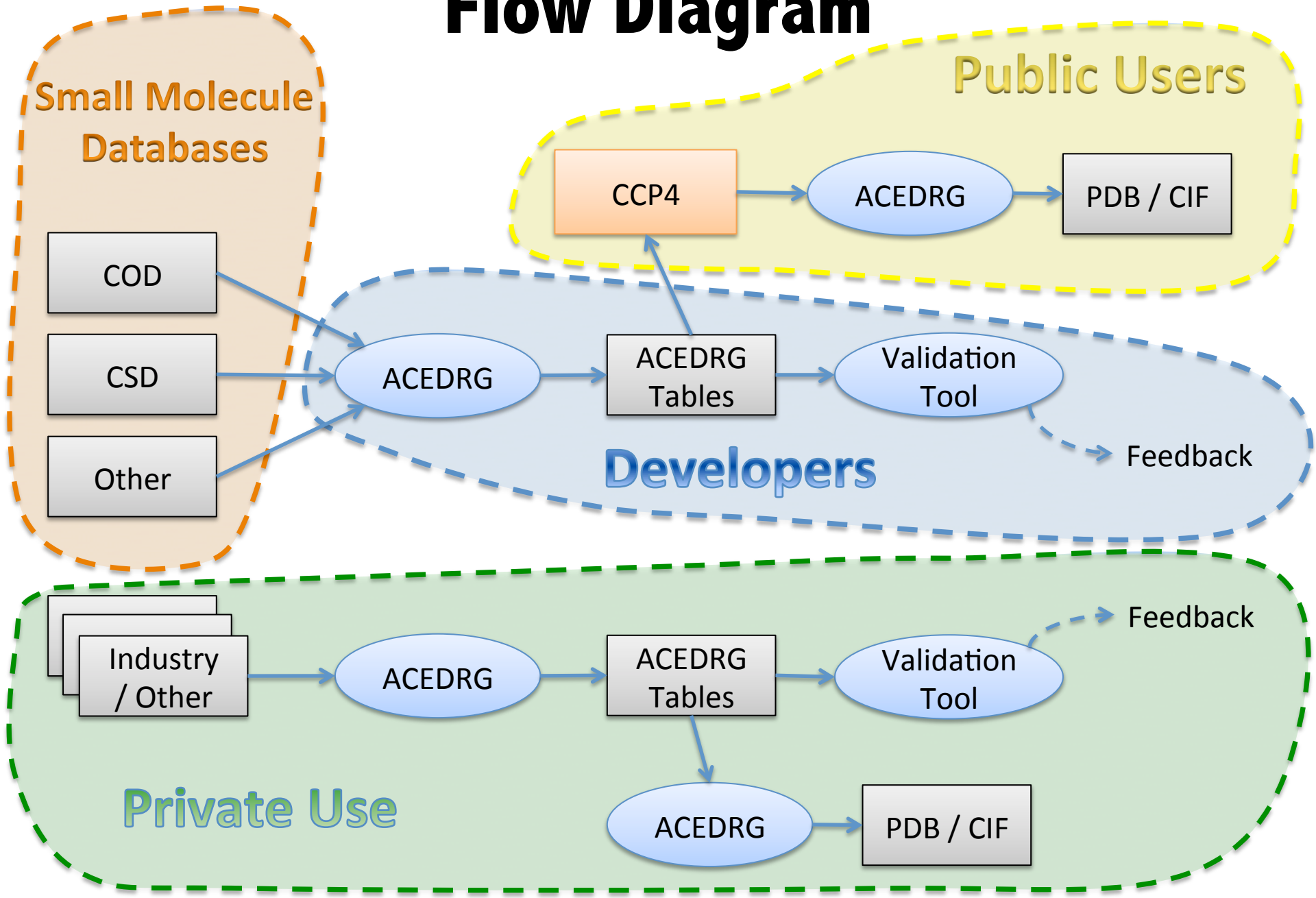
Flow Diagram



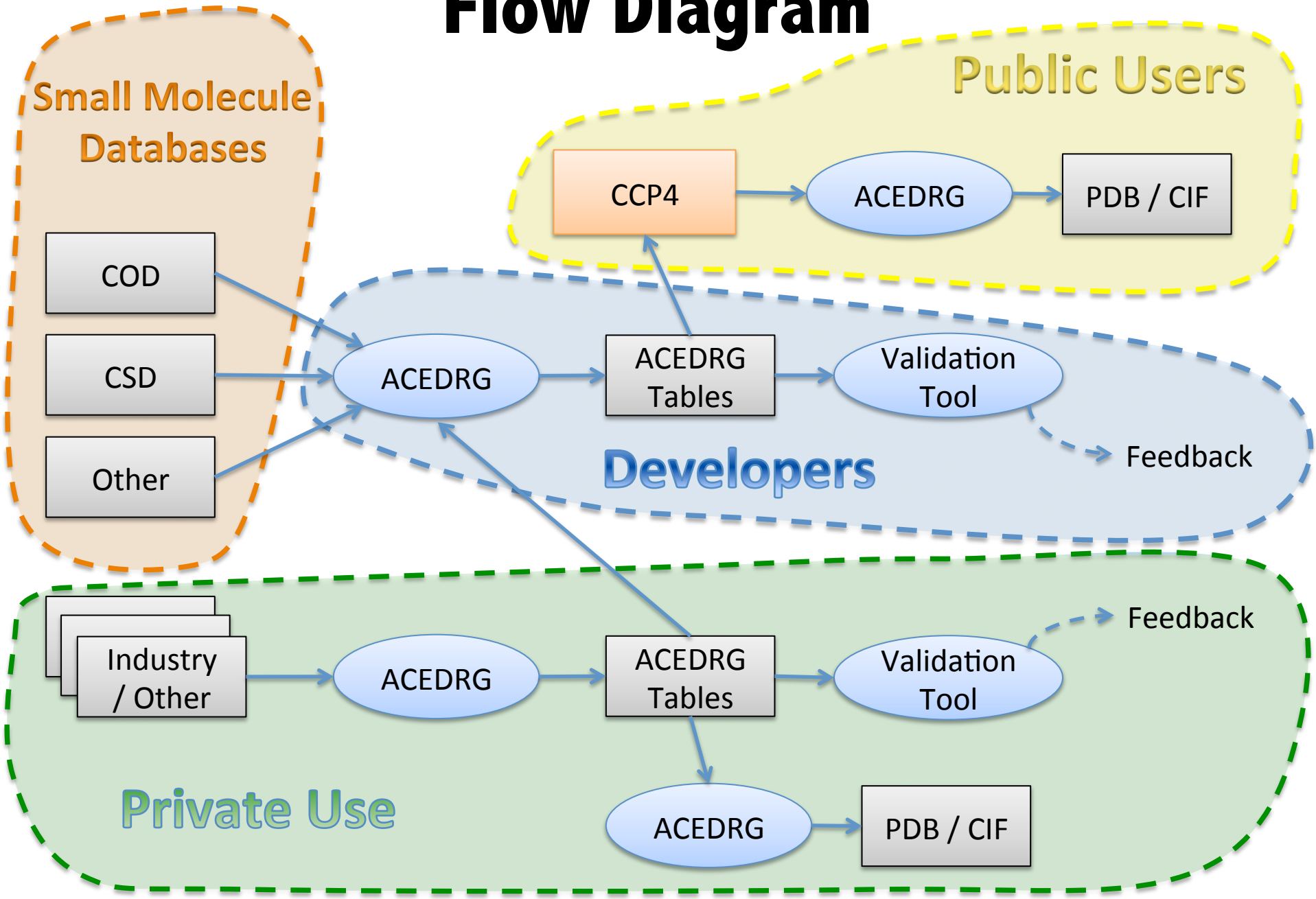
Flow Diagram



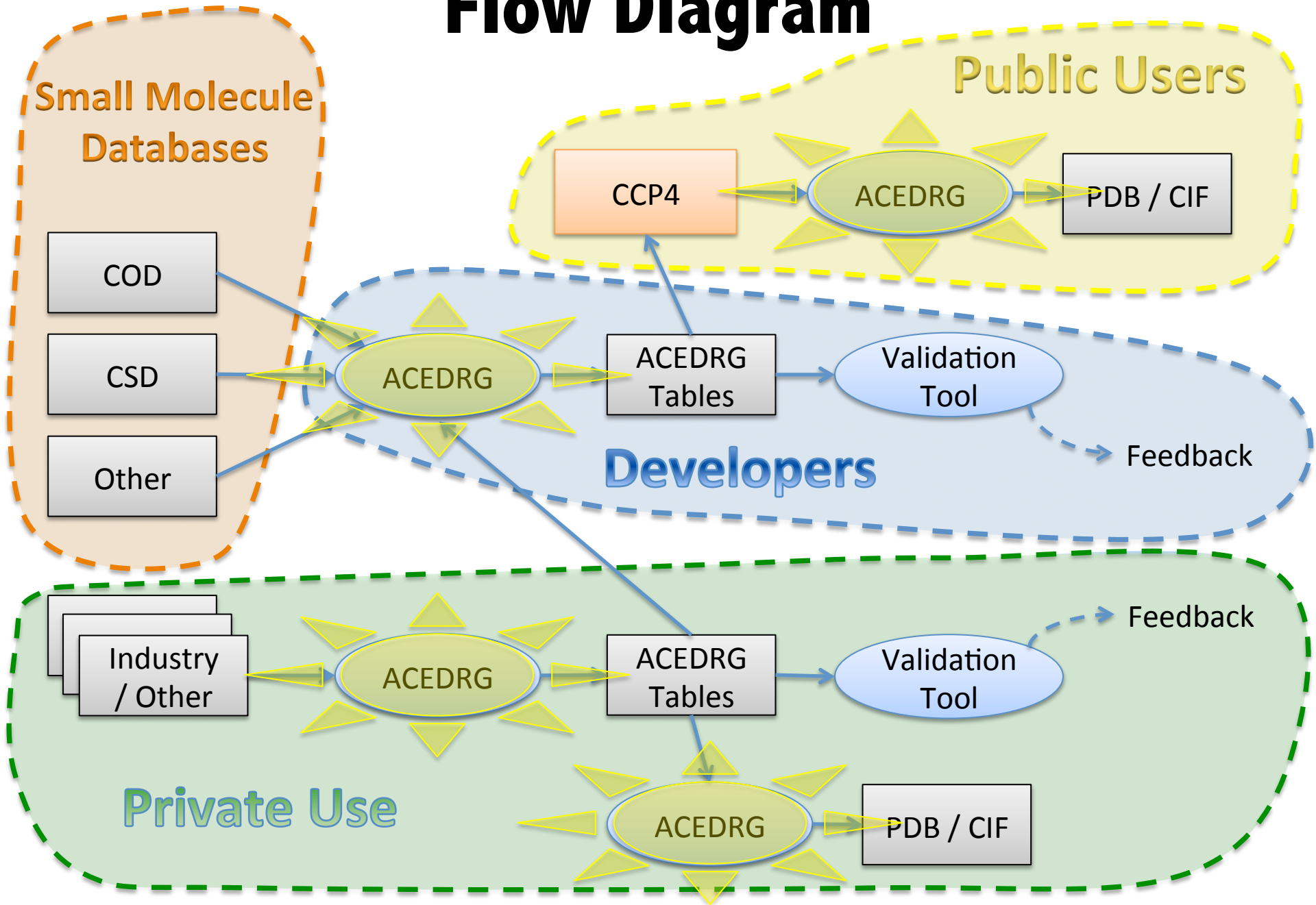
Flow Diagram



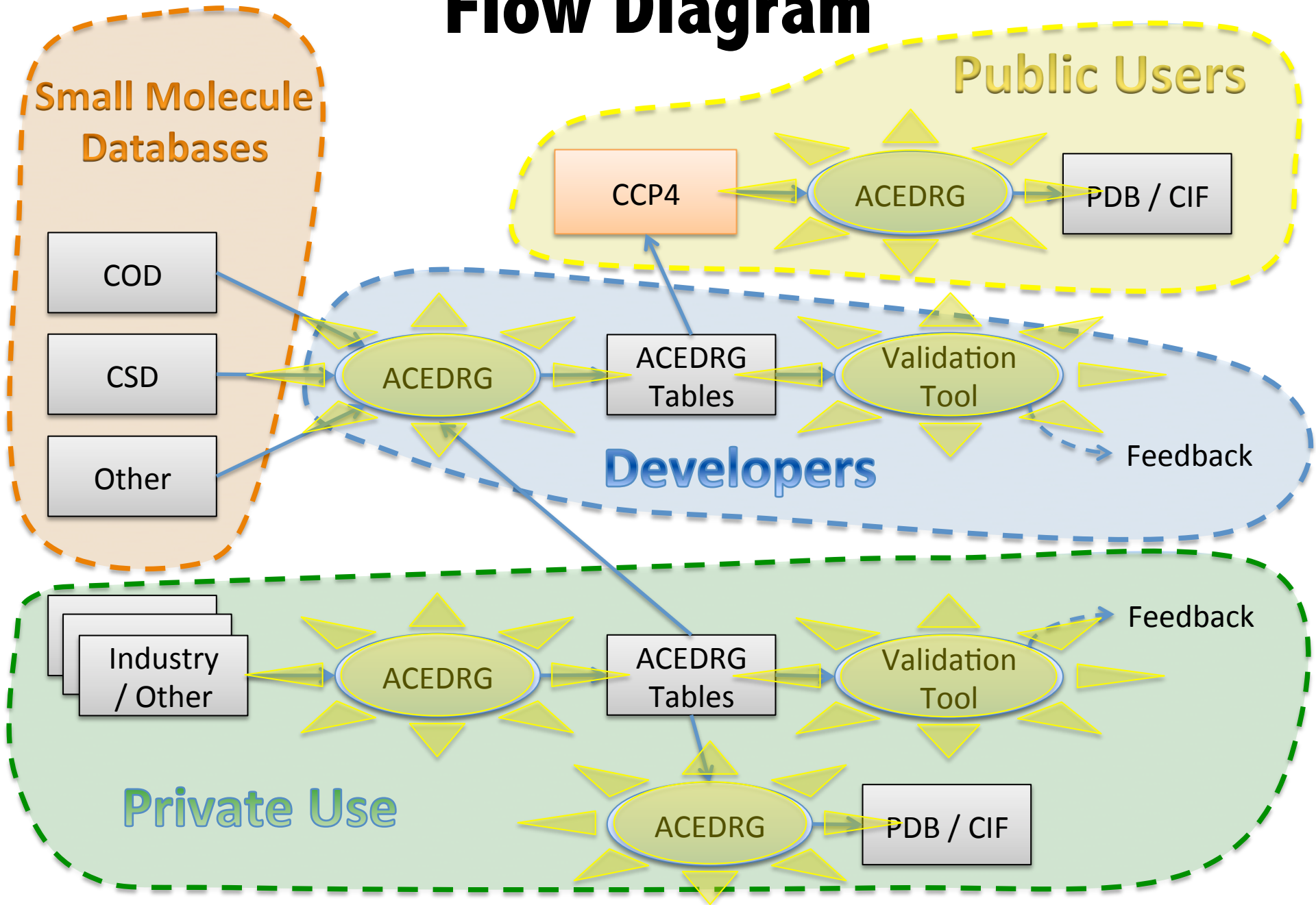
Flow Diagram



Flow Diagram



Flow Diagram



Validation & Quality Control

COD:

Chemistry

ACEDRG:

Chemistry (now takes CIF files from COD, not bond-class tables)

Independent Validation Tool:

Systematic validation
Statistical analysis

Coot:

Final ligand validation - checks consistency against reference
- can use multiple sources, e.g. resource comparison

Quality Control in ACEDRG

(1) Molecule exclusion list (manual)

(2) Automatic molecule exclusion:

- Xtal criteria:
  - Single-crystal x-ray crystallography
  - $R_{obs} < 5\%$, occupancies = 1
- Chemistry validation critieria:
  - Chemistry perception (rather than assume correct in source)
  - Chemistry validation (e.g. reasonable no. bonds, reasonable values)

Single problem in molecule -> exclude whole molecule
(and any covalently bound other molecules, e.g. via a metal)

(3) Measures taken to ensure restraint quality (automatic, on-the-fly):

- Remove bond outliers (z-score)
- Robust estimation(?)

Agree future action point – make criteria available with each version

Independent Systematic Validation

Only for bonds at present
Bond angles, torsions, chiralities, etc. need separate attention

Sanity checking & bug detection

- Internal consistency of bond-observation tables and bond-class tables
  - Resolves any confusion about table versions
- Class table:
  - Uniqueness of classes at various levels of the hierarchy
  - Consistency of atom-class indexing
- Bond observations table:
  - Check for unclassified bonds

Systematic data validation – check criteria, identify peculiarities

- Intra-molecular bond duplications

Identify peculiar molecules that should be removed – provide feedback in order to steer evolution of ACEDRG's molecule exclusion criteria

Independent Systematic Validation

Example – bond duplications within molecules:

Different atom types:

1.47334 C N N[5](C[5]CCH)2(C[5]CO)*N C[5](C[5]CHH)(C[5]CHN)(N[5]CC)(H)*N
1.47335 C C C[5](C[5]CH)(N[5]CC)(O)*Y C[5](C[5]CO)(C[5]NO)(H)*Y

Molecule removed.

Independent Systematic Validation

Example – bond duplications within molecules:

Different atom types:

1.47334	**C**	**N**	**N[5](C[5]CCH)2(C[5]CO)*N**	**C[5](C[5]CHH)(C[5]CHN)(N[5]CC)(H)*N**
1.47335	**C**	**C**	**C[5](C[5]CH)(N[5]CC)(O)*Y**	**C[5](C[5]CO)(C[5]NO)(H)*Y**

Molecule removed.

Same atom types:

1.380630	**C**	**C**	**C[6](C[6]CH)(C[6]CP)(H)*Y**	**C[6](C[6]CH)2(H)*Y**
1.380630	**C**	**C**	**C[6](C[6]CH)2(H)*Y**	**C[6](C[6]CH)2(H)*Y**
1.381280	**C**	**C**	**C[6](C[6]CH)(C[6]CP)(H)*Y**	**C[6](C[6]CH)2(H)*Y**
1.381280	**C**	**C**	**C[6](C[6]CH)2(H)*Y**	**C[6](C[6]CH)2(H)*Y**

Molecule not removed.

However, should >1 representation of a particular atom-pair from any given molecule be present in ACEDRG???

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Reading bond list: class_validate_bonds.txt
Completed in 21.1156 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 8.28468 seconds.
Number of bonds remaining: 5058712

Reading molecule exclusion list: molecules.txt
Filtering bond list...
Completed in 3.98476 seconds.
Number of molecules excluded: 2
Number of bonds remaining: 5058401

Processing bonds...
Completed in 1.95291 seconds.

Validating molecules...
206575 molecules.
Duplicate bond value within tolerance level:
Bond1: 4106551 1.47334 N[5](C[5]CCH)2(C[5]CO)*N C[5](C[5]CHH)(C[5]CHN)(N[5]CC)(H)*N C N
Bond2: 4106551 1.47335 C[5](C[5]CH)(N[5]CC)(O)*Y C[5](C[5]CO)(C[5]NO)(H)*Y C C
Duplicate bond value within tolerance level:
Bond1: 4113596 1.34752 C[5](C[5]CH)(N[5]BN)(H)*Y N[5](N[5]MoC)(C[5]CH)(BHNN)*Y C N
Bond2: 4113596 1.34753 C[5](C[5]HN)2(H)*Y C[5](C[5]CH)(N[5]BN)(H)*Y C C
Duplicate bond value within tolerance level:
Bond1: 4119196 1.35108 C[6](N[6]PdC)(C[6]CH)(H)*Y N[6](C[6]CH)2(PdN3)*Y C N
Bond2: 4119196 1.35109 C[6](C[6]CC)(C[6]HN)(H)*Y C[6](N[6]PdC)(C[6]CH)(H)*Y C C
Duplicate bond value within tolerance level:
Bond1: 7003851 1.34969 C[6](C[6]CH)2(H)*Y C[6](C[6]CH)(C[6]HN)(H)*Y C C
Bond2: 7003851 1.3497 C[6](N[6]RuC)(C[6]CH)(C[6]CN)*Y N[6](C[6]CC)(C[6]CH)(RuN5)*Y C N
Duplicate bond value within tolerance level:
