High Throughput Screening of Materials (CCP9)

Friday 20th April 2012

CXD Workshop

can
We ↑ generate lots of data too

- Classical molecular dynamics trajectory files – can produce as many Bytes per second as total machine Ops.
- Similar demands of storing and analysing data to extract useful information for longer term storage – ie sequence of uncorrelated configurations, etc...
- Many of the issues discussed at this workshop are also relevant to this problem.
- BUT now for another set of challenges.

High Throughput Screening of Materials (CCP9)

Friday 20th April 2012

CXD Workshop

Anthony Leung - CCP9 Postdoc
Dr. Jacqui Cole – Head of Structure and Dynamics Group
Cavendish Laboratory

Example of approach

Small scale project to aggregate 30 000 crystallographic data files from the literature and mine thermal properties information from them

Steps:

1) Collect CIFs - data provided from RSC

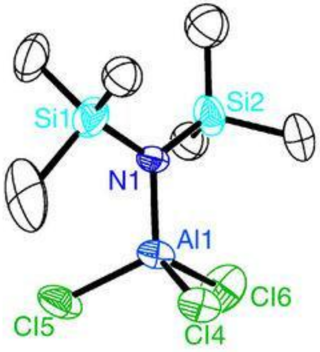
2) Parse CIFs and upload into database

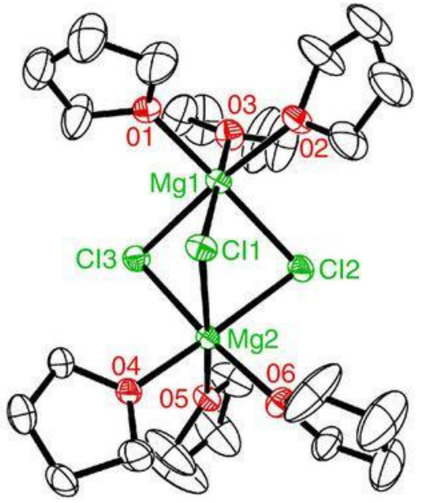
3) Mine information from this database

Example: Thermal Properties

Atomic positions within crystal structures can be accurately determined by neutron-scattering diffraction experiments. These are known as thermal properties, or Atomic Displacement Parameters (ADPs).

These factors represent the atomic position as an approximately ellipsoidal volume.



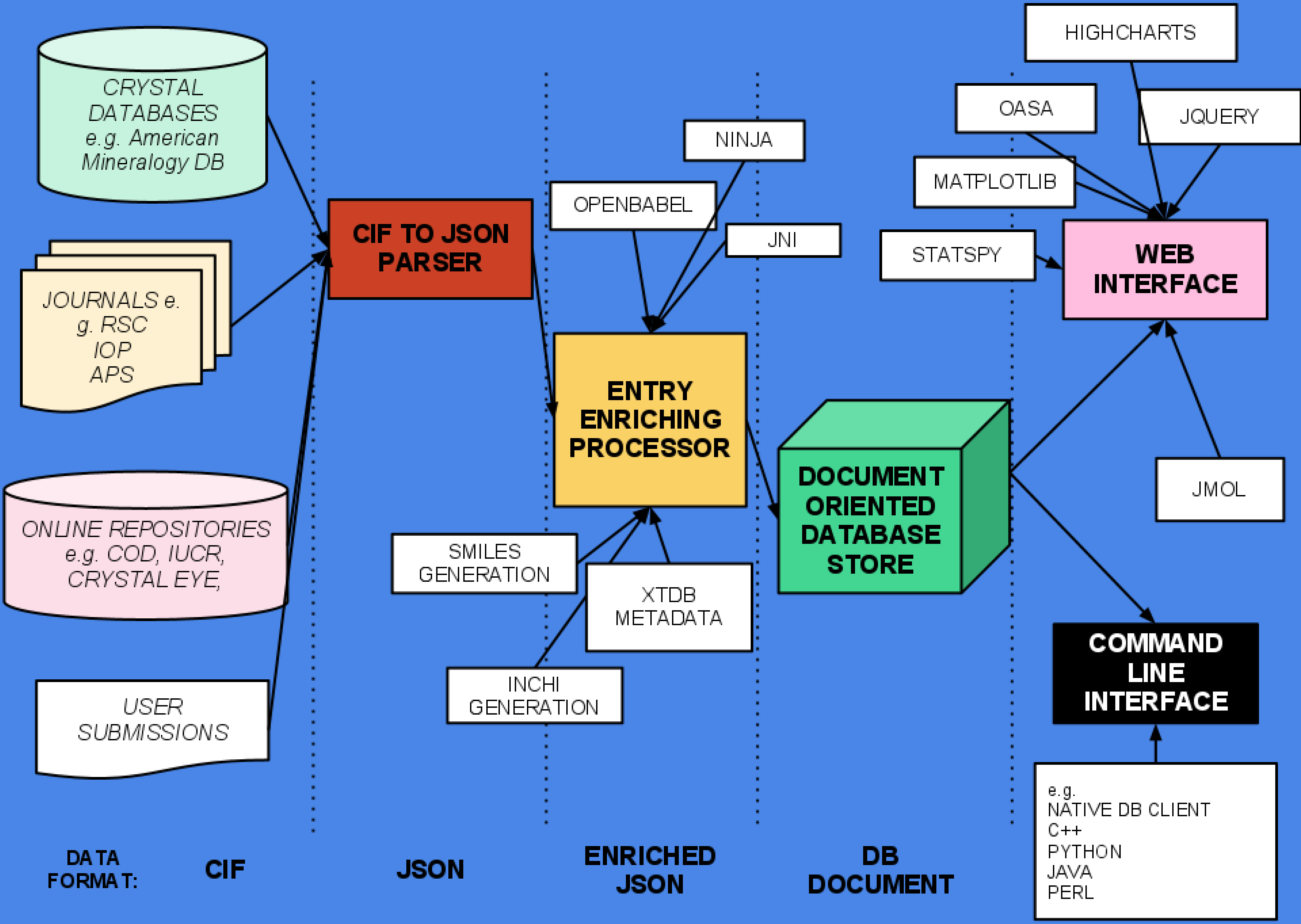


CIFs

- Crystallographic Information Files, or CIFs, are a widely adopted standard format for representing crystallographic data
- Required by crystallographic journals as part of the submission process.
- The CIFs are held as part of various journals and/or in topic specific repositories.

Data mining process

- Parse CIFs
- Enrich Data (additional information such as connectivity, SMILES/INCHI code strings for identification and searching)
- Enter into Database
- Query Database
- Find thermal properties trends over the 30,000 datasets.



CRYSTAL DATABASES e.g. American Mineralogy DB

JOURNALS e.g. RSC IOP APS

ONLINE REPOSITORIES e.g. COD, IUCR, CRYSTAL EYE,

USER SUBMISSIONS

CIF TO JSON PARSER

OPENBABEL

NINJA

JNI

ENTRY ENRICHING PROCESSOR

SMILES GENERATION

XTDB METADATA

INCHI GENERATION

DOCUMENT ORIENTED DATABASE STORE

HIGHCHARTS

OASA

JQUERY

MATPLOTLIB

STATSPY

WEB INTERFACE

JMOL

COMMAND LINE INTERFACE

e.g.
NATIVE DB CLIENT
C++
PYTHON
JAVA
PERL

DATA FORMAT: **CIF** **JSON** **ENRICHED JSON** **DB DOCUMENT**

Thermal Properties trends

Here are some graphs summarising some trends found for Atomic Displacement Parameters with respect to:

- temperature
- p-block elements
- transition metals
- coordination
- bond angle
- zero point motion estimatation