Bond1: 7153156 1.46198 C(NCuCC)(CHHN)(H)2*N N(CuCuNNOO)(CCHH)3*N C N
Bond2: 7153156 1.46199 C[6](C[6,6]CC)(C[6]CH)(H)*Y C[6,6](C[6,6]CN)(C[6]ClC)(C[6]CH)*Y C C
Number of molecules with peculiarities: 5
Validation result: Failure (0.755283 seconds).

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Read class table

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Populate Internal Structure

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Validation of class file data

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Read bond observations table

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Populate Internal Structure

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Perform validation/filtering

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Write bonds file (internal format)

Independent Validation Tool

Reading class lis
Completed in 1.4

Read re-formatted bonds file

Reading bond list: class_validate_bonds.txt
Completed in 21.1156 seconds.
Number of bonds: 10555471

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.1429
Number of hash code
Number of classes:

Exclude hydrogens (optional)

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Reading bond list: class_validate_bonds.txt
Completed in 21.1156 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 8.28468 seconds.
Number of bonds remaining: 5058712

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Su

Exclude molecules (optional)

Reading bond list: all
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond codes...
Completed in 47.0865 seconds.

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Reading bond list: class_validate_bonds.txt
Completed in 21.1156 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 8.28468 seconds.
Number of bonds remaining: 5058712

Reading molecule exclusion list: molecules.txt
Filtering bond list...
Completed in 3.98476 seconds.
Number of molecules excluded: 2
Number of bonds remaining: 5058401

Independent Validation Tool

Reading class list: allOrgBonds.table
Completed in 1.43164 seconds.

Processing classes...
Completed in 5.14298 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.014137 seconds). All hashes are unique.

Reading bond list: allOrgBonds.txt
Completed in 19.0416 seconds.
Number of bonds: 10555471

Processing bonds...
Completed in 1.92403 seconds.

Sorting bonds...
Completed in 18.5706 seconds.

Getting list of unique bond co
Completed in 47.0865 second

Perform data validation

Finding and removing unclassified bonds...
Completed in 354.018 seconds.

Finding and removing bond duplications...
Completed in 1.0439 seconds.

Getting class indexes...
Completed in 58.063 seconds.

Writing bonds file...
Completed in 47.9413 seconds.

Reading bond list: class_validate_bonds.txt
Completed in 21.1156 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 8.28468 seconds.
Number of bonds remaining: 5058712

Reading molecule exclusion list: molecules.txt
Filtering bond list...
Completed in 3.98476 seconds.
Number of molecules excluded: 2
Number of bonds remaining: 5058401

Processing bonds...
Completed in 1.95291 seconds.

Validating molecules...
206575 molecules.
Duplicate bond value within tolerance level:

Bond1: 4106551	1.47334	N[5](C[5]CCH)2C[5]CO)*N	C[5](C[5]CHH)(C[5]CHN)(N[5]CC)(H)*N	C	N
Bond2: 4106551	1.47335	C[5](C[5]CH)(N[5]CC)(O)*Y	C[5](C[5]CO)(C[5]NO)(H)*Y	C	C
Duplicate bond value within tolerance level:					
Bond1: 4113596	1.34752	C[5](C[5]CH)(N[5]BN)(H)*Y	N[5](N[5]MoC)(C[5]CH)(BHNN)*Y	C	N
Bond2: 4113596	1.34753	C[5](C[5]HN)2(H)*Y	C[5](C[5]CH)(N[5]BN)(H)*Y	C	C
Duplicate bond value within tolerance level:					
Bond1: 4119196	1.35108	C[6](N[6]PdC)(C[6]CH)(H)*Y	N[6](C[6]CH)2(PdN3)*Y	C	N
Bond2: 4119196	1.35109	C[6](C[6]CC)(C[6]HN)(H)*Y	C[6](N[6]PdC)(C[6]CH)(H)*Y	C	C
Duplicate bond value within tolerance level:					
Bond1: 7003851	1.34969	C[6](C[6]CH)2(H)*Y	C[6](C[6]CH)(C[6]HN)(H)*Y	C	C
Bond2: 7003851	1.3497	C[6](N[6]RuC)(C[6]CH)(C[6]CN)*Y	N[6](C[6]CC)(C[6]CH)(RuN5)*Y	C	N
Duplicate bond value within tolerance level:					
Bond1: 7153156	1.46198	C(NCuCC)(CHHN)(H)2*N	N(CuCuNNOO)(CCHH)3*N	C	N
Bond2: 7153156	1.46199	C[6](C[6,6]CC)(C[6]CH)(H)*Y	C[6,6](C[6,6]CN)(C[6]ClC)(C[6]CH)*Y	C	C

Number of molecules with peculiarities: 5
Validation result: Failure (0.755283 seconds).