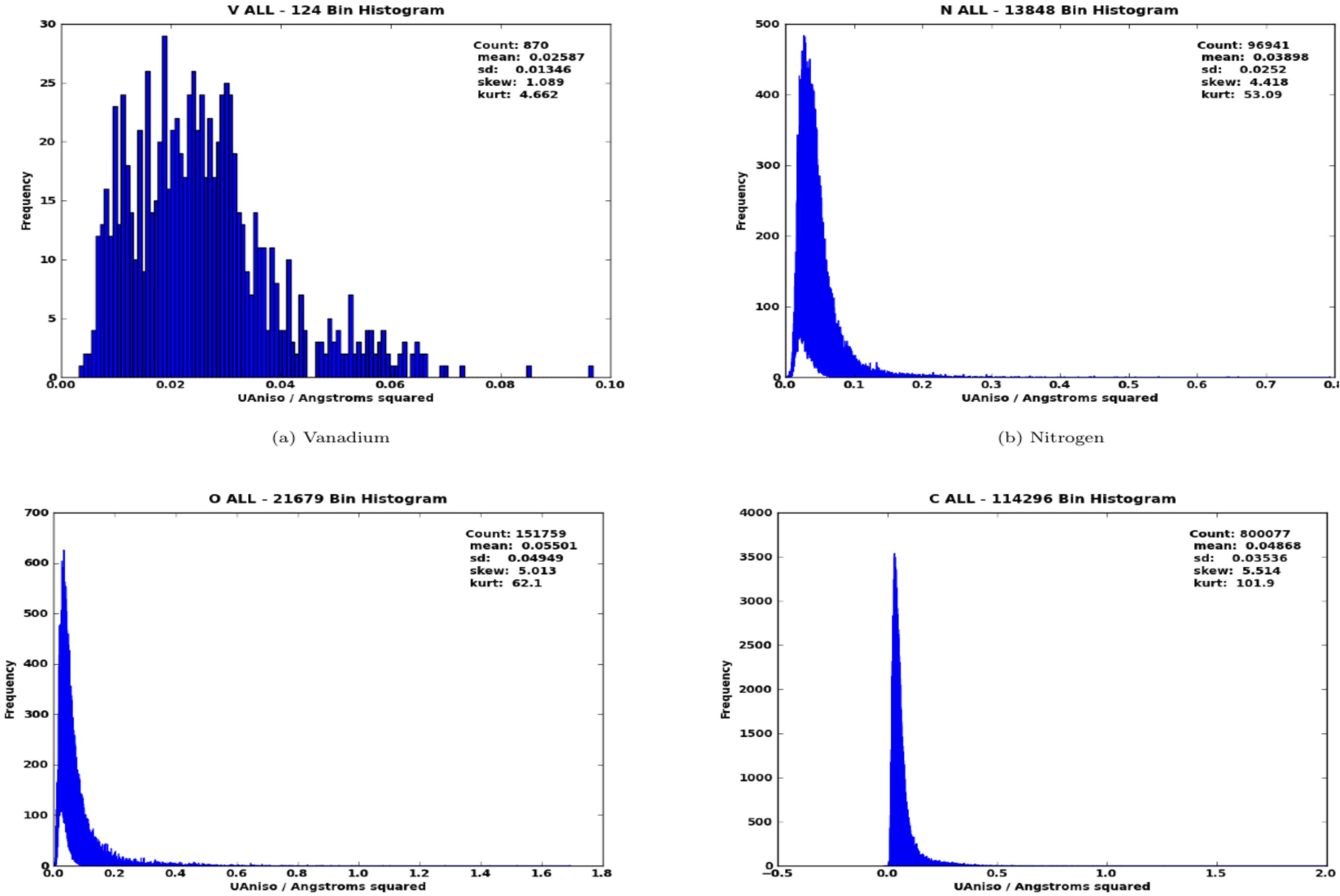
Distribution of ADP magnitudes

Following slide shows the distribution of ADP magnitudes for Carbon, Nitrogen, Oxygen (following a skewed bell curve) and Vanadium (which has fewer results).

Factors affecting the magnitude of a particular atomic ADP for a given element include:

1) Temperature of experiment

2) Bonding environment of atom (which dictates the freedom of atom to move)



(a) Vanadium

(b) Nitrogen

(c) Oxygen

(d) Carbon

Figure 2.4: Thermal properties $\mathbf{U}_{anisoequiv}$ distributions for a) Vanadium, b) Nitrogen, c) Oxygen and d) Carbon. These elements were selected for illustration because Vanadium is an example of a metal with a number of possible oxidation states; while C, N and O are the most common non-Hydrogen elements that comprise organic compounds.

Table of ADP values

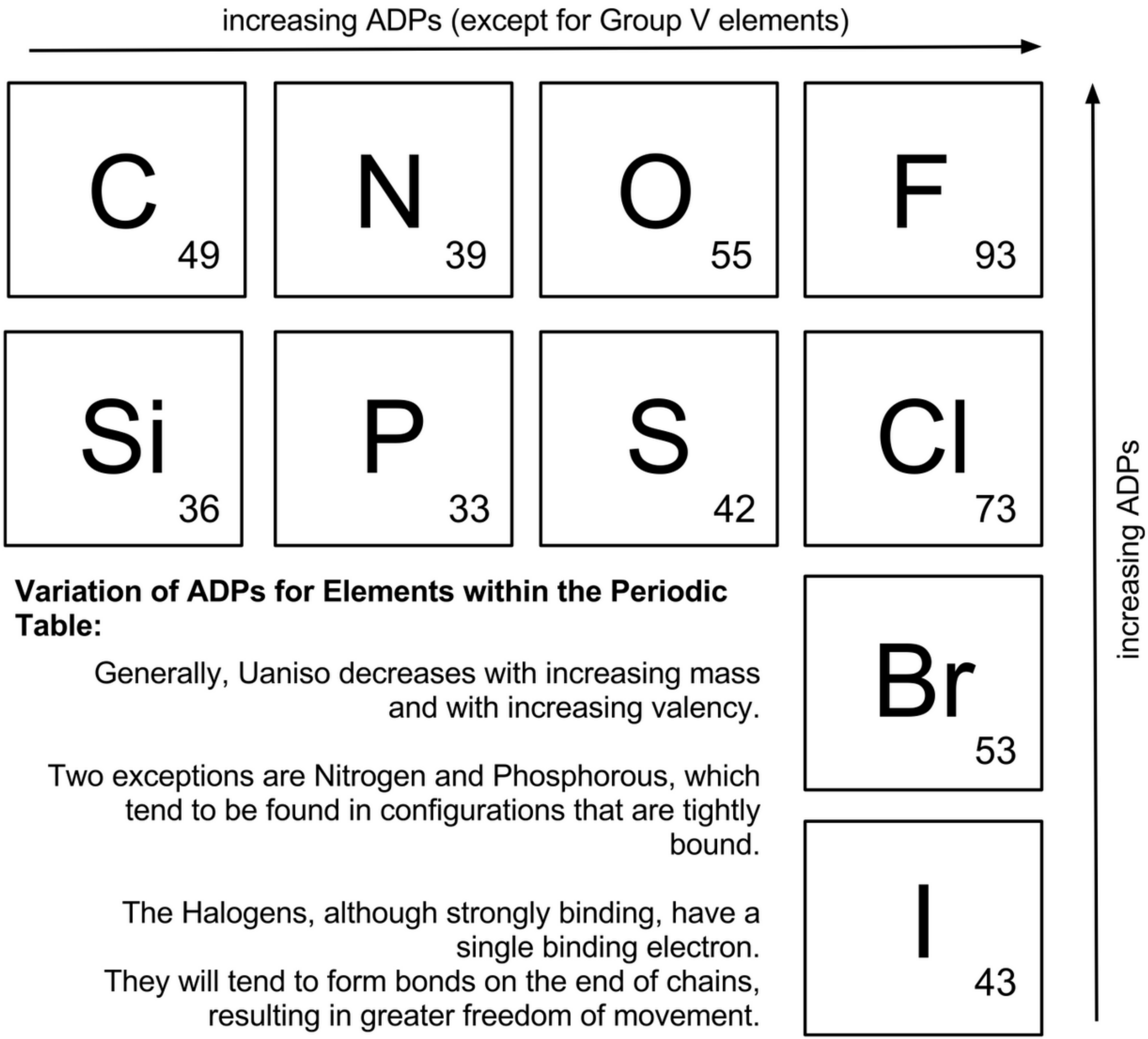
Mean ADP values for a selection of elements, and for Carbon in its various forms and bonding environments.