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Read class table

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Populate Internal Structure

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.0139799 seconds). All hashes are unique.

Validation of class file data

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.0139799 seconds). All hashes are unique.

Reading bond list: class_validate_bonds.txt
Completed in 34.8804 seconds.
Number of bonds: 10555471

Read re-formatted bonds file

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.0139799 seconds). All hashes are unique.

Reading bond list: class_validate_bonds.txt
Completed in 34.8804 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 92.3112 seconds.
Number of bonds remaining: 5058712

Exclude hydrogens (optional)

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.0139799 seconds). All hashes are unique.

Reading bond list: class_validate_bonds.txt
Completed in 34.8804 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 92.3112 seconds.
Number of bonds remaining: 5058712

Processing bonds...
Completed in 24.8393 seconds.

Link bonds to bond-classes

Independent Validation Tool

Subsequent runs…

Reading class list: allOrgBonds.table
Completed in 1.75601 seconds.

Processing classes...
Completed in 4.97165 seconds.
Number of hash codes: 1452
Number of classes: 465367

Validating uniqueness of hashes...
Validation result: Success (0.0139799 seconds). All hashes are unique.

Reading bond list: class_validate_bonds.txt
Completed in 34.8804 seconds.
Number of bonds: 10555471

Excluding hydrogens...
Completed in 92.3112 seconds.
Number of bonds remaining: 5058712

Processing bonds...
Completed in 24.8393 seconds.

Calculating bond class statistics...

Writing bond class statistics file: bond_class_stats.txt
Completed in 9.53772 seconds.

Statistical analysis

Job completed in 168.311 seconds.

Statistical Analysis

Analysis of ACEDRG bond classes

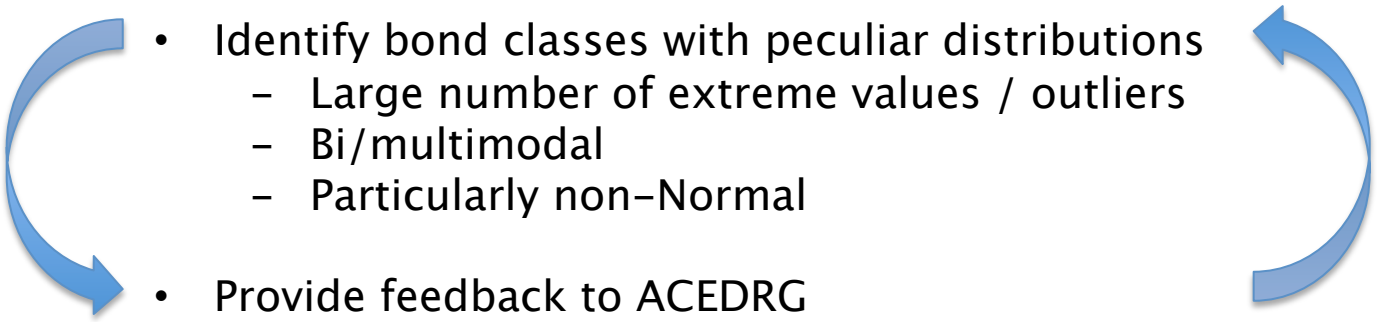
Objectives:

- Identify bond classes with peculiar distributions
  - Large number of extreme values / outliers
  - Bi/multimodal
  - Particularly non-Normal
- Provide feedback to ACEDRG
  - Improve ACEDRG atom classification
  - Improve ACEDRG molecule/bond validation/exclusion criteria
  - Guide future design and protocol

Statistical Analysis

Analysis of ACEDRG bond classes

Objectives:



- Identify bond classes with peculiar distributions
  - Large number of extreme values / outliers
  - Bi/multimodal
  - Particularly non-Normal
- Provide feedback to ACEDRG
  - Improve ACEDRG atom classification
  - Improve ACEDRG molecule/bond validation/exclusion criteria
  - Guide future design and protocol

- Future: assessing suitability of combining of bond classes up the hierarchy

Statistical Analysis

Analysis of ACEDRG bond classes

Objectives:

- Identify bond classes with peculiar distributions
  - Large number of extreme values / outliers
  - Bi/multimodal
  - Particularly non-Normal

One approach – look at moments:

- Standard deviation – is bond uncertainty reasonable?
- Skewness – distribution symmetry; are there extreme outliers?
- Kurtosis – heavy tails? Short tails – indicator of multimodality.

Statistical Analysis

Analysis of ACEDRG bond classes

Objectives:

- Identify bond classes with peculiar distributions
  - Large number of extreme values / outliers
  - Bi/multimodal
  - Particularly non-Normal

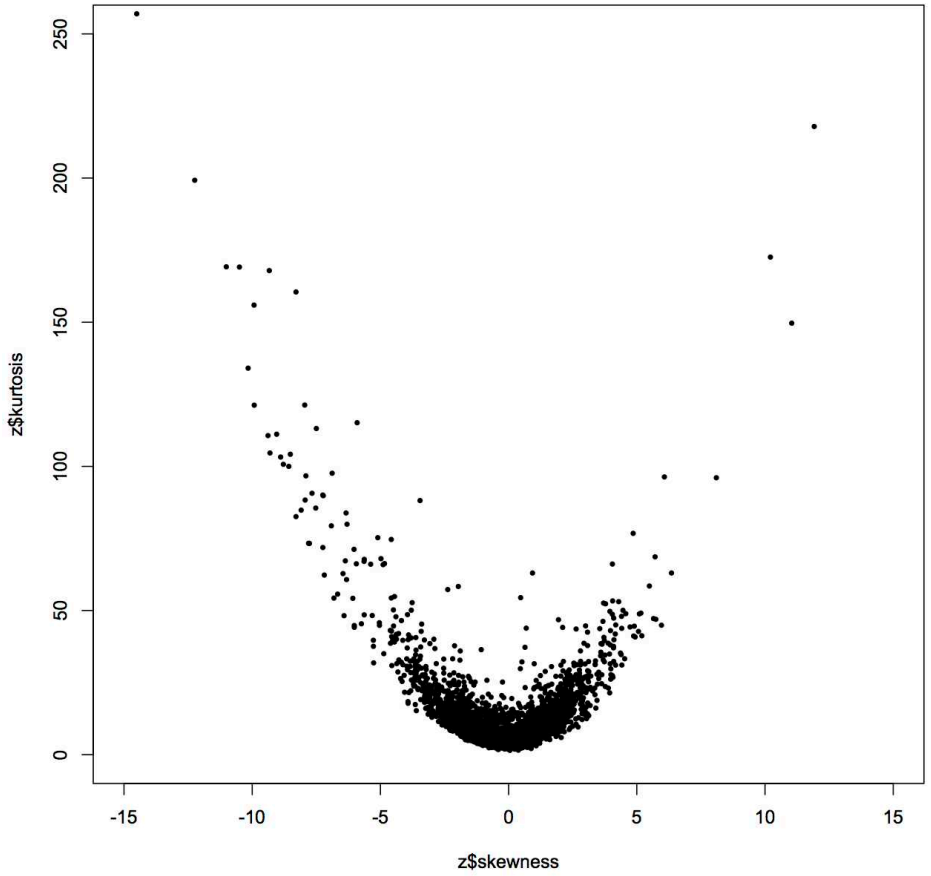
One approach – look at moments:

- Standard deviation – is bond uncertainty reasonable?
- Skewness – distribution symmetry; are there extreme outliers?
- Kurtosis – heavy tails? Short tails – indicator of multimodality.

In the presented results:

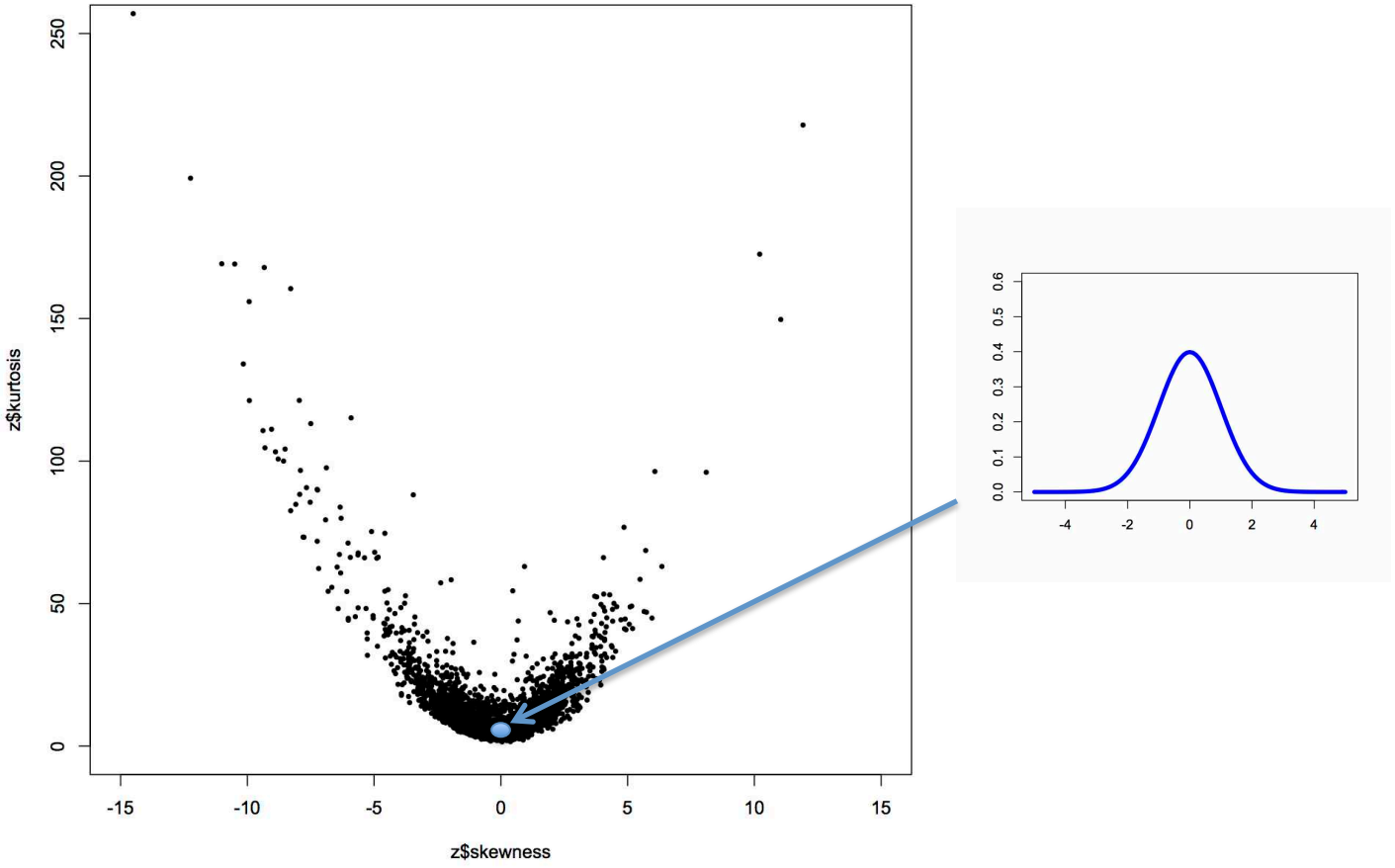
- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations

Statistical Analysis



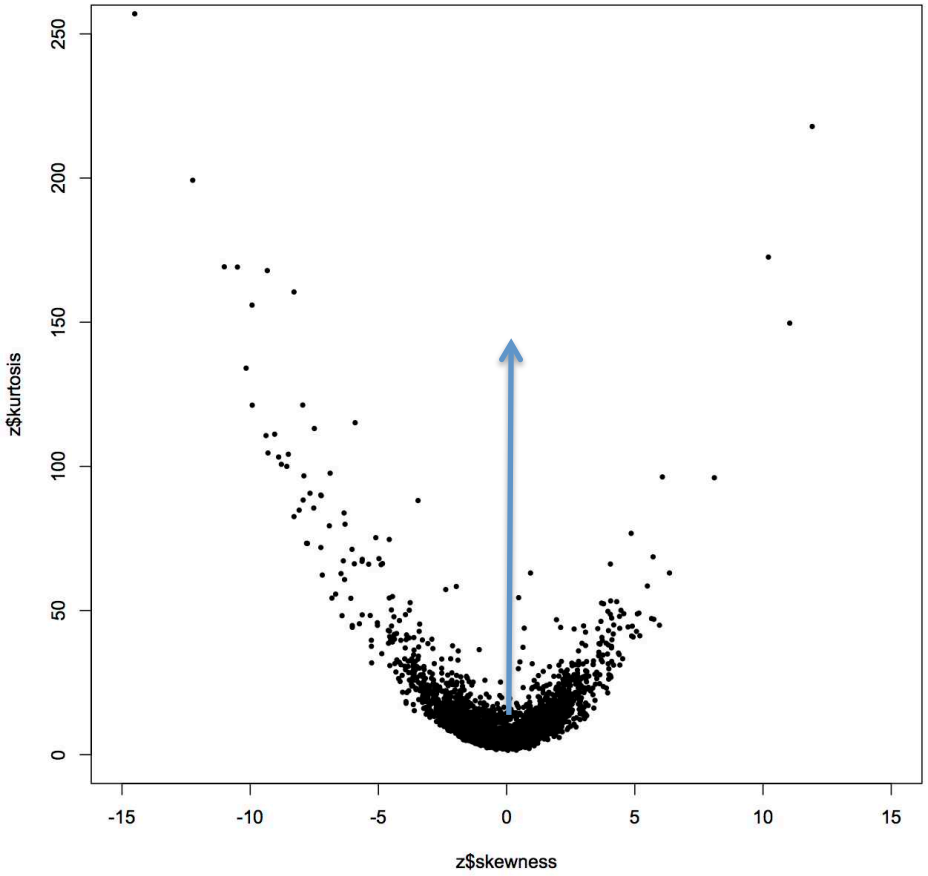
- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations

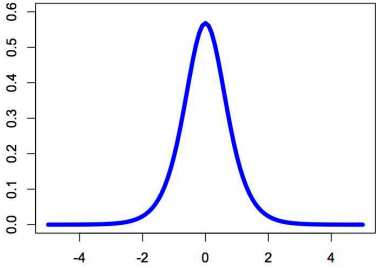
Statistical Analysis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations

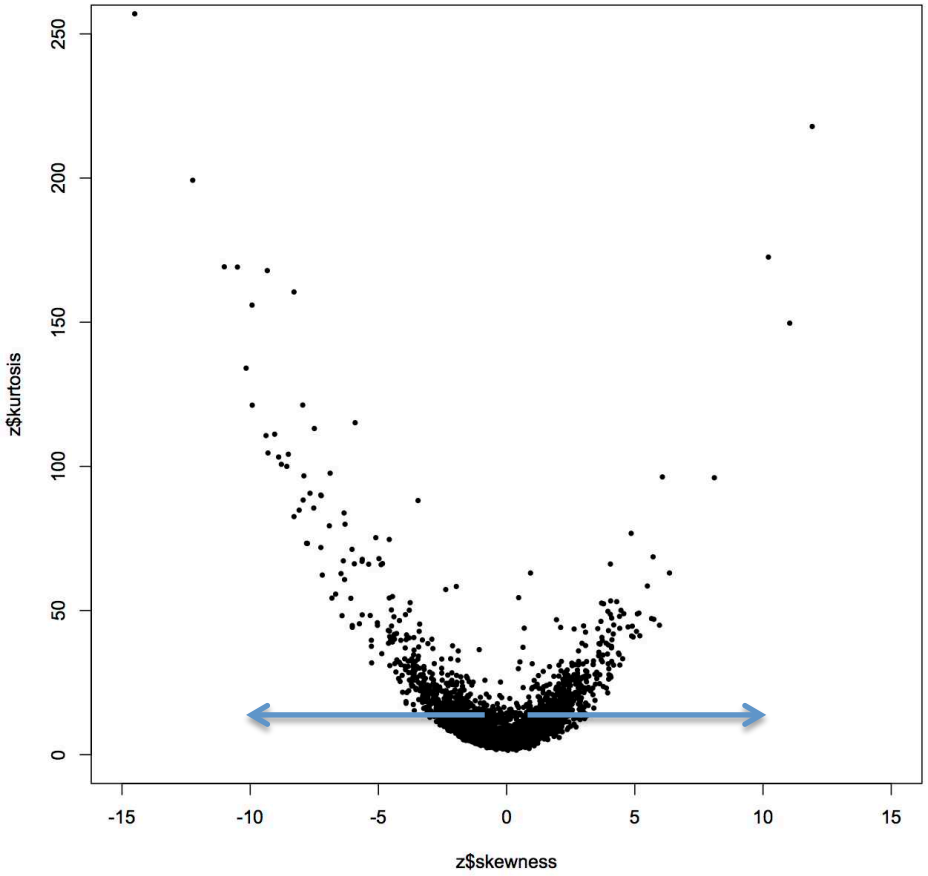
Statistical Analysis

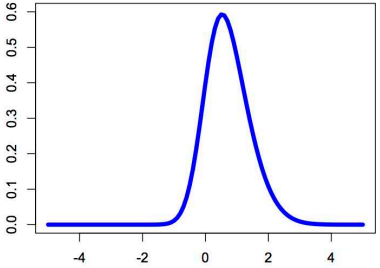




- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations

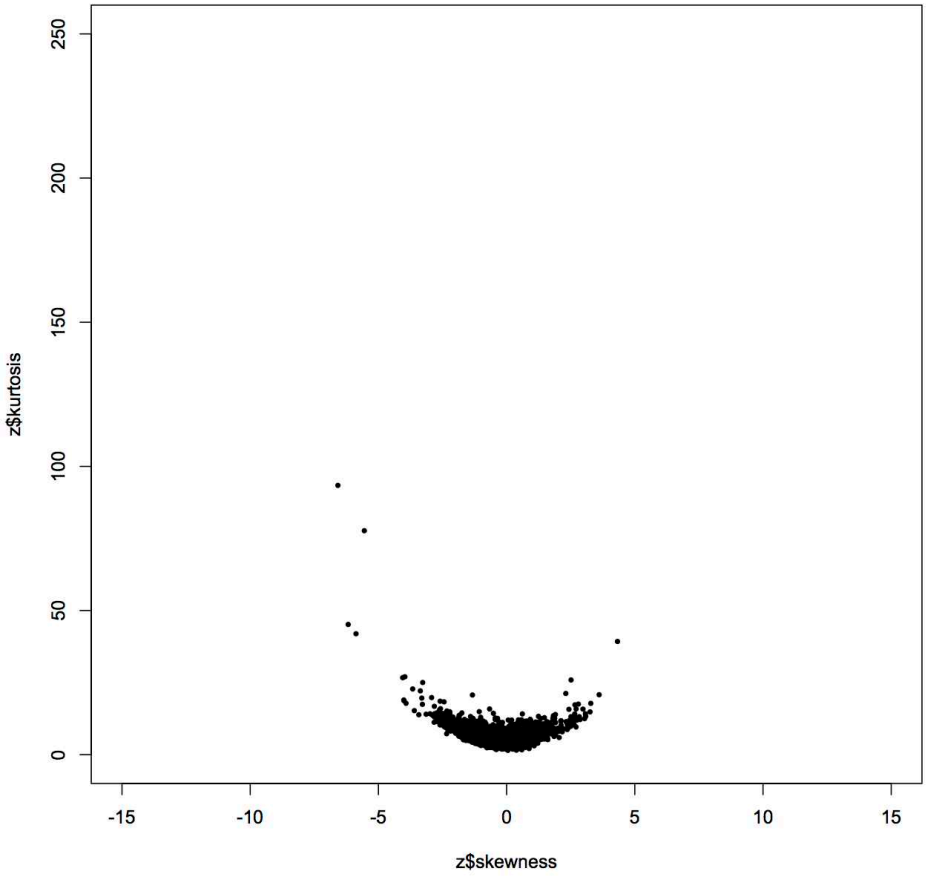
Statistical Analysis





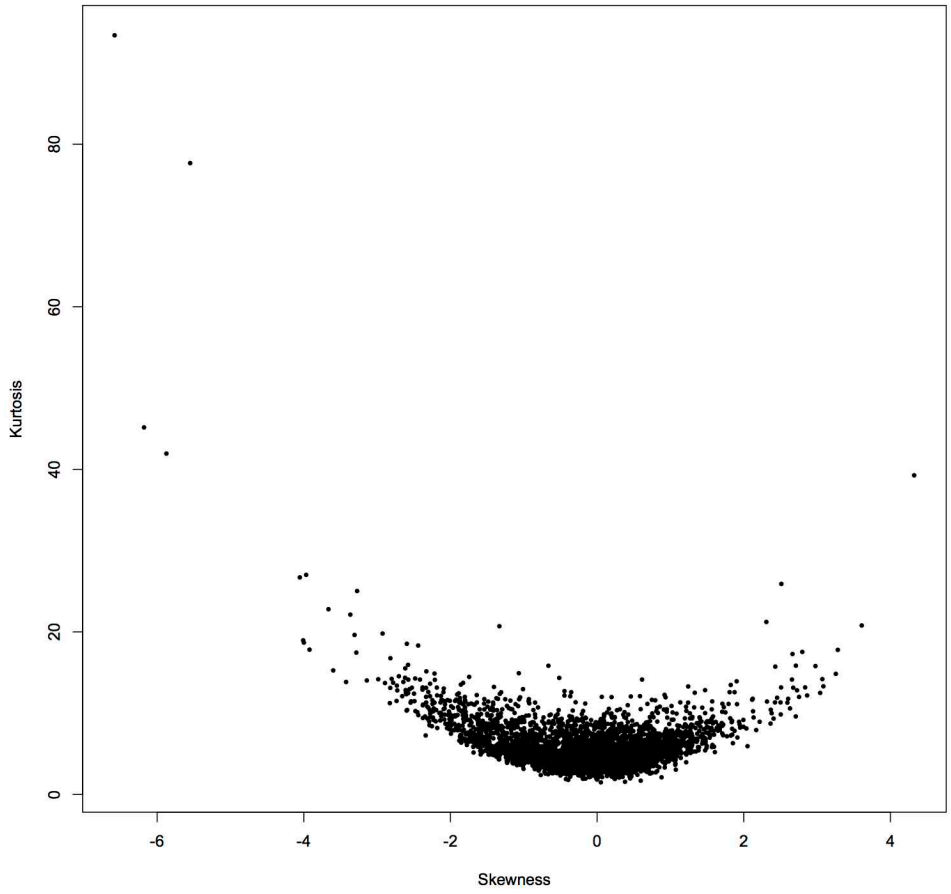
- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations

Statistical Analysis



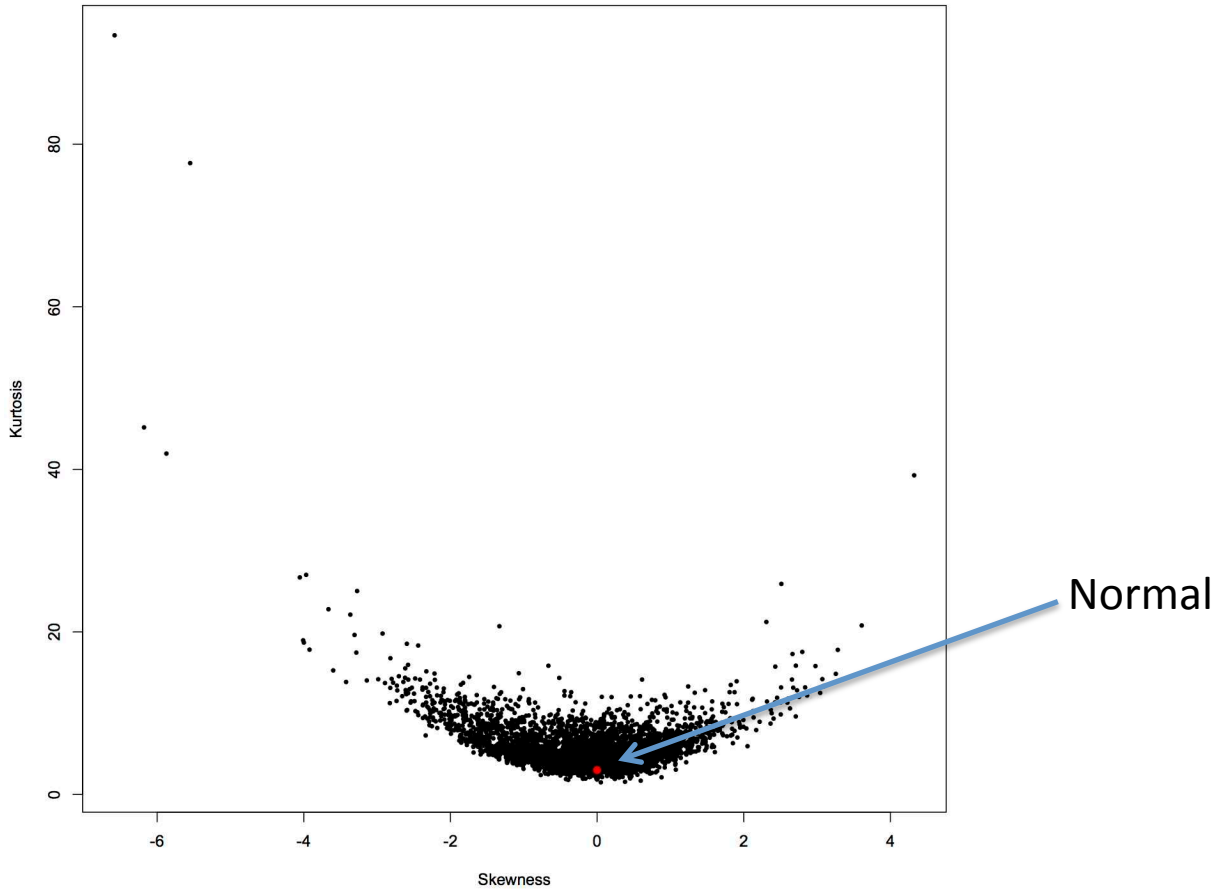
- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