Species	**Mean** / Å^2 × 10^{-3}	Number of Hits	Std / Å^2	Skew	Kurt
Carbon	**48.68**	800077	0.03536	5.514	101.9
Carbon (Aromatic)	**43.94**	373189	0.0265	4.176	85.7
Carbon (Non-Aromatic)	**52.83**	426672	0.04128	5.379	88.78
Carbon (sp^1)	**48.03**	6884	0.03438	4.114	36.9
Carbon (sp^2)	**43.55**	490332	0.02664	4.961	108.3
Carbon (sp^2, non-aromatic)	**42.32**	117153	0.0278	7.083	165.7
Carbon (sp^3)	**57.01**	302634	0.04489	5.035	78.39
Carbon (Methyl)	**68.11**	100988	0.04979	5.327	96.25
Nitrogen	**38.98**	96941	0.0252	4.418	53.09
Nitrogen (Aromatic)	**37.51**	20344	0.01942	2.413	20.96
Oxygen	**55.01**	151759	0.04949	5.013	62.1
Silicon	**35.94**	3434	0.02214	10.24	287.9
Phosphorous	**32.55**	11710	0.02304	4.885	55.31
Sulphur	**42.45**	15353	0.02769	3.307	24.55
Fluorine	**92.5**	37418	0.06608	2.784	28.3
Chlorine	**72.94**	18686	0.06678	3.814	29.91
Bromine	**52.88**	3076	0.03701	4.523	42.56
Iodine	**46.35**	2795	0.02723	3.901	36.57
Lead	**32.5**	317	0.01677	1.461	5.974
Scandium	**21.26**	40	0.01001	0.7077	2.87
Titanium	**29.67**	518	0.01523	2.215	11.95
Vanadium	**25.87**	870	0.01346	1.089	4.662
Chromium	**28.17**	493	0.01398	1.178	4.319
Manganese	**29.55**	2042	0.01541	2.092	11.04
Iron	**30.76**	2921	0.0167	1.972	10.52
Cobalt	**28.37**	2233	0.01484	1.698	8.319
Nickel	**30.83**	2163	0.0169	2.277	13.18
Copper	**33.09**	4884	0.01888	2.697	21.1
Zinc	**31.27**	2270	0.01488	1.51	7.522

Table 2.1: $\mathbf{U_{anisoeq}}$ values

Variation with atomic species

- A graphic showing variation of ADP with respect to atomic species - increasing mobility of atoms rightwards and upwards in the periodic table.
- Some anomalies thought to be due to typical bonding environment of those atoms.



Variation of ADPs for Elements within the Periodic Table:

Generally, Uaniso decreases with increasing mass and with increasing valency.

Two exceptions are Nitrogen and Phosphorous, which tend to be found in configurations that are tightly bound.

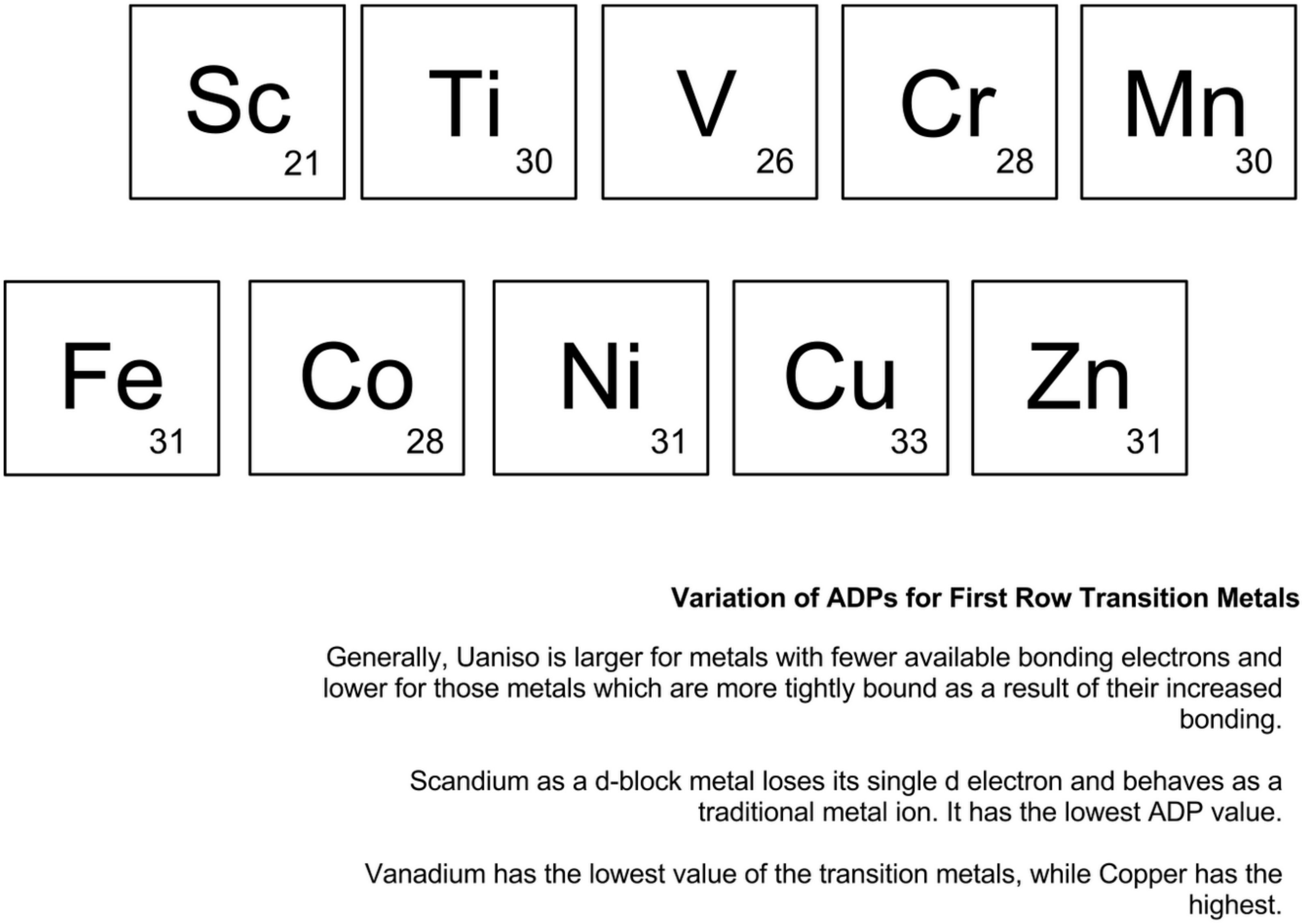
The Halogens, although strongly binding, have a single binding electron.
They will tend to form bonds on the end of chains, resulting in greater freedom of movement.