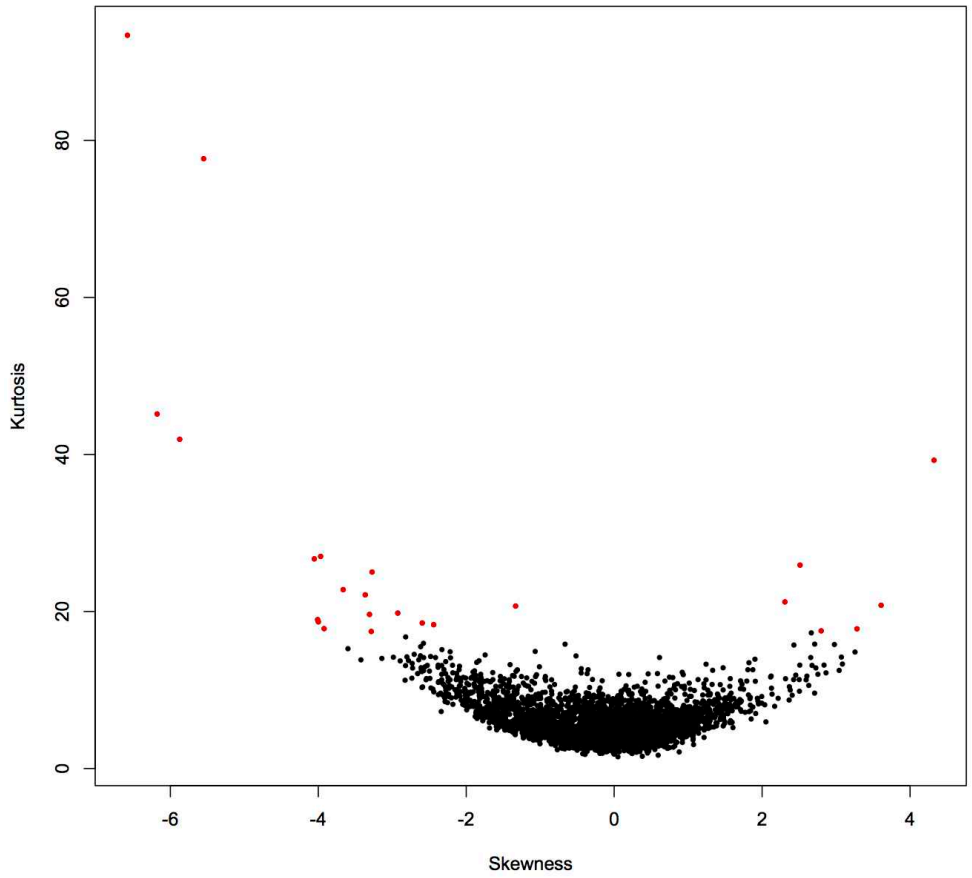
Statistical Analysis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis

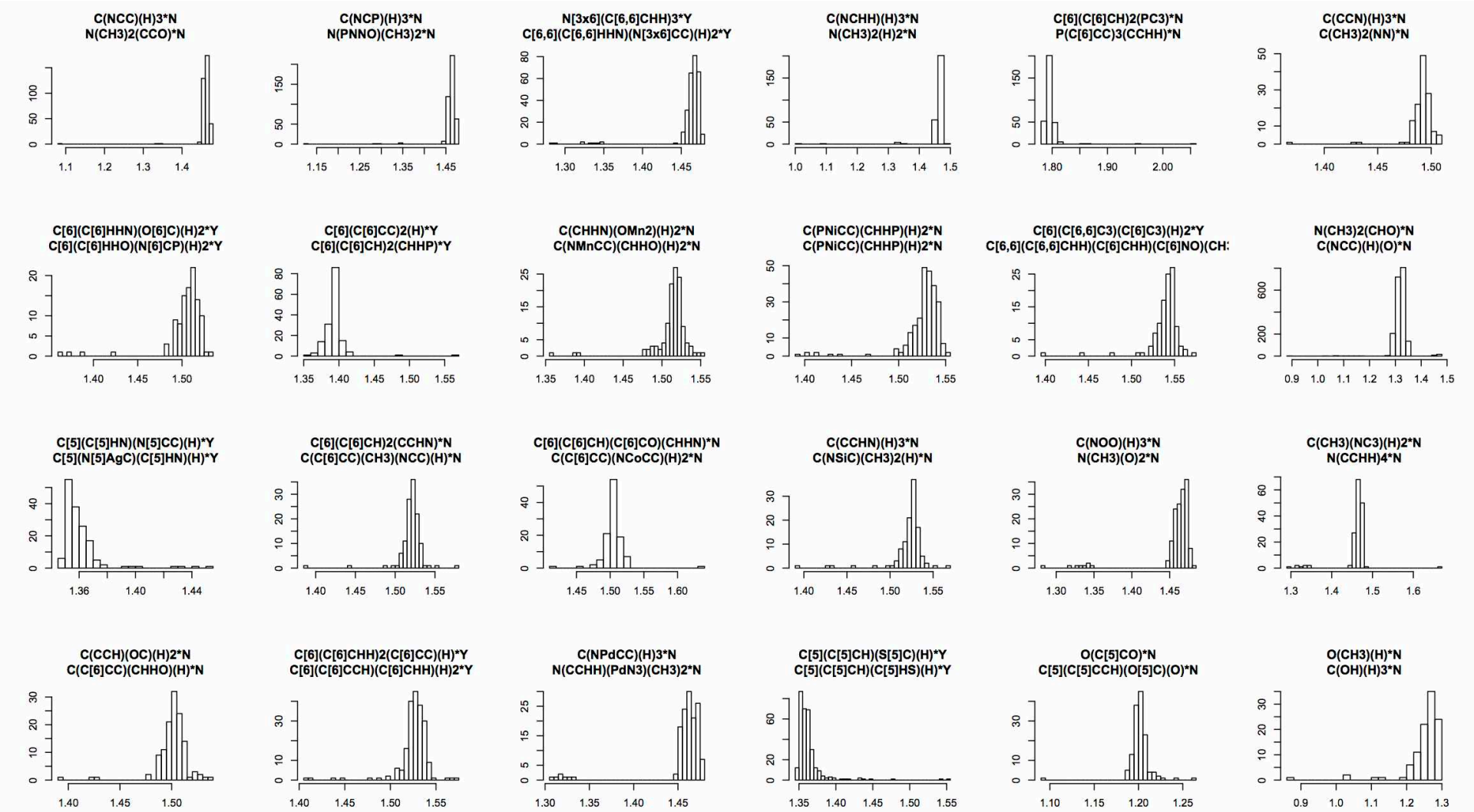
High Kurtosis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

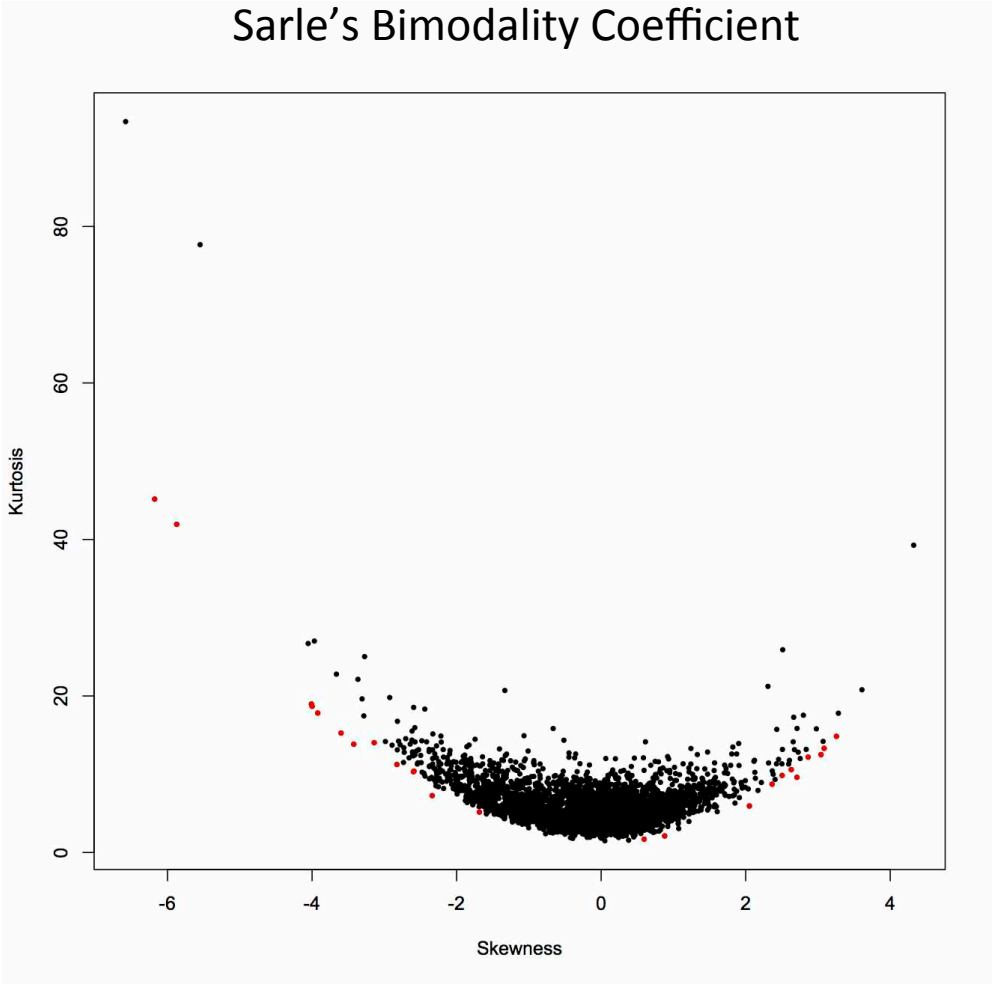
Statistical Analysis

High Kurtosis



Statistical Analysis

$$\frac{\gamma^2 + 1}{\kappa}$$

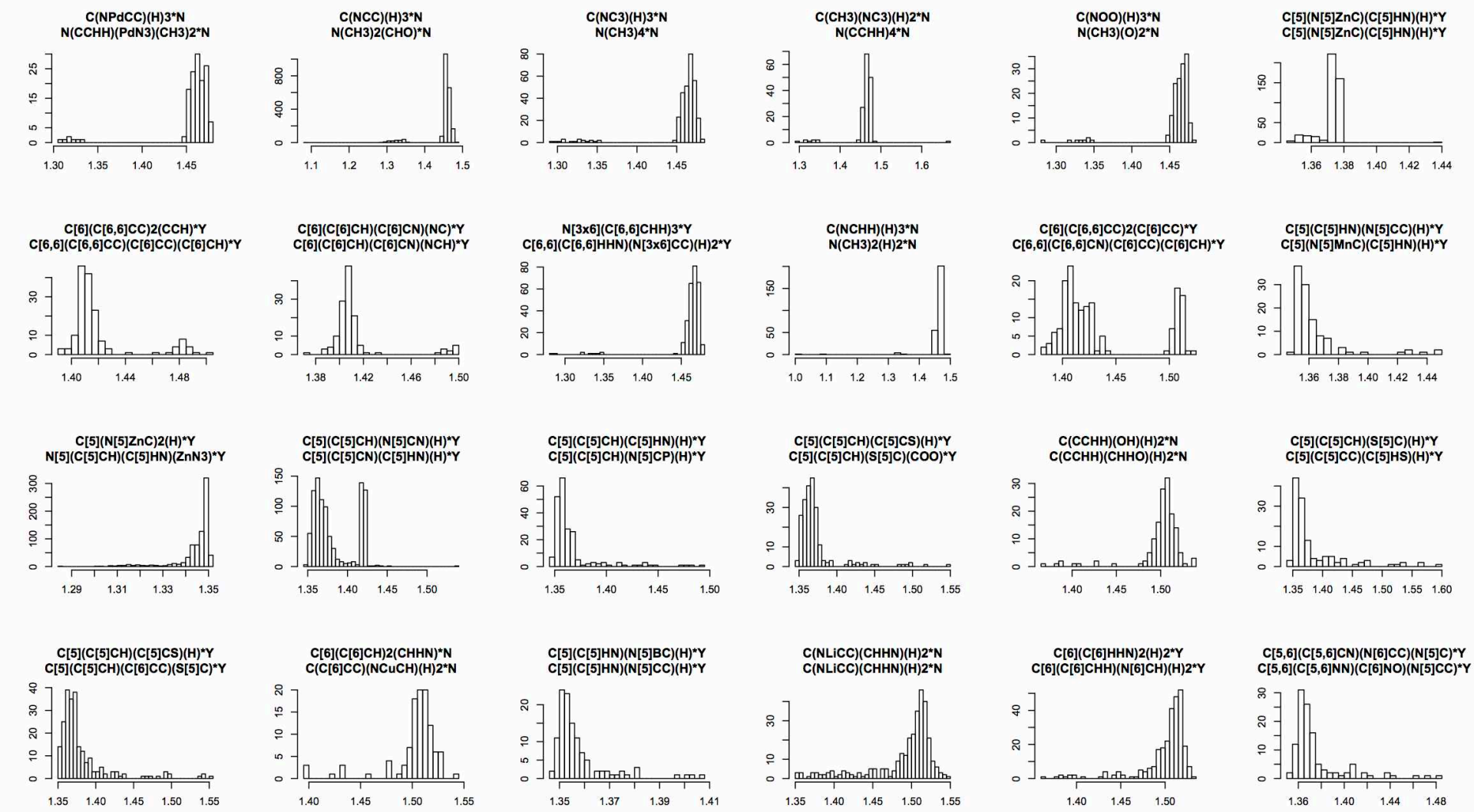


- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis

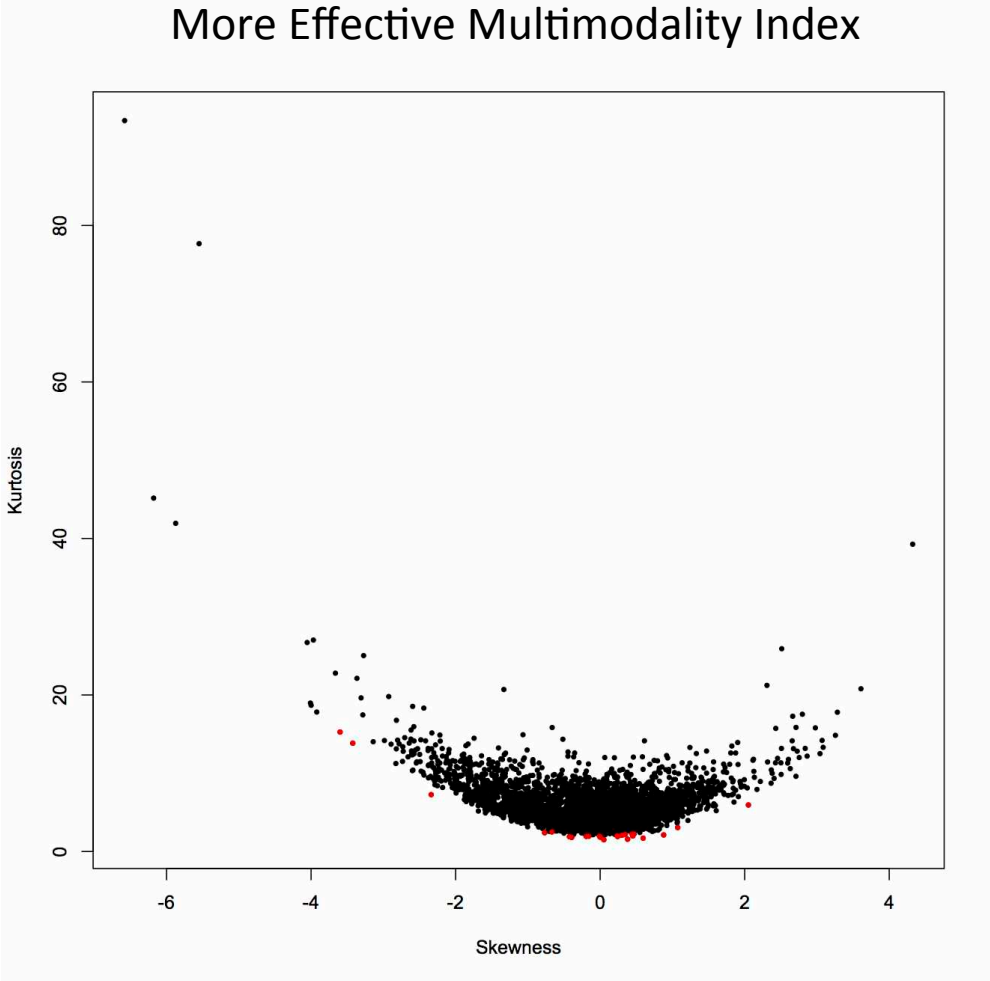
$$\frac{\gamma^2 + 1}{\kappa}$$

Sarle's Bimodality Coefficient



Statistical Analysis

$$\frac{\gamma^2 + 2}{2\kappa}$$

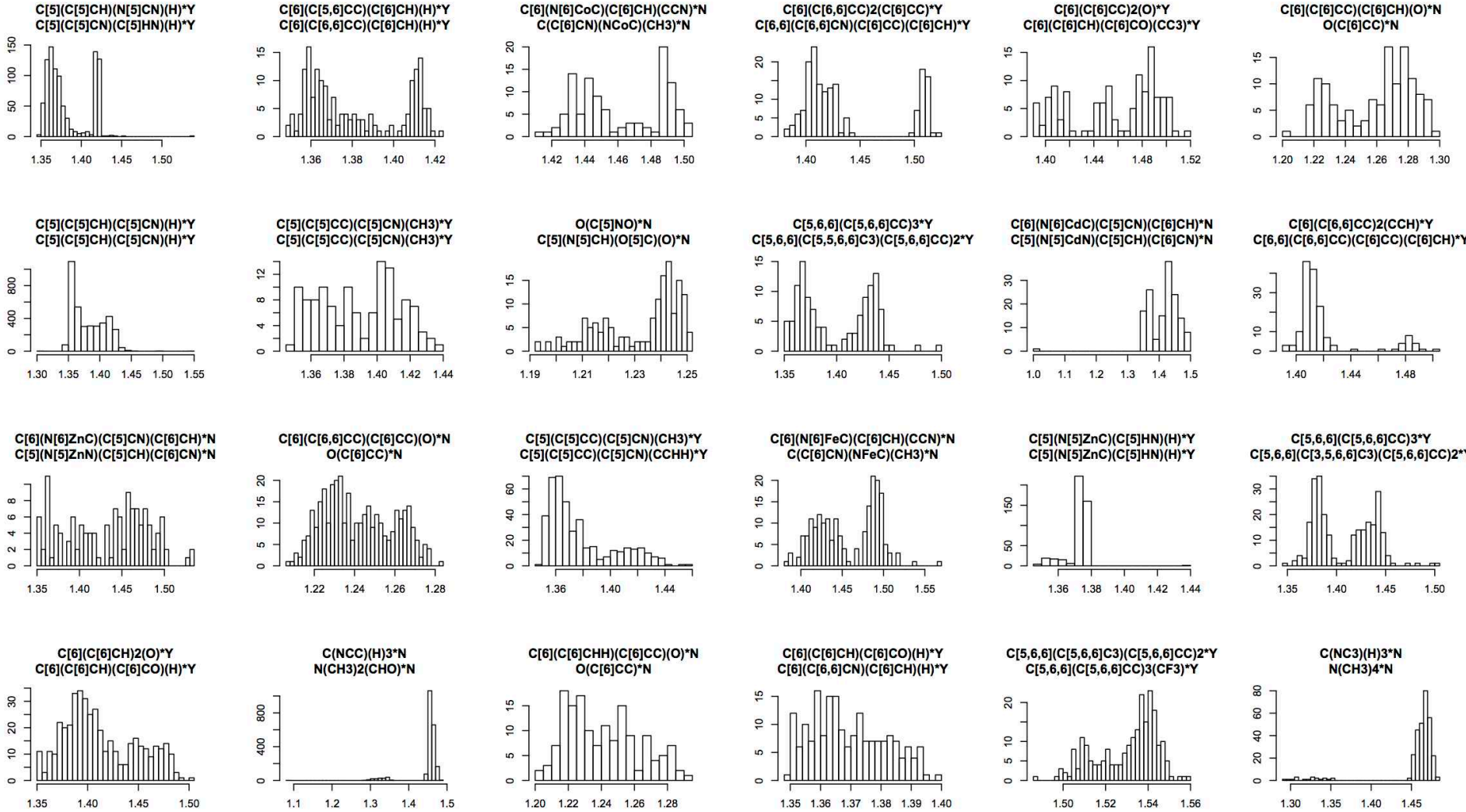


- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis

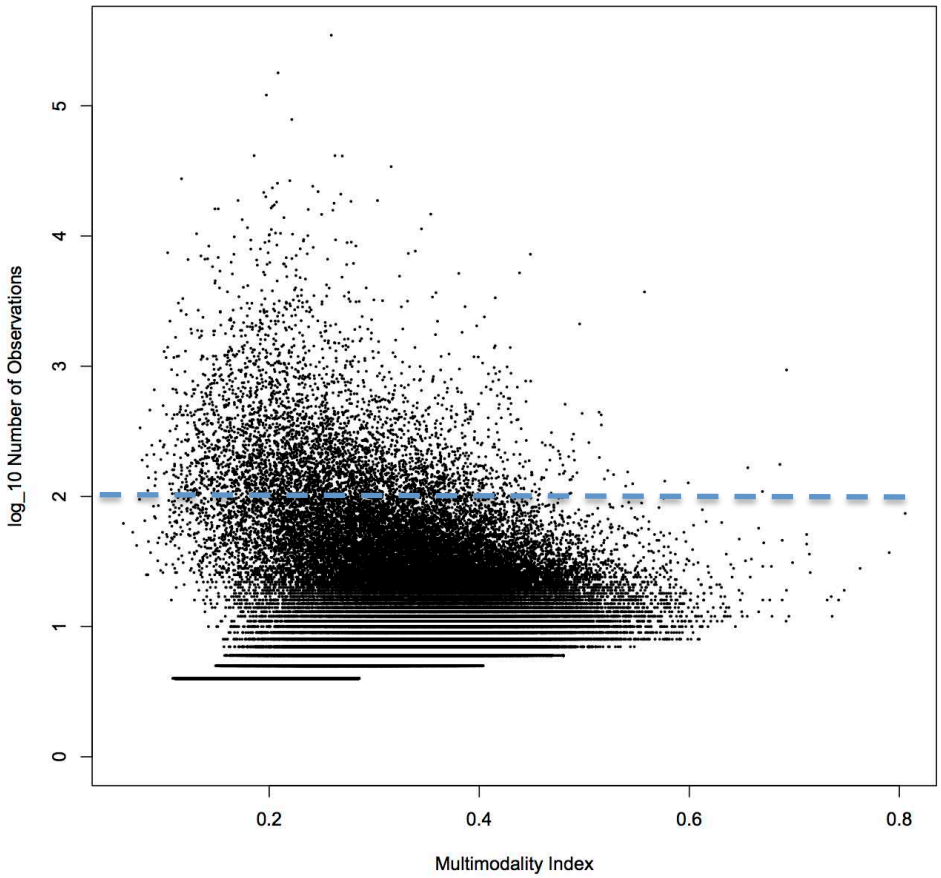