Figure 2.5: ADP variation on the right hand side of the Periodic Table: all units in $\text{Å}^2 \times 10^{-3}$

Variation with atomic species for Transition Metals

Figure showing variation in ADP magnitude for transition metals.

- Less clear cut - bonding clearly a major influence as well as mass.



Sc	Ti	V	Cr	Mn
21	30	26	28	30
Fe	Co	Ni	Cu	Zn
31	28	31	33	31

Variation of ADPs for First Row Transition Metals

Generally, Uaniso is larger for metals with fewer available bonding electrons and lower for those metals which are more tightly bound as a result of their increased bonding.

Scandium as a d-block metal loses its single d electron and behaves as a traditional metal ion. It has the lowest ADP value.

Vanadium has the lowest value of the transition metals, while Copper has the highest.

Figure 2.6: First Row Transition Metal ADPs: all units in $(\mathbf{A^2})^{-3}$

Bond angle dependence

Dependence of ADPs on bond angle. Bond angle information is stored as part of the CIF; pulling it out of the database is straightforward as it was already there due to the schemaless approach

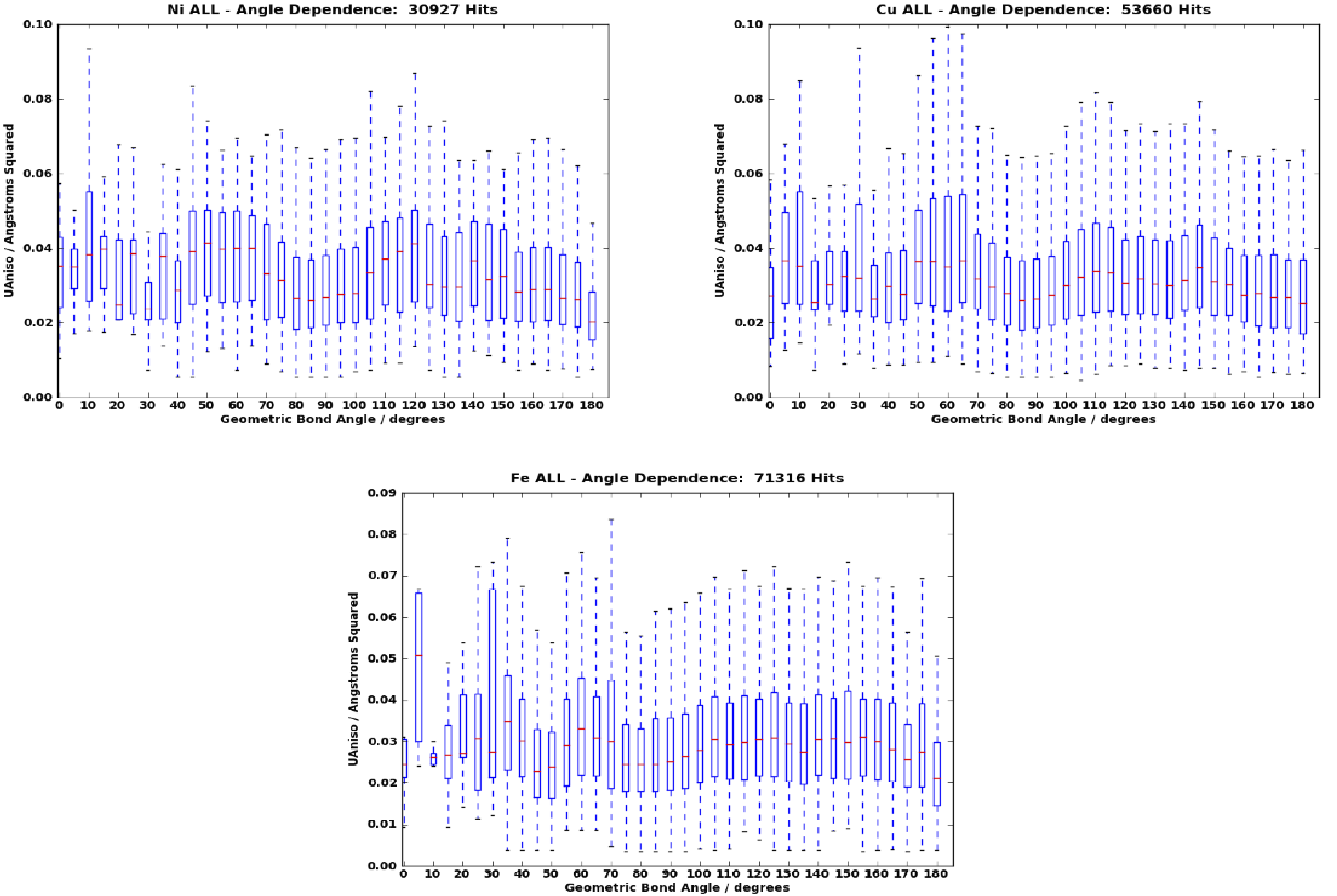


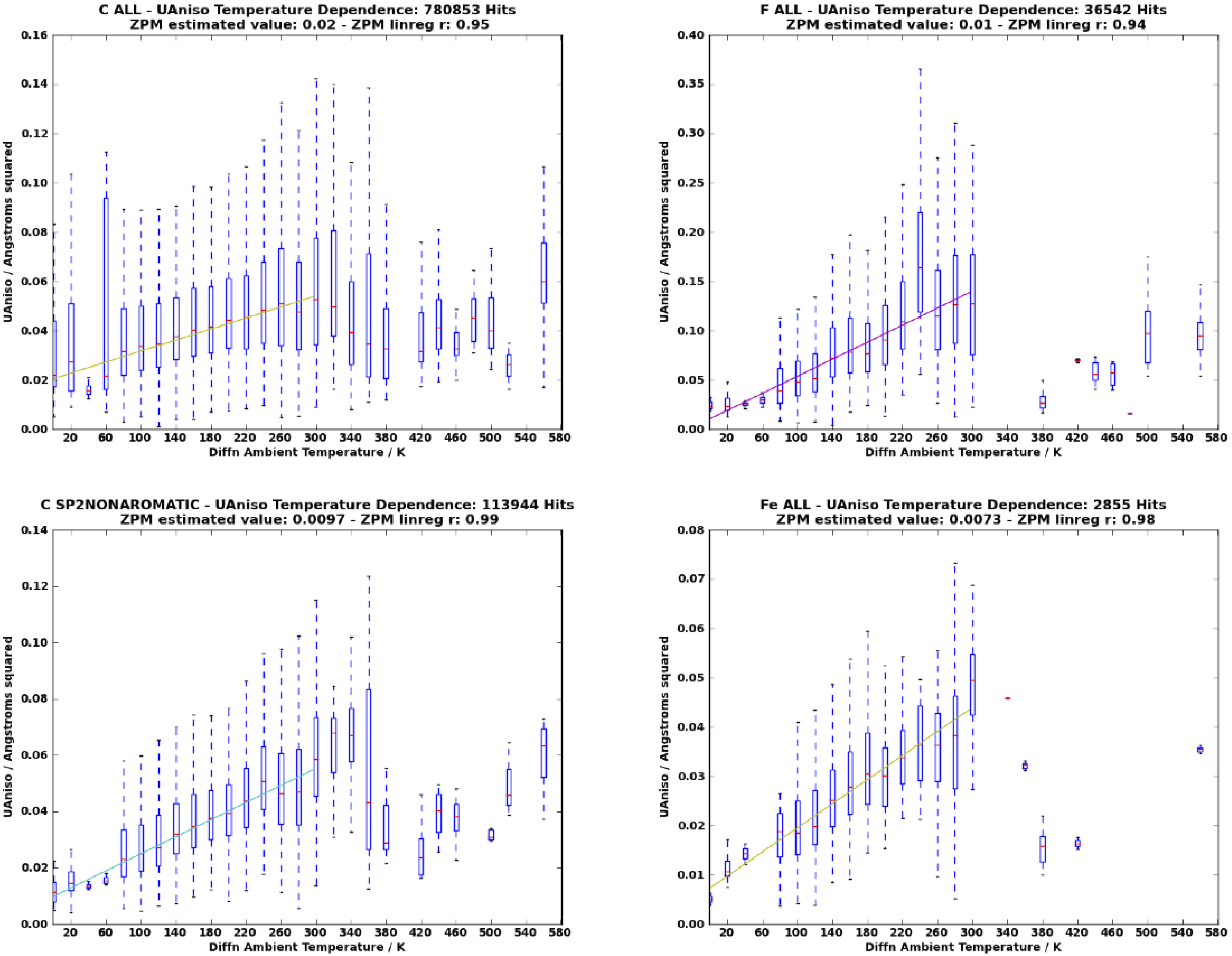