$$\frac{\gamma^2 + 2}{2\kappa}$$

More Effective Multimodality Index



Statistical Analysis

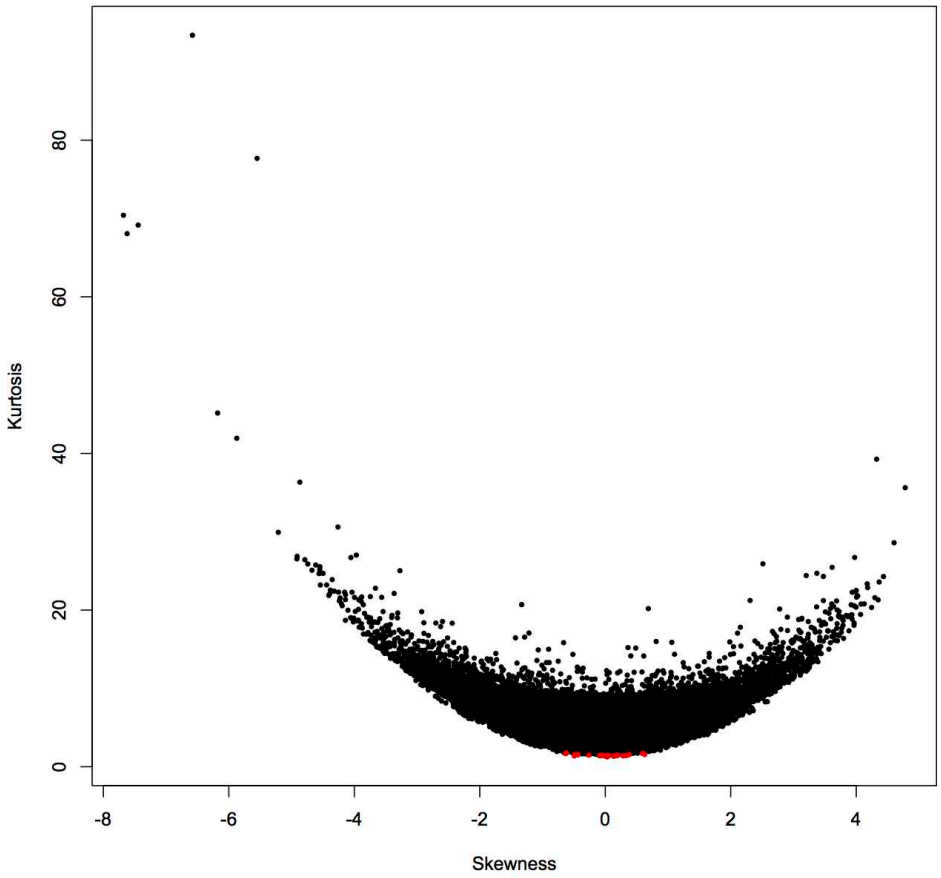
Now consider smaller classes



- Hydrogens are excluded (by default – not a restriction)
- ~~Only consider classes with >100 bond observations~~
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis

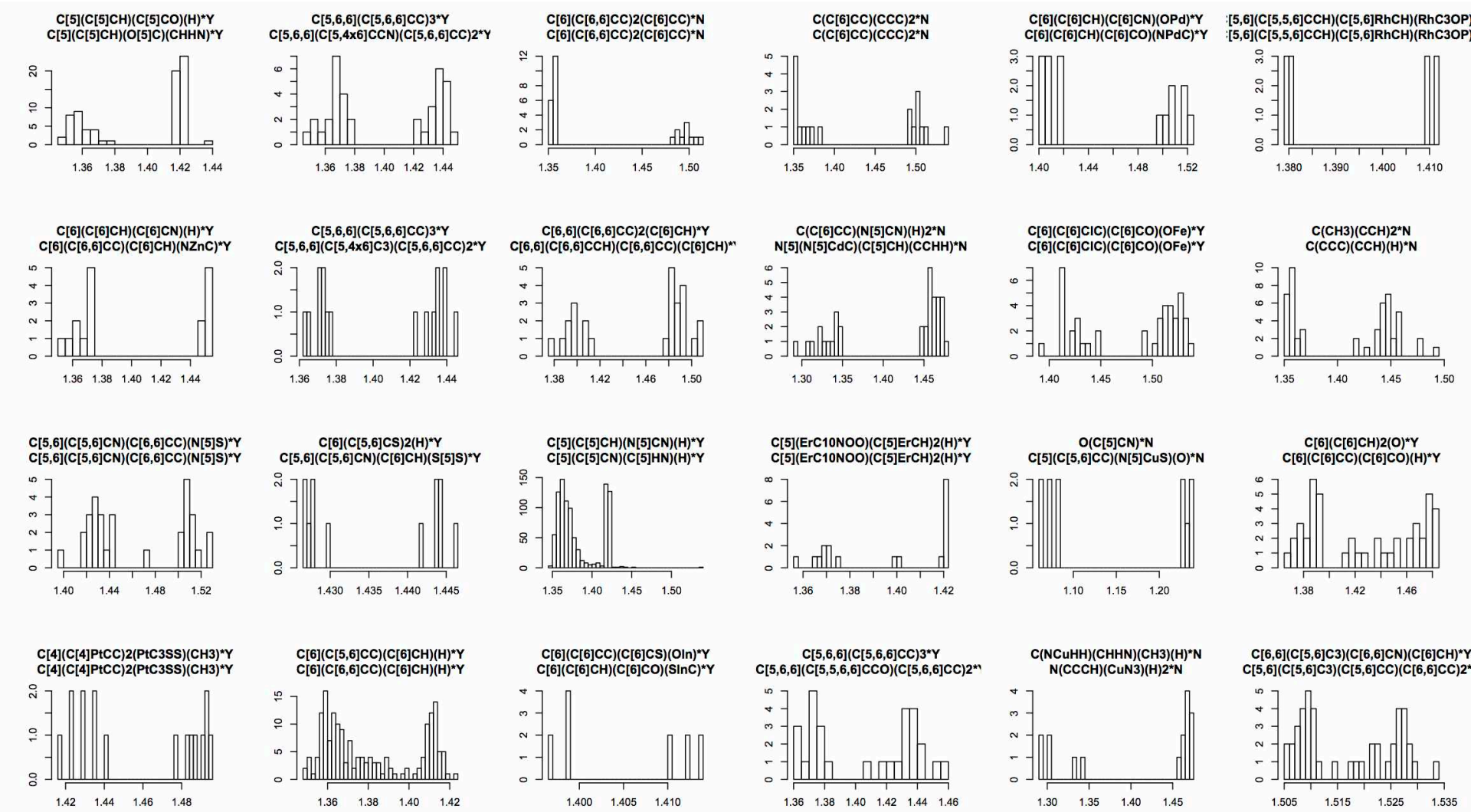
Multimodality Index



- Hydrogens are excluded (by default – not a restriction)
- ~~Only consider classes with >100 bond observations~~
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