Figure 2.9: Dependence of U_{aniso} on bond angle for Nickel, Copper and Iron. Lower values corresponding to restricted moveme at 90 degrees and a 180 degrees, with greater range of motion for tetrahedral complexes between 105 and 110 degrees. This b plot once again shows the interquartile range as the box around the red line indicating the median, while the whiskers mark t minimum and maximum value for that range of angles.

Temperature dependence

Temperature dependence of ADPs for several elements: the median ADP values for each temperature bin in this histogram exhibit linear dependence on temperature.

-Allows for estimate of zero point motion, the motion of an atom due to the Heisenberg Uncertainty Principle at 0K.

-Irregular data above 300K indicative of measurement difficulties at higher temperatures



C ALL - UAniso Temperature Dependence: 780853 Hits
ZPM estimated value: 0.02 - ZPM linreg r: 0.95

F ALL - UAniso Temperature Dependence: 36542 Hits
ZPM estimated value: 0.01 - ZPM linreg r: 0.94

C SP2NONAROMATIC - UAniso Temperature Dependence: 113944 Hits
ZPM estimated value: 0.0097 - ZPM linreg r: 0.99

Fe ALL - UAniso Temperature Dependence: 2855 Hits
ZPM estimated value: 0.0073 - ZPM linreg r: 0.98

Figure 2.10: Temperature Dependence of ADPs

Recap:

- Parsed Cifs
- Enriched Data
- Entered into Database
- Datamined 30,000 crystal structures, 300,000 atomic parameters.
- Some thermal properties trends confirmed/uncovered

CCP9 Flagship Project

- CCP9 develops quantum mechanical modelling codes for condensed matter systems.
- Using the largest supercomputers could use these codes to generate many millions (soon billions) of pieces of materials data each day.
- Data very varied: energies, structures, vibrational properties, optical spectra, excited states,.....
- Data is not restricted to known materials, or even stable structures.
- Ultimate aim of project - materials discovery.

Challenges

- Given the nature of the materials simulation community, there will be multiple databases each covering different combinations of materials and properties both theoretical and experimental.
- Will never get concensus in advance on format/contents of databases.
- The errors associated with the data in theoretical databases cannot usually be quantified except by comparison with experimental data or, possibly, predictions made with a more accurate approach.