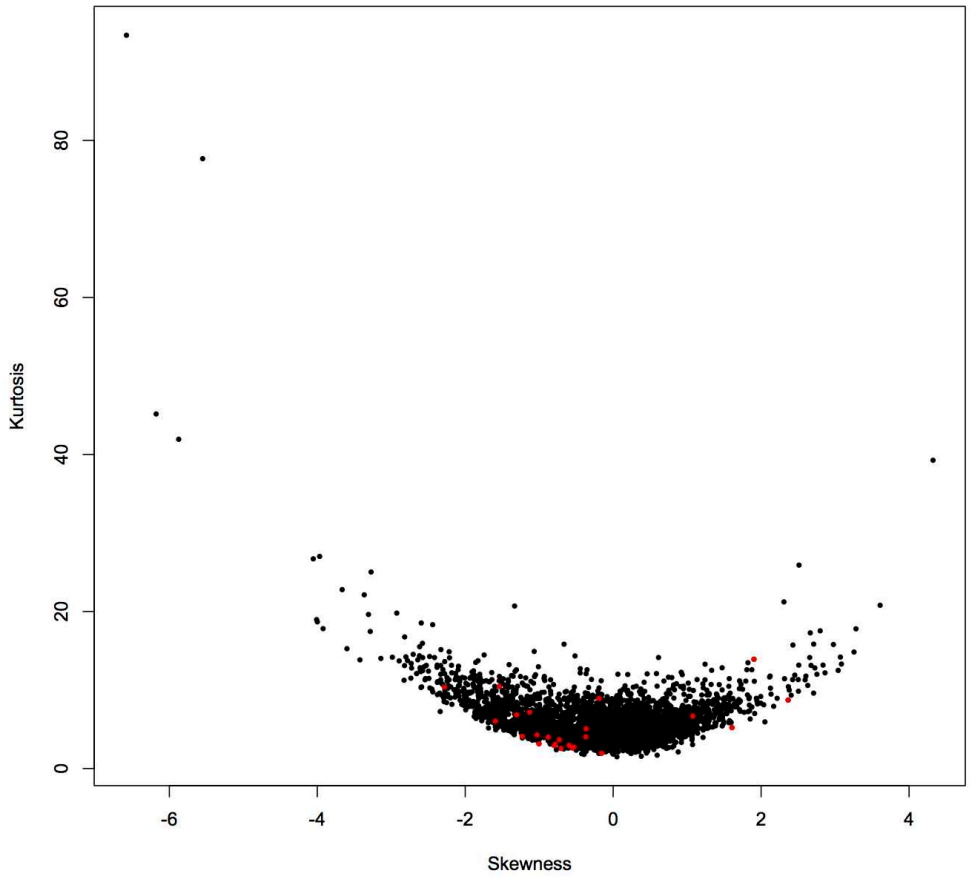
Statistical Analysis

Multimodality Index



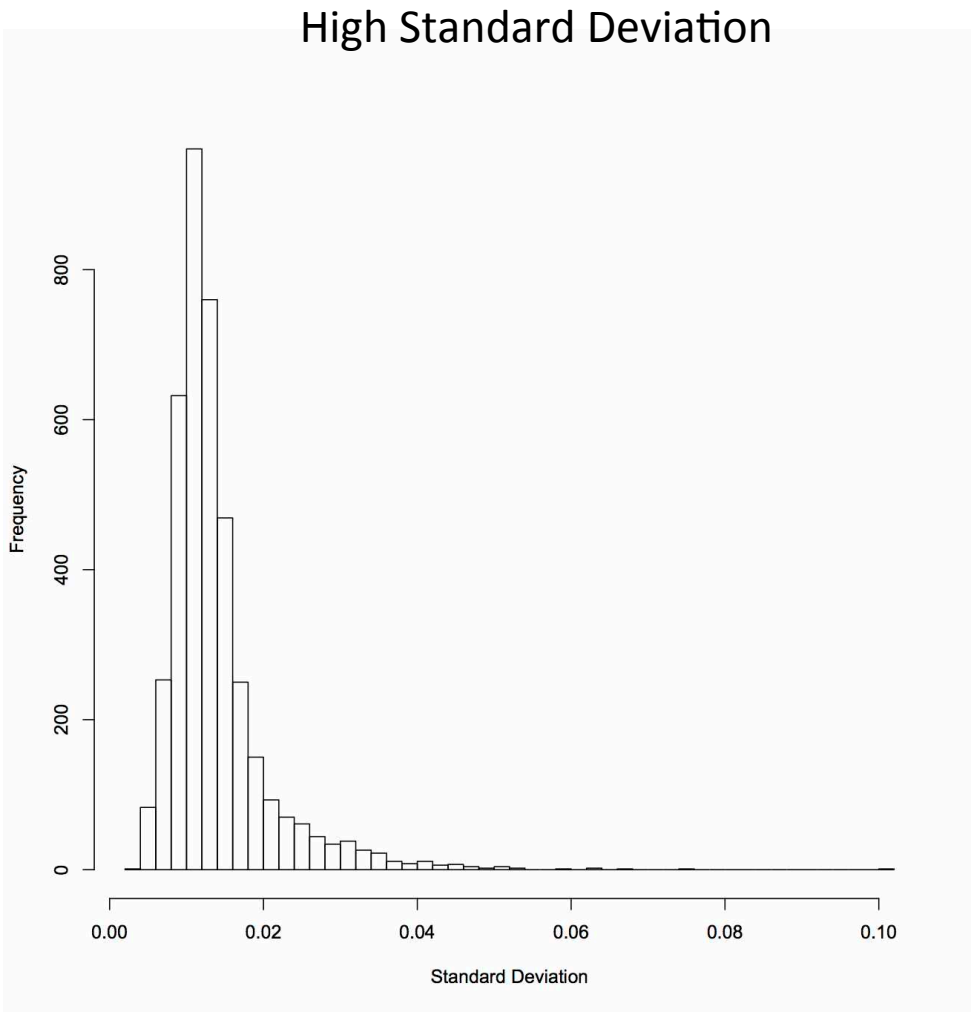
Statistical Analysis

High Standard Deviation



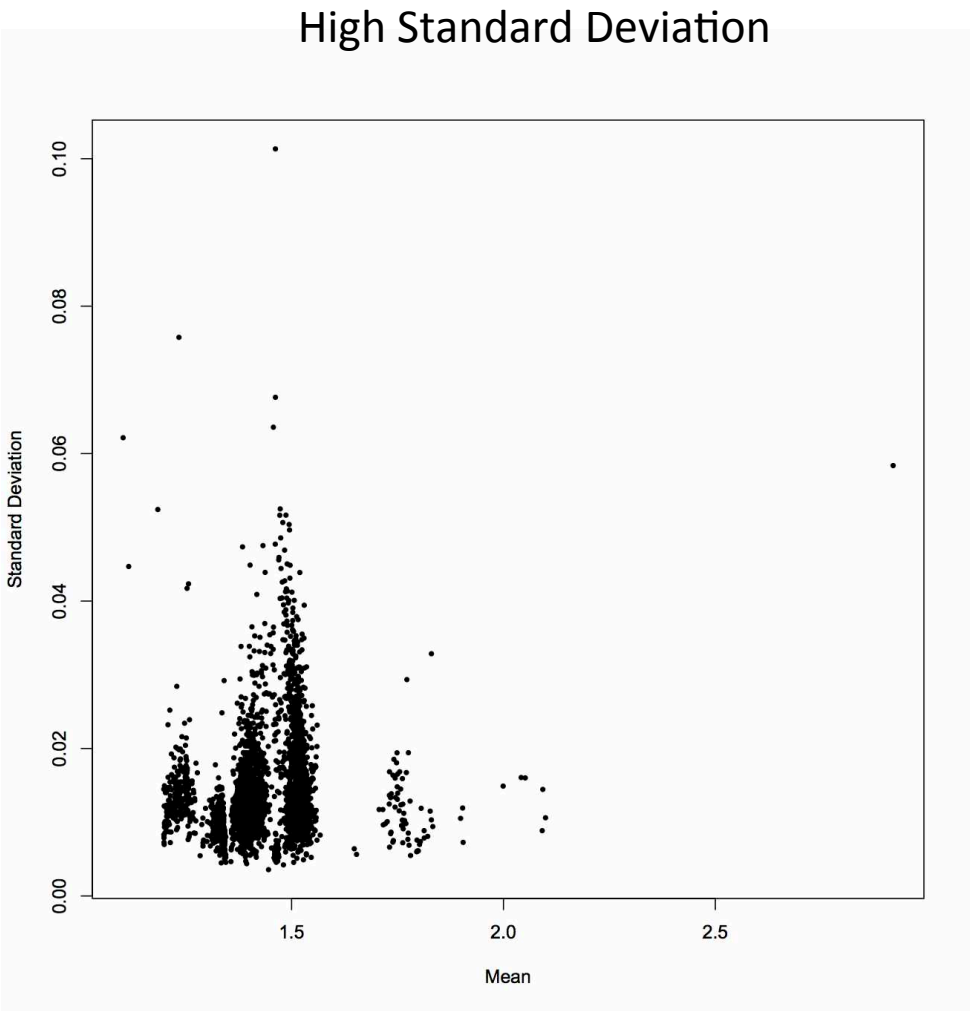
- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

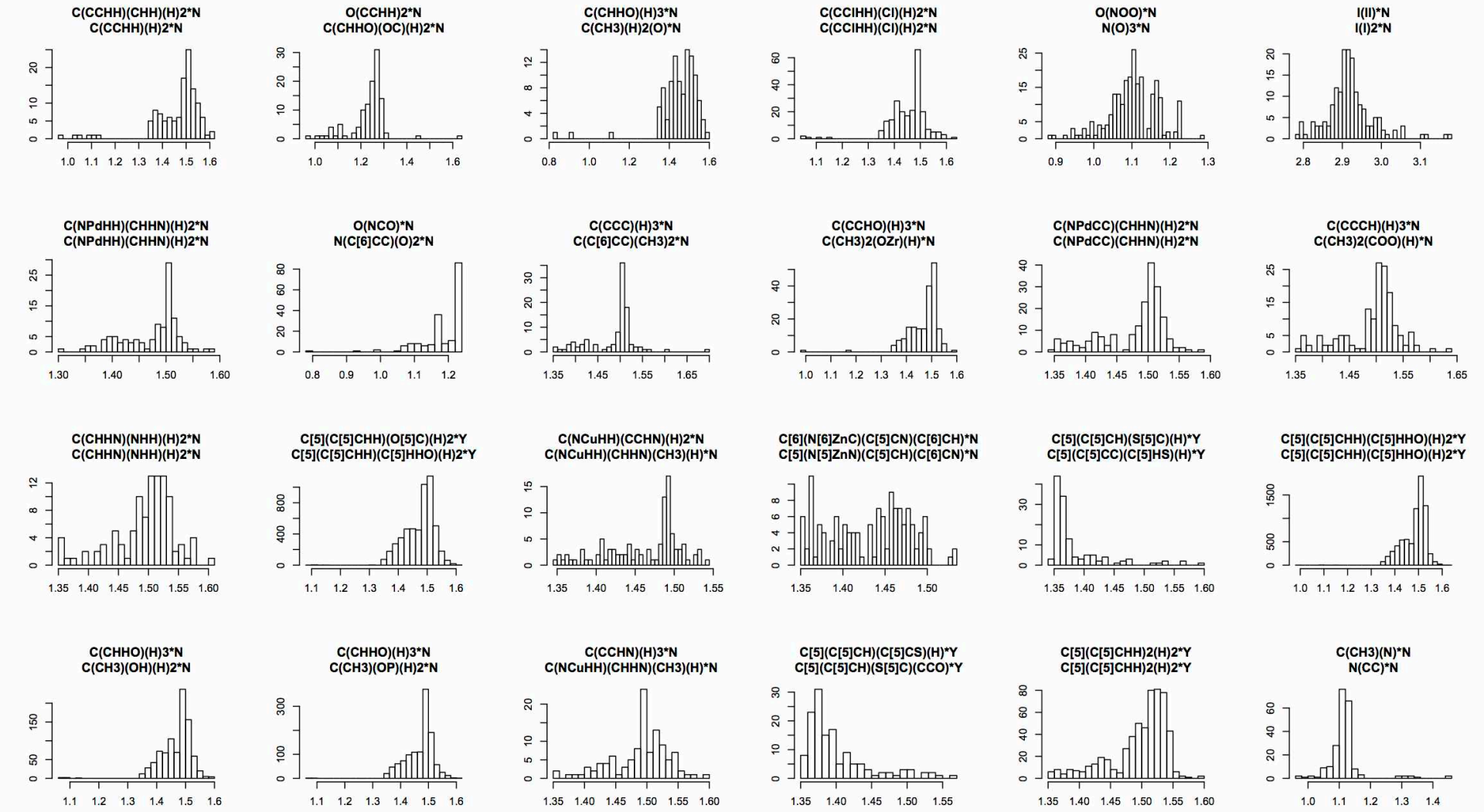
Statistical Analysis



- Hydrogens are excluded (by default – not a restriction)
- Only consider classes with >100 bond observations
- Exclude extreme outliers: z-score > 5 (ACEDRG removes outliers on-the-fly)

Statistical Analysis

High Standard Deviation



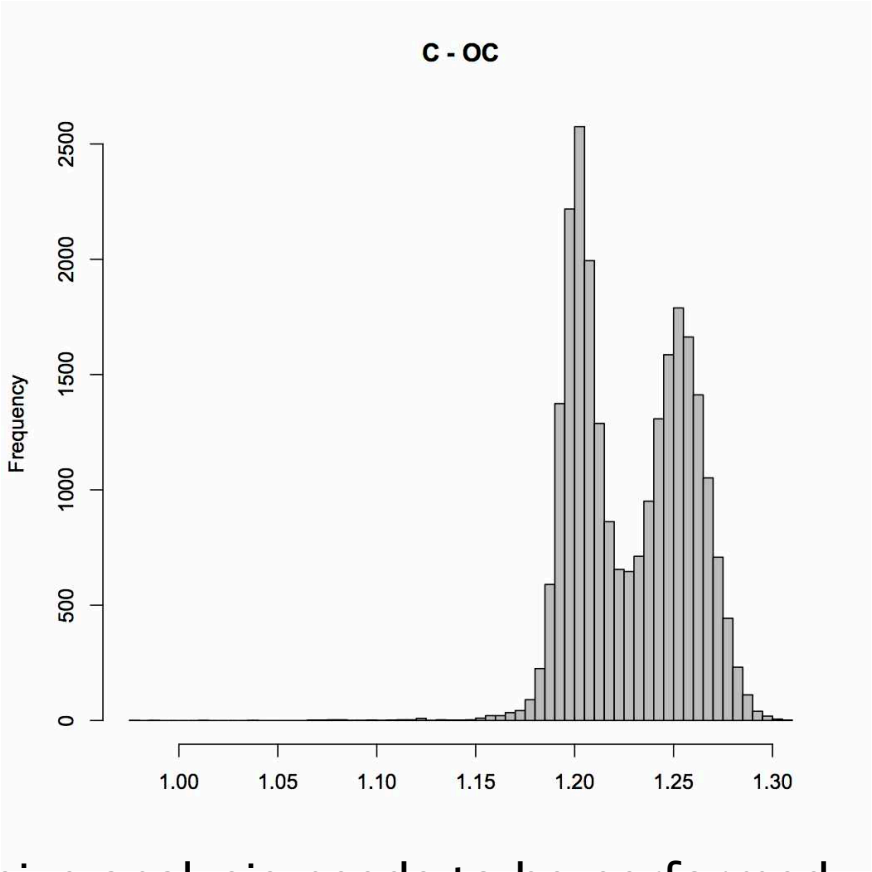
Summary

- First round of analysis (June) identified some multimodal cases.
- ACEDRG classes were updated to address some of these.
  Others will be fixed in upcoming revision(s)
- Second round of analysis has provided further feedback.
  Multimodality index dependent on N_{obs} – binning may be required.
- ACEDRG classes are being improved, and feedback cycle repeated.

Analysis of Refmac Monomer Library

We can now look at the enerlib.cif classes in light of ACEDRG & COD

Example:



Comprehensive analysis needs to be performed.

Conclusions

Further improvements to atom classes needed

Further iterations of feedback cycle needed

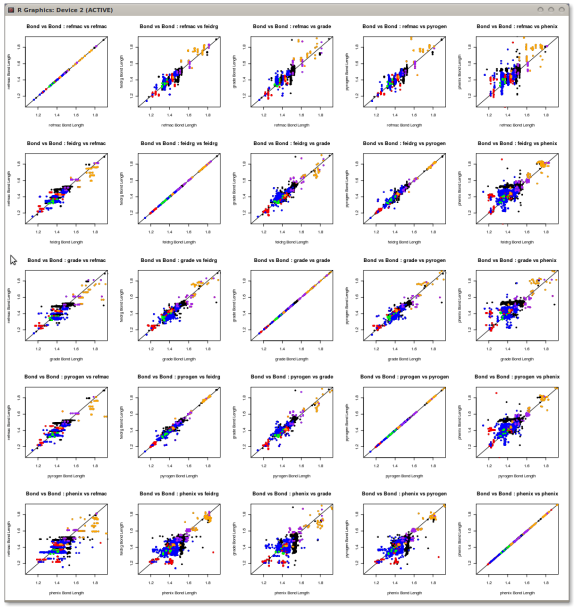
Intended Future Plans:

- Make independent validation tool available for non-dev usage; optimise functionality; streamline feedback cycle
- Test suitability of merging up the hierarchy (when appropriate)
- Further analysis of REFMAC Monomer Library versus ACEDRG – question: “how much does ACEDRG atom classes improve bond descriptions?”

Discussion

Topics for discussion:

- Should this approach be used for H's? Are small molecules reliable for this? Probably not. So why are H's included in the ACEDRG tables?
- Is it sensible to include >1 observation per molecule of a given COD atom type in the ACEDRG tables, due to the potential for bias.



Concerns Raised:

- Potential for cross-validation:
  - Is retaining a “test set” of molecules appropriate?
  - We can test existing protocol using other databases (CSD)
  - Can industrial partners provide suitably diverse test sets?
  - Of course, we can also compare against other resources…
- Suitability/scope of release/exposure at this stage:

  Pro’s:
  - probably better than Refmac Monomer Library in general.

  Con’s
  - May be worse than Refmac Monomer Library in some cases.
  - Identified that current bond classification system needs improvement.
  - Negative results would be damaging to the reputation of the resource.